
Benign overfitting in leaky ReLU networks with moderate input dimension

Kedar Karhadkar^{1*} Erin George^{1*} Michael Murray¹ Guido Montúfar^{1,2} Deanna Needell¹
{kedar, egeo, mmurray, montufar, deanna}@math.ucla.edu
¹UCLA ²Max Planck Institute for Mathematics in the Sciences
*Equal contribution

Abstract

The problem of benign overfitting asks whether it is possible for a model to perfectly fit noisy training data and still generalize well. We study benign overfitting in two-layer leaky ReLU networks trained with the hinge loss on a binary classification task. We consider input data that can be decomposed into the sum of a common signal and a random noise component, that lie on subspaces orthogonal to one another. We characterize conditions on the signal to noise ratio (SNR) of the model parameters giving rise to benign versus non-benign (or harmful) overfitting: in particular, if the SNR is high then benign overfitting occurs, conversely if the SNR is low then harmful overfitting occurs. We attribute both benign and non-benign overfitting to an approximate margin maximization property and show that leaky ReLU networks trained on hinge loss with gradient descent (GD) satisfy this property. In contrast to prior work we do not require the training data to be nearly orthogonal. Notably, for input dimension d and training sample size n , while results in prior work require $d = \Omega(n^2 \log n)$, here we require only $d = \Omega(n)$.

1 Introduction

Intuition from learning theory suggests that fitting noise during training reduces a model's performance on test data. However, it has been observed in some settings that machine learning models can interpolate noisy training data with only *nominal* cost to their generalization performance (Zhang et al., 2017; Belkin et al., 2018, 2019), a phenomenon referred to as *benign overfitting*. Establishing theory that can explain this phenomenon has attracted much interest in recent years and there is now a rich body of work on this topic particularly in the context of linear models. However, the study of benign overfitting in the context of non-linear models, in particular shallow ReLU or leaky ReLU networks, has additional technical challenges and subsequently is less well advanced.

Much of the effort in regard to theoretically characterizing benign overfitting focuses on showing, under an appropriate scaling of the dimension of the input domain d , size of the training sample n , number of corruptions k and number of model parameters p that a model can interpolate noisy training data while achieving an arbitrarily small generalization error. Such characterizations of benign overfitting position it as a *high dimensional phenomenon*¹: indeed, the decrease in generalization error is achieved by escaping to higher dimensions at some rate relative to the other aforementioned hyperparameters. However, for these mathematical results to be relevant for explaining benign overfitting as observed in practice, clearly the particular scaling of d with respect to n , k and p needs to reflect the ratios seen in practice. Although a number of works, which we discuss in Section 1.2, establish benign overfitting results for shallow neural networks, a key and significant limitation they share is the requirement that the input features of the training data are at least approximately

¹We provide a formal definition of benign overfitting as a high dimensional phenomenon in Appendix E.

orthogonal to one another. To study benign overfitting, these prior works typically assume the input features consist of a small, low-rank signal component plus an isotropic noise term. Therefore, for the near orthogonality property to hold with high probability it is required that the input dimension d scales as $d = \Omega(n^2 \log n)$ or higher. This assumption highly restricts the applicability of these results for explaining benign overfitting in practice.

In this work we assume only $d = \Omega(n)$ and establish both harmful and benign overfitting results for shallow leaky ReLU networks trained via gradient descent (GD) on the hinge loss. In particular, we consider n data point pairs $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$, where, for some vector $\mathbf{v} \in \mathbb{S}^{d-1}$ and scalar $\gamma \in [0, 1]$, the input features are drawn from a pair of Gaussian clusters $\mathbf{x}_i \sim \mathcal{N}(\pm \sqrt{\gamma} \mathbf{v}, \sqrt{1 - \gamma} \frac{1}{d} (\mathbf{I}_d - \mathbf{v} \mathbf{v}^T))$ and $y_i = \text{sign}(\langle \mathbb{E}[\mathbf{x}_i], \mathbf{v} \rangle)$. The training data is noisy in that k of the n points in the training sample have their output label flipped. We assume equal numbers of positive and negative points among clean and corrupt ones. We provide a full description of our setup and assumptions in Section 2. Our proof techniques are novel and identify a new condition allowing for the analysis of benign and harmful overfitting which we term *approximate margin maximization*, wherein the norm of the network parameters is upper bounded by a constant of the norm of the max-margin linear classifier.

1.1 Summary of contributions

Our key results are summarized as follows.

- In Theorem 3.1, we prove that a leaky ReLU network trained on linearly separable data with gradient descent and the hinge loss will attain zero training loss in finitely many iterations. Moreover, the network weight matrix $\mathbf{W}$ at convergence will be approximately max-margin in the sense that $\|\mathbf{W}\| = O\left(\frac{\|\mathbf{w}^*\|}{\alpha \sqrt{m}}\right)$, where α is the leaky parameter of the activation function, m is the width of the network, and $\mathbf{w}^*$ is the max-margin linear classifier. We apply this result to derive generalization bounds for the network on test data.
- In Theorem 3.2, we establish conditions under which benign overfitting occurs for leaky ReLU networks. If the input dimension d , number of training points n , number of corrupt points k , and signal strength γ satisfy $d = \Omega(n)$ and $\gamma = \Omega(\frac{1}{k})$, then the network will exhibit benign overfitting. We emphasize that existing works on benign overfitting require $d = \Omega(n^2 \log n)$ to ensure nearly orthogonal data.
- In Theorem 3.3, we find a complementary lower bound for the generalization error to show that, for gradient descent classifiers, the bound in Theorem 3.2 is tight up to a constant in the exponent that can depend on α .
- In Theorem 3.4, we find conditions under which non-benign overfitting occurs. If $d = \Omega(n)$ and $\gamma = O(\frac{1}{d})$, then the network will exhibit non-benign overfitting: in particular its generalization error will be at least $\frac{1}{8}$.

1.2 Related work

There is now a significant body of literature theoretically characterizing benign overfitting in the context of linear models, including linear regression (Bartlett et al., 2020; Muthukumar et al., 2020; Wu & Xu, 2020; Zou et al., 2021; Hastie et al., 2022; Koehler et al., 2021; Wang et al., 2021a; Chatterji & Long, 2022; Shamir, 2022), logistic regression (Chatterji & Long, 2021; Muthukumar et al., 2021; Wang et al., 2021b), max-margin classification with linear and random feature models (Montanari et al., 2023b,a; Mei & Montanari, 2022; Cao et al., 2021) and kernel regression (Liang & Rakhlin, 2020; Liang et al., 2020; Adlam & Pennington, 2020). However, the study of benign overfitting in non-linear models is more nascent.

Homogeneous networks trained with gradient descent and an exponentially tailed loss are known to converge in direction to a Karush-Kuhn-Tucker (KKT) point of the associated max-margin problem (Lyu & Li, 2020; Ji & Telgarsky, 2020)². This property has been widely used in prior works to prove benign overfitting results for shallow neural networks. Frei et al. (2022) consider a shallow, smooth leaky ReLU network trained with an exponentially tailed loss and assume the data is drawn from a mixture of well-separated sub-Gaussian distributions. A key result of this work is, given

²One also needs to assume initialization from a position with a low initial loss.

sufficient iterations of GD, that the network will interpolate noisy training data while also achieving minimax optimal generalization error up to constants in the exponents. Xu & Gu (2023) extend this result to more general activation functions, including ReLU, as well as relax the assumptions on the noise distribution to being centered with bounded logarithmic Sobolev constant, and finally also improve the convergence rate. George et al. (2023) also study ReLU as opposed to leaky ReLU networks but do so in the context of the hinge loss, for which, and unlike exponentially tailed losses, a characterization of the implicit bias is not known. This work also establishes transitions on the margin of the clean data driving harmful, benign and no-overfitting training outcomes. Frei et al. (2023) use the aforementioned implicit bias of GD for linear classifiers and shallow leaky ReLU networks towards solutions that satisfy the KKT conditions of the margin maximization problem to establish settings where the satisfaction of said KKT conditions implies benign overfitting. Kornowski et al. (2023) also use the implicit bias results for exponentially tailed losses to derive similar benign overfitting results for shallow ReLU networks. Cao et al. (2022); Kou et al. (2023) study benign overfitting in two-layer convolutional as opposed to feedforward neural networks: indeed, whereas in most prior works data is modeled as the sum of a signal and noise component, in these two works the signal and noise components are assumed to lie in disjoint patches. The weight vector of each neuron is applied to both patches separately and a non-linearity, such as ReLU, is applied to the resulting pre-activation. In this setting, the authors prove interpolation of the noisy training data and derive conditions on the clean margin under which the network benignly versus harmfully overfits. A follow up work (Chen et al., 2023) considers the impact of Sharpness Aware Minimization (SAM) in the same setting. Finally, and assuming $d = \Omega(n^5)$, Xu et al. (2024) establish benign overfitting results for a data distribution which, instead of being linearly separable, is separated according to an XOR function.

We emphasize that the prior work on benign overfitting in the context of shallow neural networks requires the input data to be approximately orthogonal. Under standard data models studied this equates to the requirement that the input dimension d versus the size of the training sample n satisfies $d = \Omega(n^2 \log n)$ or higher. Here we require only $d = \Omega(n)$. The weaker dimensionality requirement requires substantially different proof techniques. George et al. (2023) study a setting most similar to the one studied here, however, the techniques are very different. In particular, the results presented in this other work are derived by carefully tracking neuron activation patterns. While in high dimensions this is feasible due to the near orthogonality of the noise in low dimensions this is far more challenging as noise vectors can be highly correlated leading to coupling effects.

Finally, we remark that our proof technique for the convergence of GD to a global minimizer in the context of a shallow leaky ReLU network (Theorem 3.1) is closely related to the proof techniques used by Brutzkus et al. (2018). While this work does establish a generalization bound, the bound assumes that population dataset is linearly separable rather than just the training dataset. Hence, it cannot be applied when the training dataset has label-flipping noise, which is the setting that we are interested in for benign overfitting.

2 Preliminaries

Let $[n] = \{1, 2, \dots, n\}$ denote the set of the first n natural numbers. We remark that when using big- O notation we implicitly assume only positive constants. We use $c, C, C_1, C_2, \dots$ to denote absolute constants with respect to the input dimension d , the training sample size n , and the width of the network m . Note constants may change in value from line to line. Furthermore, when using big- O notation all variables aside from d, n, k and m are considered constants. However, for clarity we will frequently make the constants concerning the confidence δ and failure probability ϵ explicit. Moreover, for two functions $f, g : \mathbb{N} \rightarrow \mathbb{N}$, if we say $f = O(g)$ implies property p , what we mean is there exists an $N \in \mathbb{N}$ and a constant C such that if $f(n) \leq Cg(n)$ for all $n \geq N$ then property p holds. Likewise, if we say $f = \Omega(g)$ implies property p , what we mean is there exists an $N \in \mathbb{N}$ and a constant C such that if $f(n) \geq Cg(n)$ for all $n \geq N$ then property p holds. Finally, we use $\|\cdot\|$ to denote the ℓ^2 norm of the vector argument or $\ell^2 \rightarrow \ell^2$ operator norm of the matrix argument.

2.1 Data model

We study data generated as per the following data model.

Definition 2.1. Suppose $d, n, k \in \mathbb{N}$, $\gamma \in (0, 1)$ and $\mathbf{v} \in \mathbb{S}^{d-1}$. If $(\mathbf{X}, \hat{\mathbf{y}}, \mathbf{y}, \mathbf{x}, y) \sim \mathcal{D}(d, n, k, \gamma, \mathbf{v})$ then

1. $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a random matrix whose rows, which we denote $\mathbf{x}_i$, satisfy $\mathbf{x}_i = \sqrt{\gamma} y_i \mathbf{v} + \sqrt{1 - \gamma} \mathbf{n}_i$, where $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}_d, \frac{1}{d}(\mathbf{I}_d - \mathbf{v} \mathbf{v}^T))$ are mutually i.i.d..
2. $\mathbf{y} \in \{\pm 1\}^n$ is a random vector with entries y_i that are mutually independent of one another as well as the noise vectors $(\mathbf{n}_i)_{i \in [n]}$ and are uniformly distributed over $\{\pm 1\}$. This vector holds the true labels of the training set.
3. Let $\mathcal{B} \subset [n]$ be any subset chosen independently of $\mathbf{y}$ such that $|\mathcal{B}| = k$. Then $\hat{\mathbf{y}} \in \{\pm 1\}^n$ is a random vector whose entries satisfy $\hat{y}_i \neq y_i$ for all $i \in \mathcal{B}$ and $\hat{y}_i = y_i$ for all $i \in \mathcal{B}^c =: \mathcal{G}$. This vector holds the observed labels of the training set.
4. y is a random variable representing a test label which is uniformly distributed over $\{\pm 1\}$.
5. $\mathbf{x} \in \mathbb{R}^d$ is a random vector representing the input feature of a test point and satisfies $\mathbf{x} = \sqrt{\gamma} y \mathbf{v} + \sqrt{1 - \gamma} \mathbf{n}$, where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_d, \frac{1}{d}(\mathbf{I}_d - \mathbf{v} \mathbf{v}^T))$ is mutually independent of the random vectors $(\mathbf{n}_i)_{i \in [n]}$.

We refer to $(\mathbf{X}, \hat{\mathbf{y}})$ as the training data and $(\mathbf{x}, y)$ as the test data. Furthermore, for typographical convenience we define $\mathbf{y} \odot \hat{\mathbf{y}} =: \beta \in \{\pm 1\}^n$.

To provide some interpretation to Definition 2.1, the training data consists of n points of which k have their observed label flipped relative to the true label. We refer to $\mathbf{v}$ and $\mathbf{n}_i$ as the signal and noise components of the i -th data point respectively: indeed, with $\gamma > 0$ then for $i \in \mathcal{G}$ $y_i \langle \mathbf{x}_i, \mathbf{v} \rangle = \sqrt{\gamma} > 0$. The test data is drawn from the same distribution and is assumed not to be corrupted. We say that the training data $(\mathbf{X}, \hat{\mathbf{y}})$ is *linearly separable* if there exists $\mathbf{w} \in \mathbb{R}^d$ such that

$$\hat{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1, \quad \text{for all } i \in [n].$$

For finite n , this condition is equivalent to the existence of a $\mathbf{w}$ with $\hat{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0$ for all $i \in [n]$. We denote the set of linearly separable datasets as $\mathcal{X}_{lin} \subset \mathbb{R}^{n \times d} \times \{\pm 1\}^n$. For a linearly separable dataset $(\mathbf{X}, \hat{\mathbf{y}})$, the *max-margin linear classifier* is the unique solution to the optimization problem

$$\arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{w}\| \text{ such that } \hat{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \text{ for all } i \in [n].$$

Observe one may equivalently take a strictly convex objective $\|\mathbf{w}\|^2$ and the constraint set is a closed convex polyhedron that is non-empty iff the data is linearly separable. The max-margin linear classifier $\mathbf{w}^*$ has a corresponding geometric margin $2/\|\mathbf{w}^*\|$. When $d \geq n$ and $\gamma > 0$, input feature matrices $\mathbf{X}$ from our data model almost surely have linearly independent rows $\mathbf{x}_i$ and thus $(\mathbf{X}, \hat{\mathbf{y}})$ is almost surely linearly separable for any observed labels $\hat{\mathbf{y}} \in \{\pm 1\}^n$.

2.2 Architecture and learning algorithm

We study shallow leaky ReLU networks with a forward pass function $f : \mathbb{R}^{2m \times d} \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined as

$$f(\mathbf{W}, \mathbf{x}) = \sum_{j=1}^{2m} (-1)^j \sigma(\langle \mathbf{w}_j, \mathbf{x} \rangle), \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{2m \times d}$ are the parameters of the network, $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is the leaky ReLU function, defined as $\sigma(x) = \max(x, \alpha x)$, where $\alpha \in (0, 1]$ is referred to as the leaky parameter. We remark that we only train the weights of the first layer and keep the output weights of each neuron fixed. Although σ is not differentiable at 0, in the context of gradient descent we adopt a subgradient and let $\dot{\sigma}(z) = 1$ for $z \geq 0$ and let $\dot{\sigma}(z) = \alpha$ otherwise. The *hinge loss* $\ell : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is defined as

$$\ell(z) = \max\{0, 1 - z\}. \quad (2)$$

Again, ℓ is not differentiable at zero; adopting a subgradient we define for any $j \in [2m]$

$$\nabla_{\mathbf{w}_j} \ell(\hat{y}_j f(\mathbf{W}, \mathbf{x})) = \begin{cases} (-1)^{j+1} \hat{y}_j \mathbf{x} \dot{\sigma}(\langle \mathbf{w}_j, \mathbf{x} \rangle) & \hat{y}_j f(\mathbf{W}, \mathbf{x}) < 1, \\ 0 & \hat{y}_j f(\mathbf{W}, \mathbf{x}) \geq 1. \end{cases}$$

The training loss $L : \mathbb{R}^{2m \times d} \times \mathbb{R}^{n \times d} \times \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as

$$L(\mathbf{W}, \mathbf{X}, \hat{\mathbf{y}}) = \sum_{i=1}^n \ell(\hat{y}_i f(\mathbf{W}, \mathbf{x}_i)). \quad (3)$$

Let $\mathbf{W}^{(0)} \in \mathbb{R}^{2m \times d}$ denote the model parameters at initialization. For each $t \in \mathbb{N}$ we define $\mathbf{W}^{(t)}$ recursively as

$$\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)} - \eta \nabla_{\mathbf{W}} L(\mathbf{W}^{(t-1)}, \mathbf{X}, \hat{\mathbf{y}}),$$

where $\eta > 0$ is the step size. Let $\mathcal{F}^{(t)} \subseteq [n]$ denote the set of all $i \in [n]$ such that $\hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) < 1$. Then equivalently each neuron is updated according to the following rule: for $j \in [2m]$

$$\mathbf{w}_j^{(t)} = GD(\mathbf{W}^{(t-1)}, \eta) := \mathbf{w}_j^{(t-1)} + \eta(-1)^j \sum_{i \in \mathcal{F}^{(t-1)}} \hat{y}_i \mathbf{x}_i \dot{\sigma}(\langle \mathbf{w}_j^{(t-1)}, \mathbf{x}_i \rangle). \quad (4)$$

For ease of reference we now provide the following definition of the learning algorithm described above.

Definition 2.2. Let $\mathcal{A}_{GD} : \mathbb{R}^{n \times d} \times \{\pm 1\}^n \times \mathbb{R} \times \mathbb{R}^{2m \times d} \rightarrow \mathbb{R}^{2m \times d}$ return $\mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}}, \eta, \mathbf{W}^{(0)}) =: \mathbf{W}$, where the j -th row $\mathbf{w}_j$ of $\mathbf{W}$ is defined as follows: let $\mathbf{w}_j^{(0)}$ be the j -th row of $\mathbf{W}^{(0)}$ and generate the sequence $(\mathbf{w}_j^{(t)})_{t \geq 0}$ using the recurrence relation $\mathbf{w}_j^{(t)} = GD(\mathbf{W}^{(t-1)}, \eta)$ as defined in equation 4.

1. If for $j \in [2m]$, $\lim_{t \rightarrow \infty} \mathbf{w}_j^{(t)}$ does not exist then we say $\mathcal{A}_{GD}$ is undefined.
2. Otherwise we say $\mathcal{A}_{GD}$ converges and $\mathbf{w}_j = \lim_{t \rightarrow \infty} \mathbf{w}_j^{(t)}$.
3. If there exists a $T \in \mathbb{N}$ such that for all $j \in [2m]$ $\mathbf{w}_j^{(t)} = \mathbf{w}_j^{(T)}$ for all $t \geq T$, then we say $\mathcal{A}_{GD}$ converges in finite time.

We often find that all matrices in the set

$$\{\mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}}, \eta, \mathbf{W}^{(0)}) : \forall j \|\mathbf{w}_j^{(0)}\| \leq \lambda\}$$

agree on all relevant properties. In this case, we abuse notation and say that $\mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}}, \eta, \lambda) = \mathbf{W}$ where $\mathbf{W}$ is a generic element from this set.

Finally, in order to derive our results we make the following assumptions concerning the step size and initialization of the network.

Assumption 1. The step size η satisfies $\eta \leq 1/(mn \max_{i \in [n]} \|\mathbf{x}_i\|^2)$ and for all $j \in [2m]$ the network at initialization satisfies $\|\mathbf{w}_j^{(0)}\| \leq \sqrt{\alpha}/(m \min_{i \in [n]} \|\mathbf{x}_i\|)$.

Under our data model the input data points have approximately unit norm; therefore these assumptions reduce to $\eta \leq \frac{C}{mn}$ and $\|\mathbf{w}_j^{(0)}\| \leq \frac{C\sqrt{\alpha}}{m}$.

2.3 Approximate margin maximization

We now introduce the notion of an approximate margin maximizing algorithm, which plays a key role in deriving our results. Although the primary setting we consider in this work is the learning algorithm $\mathcal{A}_{GD}$ (see Definition 2.2), we derive benign overfitting guarantees more broadly for any learning algorithm which fits into this category. Recall $\mathcal{X}_{lin}$ denotes the set of linearly separable datasets $(\mathbf{X}, \hat{\mathbf{y}}) \in \mathbb{R}^{n \times d} \times \{\pm 1\}^n$.

Definition 2.3. Let $f : \mathbb{R}^p \times \mathbb{R}^d \rightarrow \mathbb{R}$ denote a predictor function with p parameters. An algorithm $\mathcal{A} : \mathbb{R}^{n \times d} \times \mathbb{R}^n \rightarrow \mathbb{R}^p$ is approximately margin maximizing with factor $M > 0$ on f if for all $(\mathbf{X}, \hat{\mathbf{y}}) \in \mathcal{X}_{lin}$

$$\hat{y}_i f(\mathcal{A}(\mathbf{X}, \hat{\mathbf{y}}), \mathbf{x}_i) \geq 1 \text{ for all } i \in [n] \quad (5)$$

and

$$\|\mathcal{A}(\mathbf{X}, \hat{\mathbf{y}})\| \leq M \|\mathbf{w}^*\|, \quad (6)$$

where $\mathbf{w}^*$ is the max-margin linear classifier of $(\mathbf{X}, \hat{\mathbf{y}})$. Moreover, if $\mathcal{A}$ is an approximate margin maximizing algorithm we define

$$|\mathcal{A}| = \inf\{M > 0 : \mathcal{A} \text{ is approximately margin maximizing with factor } M\}. \quad (7)$$

In the above definition we take the standard Euclidean norm on $\mathbb{R}^p$. In particular if $\mathbb{R}^p = \mathbb{R}^{2m \times d}$ is a space of matrices we take the Frobenius norm.

3 Main results

In order to prove benign overfitting it is necessary to show that the learning algorithm outputs a model that correctly classifies all points in the training sample. The following theorem establishes this for $\mathcal{A}_{GD}$ and bounds the margin maximizing factor $|\mathcal{A}_{GD}|$.

Theorem 3.1. *Let $f : \mathbb{R}^p \times \mathbb{R}^n \rightarrow \mathbb{R}$ be a leaky ReLU network with forward pass as defined by equation 1. Suppose the step size η and initialization condition λ satisfy Assumption 1. Then for any linearly separable data set $(\mathbf{X}, \hat{\mathbf{y}})$ $\mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}}, \eta, \lambda)$ converges after T iterations, where*

$$T \leq \frac{C \|\mathbf{w}^*\|^2}{\eta \alpha^2 m}.$$

Furthermore $\mathcal{A}_{GD}$ is approximately margin maximizing on f (Definition 2.3) with

$$|\mathcal{A}_{GD}| \leq \frac{C}{\alpha \sqrt{m}}.$$

A proof of Theorem 3.1 can be found in Appendix D.1. Note also by Definition 2.3 that the solution $\mathbf{W} = \mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}})$ for $(\mathbf{X}, \hat{\mathbf{y}}) \in \mathcal{X}_{lin}$ is a global minimizer of the training loss defined in equation 3 with $L(\mathbf{W}, \mathbf{X}, \hat{\mathbf{y}}) = 0$. Our approach to proving this result is reminiscent of the proof of convergence of the perceptron algorithm and therefore is also similar to the techniques used by Brutzkus et al. (2018).

For training and test data as per Definition 2.1 we provide an upper bound on the generalization error for approximately margin maximizing algorithms. For convenience we summarize our setting as follows.

Assumption 2. *Setting for proving generalization results.*

- $f : \mathbb{R}^{2m \times d} \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a shallow leaky ReLU network as per equation 1.
- $\mathcal{A} : \mathbb{R}^{n \times d} \times \{\pm 1\}^n \rightarrow \mathbb{R}^{2m \times d}$ is a learning algorithm that returns the weights $\mathbf{W} \in \mathbb{R}^{2m \times d}$ of the first layer of f .
- We let $\mathbf{v} \in \mathbb{S}^{d-1}$ and consider training data $(\mathbf{X}, \hat{\mathbf{y}})$ and test data $(\mathbf{x}, y)$ distributed according to $(\mathbf{X}, \hat{\mathbf{y}}, \mathbf{y}, \mathbf{x}, y) \sim \mathcal{D}(d, n, k, \gamma, \mathbf{v})$ as per Definition 2.1.

Under this setting we have the following generalization result for an approximately margin maximizing algorithm $\mathcal{A}$. Note this result requires γ , and hence the signal to noise ratio of the inputs, to be sufficiently large.

Theorem 3.2. *Under the setting given in Assumption 2, let $\delta \in (0, 1)$ and suppose $\mathcal{A}$ is approximately margin-maximizing (Definition 2.3). If $n = \Omega(\log \frac{1}{\delta})$, $d = \Omega(n)$, $k = O(\frac{n}{1+m|\mathcal{A}|^2})$, and $\gamma = \Omega(\frac{1}{k})$ then there is a fixed positive constant C such that with probability at least $1 - \delta$ over $(\mathbf{X}, \hat{\mathbf{y}})$*

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0 \mid \mathbf{X}, \hat{\mathbf{y}}) \leq \exp\left(-C \cdot \frac{d}{k(1+m|\mathcal{A}|^2)}\right).$$

A proof of Theorem 3.2 is provided in Appendix D.2. To comment informally on the relationship between k and γ , we require $\gamma = \Omega(k^{-1})$ in order to guarantee that any network which achieves zero hinge loss does so by focusing on the signal component $\mathbf{v}$ rather than the noise components $\mathbf{n}_i$. We use the projection of the model weights onto the signal subspace as a measure of the strength of the signal the model has learned and derive our generalization results based on this measure. In Section 4 we provide a proof sketch of this framework in the simpler, linear model setting. Combining Theorems 3.1, and 3.2 we arrive at the following benign overfitting result for shallow leaky ReLU networks trained with GD on hinge loss.

Corollary 3.2.1. *Under the setting given in Assumption 2, let $\delta \in (0, 1)$ and suppose $\mathcal{A} = \mathcal{A}_{GD}$ where $\eta, \lambda \in \mathbb{R}_{>0}$ satisfy Assumption 1. If $n = \Omega(\log \frac{1}{\delta})$, $d = \Omega(n)$, $k = O(\alpha^2 n)$, and $\gamma = \Omega(\frac{1}{k})$ then the following hold.*

1. The algorithm $\mathcal{A}_{GD}$ terminates almost surely after a finite number of updates. If $\mathbf{W} = \mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}})$, then $L(\mathbf{W}, \mathbf{X}, \hat{\mathbf{y}}) = 0$.
2. There is a fixed positive constant C such that, with probability at least $1 - \delta$ over the training data $(\mathbf{X}, \hat{\mathbf{y}})$,

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0 \mid \mathbf{X}, \hat{\mathbf{y}}) \leq \exp\left(-C \cdot \frac{\alpha^2 d}{k}\right).$$

We remark that the upper bound is at most $\exp(-Cd/n)$ for a different constant C as we assume $k = O(\alpha^2 n)$.

If k is large enough, this bound is tight up to constants and factors of α in the exponent. This is given by the following theorem, proven in Appendix D.2.

Theorem 3.3. *Under the setting given in Assumption 2, let $\delta \in (0, 1)$ and suppose $\mathcal{A} = \mathcal{A}_{GD}$ where $\eta, \lambda \in \mathbb{R}_{>0}$ satisfy Assumption 1. If $n = \Omega(k)$, $d = \Omega(n)$, and $k = \Omega(\log \frac{1}{\delta} + \frac{1}{\alpha})$, then there is a fixed positive constant C such that with probability at least $1 - \delta$ over $(\mathbf{X}, \hat{\mathbf{y}})$*

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0 \mid \mathbf{X}, \hat{\mathbf{y}}) \geq \exp\left(-C \cdot \frac{d}{\alpha k}\right).$$

In addition to this benign overfitting result we also provide the following non-benign overfitting result for $\mathcal{A}_{GD}$. Note that conversely this result requires γ , and hence the signal to noise ratio of the inputs, to be sufficiently small.

Theorem 3.4. *Under the setting given in Assumption 2, let $\delta \in (0, 1)$ and suppose $\mathcal{A} = \mathcal{A}_{GD}$, where $\eta, \lambda \in \mathbb{R}_{>0}$ satisfy Assumption 1. If $n = \Omega(1)$, $d = \Omega(n + \log \frac{1}{\delta})$ and $\gamma = O\left(\frac{\alpha^3}{d}\right)$ then the following hold.*

1. The algorithm $\mathcal{A}_{GD}$ terminates almost surely after finitely many updates. With $\mathbf{W} = \mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}})$, $L(\mathbf{W}, \mathbf{X}, \hat{\mathbf{y}}) = 0$.
2. With probability at least $1 - \delta$ over the training data $(\mathbf{X}, \hat{\mathbf{y}})$

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) < 0 \mid \mathbf{X}, \hat{\mathbf{y}}) \geq \frac{1}{8}.$$

A proof of Theorem 3.4 is provided in Appendix D.3.

4 Approximate margin maximization and generalization: insight from linear models

In this section we outline proofs for the analogues of Theorems 3.2 and 3.4 in the context of linear models. The arguments are thematically similar and clearer to present. We provide complete proofs of benign and non-benign overfitting for linear models in Appendix C.

An important lemma is the following, which bounds the largest and n -th largest singular values (σ_1 and σ_n respectively) of the noise matrix $\mathbf{N}$:

Lemma 4.1. *Let $\mathbf{N} \in \mathbb{R}^{n \times d}$ denote a random matrix whose rows are drawn mutually i.i.d. from $\mathcal{N}(\mathbf{0}_d, \frac{1}{d}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$. If $d = \Omega(n + \log \frac{1}{\delta})$, then there exists constants C_1 and C_2 such that, with probability at least $1 - \delta$,*

$$C_1 \leq \sigma_n(\mathbf{N}) \leq \sigma_1(\mathbf{N}) \leq C_2.$$

We prove this lemma in Appendix B using results from Vershynin (2018) and Rudelson & Vershynin (2009). A consequence of this lemma is that with probability at least $1 - \delta$, the condition number of $\mathbf{N}$ restricted to $\text{span } \mathbf{N}$ can be bounded above independently of all hyperparameters. For this reason, we refer to the noise as being well-conditioned.

Now let $\mathbf{w} = \mathcal{A}(\mathbf{X}, \hat{\mathbf{y}})$ be the linear classifier returned by the algorithm. Observe that we can decompose the weight vector into a signal and noise component

$$\mathbf{w} = a_v \mathbf{v} + \mathbf{z},$$

where $\mathbf{z} \perp \mathbf{v}$ and $a_v \in \mathbb{R}$. Based on this decomposition the proof proceeds as follows.

1. Generalization bounds based on the SNR: For test data as per the data model given in Definition 2.1 we want to bound the probability of misclassification: in particular, we want to bound the probability that

$$X := y\langle \mathbf{w}, \mathbf{x} \rangle = \sqrt{\gamma}a_v + \sqrt{1-\gamma}\langle \mathbf{n}, \mathbf{z} \rangle \leq 0.$$

As the noise is normally distributed, $X \sim \mathcal{N}(\sqrt{\gamma}a_v, \frac{1-\gamma}{d}\|\mathbf{z}\|^2)$ and the desired upper bound therefore follows from Hoeffding's inequality,

$$\mathbb{P}(X \leq 0) \leq \exp\left(-\frac{\gamma da_v^2}{2(1-\gamma)\|\mathbf{z}\|^2}\right).$$

Using Gaussian anti-concentration, we also obtain a lower bound for the probability of misclassification:

$$\mathbb{P}(y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0) \geq \max\left\{\frac{1}{2} - \sqrt{\frac{d\gamma}{2\pi(1-\gamma)}} \frac{a_v}{\|\mathbf{z}\|}, \frac{1}{4} \exp\left(-\frac{6d}{\pi} \frac{\gamma}{1-\gamma} \frac{a_v^2}{\|\mathbf{z}\|^2}\right)\right\}.$$

2. Upper bound the norm of the max-margin classifier: In order to use the approximate max-margin property we require an upper bound on $\|\mathbf{w}^*\|$. As by definition $\|\mathbf{w}^*\| \leq \|\tilde{\mathbf{w}}\|$, it suffices to construct a vector $\tilde{\mathbf{w}}$ that interpolates the data and has small norm. Using that the noise matrix of the data is well-conditioned with high probability, we achieve this by strategically constructing the signal and noise components of $\tilde{\mathbf{w}}$. This yields the bound

$$\|\mathbf{w}^*\| \leq \|\tilde{\mathbf{w}}\| \leq C \min\left(\sqrt{\frac{n}{1-\gamma}}, \sqrt{\frac{1}{\gamma} + \frac{k}{1-\gamma}}\right),$$

where the arguments of the min function originate from a small and large γ regime respectively.

3. Lower bound the SNR using the approximate margin maximization property: Based on step 1 the key quantity of interest from a generalization perspective is the ratio $a_v/\|\mathbf{z}\|$, which describes the signal to noise ratio (SNR) of the learned classifier. To lower bound this quantity we first lower bound a_v . In particular, if a_v is small, then the only way to attain zero loss on the clean data is for $\|\mathbf{z}\|$ to be large. However, under appropriate assumptions on d, n, k and γ this can be shown to contradict the bound $\|\mathbf{z}\| \leq \|\mathbf{w}\| \leq |\mathcal{A}|\|\mathbf{w}^*\|$, and thus a_v must be bounded from below. A lower bound on $a_v/\|\mathbf{z}\|$ then follows by again using $\|\mathbf{z}\| \leq \|\mathbf{w}^*\|$. Hence we obtain a lower bound for the SNR and establish benign overfitting.

4. Upper bound the SNR using the zero loss condition: For the generalization lower bound, we compute an upper bound for the ratio $a_v/\|\mathbf{z}\|$ rather than a lower bound. Since the model perfectly fits the training data with margin one,

$$1 \leq \hat{y}_i\langle \mathbf{w}, \mathbf{x}_i \rangle = \sqrt{\gamma}\beta_i a_v + \sqrt{1-\gamma}\hat{y}_i\langle \mathbf{z}, \mathbf{n}_i \rangle$$

for all $i \in [n]$. The above inequality implies that $\sqrt{1-\gamma}\hat{y}_i\langle \mathbf{n}_i, \mathbf{z} \rangle$ is at least $\sqrt{\gamma}a_v$ for all corrupt points. Since the noise is well-conditioned, this gives a lower bound on $\|\mathbf{z}\|$ in terms of a_v and hence an upper bound on the SNR $a_v/\|\mathbf{z}\|$. By the second generalization lower bound in step 1, the generalization error is bounded below at a similar exponential rate to the upper bound.

5. Upper bound the SNR using the zero loss condition and maximum margin property: To prove non-benign overfitting, we again compute an upper bound for the ratio $a_v/\|\mathbf{z}\|$. We return to the zero loss condition:

$$1 \leq \hat{y}_i\langle \mathbf{w}, \mathbf{x}_i \rangle = \sqrt{\gamma}\beta_i a_v + \sqrt{1-\gamma}\hat{y}_i\langle \mathbf{z}, \mathbf{n}_i \rangle$$

for all $i \in [n]$. If γ is small and $|\sqrt{\gamma}a_v|$ is large, then $\|\mathbf{w}\|$ will be large, contradicting the approximate margin maximization. Hence the above inequality implies that $\sqrt{1-\gamma}\hat{y}_i\langle \mathbf{n}_i, \mathbf{z} \rangle$ is large for all $i \in [n]$. Since the noise is well-conditioned, this can only happen when $\|\mathbf{z}\|$ is large. This gives us a lower bound on $\|\mathbf{z}\|$. As before, we can also upper bound a_v by $\|\mathbf{w}\|$, giving us an upper bound on the SNR $a_v/\|\mathbf{z}\|$. By the first generalization lower bound in step 1, the classifier generalizes poorly and exhibits non-benign overfitting.

4.1 From linear models to leaky ReLU networks

The proof of benign and non-benign overfitting in the linear case uses the tension between the two properties of approximate margin maximization: fitting both the clean and corrupt points with margin versus the bound on the norm. To extend this idea to a shallow leaky ReLU network as per equation 1, we consider the same decomposition for each neuron $j \in [2m]$,

$$\mathbf{w}_j = a_j \mathbf{v} + \mathbf{z}_j,$$

where $a_j \in \mathbb{R}$ and $\mathbf{z}_j \perp \mathbf{v}$. In the linear case $\pm a_v$ can be interpreted as the activation of the linear classifier on $\pm \mathbf{v}$ respectively: in terms of magnitude the signal activation is the same in either case and thus we measure the alignment of the linear model with the signal using $|a_v|$. For leaky ReLU networks we define their activation on $\pm \mathbf{v}$ respectively as $A_1 = f(\mathbf{W}, \mathbf{v})$ and $A_{-1} = -f(\mathbf{W}, -\mathbf{v})$, and then define the alignment of the network as $A_{\min} = \min\{A_1, A_{-1}\}$. Considering the alignment of the network with the noise, then if $\mathbf{Z} \in \mathbb{R}^{2m \times d}$ denotes a matrix whose j -th row is $\mathbf{z}_j$, then we measure the alignment of the network using $\|\mathbf{Z}\|_F$. As a result, analogous to $a_v/\|\mathbf{z}\|$, the key ratio from a generalization perspective in the context of a leaky ReLU network is $A_{\min}/\|\mathbf{Z}\|_F$. The proof Theorems 3.2, 3.3, and 3.4 then follow the same outline as Steps 1-3 above but with additional non-trivial technicalities.

5 Conclusion

In this work we have proven conditions under which leaky ReLU networks trained on binary classification tasks exhibit benign and non-benign overfitting. We have substantially relaxed the necessary assumptions on the input data compared with prior work; instead of requiring nearly orthogonal data with $d = \Omega(n^2 \log n)$ or higher, we only need $d = \Omega(n)$. We achieve this by using the distribution of singular values of the noise rather than specific correlations between noise vectors. Our emphasis was on networks trained by gradient descent with the hinge loss, but we establish a new framework that is general enough to accommodate any algorithm that is approximately margin maximizing.

There are a few limitations of our results which would be natural questions to address in future work. While we improve upon existing results in our dependence on the input dimension of the data, we still require that the training dataset is linearly separable. This leaves open the question of whether an overparameterized network will perfectly fit the training data and generalize well for lower dimensional data, or satisfy a similar margin maximization condition. We also focus mainly on two-layer networks with fixed outer layer weights trained with the hinge loss. It would be interesting to investigate whether analogous results hold for deeper architectures or different loss functions and data models.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1928930 and by the Alfred P. Sloan Foundation under grant G-2021-16778, while the authors EG and DN were in residence at the Simons Laufer Mathematical Sciences Institute (formerly MSRI) in Berkeley, California, during the Fall 2023 semester. EG and DN were also partially supported by NSF DMS 2011140. EG was also supported by a NSF Graduate Research Fellowship under grant DGE 2034835. GM and KK were partly supported by NSF CAREER DMS 2145630 and DFG SPP 2298 Theoretical Foundations of Deep Learning grant 464109215. GM was also partly supported by NSF grant CCF 2212520, ERC Starting Grant 757983 (DLT), and BMBF in DAAD project 57616814 (SECAI).

References

- Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 74–84. PMLR, 2020. URL <https://proceedings.mlr.press/v119/adlam20a.html>.
- Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1907378117>.
- Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 541–549. PMLR, 2018. URL <https://proceedings.mlr.press/v80/belkin18a.html>.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. URL <https://doi.org/10.1073/pnas.190307011>.
- 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. URL <https://openreview.net/forum?id=rJ33wwxRb>.
- Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8407–8418. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/46e0eae7d5217c79c3ef6b4c212b8c6f-Paper.pdf.
- Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25237–25250. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a12c999be280372b157294e72a4bbc8b-Paper-Conference.pdf.
- Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021. URL <http://jmlr.org/papers/v22/20-974.html>.
- Niladri S. Chatterji and Philip M. Long. Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):1–12, 2022. URL <http://jmlr.org/papers/v23/21-1199.html>.
- Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than SGD? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=3WAnGWLpSQ>.
- Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2668–2703. PMLR, 2022. URL <https://proceedings.mlr.press/v178/frei22a.html>.
- Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from KKT conditions for margin maximization. In Gergely Neu and Lorenzo Rosasco (eds.), *The Thirty Sixth Annual Conference on Learning Theory, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3173–3228. PMLR, 2023. URL <https://proceedings.mlr.press/v195/frei23a.html>.

- Erin George, Michael Murray, William Swartworth, and Deanna Needell. Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign? In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 35139–35189. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6e73c39cc428c7d264d9820319f31e79-Paper-Conference.pdf.
- Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. URL <https://doi.org/10.1214/21-AOS2133>.
- Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17176–17186. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf.
- Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=Fy0hThdDBM>.
- Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From tempered to benign overfitting in ReLU neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=LnZuxp3Tx7>.
- Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer ReLU convolutional neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17615–17659. PMLR, 2023. URL <https://proceedings.mlr.press/v202/kou23a.html>.
- Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel “Ridgeless” regression can generalize. *The Annals of Statistics*, 48(3):1329 – 1347, 2020. URL <https://doi.org/10.1214/19-AOS1849>.
- Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2683–2711. PMLR, 2020. URL <https://proceedings.mlr.press/v125/liang20a.html>.
- Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SJeLIgBKPS>.
- Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22008>.
- Andrea Montanari, Feng Ruan, Basil Saeed, and Youngtak Sohn. Universality of max-margin classifiers. *arXiv preprint arXiv:2310.00176*, 2023a.
- Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2023b.
- Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 67–83, 2020. URL <https://doi.org/10.1109/JSAIT.2020.2984716>.

- Vidya Muthukumar, Adhyayan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *J. Mach. Learn. Res.*, 22(1), 2021. URL <http://jmlr.org/papers/v22/20-603.html>.
- George Pólya. Remarks on computing the probability integral in one and two dimensions. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, number 1, pp. 63. University of California Press Berkeley, 1949.
- Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009. URL <https://doi.org/10.1002/cpa.20294>.
- Ohad Shamir. The implicit bias of benign overfitting. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 448–478. PMLR, 2022. URL <https://proceedings.mlr.press/v178/shamir22a.html>.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. URL <https://doi.org/10.1017/9781108231596>.
- Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum ℓ_1 -norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, 2021a. URL <https://proceedings.mlr.press/v151/wang22k.html>.
- Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24164–24179. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/caaa29eab72b231b0af62fbddf89bfce-Paper.pdf.
- Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10112–10123. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/72e6d3238361fe70f22fb0ac624a7072-Paper.pdf.
- Xingyu Xu and Yuantao Gu. Benign overfitting of non-smooth neural networks beyond lazy training. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11094–11117. PMLR, 2023. URL <https://proceedings.mlr.press/v206/xu23k.html>.
- Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. Benign overfitting and grokking in ReLU networks for XOR cluster data. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=BxHgpC6FNv>.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Sy8gdB9xx>.
- Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsizes SGD for linear regression. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4633–4635. PMLR, 2021. URL <https://proceedings.mlr.press/v134/zou21a.html>.

Appendix A Preliminaries on random vectors

Recall that the sub-exponential norm of a random variable X is defined as

$$\|X\|_{\psi_1} := \inf\{t > 0: \mathbb{E}[\exp(|X|/t)] \leq 2\}$$

(see Vershynin, 2018, Definition 2.7.5) and that the sub-Gaussian norm is defined as

$$\|X\|_{\psi_2} := \inf\{t > 0: \mathbb{E}[\exp(X^2/t^2)] \leq 2\}.$$

A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Furthermore, $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$.

Lemma A.1. *Let $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$ and suppose that $\mathbf{Z} \in \mathbb{R}^{m \times d}$. Then with probability at least $1 - \epsilon$,*

$$\|\mathbf{Z}\mathbf{n}\| \leq C\|\mathbf{Z}\|_F \sqrt{\frac{1}{d} \log \frac{1}{\epsilon}}.$$

Proof. Let $\mathbf{P} = \mathbf{I}_d - \mathbf{v}\mathbf{v}^T$ be the orthogonal projection onto $\text{span}(\{\mathbf{v}\})^\perp$, so that $\mathbf{Z}\mathbf{n}$ is identically distributed to $d^{-1/2}\mathbf{Z}\mathbf{P}\mathbf{n}'$, where $\mathbf{n}'$ has distribution $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. Following Vershynin (2018, Theorem 6.3.2),

$$\begin{aligned} \left\| \|\mathbf{Z}\mathbf{n}\| - \|d^{-1/2}\mathbf{Z}\mathbf{P}\|_F \right\|_{\psi_2} &= \left\| \|d^{-1/2}\mathbf{Z}\mathbf{P}\mathbf{n}'\| - \|d^{-1/2}\mathbf{Z}\mathbf{P}\|_F \right\|_{\psi_2} \\ &\leq C\|d^{-1/2}\mathbf{Z}\mathbf{P}\| \\ &\leq Cd^{-1/2}\|\mathbf{Z}\|\|\mathbf{P}\| \\ &= Cd^{-1/2}\|\mathbf{Z}\| \\ &\leq Cd^{-1/2}\|\mathbf{Z}\|_F, \end{aligned}$$

where we used that $\mathbf{P}$ is an orthogonal projection in the fourth line and that the operator norm is by bounded above by the Frobenius norm in the fifth line. As a result the sub-Gaussian norm of $\|\mathbf{Z}\mathbf{n}\|$ is bounded as

$$\begin{aligned} \left\| \|\mathbf{Z}\mathbf{n}\| \right\|_{\psi_2} &\leq \left\| \|\mathbf{Z}\mathbf{n}\| - \|d^{-1/2}\mathbf{Z}\mathbf{P}\|_F \right\|_{\psi_2} + \left\| \|d^{-1/2}\mathbf{Z}\mathbf{P}\|_F \right\|_{\psi_2} \\ &\leq Cd^{-1/2}(\|\mathbf{Z}\|_F + \|\mathbf{Z}\mathbf{P}\|_F) \\ &\leq Cd^{-1/2}\|\mathbf{Z}\|_F, \end{aligned}$$

where the last line follows from the calculation

$$\begin{aligned} \|\mathbf{Z}\mathbf{P}\|_F &= \|\mathbf{P}^T\mathbf{Z}^T\|_F \\ &= \|\mathbf{P}\mathbf{Z}^T\|_F \\ &\leq \|\mathbf{P}\|\|\mathbf{Z}^T\|_F \\ &= \|\mathbf{Z}^T\|_F \\ &= \|\mathbf{Z}\|_F. \end{aligned}$$

This implies a tail bound (see Vershynin, 2018, Proposition 2.5.2)

$$\mathbb{P}(\|\mathbf{Z}\mathbf{n}\| \geq t) \leq 2 \exp\left(-\frac{dt^2}{C\|\mathbf{Z}\|_F^2}\right), \quad \text{for all } t \geq 0.$$

Setting $t = C\|\mathbf{Z}\|_F \sqrt{\frac{1}{d} \log \frac{2}{\epsilon}}$, the result follows. $\square$

Lemma A.2. *Let $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$ and suppose $\mathbf{z} \in \text{span}(\{\mathbf{v}\})^\perp$. There exists a $C > 0$ such that with probability at least $1 - \delta$*

$$|\langle \mathbf{n}, \mathbf{z} \rangle| \leq C\|\mathbf{z}\| \sqrt{\frac{1}{d} \log \frac{1}{\delta}}.$$

Furthermore, there exists a $c > 0$ such that with probability at least $\frac{1}{2}$

$$|\langle \mathbf{n}, \mathbf{z} \rangle| \geq \frac{c\|\mathbf{z}\|}{\sqrt{d}}.$$

Proof. Let $X = \langle \mathbf{n}, \mathbf{z} \rangle$. Then X is Gaussian with variance

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}[\mathbf{z}^T \mathbf{n} \mathbf{n}^T \mathbf{z}] \\ &= \mathbf{z}^T \mathbb{E}[\mathbf{n} \mathbf{n}^T] \mathbf{z} \\ &= d^{-1} \mathbf{z}^T (\mathbf{I}_d - \mathbf{v} \mathbf{v}^T) \mathbf{z} \\ &= d^{-1} \mathbf{z}^T \mathbf{z} \\ &= d^{-1} \|\mathbf{z}\|^2.\end{aligned}$$

Note the third line above follows from the fact that $\mathbf{z} \in \text{span}(\{\mathbf{v}\})^\perp$. Therefore by Hoeffding's inequality, for all $t \geq 0$

$$\mathbb{P}(|\langle \mathbf{n}, \mathbf{z} \rangle| \geq t) \leq 2 \exp\left(-\frac{dt^2}{C\|\mathbf{z}\|^2}\right).$$

Setting $t = C\|\mathbf{z}\|^2 \sqrt{\frac{1}{d} \log \frac{2}{\delta}}$, we obtain

$$\mathbb{P}(|\langle \mathbf{n}, \mathbf{z} \rangle| \geq t) \leq \delta,$$

which establishes the first part of the result.

Since $d^{1/2}\|\mathbf{z}\|^{-1}X$ is a standard Gaussian, there exists a constant c such that

$$\mathbb{P}(|d^{1/2}\|\mathbf{z}\|^{-1}X| \geq c) \leq \frac{1}{2}.$$

Rearranging, we obtain

$$\mathbb{P}(|\langle \mathbf{n}, \mathbf{z} \rangle| \geq cd^{-1/2}\|\mathbf{z}\|) \leq \frac{1}{2}.$$

This establishes the second part of the result. □

Appendix B Upper bounding the norm of the max-margin classifier of the data

Here we establish key properties concerning the data model given in Definition 2.1, our main goal being to establish bounds on the norm of the max-margin classifier. To this end we first identify certain useful facts about rectangular Gaussian matrices. In what follows we index the singular values of any given matrix $\mathbf{A}$ in decreasing order as $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \dots$. Furthermore, we denote the i -th-row of a matrix $\mathbf{A}$ as $\mathbf{a}_i$.

Lemma B.1. *Let $\mathbf{G} \in \mathbb{R}^{n \times d}$ be a Gaussian matrix whose entries are mutually i.i.d. with distribution $\mathcal{N}(0, 1)$. If $d = \Omega\left(n + \log \frac{1}{\delta}\right)$, then with probability at least $1 - \delta$ the following inequalities are simultaneously true.*

1. $\sigma_1(\mathbf{G}) \leq C(\sqrt{d} + \sqrt{n})$,
2. $\sigma_n(\mathbf{G}) \geq c(\sqrt{d} - \sqrt{n})$.

Proof. We proceed by upper bounding the probability that each individual inequality does not hold.

1. To derive an upper bound on $\sigma_1(\mathbf{G})$ we use the following fact (see Vershynin, 2018, Theorem 4.4.5). For any $\epsilon > 0$,

$$\mathbb{P}(\sigma_1(\mathbf{G}) \geq C_1(\sqrt{n} + \sqrt{d} + \epsilon)) \leq 2 \exp(-\epsilon^2).$$

With $\epsilon = \sqrt{n} + \sqrt{d}$ and $d \geq \log \frac{4}{\delta}$ then

$$\begin{aligned}\mathbb{P}(\sigma_1(\mathbf{G}) \geq 2C_1(\sqrt{n} + \sqrt{d})) &\leq 2 \exp(-d) \\ &\leq \frac{\delta}{2}.\end{aligned}$$

2. To derive a lower bound on $\sigma_n(\mathbf{G})$ we use the following fact (see Rudelson & Vershynin, 2009, Theorem 1.1). There exist constants $C_1, C_2 > 0$ such that, for any $\epsilon > 0$,

$$\mathbb{P}(\sigma_n(\mathbf{G}) \leq \epsilon(\sqrt{d} - \sqrt{n-1})) \leq (C_1\epsilon)^{d-n+1} + e^{-C_2d}.$$

Let $\epsilon = \frac{1}{C_1 e}$ and let $d \geq 2n + \left(2 + \frac{1}{C_2}\right) \log \frac{4}{\delta}$. Then

$$\begin{aligned} \mathbb{P}(\sigma_n(\mathbf{A}) \leq \epsilon(\sqrt{d} - \sqrt{n})) &\leq \exp(-d/2) + \exp(-C_2d) \\ &\leq \frac{\delta}{4} + \frac{\delta}{4} \\ &= \frac{\delta}{2}. \end{aligned}$$

Hence both bounds hold simultaneously with probability at least $1 - \delta$. $\square$

The next lemma formulates lower and upper bounds on the smallest and largest singular values of a noise matrix under our data model.

Lemma 4.1. *Let $\mathbf{N} \in \mathbb{R}^{n \times d}$ denote a random matrix whose rows are drawn mutually i.i.d. from $\mathcal{N}(\mathbf{0}_d, \frac{1}{d}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$. If $d = \Omega\left(n + \log \frac{1}{\delta}\right)$, then there exists constants C_1 and C_2 such that, with probability at least $1 - \delta$,*

$$C_1 \leq \sigma_n(\mathbf{N}) \leq \sigma_1(\mathbf{N}) \leq C_2.$$

Proof. Let $\mathbb{H} = \text{span}(\{\mathbf{v}\})^\perp \cong \mathbb{R}^{d-1}$. Let $\mathbf{N}' : \mathbb{H} \rightarrow \mathbb{R}^n$ be a random matrix whose rows are drawn mutually i.i.d. from $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_{\mathbb{H}})$. Since $d = \Omega\left(n + \log \frac{1}{\delta}\right)$, with probability at least $1 - \delta$, we have

$$c(\sqrt{d} - \sqrt{n}) \leq \sigma_n(\mathbf{N}') \leq \sigma_1(\mathbf{N}') \leq C(\sqrt{d} + \sqrt{n}).$$

We denote the above event by ω . Let $\mathbf{J} : \mathbb{H} \rightarrow \mathbb{R}^d$ be the inclusion map and let $\mathbf{P} = \mathbf{I}_d - \mathbf{v}\mathbf{v}^T$. For any random vector $\mathbf{n}$ with distribution $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_{\mathbb{H}})$, $\mathbf{J}\mathbf{n}$ is a Gaussian random vector with covariance matrix $\mathbf{J}\mathbf{J}^T = \mathbf{P}$. Therefore, $d^{-1/2}\mathbf{N}'\mathbf{J}^T$ is a random matrix whose rows are drawn mutually i.i.d. from $\mathcal{N}(\mathbf{0}_d, d^{-1}\mathbf{P})$. That is, $d^{-1/2}\mathbf{N}'\mathbf{J}^T$ is identically distributed to $\mathbf{N}$. For the lower bound on the n -th largest singular value, if $d \geq \frac{4n}{c^2}$, then conditional on ω we have

$$\begin{aligned} \sigma_n(d^{-1/2}\mathbf{N}'\mathbf{J}^T) &= d^{-1/2}\sigma_{\min}(\mathbf{J}\mathbf{N}'^T) \\ &\geq d^{-1/2}\sigma_{\min}(\mathbf{J})\sigma_{\min}(\mathbf{N}'^T) \\ &= d^{-1/2}\sigma_{\min}(\mathbf{N}'^T) \\ &= d^{-1/2}\sigma_n(\mathbf{N}') \\ &\geq c - \sqrt{\frac{n}{d}} \\ &\geq \frac{c}{2}. \end{aligned}$$

Note here we define $\sigma_{\min}$ to be the smallest singular value of a matrix. In the first line we used $\mathbf{J}\mathbf{N}'^T$ is a linear map $\mathbb{R}^n \rightarrow \mathbb{R}^d$ and $d \geq n$, and in the third line we used the fact that $\mathbf{J}$ is an inclusion map. For the upper bound on the largest singular value, if $d \geq \frac{n}{C^2}$, then conditional on ω we have

$$\begin{aligned} \sigma_1(d^{-1/2}\mathbf{N}'\mathbf{J}^T) &= d^{-1/2}\sigma_1(\mathbf{J}\mathbf{N}'^T) \\ &\leq d^{-1/2}\sigma_1(\mathbf{N}'^T) \\ &= d^{-1/2}\sigma_1(\mathbf{N}') \\ &\leq C + \sqrt{\frac{n}{d}} \\ &\leq 2C. \end{aligned}$$

Note here that again we used the fact that $\mathbf{J}$ is an inclusion map in the first line. Therefore, if $d = \Omega\left(n + \log \frac{1}{\delta}\right)$

$$\begin{aligned} \mathbb{P}\left(\frac{c}{2} \leq \sigma_n(\mathbf{N}) \leq \sigma_1(\mathbf{N}) \leq 2C\right) &\geq \mathbb{P}(\omega) \\ &\geq 1 - \delta. \end{aligned}$$

□

The following lemma is useful for constructing vectors in the noise subspace with properties suitable for bounding the norm of the max-margin solution. We remark that the same approach could be used in the setting where the noise and signal are not orthogonal by considering the pseudo-inverse $([\mathbf{N}, \mathbf{v}]^T)^\dagger$ instead of $\mathbf{N}^\dagger$.

Lemma B.2. *Let $\mathcal{I} \subseteq [n]$ be an arbitrary subset such that $|\mathcal{I}| = \ell$. In the context of the data model given in Definition 2.1, assume $d = \Omega\left(n + \log \frac{1}{\delta}\right)$. Then there exists $\mathbf{z} \in \mathbb{R}^d$ such that with probability at least $1 - \delta$ the following hold simultaneously.*

1. $\hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle = 1$ for all $i \in \mathcal{I}$,
2. $\hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle = 0$ for all $i \notin \mathcal{I}$,
3. $\mathbf{z} \perp \mathbf{v}$,
4. $C_1 \leq \|\mathbf{z}\| \leq C_2$.

Proof. Recall that $\mathbf{N} \in \mathbb{R}^{n \times d}$ is a random matrix whose rows are selected i.i.d. from $\mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$. By Lemma 4.1, with probability at least $1 - \delta$ we have

$$c \leq \sigma_n(\mathbf{N}) \leq \sigma_1(\mathbf{N}) \leq C.$$

Conditioning on this event, we will construct a vector $\mathbf{z}$ which satisfies the desired properties. Let $\mathbf{w} \in \mathbb{R}^n$ satisfy $w_i = \hat{y}_i$ if $i \in \mathcal{I}$ and $w_i = 0$ otherwise. Let $\mathbf{z} = \mathbf{N}^\dagger \mathbf{w}$, where $\mathbf{N}^\dagger = \mathbf{N}^T(\mathbf{N}\mathbf{N}^T)^{-1}$ is the right pseudo-inverse of $\mathbf{N}$. Then $\mathbf{N}\mathbf{z} = \mathbf{w}$. In particular, for $i \in \mathcal{I}$, $\hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle = \hat{y}_i w_i = 1$, and for $i \notin \mathcal{I}$, $\hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle = \hat{y}_i w_i = 0$. This establishes properties 1 and 2. Since $\mathbf{N}^\dagger \mathbf{w}$ is in the span of the set $\{\mathbf{n}_i\}_{i \in [n]}$, it is orthogonal to $\mathbf{v}$. This establishes property 3. Finally, we can bound

$$\begin{aligned} \|\mathbf{z}\| &= \|\mathbf{N}^\dagger \mathbf{w}\| \\ &\leq \|\mathbf{N}^\dagger\| \|\mathbf{w}\| \\ &= \frac{\|\mathbf{w}\|}{\sigma_n(\mathbf{N})} \\ &\leq \frac{\|\mathbf{w}\|}{c} \\ &= \frac{\sqrt{\ell}}{c} \end{aligned}$$

and

$$\begin{aligned} \|\mathbf{z}\| &= \|\mathbf{N}^\dagger \mathbf{w}\| \\ &\geq \sigma_n(\mathbf{N}^\dagger) \|\mathbf{w}\| \\ &= \frac{\|\mathbf{w}\|}{\sigma_1(\mathbf{N})} \\ &\geq \frac{\|\mathbf{w}\|}{C} \\ &= \frac{\sqrt{\ell}}{C} \end{aligned}$$

which establishes property 4.

□

With Lemma B.2 in place we are now able to appropriately bound the max-margin norm.

Lemma B.3. *In the context of the data model given in Definition 2.1, let $\mathbf{w}^*$ denote the max-margin classifier of the training data $(\mathbf{X}, \hat{\mathbf{y}})$, which exists almost surely. If $d = \Omega(n + \log \frac{1}{\delta})$ then with probability at least $1 - \delta$*

$$\|\mathbf{w}^*\| \leq C \sqrt{\frac{1}{\gamma} + \frac{k}{1-\gamma}},$$

where $C > 0$ is a constant.

Proof. Under the assumptions stated, the conditions of Lemma B.2 hold with probability at least $1 - \delta$. Conditioning on this let

$$\mathbf{w} = \frac{1}{\sqrt{\gamma}} \mathbf{v} + \frac{2}{\sqrt{1-\gamma}} \mathbf{z}$$

where $\mathbf{z}$ is the vector constructed in Lemma B.2 with $\mathcal{I} = \mathcal{B}$. For any $i \in [n]$ we therefore have

$$\hat{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \beta_i + 2\hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle.$$

As a result, for $i \in \mathcal{G}$

$$\hat{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle = 1 + 2\hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle = 1,$$

while for $l \in \mathcal{B}$

$$\hat{y}_l \langle \mathbf{x}_l, \mathbf{w} \rangle = -1 + 2\hat{y}_l \langle \mathbf{n}_l, \mathbf{z} \rangle = 1.$$

As a result $\hat{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle = 1$ for all $i \in [n]$. Furthermore, observe

$$\begin{aligned} \|\mathbf{w}\|^2 &= \frac{1}{\gamma} + \frac{4}{1-\gamma} \|\mathbf{z}\|^2 \\ &\leq C \left(\frac{1}{\gamma} + \frac{k}{1-\gamma} \right) \end{aligned}$$

for a universal constant C . To conclude observe $\|\mathbf{w}^*\| \leq \|\mathbf{w}\|$ by definition of being max-margin. $\square$

Lemma B.3 constructs a classifier with margin one using an appropriate linear combination of the signal vector $\mathbf{v}$ and a vector in the noise subspace which classifies all noise components belonging to bad points correctly. This bound is useful for the benign overfitting setting in which γ is not too small. However, for small γ , as is the case in the non-benign overfitting setting, this bound behaves poorly as the only way the construction can fit all the data points is by making the coefficient in front of the $\mathbf{v}$ component large. For the non-benign overfitting setting we therefore require a different approach and instead fit all data points based on their noise components alone. In particular, the following bound behaves better than that given in Lemma B.3 when γ approaches 0.

Lemma B.4. *In the context of the data model given in Definition 2.1, let $\mathbf{w}^*$ denote the max-margin classifier of the training data $(\mathbf{X}, \hat{\mathbf{y}})$, which exists almost surely. If $d = \Omega(n + \log \frac{1}{\delta})$ then with probability at least $1 - \delta$*

$$\|\mathbf{w}^*\| \leq C \sqrt{\frac{n}{1-\gamma}}.$$

Proof. Applying Lemma B.2 with $\mathcal{I} = [n]$ then with probability $1 - \delta$ there exists $\mathbf{z} \in \mathbb{R}^d$ such that $\|\mathbf{z}\| = \Theta(\sqrt{n})$, $\hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle = 1$ for all $i \in [n]$ and $\mathbf{z} \perp \mathbf{v}$. Conditioning on this event, let

$$\mathbf{w} = \frac{1}{\sqrt{1-\gamma}} \mathbf{z}.$$

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

$$\hat{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle = 1.$$

Furthermore, there exists a constant $C > 0$ such that

$$\|\mathbf{w}\| \leq C \sqrt{\frac{n}{1-\gamma}}.$$

To conclude observe $\|\mathbf{w}^*\| \leq \|\mathbf{w}\|$ by definition of being max-margin. $\square$

Appendix C Linear models

C.1 Sufficient conditions for benign and harmful overfitting

We start by providing a lemma which characterizes the generalization properties of linear classifiers.

Lemma C.1. *In the context of the data model given in Definition 2.1, consider the linear classifier $\mathbf{w} = a_v \mathbf{v} + \mathbf{z}$, where $a_v \in \mathbb{R}$ and $\langle \mathbf{z}, \mathbf{v} \rangle = 0$. If $a_v \geq 0$, then the generalization error can be bounded as follows:*

$$\mathbb{P}(y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0) \leq \exp \left(-\frac{d}{2} \frac{\gamma}{1-\gamma} \frac{a_v^2}{\|\mathbf{z}\|^2} \right).$$

and

$$\mathbb{P}(y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0) \geq \max \left\{ \frac{1}{2} - \sqrt{\frac{d\gamma}{2\pi(1-\gamma)}} \frac{a_v}{\|\mathbf{z}\|}, \frac{1}{4} \exp \left(-\frac{6d}{\pi} \frac{\gamma}{1-\gamma} \frac{a_v^2}{\|\mathbf{z}\|^2} \right) \right\}.$$

Proof. Recall from Definition 2.1 that a test pair $(\mathbf{x}, y)$ satisfies

$$\mathbf{x} = y(\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n}),$$

where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$ is a random vector. Let $X = y\langle \mathbf{w}, \mathbf{x} \rangle$, so

$$\begin{aligned} X &= \langle a_v \mathbf{v} + \mathbf{z}, \sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n} \rangle \\ &= \sqrt{\gamma}a_v + \sqrt{1-\gamma}\langle \mathbf{n}, \mathbf{z} \rangle. \end{aligned}$$

Then X is a Gaussian random variable with expectation

$$\begin{aligned} \mathbb{E}[X] &= \sqrt{\gamma}a_v + \sqrt{1-\gamma}\mathbb{E}[\langle \mathbf{n}, \mathbf{z} \rangle] \\ &= \sqrt{\gamma}a_v \end{aligned}$$

and variance

$$\begin{aligned} \text{Var}(X) &= (1-\gamma)\text{Var}(\langle \mathbf{n}, \mathbf{z} \rangle) \\ &= (1-\gamma)\mathbb{E}[\mathbf{z}^T \mathbf{n} \mathbf{n}^T \mathbf{z}] \\ &= \frac{1-\gamma}{d} \mathbf{z}^T (\mathbf{I}_d - \mathbf{v}\mathbf{v}^T) \mathbf{z} \\ &= \frac{1-\gamma}{d} \mathbf{z}^T \mathbf{z} \\ &= \frac{(1-\gamma)\|\mathbf{z}\|^2}{d}. \end{aligned}$$

By Hoeffding's inequality, for all $t \geq 0$,

$$\mathbb{P}(X \leq \sqrt{\gamma}a_v - t) \leq \exp \left(-\frac{t^2 d}{2(1-\gamma)\|\mathbf{z}\|^2} \right).$$

Setting $t = \sqrt{\gamma}a_v$, we obtain

$$\mathbb{P}(X \leq 0) \leq \exp \left(-\frac{\gamma d a_v^2}{2(1-\gamma)\|\mathbf{z}\|^2} \right),$$

which establishes the upper bound on the generalization error.

To prove the lower bound, we integrate a standard Gaussian pdf:

$$\begin{aligned} \mathbb{P}(X \leq 0) &= \mathbb{P} \left(\frac{X - \sqrt{\gamma}a_v}{\sqrt{1-\gamma}\|\mathbf{z}\|/\sqrt{d}} \leq -\sqrt{\frac{\gamma d}{1-\gamma}} \frac{a_v}{\|\mathbf{z}\|} \right) \\ &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{\frac{\gamma d}{1-\gamma}} \frac{a_v}{\|\mathbf{z}\|}}^0 e^{-t^2/2} dt \\ &\geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \left(\sqrt{\frac{\gamma d}{1-\gamma}} \frac{a_v}{\|\mathbf{z}\|} \right). \end{aligned}$$

Another bound can be obtained using the following inequality (Pólya, 1949, (1.5)):

$$\frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt \leq \sqrt{1 - e^{-2x^2/\pi}}.$$

We proceed

$$\begin{aligned} \mathbb{P}(X \leq 0) &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{\frac{\gamma d}{1-\gamma}} \frac{a_v}{\|\mathbf{z}\|}}^0 e^{-t^2/2} dt \\ &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^{\sqrt{\frac{\gamma d}{1-\gamma}} \frac{a_v}{\|\mathbf{z}\|}} e^{-t^2/2} dt \\ &\geq \frac{1}{2} - \frac{1}{2} \sqrt{1 - \exp\left(-\frac{2\gamma d a_v^2}{\pi(1-\gamma)\|\mathbf{z}\|^2}\right)} \\ &\geq \frac{1}{2} - \frac{1}{2} \left(1 - \frac{1}{2} \exp\left(-\frac{2\gamma d a_v^2}{\pi(1-\gamma)\|\mathbf{z}\|^2}\right)\right) \\ &\geq \frac{1}{4} \exp\left(-\frac{2\gamma d a_v^2}{\pi(1-\gamma)\|\mathbf{z}\|^2}\right). \quad \square \end{aligned}$$

The following result establishes benign and non-benign overfitting for linear models and data as per Definition 2.1, with a phase transition between these outcomes depending on the signal to noise parameter γ .

Theorem C.2. *In the context of the data model described in Section 2, let $\mathbf{w}^*$ be a max-margin linear classifier of the training data. Let $\mathcal{A} : \mathbb{R}^{n \times d} \times \{\pm 1\}^n \rightarrow \mathbb{R}^d$ be a learning algorithm which is approximately margin-maximizing, Definition 2.3. For $\delta \in (0, 1]$, let $d = \Omega\left(n + \log \frac{1}{\delta}\right)$. Then with probability at least $1 - \delta$ over the randomness of the training data $(\mathbf{X}, \hat{\mathbf{y}})$, the following hold with ϵ denoting the generalization error of $\mathcal{A}(\mathbf{X}, \hat{\mathbf{y}})$.*

(A) *If $\gamma = \Omega(|\mathcal{A}|^2 n^{-1})$ and $k = O(|\mathcal{A}|^{-2} n)$, then $\exp(-C_1 d k^{-1}) \leq \epsilon \leq \exp(-C_2 d k^{-1} |\mathcal{A}|^{-2})$ for fixed positive constants C_1 and C_2 .*

(B) *If $\gamma = O(|\mathcal{A}|^{-2} d^{-1})$, then $\epsilon \geq \frac{1}{2} - \sqrt{C d \gamma |\mathcal{A}|^2}$ for a fixed positive constant C .*

Proof. For training data $(\mathbf{X}, \hat{\mathbf{y}})$ let $\mathbf{w} = \mathcal{A}(\mathbf{X}, \hat{\mathbf{y}})$ be the learned linear classifier. First, recall $\mathbf{x}_i = \sqrt{\gamma} y_i \mathbf{v} + \sqrt{1 - \gamma} \mathbf{n}_i$ for all $i \in [n]$, $\|\mathbf{v}\| = 1$, and observe that we can decompose the vector $\mathbf{w}$ as $\mathbf{w} = a_v \mathbf{v} + \mathbf{z}$, where $a_v \in \mathbb{R}$, and $\mathbf{z} \perp \mathbf{v}$. As a result, for each $i \in [n]$,

$$\hat{y}_i \langle \mathbf{x}_i, \mathbf{w} \rangle = \sqrt{\gamma} a_v \beta_i + \sqrt{1 - \gamma} \hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle. \quad (8)$$

First we establish (A). As $d = \Omega\left(n + \log \frac{1}{\delta}\right)$, Lemmas 4.1 and B.3 show that with probability at least $1 - \frac{\delta}{2}$, $\|\mathbf{N}\|^2, \|\mathbf{N}_{\mathcal{G}}\|^2, \|\mathbf{N}_{\mathcal{B}}\| \leq C$ and $\|\mathbf{w}^*\| \leq C \sqrt{\frac{1}{\gamma} + \frac{k}{1-\gamma}}$. Here $\mathbf{N}_{\mathcal{G}}$ and $\mathbf{N}_{\mathcal{B}}$ denote the matrices formed by taking only the rows of $\mathbf{N}$ which satisfy $\beta = 1$ and $\beta = -1$, respectively. We denote this event ω and condition on it in all that follows for the proof of (A). As $\mathcal{A}$ is approximately max margin then given equation 8 we have for all $i \in \mathcal{G}$

$$1 \leq \sqrt{\gamma} a_v + \sqrt{1 - \gamma} \hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle.$$

Suppose that $\sqrt{\gamma} a_v < \frac{1}{2}$. Then the above inequality implies $\sqrt{1 - \gamma} \hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle \geq \frac{1}{2}$ for all $i \in \mathcal{G}$. Squaring and then summing this expression over all $i \in \mathcal{G}$ it follows that

$$\begin{aligned} \frac{n - k}{4} &\leq (1 - \gamma) \sum_{i \in \mathcal{G}} |\langle \mathbf{n}_i, \mathbf{z} \rangle|^2 \\ &\leq (1 - \gamma) \|\mathbf{N}_{\mathcal{G}} \mathbf{z}\|^2 \\ &\leq (1 - \gamma) \|\mathbf{N}_{\mathcal{G}}\|^2 \|\mathbf{z}\|^2. \end{aligned}$$

Since $\mathcal{A}$ is approximately margin-maximizing,

$$\begin{aligned}\frac{n-k}{4} &\leq (1-\gamma)C\|\mathbf{z}\|^2 \\ &\leq (1-\gamma)C\|\mathbf{w}\|^2 \\ &\leq C|\mathcal{A}|^2(1-\gamma)\|\mathbf{w}^*\|^2 \\ &\leq C|\mathcal{A}|^2(1-\gamma)\left(\frac{1}{\gamma} + \frac{k}{1-\gamma}\right) \\ &\leq \frac{C|\mathcal{A}|^2}{\gamma} + C|\mathcal{A}|^2k,\end{aligned}$$

which further implies $n \leq \frac{C|\mathcal{A}|^2}{\gamma} + C|\mathcal{A}|^2k$ for some other constant C . For this inequality to hold, either $\frac{C|\mathcal{A}|^2}{\gamma} \geq \frac{n}{2}$ or $C|\mathcal{A}|^2k \geq \frac{n}{2}$. With $k \leq \frac{n}{4C|\mathcal{A}|^2}$ and $\gamma \geq \frac{1}{k}$, neither of these can be true and therefore we conclude that $\sqrt{\gamma}a_v \geq \frac{1}{2}$. Then

$$\begin{aligned}\frac{a_v^2}{\|\mathbf{z}\|^2} &\geq \frac{1}{4\gamma\|\mathbf{z}\|^2} \\ &\geq \frac{1}{4\gamma\|\mathbf{w}\|^2} \\ &\geq \frac{1}{4\gamma|\mathcal{A}|^2\|\mathbf{w}^*\|^2} \\ &\geq \frac{C}{|\mathcal{A}|^2\gamma} \frac{1}{\frac{1}{\gamma} + \frac{k}{1-\gamma}} \\ &\geq \frac{C}{2|\mathcal{A}|^2\gamma} \frac{1-\gamma}{k}\end{aligned}$$

for a positive constant C . Letting $(\mathbf{x}, y)$ denote a test point pair, then by Lemma C.1 it follows that for a different constant C

$$\begin{aligned}\mathbb{P}(y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0 \mid \omega) &\leq \exp\left(-\frac{d}{2} \frac{\gamma}{1-\gamma} \frac{a_v^2}{\|\mathbf{z}\|^2}\right) \\ &\leq \exp\left(-\frac{Cd}{|\mathcal{A}|^2k}\right).\end{aligned}$$

Hence the generalization error is at most ϵ when ω occurs, which happens with probability at least $1 - \delta$. This establishes the upper bound of (A).

For the lower bound of (A), since $\mathbf{w}$ is a linear classifier,

$$\langle \mathbf{n}_i, \mathbf{z} \rangle \geq \frac{1}{\sqrt{1-\gamma}}(1 + a_v\sqrt{\gamma}) \geq a_v\sqrt{\frac{\gamma}{1-\gamma}}$$

for all $i \in \mathcal{B}$, from which we conclude $|\langle \mathbf{n}_i, \mathbf{z} \rangle| \geq a_v\sqrt{\gamma}$. This implies

$$\|\mathbf{N}_{\mathcal{B}}\mathbf{z}\| \geq a_v\sqrt{\frac{k\gamma}{1-\gamma}}$$

Along with $\|\mathbf{N}_{\mathcal{B}}\mathbf{z}\| \leq \|\mathbf{N}_{\mathcal{B}}\|\|\mathbf{z}\| \leq C\|\mathbf{z}\|$ we conclude

$$\|\mathbf{z}\| \geq \frac{a_v}{C}\sqrt{\frac{k\gamma}{1-\gamma}}.$$

With this bound we then bound

$$\frac{a_v}{\|\mathbf{z}\|} \leq C \cdot \sqrt{\frac{1-\gamma}{k\gamma}}.$$

By Lemma C.1 we then can bound

$$\begin{aligned}\mathbb{P}(y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0 \mid \omega) &\geq \frac{1}{4} \exp\left(-\frac{6d}{\pi} \frac{\gamma}{1-\gamma} \frac{a_v^2}{\|\mathbf{z}\|^2}\right) \\ &\geq \exp\left(-\frac{Cd}{k}\right)\end{aligned}$$

for a new constant C , provided a_v is positive. In the last line we can bound $\frac{d}{k}$ below as $d \geq n$ and $k = O(n)$. If a_v is negative then the generalization error is at least $\frac{1}{2}$, which is still bounded below by $\exp(-Cd/k)$.

We now turn our attention to (B). As $d = \Omega(n + \log \frac{1}{\delta})$, from Lemmas 4.1 and B.4 with probability at least $1 - \delta$ it holds that $\|\mathbf{N}_{\mathcal{G}}\|^2 \leq C$ and $\|\mathbf{w}^*\| \leq \frac{C\sqrt{n}}{\sqrt{1-\gamma}}$. We denote this event ω' and condition on it in all that follows for the proof of (B). In particular,

$$\begin{aligned}a_v^2 + \|\mathbf{z}\|^2 &= \|\mathbf{w}\|^2 \\ &\leq |\mathcal{A}|^2 \|\mathbf{w}^*\|^2 \\ &\leq \frac{C|\mathcal{A}|^2 n}{1-\gamma}.\end{aligned}\tag{9}$$

For all $i \in [n]$,

$$1 \leq \sqrt{\gamma} \beta_i a_v + \sqrt{1-\gamma} \hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle.$$

For this inequality to hold, either $|\sqrt{\gamma} a_v| \geq 1/2$ or $\sqrt{1-\gamma} \hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle \geq 1/2$ for all $i \in [n]$. If $|\sqrt{\gamma} a_v| \geq 1/2$, then with $\gamma \leq \frac{1}{4(C+1)|\mathcal{A}|^2 d}$ we have

$$a_v^2 \geq 2C|\mathcal{A}|^2 d \geq 2C|\mathcal{A}|^2 n.$$

However, from equation 9 we have

$$2C|\mathcal{A}|^2 n > \frac{C|\mathcal{A}|^2 n}{1-\gamma} \geq a_v^2$$

which is a contradiction. Therefore, under the regime specified there exists a $C > 0$ such that with $\gamma \leq \frac{1}{C|\mathcal{A}|^2 d}$, we have $\sqrt{1-\gamma} \hat{y}_i \langle \mathbf{n}_i, \mathbf{z} \rangle \geq 1/2$ for all $i \in [n]$. Rearranging, squaring and summing over all $i \in [n]$ yields

$$\frac{n}{4(1-\gamma)} \leq \sum_{i=1}^n |\langle \mathbf{n}_i, \mathbf{z} \rangle|^2 \leq \|\mathbf{N}\mathbf{z}\|^2 \leq C\|\mathbf{z}\|^2,$$

where the final inequality follows from conditioning on ω . Then

$$\begin{aligned}\frac{a_v^2}{\|\mathbf{z}\|^2} &\leq \frac{\|\mathbf{w}\|^2}{\|\mathbf{z}\|^2} \\ &\leq \frac{|\mathcal{A}|^2 \|\mathbf{w}^*\|^2}{\|\mathbf{z}\|^2} \\ &\leq \frac{C|\mathcal{A}|^2 \frac{n}{1-\gamma}}{\frac{n}{1-\gamma}} \\ &\leq C|\mathcal{A}|^2,\end{aligned}$$

where the constant $C > 0$ may vary between inequalities. Letting $(\mathbf{x}, y)$ denote a test point pair, by Lemma C.1 it follows that

$$\begin{aligned}\mathbb{P}(y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0 \mid \omega) &\geq \frac{1}{2} - \sqrt{\frac{d\gamma}{2\pi(1-\gamma)}} \frac{a_v}{\|\mathbf{z}\|} \\ &\geq \frac{1}{2} - \sqrt{\frac{Cd\gamma|\mathcal{A}|^2}{1-\gamma}} \\ &\geq \frac{1}{2} - \sqrt{Cd\gamma|\mathcal{A}|^2}.\end{aligned}$$

Hence the generalization error is at least $\frac{1}{2} - \sqrt{Cd\gamma|\mathcal{A}|^2}$ when ω' occurs, which happens with probability at least $1 - \delta$. This establishes (B). $\square$

Appendix D Leaky ReLU Networks

In this section we consider a leaky ReLU network $f : \mathbb{R}^{2m} \times \mathbb{R}^d \rightarrow \mathbb{R}$ with forward pass given by

$$f(\mathbf{W}, \mathbf{x}) = \sum_{j=1}^{2m} (-1)^j \sigma(\langle \mathbf{w}_j, \mathbf{x} \rangle),$$

where $\sigma(z) = \max(\alpha z, z)$ for some $\alpha \in (0, 1)$. For any such network, we may decompose the neuron weights $\mathbf{w}_j$ into a signal component and a noise component,

$$\mathbf{w}_j = a_j \mathbf{v} + \mathbf{z}_j,$$

where $a_j \in \mathbb{R}$ and $\mathbf{z}_j \in \mathbb{R}^d$ satisfies $\mathbf{z}_j \perp \mathbf{v}$. The ratio $a_j / \|\mathbf{z}_j\|$ therefore grows with the alignment of $\mathbf{w}_j$ with the signal and shrinks if $\mathbf{w}_j$ instead aligns more with the noise. Collecting the noise components of the weight vectors, let $\mathbf{Z} \in \mathbb{R}^{(2m) \times d}$ be the matrix whose j -th row is $\mathbf{z}_j$. In order to track the alignment of the network as a whole with the signal versus noise subspaces we introduce the following quantities. Let

$$A_1 = f(\mathbf{W}, \mathbf{v}) = \sum_{j=1}^{2m} (-1)^j \sigma(a_j),$$

$$A_{-1} = f(\mathbf{W}, -\mathbf{v}) = \sum_{j=1}^{2m} (-1)^{j+1} \sigma(-a_j)$$

be referred to as the positive and negative signal activation of the network respectively. Moreover, define

$$A_{\min} = \min(A_1, A_{-1})$$

as the worst-case signal activation of the network, and

$$A_{\text{lin}} = \sum_{j=1}^{2m} (-1)^j a_j$$

as the linearized network activation. To measure the amount of noise the network learns we define

$$\mathbf{z}_{\text{lin}} = \sum_{j=1}^{2m} (-1)^j \mathbf{z}_j.$$

D.1 Training dynamics

Theorem 3.1. *Let $f : \mathbb{R}^p \times \mathbb{R}^n \rightarrow \mathbb{R}$ be a leaky ReLU network with forward pass as defined by equation 1. Suppose the step size η and initialization condition λ satisfy Assumption 1. Then for any linearly separable data set $(\mathbf{X}, \hat{\mathbf{y}})$ $\mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}}, \eta, \lambda)$ converges after T iterations, where*

$$T \leq \frac{C \|\mathbf{w}^*\|^2}{\eta \alpha^2 m}.$$

Furthermore $\mathcal{A}_{GD}$ is approximately margin maximizing on f (Definition 2.3) with

$$|\mathcal{A}_{GD}| \leq \frac{C}{\alpha \sqrt{m}}.$$

Proof. Our approach is to adapt a classical technique used for the proof of convergence of the Perceptron algorithm for linearly separable data. This is also the approach adopted by Brutkus et al. (2018). The key idea of the proof is to bound in terms of the number of updates both the norm of the learned vector $\mathbf{w}$ as well as its alignment with any linear separator of the data. From the Cauchy-Schwarz inequality these bounds cannot cross, and this in turn bounds the number of updates that can occur. Analogously, we track the alignment of $\mathbf{W}^{(t)}$ with the max-margin classifier along with the Frobenius norm of the $\mathbf{W}^{(t)}$. To this end denote

$$G(t) = \|\mathbf{W}_j^{(t)}\|_F^2$$

and

$$F(t) = \sum_{j=1}^{2m} (-1)^j \langle \mathbf{w}_j^{(t)}, \mathbf{w}^* \rangle,$$

where $\mathbf{w}^*$ is a max-margin linear classifier of the dataset. Recall that $\mathcal{F}^{(t)} = \{i \in [n] : \hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) < 1\}$ denotes the number of active data points at training step t . We also define $U(t) = \sum_{s=0}^{t-1} |\mathcal{F}^{(s)}|$ to be the number of data point updates between iterations 0 and t . First, by Cauchy-Schwarz

$$1 \leq \langle \mathbf{w}^*, \mathbf{x}_i \rangle \leq \|\mathbf{w}^*\| \cdot \|\mathbf{x}_i\|$$

for all $i \in [n]$. Therefore,

$$\|\mathbf{w}^*\| \geq \frac{1}{\min_{i \in [n]} \|\mathbf{x}_i\|}.$$

By Assumption 1, for all $j \in [2m]$,

$$\|\mathbf{w}_j^{(0)}\| \leq \frac{\sqrt{\alpha}}{m \min_{i \in [n]} \|\mathbf{x}_i\|} \leq \frac{\|\mathbf{w}^*\|}{\alpha m}. \quad (10)$$

For all $t \geq 0$, the update rule of GD implies

$$\begin{aligned} G(t+1) &= \sum_{j=1}^{2m} \|\mathbf{w}_j^{(t+1)}\|^2 \\ &= \sum_{j=1}^{2m} \left\| \mathbf{w}_j^{(t)} + \eta(-1)^j \sum_{i \in \mathcal{F}^{(t)}} \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \hat{y}_i \mathbf{x}_i \right\|^2 \\ &= \sum_{j=1}^{2m} \|\mathbf{w}_j^{(t)}\|^2 + 2\eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} (-1)^j \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \hat{y}_i \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle + \eta^2 \sum_{j=1}^{2m} \sum_{i, l \in \mathcal{F}^{(t)}} \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \langle \hat{y}_i \mathbf{x}_i, \hat{y}_l \mathbf{x}_l \rangle \\ &\leq \sum_{j=1}^{2m} \|\mathbf{w}_j^{(t)}\|^2 + 2\eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} (-1)^j \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \hat{y}_i \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle + 2m\eta^2 |\mathcal{F}^{(t)}|^2 \max_{i \in [n]} \|\mathbf{x}_i\|^2. \end{aligned}$$

Observe that for all $z \in \mathbb{R}$, $\sigma(s) = \dot{\sigma}(z)z$, so can rewrite the second term of the above expression as

$$\begin{aligned} 2\eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} (-1)^j \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \hat{y}_i \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle &= 2\eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} (-1)^j \sigma(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \hat{y}_i \\ &= 2\eta \sum_{i \in \mathcal{F}^{(t)}} \hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) \\ &< 2\eta \sum_{i \in \mathcal{F}^{(t)}} 1 \\ &= 2\eta |\mathcal{F}^{(t)}| \end{aligned}$$

where the inequality in the second-to-last line follows as we are summing over $\mathcal{F}^{(t)}$, which by definition consists of the $i \in [n]$ such that $\hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) < 1$. As a result we obtain

$$\begin{aligned} G(t+1) &\leq \sum_{j=1}^{2m} \|\mathbf{w}_j^{(t)}\|^2 + 2\eta |\mathcal{F}^{(t)}| + 2m\eta^2 |\mathcal{F}^{(t)}|^2 \max_{i \in [n]} \|\mathbf{x}_i\|^2 \\ &= G(t) + 2\eta |\mathcal{F}^{(t)}| + 2m\eta^2 |\mathcal{F}^{(t)}|^2 \max_{i \in [n]} \|\mathbf{x}_i\|^2 \\ &\leq G(t) + 4\eta |\mathcal{F}^{(t)}|, \end{aligned}$$

where the last line follows since

$$\eta \leq \frac{1}{mn \max_{i \in [n]} \|\mathbf{x}_i\|^2} \leq \frac{1}{|\mathcal{F}^{(t)}| m \max_{i \in [n]} \|\mathbf{x}_i\|^2}.$$

By equation 10, the initialization satisfies

$$\begin{aligned} G(0) &= \sum_{j=1}^{2m} \|\mathbf{w}_j^{(0)}\|^2 \\ &\leq \sum_{j=1}^{2m} \frac{\|\mathbf{w}^*\|^2}{\alpha^2 m^2} \\ &= \frac{2\|\mathbf{w}^*\|^2}{\alpha^2 m} \end{aligned}$$

So by induction, for all $t \geq 0$

$$G(t) \leq \frac{2\|\mathbf{w}^*\|^2}{\alpha^2 m} + 3\eta \sum_{s=0}^{t-1} |\mathcal{F}^{(s)}| = \frac{2\|\mathbf{w}^*\|^2}{\alpha^2 m} + 3\eta U(t). \quad (11)$$

Next we find a bound for $F(t)$. For all $t \geq 0$ then by definition of the GD update

$$\begin{aligned} F(t+1) &= \sum_{j=1}^{2m} (-1)^j \langle \mathbf{w}_j^{(t+1)}, \mathbf{w}^* \rangle \\ &= \sum_{j=1}^{2m} (-1)^j \langle \mathbf{w}_j^{(t)}, \mathbf{w}^* \rangle + \eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \hat{y}_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle. \end{aligned}$$

Since $\hat{y}_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq 1$ for all $i \in [n]$, the above expression is bounded below by

$$\begin{aligned} \sum_{j=1}^{2m} (-1)^j \langle \mathbf{w}_j^{(t)}, \mathbf{w}^* \rangle + \eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \hat{y}_i &= F(t) + \eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} \dot{\sigma}(\langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle) \\ &\geq F(t) + \eta \sum_{j=1}^{2m} \sum_{i \in \mathcal{F}^{(t)}} \alpha \\ &\geq F(t) + 2\eta m \alpha |\mathcal{F}^{(t)}|. \end{aligned}$$

Hence unrolling the update for GD for all $t \geq 0$ it follows that

$$F(t+1) \geq F(0) + 2\eta m \alpha \sum_{s=0}^{t-1} |\mathcal{F}^{(s)}|.$$

At initialization, by equation 10 then

$$\begin{aligned} F(0) &= \sum_{j=1}^{2m} (-1)^j \langle \mathbf{w}_j^{(0)}, \mathbf{w}^* \rangle \\ &\geq - \sum_{j=1}^{2m} \|\mathbf{w}_j^{(0)}\| \cdot \|\mathbf{w}^*\| \\ &\geq - \sum_{j=1}^{2m} \frac{\|\mathbf{w}^*\|^2}{\alpha m} \\ &= - \frac{2\|\mathbf{w}^*\|^2}{\alpha}. \end{aligned}$$

Therefore by induction, for all $t \geq 0$ we have

$$\begin{aligned} F(t) &\geq - \frac{2\|\mathbf{w}^*\|^2}{\alpha} + 2\eta m \alpha \sum_{s=0}^{t-1} |\mathcal{F}^{(s)}| \\ &= - \frac{2\|\mathbf{w}^*\|^2}{\alpha} + 2\eta m \alpha U(t). \end{aligned}$$

Combining our bounds for $F(t)$ and $G(t)$, we obtain

$$\begin{aligned}
-\frac{2\|\mathbf{w}^*\|^2}{\alpha} + 2\eta m \alpha U(t) &\leq F(t) \\
&= \sum_{j=1}^{2m} (-1)^j \langle \mathbf{w}_j^{(t)}, \mathbf{w}^* \rangle \\
&\leq \|\mathbf{w}^*\| \sum_{j=1}^{2m} \|\mathbf{w}_j^{(t)}\| \\
&\leq \|\mathbf{w}^*\| \left(2m \sum_{j=1}^{2m} \|\mathbf{w}_j^{(t)}\|^2 \right)^{1/2} \\
&= \|\mathbf{w}^*\| (2mG(t))^{1/2} \\
&\leq \|\mathbf{w}^*\| \left(\frac{4\|\mathbf{w}^*\|^2}{\alpha^2} + 6m\eta U(t) \right)^{1/2}.
\end{aligned}$$

This implies that either

$$-\frac{2\|\mathbf{w}^*\|^2}{\alpha} + 2\eta m \alpha U(t) \leq 0 \quad (12)$$

or

$$\left(-\frac{2\|\mathbf{w}^*\|^2}{\alpha} + 2\eta m \alpha U(t) \right)^2 \leq \|\mathbf{w}^*\|^2 \left(\frac{4\|\mathbf{w}^*\|^2}{\alpha^2} + 6m\eta U(t) \right). \quad (13)$$

If (12) holds, then

$$U(t) \leq \frac{\|\mathbf{w}^*\|^2}{\eta \alpha^2 m}.$$

If (13) holds, then rearranging yields

$$\begin{aligned}
4\eta^2 m^2 \alpha^2 U(t)^2 &\leq 14\|\mathbf{w}^*\|^2 \eta m U(t) \\
U(t) &\leq \frac{7\|\mathbf{w}^*\|^2}{2\eta \alpha^2 m}.
\end{aligned}$$

Therefore, in both cases there exists a constant C such that

$$U(t) \leq \frac{C\|\mathbf{w}^*\|^2}{\eta \alpha^2 m}. \quad (14)$$

This holds for all $t \in \mathbb{N}$ and therefore

$$\sum_{t=0}^{\infty} |\mathcal{F}^{(t)}| \leq \frac{C\|\mathbf{w}^*\|^2}{\eta \alpha^2 m} < \infty.$$

This implies that there exists $s \in \mathbb{N}$ such that $|\mathcal{F}^{(s)}| = 0$. Let $T \in \mathbb{N}$ be the minimal iteration such that $|\mathcal{F}^{(T)}| = 0$. Then for all $i \in [n]$ $\hat{y}_i f(\mathbf{W}^{(T)}, \mathbf{x}_i) \geq 1$. So the network achieves zero loss and also has zero gradient at iteration T . In particular,

$$T = \sum_{t=0}^{T-1} 1 \leq \sum_{t=0}^{T-1} |\mathcal{F}^{(t)}| \leq \frac{C\|\mathbf{w}^*\|^2}{\eta \alpha^2 m}.$$

To bound $|\mathcal{A}_{GD}|$ we combine equations (14) and (11) to obtain

$$\begin{aligned}
G(T) &\leq \frac{2\|\mathbf{w}^*\|^2}{\alpha^2 m} + 3\eta U(t) \\
&\leq \frac{2\|\mathbf{w}^*\|^2}{\alpha^2 m} + \frac{C\|\mathbf{w}^*\|^2}{\eta \alpha^2 m} \\
&\leq \frac{C\|\mathbf{w}^*\|^2}{\alpha^2 m}.
\end{aligned}$$

As a result for all linearly separable datasets $(\mathbf{X}, \hat{\mathbf{y}})$

$$\frac{\|\mathbf{W}\|_F}{\|\mathbf{w}^*\|} = \frac{C}{\alpha\sqrt{m}}$$

and therefore

$$|\mathcal{A}_{GD}| \leq \frac{C}{\alpha\sqrt{m}}$$

as claimed. $\square$

The training dynamics of gradient descent also give us the following result relating the linearization of the noise component of the network to the noise component of the network itself.

Lemma D.1. *Let $\lambda, \delta > 0$. Suppose that $d \geq \Omega(n + \log \frac{1}{\delta})$. In the context of training data $(\mathbf{X}, \hat{\mathbf{y}})$ sampled under the data model given in Definition 2.1, let $\mathbf{W} = \mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}}, \eta, \lambda)$. Then with probability at least $1 - \delta$ over the randomness of $(\mathbf{X}, \hat{\mathbf{y}})$*

$$\|\mathbf{Z}\|_F^2 - 2\lambda\sqrt{2m}\|\mathbf{Z}\|_F - 2m\lambda^2 \leq \frac{C}{\alpha m} (\|\mathbf{z}_{\text{lin}}\| + 2m\lambda)^2.$$

Proof. At each iteration of gradient descent,

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} + \eta(-1)^j \sum_{i=1}^n b_{ij}^{(t)} \hat{y}_i \mathbf{x}_i,$$

where

$$b_{ij}^{(t)} = \begin{cases} 0 & \text{if } \hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) \geq 1 \\ 1 & \text{if } \hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) < 1 \text{ and } \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i \rangle \geq 0 \\ \alpha & \text{otherwise.} \end{cases}$$

Let T be the iteration at which gradient descent terminates. Then for each $j \in [2m]$,

$$\mathbf{w}_j = \mathbf{w}_j^{(T)} = \mathbf{w}_j^{(0)} + \eta(-1)^j \sum_{t=0}^{T-1} \sum_{i=1}^n b_{ij}^{(t)} \hat{y}_i \mathbf{x}_i.$$

Then the noise component of $\mathbf{w}_j$ is given by

$$\begin{aligned} \mathbf{z}_j &= \mathbf{w}_j - \langle \mathbf{w}_j, \mathbf{v} \rangle \mathbf{v} \\ &= \mathbf{w}_j^{(0)} - \langle \mathbf{w}_j^{(0)}, \mathbf{v} \rangle \mathbf{v} + \eta(-1)^j \sum_{t=0}^{T-1} \sum_{i=1}^n b_{ij}^{(t)} \hat{y}_i (\mathbf{x}_i - \langle \mathbf{x}_i, \mathbf{v} \rangle \mathbf{v}) \\ &= \mathbf{w}_j^{(0)} - \langle \mathbf{w}_j^{(0)}, \mathbf{v} \rangle \mathbf{v} + \eta(-1)^j \sum_{t=0}^{T-1} \sum_{i=1}^n b_{ij}^{(t)} \hat{y}_i \mathbf{n}_i. \end{aligned}$$

Define

$$\hat{\mathbf{z}}_j = \mathbf{z}_j - \mathbf{w}_j^{(0)} + \langle \mathbf{w}_j^{(0)}, \mathbf{v} \rangle \mathbf{v}$$

and let

$$\hat{\mathbf{z}}_{\text{lin}} = \sum_{j=1}^{2m} (-1)^j \hat{\mathbf{z}}_j,$$

Then for all $j \in [2n]$,

$$\begin{aligned} (\|\hat{\mathbf{z}}_j\| - \|\mathbf{z}_j\|)^2 &\leq \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|^2 \\ &= \|\mathbf{w}_j^{(0)} - \langle \mathbf{w}_j^{(0)}, \mathbf{v} \rangle \mathbf{v}\|^2 \\ &\leq \|\mathbf{w}_j^{(0)}\|^2 \\ &\leq \lambda^2. \end{aligned}$$

Furthermore, if $\|\mathbf{z}_j\| \leq \|\hat{\mathbf{z}}_j\|$ then the above implies $\|\hat{\mathbf{z}}_j\| \leq \|\mathbf{z}_j\| + \lambda$ while if $\|\mathbf{z}_j\| \geq \|\hat{\mathbf{z}}_j\|$ then this inequality holds trivially. As a result,

$$\begin{aligned} |\|\hat{\mathbf{z}}_j\|^2 - \|\mathbf{z}_j\|^2| &= |(\|\hat{\mathbf{z}}_j\| + \|\mathbf{z}_j\|) \cdot (\|\hat{\mathbf{z}}_j\| - \|\mathbf{z}_j\|)| \\ &\leq \|\hat{\mathbf{z}}_j\| + \|\mathbf{z}_j\| \cdot |\|\hat{\mathbf{z}}_j\| - \|\mathbf{z}_j\|| \\ &\leq (2\|\mathbf{z}_j\| + \lambda)(\lambda). \end{aligned}$$

If $\|\mathbf{z}_j\| \geq \|\hat{\mathbf{z}}_j\|$ then the above implies $\|\hat{\mathbf{z}}_j\|^2 \geq \|\mathbf{z}_j\|^2 - \lambda(2\|\mathbf{z}_j\| + \lambda)$, if $\|\mathbf{z}_j\| \leq \|\hat{\mathbf{z}}_j\|$ this inequality is trivially true. As a result,

$$\begin{aligned} \sum_{j=1}^{2m} \|\hat{\mathbf{z}}_j\|^2 &\geq \sum_{j=1}^{2m} (\|\mathbf{z}_j\|^2 - \lambda(2\|\mathbf{z}_j\| + \lambda)) \\ &= \sum_{j=1}^{2m} \|\mathbf{z}_j\|^2 - 2\lambda \sum_{j=1}^{2m} \|\mathbf{z}_j\| - 2m\lambda^2 \\ &\geq \sum_{j=1}^{2m} \|\mathbf{z}_j\|^2 - 2\lambda\sqrt{2m} \left(\sum_{j=1}^{2m} \|\mathbf{z}_j\|^2 \right)^{1/2} - 2m\lambda^2 \\ &= \|\mathbf{Z}\|_F^2 - 2\lambda\sqrt{2m}\|\mathbf{Z}\|_F - 2m\lambda^2, \end{aligned} \tag{15}$$

where the third line is an application of Cauchy-Schwarz. Moreover,

$$\begin{aligned} \|\hat{\mathbf{z}}_{\text{lin}} - \mathbf{z}_{\text{lin}}\| &= \left\| \sum_{j=1}^{2m} (-1)^j (\hat{\mathbf{z}}_j - \mathbf{z}_j) \right\| \\ &\leq \sum_{j=1}^{2m} \|\hat{\mathbf{z}}_j - \mathbf{z}_j\| \\ &\leq \sum_{j=1}^{2m} \lambda \\ &\leq 2m\lambda, \end{aligned}$$

so

$$\|\hat{\mathbf{z}}_{\text{lin}}\| \geq \|\mathbf{z}_{\text{lin}}\| - 2m\lambda. \tag{16}$$

Let $\mathbf{N}' \in \mathbb{R}^{d \times n}$ to be the matrix whose i -th column is $\hat{y}_i \mathbf{n}_i$, equivalently $\mathbf{N}' = \mathbf{N} \text{diag}(\hat{\mathbf{y}})$. Then

$$\hat{\mathbf{z}}_j = \eta(-1)^j \mathbf{N}' \mathbf{c}_j,$$

where $\mathbf{c}_j \in \mathbb{R}^n$ is given by

$$(\mathbf{c}_j)_i = \sum_{t=0}^{T-1} b_{ij}^{(t)}.$$

Due to symmetry of the noise distribution then the columns of $\mathbf{N}'$ are i.i.d. with distribution $\mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$. Therefore by Lemma 4.1 (and the assumptions $d = \Omega(n + \log \frac{1}{\delta})$), with probability at least $1 - \delta$ over the randomness of the training data there exist positive constants C', C such that $C' \leq \sigma_{\min}(\mathbf{N}') \leq \sigma_{\max}(\mathbf{N}') \leq C$. As a result

$$C' \eta \|\mathbf{c}_j\| \leq \|\hat{\mathbf{z}}_j\| \leq C \eta \|\mathbf{c}_j\|. \tag{17}$$

We claim that for any $j, j' \in [2m]$ and $i \in [n]$, $(\mathbf{c}_j)_i \geq \alpha(\mathbf{c}_{j'})_i$. Indeed, if $\hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) \geq 1$, then $b_{ij}^{(t)} = b_{ij'}^{(t)} = 0$, and if $\hat{y}_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) < 1$, then both $b_{ij}^{(t)}$ and $b_{ij'}^{(t)}$ are elements of $\{\alpha, 1\}$. This in particular implies that

$$\langle \mathbf{c}_j, \mathbf{c}_{j'} \rangle \geq \alpha \langle \mathbf{c}_j, \mathbf{c}_{j'} \rangle.$$

Let us define

$$\mathbf{c}_{\text{lin}} = \sum_{j=1}^{2m} \mathbf{c}_j.$$

Then

$$\begin{aligned} \|\hat{\mathbf{z}}_{\text{lin}}\|^2 &= \left\| \sum_{j=1}^{2m} (-1)^j \hat{\mathbf{z}}_j \right\|^2 \\ &= \left\| \sum_{j=1}^{2m} \eta \mathbf{N}' \mathbf{c}_j \right\|^2 \\ &\leq C\eta^2 \left\| \sum_{j=1}^{2m} \mathbf{c}_j \right\|^2 \\ &= C\eta^2 \|\mathbf{c}_{\text{lin}}\|^2, \end{aligned} \tag{18}$$

where we used that $\|\mathbf{N}'\| \leq C$ in the third line. We also have

$$\begin{aligned} \|\mathbf{c}_{\text{lin}}\|^2 &= \sum_{j=1}^{2m} \sum_{j'=1}^{2m} \langle \mathbf{c}_j, \mathbf{c}_{j'} \rangle \\ &\geq \alpha \sum_{j=1}^{2m} \sum_{j'=1}^{2m} \langle \mathbf{c}_j, \mathbf{c}_j \rangle \\ &= 2\alpha m \sum_{j=1}^{2m} \|\mathbf{c}_j\|^2. \end{aligned} \tag{19}$$

Finally we combine our bounds for $\mathbf{c}$, $\mathbf{z}$, and $\hat{\mathbf{z}}$:

$$\begin{aligned} \|\mathbf{Z}\|_F^2 - 2\lambda\sqrt{2m}\|\mathbf{Z}\|_F - 2m\lambda^2 &\leq \sum_{j=1}^{2m} \|\hat{\mathbf{z}}_j\|^2 \\ &\leq C\eta^2 \sum_{j=1}^{2m} \|\mathbf{c}_j\|^2 \\ &\leq \frac{C\eta^2}{\alpha m} \|\mathbf{c}_{\text{lin}}\|^2 \\ &\leq \frac{C}{\alpha m} \|\hat{\mathbf{z}}_{\text{lin}}\|^2 \\ &\leq \frac{C}{\alpha m} (\|\mathbf{z}_{\text{lin}}\| + 2m\lambda)^2. \end{aligned}$$

Here we applied equations (15) in the first line, (17) in the second line, (19) in the third line, (18) in the fourth line, and (16) in the fifth line. This establishes both the bounds claimed. $\square$

D.2 Benign overfitting

To establish benign overfitting in leaky ReLU networks, we first determine an upper bound on the generalization error of the model in terms of the signal-to-noise ratio of the network weights.

Lemma D.2. *Let $\epsilon \in (0, 1)$. Suppose that*

$$\frac{A_{\min}}{\|\mathbf{Z}\|_F} \geq C_2 \sqrt{\frac{(1-\gamma)m \log \frac{1}{\epsilon}}{\gamma d}}.$$

Then for test data $(\mathbf{x}, y)$ as per Definition 2.1,

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0) \leq \epsilon.$$

Proof. Recall that a test point $(\mathbf{x}, y)$ satisfies

$$\mathbf{x} = y(\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n}),$$

where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_d, \frac{1}{d}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$. If $yf(\mathbf{W}, \mathbf{x}) \leq 0$, then

$$\begin{aligned} 0 &\geq yf(\mathbf{W}, \mathbf{x}) \\ &= \sum_{j=1}^{2m} (-1)^j y \sigma(\langle \mathbf{w}_j, \mathbf{x} \rangle) \\ &= \sum_{j=1}^{2m} (-1)^j y \sigma(\langle a_j \mathbf{v} + \mathbf{z}_j, y(\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n}) \rangle) \\ &= \sum_{j=1}^{2m} (-1)^j y \sigma(y(\sqrt{\gamma}a_j + \sqrt{1-\gamma}\langle \mathbf{z}_j, \mathbf{n} \rangle)) \\ &\geq \sum_{j=1}^{2m} (-1)^j y \sigma(y\sqrt{\gamma}a_j) - \sum_{j=1}^{2m} \sqrt{1-\gamma} |\langle \mathbf{z}_j, \mathbf{n} \rangle| \\ &= \sqrt{\gamma}A_y - \sum_{j=1}^{2m} \sqrt{1-\gamma} |\langle \mathbf{z}_j, \mathbf{n} \rangle|. \end{aligned}$$

When $A_{\min} \geq 0$, this implies that

$$\begin{aligned} \gamma A_{\min}^2 &\leq (1-\gamma) \left(\sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n} \rangle| \right)^2 \\ &\leq 2m(1-\gamma) \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n} \rangle|^2 \\ &= 2m(1-\gamma) \|\mathbf{Z}\mathbf{n}\|^2 \\ &\leq 2m(1-\gamma) \|\mathbf{Z}\|_F^2 \|\mathbf{n}\|^2, \end{aligned}$$

where the second inequality is an application of Cauchy-Schwarz. So

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0) \leq \mathbb{P}\left(\|\mathbf{Z}\mathbf{n}\|^2 \geq \frac{\gamma A_{\min}^2}{2m(1-\gamma)}\right).$$

By Lemma A.1, the above probability is less than ϵ if

$$\sqrt{\frac{\gamma}{2m(1-\gamma)}} A_{\min} \geq C \|\mathbf{Z}\|_F \sqrt{\frac{1}{d} \log \frac{1}{\epsilon}},$$

or equivalently,

$$\frac{A_{\min}}{\|\mathbf{Z}\|_F} \geq C_2 \sqrt{\frac{(1-\gamma)m \log \frac{1}{\epsilon}}{\gamma d}}.$$

□

We will also need the number of positive labels to be (mildly) balanced with the number of negative labels.

Lemma D.3. *Let $\delta > 0$ and suppose that $\ell = \Omega(\log \frac{1}{\delta})$. Let $\mathcal{I} \subseteq [n]$ be an arbitrary subset such that $|\mathcal{I}| = \ell$. Consider training data $(\mathbf{X}, \mathbf{y})$ as per the data model given in Definition 2.1. Then with probability at least $1 - \delta$,*

$$\frac{\ell}{4} \leq |\{i \in \mathcal{S} : y_i = 1\}| \leq \frac{3\ell}{4}.$$

Proof. For $i \in \mathcal{I}$ let Y_i be a random variable taking the value 1 if $y_i = 1$ and 0 if $y_i = -1$. Then the Y_i are i.i.d. Bernoulli random variables with $\mathbb{P}(Y_i = 1) = \frac{1}{2}$. Let

$$Y = \sum_{i \in \mathcal{I}} Y_i = |\{i \in \mathcal{S} : y_i = 1\}|$$

so that $\mathbb{E}[Y] = \frac{\ell}{2}$. By Chernoff's inequality, for all $t \in (0, 1)$,

$$\mathbb{P}\left(\left|Y - \frac{\ell}{2}\right| \geq t \frac{\ell}{2}\right) \leq 2e^{-C\ell t^2}.$$

Setting $t = \frac{1}{2}$, we see that $\frac{\ell}{4} \leq Y \leq \frac{3\ell}{4}$ with probability at least

$$1 - 2 \exp\left(-\frac{C\ell}{4}\right) \geq 1 - \delta$$

when $\ell = \Omega\left(\log \frac{1}{\delta}\right)$. □

We are now able to prove our main benign overfitting result for leaky ReLU networks.

Theorem 3.2. *Under the setting given in Assumption 2, let $\delta \in (0, 1)$ and suppose $\mathcal{A}$ is approximately margin-maximizing (Definition 2.3). If $n = \Omega\left(\log \frac{1}{\delta}\right)$, $d = \Omega(n)$, $k = O\left(\frac{n}{1+m|\mathcal{A}|^2}\right)$, and $\gamma = \Omega\left(\frac{1}{k}\right)$ then there is a fixed positive constant C such that with probability at least $1 - \delta$ over $(\mathbf{X}, \hat{\mathbf{y}})$*

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0 \mid \mathbf{X}, \hat{\mathbf{y}}) \leq \exp\left(-C \cdot \frac{d}{k(1+m|\mathcal{A}|^2)}\right).$$

Proof. Since $d = \Omega(n) = \Omega\left(n + \log \frac{1}{\delta}\right)$, by Lemma B.3, with probability at least $1 - \frac{\delta}{3}$ over the randomness of the data, the max-margin classifier $\mathbf{w}^*$ satisfies

$$\|\mathbf{w}^*\| \leq C \sqrt{\frac{1}{\gamma} + \frac{k}{1-\gamma}}.$$

We denote this event by ω_1 . For $s \in \{1, -1\}$, let $\mathcal{G}_s$ denote the set of $i \in \mathcal{G}$ such that $\langle \mathbf{v}, \mathbf{x}_i \rangle = s$. If $n = \Omega\left(\frac{1}{\delta}\right)$ and $k = O(n)$, then $|\mathcal{G}| = \Omega\left(\log \frac{1}{\delta}\right)$. Under these assumptions, by Lemma D.3,

$$|\mathcal{G}_s| \geq \frac{1}{4}|\mathcal{G}| \geq Cn$$

for both $s \in \{1, -1\}$ with probability at least $1 - \frac{\delta}{3}$. We denote this event by ω_2 . For $s \in \{1, -1\}$, let $\mathbf{N}_{\mathcal{G}_s} \in \mathbb{R}^{|\mathcal{G}_s| \times d}$ be the matrix whose rows are indexed by $\mathcal{G}_s$ and are given by the vectors $\mathbf{n}_i$ for $i \in \mathcal{G}_s$. As $d = \Omega(n) = \Omega\left(n + \log \frac{1}{\delta}\right)$ and the rows of $\mathbf{N}_{\mathcal{G}_s}$ are drawn mutually i.i.d. from $\mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}^T))$, the following holds by Lemma 4.1. With probability at least $1 - \frac{\delta}{3}$ over the randomness of the training data, $\|\mathbf{N}_{\mathcal{G}_s}\| \leq C$ for both $s \in \{1, -1\}$. We denote this event by ω_3 . Let $\omega = \omega_1 \cap \omega_2 \cap \omega_3$. By the union bound $\mathbb{P}(\omega) \geq 1 - \delta$. We condition on ω for the remainder of this proof.

Since $\mathbf{W} = \mathcal{A}(\mathbf{X}, \hat{\mathbf{y}})$ and $\mathcal{A}$ is approximately margin maximizing,

$$\begin{aligned} \|\mathbf{W}\|_F &\leq |\mathcal{A}| \cdot \|\mathbf{w}^*\| \\ &\leq C|\mathcal{A}| \sqrt{\frac{1}{\gamma} + \frac{k}{1-\gamma}}. \end{aligned} \tag{20}$$

Let $s \in \{-1, 1\}$ be such that $A_s = A_{\min}$. Since the network attains zero loss, for all $i \in \mathcal{G}_s$,

$$\begin{aligned}
1 &\leq \hat{y}_i f(\mathbf{W}, \mathbf{x}_i) \\
&= \sum_{j=1}^{2m} (-1)^j \hat{y}_i \sigma(\langle \mathbf{w}_j, \mathbf{x}_i \rangle) \\
&= \sum_{j=1}^{2m} (-1)^j y_i \sigma(\langle a_j \mathbf{v} + \mathbf{z}_j, \sqrt{\gamma} y_i \mathbf{v} + \sqrt{1-\gamma} \mathbf{n}_i \rangle) \\
&= \sum_{j=1}^{2m} (-1)^j y_i \sigma(\sqrt{\gamma} a_j y_i + \sqrt{1-\gamma} \langle \mathbf{z}_j, \mathbf{n}_i \rangle) \\
&\leq \sum_{j=1}^{2m} (-1)^j y_i \sigma(\sqrt{\gamma} a_j y_i) + \sum_{j=1}^{2m} |\sqrt{1-\gamma} \langle \mathbf{z}_j, \mathbf{n}_i \rangle| \\
&= \sqrt{\gamma} \sum_{j=1}^{2m} (-1)^j s \sigma(s a_j) + \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \\
&= \sqrt{\gamma} A_s + \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \\
&= \sqrt{\gamma} A_{\min} + \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|.
\end{aligned}$$

Hence, we have either $\sqrt{\gamma} A_s \geq \frac{1}{2}$ or $\sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \geq \frac{1}{2}$ for all $i \in \mathcal{G}_s$. We consider these two cases separately.

If $\sqrt{\gamma} A_{\min} \geq \frac{1}{2}$, then

$$\begin{aligned}
\frac{A_{\min}}{\|\mathbf{Z}\|_F} &\geq \frac{A_{\min}}{\|\mathbf{W}\|_F} \\
&\geq \frac{1}{2\sqrt{\gamma} \|\mathbf{W}\|_F} \\
&\geq C \frac{1}{\sqrt{\gamma} |\mathcal{A}| \sqrt{\frac{1}{\gamma} + \frac{k}{1-\gamma}}} \\
&= C \frac{1}{|\mathcal{A}| \sqrt{1 + \frac{k\gamma}{1-\gamma}}} \\
&\geq C \frac{1}{|\mathcal{A}| + |\mathcal{A}| \sqrt{\frac{k\gamma}{1-\gamma}}}.
\end{aligned}$$

Then by Lemma D.2, the network has generalization error less than ϵ when

$$\frac{1}{|\mathcal{A}| + |\mathcal{A}| \sqrt{\frac{k\gamma}{1-\gamma}}} \geq C \sqrt{\frac{(1-\gamma)m \log \frac{1}{\epsilon}}{\gamma d}}$$

or equivalently

$$\sqrt{\frac{(1-\gamma)m \log \frac{1}{\epsilon}}{\gamma d}} + \sqrt{\frac{mk \log \frac{1}{\epsilon}}{d}} \leq \frac{C}{|\mathcal{A}|}.$$

This is satisfied for $\epsilon = \exp(-C \cdot \frac{d}{k(1+m|\mathcal{A}|^2)})$ for some different constant C when $\gamma = \Omega(\frac{1}{k})$, which is true by assumption. So if $\sqrt{\gamma} A_{\min} \geq \frac{1}{2}$, then the network has generalization error less than ϵ whenever ω occurs, which happens with probability at least $1 - \delta$.

Now suppose that $\sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \geq \frac{1}{2}$ for all $i \in \mathcal{G}_s$. Squaring both sides of the inequality and applying Cauchy-Schwarz, we obtain

$$\begin{aligned} \frac{1}{4} &\leq (1-\gamma) \left(\sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \right)^2 \\ &\leq 2m(1-\gamma) \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|^2. \end{aligned}$$

Summing over all $i \in \mathcal{G}_s$, we obtain

$$\begin{aligned} \frac{|\mathcal{G}_s|}{4} &\leq 2m(1-\gamma) \sum_{i \in \mathcal{G}_s} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|^2 \\ &= 2m(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{N}_{\mathcal{G}_s} \mathbf{z}_j\|^2, \end{aligned}$$

Applying ω_2 and ω_3 , we obtain the bound

$$\begin{aligned} n &\leq Cm(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{N}_{\mathcal{G}_s} \mathbf{z}_j\|^2 \\ &\leq Cm(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{N}_{\mathcal{G}_s}\|^2 \|\mathbf{z}_j\|^2 \\ &\leq Cm(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{z}_j\|^2 \\ &= Cm(1-\gamma) \|\mathbf{Z}\|_F^2 \\ &\leq Cm(1-\gamma) \|\mathbf{W}\|_F^2. \end{aligned}$$

Then applying (20),

$$\begin{aligned} n &\leq Cm(1-\gamma) \|\mathbf{W}\|_F^2 \\ &\leq Cm(1-\gamma) |\mathcal{A}|^2 \left(\frac{1}{\gamma} + \frac{k}{1-\gamma} \right) \\ &\leq Cm |\mathcal{A}|^2 \left(\frac{1}{\gamma} + k \right). \end{aligned}$$

This implies that

$$n \leq \frac{Cm |\mathcal{A}|^2}{\gamma}$$

or

$$k \geq \frac{Cn}{m |\mathcal{A}|^2}.$$

Neither of these conditions can occur if $\gamma = \Omega\left(\frac{1}{k}\right)$ and $k = O\left(\frac{n}{|\mathcal{A}|^2 m}\right)$. Thus, in all cases, the network has generalization error less than $\exp(-C \cdot \frac{d}{k(1+m|\mathcal{A}|^2)})$ when ω occurs, which happens with probability at least $1 - \delta$. $\square$

We are also able to show the lower bound for the generalization error stated in the main text.

Theorem 3.3. *Under the setting given in Assumption 2, let $\delta \in (0, 1)$ and suppose $\mathcal{A} = \mathcal{A}_{GD}$ where $\eta, \lambda \in \mathbb{R}_{>0}$ satisfy Assumption 1. If $n = \Omega(k)$, $d = \Omega(n)$, and $k = \Omega(\log \frac{1}{\delta} + \frac{1}{\alpha})$, then there is a fixed positive constant C such that with probability at least $1 - \delta$ over $(\mathbf{X}, \hat{\mathbf{y}})$*

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0 \mid \mathbf{X}, \hat{\mathbf{y}}) \geq \exp\left(-C \cdot \frac{d}{\alpha k}\right).$$

Proof. We proceed along the lines of Theorem 3.2. For $s \in \{1, -1\}$, let $\mathcal{B}_s$ denote the set of $i \in \mathcal{B}$ such that $\langle \mathbf{v}, \mathbf{x}_i \rangle = s$. Note $|\mathcal{B}| = \Omega(\log \frac{1}{\delta})$. Under these assumptions, by Lemma D.3,

$$|\mathcal{B}_s| \geq \frac{1}{4}|\mathcal{B}| \geq Ck$$

for both $s \in \{1, -1\}$ with probability at least $1 - \frac{\delta}{3}$. We denote this event by ω_1 . For $s \in \{1, -1\}$, let $\mathbf{N}_{\mathcal{B}_s} \in \mathbb{R}^{|\mathcal{B}_s| \times d}$ be the matrix whose rows are indexed by $\mathcal{B}_s$ and are given by the vectors $\mathbf{n}_i$ for $i \in \mathcal{B}_s$. As $d = \Omega(n) = \Omega(k + \log \frac{1}{\delta})$ and the rows of $\mathbf{N}_{\mathcal{B}_s}$ are drawn mutually i.i.d. from $\mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$, the following holds by Lemma 4.1. With probability at least $1 - \frac{\delta}{3}$ over the randomness of the training data, $\|\mathbf{N}_{\mathcal{B}_s}\| \leq C$ for both $s \in \{1, -1\}$. We denote this event by ω_2 . By Lemma D.1, there is a constant C such that

$$\|\mathbf{Z}\|_F^2 - 2\lambda\sqrt{2m}\|\mathbf{Z}\|_F - 2m\lambda^2 \leq \frac{C}{\alpha m} (\|\mathbf{z}_{\text{lin}}\| + 2m\lambda)^2. \quad (21)$$

with probability at least $1 - \frac{\delta}{3}$. We denote this event by ω_3 . Let $\omega = \omega_1 \cap \omega_2 \cap \omega_3$. By the union bound $\mathbb{P}(\omega) \geq 1 - \delta$. We condition on ω for the remainder of this proof.

Let $s \in \{1, -1\}$ be such that $A_s = \max\{A_1, A_{-1}\}$. Since the network attains zero loss, for all $i \in \mathcal{B}_s$,

$$\begin{aligned} 1 &\leq \hat{y}_i f(\mathbf{W}, \mathbf{x}_i) \\ &= \sum_{j=1}^{2m} (-1)^j \hat{y}_i \sigma(\langle \mathbf{w}_j, \mathbf{x}_i \rangle) \\ &= \sum_{j=1}^{2m} (-1)^j y_i \sigma(\langle a_j \mathbf{v} + \mathbf{z}_j, -\sqrt{\gamma} y_i \mathbf{v} + \sqrt{1-\gamma} \mathbf{n}_i \rangle) \\ &= \sum_{j=1}^{2m} (-1)^j y_i \sigma(-\sqrt{\gamma} a_j y_i + \sqrt{1-\gamma} \langle \mathbf{z}_j, \mathbf{n}_i \rangle) \\ &\leq \sum_{j=1}^{2m} (-1)^j y_i \sigma(-\sqrt{\gamma} a_j y_i) + \sum_{j=1}^{2m} |\sqrt{1-\gamma} \langle \mathbf{z}_j, \mathbf{n}_i \rangle| \\ &= \sqrt{\gamma} \sum_{j=1}^{2m} (-1)^{j+1} s \sigma(s a_j) + \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \\ &= -\sqrt{\gamma} A_s + \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|. \end{aligned}$$

From which we conclude

$$\sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \geq 1 + \sqrt{\gamma} A_s \geq \sqrt{\gamma} A_s$$

for all such i . Squaring both sides of the inequality and applying Cauchy-Schwarz, we obtain

$$\begin{aligned} \gamma A_s &\leq (1-\gamma) \left(\sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \right)^2 \\ &\leq 2m(1-\gamma) \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|^2. \end{aligned}$$

Summing over all $i \in \mathcal{B}_s$, we obtain

$$\begin{aligned} |\mathcal{B}_s| \gamma A_s &\leq 2m(1-\gamma) \sum_{i \in \mathcal{B}_s} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|^2 \\ &= 2m(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{N}_{\mathcal{B}_s} \mathbf{z}_j\|^2, \end{aligned}$$

Applying ω_2 and ω_3 , we obtain the bound

$$\begin{aligned} k\gamma A_s &\leq Cm(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{N}_{\mathcal{B}_s} \mathbf{z}_j\|^2 \\ &\leq Cm(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{N}_{\mathcal{B}_s}\|^2 \|\mathbf{z}_j\|^2 \\ &\leq Cm(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{z}_j\|^2 \\ &= Cm(1-\gamma) \|\mathbf{Z}\|_F^2. \end{aligned}$$

For $k = \Omega(\frac{1}{\alpha})$, this inequality along with Assumption 1 implies that

$$C\|\mathbf{Z}\|_F^2 \leq \|\mathbf{Z}\|_F^2 + 2\lambda\sqrt{2m}\|\mathbf{Z}\|_F + 2m\lambda^2$$

for a different constant C . With the last two inequalities and equation 21, we obtain the bound, for a new constant C .

$$k\gamma A_s \leq C \frac{1-\gamma}{\alpha} (\|\mathbf{z}_{\text{lin}}\| + 2m\lambda)^2.$$

We then apply $k = \Omega(\frac{1}{\alpha})$ and Assumption 1 again to conclude that for some C ,

$$\|\mathbf{z}_{\text{lin}}\| \geq C \sqrt{\frac{k\gamma A_s \alpha}{1-\gamma}}.$$

Note that

$$A_{\text{lin}} = \frac{A_1 + A_{-1}}{1+\alpha} \leq 2A_s.$$

We then bound

$$\begin{aligned} \frac{A_{\text{lin}}}{\|\mathbf{z}_{\text{lin}}\|} &\leq C \frac{A_s}{\sqrt{\frac{k\gamma A_s \alpha}{1-\gamma}}} \\ &\leq C \sqrt{\frac{1-\gamma}{k\gamma \alpha}} \end{aligned}$$

for some constant C . Now consider a test point $(\mathbf{x}, y)$, which satisfies

$$\mathbf{x} = y(\sqrt{\gamma}\mathbf{v} + \sqrt{1-\gamma}\mathbf{n}),$$

where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_d, \frac{1}{d}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$. Since the data distribution is symmetric,

$$\begin{aligned} \mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0) &\geq \frac{1}{2} \mathbb{P}(yf(\mathbf{W}, \mathbf{x}) \leq 0 \text{ or } -yf(\mathbf{W}, -\mathbf{x}) \leq 0) \\ &\geq \frac{1}{2} \mathbb{P}(yf(\mathbf{W}, \mathbf{x}) - yf(\mathbf{W}, -\mathbf{x}) \leq 0). \end{aligned}$$

We see that

$$yf(\mathbf{W}, \mathbf{x}) - yf(\mathbf{W}, -\mathbf{x}) = (1+\alpha) \left(yA_{\text{lin}}\sqrt{\gamma} + \langle \mathbf{z}_{\text{lin}}, \mathbf{n} \rangle \sqrt{1-\gamma} \right)$$

By Lemma C.1 we then can bound

$$\begin{aligned} \mathbb{P}(y\langle \mathbf{w}, \mathbf{x} \rangle \leq 0 \mid \omega) &\geq \frac{1}{8} \exp \left(-\frac{6d}{\pi} \frac{\gamma}{1-\gamma} \frac{A_{\text{lin}}^2}{\|\mathbf{z}_{\text{lin}}\|^2} \right) \\ &\geq \exp \left(-\frac{Cd}{\alpha k} \right) \end{aligned}$$

for a new constant C , provided A_{lin} is positive. In the last line we can bound $\frac{d}{k}$ below as $d = \Omega(n) = \Omega(k)$. If A_{lin} is negative, then the generalization error is at least $\frac{1}{4}$ which is also at least $\exp(-Cd/(\alpha k))$. $\square$

D.3 Non-benign overfitting

In this section we show that leaky ReLU networks trained on low-signal data exhibit non-benign overfitting. As in the case of benign overfitting, we will rely on a generalization bound which depends on the signal-to-noise ratio of the network.

Lemma D.4. *Let $\mathbf{W} \in \mathbb{R}^{2m \times d}$ be the first layer weight matrix of a shallow leaky ReLU network given by equation 1. Suppose $(\mathbf{x}, y)$ is a random test point sampled under the data model given in Definition 2.1. If $\mathbf{W}$ is such that $A_{\text{lin}} \geq 0$ and*

$$\frac{A_{\text{lin}}}{z_{\text{lin}}} = O\left(\sqrt{\frac{1-\gamma}{\gamma d}}\right)$$

then

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) < 0) \geq \frac{1}{8}.$$

Alternatively, if $A_{\text{lin}} \leq 0$ then

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) < 0) \geq \frac{1}{4}.$$

Proof. By Definition 2.1 $(-\mathbf{x}, -y)$ is identically distributed to $(\mathbf{x}, y)$, therefore

$$\begin{aligned} \mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x})) &= \frac{1}{2} (\mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x})) + \mathbb{P}(0 > -yf(\mathbf{W}, -\mathbf{x}))) \\ &\geq \frac{1}{2} \mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x}) \cup 0 > -yf(\mathbf{W}, -\mathbf{x})) \\ &\geq \frac{1}{2} \mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x}) - yf(\mathbf{W}, -\mathbf{x})). \end{aligned}$$

Next we compute

$$\begin{aligned} &yf(\mathbf{W}, \mathbf{x}) - yf(\mathbf{W}, -\mathbf{x}) \\ &= \sum_{j=1}^{2m} (-1)^j y(\sigma(\langle \mathbf{w}_j, \mathbf{x} \rangle) - \sigma(\langle \mathbf{w}_j, -\mathbf{x} \rangle)) \\ &= (1 + \alpha) \sum_{j=1}^{2m} (-1)^j y \langle \mathbf{w}_j, \mathbf{x} \rangle \\ &= (1 + \alpha) \sum_{j=1}^{2m} (-1)^j \langle a_j \mathbf{v} + \mathbf{z}_j, \sqrt{\gamma} \mathbf{v} + \sqrt{1-\gamma} \mathbf{n} \rangle \\ &= (1 + \alpha) \sqrt{\gamma} \sum_{j=1}^{2m} (-1)^j a_j + (1 + \alpha) \sqrt{1-\gamma} \left\langle \mathbf{n}, \sum_{j=1}^{2m} (-1)^j \mathbf{z}_j \right\rangle \\ &= (1 + \alpha) \sqrt{\gamma} A_{\text{lin}} + (1 + \alpha) \sqrt{1-\gamma} \langle \mathbf{n}, \mathbf{z}_{\text{lin}} \rangle. \end{aligned}$$

The above two calculations imply that

$$\begin{aligned} \mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x})) &\geq \frac{1}{2} \mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x}) - yf(\mathbf{W}, -\mathbf{x})) \\ &= \frac{1}{2} \mathbb{P}(0 > (1 + \alpha) \sqrt{\gamma} A_{\text{lin}} + (1 + \alpha) \sqrt{1-\gamma} \langle \mathbf{n}, \mathbf{z}_{\text{lin}} \rangle) \\ &= \frac{1}{2} \mathbb{P}\left(\langle -\mathbf{n}, \mathbf{z}_{\text{lin}} \rangle > \sqrt{\frac{\gamma}{1-\gamma}} A_{\text{lin}}\right). \end{aligned}$$

Suppose that $A_{\text{lin}} \geq 0$. As the noise distribution is symmetric $\langle \mathbf{n}, \mathbf{z}_{\text{lin}} \rangle \stackrel{d}{=} \langle -\mathbf{n}, \mathbf{z}_{\text{lin}} \rangle$. Therefore,

$$\frac{1}{4} \mathbb{P}\left(|\langle \mathbf{n}, \mathbf{z}_{\text{lin}} \rangle| > \sqrt{\frac{\gamma}{1-\gamma}} A_{\text{lin}}\right) = \frac{1}{4} \mathbb{P}\left(|\langle \mathbf{n}, \mathbf{u} \rangle| > \sqrt{\frac{\gamma}{1-\gamma}} \frac{A_{\text{lin}}}{\|\mathbf{z}_{\text{lin}}\|}\right),$$

where $\mathbf{u} = \frac{\mathbf{z}_{\text{lin}}}{\|\mathbf{z}_{\text{lin}}\|}$ is the unit vector pointing in the direction of $\mathbf{z}_{\text{lin}}$. Note by construction $\mathbf{u} \in \text{span}(\{\mathbf{v}\})^\perp$. If

$$\frac{A_{\text{lin}}}{\|\mathbf{z}_{\text{lin}}\|} = O\left(\sqrt{\frac{1-\gamma}{\gamma d}}\right),$$

then by Lemma A.2,

$$\mathbb{P}\left(|\langle \mathbf{n}, \mathbf{u} \rangle| > \sqrt{\frac{\gamma}{1-\gamma}} \frac{A_{\text{lin}}}{\|\mathbf{z}_{\text{lin}}\|}\right) \geq \frac{1}{2}$$

and therefore

$$\mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x})) \geq \frac{1}{4} \mathbb{P}\left(|\langle \mathbf{n}, \mathbf{u} \rangle| > \sqrt{\frac{\gamma}{1-\gamma}} \frac{A_{\text{lin}}}{\|\mathbf{z}_{\text{lin}}\|}\right) \geq \frac{1}{8}.$$

If $A_{\text{lin}} < 0$, then again by the symmetry of the noise

$$\begin{aligned} \mathbb{P}(0 > yf(\mathbf{W}, \mathbf{x})) &\geq \frac{1}{2} \mathbb{P}\left(\langle -\mathbf{n}, \mathbf{z}_{\text{lin}} \rangle > \sqrt{\frac{\gamma}{1-\gamma}} A_{\text{lin}}\right) \\ &\geq \frac{1}{2} \mathbb{P}(\langle -\mathbf{n}, \mathbf{z}_{\text{lin}} \rangle > 0) \\ &= \frac{1}{4}. \end{aligned}$$

This establishes the result. □

Theorem 3.4. *Under the setting given in Assumption 2, let $\delta \in (0, 1)$ and suppose $\mathcal{A} = \mathcal{A}_{GD}$, where $\eta, \lambda \in \mathbb{R}_{>0}$ satisfy Assumption 1. If $n = \Omega(1)$, $d = \Omega\left(n + \log \frac{1}{\delta}\right)$ and $\gamma = O\left(\frac{\alpha^3}{d}\right)$ then the following hold.*

1. *The algorithm $\mathcal{A}_{GD}$ terminates almost surely after finitely many updates. With $\mathbf{W} = \mathcal{A}_{GD}(\mathbf{X}, \hat{\mathbf{y}})$, $L(\mathbf{W}, \mathbf{X}, \hat{\mathbf{y}}) = 0$.*
2. *With probability at least $1 - \delta$ over the training data $(\mathbf{X}, \hat{\mathbf{y}})$*

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) < 0 \mid \mathbf{X}, \hat{\mathbf{y}}) \geq \frac{1}{8}.$$

Proof. If $A_{\text{lin}} < 0$, then by Lemma D.4,

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) < 0) \geq \frac{1}{4}.$$

So it suffices to consider the case $A_{\text{lin}} \geq 0$. Since $d = \Omega\left(n + \log \frac{1}{\delta}\right)$, by Lemma B.4, the max-margin classifier $\mathbf{w}^*$ satisfies

$$\|\mathbf{w}^*\| \leq C \sqrt{\frac{n}{1-\gamma}}$$

with probability at least $1 - \frac{\delta}{3}$ over the randomness of the input dataset. We denote this event by ω_1 and condition on it for the rest of this proof. By Theorem 3.1,

$$\begin{aligned} \|\mathbf{W}\| &\leq \frac{C\|\mathbf{w}^*\|}{\alpha\sqrt{m}} \\ &\leq \frac{C}{\alpha} \sqrt{\frac{n}{m(1-\gamma)}}. \end{aligned}$$

By Theorem 3.1, the network perfectly fits the training data, so for all $i \in [n]$, $\hat{y}_i f(\mathbf{W}, \mathbf{x}_i) \geq 1$, and therefore

$$\begin{aligned}
1 &\leq |f(\mathbf{W}, \mathbf{x}_i)| \\
&= \left| \sum_{j=1}^{2m} (-1)^j \sigma(\langle \mathbf{w}_j, \mathbf{x}_i \rangle) \right| \\
&\leq \sum_{j=1}^{2m} |\langle \mathbf{w}_j, \mathbf{x}_i \rangle| \\
&= \sum_{j=1}^{2m} |\langle a_j \mathbf{v} + \mathbf{z}_j, \sqrt{\gamma} y_i \mathbf{v} + \sqrt{1-\gamma} \mathbf{n}_i \rangle| \\
&= \sum_{j=1}^{2m} |a_j y_i \sqrt{\gamma} + \sqrt{1-\gamma} \langle \mathbf{z}_j, \mathbf{n}_i \rangle| \\
&\leq \sqrt{\gamma} \sum_{j=1}^{2m} |a_j| + \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|.
\end{aligned}$$

This implies that either $\frac{1}{2} \leq \sqrt{\gamma} \sum_{j=1}^{2m} |a_j|$ or $\frac{1}{2} \leq \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|$ for all $i \in [n]$. We consider both cases separately.

Suppose that $\frac{1}{2} \leq \sqrt{\gamma} \sum_{j=1}^{2m} |a_j|$. Then squaring both sides and applying Cauchy-Schwarz, we obtain

$$\begin{aligned}
\frac{1}{4} &\leq \gamma \left(\sum_{j=1}^{2m} |a_j| \right)^2 \\
&\leq 2m\gamma \sum_{j=1}^{2m} |a_j|^2 \\
&\leq 2m\gamma \sum_{j=1}^{2m} \|\mathbf{w}_j\|^2 \\
&= 2m\gamma \|\mathbf{W}\|_F^2 \\
&\leq \frac{C\gamma n}{\alpha^2(1-\gamma)}.
\end{aligned}$$

This cannot occur if $\gamma = O\left(\frac{\alpha^2}{n}\right)$, and in particular it cannot occur if $d = \Omega(n)$ and $\gamma = O\left(\frac{\alpha^3}{d}\right)$.

Now suppose that $\frac{1}{2} \leq \sqrt{1-\gamma} \sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle|$ for all $i \in [n]$. Squaring both sides and applying Cauchy-Schwarz, we obtain

$$\begin{aligned}
\frac{1}{4} &\leq (1-\gamma) \left(\sum_{j=1}^{2m} |\langle \mathbf{z}_j, \mathbf{n}_i \rangle| \right)^2 \\
&\leq 2m(1-\gamma) \sum_{j=1}^{2m} \|\langle \mathbf{z}_j, \mathbf{n}_i \rangle\|^2.
\end{aligned}$$

Summing over all $i \in [n]$, we obtain

$$\begin{aligned}
\frac{n}{4} &\leq 2m(1-\gamma) \sum_{i=1}^n \sum_{j=1}^{2m} \|\langle \mathbf{z}_j, \mathbf{n}_i \rangle\|^2 \\
&= 2m(1-\gamma) \sum_{j=1}^{2m} \|\mathbf{N} \mathbf{z}_j\|^2 \\
&\leq 2m(1-\gamma) \|\mathbf{N}\|^2 \sum_{j=1}^{2m} \|\mathbf{z}_j\|^2 \\
&= 2m(1-\gamma) \|\mathbf{N}\|^2 \|\mathbf{Z}\|_F^2.
\end{aligned}$$

Recall that $d = \Omega(n + \log \frac{1}{\delta})$, and that the rows of $\mathbf{N}$ are i.i.d. with distribution $\mathcal{N}(\mathbf{0}_d, d^{-1}(\mathbf{I}_d - \mathbf{v}\mathbf{v}^T))$. So by Lemma 4.1, with probability at least $1 - \frac{\delta}{3}$ over the randomness of the dataset, $\|\mathbf{N}\| \leq C$. We denote this event by ω_2 and condition on it for the rest of this proof. So

$$\|\mathbf{Z}\|_F^2 \geq \frac{Cn}{m(1-\gamma)}. \quad (22)$$

Let $\lambda = \frac{\sqrt{\alpha}}{m}$. By Assumption 1, $\|\mathbf{w}_j^{(0)}\| \leq \lambda$ for all $j \in [2m]$. So by Lemma D.1,

$$\|\mathbf{Z}\|_F^2 - 2\lambda\sqrt{2m}\|\mathbf{Z}\|_F - 2m\lambda^2 \leq \frac{C}{\alpha m}(\|\mathbf{z}_{\text{lin}}\| + 2m\lambda)^2. \quad (23)$$

By (22),

$$\begin{aligned}
\|\mathbf{Z}\|_F &\geq \frac{C\sqrt{n}}{\sqrt{m(1-\gamma)}} \\
&\geq \frac{C\sqrt{n}}{\sqrt{m}} \\
&\geq C\lambda\sqrt{nm} \\
&\geq 8\lambda\sqrt{2m},
\end{aligned}$$

where the last line holds if $n = \Omega(1)$. Then by (23),

$$\begin{aligned}
\frac{1}{2}\|\mathbf{Z}\|_F^2 &= \|\mathbf{Z}\|_F^2 - \frac{1}{4}\|\mathbf{Z}\|_F^2 - \frac{1}{4}\|\mathbf{Z}\|_F^2 \\
&\leq \|\mathbf{Z}\|_F^2 - 2\lambda\sqrt{2m}\|\mathbf{Z}\|_F - 2m\lambda^2 \\
&\leq \frac{C}{\alpha m}(\|\mathbf{z}_{\text{lin}}\| + 2m\lambda)^2 \\
&= \frac{C}{\alpha m}(\|\mathbf{z}_{\text{lin}}\| + 2\sqrt{\alpha})^2.
\end{aligned}$$

Taking the square root of both sides and recalling that $\alpha \in (0, 1)$ is a constant, we obtain

$$\|\mathbf{Z}\|_F \leq \frac{C\|\mathbf{z}_{\text{lin}}\|}{\sqrt{\alpha m}} + \frac{C}{\sqrt{m}}.$$

This implies that either $\|\mathbf{Z}\|_F \leq \frac{2C}{\sqrt{m}}$ or $\|\mathbf{Z}\|_F \leq \frac{2C\|\mathbf{z}_{\text{lin}}\|}{\sqrt{\alpha m}}$. The case $\|\mathbf{Z}\|_F \leq \frac{2C}{\sqrt{m}}$ cannot happen, since by (22),

$$\begin{aligned}
\|\mathbf{Z}\|_F &\geq \frac{C'\sqrt{n}}{\sqrt{m(1-\gamma)}} \\
&\geq \frac{C'\sqrt{n}}{\sqrt{m}} \\
&\geq \frac{2C}{\sqrt{m}}
\end{aligned}$$

when $n = \Omega(1)$. So we have $\|\mathbf{Z}\|_F \leq \frac{2C\|\mathbf{z}_{\text{lin}}\|}{\sqrt{\alpha m}}$. Again applying (22), we obtain

$$\begin{aligned}\|\mathbf{z}_{\text{lin}}\| &\geq C\sqrt{\alpha m}\|\mathbf{Z}\|_F \\ &\geq \frac{C\sqrt{\alpha n}}{\sqrt{1-\gamma}}.\end{aligned}$$

So

$$\begin{aligned}\frac{A_{\text{lin}}}{\|\mathbf{z}_{\text{lin}}\|} &\leq C \frac{A_{\text{lin}}\sqrt{1-\gamma}}{\sqrt{\alpha n}} \\ &= \frac{C\sqrt{1-\gamma}}{\sqrt{\alpha n}} \sum_{j=1}^{2m} (-1)^j a_j \\ &\leq \frac{C\sqrt{1-\gamma}}{\sqrt{\alpha n}} \sqrt{2m} \left(\sum_{j=1}^{2m} |a_j|^2 \right)^{1/2} \\ &\leq \frac{C\sqrt{m(1-\gamma)}}{\sqrt{\alpha n}} \left(\sum_{j=1}^{2m} \|\mathbf{w}_j\|^2 \right)^{1/2} \\ &= \frac{C\|\mathbf{W}\|_F \sqrt{m(1-\gamma)}}{\sqrt{\alpha n}} \\ &\leq \frac{C}{\alpha^{3/2}}.\end{aligned}$$

Here we used that $A_{\text{lin}} \geq 0$ and applied Cauchy-Schwarz in the third line. Then by Lemma D.4, if

$$\frac{C}{\alpha^{3/2}} \leq O\left(\sqrt{\frac{1-\gamma}{\gamma d}}\right),$$

then

$$\mathbb{P}(yf(\mathbf{W}, \mathbf{x}) < 0) \geq \frac{1}{8}.$$

This occurs if $\gamma = O\left(\frac{\alpha^3}{d}\right)$. Hence, in all cases, we have shown that with the appropriate scaling, the generalization error is at least $\frac{1}{8}$ when both ω_1 and ω_2 occur. This happens with probability at least $1 - \delta$. $\square$

Appendix E Formalizing benign overfitting as a high dimensional phenomenon

To formalize benign overfitting as a high dimensional phenomenon we first introduce the notion of a *regime*. Informally, a regime is a subset of the hyperparameters $\Omega \in \mathbb{N}^4$ which describes accepted combinations of the input data dimension d , the number of points in the training sample n , the number of corrupt points k and the number of trainable model parameters p .

Definition E.1. A regime is a subset $\Omega \subset \mathbb{N}^4$ which satisfies the following properties.

1. For any tuple $(d, n, k, p) \in \Omega$ the number of corrupt points is at most the total number of points, $k \leq n$.
2. There is no upper bound on the number of points,

$$\sup_{(d,n,k,p) \in \Omega} n = \infty.$$

A non-trivial regime is a regime which satisfies the following additional condition.

3. Define the set of increasing sequences of Ω as $\Omega^* = \{(n_l, d_l, k_l, p_l)_{l \in \mathbb{N}} \subset \Omega \text{ s.t. } \lim_{l \rightarrow \infty} n_l = \infty\}$. For any $(n_l, d_l, k_l, p_l)_{l \in \mathbb{N}} \in \Omega^*$ it holds that

$$\liminf_{l \rightarrow \infty} \frac{k_l}{n_l} > 0.$$

Intuitively, a regime defines how the four hyperparameters (d, n, k, p) can grow in relation to one another as n goes to infinity. A non-trivial regime is one in which the fraction of corrupt points in the training sample is non-vanishing. In order to make a formal definition of benign overfitting as high dimensional phenomenon we introduce the following additional concepts.

- A learning algorithm $\mathcal{A} = (\mathcal{A}_{d,n,p})_{(d,n,p) \in \mathbb{N}^3}$ is a triple indexed sequence of measurable functions $\mathcal{A}_{d,n,p} : \mathbb{R}^{n \times d} \times \mathbb{R}^n \rightarrow \mathbb{R}^p$.
- An architecture $\mathcal{M} = (f_{d,p})_{d,p \in \mathbb{N}^2}$ is a double indexed sequence of measurable functions $f_{d,p} : \mathbb{R}^d \times \mathbb{R}^p \rightarrow \mathbb{R}$.
- A data model $\mathcal{D} = (D_{d,n,k})_{(d,n,k) \in \mathbb{N}^3}$ is a triple indexed sequence of Borel probability measures $D_{d,n,k}$ defined over $\mathbb{R}^{n \times d} \times \{\pm 1\} \times \mathbb{R}^d \times \{\pm 1\}$.

With these notions in place we are ready to provide a definition of benign overfitting in high dimensions.

Definition E.2. Let $(\epsilon, \delta) \in (0, 1]^2$, $\mathcal{A}$ be a learning algorithm, $\mathcal{M}$ an architecture, $\mathcal{D}$ a data model and Ω a regime. If for every increasing sequence $(d_l, n_l, k_l, p_l)_{l \in \mathbb{N}} \in \Omega^*$ there exists an $L \in \mathbb{N}$ such that for all $l \geq L$ with probability at least $1 - \delta$ over $(\mathbf{X}, \hat{\mathbf{y}})$, where $(\mathbf{X}, \hat{\mathbf{y}}, \mathbf{x}, y) \sim D_{d_l, n_l, k_l}$, it holds that

1. $y_i f(\mathcal{A}_{d_l, n_l, p_l}(\mathbf{X}, \hat{\mathbf{y}}), \mathbf{x}_i) > 0 \quad \forall i \in [n_l]$,
2. $\mathbb{P}(y f(\mathcal{A}_{d_l, n_l, p_l}(\mathbf{X}, \hat{\mathbf{y}}), \mathbf{x}) \leq 0) \leq \inf_{\mathbf{W} \in \mathbb{R}^{n_l \times d_l}} \mathbb{P}(y f(\mathbf{W}, \mathbf{x}) \leq 0) + \epsilon$,

then the quadruplet $(\mathcal{A}, \mathcal{M}, \mathcal{D}, \Omega)$ (ϵ, δ) -benignly overfits. If $(\mathcal{A}, \mathcal{M}, \mathcal{D}, \Omega)$ (ϵ, δ) -benignly overfits for any $(\epsilon, \delta) \in (0, 1]^2$ then we say $(\mathcal{A}, \mathcal{M}, \mathcal{D}, \Omega)$ benignly overfits.

Analogously, we define non-benign overfitting as follows.

Definition E.3. Let $(\epsilon, \delta) \in (0, 1]^2$, $\mathcal{A}$ be a learning algorithm, $\mathcal{M}$ an architecture, $\mathcal{D}$ a data model and Ω a regime. If for every increasing sequence $(d_l, n_l, k_l, p_l)_{l \in \mathbb{N}} \in \Omega^*$ there exists an $L \in \mathbb{N}$ such that for all $l \geq L$ with probability at least $1 - \delta$ over $(\mathbf{X}, \hat{\mathbf{y}})$, where $(\mathbf{X}, \hat{\mathbf{y}}, \mathbf{x}, y) \sim D_{d_l, n_l, k_l}$, it holds that

1. $y_i f(\mathcal{A}_{d_l, n_l, p_l}(\mathbf{X}, \hat{\mathbf{y}}), \mathbf{x}_i) > 0 \quad \forall i \in [n_l]$,
2. $\mathbb{P}(y f(\mathcal{A}_{d_l, n_l, p_l}(\mathbf{X}, \hat{\mathbf{y}}), \mathbf{x}) \leq 0) \geq \inf_{\mathbf{W} \in \mathbb{R}^{n_l \times d_l}} \mathbb{P}(y f(\mathbf{W}, \mathbf{x}) \leq 0) + \epsilon$,

then the quadruplet $(\mathcal{A}, \mathcal{M}, \mathcal{D}, \Omega)$ (ϵ, δ) -non-benignly overfits. If $(\mathcal{A}, \mathcal{M}, \mathcal{D}, \Omega)$ (ϵ, δ) -non-benignly overfits for any $(\epsilon, \delta) \in (0, 1]^2$ then we say $(\mathcal{A}, \mathcal{M}, \mathcal{D}, \Omega)$ non-benignly overfits.

One of the key contributions of this paper is proving (ϵ, δ) -benign and non-benign overfitting when the architecture is a two-layer leaky ReLU network (equation 1), the learning algorithm returns the inner layer weights of the network by minimizing the hinge loss over the training data using gradient descent (Definition 2.2), and the regime satisfies the conditions $d = \Omega(n \log 1/\epsilon)$, $n = \Omega(1/\delta)$, $k = O(n)$ and $p = 2dm$ for some network width $2m$, $m \in \mathbb{N}$.

Appendix F Experiments

To further support our theory, we train shallow neural networks on the data model described in Definition 2.1 and record the numerical results. Scripts to reproduce these experiments can be found at https://github.com/kedar2/benign_overfitting. These experiments were run on the CPU of a MacBook Pro M2 with 8GB of RAM. For our first experiment, we investigate the effect of the ratio $\frac{d}{n}$ on the generalization error of the network. Recall that by Theorem 3.2, the generalization

error is bounded above by $\exp(-Cn)$ when $\frac{d}{n} = \Omega(1)$. In other words, if $\frac{d}{n}$ is larger than a critical threshold, then the generalization error decays quickly to 0 as n increases. We empirically confirm this prediction in Figure 1, where we train several networks while varying $\frac{d}{n}$ and n , and estimate the generalization error for each configuration by averaging over 20 trials. Within each trial, we trained the inner layer of the network with gradient descent using the hinge loss until the training loss reached 0. For $\frac{d}{n}$ greater than around 7, the generalization error rapidly decays to 0 as $n \rightarrow \infty$.

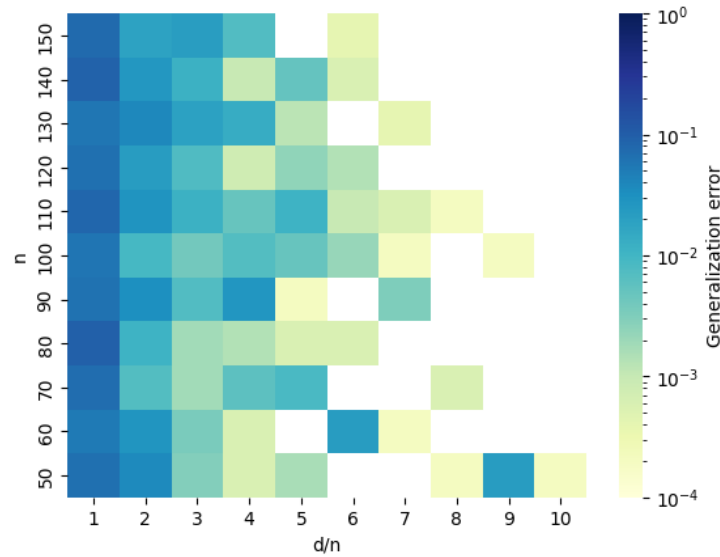


Figure 1: Generalization error of a two-layer leaky ReLU network trained to 0 hinge loss varying n and d . Parameter settings: $\alpha = 0.1$, $\gamma = 5/n$, $m = 64$, $k = 0.1n$, number of trials = 5, size of validation sample = 1000.

Next, we train a two-layer network varying n and γ (Figure 2), holding constant the ratio $\frac{d}{n}$. Since γ controls the signal-to-noise ratio of the data, the generalization error of the learned network decreases as γ increases. For each value of n , the generalization error falls off steeply as γ reaches a certain threshold. This threshold decreases as n increases, indicating that the network has higher noise tolerance as n increases. This is in agreement with our theoretical results where we found that benign overfitting occurs at the threshold $\gamma = \Omega\left(\frac{1}{k}\right)$ (which is in this case $\Omega\left(\frac{1}{n}\right)$). We also see that the generalization error for large values of γ is similar across different values of n . This effect is also predicted by Corollary 3.2.1, since we scale both d and k proportionally to n .

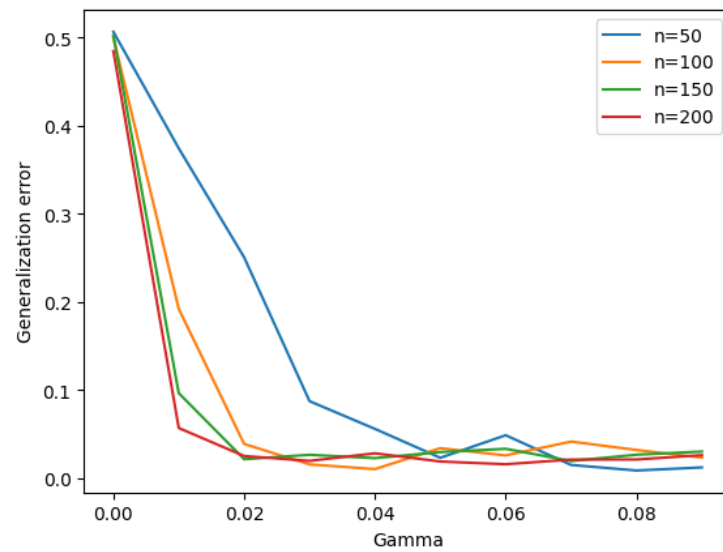


Figure 2: Generalization error of a two-layer leaky ReLU network trained to 0 hinge loss varying γ and n . Parameter settings: $\alpha = 0.1$, $d = 2n$, $m = 64$, $k = 0.1n$, number of trials = 10, size of validation sample = 1000.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The results stated in the abstract and introduction are stated formally in Section 3 and then proven in Appendix D.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss limitations of the work in the conclusion of the paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All necessary assumptions are stated in the theorems. All theorems are proven in full detail in Appendices C and D, with proof sketches appearing in Section 4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We describe our experimental setup in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a link to our code to reproduce our experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We describe our setup in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our experiments consist of a heatmap and a line chart with multiple plots. It was not possible to add error bars without crowding the plots. In our description of our experimental setup we describe the sample size, from which standard errors can be computed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We describe the resources used in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper did not use data, assets, or human participants; only studied existing models; and was conducted in accordance to the Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper is a purely theoretical study explaining behaviors seen in neural networks in practice. There are no foreseeable societal impacts of this work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No new model or data is presented in this paper. The paper is a theoretical study of neural networks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not use new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.