Provable Generalization of SGD-trained Neural Networks of Any Width in the Presence of Adversarial Label Noise

Spencer Frei [1] **Yuan Cao** [2] **Quanquan Gu** [2]

Abstract

We consider a one-hidden-layer leaky ReLU network of arbitrary width trained by stochastic gradient descent (SGD) following an arbitrary initialization. We prove that SGD produces neural networks that have classification accuracy competitive with that of the best halfspace over the distribution for a broad class of distributions that includes log-concave isotropic and hard margin distributions. Equivalently, such networks can generalize when the data distribution is linearly separable but corrupted with adversarial label noise, despite the capacity to overfit. To the best of our knowledge, this is the first work to show that overparameterized neural networks trained by SGD can generalize when the data is corrupted with adversarial label noise.

1. Introduction

The remarkable ability of neural networks trained by stochastic gradient descent (SGD) to generalize, even when trained on data that has been substantially corrupted with random noise, seems at ends with much of contemporary statistical learning theory (Zhang et al., 2017). How can a model class which is rich enough to fit randomly labeled data fail to overfit when a significant amount of random noise is introduced into the labels? And how is it that a local optimization method like SGD is so successful at learning such model classes, even when the optimization problem is highly non-convex?

In this paper, we approach these questions by analyzing the performance of SGD-trained networks on distributions which can have substantial amounts of label noise. For a distribution $\mathcal{D}$ over features $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$, let us define

$$\mathsf{OPT} := \min_{v \in \mathbb{R}^d,\ \|v\|=1} \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \operatorname{sgn}\big(\langle v, x\rangle\big)\Big) \quad (1)$$

as the optimal classification error achieved by a halfspace $\langle v, \cdot\rangle$. We prove that for a broad class of distributions, SGD-trained one-hidden-layer neural networks achieve classification error at most $\tilde{O}(\sqrt{\mathsf{OPT}})$ in polynomial time. Equivalently, one-hidden-layer neural networks can learn halfspaces up to risk $\tilde{O}(\sqrt{\mathsf{OPT}})$ in the distribution-specific agnostic PAC learning setting. Our result holds for neural networks with leaky-ReLU activations trained on the cross-entropy loss and, importantly, hold for any initialization, and for networks of arbitrary width.

By comparing the generalization of the neural network with that of the *best* linear classifier over the distribution, we can make two different but equally important claims about the training of overparameterized neural networks. The first view is that SGD produces neural networks with classification error that is competitive with that of the best linear classifier over the distribution, and that this behavior can occur for neural networks of any width and any initialization. In this view, our work provides theoretical support for the hypothesis put forward by Nakkiran et al. (2019) that the performance of SGD-trained networks in the early epochs of training can be explained by that of a linear classifier.

The second view is that of the problem of learning halfspaces in the presence of adversarial label noise. (Note that adversarial *label noise* is distinct from the notions of adversarial examples or adversarial training (Goodfellow et al., 2014; Madry et al., 2018), where the features x are perturbed rather than the labels y.) In this setting, one views the (clean) data as initially coming from a linearly separable distribution but for which each sample $(x, y) \sim \mathcal{D}$ has its label flipped $y \mapsto -y$ with some sample-dependent probability $p(x) \in [0, 1]$. Then the best error achieved by a halfspace is $\mathbb{E}_{x\sim\mathcal{D}_x}[p(x)] = \mathsf{OPT}$. Viewed from this perspective, our result shows that despite the clear capacity of an overparameterized neural network to overfit to corrupted labels, when trained by SGD, such networks can still generalize (albeit achieving the suboptimal risk $\sqrt{\mathsf{OPT}}$). We note that the optimization algorithm we consider is vanilla online SGD without any explicit regularization methods such

[1]Department of Statistics, UCLA [2]Department of Computer Science, UCLA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

as weight decay or dropout. This suggests that the ability of neural networks to generalize in the presence of noise is not solely due to explicit regularization, but that some forms of *implicit* regularization induced by gradient-based optimization play an important role.

1.1. Related Work

We discuss here a number of works related to the questions of optimization and generalization in deep learning. An approach that has attracted significant attention recently is the neural tangent kernel (NTK) approximation (Jacot et al., 2018). This approximation relies upon the fact that for a specific initialization scheme, extremely wide neural networks are well-approximated by the behavior of the neural network at initialization, which in the infinite width limit produces a kernel (the NTK) (Du et al., 2019; 2018; Allen-Zhu et al., 2019; Zou et al., 2019; Cao & Gu, 2020; Arora et al., 2019a;b; Cao & Gu, 2019; Frei et al., 2019; Zou & Gu, 2019; Ji & Telgarsky, 2020b; Chen et al., 2019). Using an assumption on separability of the training data, it is commonly shown that SGD-trained neural networks in the NTK regime can perfectly fit any training data. Under certain conditions, one can also derive generalization bounds for the performance of SGD-trained networks for distributions that can be perfectly classified by functions related to the NTK.

Although significant insights into the training dynamics of SGD-trained networks have come from this approach, it is known that neural networks deployed in practice can traverse far enough from their initialization such that the NTK approximation no longer holds (Fort et al., 2020). A line of work known as the mean field approximation allows for ultra-wide networks to be far from their initialization by connecting the trajectory of the weights of the neural network to the solution of an associated partial differential equation (Mei et al., 2019; Chizat et al., 2019; Chen et al., 2020). A separate line of work has sought to demonstrate that the concept classes that can be learned by neural networks trained by gradient descent are a strict superset of those that can be learned by the NTK (Allen-Zhu & Li, 2019; Wei et al., 2019; Li et al., 2019b; Woodworth et al., 2020; Li et al., 2020a).

More relevant to our work is understanding the generalization of neural network classifiers when the data distribution has some form of label noise. Works that explicitly derive generalization bounds for SGD-trained neural networks in the presence of label noise are scarce. Even for the simple concept class of halfspaces $x \mapsto \mathrm{sgn}(\langle v, x\rangle)$, there are often tremendous difficulties in determining whether or not *any* algorithm can efficiently learn in the presence of noise. For this reason let us take a small detour to detail some of the difficulties in learning halfspaces in the presence of noise, to emphasize the difficulty of learning more complicated function classes in the presence of noise.

The most general (and most difficult) noise class is that of adversarial label noise, which is equivalent to the agnostic PAC learning framework (Kearns et al., 1994). In this setting, one makes no assumption on the relationship between the features and the labels, and so continuing with the notation from (1), the optimal risk OPT achieved by a halfspace is strictly positive in general. It is known that learning up to classification error $O(\mathsf{OPT}) + \varepsilon$ cannot be done in $\mathrm{poly}(d, \varepsilon^{-1})$ time without assumptions on the marginal distribution of $\mathcal{D}$ (Daniely, 2016). For this reason it is common to assume some type of structure on the noise or the distribution to get tractable guarantees.

One relaxation of the noise condition is known as the Massart noise (Massart et al., 2006) where one assumes that each sample has its label flipped with some instance-dependent probability $p(x) \leq p < 1/2$. Under this noise model, it was recently shown that there are efficient algorithms that can learn up to risk $p + \varepsilon$ (Diakonikolas et al., 2019). A more simple noise setting is that of random classification noise (RCN) (Angluin & Laird, 1988), where the labels of each sample are flipped with probability p. Polynomial time algorithms for learning under this model were first shown by Blum et al. (1998). Previous theoretical works on the ability of neural network classifiers to generalize in the presence of label noise were restricted to the RCN setting (Hu et al., 2020a) or Massart noise setting (Li et al., 2019a). In this paper, we consider the most general setting of adversarial label noise.

In terms of distribution-specific learning guarantees in the presence of noise, polynomial time algorithms for learning halfspaces under Massart noise for the uniform distribution on the sphere were first shown by Awasthi et al. (2015), and for log-concave isotropic distributions by Awasthi et al. (2016). Awasthi et al. (2017) constructed a localization-based algorithm that efficiently learns halfspaces up to risk $O(\mathsf{OPT})$ when the marginal is log-concave isotropic. For more background on learning halfspaces in the presence of noise, we refer the reader to Balcan & Haghtalab (2021).

Returning to the neural network literature, in light of the above it should not be surprising that computational tractability issues arise even for the case of neural networks consisting of a single neuron. Goel et al. (2019) showed that learning a single ReLU neuron up to the best-possible risk $\mathsf{OPT}_{\mathrm{ReLU}}$ (under the squared loss) is computationally intractable, even when the marginal is a standard Gaussian. By contrast, Frei et al. (2020) showed that gradient descent on the empirical risk can learn single ReLUs up to risk $O(\sqrt{\mathsf{OPT}_{\mathrm{ReLU}}})$ efficiently for many distributions. Two recent works have shown that even in the realizable setting—i.e., when the labels are generated by a neural net-

work without noise—it is computationally hard to learn one-hidden-layer neural networks with (non-stochastic) gradient descent when the marginal distribution is Gaussian (Goel et al., 2020; Diakonikolas et al., 2020a).

In terms of results that show neural networks can generalize in the presence of noise, Li et al. (2019a) considered clustered distributions with real-valued labels (using the squared loss) and analyzed the performance of GD-trained one-hidden-layer neural networks when a fraction of the labels are switched. They derived guarantees for the empirical risk but did not derive a generalization bound for the resulting classifier. Hu et al. (2020a) analyzed the performance of regularized neural networks in the NTK regime when trained on data with labels corrupted by RCN, and argued that regularization was helpful for generalization. By contrast, our work shows that neural networks can generalize for linearly separable distributions corrupted by adversarial label noise without any explicit regularization, suggesting that certain forms of implicit regularization in the choice of the algorithm plays an important role. We note that a number of researchers have sought to understand the implicit bias of gradient descent (Soudry et al., 2018; Ji & Telgarsky, 2019; Lyu & Li, 2020; Ji & Telgarsky, 2020a; Moroshko et al., 2020; Li et al., 2020b). Such works assume that the distribution is linearly separable by a large margin, and characterize the solutions found by gradient descent (or gradient flow) in terms of the maximum margin solution.

Finally, we note some recent works that connected the training dynamics of SGD-trained neural networks with linear models. Brutzkus et al. (2018) showed that SGD-trained one-hidden-layer leaky ReLU networks can generalize on linearly separable data. Shamir (2018) compared the performance of residual networks with those of linear predictors in the regression setting. They showed that there exist weights for residual networks with generalization performance competitive with linear predictors, and they proved that SGD is able to find those weights when there is a residual connection from the input layer to the output layer. Nakkiran et al. (2019) provided experimental evidence for the hypothesis that much of the performance of SGD-trained neural networks in the early epochs of training can be explained by linear classifiers. Hu et al. (2020b) provided theoretical evidence for this hypothesis by showing that overparameterized neural networks with the NTK initialization and scaling have similar dynamics to a linear predictor defined in terms of the network's NTK. Shah et al. (2020) showed that neural networks are biased towards simple classifiers even when more complex classifiers are capable of improving generalization.

2. Problem Description and Results

In this section we study the problem we consider and our main results.

2.1. Notation

For a vector v, we denote $\|v\|$ as its Euclidean norm. For a matrix W, we use $\|W\|_F$ to denote its Frobenius norm. We use the standard $O(\cdot)$ and $\Omega(\cdot)$ notations to ignore universal constants when describing growth rates of functions. The notation $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ further ignores logarithmic factors. We use $a \vee b$ to denote the maximum of $a, b \in \mathbb{R}$, and $a \wedge b$ their minimum. The notation $\mathbb{1}(E)$ denotes the indicator function of the set E, which is one on the set and zero outside of it.

2.2. Problem Setup

Consider a distribution $\mathcal{D}$ over $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ with marginal distribution $\mathcal{D}_x$ over x. Let $m \in \mathbb{N}$, and consider a one-hidden-layer leaky ReLU network with m neurons,

$$f_x(W) := \sum_{j=1}^{m} a_j \sigma(\langle w_j, x \rangle), \tag{2}$$

where $\sigma(z) = \max(\alpha z, z)$ is the leaky-ReLU activation with $\alpha \in (0, 1]$. Assume that $a_j \overset{\text{i.i.d.}}{\sim} \text{Unif}(\pm a)$ for some $a > 0$ and that the $\{a_j\}$ are randomly initialized and not updated throughout training, as is commonly assumed in theoretical analyses of SGD-trained neural networks (Du et al., 2019; Arora et al., 2019b; Ji & Telgarsky, 2020b).[1] We are interested in the classification error for the neural network,

$$\text{err}(W) := \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \text{sgn}\left(f_x(W)\right)\Big),$$

where $\text{sgn}(z) = 1$ if $z > 0$, $\text{sgn}(0) = 0$, and $\text{sgn}(z) = -1$ otherwise. We will seek to minimize $\text{err}(W)$ by minimizing,

$$L(W) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(y f_x(W)),$$

where ℓ is a convex loss function. We will use the fact that for any convex, twice differentiable and decreasing function ℓ, the function $-\ell'$ is non-negative and decreasing, and thus $-\ell'$ can also serve as a loss function. In particular, by Markov's inequality, these properties allow us to bound

[1] The specific choice of the initialization of the second layer is immaterial; our analysis holds for any second-layer weights that are fixed at a random initialization. The only difference that may arise is in the sample complexity: if with high probability $\|a\| = \Theta(1)$ then the sample complexity requirement will be the same within constant factors, while for initializations satisfying $\|a\| = \omega(1)$ or $\|a\| = o(1)$ our upper bound for the sample complexity will become worse as the network becomes larger.

the classification error by the population risk under $-\ell'$:

$$\begin{aligned}&\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \operatorname{sgn}\big(f_x(W)\big)\Big) = \mathbb{P}\Big(y \cdot f_x(W) \leq 0\Big)\\&= \mathbb{P}\Big(-\ell'\big(yf_x(W)\big) \geq -\ell'(0)\Big)\\&\leq \frac{\mathbb{E}_{(x,y)\sim\mathcal{D}} - \ell'\big(yf_x(W)\big)}{-\ell'(0)} \qquad (3)\end{aligned}$$

Thus, provided $-\ell'(0) > 0$, upper bounds for the population risk under $-\ell'$ yield guarantees for the classification error. This property has previously been used to derive generalization bounds for deep neural networks trained by gradient descent (Cao & Gu, 2020; Frei et al., 2019; Ji & Telgarsky, 2020b; Chen et al., 2019). To this end, we make the following assumptions on the loss throughout this paper.

Assumption 2.1. *The loss $\ell(\cdot) : \mathbb{R} \to \mathbb{R}$ is convex, twice differentiable, decreasing, 1-Lipschitz, and satisfies $-\ell'(0) > 0$. Moreover, for $z \geq 1$, ℓ satisfies $-\ell'(z) \leq 1/z$.*

The assumption that $-\ell'(z) \leq 1/z$ for $z \geq 1$ is to ensure that the surrogate loss $-\ell'$ is not too large on samples that are classified correctly. Note that the standard loss used for training neural networks in binary classification tasks—the binary cross-entropy loss $\ell(z) = \log(1 + \exp(-z))$—satisfies all of the conditions in Assumption 2.1. We denote the population risk under the surrogate loss $-\ell'$ as follows,

$$\mathcal{E}(W) := \mathbb{E}_{(x,y)\sim\mathcal{D}} - \ell'(yf_x(W)).$$

We seek to minimize the population risk by minimizing the empirical risk induced by a set of i.i.d. examples $\{(x_t, y_t)\}_{t\geq 1}$ using the online stochastic gradient descent algorithm. Denote $f_t(W) = f_{x_t}(W)$ as the neural network output for sample x_t, and denote the loss under ℓ and $-\ell'$ for sample x_t by

$$\widehat{L}_t(W) := \ell(y_t f_t(W)), \quad \widehat{\mathcal{E}}_t(W) := -\ell'(y_t f_t(W)). \quad (4)$$

The updates of online stochastic gradient descent are given by

$$W^{(t+1)} := W^{(t)} - \eta\nabla\widehat{L}_t(W^{(t)}).$$

Before proceeding with our main theorem we will introduce some of the definitions and assumptions which will be used in our analysis. The first is that of sub-exponential distributions.

Definition 2.2 (Sub-exponential distributions)**.** *We say $\mathcal{D}_x$ is C_m-*sub-exponential *if every $x \sim \mathcal{D}_x$ is a sub-exponential random vector with sub-exponential norm at most C_m. In particular, for any $\bar{v} \in \mathbb{R}^d$ with $\|\bar{v}\| = 1$, $\mathbb{P}_{\mathcal{D}_x}(|\bar{v}^\top x| \geq t) \leq \exp(-t/C_m)$.*

We note that every sub-Gaussian distribution is sub-exponential. The next property we introduce is that of a *soft margin*. This condition was recently utilized by Frei et al. (2021) and Zou et al. (2021) for the agnostic learning of halfspaces using convex surrogates for the zero-one loss.

Definition 2.3. *Let $\bar{v} \in \mathbb{R}^d$ satisfy $\|\bar{v}\| = 1$. We say $\bar{v}$ satisfies the* soft margin condition *with respect to a function $\phi_{\bar{v}} : \mathbb{R} \to \mathbb{R}$ if for all $\gamma \in [0, 1]$, it holds that*

$$\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathbb{1}\left(|\bar{v}^\top x| \leq \gamma\right)\right] \leq \phi_{\bar{v}}(\gamma).$$

The soft margin can be seen as a probabilistic analogue of the standard hard margin, where we relax the typical requirement for a margin-based condition from holding almost surely to holding with some controlled probability. As written above, the soft margin condition can hold for a specific vector $\bar{v} \in \mathbb{R}^d$, and our final generalization bound below will only care about the soft margin function for a halfspace $\bar{v}$ that achieves population risk OPT. However, for many distributions, one can show that *all* unit norm vectors $\bar{v}$ satisfy a soft margin of the form $\phi_{\bar{v}}(\gamma) = O(\gamma)$. One important class of such distributions are those satisfying a type of anti-concentration property.

Definition 2.4 (Anti-concentration)**.** *For $\bar{v} \in \mathbb{R}^d$, denote by $p_{\bar{v}}(\cdot)$ the marginal distribution of $x \sim \mathcal{D}_x$ on the subspace spanned by $\bar{v}$. We say $\mathcal{D}_x$ satisfies U-*anti-concentration *if there is some $U > 0$ such that for all unit norm $\bar{v}$, $p_{\bar{v}}(z) \leq U$ for all $z \in \mathbb{R}$.*

Anti-concentration is a typical assumption used for deriving distribution-specific agnostic PAC learning guarantees (Klivans et al., 2009; Diakonikolas et al., 2020b;c; Frei et al., 2021) as it allows for one to ignore pathological distributions where arbitrarily large probability mass can be concentrated in tiny regions of the domain. Below, we collect some examples of soft margin function behavior for different distributions, including those satisfying the above anti-concentration property. We shall see in Theorem 2.6 that the behavior of $\phi(\gamma)$ for $\gamma \ll 1$ will be the determining factor in our generalization bound, and thus in the below examples one only needs to pay attention to the behavior of $\phi(\gamma)$ for γ sufficiently small.

Example 2.5.

1. *If $|\bar{v}^\top x| > \gamma^*$ a.s., then $\phi_{\bar{v}}(\gamma) = 0$ for $\gamma < \gamma^*$.*
2. *If $\mathcal{D}_x$ satisfies U-anti-concentration, then for any $\bar{v}$ with $\|\bar{v}\| = 1$, $\phi_{\bar{v}}(\gamma) \leq 2U\gamma$ holds.*
3. *If $\mathcal{D}_x$ is isotropic and log-concave (i.e. its probability density function is log-concave), then $\mathcal{D}_x$ satisfies 1-anti-concentration and hence $\phi_{\bar{v}}(\gamma) \leq 2\gamma$ for all $\bar{v}$.*

The proofs for the properties described in Example 2.5 can be found in Frei et al. (2021, Section 3).

2.3. Main Results

With the above in place, we can provide our main result.

Theorem 2.6. *Assume $\mathcal{D}_x$ is C_m-subexponential and there exists $B_X > 0$ such that $\mathbb{E}[\|x\|^2] \leq B_X^2 < \infty$. Denote $\mathsf{OPT} := \min_{\|w\|=1} \mathbb{P}_{(x,y)\sim\mathcal{D}}(y\langle w, x\rangle < 0)$ as the best classification error achieved by a unit norm halfspace v^*. Let $m \in \mathbb{N}$ be arbitrary, and consider a leaky-ReLU network of the form (2) where $a = 1/\sqrt{m}$. Let $W^{(0)}$ be an arbitrary initialization and denote $G_0 := \|W^{(0)}\|_F$. Let the step size satisfy $\eta \leq B_X^{-2}$. Then for any $\gamma > 0$, by running online SGD for $T = O(\eta^{-1}\gamma^{-2}[\phi_{v^*}(\gamma) + \mathsf{OPT}]^{-2}[1 \vee G_0])$ iterations, there exists a point $t^* < T$ such that in expectation over $(x_1, \ldots, x_T) \sim \mathcal{D}^T$,*

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \operatorname{sgn}\left(f_x(W^{(t^*)})\right)\Big) \leq 2|\ell'(0)|^{-1}\alpha^{-1}\phi_{v^*}(\gamma) + \left[\Big(1 + \gamma^{-1}C_m + \gamma^{-1}C_m \log(1/\mathsf{OPT})\Big)\mathsf{OPT}\right].$$

To concretize the generalization bound in Theorem 2.6 we need to analyze the properties of the soft margin function ϕ_{v^*} at the best halfspace and then optimize over the choice of γ. But before doing so, let us make a few remarks on Theorem 2.6 that hold in general. The sample complexity (number of SGD iterations) T, and the resulting generalization bound, are independent of the number of neurons m, showing that the neural network can generalize despite the capacity to overfit.[2] If $\|x\| \leq B_X$ a.s. for some absolute constant B_X, then the sample complexity is dimension-independent, while if $\mathcal{D}_x$ is isotropic, $\mathbb{E}[\|x\|^2] = d$ and so the sample complexity is linear in d. Finally, we note that large learning rates and arbitrary initializations are allowed.

In the remainder of the section, we will discuss the implications of Theorem 2.6 for common distributions. The first distribution we consider is a hard margin distribution.

Corollary 2.7 (Hard margin distributions)**.** *Suppose there exists some $v^* \in \mathbb{R}^d$, $\|v^*\| = 1$, and $\gamma_0 > 0$ such that $\mathbb{P}\big(y \neq \operatorname{sgn}(\langle v^*, x\rangle)\big) = \mathsf{OPT}$ and $|\langle v^*, x\rangle| \geq \gamma_0 > 0$ almost surely over $\mathcal{D}_x$. Assume for simplicity that ℓ is the binary cross-entropy loss, $\ell(z) = \log(1 + \exp(-z))$. Then under the settings of Theorem 2.6, there exists some $t^* < T = O(\eta^{-1}\gamma_0^{-2}\mathsf{OPT}^{-2}[1 \vee G_0])$ such that in expectation over $(x_1, \ldots, x_T) \sim \mathcal{D}^T$,*

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \operatorname{sgn}\left(f_x(W^{(t^*)})\right)\Big) \leq \tilde{O}(\gamma_0^{-1}\mathsf{OPT}).$$

The proof of Corollary 2.7 can be found in Appendix A and follows from Theorem 2.6 $\gamma = \gamma_0$ and using $\phi_{v^*}(\gamma_0) = 0$.

[2] Brutzkus et al. (2018, Theorem 7) showed that if there are T samples and $m = \Omega(T/d)$, then for any set of labels $(y_1, \ldots, y_T) \in \{\pm 1\}^T$ and for almost every $(x_1, \ldots, x_T) \sim \mathcal{D}_x^T$, there exist hidden layer weights W^* and outer layer weights $\vec{a} \in \mathbb{R}^m$ such that $f_t(W^*) = y_t$ for all $t \in [T]$. In contrast, Theorem 2.6 shows that when m is sufficiently large there exist neural networks that can fit random labels of the data but SGD training avoids these networks.

The above result shows that if the data comes from a linearly separable data distribution with margin γ_0 but is then corrupted by adversarial label noise, then SGD-trained networks will still find weights that can generalize with classification error at most $\tilde{O}(\gamma_0^{-1}\mathsf{OPT})$. In the next corollary we show that for distributions satisfying U-anti-concentration we get a generalization bound of the form $\tilde{O}(\sqrt{\mathsf{OPT}})$.

Corollary 2.8 (Distributions satisfying anti-concentration)**.** *Assume $\mathcal{D}_x$ satisfies U-anti-concentration. Assume for simplicity that ℓ is the binary cross-entropy loss, $\ell(z) = \log(1 + \exp(-z))$. Then under the settings of Theorem 2.6, there exists some $t^* < T = O(\eta^{-1}\mathsf{OPT}^{-3}[1 \vee G_0])$ such that in expectation over $(x_1, \ldots, x_T) \sim \mathcal{D}^T$,*

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \operatorname{sgn}\left(f_x(W^{(t^*)})\right)\Big) \leq \tilde{O}(\sqrt{\mathsf{OPT}}).$$

The proof of Corollary 2.8 can be foudn in Appendix A and follows by taking $\gamma = \mathsf{OPT}^{1/2}$ in Theorem 2.6 and using that $\phi_v(\gamma) = O(\gamma)$ for distributions satisfying U-anti-concentration. The above corollary covers, for instance, log-concave isotropic distributions like the Gaussian or the uniform distribution over a convex set by Example 2.5.

Taken together, Corollaries 2.7 and 2.8 demonstrate that despite the capacity for overparameterized neural networks to overfit to the data, SGD-trained neural networks are fairly robust to adversarial label noise. We emphasize that our results hold for SGD-trained neural networks of arbitrary width and following an arbitrary initialization, and that the resulting generalization and sample complexity do not depend on the number of neurons m. In particular, the above phenomenon cannot be explained by the neural tangent kernel approximation, which is highly dependent on assumptions about the initialization, learning rate, and number of neurons.

2.4. Comparisons with Related Work

We now discuss how our result relates to others appearing in the literature. First, Brutzkus et al. (2018) showed that by running multiple-pass SGD on the hinge loss one can learn linearly separable data. They assume a noiseless ($\mathsf{OPT} = 0$) model over a norm-bounded domain and assume a hard margin distribution, so that $y\langle v^*, x\rangle > \gamma_0$ for some $\gamma_0 > 0$. In the noiseless setting, Corollaries 2.7 and 2.8 generalize their result to include unbounded, linearly separable (marginal) distributions without a hard margin like log-concave isotropic distributions. More significantly, our results hold in the adversarial label noise setting (a.k.a., agnostic PAC learning). This allows for us to compare the generalization of an SGD-trained neural network with that of the *best* linear classifier over the distribution, and make a much more general claim about the dynamics of SGD-trained neural networks.

Hu et al. (2020b) showed that for sufficiently wide neural networks with the NTK initialization scheme, and under the assumption that the components of the input distribution are independent, the dynamics in the early stages of SGD-training are closely related to that of a linear predictor defined in terms of the NTK of the neural network. By contrast, our result holds for any initialization and neural networks of any width and covers a larger class of distributions. Their result was for the squared loss, while ours holds for the standard losses used for classification problems. Our results can be understood as a claim about the 'early training dynamics' of SGD, since we show that there exists *some* iterate of SGD that performs almost as well as the best linear classifier over the distribution, and we provide an upper bound on the number of iterations required to reach this point. One might expect that under more stringent assumptions (on, say, the initialization, learning rate schedule, and/or network architecture), stronger guarantees for the classification error could hold in the later stages of training; we will revisit this question with experimental results in Section 4.

Li et al. (2019a) considered a handcrafted distribution consisting of noisy clusters and showed that sufficiently wide one-hidden-layer neural networks trained by GD on the squared loss with the NTK initialization have favorable properties in the early training dynamics. A direct comparison of our results is difficult as they do not provide a guarantee for the generalization error of the resulting neural network. But at a high level, their analysis focused on a noise model akin to Massart noise (a more restrictive setting than the agnostic noise considered in this paper), and they made a number of assumptions—a particular (large) initialization, sufficiently wide network, and the use of the squared loss for classification—that were not used in this work. The results of Li et al. (2019a) covered general, smooth activation functions (but not leaky-ReLU).

Hu et al. (2020a) showed that ultra-wide networks with NTK scaling and initialization trained by SGD with various forms of regularization can generalize when the labels are corrupted with random classification noise. Their generalization bound was given in terms of the classification error on the 'clean' data distribution (without any noise) and allowed for general activation functions (including leaky-ReLU). In comparison, we assume that the training data and the test data come from the same distribution, and our generalization bound is given in terms of the performance of the best linear classifier over the distribution. Our generalization guarantee holds without any explicit forms of regularization, suggesting that the mechanism responsible for the lack of overfitting is not explicit regularization, but forms of regularization that are *implicit* to the SGD algorithm.

3. Proof Outline of the Main Results

We will show that stochastic gradient descent achieves small classification error by using a proof technique similar to that of Brutzkus et al. (2018), who showed the convergence and generalization of gradient descent on the hinge loss for one-hidden-layer leaky ReLU networks on linearly separable data.[3] Their proof relies upon the fact that both the classification error and the hinge loss for the best halfspace are zero. In our setting—without the assumption of linear separability, and with more general loss functions—their strategy for showing that the empirical risk can be driven to zero will not work. (We remind the reader that our goal is to show that the neural network will generalize when it is of *arbitrary* width, and when significant noise is present, and thus we cannot guarantee the smallest empirical or population loss is arbitrarily close to zero.) Instead, we need to compare the performance of the neural network with that of the best linear classifier over the data, which will in general have error (both classification and loss value) bounded away from zero. To do so, we use some of the ideas used in Frei et al. (2021) to derive generalization bounds for the classification error when the surrogate loss is bounded away from zero.

To begin, let us introduce some notation. Let $v^* \in \mathbb{R}^d$ be a unit norm halfspace that minimizes the halfspace error, so that

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \operatorname{sgn}\big(\langle v^*, x\rangle\big)\Big) = \mathsf{OPT}.$$

Denote the matrix $V \in \mathbb{R}^{m\times d}$ as having rows $v_j^\top \in \mathbb{R}^d$ defined by

$$v_j = \frac{1}{\sqrt{m}} \operatorname{sgn}(a_j) v^*. \tag{5}$$

The scaling of each row of the matrix V ensures that $\|V\|_F = 1$. For $\gamma > 0$, denote

$$\widehat{\xi}_t(\gamma) := \mathbb{1}(y_t\langle v^*, x_t\rangle \in [0,\gamma)) + (1+\gamma^{-1}|\langle v^*, x_t\rangle|)\mathbb{1}(y_t\langle v^*, x_t\rangle < 0).$$

The expected value of the above quantity will be an important quantity in our proof. To give some idea of how this quantity will fit in to our analysis, assume for the moment that $\|x\| \leq 1$ a.s. Then taking expectations of the above and using Cauchy–Schwarz, we get

$$\begin{aligned}
&\mathbb{E}\widehat{\xi}_t(\gamma) \\
&\leq \phi_{v^*}(\gamma) + (1+\gamma^{-1})\mathbb{E}[|\langle v^*, x_t\rangle|\mathbb{1}(y_t\langle v^*, x_t\rangle < 0)] \\
&\leq \phi_{v^*}(\gamma) + (1+\gamma^{-1})\mathsf{OPT}.
\end{aligned} \tag{6}$$

The above appears (in a more general form) in the bound for the classification error presented in Theorem 2.6. In particular, the goal below will be to show that the classification

[3] This proof technique can be viewed as an extension of the Perceptron proof presented in Shalev-Shwartz & Ben-David (2014, Theorem 9.1).

error can be bounded by a constant multiple of $\mathbb{E}[\widehat{\xi}_t(\gamma)]$. Continuing, let us denote

$$\widehat{H}_t := \langle W^{(t)}, V\rangle, \quad \widehat{G}_t^2 = \|W^{(t)}\|_F^2. \tag{7}$$

The quantity $\widehat{H}_t$ measures the correlation between the weights found by SGD and those of the best linear classifier over the distribution. We define the population-level versions of each of the random variables above by replacing the $\widehat{\cdot}$ with their expectation $\mathbb{E}_{\text{sgd}}(\cdot)$ over the randomness of the draws $(x_1, \dots, x_t)$ of the distribution used for SGD. That is, we denote $L_t := \mathbb{E}_{\text{sgd}}\widehat{L}_t(W^{(t)})$, $\mathcal{E}_t := \mathbb{E}_{\text{sgd}}\widehat{\mathcal{E}}_t(W^{(t)})$, $H_t := \mathbb{E}_{\text{sgd}}\widehat{H}_t$, $G_t^2 := \mathbb{E}_{\text{sgd}}[\widehat{G}_t^2]$, and $\xi(\gamma) := \mathbb{E}_{(x_t,y_t)\sim\mathcal{D}}\widehat{\xi}_t(\gamma)$.

Our proof strategy will be to show that until gradient descent finds weights with small risk, the correlation H_T between the weights found by SGD and those of the best linear predictor will grow at least as fast as $\Omega(T)$, while G_T always grows at a rate of at most $O(\sqrt{T})$. Since $\|V\|_F = 1$, by Cauchy–Schwarz we have the bound $H_T \le G_T$, and so the growth rates $H_T = \Omega(T)$ and $G_T = O(\sqrt{T})$ can only be satisfied for a small number of iterations. In particular, there can only be a small number of iterations until SGD finds weights with small risk.

To see how we might be able to show that the correlation H_T is increasing, note that we have the identity

$$\widehat{H}_{t+1} - \widehat{H}_t = -\eta\ell'(y_t f_t(W^{(t)}))y_t\langle\nabla f_t(W^{(t)}), V\rangle.$$

Since $-\ell' \ge 0$, the inequality $\widehat{H}_{t+1} > \widehat{H}_t$ holds if we can show $y_t\langle\nabla f_t(W^{(t)}), V\rangle > 0$, i.e. if we can show that the gradient of the neural network is correlated with the weights of the best linear predictor. For this reason, the following technical lemma is a key ingredient in our proof.

Lemma 3.1. *For V defined in (5), for any $(x_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$, for any $W \in \mathbb{R}^{m\times d}$, and any $\gamma \in (0, 1)$,*

$$y_t\langle\nabla f_t(W), V\rangle \ge a\gamma\sqrt{m}\Big[\alpha - \widehat{\xi}_t(\gamma)\Big]. \tag{8}$$

The proof of the above lemma is in Appendix B.1. As alluded to above, with this technical lemma we can show that until the surrogate risk is as small as a constant factor of $\xi(\gamma)$, the correlation of the weights found by SGD and those of the best linear predictor is increasing.

Lemma 3.2. *For any $t \in \mathbb{N}\cup\{0\}$, for any $\gamma > 0$, it holds that*

$$H_{t+1} \ge H_t + \eta a\gamma\sqrt{m}\big[\alpha\mathcal{E}_t - \xi(\gamma)\big].$$

Proof. We can write

$$\begin{aligned}
\widehat{H}_{t+1} &= \widehat{H}_t - \eta\langle\nabla\widehat{L}_t(W^{(t)}), V\rangle \\
&= \widehat{H}_t - \eta\ell'(y_t f_t(W^{(t)}))y_t\langle\nabla f_t(W^{(t)}), V\rangle \\
&\ge \widehat{H}_t - \eta\ell'(y_t f_t(W^{(t)}))a\gamma\sqrt{m}[\alpha - \widehat{\xi}_t(\gamma)] \\
&\ge \widehat{H}_t + \eta a\gamma\sqrt{m}\left[\alpha\widehat{\mathcal{E}}_t(W^{(t)}) - \widehat{\xi}_t(\gamma)\right].
\end{aligned}$$

In the first inequality we have used Lemma 3.1 and that $-\ell' \ge 0$, and in the second inequality we have used that $-\ell' \le 1$. Taking expectations over the draws of the distribution on both sides completes the proof. □

Notice that if $\alpha\mathcal{E}_t > \xi(\gamma)$, Lemma 3.2 shows that $H_{t+1} - H_t > 0$. We will later repeat this argument for T iterations to show that until we find a point with $\alpha\mathcal{E}_t \le 2\xi(\gamma)$, H_T will grow at least as fast as $\Omega(T)$.

All that remains is to show that $G_T = O(\sqrt{T})$. We will accomplish this by first demonstrating a bound on $G_{t+1}^2 - G_t^2$.

Lemma 3.3. *For any $t \in \mathbb{N}\cup\{0\}$, $\eta > 0$, and if $\mathbb{E}[\|x\|^2] \le B_X^2$,*

$$G_{t+1}^2 \le G_t^2 + 2\eta + \eta^2 m a^2 B_X^2.$$

The proof of Lemma 3.3 is provided in Appendix B. We now have all of the ingredients needed to prove Theorem 2.6. We provide a proof sketch below and leave the complete proof for Appendix B.3.

Sketch of Proof of Theorem 2.6. For V defined as (5) (satisfying $\|V\|_F = 1$), we have by Cauchy–Schwarz,

$$H_t^2 = (\mathbb{E}[\langle W^{(t)}, V\rangle])^2 \le \mathbb{E}\|W^{(t)}\|_F^2\mathbb{E}\,\|V\|_F^2 = G_t^2 \tag{9}$$

For $a = 1/\sqrt{m}$, and for $\eta \le (ma^2B_X^2)^{-1} = B_X^{-2}$, Lemma 3.3 becomes

$$G_{t+1}^2 \le G_t^2 + 2\eta + \eta^2 m a^2 B_X^2 \le G_t^2 + 3\eta.$$

Summing the above from $t = 0, \dots, T-1$, we get

$$G_T^2 \le G_0^2 + 3\eta T. \tag{10}$$

Similarly, Lemma 3.2 becomes $H_{t+1} \ge H_t + \eta\gamma[\alpha\mathcal{E}_t - \xi]$. (Note that $\xi = \xi(\gamma)$ depends on γ, but we have dropped the notation for simplicity.) Summing this from $t = 0$ to $T-1$, we get

$$H_T \ge H_0 + \eta\gamma\sum_{t=0}^{T-1}[\alpha\mathcal{E}_t - \xi]. \tag{11}$$

We can therefore bound

$$\begin{aligned}
-G_0 + \eta\gamma\sum_{t=0}^{T-1}[\alpha\mathcal{E}_t - \xi] &\le H_0 + \eta\gamma\sum_{t=0}^{T-1}[\alpha\mathcal{E}_t - \xi] \\
&\le H_T \le G_T \\
&\le G_0 + \sqrt{T}\cdot 2\sqrt{\eta}.
\end{aligned} \tag{12}$$

The first inequality uses (9). The second inequality uses (11). The third inequality again uses (9). The final inequality uses (10) together with $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$.

We claim now that this implies that within a polynomial number of samples, SGD finds weights satisfying $\mathcal{E}_t \le 2\alpha^{-1}\xi$. Suppose that for every iteration $t = 1, \dots, T$, we have $\mathcal{E}_t > 2\alpha^{-1}\xi$. Then (12) gives

$$\eta\alpha\gamma\xi \cdot T - 2\sqrt{\eta} \cdot \sqrt{T} - 2G_0 \le 0.$$

This is an equation of the form $\beta_2(\sqrt{T})^2 - \beta_1\sqrt{T} - \beta_0 \le 0$ where $\beta_i > 0$, which can only be satisfied when T is smaller than some polynomial function of the β_i (see the appendix for the exact details). In particular, within $T = O(\eta^{-1}\gamma^{-2}\xi^{-2}[G_0 \vee 1])$ iterations, gradient descent finds a point satisfying

$$\mathcal{E}_t = \mathbb{E}_{\text{sgd}}\big[-\ell'\big(yf_x(W^{(t)})\big)\big] \le 2\alpha^{-1}\xi. \tag{13}$$

By Markov's inequality (see (3)) this implies

$$\mathbb{P}(yf_x(W^{(t)}) < 0) \le 2|\ell'(0)|^{-1}\alpha^{-1}\xi.$$

To complete the proof, we want to bound ξ. Recall from the calculation (6) that

$$\xi = \phi_{v^*}(\gamma) + \mathsf{OPT} + \gamma^{-1}\mathbb{E}\Big[|\langle v^*, x\rangle|\mathbb{1}(y\langle v^*, x\rangle < 0)\Big].$$

For simplicity consider the case that $\|x\| \le 1$ a.s. (the sub-exponential case follows using a truncation argument; details given in Appendix B.3). Then by Cauchy–Schwarz,

$$\mathbb{E}[|\langle v^*, x\rangle|\mathbb{1}(y\langle v^*, x\rangle < 0)] \le \mathbb{P}(y\langle v^*, x\rangle < 0) = \mathsf{OPT}.$$

Substituting the above into (13), we get

$$\mathbb{P}(yf_x(W^{(t)}) < 0) \le 2|\ell'(0)|^{-1}\alpha^{-1}\phi_{v^*}(\gamma) + 2|\ell'(0)|^{-1}\alpha^{-1}(1+\gamma^{-1})\mathsf{OPT}.$$

□

4. Experiments

In this section, we provide some experimental verification of our theoretical results. We consider a distribution $\mathcal{D}_{b,\gamma_0}$ that is a mixture of two 2D Gaussians perturbed by both random classification noise and deterministic (adversarial) label noise. The distribution is constructed as follows. We first take two independent Gaussians with independent components of unit variance and means $(-3, 0)$ and $(3, 0)$, and assign the label -1 to the left cluster and $+1$ to the right cluster. We remove all samples with first component x_1 satisfying $|x_1| \le \gamma_0 = 0.5$, so that we have a hard margin distribution with margin γ_0. We then introduce a boundary factor $b > \gamma_0$, and for samples with first component satisfying $|x_1| \le b$ we deterministically flip the label to the opposite sign. Finally, for samples with $|x_1| > b$, we introduce random classification noise at level 10%, flipping the labels in those regions with probability 0.1 each. The symmetry of the distribution implies that an optimal halfspace is the vector $v^* = (1, 0)$.

The boundary factor b can be tweaked to incorporate more deterministic label noise which will affect the best linear classifier: if b is larger, OPT is larger as well. We give details on the precise relationship of b and OPT in Appendix C. But because this 'noise' is deterministic, the best classifier over $\mathcal{D}_{b,\gamma_0}$ (the Bayes optimal classifier) can always achieve accuracy of at least 90% by using the decision rule

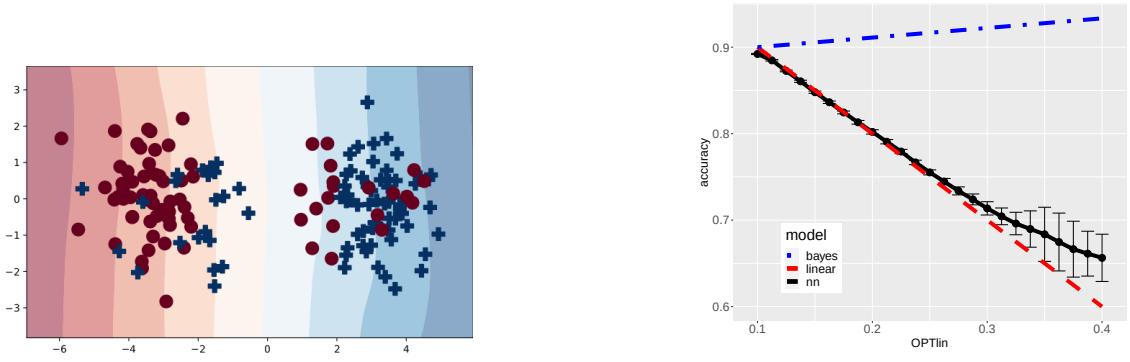
$$y_{\text{Bayes}} = \begin{cases} +1, & x_1 \in (-b, 0) \cup (b, \infty), \\ -1, & x_1 \in (-\infty, b] \cup [0, b]. \end{cases} \tag{14}$$

Since the error for the Bayes decision rule corresponds to the region $\{|x_1| > b\}$ with random classification noise, we can exactly calculate the error for the Bayes classifier as well as OPT. As b increases, the region with random classification noise becomes smaller, and thus the Bayes classifier gets better as the linear classifier becomes worse on $\mathcal{D}_{b,\gamma_0}$. This makes $\mathcal{D}_{b,\gamma_0}$ a good candidate for understanding the performance of SGD-trained one-hidden-layer networks in comparison to linear classifiers. Further, to our knowledge no previous work has been able to show that neural networks can provably generalize if the data distribution is $\mathcal{D}_{b,\gamma_0}$.[4]

Since $\mathcal{D}_{b,\gamma_0}$ is a subexponential hard margin distribution, Corollary 2.7 shows that we can expect an SGD-trained leaky ReLU network on $\mathcal{D}_{b,0.5}$ to achieve a test set accuracy of at least $1 - C \cdot \mathsf{OPT}\log(1/\mathsf{OPT})$ for some constant $C \ge 1$. We ran experiments on such a neural network with $m = 1000$ neurons and learning rate $\eta = 0.01$ and first layer weights initialized as independent normal random variables with variance $1/m$ (see Appendix C for more details on the experiment setup). In Figure 1a we plot the decision boundary for the SGD-trained neural network on the distribution $\mathcal{D}_{2.04,0.5}$, where $b = 2.04$ is chosen so that $\mathsf{OPT} = 0.25$. We notice that the decision boundary is almost exactly linear and is essentially the same as that of the best linear classifier $(x_1, x_2) \mapsto \text{sgn}(x_1)$. And in Figure 1b, we see that the neural network accuracy is almost exactly equal to $1 - \mathsf{OPT}$ when $\mathsf{OPT} \le 0.30$ and that the network slightly outperforms the best linear classifier when $\mathsf{OPT} > 0.30$.

In Appendix C we conduct additional experiments to better understand whether this behavior is consistent across hyperparameter and architectural modifications to the net-

[4] There are two reasons that no other work can show generalization bounds in the settings we consider. The first is the presence of adversarial label noise. The second is that our generalization bound holds for neural networks with finite width and any initialization. All previous works fail to allow at least one of these conditions.



(a) (b)

Figure 1. (a) Samples from $\mathcal{D}_{2.04,0.5}$ with random classification noise of 10% on $\{|x_1| > 2.04\}$ with the boundary term $b = 2.04$ chosen so that $\mathsf{OPT} = 0.25$. Blue plus signs correspond to $y = +1$ and red circles to $y = -1$. The contour plot displays the class probability for the output of a leaky ReLU network trained by online SGD and has dark hues when the neural network is more confident in its predictions. (b) Test classification accuracy for data coming $\mathcal{D}_{b,0.5}$. The red dashed line is the accuracy of the best linear classifier, and the black solid line is the average accuracy of the neural network with error bars over ten random initializations of the first layer weights (experimental details can be found in Appendix C). The blue dash-dotted line is the Bayes optimal classifier accuracy.

work. When using the bias-free networks of the form (2) we consider in this paper, we found that one-hidden-layer SGD-trained networks failed to generalize better than a linear classifier when using $\tanh$ activations (Figure 3), using different learning rates (Figure 4), different initialization variances (Figure 5), and using multiple-pass SGD rather than online SGD (Figure 6). On the other hand, we found that introducing bias terms can lead to decision boundaries closer to that of the Bayes-optimal classifier (Figure 7). Interestingly, this behavior was strongly dependent on the initialization scheme used: when using an initialization variance of $1/m^4$, a linear decision boundary was consistently learned, while using an initialization variance of $1/m$ lead to approximately Bayes-optimal decision boundaries. By contrast, the result we present in Theorem 2.6 holds for *arbitrary* initialization schemes. This suggests that a new analytical approach would be needed in order to guarantee neural network generalization performance better than that of a linear classifier on $\mathcal{D}_{\gamma_0,b}$.

5. Discussion

We have shown that overparameterized one-hidden-layer networks can generalize almost as well as the best linear classifier over the distribution for a broad class of distributions. Our results imply two related but distinct insights on SGD-trained neural networks. First, regardless of the initialization scheme and number of neurons, SGD training will produce neural networks that are competitive with the best linear predictor over the data, providing theoretical support for the hypothesis presented by Nakkiran et al. (2019) that the performance of SGD-trained networks in the early stages of training can be explained by that of a linear classifier. Second, a linearly separable dataset can be corrupted by adversarial label noise and overparameterized neural networks will still be able to generalize, despite the capacity to overfit to the label noise.

A number of extensions and open questions remain. First, our analysis was specific to one-hidden-layer networks with the leaky-ReLU activation. We are interested in extending our results to more general neural network architectures. Second, a natural question is whether or not there are concept classes that are more expressive than halfspaces for which overparameterized neural networks can generalize for noisy data. We are particularly keen on understanding this question for finite width neural networks that are not well-approximated by the NTK.

Acknowledgements

We thank James-Michael Leahy for a number of helpful discussions. We thank Maria-Florina Balcan for pointing us to a number of works on learning halfspaces in the presence of noise. We also thank the anonymous reviewers for their helpful comments. SF is supported by the UCLA Dissertation Year Fellowship. YC and QG are partially supported by the National Science Foundation IIS-2008981. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Allen-Zhu, Z. and Li, Y. What can resnet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2019b.

Awasthi, P., Balcan, M.-F., Haghtalab, N., and Urner, R. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory (COLT)*, 2015.

Awasthi, P., Balcan, M.-F., Haghtalab, N., and Zhang, H. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory (COLT)*, 2016.

Awasthi, P., Balcan, M. F., and Long, P. M. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6), January 2017. ISSN 0004-5411.

Balcan, M.-F. and Haghtalab, N. Noise in classification. In Roughgarden, T. (ed.), *Beyond Worst Case Analysis of Algorithms*, chapter 16. Cambridge University Press, 2021.

Blum, A., Frieze, A., Kannan, R., and Vempala, S. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1-2):35–52, 1998.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations (ICLR)*, 2018.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Cao, Y. and Gu, Q. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Chen, Z., Cao, Y., Zou, D., and Gu, Q. How much over-parameterization is sufficient to learn deep relu networks? *arXiv*, abs/1911.12360, 2019. URL http://arxiv.org/abs/1911.12360.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A generalized neural tangent kernel analysis for two-layer neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Daniely, A. Complexity theoretic limitations on learning halfspaces. In *ACM Symposium on Theory of Computing (STOC)*, pp. 105–117, 2016.

Diakonikolas, I., Gouleakis, T., and Tzamos, C. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Diakonikolas, I., Kane, D. M., Kontonis, V., and Zarifis, N. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory (COLT)*, 2020a.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory (COLT)*, 2020b.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Non-convex sgd learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020c.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

Du, S. S., Zhai, X., Póczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Fort, S., Dziugaite, G. K., Paul, M., Kharaghani, S., Roy, D. M., and Ganguli, S. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Frei, S., Cao, Y., and Gu, Q. Algorithm-dependent generalization bounds for overparameterized deep residual networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning (ICML)*, 2021.

Goel, S., Karmalkar, S., and Klivans, A. R. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Goel, S., Gollakota, A., Jin, Z., Karmalkar, S., and Klivans, A. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning (ICML)*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hu, W., Li, Z., and Yu, D. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations (ICLR)*, 2020a.

Hu, W., Xiao, L., Adlam, B., and Pennington, J. The surprising simplicity of the early-time learning dynamics of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ji, Z. and Telgarsky, M. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory (COLT)*, 2019.

Ji, Z. and Telgarsky, M. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations (ICLR)*, 2020b.

Kearns, M. J., Schapire, R. E., and Sellie, L. M. Toward efficient agnostic learning. *Machine Learning*, 17(2-3): 115–141, 1994.

Klivans, A. R., Long, P. M., and Servedio, R. A. Learning halfspaces with malicious noise. *Journal of Machine Learning Research (JMLR)*, 10(94):2715–2740, 2009.

Li, M., Soltanolkotabi, M., and Oymak, S. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019a.

Li, Y., Wei, C., and Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

Li, Y., Ma, T., and Zhang, H. R. Learning over-parametrized two-layer relu neural networks beyond ntk. In *Conference on Learning Theory (COLT)*, 2020a.

Li, Y., X.Fang, E., Xu, H., and Zhao, T. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations (ICLR)*, 2020b.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Massart, P., Nédélec, É., et al. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory (COLT)*, 2019.

Moroshko, E., Gunasekar, S., Woodworth, B., Lee, J. D., Srebro, N., and Soudry, D. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Nakkiran, P., Kaplun, G., Kalimeris, D., Yang, T., Edelman, B. L., Zhang, F., and Barak, B. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135.

Shamir, O. Are resnets provably better than linear predictors? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*, 19(70):1–57, 2018.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 2019.

Zou, D., Frei, S., and Gu, Q. Provable robustness of adversarial training for learning halfspaces with noise. In *International Conference on Machine Learning (ICML)*, 2021.