Journal of Machine Learning Research 25 (2024) 1-41 Submitted 4/23; Revised 2/24; Published 4/24

Generalization and Stability of Interpolating Neural Networks with Minimal Width

Hossein Taheri HOSSEIN@UCSB.EDU
Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA, USA

Christos Thrampoulidis CTHRAMPO@ECE.UBC.CA
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada

Editor: Pradeep Ravikumar

Abstract

We investigate the generalization and optimization properties of shallow neural-network classifiers trained by gradient descent in the interpolating regime. Specifically, in a realizable scenario where model weights can achieve arbitrarily small training error ϵ and their distance from initialization is $g(\epsilon)$, we demonstrate that gradient descent with n training data achieves training error $O(g(1/T)^2/T)$ and generalization error $O(g(1/T)^2/n)$ at iteration T, provided there are at least $m = \Omega(g(1/T)^4)$ hidden neurons. We then show that our realizable setting encompasses a special case where data are separable by the model's neural tangent kernel. For this and logistic-loss minimization, we prove the training loss decays at a rate of $\tilde{O}(1/T)$ given polylogarithmic number of neurons $m = \Omega(\log^4(T))$. Moreover, with $m = \Omega(\log^4(n))$ neurons and $T \approx n$ iterations, we bound the test loss by $\tilde{O}(1/n)$. Our results differ from existing generalization outcomes using the algorithmic-stability framework, which necessitate polynomial width and yield suboptimal generalization rates. Central to our analysis is the use of a new self-bounded weak-convexity property, which leads to a generalized local quasi-convexity property for sufficiently parameterized neural-network classifiers. Eventually, despite the objective's non-convexity, this leads to convergence and generalization-gap bounds that resemble those found in the convex setting of linear logistic regression.

Keywords: Generalization Error, Neural Networks, Optimization, Over-parameterization, Interpolation.

1. Introduction

Neural networks have remarkable expressive capabilities and can memorize a complete dataset even with mild overparameterization. In practice, using gradient descent (GD) on neural networks with logistic or cross-entropy loss can result in the objective reaching zero training error and close to zero training loss. Zero training error, often referred to as "interpolating" the data, indicates perfect classification of the dataset. Despite their strong memorization ability, these networks also exhibit remarkable generalization capabilities to new data. This has motivated a surge of studies in recent years exploring the optimization

and generalization properties of first-order gradient methods in overparameterized neural networks, with a specific focus in the so-called Neural Tangent Kernel (NTK) regime. In the NTK regime, the model operates as the first-order approximation of the network at a sufficiently large initialization or at the large-width limit (Jacot et al., 2018; Chizat et al., 2019). Prior works on this topic mostly focused on quadratic-loss minimization and their optimization/generalization guarantees required network widths that increased polynomially with the sample size n. This, however, is not in line with practical experience. Improved results were obtained more recently by Ji and Telgarsky (2020a); Chen et al. (2020) who have investigated the optimization and generalization of ReLU neural networks with logistic loss, which is more suitable for classification tasks. Assuming that the NTK with respect to the model can interpolate the data (i.e. separate them with positive margin), they showed through a Rademacher complexity analysis that GD on neural networks with polylogarithmic width can achieve generalization guarantees that decrease with the sample size n at a rate of $\tilde{O}(\frac{1}{\sqrt{n}})$.

In this paper, we provide rate-optimal optimization and generalization analyses of GD for shallow neural networks of minimal width assuming that the model itself can interpolate the data. We focus on two-layer networks with smooth activations that can almost surely separate n training samples from the data distribution. Concretely, we consider a realizability condition where data and initialization are such that model weights can achieve arbitrarily small training error ε while their distance from initialization is $g(\varepsilon)$ for some function $g : \mathbb{R}_+ \to \mathbb{R}_+$. Under this condition, we demonstrate generalization guarantees of order $O(\frac{g(\frac{1}{T})^2}{n})$. More generally, for any iteration T of GD and assuming network width $m = \Omega(g(\frac{1}{T})^4)$, we obtain an expected test-loss rate $O(\frac{g(\frac{1}{T})^2}{T} + \frac{g(\frac{1}{T})^2}{n})$. Additional to the generalization bounds, we provide optimization guarantees under the same setting by showing that the training loss approaches zero at rate $O(\frac{g(\frac{1}{T})^2}{T})$. We note that these results are derived without NTK-type analyses. For demonstration and also for connection to prior works on neural-tangent data models, we specialize our generalization and optimization results to the class of NTK-separable data. We show this is possible because the NTK-data separability assumption implies our realizability condition holds. Thus, for logistic-loss minimization on NTK-separable data, we show that the expected test loss of GD is $\tilde{O}(\frac{1}{T} + \frac{1}{n})$ provided polylogarithmic number of neurons $m = \Omega(\log^4(T))$. This further suggests that a network of width $m = \Omega(\log^4(n))$, attains expected test loss $\tilde{O}(\frac{1}{n})$ after $T \approx n$ iterations.

In contrast to prior optimization and generalization analyses that often depend on the NTK framework, which requires the first-order approximation of the model, we build on the algorithmic stability approach (Bousquet and Elisseeff, 2002) for shallow neural-network models of finite width. Although the stability analysis has been utilized in previous studies to derive generalization bounds for (stochastic) gradient descent in various models, most results that are rate-optimal heavily rely on the convexity assumption. Specifically, the stability-analysis framework has been successful in achieving optimal generalization bounds for convex objectives in (Lei and Ying, 2020a; Bassily et al., 2020; Schliserman and Koren, 2022). On the other hand, previous studies on non-convex objectives either resulted in suboptimal bounds or relied on assumptions that are not in line with the actual practices of neural network training. For instance, Hardt et al. (2016) derived a generalization bound of $O(\frac{T^{\beta c/(\beta c+1)}}{n})$ for general β-smooth and non-convex objectives, but this required

a time-decaying step-size $\eta_t \leq c/t$, which can degrade the training performance. More recently, Richards and Rabbat (2021) explored the use of the stability approach specifically for logistic-loss minimization of a two-layer network. By refining the model-stability analysis framework introduced by (Lei and Ying, 2020a), they derived generalization-error bounds provided the hidden width increases polynomially with the sample size. In comparison, our analysis leads to significantly improved generalization and optimization rates and under standard separability conditions such as NTK-separability, only requires a polylogarithmic width for both global convergence and generalization.

Notation

We define $[n] := \{1, 2, \cdots, n\}$. We use the standard notation $O(\cdot), \Omega(\cdot)$ and use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ to hide polylogarithmic factors. Occasionally we use $\lesssim$ to hide numerical constants. The Gradient and Hessian of a function $\Phi : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ with respect to the ith input $(i = 1, 2)$ are denoted by $\nabla_i \Phi$ and $\nabla_i^2 \Phi$, respectively. All logarithms are in base e. We use $\|\cdot\|$ for the ℓ_2 norm of vectors and the operator norm of matrices. We denote $[w_1, w_2] := \{w : w = \alpha w_1 + (1 - \alpha) w_2, \alpha \in [0, 1]\}$ the line segment between $w_1, w_2 \in \mathbb{R}^{d'}$.

2. Problem Setup

Given n i.i.d. samples $(x_i, y_i) \sim \mathcal{D}, i \in [n]$ from data distribution $\mathcal{D}$, we study unconstrained empirical risk minimization with objective $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$:

$$\min_{w \in \mathbb{R}^{d'}} \Big\{ \widehat{F}(w) := \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_i(w) = \frac{1}{n} \sum_{i=1}^{n} f\left(y_i \Phi\left(w, x_i\right)\right) \Big\}. \tag{1}$$

This serves as a proxy for minimizing the *test loss* $F : \mathbb{R}^{d'} \to \mathbb{R}$:

$$F(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[f\left(y \Phi(w, x)\right) \right]. \tag{2}$$

We introduce our assumptions on the data (x, y), the model $\Phi(\cdot, x)$, and the loss function $f(\cdot)$, below. We start by imposing the following mild assumption on the data distribution.

Assumption 1 (Bounded features)**.** *Assume any $(x, y) \sim \mathcal{D}$ has almost surely bounded features, i.e. $\|x\| \leq R$, and binary label $y \in \{\pm 1\}$.*

The model $\Phi : \mathbb{R}^{d'} \times \mathbb{R}^{d} \to \mathbb{R}$ is parameterized by trainable weights $w \in \mathbb{R}^{d'}$ and takes input $x \in \mathbb{R}^{d}$. For our main results, we assume Φ is a one-hidden layer neural-net of m neurons, i.e.

$$\Phi(w, x) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \, \sigma(\langle w_j, x \rangle), \tag{3}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, $w_j \in \mathbb{R}^{d}$ denotes the weight vector of the jth hidden neuron and $\frac{a_j}{\sqrt{m}}, j \in [m]$ are the second-layer weights. For the second layer weights, we assume that they are fixed during training taking values $a_j \in \{\pm 1\}$. We assume that for half of second layer weights $a_j = 1$ and for the other half $a_j = -1$. On the other hand,

all the first-layer weights are updated during training. Thus, the total number of trainable parameters is $d' = md$ and we denote $w = [w_1; w_2; \ldots; w_m] \in \mathbb{R}^{d'}$ the vector of trainable weights. Throughout, we make the following assumptions on the activation function.

Assumption 2 (Lipschitz and smooth activation)**.** *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies the following for non-negative constants ℓ, L:*

$$|\sigma'(u)| \leq \ell, \quad |\sigma''(u)| \leq L, \qquad \forall u \in \mathbb{R}.$$

We note that the smoothness assumption which is required by our framework excludes the use of ReLU. Examples of activation functions that satisfy the smoothness condition include Softplus $\sigma(u) = \log(1 + e^u)$, Gaussian error linear unit (GELU) $\sigma(u) = \frac{1}{2}u(1 + \mathrm{erf}(\frac{u}{\sqrt{2}}))$, and Hyperbolic-Tangent where $\sigma(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$. On the other hand, Lipschitz assumption is rather mild, since it is possible to restrict the parameter space to a bounded domain.

Next, we discuss conditions on the loss function. Of primal interest is the commonly used logistic loss function $f(u) = \log(1 + e^{-u})$. However, our results hold for a broader class of convex, non-negative and monotonically decreasing functions ($\lim_{u\to\infty} f(u) = 0$) that satisfy the following:

Assumption 3 (Lipschitz and smooth loss)**.** *The convex loss function $f : \mathbb{R} \to \mathbb{R}_+$ satisfies for all $u \in \mathbb{R}$:*

3.A: *Lipschitzness: $|f'(u)| \leq G_f$.*

3.B: *Smoothness: $f''(u) \leq L_f$.*

Assumption 4 (Self-bounded loss)**.** *The convex loss function $f : \mathbb{R} \to \mathbb{R}_+$ is self-bounded with some constant $\beta_f > 0$, i.e., $|f'(u)| \leq \beta_f f(u), \forall u \in \mathbb{R}$.*

The self-boundedness Assumption 4 is the key property of the loss that drives our analysis and justifies the polylogarithmic width requirement, as will become evident. Note that the logistic loss naturally satisfies Assumptions 3.A and 3.B (with $G_f = 1, L_f = 1/4$), as well as, Assumption 4 with $\beta_f = 1$. Other interesting examples of loss functions satisfying those assumptions include polynomial losses, with the tail behavior $f(u) = 1/u^\beta$ for $\beta > 0$, which we discuss in Remark 2. To lighten the notation and without loss of generality, we set $G_f = L_f = \beta_f = 1$ for the rest of the paper. We remark that our training-loss results also hold for the exponential loss e^{-u}. The exponential loss is self-bounded and while it is not Lipschitz or smooth it satisfies a second-order self-bounded property $f''(u) \leq f(u)$, which we can leverage instead; see Appendix A for details.

3. Main Results

We present bounds on the train loss and generalization gap of gradient-descent (GD) under the setting of Section 2. Formally, GD with step-size $\eta > 0$ optimizes (1) by performing the following updates starting from an initialization w_0:

$$\forall t \geq 0 : \; w_{t+1} = w_t - \eta \nabla \widehat{F}(w_t).$$

3.1 Key properties

The key challenge in both the optimization and generalization analysis is the non-convexity of $f(y\Phi(\cdot, x))$, and consequently of the train loss $\widehat{F}(\cdot)$. Despite non-convexity, we derive bounds analogous to the convex setting, e.g. corresponding bounds on linear logistic regression in (Ji and Telgarsky, 2018; Shamir, 2021; Schliserman and Koren, 2022). We show this is possible provided the loss satisfies the following key property, which we call *self-bounded weak convexity*.

Definition 1 (Self-bounded weak convexity)**.** *We say a function $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$ is self-bounded weakly convex if there exists constant $\kappa > 0$ such that for all w,*

$$\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\kappa\, \widehat{F}(w)\,. \tag{4}$$

Recall a function $G : \mathbb{R}^{d'} \to \mathbb{R}$ is weakly convex if $\exists \kappa \geq 0$ such that uniformly over all $w \in \mathbb{R}^{d'}$, $\lambda_{\min}\left(\nabla^2 G(w)\right) \geq -\kappa$. If $\kappa = 0$, the function is convex. Instead, property (4) lower bounds the curvature by $-\kappa\, G(w)$ that changes proportionally with the function value $G(w)$. We explain below how this is exploited in our setting.

To begin with, the following lemma shows that property (4) holds for the train loss under the setting of Section 2: training of a two-layer net with smooth activation and self-bounded loss. The lemma also shows that the gradient of the train loss is self bounded. Those two properties together summarize the key ingredients for which our analysis applies.

Lemma 3.1 (Key self-boundedness properties)**.** *Consider the setup of Section 2 and let Assumptions 1-2 hold. Further assume the loss is self-bounded as per Assumption 4. Then, the objective satisfies the following self-boundedness properties for its Gradient and Hessian:*

1. *Self-bounded gradient:* $\left\|\nabla \widehat{F}_i(w)\right\| \leq \ell R\, \widehat{F}_i(w), \;\; \forall i \in [n]$.
2. *Self-bounded weak convexity:* $\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\frac{LR^2}{\sqrt{m}} \widehat{F}(w)$.

Both of these properties follow from the self-boundedness of the convex loss f combined with Lipshitz and smoothness of σ. The self-boundedness of the gradient is used for generalization analysis and in particular in obtaining the model stability bound. The self-bounded weak convexity plays an even more critical role for our optimization and generalization results. In particular, the wider the network the closer the loss to having convex-like properties. Moreover, the "self-bounded" feature of this property provides another mechanism that favors convex-like optimization properties of the loss. To see this, consider the minimum Hessian eigenvalue $\lambda_{\min}(\nabla^2 \widehat{F}(w_t))$ at gradient descent iterates $\{w_t\}_{t\geq 1}$: As training progresses, the train loss $\widehat{F}(w_t)$ decreases, and thanks to the self-bounded weak convexity property, the gap to convexity also decreases. We elaborate on the role of self-bounded weak convexity in our proofs in Section 5.

3.2 Training loss

We begin with a general bound on the training loss and the parameter's norm, which is also required for our generalization analysis.

Theorem 3.2 (Training loss – General bound). *Suppose Assumptions 1-4 hold. Fix any training horizon $T \geq 0$ and any step-size $\eta \leq 1/L_{\widehat{F}}$ where $L_{\widehat{F}}$ is the objective's smoothness parameter. Assume any $w \in \mathbb{R}^{d'}$ and hidden-layer width m such that $\|w - w_0\|^2 \geq \max\{\eta T \widehat{F}(w), \eta \widehat{F}(w_0)\}$ and $m \geq 18^2 L^2 R^4 \|w - w_0\|^4$. Then, the training loss and the parameters' norm satisfy*

$$\widehat{F}(w_T) \;\leq\; \frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \;\leq\; 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T}, \tag{5}$$
$$\forall t \in [T] \;\;:\;\; \|w_t - w_0\| \;\leq\; 4\|w - w_0\|.$$

A few remarks are in place regarding the theorem. First, Eq. (5) upper bounds the running average (also known as regret) of train loss for iterations $1, \ldots, T$ by the value, at an arbitrarily chosen point w, of a ridge-regularized objective with regularization parameter inversely proportional to ηT. Because of smoothness and Lipschitz Assumption 3 of f, it turns out that the training objective is $L_{\widehat{F}}$-smooth. Hence, by the descent lemma of GD for smooth functions, the same upper bound holds in Eq. (5) for the value of the loss at time T, as well. Moreover, the theorem provides a uniform upper bound of the norm of all GD iterates in terms of $\|w - w_0\|$. Notably, and despite the non-convexity in our setting, our bounds are same up to constants to analogous bounds for logistic linear regression in Shamir (2021); Schliserman and Koren (2022). As discussed in Sec. 3.1 this is possible thanks to the self-bounded weak convexity property.

The condition $m \gtrsim \|w - w_0\|^4$ on the norm of the weights controls the maximum deviations of weights w from initialization (with respect to network width) required for our results to guarantee arbitrarily small train loss. Specifically, to get the most out of Theorem 3.2 we need to choose appropriate w that satisfies both the condition $m \gtrsim \|w - w_0\|^4$ and keeps the associated ridge-regularized loss $\widehat{F}(w) + \|w - w_0\|^2/(\eta T)$ small. This combined requirement is formalized in the neural-net realizability Assumption 5 below. As we will discuss later in Section 4, this assumption translates into an assumption on the underlying data distribution that ultimately enables the application of Theorem 3.2 to achieve vanishing training error.

Assumption 5 (NN–Realizability). *There exists a decreasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$ which measures the norm of deviations from initialization of models that achieve arbitrarily small training error.*

Formally, for almost surely all n training samples and for any sufficiently small $\varepsilon > 0$ there exists $w^{(\varepsilon)} \in \mathbb{R}^{d'}$ such that

$$\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon, \quad \textit{and} \quad g(\varepsilon) = \left\| w^{(\varepsilon)} - w_0 \right\|.$$

Since Assumption 5 holds for arbitrarily small ε, it guarantees that the model has enough capacity to interpolate the data, i.e., attain train error that is arbitrarily small (ε). Additionally, this is accomplished for model weights whose distance from initialization is managed by the function $g(\varepsilon)$. By using these model weights to select w in Theorem 3.2 we obtain train loss bounds for interpolating models.

Theorem 3.3 (Training loss under interpolation). *Let Assumptions 1-5 hold. Let* $\eta \leq \min\{\frac{1}{L_{\widehat{F}}}, g(1)^2, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ *and assume the width satisfies* $m \geq 18^2 L^2 R^4\, g(\frac{1}{T})^4$ *for a fixed training horizon* T*. Then,*

$$\widehat{F}(w_T) \;\leq \frac{2}{T} + \frac{5\, g(\frac{1}{T})^2}{2\eta T}, \tag{6}$$
$$\forall t \in [T] \;:\; \|w_t - w_0\| \;\leq\; 4\, g(\frac{1}{T}).$$

To interpret the theorem's conclusions suppose that the function $g(\cdot)$ of Assumption 5 is at most logarithmic; i.e., $g(\frac{1}{T}) = O(\log(T))$. Then, Theorem 3.3 implies that $m = \Omega(\log^4(T))$ neurons suffice to achieve train loss $\tilde{O}(\frac{1}{T})$ while GD iterates at all iterations satisfy $\|w_t - w_0\| = O(\log(T))$. In Section 4 (see also Remark 1), we will give examples of data separability conditions that guarantee the desired logarithmic growth of $g(\cdot)$ for logistic loss minimization, which in turn imply the favorable convergence guarantees described above. Under the same conditions we will show that the step-size requirement simplifies to $\eta \leq \min\{3, 1/L_{\widehat{F}}\}$ (see Corollary 4.1.1). Finally, we remark that Theorem 3.3 provides sufficient parameterization conditions under which GD with $T = \tilde{\Omega}(n)$ iterations finds weights w_T that yield an interpolating classifier and thus, achieve zero training error. To see this, assume logistic loss and observe setting $T \gtrsim n$ in Eq. (6) gives $\widehat{F}(w_T) \leq \log(2)/n$. This in turn implies that every sample loss satisfies $\widehat{F}_i(w_T) \leq \log(2)$, equivalently $y_i = \operatorname{sign}(\Phi(w_T, x_i))$.

3.3 Generalization

Our main result below bounds the generalization gap of GD for training two-layer nets with self-bounded loss functions. We remark that all expectations that appear below are over the training set.

Theorem 3.4 (Generalization gap – General bound). *Suppose Assumptions 1-4 hold. Fix any time horizon* $T \geq 1$ *and any step size* $\eta \leq 1/L_{\widehat{F}}$ *where* $L_{\widehat{F}}$ *is the objective's smoothness parameter. Let any* $w \in \mathbb{R}^{d'}$ *such that* $\|w - w_0\|^2 \geq \max\{\eta T\, \widehat{F}(w), \eta \widehat{F}(w_0)\}$*. Suppose hidden-layer width* m *satisfies* $m \geq 64^2 L^2 R^4 \|w - w_0\|^4$*. Then, the generalization gap of GD at iteration* T *is bounded as*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{8\ell^2 R^2}{n}\, \mathbb{E}\left[\eta T\, \widehat{F}(w) + 2\|w - w_0\|^2\right].$$

A few remarks regarding the theorem are in place. The theorem's assumptions are similar to those in Theorem 3.2, which bounds the training loss. The condition $\|w - w_0\|^2 \geq \max\{\eta T \widehat{F}(w), \eta \widehat{F}(w_0)\}$ needs to hold almost surely over the training data, which is non-restrictive, as in later applications of the theorem, the choice of w arises from Assumption 5. The condition $m \geq 64^2 L^2 R^4 \|w - w_0\|^4$ on the width of the network, is also the same as that of Theorem 3.2 but with a larger constant. This means that the last-iterate train loss bound from Theorem 3.2 (Eq. (5)) holds under the setting of Theorem 3.4. Hence, it applies to the expected train loss $\mathbb{E}[\widehat{F}(w_T)]$ and, combined with the generalization-gap bound, yields a bound on the expected test loss $\mathbb{E}[F(w_T)]$.

To optimize the bound, a proper w must be selected by minimizing the population version of a ridge-regularized training objective. In interpolation settings, the procedure

for selecting w follows the same guidelines as in Assumption 5 and in a similar style as obtaining Theorem 3.3.

Theorem 3.5 (Generalization gap under interpolation)**.** *Let Assumptions 1-5 hold. Fix $T \geq 1$ and let $m \geq 64^2 L^2 R^4 \, g(\frac{1}{T})^4$. Then, for any $\eta \leq \min\{\frac{1}{L_{\widehat{F}}}, g(1)^2, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ the expected generalization gap at iteration T satisfies*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{24\ell^2 R^2 \, g(\frac{1}{T})^2}{n}\,. \tag{7}$$

Note the width condition is similar in order to that of Theorem 3.3. Thus, provided $g(\frac{1}{T}) \lesssim \log(T)$ (see Remark 1 and Section 4 for examples), we have generalization gap of order $\tilde{O}(\frac{1}{n})$ with $m = \Omega(\log^4(T))$ neurons. Combined with the training loss guarantees from Theorem 3.3, we have test loss rate $\tilde{O}(\frac{1}{T} + \frac{1}{n})$. This further implies that with $m \approx \log^4(n)$ neurons and $T = n$ iterations, the test loss reaches the optimal rate of $\tilde{O}(\frac{1}{n})$. On the other hand, previous stability-based generalization bounds (e.g., Richards and Rabbat (2021)) required polynomial width $m \gtrsim T^2$ and eventually obtained sub-optimal generalization rates of order $O(\frac{T}{n})$. We further discuss the technical novelties resulting in these improvements in Section 5.

Remark 1 (Example: Linearly-separable data)**.** *Consider logistic-loss minimization,* tanh *activation $\sigma(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$ and data distribution that is linearly separable with margin γ, i.e., for almost surely all n samples there exists unit-norm vector $v^\star \in \mathbb{R}^d$ such that $\min_{i\in[n]} y_i \langle v^\star, x_i\rangle = \gamma$. We initialize the weights to zero, i.e. $w_0 = 0$ and show that the realizability Assumption 5 naturally holds in this setting. To see this, for any fixed $\varepsilon > 0$, set $\alpha = 2\log(1/\varepsilon)/\gamma\sqrt{m}$ and assume $m \geq 4\log^2(1/\varepsilon)$. With this choice, select weights $w_j^{(\varepsilon)} := \alpha v^\star, a_j = \frac{1}{\sqrt{m}}$ for $j \in [1, \cdots, \frac{m}{2}]$ and $w_j^{(\varepsilon)} := -\alpha v^\star, a_j = \frac{-1}{\sqrt{m}}$ for $j \in \{\frac{m}{2} + 1, \cdots, m\}$. Then, the model output for any sample (x_i, y_i) satisfies*

$$\begin{aligned} y_i \Phi(w^{(\varepsilon)}, x_i) &= \frac{y_i\sqrt{m}}{2}\left(\sigma(\alpha\langle v^\star, x_i\rangle) - \sigma(-\alpha\langle v^\star, x_i\rangle)\right) \\ &= y_i\sqrt{m}\,\sigma(\alpha\langle v^\star, x_i\rangle) \geq \sqrt{m}\,\sigma(\alpha\gamma) \geq \frac{\sqrt{m}}{2}\alpha\gamma = \log(1/\varepsilon), \end{aligned}$$

where the second equality uses the fact that tanh is odd, the first inequality follows by the increasing nature of tanh and data separability, and the last inequality follows since $\alpha\gamma \leq 1$ and $\sigma(u) \geq u/2$ for all $u \in [0, 1]$. Thus, the loss satisfies $\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon$ since for the logistic function $\log(1 + e^u) \leq e^u$. Moreover, our choice of α implies $g(\varepsilon) = \|w^{(\varepsilon)} - w_0\| = \|w^{(\varepsilon)}\| = \alpha\sqrt{m} = 2\log(1/\varepsilon)/\gamma$. To conclude, the NN-Realizability Assumption 5 holds with $g(\varepsilon) = 2\log(1/\varepsilon)/\gamma$ and thus applying Theorems 3.3 and 3.5 shows that with $m = \Omega(\log^4(T))$ neurons, the training loss and generalization gap are respectively bounded by $\tilde{O}(\frac{1}{\gamma^2 T})$ and $\tilde{O}(\frac{1}{\gamma^2 n})$, which are known to be optimal under this data assumption. We note that the same conclusion as above holds for other smooth activations such as Softmax or GELU.

4. On Realizability of NTK-Separable Data

In this section, we interpret our results for NTK-separable data by showing that our realizability condition holds for this class. We recall the definition of NTK-separability below Nitanda et al. (2019); Chen et al. (2020); Cao and Gu (2020).

Assumption 6 (Separability by NTK)**.** *For almost surely all n training samples from the data distribution there exists $w^\star \in \mathbb{R}^{d'}$ and $\gamma > 0$ such that $\|w^\star\| = 1$ and for all $i \in [n]$,*

$$y_i \left\langle \nabla_1 \Phi(w_0, x_i), w^\star \right\rangle \geq \gamma. \tag{8}$$

We also assume a bound on the model's output at initialization. Similar assumptions, but for the value of the loss, also appear in prior works that study generalization using the algorithmic stability framework Richards and Kuzborskij (2021); Lei et al. (2022).

Assumption 7 (Initialization bound)**.** *There exists parameter C such that $\forall i \in [n]$: $|\Phi(w_0, x_i)| \leq C$, for almost surely all n training samples from the data distribution*

The next proposition relates the NTK-separability assumption to our realizability assumption. The proofs for this section are given in Appendix C.

Proposition 4.1 (Realizability of NTK-separable data)**.** *Let Assumptions 1-2,6-7 hold. Assume $f(\cdot)$ to be the logistic loss. Fix $\varepsilon > 0$ and let $m \geq \frac{L^2 R^4}{4\gamma^4 C^2}(2C + \log(1/\varepsilon))^4$. Then the realizability Assumption 5 holds with $g(\varepsilon) = \frac{1}{\gamma}(2C + \log(1/\varepsilon))$. In other words, there exists $w^{(\varepsilon)}$ such that*

$$\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon, \quad \textit{and} \quad \left\| w^{(\varepsilon)} - w_0 \right\| = \frac{1}{\gamma} \left(2C + \log(1/\varepsilon)\right). \tag{9}$$

Having established realizability, the following is an immediate corollary of the general results presented in the last section.

Corollary 4.1.1 (Results under NTK-separability)**.** *Let Assumptions 1-2,6-7 hold and assume logistic loss. Suppose $m \geq \frac{64^2 L^2 R^4}{\gamma^4}(2C + \log(T))^4$ for a fixed training horizon T. Then for any $\eta \leq \min\{3, \frac{1}{L_{\widehat{F}}}\}$, the training loss and generalization gap are bounded as follows:*

$$\widehat{F}(w_T) \leq \frac{5(2C + \log(T))^2}{\gamma^2 \eta T},$$
$$\mathbb{E}\left[F(w_T) - \widehat{F}(w_T)\right] \leq \frac{24\ell^2 R^2}{\gamma^2 n} \left(2C + \log(T)\right)^2.$$

A few remarks are in place regarding the corollary. By Corollary 4.1.1, we can conclude that the expected generalization rate of GD on logistic loss and NTK-separable data as per Assumption 6 is $\tilde{O}(\frac{1}{n})$ provided width $m = \Omega(\log^4(T))$. Moreover, the expected training loss is $\mathbb{E}[\widehat{F}(w_T)] = \tilde{O}(\frac{1}{T})$. Thus, the expected test loss after T steps is $\tilde{O}(\frac{1}{T} + \frac{1}{n})$. In particular for $T = \Omega(n)$, the expected test loss becomes $\tilde{O}(\frac{1}{n})$. This rate is optimal with respect to sample size and only requires polylogarithmic hidden width with respect to n, specifically, $m = \Omega(\log^4(n))$. Notably, it represents an improvement over prior stability

results, e.g., (Richards and Rabbat, 2021) which required polynomial width and yielded suboptimal generalization rates of order $O(T/n)$. It is worth noting that the test loss bound's dependence on the margin, particularly the $\frac{1}{\gamma^2 n}$-rate obtained in our analysis, bears similarity to the corresponding results in the convex setting of linearly separable data recently established in (Shamir, 2021; Schliserman and Koren, 2022). Additionally, our results improve upon corresponding bounds for neural networks obtained via Rademacher complexity analysis (Ji and Telgarsky, 2020a; Chen et al., 2020) which yield generalization rates $\tilde{O}(\frac{1}{\sqrt{n}})$. Moreover, these works have a γ^{-8} dependence on margin for the minimum network width, whereas in Corollary 4.1.1 this is reduced to γ^{-4}. We also note that in general, both γ and C may depend on the data distribution, the data dimension, or the nature of initialization. This is demonstrated in the next section where we apply the corollary above to the noisy XOR data distribution and Gaussian initialization.

Remark 2 (Benefits of exponential tail)**.** *We have stated Corollary 4.1.1 for the logistic loss, which has an exponential tail behavior. For general self-bounded loss functions and by following the same steps, we can show a bound on generalization gap of order* $O(\frac{1}{n}(f^{-1}(\frac{1}{T}))^2)$ *provided* $m = \Omega((f^{-1}(\frac{1}{T}))^4)$*. Hence, the tail behavior of* f *controls both the generalization gap and minimum width requirement. In particular, under Assumption 6, polynomial losses with tail behavior* $f(u) \sim 1/u^\beta$ *result in generalization gap* $O(T^{2/\beta}/n)$ *for* $m = \Omega(T^{4/\beta})$*. Thus, increasing the rate of decay* β *for the loss, improves both bounds on generalization and width. This suggests the benefits of self-bounded fast-decaying losses such as exponentially-tailed loss functions for which the dependence on* T *is indeed only logarithmic.*

Example: Noisy XOR data

Next, we specialize the results of the last section to the noisy XOR data distribution Wei et al. (2019) and derive the corresponding margin and test-loss bounds. Consider the following 2^d points,

$$x_i = (x_i^1, x_i^2, \cdots, x_i^d) \in \{(1,0),(0,1),(-1,0),(0,-1)\} \times \{-1,1\}^{d-2},$$

where $\times$ denotes the Cartesian product and the labels are determined as $y_i = -1$ if $x_i^1 = 0$ and $y_i = 1$ if $x_i^1 = \pm 1$. Moreover, consider normalization $\overline{x}_i = \frac{1}{\sqrt{d-1}} x_i$ so that $R = 1$. The noisy XOR data distribution is the uniform distribution over the set with elements $(\overline{x}_i, y_i)$. For this dataset and Gaussian initialization, Ji and Telgarsky (2020a) have shown for ReLU activation that the NTK-separability assumption holds with margin $\gamma = \Omega(1/d)$. In the next result, we compute the margin for activation functions that are convex, Lipshitz and locally strongly convex.

Proposition 4.2 (Margin)**.** *Consider the noisy XOR data* $(\overline{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$*. Assume the activation function is convex,* ℓ*-Lipschitz and* μ*-strongly convex in the interval* $[-2,2]$ *for some* $\mu > 0$*, i.e.,* $\min_{t\in[-2,2]} \sigma''(t) \geq \mu$*. Moreover, assume Gaussian initialization* $w_0 \in \mathbb{R}^{d'}$ *with entries iid* $N(0,1)$*. If* $m \geq \frac{80^2 d^3 \ell^2}{2\mu^2} \log(2/\delta)$*, then with probability at least* $1-\delta$ *over the initialization, the NTK-separability Assumption 6 is satisfied with margin* $\gamma = \frac{\mu}{80d}$*.*

An interesting example of an activation function that satisfies the mentioned assumptions is the Softplus activation where $\sigma(u) = \log(1+e^u)$. This activation function has $\mu = 0.1$

and $\ell = 1$, and it is also smooth with $L = 1/4$. Therefore, the results on generalization and training loss presented in Corollary 4.1.1 hold for it. For noisy XOR data, Proposition 4.2 shows the margin in Assumption 6 is $\gamma \gtrsim 1/d$. Additionally, for standard Gaussian initialization we have by Lemma C.5 that with high-probability the initialization bound in Assumption 7 satisfies $C \lesssim \sqrt{d}$. Putting these together, and applying Corollary 4.1.1 shows that GD with n training samples reaches test loss rate $\tilde{O}(\frac{d^3}{n})$ after $T \approx n$ iterations and given $m = \tilde{\Omega}(d^6)$ neurons. It is worth noting that the number of training samples can be exponentially large with respect to d. In this case the minimum width requirement is only polylogarithmic in n.

5. Proof Sketches

We discuss here high-level proof ideas for both optimization and generalization bounds of Theorems 3.2 and 3.4. Formal proofs are deferred to Appendices A and B.

5.1 Training loss

As already discussed in Section 3.1, the key insight we use to obtain bounds that are analogous to results for optimizing convex objectives, is to exploit the self-bounded weak convexity property of the objective in Eq. (4). Thanks to this property, the Hessian minimum eigenvalue $\lambda_{\min}(\nabla^2 \widehat{F}(w_t))$ becomes less negative at the same rate at which the train loss $\widehat{F}(w_t)$ decreases.

The technical challenge at formalizing this intuition arises as follows. Controlling the rate at which $\widehat{F}(w_t)$ converges to $\widehat{F}(w)$ for the theorem's w requires controlling the Hessian at *all* intermediate points $w_{\alpha t} := \alpha w_t + (1-\alpha)w, \alpha \in [0,1]$ between w and GD iterates w_t. This is due to Taylor's theorem used to relate $\widehat{F}(w_t)$ to the target value $\widehat{F}(w)$ as follows:

$$\widehat{F}(w) \geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w - w_t \right\rangle + \frac{1}{2}\lambda_{\min}\left(\nabla^2 \widehat{F}(w_{\alpha t})\right) \left\| w - w_t \right\|^2.$$

Thus from self-bounded weak convexity, to control the last term above we need to control $\widehat{F}(w_{\alpha t})$ for any intermediate point $w_{\alpha t}$ along the GD trajectory. This is made possible by establishing the following generalized local quasi-convexity property.

Proposition 5.1 (Generalized Local Quasi-Convexity)**.** *Suppose $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$ satisfies the self-bounded weak convexity property in Eq. (4) with parameter κ. Let $w_1, w_2 \in \mathbb{R}^{d'}$ be two arbitrary points with distance $\|w_1 - w_2\| \leq D < \sqrt{2/\kappa}$. Set $\tau := \left(1 - \kappa D^2/2\right)^{-1}$. Then,*

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) \leq \tau \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{10}$$

Recall that quasi-convex functions satisfy Eq. (10) with $\tau = 1$ and D can be unboundedly large. The Proposition 5.1 indicates that our neural-net objective function is approximately quasi-convex (since $\tau > 1$) and this property holds locally, i.e. provided that w_1, w_2 are sufficiently close.

Applying (10) for $w_1 = w_t, w_2 = w$ allows controlling $\widehat{F}(w_{\alpha t})$ in terms of the train loss $\widehat{F}(w_t)$ and the target loss $\widehat{F}(w)$. The only additional requirement in Proposition 5.1 for this

to hold is that

$$1/\kappa \propto \sqrt{m} \gtrsim \|w_t - w\|^2. \tag{11}$$

This condition exactly determines the required neural-net width. Formally, we have the following.

Corollary 5.1.1 (GLQC of sufficiently wide neural nets)**.** *Let Assumptions 1,2, 4 hold. Fix arbitrary $w_1, w_2 \in \mathbb{R}^{d'}$, any constant $\lambda > 1$, and m large enough such that $\sqrt{m} \geq \lambda \frac{LR^2}{2}\|w_1 - w_2\|^2$. Then,*

$$\max_{v\in[w_1,w_2]} \widehat{F}(v) \leq (1 - 1/\lambda)^{-1} \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{12}$$

To conclude, using Corollary 5.1.1, we can show the regret bound in Eq. (5) provided (by (11)) that $\sqrt{m} \gtrsim \|w_t - w\|^2$ is true for all $t \in [T]$. To make the width requirement independent of w_t, we then use a recursive argument to prove that $\|w_t - w\| \leq 3\|w - w_0\|$. These things put together, lead to the parameter bound $\|w_t - w_0\| \leq 4\|w - w_0\|$ and the width requirement $\sqrt{m} \gtrsim \|w - w_0\|^2$ in the theorem's statement. We note that the GLQC property is also crucially required for the generalization analysis which we discuss next.

5.2 Generalization gap

We bound the generalization gap using stability analysis Bousquet and Elisseeff (2002); Hardt et al. (2016). In particular, we use (Lei and Ying, 2020a, Thm. 2) that relates the generalization gap to the "on average model stability". Formally, let $w_t^{\neg i}$ denote the t-th iteration of GD on the leave-one-out loss $\widehat{F}^{\neg i}(w) := \frac{1}{n}\sum_{j\neq i} \widehat{F}_j(w)$. As before, w_t denotes the GD output on full-batch loss $\widehat{F}$. We will use the fact (see Corollary D.2.1) that $f(y\Phi(\cdot, x))$ is $G_{\widehat{F}}$-Lipschitz with $G_{\widehat{F}} = \ell R$ under Assumptions 2 and 3.A. Then, using (Lei and Ying, 2020a, Thm. 2(a)) (cf. Lemma B.3) it holds that

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq 2G_{\widehat{F}}\, \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} \|w_T - w_T^{\neg i}\|\Big]. \tag{13}$$

In order to bound the on-average model-stability term on the right-hand side above we need to control the degree of expansiveness of GD. Recall that for convex objectives GD is non-expansive (e.g. Hardt et al. (2016)), that is $\|(w - \eta\nabla\widehat{F}(w)) - (w' - \eta\nabla\widehat{F}(w'))\| \leq \|w - w'\|$ for any w, w'. For the non-convex objective in our setting, the lemma below establishes a generalized non-expansiveness property via leveraging the structure of the objective's Hessian for the two-layer net.

Lemma 5.2 (GD-Expansiveness)**.** *Let Assumptions 1 and 2 hold. For any $w, w' \in \mathbb{R}^{d'}$, any step-size $\eta > 0$, and $w_\alpha := \alpha w + (1 - \alpha)w'$ it holds for $H(w) := \eta\frac{LR^2}{\sqrt{m}}\widehat{F}'(w) + \max\Big\{1, \eta\ell^2R^2\widehat{F}''(w)\Big\}$ that*

$$\Big\|\Big(w - \eta\nabla\widehat{F}(w)\Big) - \Big(w' - \eta\nabla\widehat{F}(w')\Big)\Big\| \leq \max_{\alpha\in[0,1]} H(w_\alpha)\ \big\|w - w'\big\|,$$

where we define $\widehat{F}'(w) := \frac{1}{n}\sum_{i=1}^{n}|f'(y_i\Phi(w, x_i))|$ and $\widehat{F}''(w) := \frac{1}{n}\sum_{i=1}^{n} f''(y_i\Phi(w, x_i))$.

This lemma can be further simplified for the class of self-bounded loss functions. Specifically, using $|f'(u)| \leq f(u)$ and $f''(u) \leq 1$ from Assumptions 4 and 3.B, we immediately deduce the following.

Corollary 5.2.1 (Expansiveness for self-bounded losses)**.** *In the setting of Lemma 5.2, further assume the loss satisfies Assumptions 3.B and 4. Provided $\eta \leq 1/(\ell^2 R^2)$, it holds for all $w, w' \in \mathbb{R}^{d'}$ that*

$$\left\| \left(w - \eta \nabla \widehat{F}(w)\right) - \left(w' - \eta \nabla \widehat{F}(w')\right) \right\| \leq \left(1 + \eta \frac{LR^2}{\sqrt{m}} \max_{\alpha \in [0,1]} \widehat{F}(w_\alpha)\right) \left\| w - w' \right\|. \tag{14}$$

In Eq. (14) the expansiveness is weaker than in a convex scenario, where the coefficient would be 1 instead of $1 + \frac{\eta LR^2}{\sqrt{m}} \max_{\alpha \in [0,1]} \widehat{F}(w_\alpha)$. However, for self-bounded losses (i.e. $|f'(u)| \leq f(u)$) the "gap to convexity" $\frac{\eta LR^2}{\sqrt{m}} \max_{\alpha \in [0,1]} \widehat{F}(w_\alpha)$ in Corollary 5.2.1 is better than the gap from Lemma 5.2 for 1-Lipschitz losses (i.e. $|f'(u)| \leq 1$), which would be $\frac{\eta LR^2}{\sqrt{m}}$. Indeed, after unrolling the GD iterates, the latter eventually leads to polynomial width requirements Richards and Rabbat (2021).

Instead, to obtain a polylogarithmic width, we use the expansiveness bound in Eq. (14) for self-bounded losses together with the generalized-local quasi-convexity property in Corollary 5.1.1 as follows. From Corollary 5.1.1, if m is large enough such that

$$\sqrt{m} \geq LR^2 \|w_t - w_t^{\neg i}\|^2, \qquad \forall t \in [T],\ \forall i \in [n],$$

then Eq. (12) holds on the GD path. This further simplifies the result of Corollary 5.2.1 applied for $w = w_t, w' = w_t^{\neg i}$ into

$$\left\| (w_t - \eta \nabla \widehat{F}^{\neg i}(w_t)) - (w_t^{\neg i} - \eta \nabla \widehat{F}^{\neg i}(w_t^{\neg i})) \right\| \leq \widetilde{H}_t^i \left\| w_t - w_t^{\neg i} \right\|,$$

where $\widetilde{H}_t^i := 1 + \frac{2\eta LR^2}{\sqrt{m}} \max\{\widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i})\}$. Now from the optimization analyses in Sec. 5.1, we know intuitively that $\widehat{F}^{\neg i}(w_t) \leq \widehat{F}(w_t)$ decays at rate $\tilde{O}(1/t)$; thus, so does $\widehat{F}^{\neg i}(w_t^{\neg i})$. Therefore, for all $i \in [n]$ the expansivity coefficient $\widetilde{H}_t^i$ in the above display is decaying to 1 as GD progresses.

To formalize all these and connect them to the model-stability term in (13), note using triangle inequality and the Gradient Self-boundedness property of Lemma 3.1 that

$$\left\| w_{t+1} - w_{t+1}^{\neg i} \right\| \leq \left\| (w_t - \eta \nabla \widehat{F}^{\neg i}(w_t)) - (w_t^{\neg i} - \eta \nabla \widehat{F}^{\neg i}(w_t^{\neg i})) \right\| + \frac{\eta \ell R}{n} \widehat{F}_i(w_t)\,.$$

Unrolling this display over $t \in [T]$, averaging over $i \in [n]$, and using our expansiveness bound above we show in Appendix B the following bound for the model stability term

$$\frac{1}{n} \sum_{i=1}^{n} \left\| w_T - w_T^{\neg i} \right\| \leq \frac{\eta \ell R e^{\beta}}{n} \sum_{t=0}^{T-1} \widehat{F}(w_t)\,, \tag{15}$$

where $\beta \lesssim \left(\sum_{t=1}^{T} \widehat{F}(w_t) + \sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t^{\neg i})\right) / \sqrt{m}$. But, we know from training-loss bounds in Theorem 3.2 that $\sum_{t=1}^{T} \widehat{F}(w_t) \lesssim \|w - w_0\|^2$ (and similar for $\sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t^{\neg i})$). Thus,

$\beta \lesssim \|w - w_0\|^2/\sqrt{m}$. At this point, the theorem's conditions guarantees $\sqrt{m} \gtrsim \|w - w_0\|^2$, so that $\beta = O(1)$. Plugging back in (15) we conclude with the following stability bound: $\frac{1}{n}\sum_{i=1}^n \|w_T - w_T^{\neg i}\| \lesssim \sum_{t=0}^T \widehat{F}(w_t)/n$. Applying the train-loss bounds of Theorem 3.2 once more completes the proof.

6. Prior Works

The theoretical study of generalization properties of neural networks (NN) is more than two decades old (Bartlett, 1996; Bartlett et al., 1998). Recently, there has been an increased interest in understanding and improving generalization of SGD/GD on over-parameterized neural networks, e.g. (Allen-Zhu et al., 2019a; Oymak and Soltanolkotabi, 2020; Javanmard et al., 2020; Richards and Rabbat, 2021). These results however typically require very large width where $m = \text{poly}(n)$. We discuss most-closely related-works below.

Quadratic loss. For quadratic loss, Li and Liang (2018); Soltanolkotabi et al. (2018); Allen-Zhu et al. (2019b); Zou and Gu (2019); Liu et al. (2022) showed that sufficiently over-parameterized neural networks of polynomial width satisfy a local Polyak-Łojasiewicz (PL) condition $\|\nabla \widehat{F}(w)\|^2 \geq 2\mu(\widehat{F}(w) - \widehat{F}^\star)$, where μ is at least the smallest eigenvalue of the neural tangent kernel matrix. The PL property in this case implies that the training loss converges linearly with the rate $\widehat{F}(w_t) = O((1-\eta\mu)^t)$ if the GD iterates remain in the PL region. Moreover, Charles and Papailiopoulos (2018); Lei and Ying (2020b), have used the PL condition to further characterize stability properties of corresponding non-convex models. Notably, Lei and Ying (2020b) derived order-optimal rates $O(\frac{1}{\mu n})$ for the generalization loss. However these rates only apply to quadratic loss. Models trained with logistic or exponential loss on separable data do *not* satisfy the PL condition even for simple interpolating linear models. Aside from the PL condition-related results, but again for quadratic loss, Oymak et al. (2019) showed under specific assumptions on the data translating to low-rank NTK, that logarithmic width is sufficient to obtain classification error of order $O(n^{-1/4})$. In general, they achieve error rate $O(n^{-1/2})$, but for $m = \tilde{\Omega}(n^2)$.

Logistic-loss minimization with linear models. Logistic-loss minimization is more appropriate for classification and rate-optimal generalization bounds for GD have been obtained recently in the linear setting, where the training objective is convex. In particular, for linear logistic regression on data that are linearly separable with margin $\gamma > 0$, Shamir (2021) proved a finite-time test-error bound $O(\frac{\log^2 T}{\gamma^2 T} + \frac{\log^2 T}{\gamma^2 n})$. Ignoring log factors, this is order-optimal with the sample size n and training horizon T. Their proof uses exponential-decaying properties of the logistic loss to control the norm of gradient iterates, which it cleverly combines with Markov's inequality to bound the fraction of well-separated datapoints at any iteration. This in turn translates to a test-error bound by standard margin-based generalization bounds. More recently, Schliserman and Koren (2022) used algorithmic-stability analysis proving same rates (up to log factors) for the test loss. Their results hold for general convex, smooth, self-bounded and decreasing objectives under a realizability assumption suited for convex objectives (analogous to Assumption 5). Specifically, this includes linear logistic regression with linearly separable data. Here, we show that analogous rates on the test loss hold true for more complicated nonconvex settings where data are separable by shallow neural networks.

Stability of GD in NN. State-of-the-art generalization bounds on shallow neural networks via the stability-analysis framework have appeared very recently in (Richards and Rabbat, 2021; Richards and Kuzborskij, 2021; Lei et al., 2022). For Lipschitz losses, Richards and Rabbat (2021) shows that the empirical risk is weakly convex with a weak-convexity parameter that improves as the neural-network width m increases. Leveraging this observation, they establish stability bounds for GD iterates at time T provided sufficient parameterization $m = \tilde{\Omega}(T^2)$. Since the logistic loss is Lipschitz, these bounds also apply to our setting. Nevertheless, our work improves upon Richards and Rabbat (2021) in that: (i) we require significantly smaller width, poly-logarithmic rather than polynomial, and (ii) we show $\tilde{O}(1/n)$ test loss bounds in the realizable setting, while their bounds are $O(T/n)$. Central to our improvements is a largely refined analysis of the curvature of the loss via identifying and proving a generalized quasi-convexity property for neural networks of polylogarithmic width trained with self-bounded losses (see Section 5 for details). Our results also improve upon the other two works Richards and Kuzborskij (2021); Lei et al. (2022), which both require polynomial widths. However, we note that these results are not directly comparable since Richards and Kuzborskij (2021); Lei et al. (2022) focus on quadratic-loss minimization. See also Appendix E.

Uniform convergence in NN. Uniform bounds on the generalization loss have been derived in literature via Rademacher complexity analysis (Bartlett and Mendelson, 2002); see for example (Neyshabur et al., 2015; Arora et al., 2019; Golowich et al., 2020; Vardi et al., 2022; Frei et al., 2022a) for a few results in this direction. These works typically obtain the bounds of order $O(\frac{\mathcal{R}}{\sqrt{n}})$, where $\mathcal{R}$ depends on the Rademacher complexity of the hypothesis space. Recent works by Ji and Telgarsky (2020a); Chen et al. (2020) also utilized Rademacher complexity analysis to obtain test loss rates of $O(1/\sqrt{n})$ under an NTK separability assumption (see also (Nitanda et al., 2019)) with polylogarithmic width requirement for shallow and deep networks, respectively. Instead, while maintaining minimal width requirements, we obtain test-loss rates $\tilde{O}(1/n)$, which are order-optimal. Our approach, which is based on algorithmic-stability, is also different and uncovers new properties of the optimization landscape, including a generalized local quasi-convexity property. On the other hand, the analysis in (Ji and Telgarsky, 2020a; Chen et al., 2020) applies to ReLU activation and bounds the test loss with high-probability over the sampling of the training set. Instead, we require smooth activations similar to other studies such as (Oymak et al., 2019; Chatterji et al., 2021; Bai and Lee, 2020; Nitanda et al., 2019; Richards and Rabbat, 2021; Richards and Kuzborskij, 2021; Lei et al., 2022) and we bound the test loss in expectation over the training set. Finally, we also note that data-specific generalization bounds for two-layer nets have also appeared recently in (Cao et al., 2022; Frei et al., 2022b). However, those results require that data are nearly-orthogonal.

Convergence/implicit bias of GD. Convergence and implicit bias of GD for logistic/exponential loss functions on linear models and neural networks have been investigated in (Ji and Telgarsky, 2018; Soudry et al., 2018; Nacson et al., 2019; Lyu and Li, 2020; Chizat and Bach, 2020; Chatterji et al., 2021; Taheri and Thrampoulidis, 2023). Early works by (Zou et al., 2020; Cao and Gu, 2019) have proved convergence and generalization of deep networks trained by logistic loss under polynomial width conditions. Moreover, Lyu and Li (2020); Ji and Telgarsky (2020b) have shown for homogeneous neural-networks that GD converges in direction to a max-margin solution. While certainly powerful, this implicit-bias

convergence characterization becomes relevant only when the number T of GD iterations is exponentially large. Instead, our convergence bounds apply for finite T (on the order of sample size), thus are more practically relevant. Moreover, their results assume a GD iterate t_0 such that $\widehat{F}(w_{t_0}) \leq \log(2)/n$. Similar assumption appears in (Chatterji et al., 2021), which require initialization $\widehat{F}(w_0) \leq 1/n^{1+C}$ for constant $C > 0$. Our approach is entirely different: we prove that sufficient parameterization benefits the loss curvature and suffices for GD steps to find an interpolating model and attain near-zero training loss, provided data satisfy an appropriate realizability condition.

7. Conclusions

In this paper we study smooth shallow neural networks trained with self-bounded loss functions, such as logistic loss. Under interpolation, we provide minimal sufficient parameterization conditions to achieve rate-optimal generalization and optimization bounds. These bounds improve upon prior results which require substantially large over-parameterization or obtain sub-optimal generalization rates. Specifically, we significantly improve previous stability-based analyses in terms of both relaxing the parameterization requirements and obtaining improved rates. Although our focus was on binary classification with shallow networks, our approach can potentially be extended to other architectures such as transformers; for preliminary results in this direction see Deora et al. (2023). Extending our results to the stochastic case by analyzing SGD is another important future direction. Moreover, as the width condition $m = \Omega(\log^4(T))$ depends on the time horizon, early stopping is necessary for obtaining bounded width conditions. It is interesting to investigate whether this temporal dependence can be removed. Furthermore, while our current treatment relies on smoothness of the activation function to exploit properties of the curvature of the training objective, we aim to examine the potential of our results to extend to non-smooth activations. Finally, our generalization analysis bounds the expectation of the test loss (over data sampling) and it is an important future direction extending these guarantees to a high-probability setting.

Acknowledgments

This work was partially supported by NSF Grant CCF-2009030.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019b.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*, 2020.

Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L. Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc dimension bounds for piecewise polynomial networks. NIPS'98, page 190–196. MIT Press, 1998.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3349–3356, 2020.

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in Neural Information Processing Systems*, 2022.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

Niladri S Chatterji, Philip M Long, and Peter L Bartlett. When does gradient descent with logistic loss find interpolating two-layer networks? *The Journal of Machine Learning Research*, 22(1):7135–7182, 2021.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? In *International Conference on Learning Representations*, 2020.

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

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*, 2023.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *arXiv preprint arXiv:2202.07626*, 2022a.

Spencer Frei, Gal Vardi, Peter L Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*, 2022b.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9(2):473–504, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6), 2020.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020a.

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

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020a.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2020b.

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and generalization analysis of gradient methods for shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in Neural Information Processing Systems*, 34:8609–8621, 2021.

Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR, 2021.

Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3380–3394. PMLR, 02–05 Jul 2022.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

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

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Hossein Taheri and Christos Thrampoulidis. On generalization of decentralized learning with separable data. In *International Conference on Artificial Intelligence and Statistics*, pages 4917–4945. PMLR, 2023.

Gal Vardi, Ohad Shamir, and Nathan Srebro. The sample complexity of one-hidden-layer neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

Appendix A. Training Loss Analysis

This section includes the proofs of the results stated in Section 3.2.

A.1 Proof of Theorem 3.2

We begin with proving the general train-loss and parameter-norm bounds of Theorem 3.2. In fact, we state and prove a slightly more general statement of the theorem which includes non-smooth and non-Lipschitz losses (such as expoential loss) that satisfy a second order self-bounded property described below.

Assumption 8 (2nd order self-boundedness). *The convex loss function $f : \mathbb{R} \to \mathbb{R}_+$ satisfies the 2nd order self-boundedness property, i.e.*

$$f''(u) \leq f(u), \forall u \in \mathbb{R}.$$

Theorem A.1 (General statement of Theorem 3.2). *Let Assumptions 1-2 hold. Assume the loss function satisfies self-bounded Assumption 4. Moreover, suppose either Assumption 3 or Assumption 8 hold. Fix any $T \geq 0$. Let the step-size satisfy the assumptions of the descent lemma (Lemma A.2). Assume any w and m such that $\|w - w_0\|^2 \geq \max\left\{\eta T \widehat{F}(w), \eta \widehat{F}(w_0)\right\}$ and $m \geq 18^2 L^2 R^4 \|w - w_0\|^4$. Then, the training loss and the parameters' norm satisfy*

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \leq 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T}, \tag{16}$$
$$\forall t \in [T] \quad : \quad \|w_t - w_0\| \leq 4\|w - w_0\|.$$

To prove Theorem A.1, we first state our descent lemma for both self-bounded losses and lipschitz-smooth losses.

Lemma A.2 (Descent lemma). *Let Assumptions 1-2 hold. Assume the loss function satisfies self-boundedness Assumptions 4,8. Then, for any $\eta < \frac{1}{R^2 \widehat{F}(w_t)} \min\{\frac{1}{\ell^2 + L}, \frac{1}{\sqrt{L}\ell}\}$ the descent property holds, i.e.,*

$$\widehat{F}(w_{t+1}) \leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla \widehat{F}(w_t)\|^2.$$

Moreover, if f satisfies Assumption 3 then the descent property holds for any $\eta \leq 1/L_{\widehat{F}}$ where $L_{\widehat{F}} := \ell^2 R^2 + \frac{LR^2}{\sqrt{m}}$ is the smoothness parameter of the training objective.

Proof Due to self-boundedness Assumption 8, as well as Assumptions 1-2 the objective is also self-bounded according to Corollary D.2.1, i.e., $\|\nabla^2 \widehat{F}(w)\| \leq \left(\ell^2 R^2 + \frac{LR^2}{\sqrt{m}}\right) \widehat{F}(w)$, $\|\nabla \widehat{F}(w)\| \leq \ell R \, \widehat{F}(w)$.

By Taylor's expansion, there exists a $w' \in [w_t, w_{t+1}]$ such that,

$$\begin{aligned}
\widehat{F}(w_{t+1}) &= \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w_{t+1} - w_t \right\rangle + \frac{1}{2}\left\langle w_{t+1} - w_t, \nabla^2 \widehat{F}(w')\,(w_{t+1} - w_t) \right\rangle \\
&\leq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w_{t+1} - w_t \right\rangle + \frac{1}{2} \max_{v \in [w_t, w_{t+1}]} \left\| \nabla^2 \widehat{F}(v) \right\| \cdot \|w_{t+1} - w_t\|^2 \\
&\leq \widehat{F}(w_t) - \eta \|\nabla \widehat{F}(w_t)\|^2 + \frac{\eta^2 \left(\ell^2 R^2 + \frac{LR^2}{\sqrt{m}}\right)}{2} \max_{v \in [w_t, w_{t+1}]} \widehat{F}(v) \cdot \left\| \nabla \widehat{F}(w_t) \right\|^2.
\end{aligned}$$

By Corollary A.7.1, for $\sqrt{m} \geq \eta^2 L\ell^2 R^4 \widehat{F}^2(w_t) \geq LR^2\|\eta\nabla\widehat{F}(w_t)\|^2 = LR^2\|w_{t+1}-w_t\|^2$ it holds that

$$\max_{v\in[w_t,w_{t+1}]} \widehat{F}(v) \leq 2\max\{\widehat{F}(w_t),\widehat{F}(w_{t+1})\},$$

which yields

$$\widehat{F}(w_{t+1}) \leq \widehat{F}(w_t) - \eta\|\nabla\widehat{F}(w_t)\|^2 + \eta^2\left(\ell^2R^2 + \frac{LR^2}{\sqrt{m}}\right)\max\left\{\widehat{F}(w_t),\widehat{F}(w_{t+1})\right\}\cdot\|\nabla\widehat{F}(w_t)\|^2. \tag{17}$$

We note that the condition on m simplifies to $m \geq 1$ if $\eta \leq \frac{1}{\sqrt{L}\ell R^2}\frac{1}{\widehat{F}(w_t)}$.

Back to (17), if $\widehat{F}(w_{t+1}) \geq \widehat{F}(w_t)$ by our condition $\eta < \frac{1}{\ell^2R^2+LR^2/\sqrt{m}}\frac{1}{\widehat{F}(w_t)}$ it holds that

$$\begin{aligned}\widehat{F}(w_{t+1}) &\leq \widehat{F}(w_t) + \eta\|\nabla\widehat{F}(w_t)\|^2\left(\frac{\widehat{F}(w_{t+1})}{\widehat{F}(w_t)} - 1\right)\\ &\leq \widehat{F}(w_t) + \eta\ell^2R^2\widehat{F}^2(w_t)\left(\frac{\widehat{F}(w_{t+1})}{\widehat{F}(w_t)} - 1\right).\end{aligned}$$

Since $\eta < \frac{1}{\ell^2R^2}\frac{1}{\widehat{F}(w_t)}$,

$$\begin{aligned}\widehat{F}(w_{t+1}) &< \widehat{F}(w_t) + \widehat{F}(w_t)\left(\frac{\widehat{F}(w_{t+1})}{\widehat{F}(w_t)} - 1\right)\\ &= \widehat{F}(w_{t+1}),\end{aligned}$$

which is a contradiction. Thus it holds that $\widehat{F}(w_{t+1}) < \widehat{F}(w_t)$. Continuing from Eq. (17) with the assumption $\eta < \frac{1}{\ell^2R^2+LR^2/\sqrt{m}}\frac{1}{\widehat{F}(w_t)}$, we conclude that

$$\begin{aligned}\widehat{F}(w_{t+1}) &\leq \widehat{F}(w_t) - \eta\|\nabla\widehat{F}(w_t)\|^2 + \frac{1}{2}\eta^2\left(\ell^2R^2 + \frac{LR^2}{\sqrt{m}}\right)\widehat{F}(w_t)\cdot\|\nabla\widehat{F}(w_t)\|^2\\ &\leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla\widehat{F}(w_t)\|^2.\end{aligned}$$

This completes the proof for self-bounded losses.

Next, suppose f is 1-smooth and 1-Lipschitz. Then, as per Corollary D.2.1, $\widehat{F}$ is smooth with the constant

$$L_{\widehat{F}} := \ell^2R^2 + \frac{LR^2}{\sqrt{m}}.$$

Following similar steps as in the beginning of proof and assuming step-size $\eta \leq 1/L_{\widehat{F}}$ we immediately conclude that,

$$\begin{aligned}\widehat{F}(w_{t+1}) &\leq \widehat{F}(w_t) - \eta\|\nabla\widehat{F}(w_t)\|^2 + \frac{\eta^2L_{\widehat{F}}}{2}\|\nabla\widehat{F}(w_t)\|^2\\ &\leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla\widehat{F}(w_t)\|^2.\end{aligned}$$

This completes the proof. ■

As a remark, the descent property implies that the loss decreases by each step, i.e., $\widehat{F}(w_t) \leq \widehat{F}(w_0)$. Thus for self-bounded losses the condition $\eta < \frac{1}{R^2\,\widehat{F}(w_0)} \min\{\frac{1}{\ell^2+L}, \frac{1}{\sqrt{L}\ell}\}$ is sufficient. We also note that the Lipschitz-smoothness and 2nd order self-bounded assumptions are only required for the descent lemma above, which results in conditions on the step-size based on the properties of loss. In the rest of the proof we only use the self-bounded Assumption 4 in order to use the self-bounded weak convexity property of the objective (see Def. 1).

Next lemma finds a general relation for the training loss in terms of an arbitrary point $w \in \mathbb{R}^{d'}$ and the fluctuations of loss between w and GD iterates w_t.

Lemma A.3. *Let Assumptions 1-2 hold. Assume the loss function satisfies the self-bounded Assumption 4. Moreover, suppose $\widehat{F}$ and step-size η are such that the following descent condition is satisfied for all $t \geq 0$:*

$$\widehat{F}(w_{t+1}) \leq \widehat{F}(w_t) - \frac{\eta}{2}\|\nabla \widehat{F}(w_t)\|^2. \tag{18}$$

Then, for any $w \in \mathbb{R}^{d'}$ it holds that

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \leq \widehat{F}(w) + \frac{\|w - w_0\|^2}{\eta T} + \frac{1}{2}\frac{LR^2}{\sqrt{m}}\frac{1}{T}\sum_{t=0}^{T-1} \max_{\alpha\in[0,1]} \widehat{F}(w_{\alpha t})\, \|w - w_t\|^2,$$

where we set $w_{\alpha t} := \alpha w_t + (1-\alpha)w$.

Proof

Fix any w. By Taylor, there exists $w_{\alpha t}, \alpha \in [0,1]$ such that

$$\begin{aligned}
\widehat{F}(w) &= \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w - w_t \right\rangle + \frac{1}{2}\left\langle w - w_t, \nabla^2 \widehat{F}(w_{\alpha t})\,(w - w_t) \right\rangle \\
&\geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w - w_t \right\rangle + \frac{1}{2}\lambda_{\min}\left(\nabla^2 \widehat{F}(w_{\alpha t})\right)\|w - w_t\|^2 \\
&\geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w - w_t \right\rangle - \frac{1}{2}\frac{LR^2}{\sqrt{m}}\widehat{F}(w_{\alpha t})\,\|w - w_t\|^2.
\end{aligned}$$

The last line is true by Corollary D.2.1. Thus, for any w,

$$\widehat{F}(w) \geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w - w_t \right\rangle - \frac{1}{2}\frac{LR^2}{\sqrt{m}} \max_{\alpha\in[0,1]} \widehat{F}(w_{\alpha t})\,\|w - w_t\|^2.$$

Plugging this in (18) gives

$$\begin{aligned}
\widehat{F}(w_{t+1}) &\leq \widehat{F}(w) - \left\langle \nabla \widehat{F}(w_t), w - w_t \right\rangle - \frac{\eta}{2}\left\|\nabla \widehat{F}(w_t)\right\|^2 + \frac{1}{2}\frac{LR^2}{\sqrt{m}} \max_{\alpha\in[0,1]} \widehat{F}(w_{\alpha t})\,\|w - w_t\|^2 \\
&= \widehat{F}(w) + \frac{1}{\eta}\left(\|w - w_t\|^2 - \|w - w_{t+1}\|^2\right) + \frac{1}{2}\frac{LR^2}{\sqrt{m}} \max_{\alpha\in[0,1]} \widehat{F}(w_{\alpha t})\,\|w - w_t\|^2.
\end{aligned} \tag{19}$$

where the second line follows by completion of squares using $w_{t+1} - w_t = -\eta \nabla \widehat{F}(w_t)$.

Telescoping the above display for $t = 0, \dots, T-1$, we arrive at the desired. ■

Next, when m is large enough so that we can invoke the generalized-local quasi-convexity property, the bound of Lemma A.3 takes the following convenient form

Lemma A.4. *Let the assumptions of Lemma A.3 hold. Assume w and m such that $\sqrt{m} \geq 2LR^2\|w - w_t\|^2$ for all $t \in [T-1]$ then*

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \leq 2\widehat{F}(w) + \frac{2\|w - w_0\|^2}{\eta T} + \frac{\widehat{F}(w_0)}{2T}. \tag{20}$$

Proof We invoke Corollary A.7.1 with $\lambda = 4$ to deduce that for all $t \in [T-1]$

$$\max_{\alpha \in [0,1]} \widehat{F}(w_{\alpha t}) \leq \frac{4}{3}\max\{\widehat{F}(w), \widehat{F}(w_t)\} < \frac{4}{3}\widehat{F}(w_t) + \frac{4}{3}\widehat{F}(w). \tag{21}$$

Noting the assumption on m and recalling Lemma A.3,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) &\leq \widehat{F}(w) + \frac{\|w - w_0\|^2}{\eta T} + \frac{1}{2}\frac{LR^2}{\sqrt{m}}\frac{1}{T}\sum_{t=0}^{T-1} \max_{\alpha \in [0,1]} \widehat{F}(w_{\alpha t})\, \|w - w_t\|^2 \\
&\leq \frac{4}{3}\widehat{F}(w) + \frac{\|w - w_0\|^2}{\eta T} + \frac{1}{3T}\sum_{t=0}^{T-1} \widehat{F}(w_t) \\
&\leq \frac{4}{3}\widehat{F}(w) + \frac{\|w - w_0\|^2}{\eta T} + \frac{1}{3T}\sum_{t=0}^{T} \widehat{F}(w_t).
\end{aligned}$$

Arranging terms yields the desired result. ■

Finally, using the about bounds on the training loss, we can bound the parameter-norm using a recursive argument presented in the lemma below.

Lemma A.5 (Iterates-norm bound)**.** *Suppose the assumptions of Lemma A.3 hold. Fix any $T \geq 0$ and assume any w and m such that*

$$\|w - w_0\|^2 \geq \max\{\eta T \widehat{F}(w), \eta \widehat{F}(w_0)\}. \tag{22}$$

and

$$\sqrt{m} \geq 18LR^2\|w - w_0\|^2, \tag{23}$$

Then, for all $t \in [T]$,

$$\|w_t - w\| \leq 3\|w - w_0\|. \tag{24}$$

Proof Denote $A_t = \|w_t - w\|$. Start by recalling from (19) that for all t:

$$A_{t+1}^2 \leq A_t^2 + \eta \widehat{F}(w) - \eta \widehat{F}(w_{t+1}) + \eta \frac{LR^2}{2\sqrt{m}} \max_{\alpha \in [0,1]} \widehat{F}(w_{\alpha t})\, A_t^2. \tag{25}$$

We will prove the desired statement (24) using induction. For $t = 0$, $A_0 = \|w - w_0\|$. Thus, the assumption of induction holds. Now assume (24) is correct for $t \in [T-1]$, i.e. $A_t \leq 3\|w - w_0\|, \forall t \in [T-1]$. We will then prove it holds for $t = T$.

The first observation is that by induction hypothesis $\sqrt{m} \geq 18LR^2\|w - w_0\|^2 \geq 2LR^2 A_t^2$ for all $t \in [T-1]$. Thus, for all $t \in [T-1]$, the condition of the generalized local quasi-convexity Corollary 5.1.1 holds for $\lambda = 4$ implying (see also (21))

$$\forall t \in [T-1] \,:\, \max_{\alpha \in [0,1]} \widehat{F}(w_{\alpha t}) \leq \frac{4}{3}\widehat{F}(w_t) + \frac{4}{3}\widehat{F}(w).$$

Using this in (25) we find for all $t \in [T-1]$ that

$$\begin{aligned} A_{t+1}^2 &\leq A_t^2 + \eta \widehat{F}(w) - \eta \widehat{F}(w_{t+1}) + \eta \frac{LR^2 \cdot A_t^2}{2\sqrt{m}} \left(\frac{4}{3}\widehat{F}(w_t) + \frac{4}{3}\widehat{F}(w) \right) \\ &\leq A_t^2 + \eta \widehat{F}(w) - \eta \widehat{F}(w_{t+1}) + \eta \left(\frac{1}{3}\widehat{F}(w_t) + \frac{1}{3}\widehat{F}(w) \right) \end{aligned}$$

where in the second inequality we used again that $\sqrt{m} \geq 2LR^2 A_t^2$. We proceed by telescoping the above display over $t = 0, 1, \ldots, T-1$ to get

$$\begin{aligned} A_T^2 &\leq A_0^2 + \frac{4}{3}\eta T \widehat{F}(w) + \frac{1}{3}\eta \widehat{F}(w_0) + \frac{1}{3}\eta \sum_{t=0}^{T-1} \widehat{F}(w_t) - \eta \widehat{F}(w_T) \\ &\leq A_0^2 + \frac{4}{3}\eta T \widehat{F}(w) + \frac{2}{3}\eta \widehat{F}(w_0) + \frac{1}{3}\eta \sum_{t=1}^{T} \widehat{F}(w_t), \end{aligned}$$

where the second line follows by nonegativity of the loss.

Now, to bound the last term above, observe that the condition of Lemma A.4 holds since $\sqrt{m} \geq 2LR^2 A_t^2$ for all $t \in [T-1]$ by induction hypothesis. Hence, using (20), we conclude that

$$\begin{aligned} A_T^2 &\leq A_0^2 + \frac{4}{3}\eta T \widehat{F}(w) + \frac{2}{3}\eta \widehat{F}(w_0) + \frac{1}{3}\eta T \left(2\widehat{F}(w) + \frac{2A_0^2}{\eta T} + \frac{\widehat{F}(w_0)}{2T} \right) \\ &= \frac{5}{3}A_0^2 + 2\eta T \widehat{F}(w) + \frac{5}{6}\eta \widehat{F}(w_0) \\ &\leq \frac{5}{3}\|w - w_0\|^2 + 2\|w - w_0\|^2 + \frac{5}{6}\|w - w_0\|^2 = \frac{9}{2}\|w - w_0\|^2 \qquad \implies A_T \leq 3\|w - w_0\|. \end{aligned} \tag{26}$$

In the last inequality, we used the assumptions of the lemma on $\|w - w_0\|$ and $A_0 = \|w - w_0\|$. This completes the proof. ■

Completing the proof of Theorem A.1.

The proof follows from combining the bounds on the training loss and parameters' growth from Lemmas A.4-A.5 and noting that with condition on $\|w-w_0\|^2$ from Lemma A.5 we have $\widehat{F}(w_0) \leq \|w-w_0\|^2/\eta$ to derive (16). Moreover, we have $\|w_t - w_0\| \leq \|w_t - w\| + \|w - w_0\| \leq 4\|w - w_0\|$.

A.2 Proof of Theorem 3.3

Here we prove training loss bound for interpolating NN as asserted by Theorem 3.3. Similar to the previous section, we prove a more general result where the loss is not necessarily Lipschitz or smooth. We are now ready to prove Theorem 3.3 for general self-bounded losses. In particular, Theorem 3.3 follows directly from the next result by choosing f to be Lipschitz and smooth.

Theorem A.6 (General statement of Theorem 3.3)**.** *Suppose Assumptions 1-2, 4 hold. Moreover, assume the objective and data satisfy the Assumption 5. Let the step-size satisfy the assumptions of Descent Lemma A.2. Moreover, assume* $\eta \leq \min\{g(1)^2, \frac{1}{L_{\widehat{F}}}, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ *and* $m \geq 18^2 L^2 R^4\, g(\frac{1}{T})^4$ *for a fixed training horizon* T*. Then,*

$$\widehat{F}(w_T) \;\leq \frac{2}{T} + \frac{5\, g(\frac{1}{T})^2}{2\eta T},$$
$$\forall t \in [T] \;:\; \|w_t - w_0\| \;\leq\; 4\, g(\frac{1}{T}).$$

Proof According to Assumption 5, for any sufficiently small $\varepsilon > 0$, there exists a $w^{(\varepsilon)}$ such that $\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon$ and $\|w^{(\varepsilon)} - w_0\| = g(\varepsilon)$. Pick $\varepsilon = 1/T$. With the condition $\eta \leq \min\{g(1)^2, g(1)^2/\widehat{F}(w_0)\}$ we have

$$\max\Big\{\eta T \widehat{F}(w^{(1/T)}), \eta \widehat{F}(w_0)\Big\} \leq g(1)^2 \leq g(\frac{1}{T})^2 = \|w^{(1/T)} - w_0\|^2,$$

where in the second inequality we used the fact that g is a decreasing function. The desired result is obtained by Theorem A.1. ■

A.3 Generalized local quasi-convexity property

In the remainder of this section, we prove the generalized local quasi-convexity property.

Proposition A.7 (Restatement of Proposition 5.1)**.** *Suppose* $\widehat{F} : \mathbb{R}^{d'} \to \mathbb{R}$ *satisfies the self-bounded weak convexity property in Eq. 4 with parameter* κ*. Let* $w_1, w_2 \in \mathbb{R}^{d'}$ *be two arbitrary points with distance* $\|w_1 - w_2\| \leq D < \sqrt{2/\kappa}$ *. Set* $\tau := \big(1 - \kappa D^2/2\big)^{-1}$*. Then,*

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) \leq \tau \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{27}$$

Proof Assume the claim of the proposition is incorrect, then

$$\max_{v \in [w_1, w_2]} \widehat{F}(v) > \tau \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\} > \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{28}$$

Define $w_\star := \arg\max_{v\in[w_1,w_2]} \widehat{F}(v)$. Note that $w_\star$ is an interior point. Thus by the optimality condition it holds

$$\left\langle \nabla \widehat{F}(w_\star), w_1 - w_2 \right\rangle = 0. \tag{29}$$

By Taylor's approximation theorem for two points $w_1, w \in \mathbb{R}^{d'}$, there exists a $w_\beta \in [w, w_1]$, such that

$$\widehat{F}(w_1) = \widehat{F}(w) + \left\langle \nabla \widehat{F}(w), w_1 - w \right\rangle + \frac{1}{2}\left\langle w - w_1, \nabla^2 \widehat{F}(w_\beta)\,(w - w_1) \right\rangle \tag{30}$$

Pick $w = w_\star = \alpha_\star w_1 + (1-\alpha_\star)w_2$ in Eq. (30), and note that

$$\left\langle \nabla \widehat{F}(w_\star), w_1 - w_\star \right\rangle = -(1-\alpha_\star)\left\langle \nabla \widehat{F}(w_\star), w_1 - w_2 \right\rangle = 0.$$

Therefore,

$$\begin{aligned}
\widehat{F}(w_1) &= \widehat{F}(w_\star) + \frac{1}{2}\left\langle w_\star - w_1, \nabla^2 \widehat{F}(w_\beta)\,(w_\star - w_1) \right\rangle \\
&\geq \widehat{F}(w_\star) + \frac{1}{2}\lambda_{\min}(\nabla^2 \widehat{F}(w_\beta)) \left\| w_\star - w_1 \right\|^2 \\
&\geq \widehat{F}(w_\star) - \frac{1}{2}\kappa\, \widehat{F}(w_\beta) \left\| w_\star - w_1 \right\|^2.
\end{aligned}$$

where in the last line we used the self-bounded weak convexity property i.e., $\lambda_{\min}\left(\nabla^2 \widehat{F}(w_\beta)\right) \geq -\kappa \widehat{F}(w_\beta)$.

This leads to

$$\begin{aligned}
\widehat{F}(w_1) &\geq \widehat{F}(w_\star) - \frac{(1-\alpha_\star)^2}{2}\kappa\, \widehat{F}(w_\beta) \left\| w_1 - w_2 \right\|^2 \\
&> \widehat{F}(w_\star) - \frac{1}{2}\kappa\, \widehat{F}(w_\beta) \left\| w_1 - w_2 \right\|^2.
\end{aligned}$$

Note that $w_\beta \in [w_\star, w_1] \subset [w_1, w_2]$, thus $\widehat{F}(w_\beta) \leq \widehat{F}(w_\star)$ by definition of $w_\star$. Therefore,

$$\begin{aligned}
\widehat{F}(w_\star) &< \frac{1}{1 - \frac{1}{2}\kappa \left\| w_1 - w_2 \right\|^2} \widehat{F}(w_1) \\
&\leq \frac{1}{1 - \frac{1}{2}\kappa D^2} \widehat{F}(w_1),
\end{aligned}$$

which is in contradiction with (28). This proves the statement of the proposition. ■

Specializing this property to two-layer neural networks yields the following.

Corollary A.7.1 (Restatement of Corollary 5.1.1)**.** *Let Assumptions 1,2, 4 hold. Fix arbitrary $w_1, w_2 \in \mathbb{R}^{d'}$, any constant $\lambda > 1$, and m large enough such that $\sqrt{m} \geq \lambda \frac{LR^2}{2}\|w_1 - w_2\|^2$. Then,*

$$\max_{v\in[w_1,w_2]} \widehat{F}(v) \leq (1 - 1/\lambda)^{-1} \cdot \max\{\widehat{F}(w_1), \widehat{F}(w_2)\}. \tag{31}$$

Proof By our assumptions and Corollary D.2.1 the objective's Hessian satisfies

$$\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \geq -\frac{LR^2}{\sqrt{m}}\widehat{F}(w).$$

Invoking Proposition A.7 with $\kappa := \frac{LR^2}{\sqrt{m}}$ concludes the claim. ■

Appendix B. Generalization Analysis

This section includes the proofs of the generalization results stated in Section 3.3.

B.1 Proof of Theorem 3.4

We prove the generalization gap of Theorem 3.4 for Lipshitz-smooth losses. The proof follows the steps of our proof sketch in Sec. 5.2.

First, the proofs of exansiveness of GD in NN (Lemma 5.2) and the corresponding model stability bound are given next.

Lemma B.1 (GD-Expansiveness)**.** *Let Assumptions 1-2 hold. For any w, w' and $w_\alpha = \alpha w + (1-\alpha)w'$ it holds that*

$$\left\|\left(w - \eta\nabla\widehat{F}(w)\right) - \left(w' - \eta\nabla\widehat{F}(w')\right)\right\| \leq \max_{\alpha\in[0,1]} H(w_\alpha)\,\left\|w - w'\right\|,$$
$$H(w) := \eta\frac{LR^2}{\sqrt{m}}\widehat{F}'(w) + \max\left\{1, \eta\ell^2R^2\widehat{F}''(w)\right\},$$

where we define $\widehat{F}'(w) := \frac{1}{n}\sum_{i=1}^n |f'(y_i\Phi(w,x_1))|$ and $\widehat{F}''(w) := \frac{1}{n}\sum_{i=1}^n f''(y_i\Phi(w,x_1))$.

Proof Fix $u : \|u\| = 1$ and define $g_u : \mathbb{R}^{d'} \to \mathbb{R}$:

$$g_u(w) := \langle u, w\rangle - \eta\langle u, \nabla\widehat{F}(w)\rangle.$$

Note

$$\left\|w - \nabla\widehat{F}(w) - (w' - \nabla\widehat{F}(w'))\right\| = \max_{\|u\|=1}\left|g_u(w) - g_u(w')\right|.$$

For any w, w', we have

$$\begin{aligned} g_u(w) - g_u(w') &= \int_0^1 u^\top\left(I - \eta\nabla^2\widehat{F}(w' + \alpha(w-w'))\right)(w-w')\mathrm{d}\alpha \\ &\leq \max_{\alpha\in[0,1]}\left\|\left(I - \eta\nabla^2\widehat{F}(w' + \alpha(w-w'))\right)\right\|\left\|w - w'\right\|. \end{aligned} \tag{32}$$

For convenience denote $w_\alpha := \alpha w + (1-\alpha)w'$ and $A_\alpha := \nabla^2\widehat{F}(w_\alpha)$. Then, for any $\alpha \in [0,1]$ we have that

$$\left\|I - \eta\nabla^2\widehat{F}(w_\alpha)\right\| = \max\left\{|1 - \eta\lambda_{\min}(A_\alpha)|, |1 - \eta\lambda_{\max}(A_\alpha)|\right\}. \tag{33}$$

For convenience, let $\beta := \frac{1}{\sqrt{m}} LR^2 \widehat{F}'(w_\alpha) \geq 0$ and note from Lemma D.2 that $\lambda_{\min}(A_\alpha) \geq -\beta$. Using this, we will show that

$$|1 - \eta\lambda_{\min}(A_\alpha)| \leq \max\{1 + \eta\beta, \eta\lambda_{\max}(A_\alpha)\}. \tag{34}$$

To show this consider two cases. First, if $\eta\lambda_{\min}(A_\alpha) \in [-\eta\beta, 1]$, then

$$|1 - \eta\lambda_{\min}(A_\alpha)| = 1 - \eta\lambda_{\min}(A_\alpha) \leq 1 + \eta\beta.$$

On the other hand, if $\eta\lambda_{\min}(A_\alpha) \geq 1$, then

$$|1 - \eta\lambda_{\min}(A_\alpha)| = \eta\lambda_{\min}(A_\alpha) - 1 \leq \eta\lambda_{\min}(A_\alpha) \leq \eta\lambda_{\max}(A_\alpha),$$

which shows (34).

Next, we will show that

$$|1 - \eta\lambda_{\max}(A_\alpha)| \leq \max\{1 + \eta\beta, \eta\lambda_{\max}(A_\alpha)\}. \tag{35}$$

We consider again three cases. First, if $\eta\lambda_{\max}(A_\alpha) \in [0, 1]$, then

$$|1 - \eta\lambda_{\max}(A_\alpha)| = 1 - \eta\lambda_{\max}(A_\alpha) \leq 1.$$

Second, if $\eta\lambda_{\max}(A_\alpha) \geq 1$

$$|1 - \eta\lambda_{\max}(A_\alpha)| = \eta\lambda_{\max}(A_\alpha) - 1 \leq \eta\lambda_{\max}(A_\alpha).$$

Otherwise, it must be that $-\beta \leq \lambda_{\min}(A_\alpha) \leq \lambda_{\max}(A_\alpha) \leq 0$. Thus,

$$|1 - \eta\lambda_{\max}(A_\alpha)| = 1 - \eta\lambda_{\max}(A_\alpha) \leq 1 - \eta\lambda_{\min}(A_\alpha) \leq 1 + \eta\beta.$$

To complete the proof of the lemma combine (33) with (34) and (35):

$$\|I - \eta\nabla^2 \widehat{F}(w_\alpha)\| \leq \max\{1 + \eta\beta, \eta\lambda_{\max}(A_\alpha)\},$$

and further use from Lemma D.2 that $\eta\lambda_{\max}(A_\alpha) \leq \eta\ell^2 R^2 \widehat{F}''(w) + \eta\beta$. ■

For the stability analysis below, recall the definition of the leave-one-out (loo) training loss for $i \in [n]$: $\widehat{F}^{\neg i}(w) := \frac{1}{n}\sum_{j \neq i} \widehat{F}_j(w)$. With these, define the loo model updates of GD on the loo loss:

$$w_{t+1}^{\neg i} := w_t^{\neg i} - \eta\nabla\widehat{F}^{\neg i}(w_t^{\neg i}),\ t \geq 0, \qquad w_0^{\neg i} = w_0.$$

Theorem B.2 (Model stability bound)**.** *Suppose Assumptions 1, 2, 3, 4 hold. Fix any time horizon $T \geq 1$ and any step size $\eta > 0$. Set the regret and the leave-one-out regrets of GD updates as follows:*

$$\texttt{Reg} := \frac{1}{T}\sum_{t=1}^{T} \widehat{F}(w_t) \qquad \textit{and} \qquad \texttt{Reg}_{\texttt{loo}} := \frac{1}{T}\max_{i \in [n]} \sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t^{\neg i}).$$

Suppose that the width m is large enough so that it satisfies the following two conditions:

$$\sqrt{m} \geq 4LR^2 \max\left\{\|w_t - w_0\|^2, \|w_t^{\neg i} - w_0\|^2\right\}, \quad \forall i \in [n], t \in [T], \tag{36}$$

and

$$\sqrt{m} \geq 6LR^2\eta T \max\left\{\texttt{Reg}, \texttt{Reg}_{loo}\right\}. \tag{37}$$

Then, the leave-one-out model stability is bounded as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\left\|w_T - w_T^{\neg i}\right\| \leq \frac{2\eta\ell R}{n}\left(\widehat{F}(w_0) + T\cdot\texttt{Reg}\right).$$

Proof Using self-boundedness Assumption 4 together with Corollary 5.2.1 it holds for all $i \in [n]$:

$$\begin{aligned}
\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| &\leq \left\|\left(w_t - \eta\nabla\widehat{F}^{\neg i}(w_t)\right) - \left(w_t^{\neg i} - \eta\nabla\widehat{F}^{\neg i}(w_t^{\neg i})\right)\right\| + \frac{\eta}{n}\left\|\nabla\widehat{F}_i(w_t)\right\| \\
&\leq \left\|\left(w_t - \eta\nabla\widehat{F}^{\neg i}(w_t)\right) - \left(w_t^{\neg i} - \eta\nabla\widehat{F}^{\neg i}(w_t^{\neg i})\right)\right\| + \frac{\eta\ell R}{n}\widehat{F}_i(w_t) \\
&\leq \left(1 + \eta\frac{LR^2}{\sqrt{m}}\max_{\alpha\in[0,1]}\widehat{F}^{\neg i}(w_{\alpha t}^{\neg i})\right)\left\|w_t - w_t^{\neg i}\right\| + \frac{\eta\ell R}{n}\widehat{F}_i(w_t),
\end{aligned} \tag{38}$$

where we denote for convenience $w_{\alpha t}^{\neg i} = \alpha w_t + (1-\alpha)w_t^{\neg i}$.

Moreover, by the theorem's condition in Eq. (36), it holds for all $t \in [T]$ and all $i \in [n]$ that

$$\sqrt{m} \geq 2LR^2(\|w_t - w_0\|^2 + \|w_t^{\neg i} - w_0\|^2) \geq LR^2\left\|w_t - w_t^{\neg i}\right\|^2.$$

Thus, we can apply Corollary 5.1.1 for $\lambda = 2$, which gives the following generalized-local quasi-convexity property for the loo objective:

$$\max_{\alpha\in[0,1]}\widehat{F}^{\neg i}(w_{\alpha t}^{\neg i}) \leq 2\max\left\{\widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i})\right\}.$$

In turn applying this back in (38) we have shown that

$$\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| \leq \left(1 + \eta\frac{2LR^2}{\sqrt{m}}\max\left\{\widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i})\right\}\right)\left\|w_t - w_t^{\neg i}\right\| + \frac{\eta\ell R}{n}\widehat{F}_i(w_t) \tag{39}$$

To continue, denote for convenience

$$\beta_t^i := \eta\frac{2LR^2}{\sqrt{m}}\max\left\{\widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{\neg i})\right\} \quad \text{and} \quad \rho := \eta\ell R,$$

so that:

$$\left\|w_{t+1} - w_{t+1}^{\neg i}\right\| \leq \left(1 + \beta_t^i\right)\left\|w_t - w_t^{\neg i}\right\| + \frac{\rho}{n}\widehat{F}_i(w_t), \qquad \forall i \in [n], t \in [T].$$

By unrolling the iterations over $t \in [T]$ and noting $w_0 = w_0^{\neg i}$, we obtain the following for the leave-one-out parameter distance at iteration T:

$$\begin{aligned}
\left\| w_T - w_T^{\neg i} \right\| &\leq \frac{\rho}{n} \sum_{t=0}^{T-1} \left(\prod_{\tau=t+1}^{T-1} (1 + \beta_\tau^i) \right) \widehat{F}_i(w_t) \\
&\leq \frac{\rho}{n} \sum_{t=0}^{T-1} \exp\left(\sum_{\tau=t+1}^{T-1} \beta_\tau^i \right) \widehat{F}_i(w_t) \\
&\leq \frac{\rho}{n} \sum_{t=0}^{T-1} \exp\left(\sum_{\tau=1}^{T-1} \beta_\tau^i \right) \widehat{F}_i(w_t) = \exp\left(\sum_{\tau=1}^{T-1} \beta_\tau^i \right) \frac{\rho}{n} \sum_{t=0}^{T-1} \widehat{F}_i(w_t) \\
&\leq \frac{\rho}{n} \exp\left(\max_{j \in [n]} \sum_{\tau=1}^{T-1} \beta_\tau^j \right) \sum_{t=0}^{T-1} \widehat{F}_i(w_t), \qquad \forall i \in [n] \,. \qquad (40)
\end{aligned}$$

It remains to bound $\beta := \max_{i \in [n]} \sum_{\tau=1}^{T-1} \beta_\tau^i$. We do this as follows:

$$\begin{aligned}
\beta &= \frac{2\eta L R^2}{\sqrt{m}} \max_{i \in [n]} \left\{ \max \left\{ \sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t) \,, \sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t^{\neg i}) \right\} \right\} \\
&\leq \frac{2\eta L R^2}{\sqrt{m}} \max_{i \in [n]} \left\{ \max \left\{ \sum_{t=1}^{T} \widehat{F}(w_t) \,, \sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t^{\neg i}) \right\} \right\} \\
&= \frac{2\eta L R^2}{\sqrt{m}} \max \left\{ \sum_{t=1}^{T} \widehat{F}(w_t) \,, \max_{i \in [n]} \sum_{t=1}^{T} \widehat{F}^{\neg i}(w_t^{\neg i}) \right\} \\
&= \frac{2\eta L R^2}{\sqrt{m}} T \, \max \left\{ \texttt{Reg} \,, \texttt{Reg}_{\texttt{1oo}} \right\} \leq 2/3 \,,
\end{aligned}$$

where: (i) in the first inequality we used nonnegativity of $f(\cdot)$ to conclude for any $i \in [n]$ and any w that $\widehat{F}^{\neg i}(w) \leq \widehat{F}(w)$; (ii) in the last line, we recalled the definition of the regret terms and we used the theorem's condition (43) on large enough m.

Using this in (40) and averaging over $i \in [n]$ yields

$$\begin{aligned}
\frac{1}{n} \sum_{i \in [n]} \left\| w_T - w_T^{\neg i} \right\| &\leq \frac{\rho e^\beta}{n} \sum_{t=0}^{T-1} \frac{1}{n} \sum_{i=1}^{n} \widehat{F}_i(w_t) \\
&\leq \frac{\eta \ell R e^{2/3}}{n} \sum_{t=0}^{T-1} \widehat{F}(w_t) \,.
\end{aligned}$$

The advertised bound follows by using $e^{2/3} \leq 2$ and writing

$$\frac{1}{T} \sum_{t=0}^{T-1} \widehat{F}(w_t) \leq \frac{1}{T} \sum_{t=0}^{T} \widehat{F}(w_t) = \frac{\widehat{F}(w_0)}{T} + \texttt{Reg}.$$

■

To bound the generalization gap in terms of model stability we rely on the following result.

Lemma B.3 (Lei and Ying (2020a)). *Suppose the sample loss $f(\cdot, z)$ is $G_{\widehat{F}}$-Lipschitz for almost surely all data points $z \sim \mathcal{D}$. Then, the following relation holds between expected generalization loss and model stability at any iterate T,*

$$\mathbb{E}\Big[F(w_T)\Big] - \mathbb{E}\Big[\widehat{F}(w_T)\Big] \leq 2G_{\widehat{F}}\, \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\|w_T - w_T^{\neg i}\|\Big]. \tag{41}$$

With the two results above, we are ready to prove Theorem 3.4.

Theorem B.4 (Restatement of Theorem 3.4). *Suppose Assumptions 1- 4 hold. Fix any time horizon $T \geq 1$ and any step size $\eta \leq 1/L_{\widehat{F}}$ where $L_{\widehat{F}}$ is the objective's smoothness parameter. Let any $w \in \mathbb{R}^{d'}$ such that $\|w - w_0\|^2 \geq \max\{\eta T\, \widehat{F}(w), \eta \widehat{F}(w_0)\}$. Suppose hidden-layer width m satisfies $m \geq 64^2 L^2 R^4 \|w - w_0\|^4$. Then, the generalization gap of GD at iteration T is bounded as*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{8\ell^2 R^2}{n}\, \mathbb{E}\left[\eta T\, \widehat{F}(w) + 2\|w - w_0\|^2\right],$$

where all expectations are over the training set.

Proof The proof essentially follows by combining Theorem B.2 with Theorem 3.2. Note that the assumptions of Theorem 3.2 are met. Thus, the regret and parameter-norm are bounded as follows:

$$\texttt{Reg} \leq 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T} \qquad \text{and} \qquad \max_{t\in[T]} \|w_t - w_0\| \leq 4\|w - w_0\|\,. \tag{42}$$

We can also use Theorem 3.2 to the leave-one-out objective $\widehat{F}^{\neg i}$ and the corresponding loo GD updates $w_t^{\neg i}$. This bounds the loo regret and the norm of the loo parameter, as follows:

$$\texttt{Reg}_{\texttt{loo}} \leq 2\widehat{F}(w) + \frac{5\|w - w_0\|^2}{2\eta T} \qquad \text{and} \qquad \max_{i\in[n]} \max_{t\in[T]} \|w_t^{\neg i} - w_0\| \leq 4\|w - w_0\|\,.$$

We use these two displays to show that m is by assumption large enough so that Eqs. (36) and (43) hold. Indeed, we have

$$\sqrt{m} \geq 64LR^2\|w - w_0\|^2 = 4LR^2\,(4\|w - w_0\|)^2 \geq 4LR^2 \max\left\{\|w_t - w_0\|^2, \|w_t^{\neg i} - w_0\|^2\right\}$$

and

$$\begin{aligned}\sqrt{m} \geq 64LR^2\|w - w_0\|^2 &> 6LR^2 \cdot 5\|w - w_0\|^2 \\ &> 6LR^2 \cdot (2\eta T \widehat{F}(w) + 5\|w - w_0\|^2/2) \\ &\geq 6LR^2\eta T \max\left\{\texttt{Reg}, \texttt{Reg}_{\texttt{loo}}\right\}\,.\end{aligned}$$

In the second display we also used the theorem's assumption that $\|w - w_0\|^2 \geq \eta T \widehat{F}(w)$.

Thus, we can apply Theorem B.2 to find that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left\|w_T - w_T^{\neg i}\right\| &\leq \frac{2\ell R}{n}\left(\eta\widehat{F}(w_0) + \eta T\cdot \texttt{Reg}\right)\\ &\leq \frac{2\ell R}{n}\left(\eta\widehat{F}(w_0) + 2\eta T\widehat{F}(w) + 5\|w-w_0\|^2/2\right)\\ &\leq \frac{2\ell R}{n}\left(2\eta T\widehat{F}(w) + 7\|w-w_0\|^2/2\right)\end{aligned}$$

where in the penultimate line we used (42) and in the last line we used the theorem's assumption that $\|w-w_0\|^2 \geq \eta\widehat{F}(w_0)$.

To conclude the proof, simply take expectations over the train set on the above display and apply Lemma B.3 recalling $G_{\widehat{F}} = \ell R$. ■

B.2 Proof of Theorem 3.5

Here we prove the generalization gap for interpolating neural networks as per Theorem 3.5.

Theorem B.5 (Restatement of Theorem 3.5)**.** *Let Assumptions 1-5 hold. Fix $T \geq 1$ and let $m \geq 64^2 L^2 R^4\, g(\frac{1}{T})^4$. Then, for any $\eta \leq \min\{\frac{1}{L_{\widehat{F}}}, g(1)^2, \frac{g(1)^2}{\widehat{F}(w_0)}\}$ the expected generalization gap at iteration T satisfies*

$$\mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] \leq \frac{24\ell^2 R^2\, g(\frac{1}{T})^2}{n}\,. \tag{43}$$

Proof According to Assumption 5, for any sufficiently small $\varepsilon > 0$, there exists $w^{(\varepsilon)}$ such that $\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon$ and $\|w^{(\varepsilon)} - w_0\| = g(\varepsilon)$. Recall from Theorem 3.4 that,

$$\mathbb{E}\left[F(w_T) - \widehat{F}(w_T)\right] \leq \frac{8\ell^2 R^2}{n}\,\left(\eta T\widehat{F}(w) + 2\|w-w_0\|^2\right)\,. \tag{44}$$

In particular let $\varepsilon = 1/T$ and replace w with $w^{(\varepsilon)}$. This is possible since after $T \geq 1$ steps and with the decreasing nature of g and the condition on step-size it holds that $\|w^{(1/T)} - w_0\|^2 = g(1/T)^2 \geq g(1)^2 \geq \max\{\eta T\widehat{F}(w^{(1/T)}), \eta\widehat{F}(w_0)\}$. Thus continuing from (44) we have,

$$\mathbb{E}\left[F(w_T) - \widehat{F}(w_T)\right] \leq \frac{8\ell^2 R^2}{n}\,\left(\eta + 2g(\frac{1}{T})^2\right)\,.$$

Recalling $\eta \leq g(1)^2 \leq g(\frac{1}{T})^2$ leads to the claim of the theorem. ■

Appendix C. Proofs for Section 4

We first prove proposition 4.1, which we repeat here for convenience.

Proposition C.1 (Restatement of Proposition 4.1)**.** *Let Assumptions 1-2,6-7 hold. Assume $f(\cdot)$ to be the logistic loss. Fix $\varepsilon > 0$ and let $m \geq \frac{L^2R^4}{4\gamma^4C^2}(2C + \log(1/\varepsilon))^4$. Then the realizability Assumption 5 holds with $g(\varepsilon) = \frac{1}{\gamma}(2C + \log(1/\varepsilon))$. In other words, there exists $w^{(\varepsilon)}$ such that*

$$\widehat{F}(w^{(\varepsilon)}) \leq \varepsilon, \quad \textit{and} \quad \left\|w^{(\varepsilon)} - w_0\right\| = \frac{1}{\gamma}\left(2C + \log(1/\varepsilon)\right). \tag{45}$$

Proof By Taylor there exists $w' \in [w, w_0]$ such that,

$$y_i\Phi(w, x_i) = y_i\Phi(w_0, x_i) + y_i\Big\langle \nabla_1\Phi(w_0, x_i), w - w_0\Big\rangle + \frac{1}{2}y_i\Big\langle w - w_0, \nabla_1^2\Phi(w', x_i)(w - w_0)\Big\rangle \tag{46}$$

Pick $w = w^{(\varepsilon)} := w_0 + \frac{w^\star}{\gamma}(2C + \log(1/\varepsilon))$ for $w^\star$ defined in Assumption 6. Since $\|w^\star\| = 1$, we automatically derive the desired for $\|w^{(\varepsilon)} - w_0\|$. Next, we show that $\widehat{F}_i(w^{(\varepsilon)}) \leq \varepsilon$. Based on Lemma D.1, $\|\nabla_1^2\Phi(w', x_i)\| \leq \frac{LR^2}{\sqrt{m}}$. Continuing from Eq. (46), we deduce the following,

$$\begin{aligned} y_i\Phi(w, x_i) &\geq -\left|y_i\Phi(w_0, x_i)\right| + y_i\left\langle \nabla_1\Phi(w_0, x_i), w^{(\varepsilon)} - w_0\right\rangle - \frac{1}{2}\left\|\nabla_1^2\Phi(w', x_i)\right\|\left\|w^{(\varepsilon)} - w_0\right\|^2 \\ &\geq -C + 2C + \log(1/\varepsilon) - \frac{LR^2}{2\gamma^2\sqrt{m}}(2C + \log(1/\varepsilon))^2 \\ &\geq \log(1/\varepsilon). \end{aligned}$$

The last step is due to the condition on m. The inequality above implies that $\widehat{F}_i(w) := f(y_i\Phi(w, x_i)) \leq \log(1 + \varepsilon) \leq \varepsilon$, and thus $\widehat{F}(w) \leq \varepsilon$ as desired. This completes the proof. ■

With this, we many now prove Corollary 4.1.1.

Corollary C.1.1 (Restatement of Corollary 4.1.1)**.** *Let Assumptions 1-2,6-7 hold and assume logistic loss. Suppose $m \geq \frac{64^2L^2R^4}{\gamma^4}(2C + \log(T))^4$ for a fixed training horizon T. Then, for any $\eta \leq \min\{3, \frac{1}{L_{\widehat{F}}}\}$ the training loss and generalization gap are bounded as follows:*

$$\begin{aligned} \widehat{F}(w_T) &\leq \frac{5(2C + \log(T))^2}{\gamma^2\eta T}, \\ \mathbb{E}\Big[F(w_T) - \widehat{F}(w_T)\Big] &\leq \frac{24\ell^2R^2}{\gamma^2 n}(2C + \log(T))^2. \end{aligned}$$

Proof The given assumption on m satisfies the conditions of Proposition 4.1 for $\varepsilon = \frac{1}{T}$, $g(1/T) = \frac{1}{\gamma}(2C + \log(T))$. We can apply the results of our optimization and generalization results from Theorems 3.3 and 3.5 for a fixed T which satisfies $T \geq 1$. Note that we can assume without loss of generality that $\gamma \leq 1$ which implies that $g(1)^2 = 4C^2/\gamma^2 \geq 4$. Moreover, for logistic loss it holds $g(1)^2/\widehat{F}(w_0) \geq \frac{4C^2}{\gamma^2\log(1+e^C)} \geq 3$ for all $C \geq 1$. Therefore the condition on step-size simplifies to $\eta \leq \min\{3, 1/L_{\widehat{F}}\}$. This completes the proof. ■

C.1 Proof of Proposition 4.2

The proof of Proposition 4.2 has the following steps: First, we consider an infinite-width NTK separability assumption (Assumption 9) and show in Lemma C.2 that it is equivalent with high-probability to the NTK-separability in Assumption 6 given logarithmic number of neurons. We then prove that the noisy-XOR dataset satisfies Assumption 9 for convex and locally strongly-convex activations. The result of Proposition 4.2 then follows by combining the two lemmas.

Assumption 9 (Infinite-width NTK-separability)**.** *There exists $\overline{w}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ and $\gamma > 0$ such that $\|\overline{w}(z)\|_2 \leq 1$ for all $z \in \mathbb{R}^d$, and for all $(x, y) \sim \mathcal{D}$,*

$$y \int_{\mathbb{R}^d} \sigma'\left(\langle z, x\rangle\right) \cdot \langle \overline{w}(z), x\rangle \, \mathrm{d}\mu_{\mathrm{N}}(z) \geq \gamma,$$

where $\mu_N(\cdot)$ denotes the standard Gaussian measure.

Lemma C.2. *Let $\{(x_i, y_i)\}$ be any dataset of size $\tilde{n}$ under Assumption 1, satisfying the separability condition of Assumption 9 with some margin $\tilde{\gamma} > 0$. Consider initialization $w_0 \in \mathbb{R}^{d'}$ where $w_0 \sim N(0, I_{d'})$. Then, with probability at least $1 - \delta$ the dataset is separable under Assumption 6 with margin at least $\gamma = \tilde{\gamma} - \frac{\ell R}{\sqrt{2m}} \log^{1/2}(\tilde{n}/\delta)$, i.e., there exists unit norm $w^\star$ such that for all $i \in [\tilde{n}] : y_i \langle \nabla_1 \Phi(w_0, x_i), w^\star \rangle \geq \gamma$.*

Proof By the model's gradient we have for any $w^\star \in \mathbb{R}^{d'}$,

$$\phi_i := y_i \Big\langle \nabla_1 \Phi(w_0, x_i), w^\star \Big\rangle = y_i \sum_{j=1}^{m} \frac{a_j}{\sqrt{m}} \sigma'(\langle w_{0,j}, x_i\rangle) \langle x_i, w_j^\star \rangle. \tag{47}$$

Let $w_j^\star = \frac{a_j}{\sqrt{m}} \overline{w}(w_{0,j})$. Then $\|w^\star\| \leq 1$ and by Hoeffding's inequality it holds for all $t \geq 0$,

$$\Pr\left(\phi_i \geq \tilde{\gamma} - t\right) \geq 1 - \exp\left(\frac{-2t^2 m}{\ell^2 R^2}\right). \tag{48}$$

This leads to the desired result with an extra union bound over $i \in [\tilde{n}]$. ■

Lemma C.3. *Consider the noisy XOR data distribution $\{(\bar{x}_i, y_i)\}$ and two-layer neural network with a convex activation which is μ-strongly convex in $[-2, 2]$ i.e., $\min_{t \in [-2,2]} \sigma''(t) \geq \mu$ for some $\mu > 0$. Then the separability assumption 9 is satisfied with margin $\gamma = \frac{\mu}{40d}$.*

Proof The proof is essentially similar to (Ji and Telgarsky, 2020a, Prop. 5.3) and thus we follow their notation and omit the details for brevity. While their proof relies rather crucially on the ReLU activation, it can be appropriately modified to obtain a similar margin bound under our different assumptions on the activation function. To see this, note that due to convexity of activation function, the integrand in the line above Eq. (D.4) is non-negative. Therefore, we can lower-bound the integral (which evaluates the margin) by restricting A_1 to $|p_1| < 1$. With this restriction we can use the local strong convexity of activation function to lower-bound the margin, i.e., to uniformly lower-bound $y_i \int_{\mathbb{R}^d} \sigma'\left(\langle z, x_i\rangle\right) \cdot \langle \bar{w}(z), x_i\rangle \, \mathrm{d}\mu_{\mathrm{N}}(z)$ for

all $i \in [n]$. Specifically, note that with strong convexity in $[-2,2]$, Eq. (D.4) in Ji and Telgarsky (2020a) changes to $\geq \frac{2p_1}{d-1} U(p_1) \min_{t\in[-2,2]} \sigma''(t) \geq \frac{2p_1\mu}{d-1} U(p_1)$ where $U(t) := \int_{-t}^{t} \varphi(\tau) \mathrm{d}\tau$ is the probability that a standard Gaussian random variable falls in $[-t,t]$. This leads to the final value for margin being $\frac{2\mu}{d-1} \int p_1 \, U(p_1) \, \mathbf{1}\left[p \in A_1\right] \mathrm{d}\mu_N(p) \geq \frac{8\mu}{(2\pi e)^{3/2}(d-1)} \int_0^1 p_1^3 \mathrm{d}p_1 \geq \frac{\mu}{40d}$, as desired. ■

Proposition C.4 (Restatement of Proposition 4.2)**.** *Consider the noisy XOR data distribution $\{(\bar{x}_i, y_i)\}$. Assume the activation function is convex, ℓ-Lipschitz and μ-strongly convex in the interval $[-2,2]$ for some $\mu > 0$, i.e., $\min_{t\in[-2,2]} \sigma''(t) \geq \mu$. Moreover, assume Gaussian initialization $w_0 \in \mathbb{R}^{d'}$ with entries iid $N(0,1)$. If $m \geq \frac{80^2 d^3 \ell^2}{2\mu^2} \log(2/\delta)$, then with probability at least $1-\delta$ over the initialization, the NTK-separability Assumption 6 is satisfied with margin $\gamma = \frac{\mu}{80d}$.*

Proof The claim follows by combining the last two lemmas. In particular, we derive the infinite width NTK-separability for the entire data distribution (of size 2^d) with margin $\tilde{\gamma} = \frac{\mu}{40d}$ and by the assumption on width and noting $\tilde{n} = 2^d$, we have γ-separability by NTK for the entire distribution with probability $1-\delta$ where $\gamma = \tilde{\gamma} - \frac{\ell R}{\sqrt{2m}} \log^{1/2}(\tilde{n}/\delta) = \frac{\mu}{40d} - \frac{\ell R\sqrt{d}}{\sqrt{2m}} \log^{1/2}(1/\delta) \geq \frac{\mu}{80d}$. This completes the proof. ■

Finally, we show how to control the parameter C that bounds the model output at Gaussian initialization.

Lemma C.5 (Initialization bound)**.** *Let Assumption 1 hold and assume the activation function to be ℓ-Lipschitz. Consider initialization $w_0 \in \mathbb{R}^{d'}$ where $w_0 \sim N(0, I_{d'})$. Given any $\delta \in (0,1)$, then with probability at least $1-\delta$, it holds for all $i \in [\tilde{n}]$ that*

$$\left|\Phi\left(w_0, x_i\right)\right| \leq \ell R \sqrt{2\log(2\tilde{n}/\delta)}. \tag{49}$$

Proof Recall that if a function $\phi : \mathbb{R}^{d'} \to \mathbb{R}$ is G-Lipschitz then for Gaussian vector $Z = (Z_1, Z_2, \cdots, Z_{d'})$ where each component is i.i.d. standard Gaussian $Z_i \sim N(0,1)$, it holds for all $t \geq 0$ that $\Pr[|\phi(Z) - \mathbb{E}[\phi(Z)]| \geq t] \leq 2\exp(-\frac{t^2}{2G^2})$. Note that according to Lemma D.1, $\Phi(\cdot, x_i)$ is (ℓR)-Lipschitz for any data point x_i. Therefore, with the given initialization for w_0, we have

$$\Pr\left[\left|\Phi(w_0, x_i) - \mathbb{E}[\Phi(w_0, x_i)]\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\ell^2 R^2}\right).$$

It also holds that $\mathbb{E}[\Phi(w_0, x_i)] = 0$. This is true since for half of second layer weights $a_j = 1$ and for the rest $a_j = -1$. Thus, we have $\Pr\left[|\Phi(w_0, x_i)| \geq t\right] \leq 2\exp(-\frac{t^2}{2\ell^2 R^2})$. A union bound yields that uniformly over $i \in [\tilde{n}]$, we have $\Pr\left[|\Phi(w_0, x_i)| \geq t\right] \leq 2\tilde{n} \cdot \exp(-\frac{t^2}{2\ell^2 R^2})$ which concludes the claim of lemma. ■

Appendix D. Gradients and Hessian calculations

D.1 Definitions

Assume IID data $(x, y) \sim \mathcal{D}$, $x \in \mathbb{R}^d, y \in \{\pm 1\}$. Denote for convenience $z := yx$. Suppose two-layer neural network model

$$\Phi(w, x_i) = \frac{1}{\sqrt{m}} \sum_{j \in [m]} a_j \sigma(\langle w_j, x \rangle) \tag{50}$$

$a_j \in \{\pm 1\}, j \in [m]$ and first-layer weights trained by GD on

$$\widehat{F}(w) = \frac{1}{n} \sum_{i \in [n]} f(y_i \Phi(w, x_i)) =: \frac{1}{n} \sum_{i \in [n]} f(w, z_i)\,. \tag{51}$$

for loss function $f : \mathbb{R} \to \mathbb{R}$.

For convenience define

$$\widehat{F}'(w) = \frac{1}{n} \sum_{i \in [n]} |f'(y_i \Phi(w, x_i))| \tag{52a}$$

$$\widehat{F}''(w) = \frac{1}{n} \sum_{i \in [n]} |f''(y_i \Phi(w, x_i))| \tag{52b}$$

D.2 Model's Gradient/Hessian

Lemma D.1. *The following are true for the model* (50) *under Assumption 2.*

1. $\|\nabla_1 \Phi(w, x)\| \leq \ell R$.
2. $\|\nabla_1^2 \Phi(w, x)\| \leq \frac{LR^2}{\sqrt{m}}$.

Proof Direct calculation yields that,

$$\nabla_1 \Phi(w, x) = \frac{1}{\sqrt{m}} \begin{bmatrix} a_1 \sigma'(\langle w_1, x \rangle) x \\ \cdot \\ \cdot \\ a_m \sigma'(\langle w_m, x \rangle) x \end{bmatrix}$$

Noting that $\sigma'(\cdot) \leq \ell$,

$$\begin{aligned} \|\nabla_1 \Phi(w, x)\|^2 &= \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{d} (x(i) \sigma'(\langle w_j, x \rangle))^2 \\ &\leq \ell^2 \|x\|^2 \\ &\leq \ell^2 R^2. \end{aligned} \tag{53}$$

For the Hessian,

$$\frac{\partial^2 \Phi(w, x)}{\partial w_{ij} \partial w_{k\ell}} = \frac{1}{\sqrt{m}} x(j) x(\ell) a_i \sigma''(\langle w_i, x \rangle) \mathbf{1}_{\{i=k\}}. \tag{54}$$

Thus,

$$\nabla_1^2\Phi(w,x)=\frac{1}{\sqrt{m}}\operatorname{diag}\left(a_1\sigma''(\langle w_1,x\rangle)xx^T,\ldots,a_m\sigma''(\langle w_m,x\rangle)xx^T\right)$$

for any unit norm vector $u\in\mathbb{R}^{md}$, define $\bar{u}_i:=[u_{(i-1)m+1}:u_{im}]\in\mathbb{R}^d$. Moreover, define the matrix $\nabla^2_{w_i}\Phi(w,x)\in\mathbb{R}^{d\times d}$ such that $[\nabla^2_{w_i}\Phi(w,x)]_{j\ell}=\frac{\partial^2\Phi(w,x)}{\partial w_{ij}\partial w_{i\ell}}$

$$\begin{aligned}\left\|u^\top\nabla_1^2\Phi(w,x)\right\|^2&=\sum_{i=1}^m\left\|u_i^\top\nabla^2_{w_i}\Phi(w,x)\right\|^2\\&\le\sum_{i=1}^m\left\|\nabla^2_{w_i}\Phi(w,x)\right\|^2\|\bar{u}_i\|^2\\&\le\sum_{i=1}^m\frac{L^2}{m}\|x\|^4\|\bar{u}_i\|^2\\&\le\frac{L^2R^4}{m}.\end{aligned}$$

This completes the proof. ■

D.3 Objective's Gradient/Hessian

Lemma D.2. *Let Assumption 2 hold. Then, the following are true for the loss gradient and Hessian:*

1. $\|\nabla\widehat{F}(w)\|\le\ell R\,\widehat{F}'(w)$.
2. $\|\nabla^2\widehat{F}(w)\|\le\ell^2R^2\widehat{F}''(w)+\frac{LR^2}{\sqrt{m}}\widehat{F}'(w)$.
3. $\lambda_{\min}\left(\nabla^2\widehat{F}(w)\right)\ge-\frac{LR^2}{\sqrt{m}}\widehat{F}'(w)$.

Proof The loss gradient is derived as follows,

$$\nabla\widehat{F}(w)=\frac{1}{n}\sum_{i=1}^n f'(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i)$$

Recalling that $y_i\in\{\pm1\}$, we can write

$$\begin{aligned}\left\|\nabla\widehat{F}(w)\right\|&=\frac{1}{n}\left\|\sum_{i=1}^n f'(y_i\Phi(w,x_i))y_i\nabla_1\Phi(w,x_i)\right\|\\&\le\frac{1}{n}\sum_{i=1}^n|f'(y_i\Phi(w,x_i))|\left\|\nabla_1\Phi(w,x_i)\right\|.\\&\le\ell R\,F'(w).\end{aligned}\tag{55}$$

For the Hessian of loss, note that

$$\nabla^2 \widehat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} f''(y_i\Phi(w,x_i))\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top + f'(y_i\Phi(w,x_i))y_i\nabla_1^2\Phi(w,x_i). \tag{56}$$

It follows that

$$\begin{aligned}
\left\|\nabla^2 \widehat{F}(w)\right\| &= \left\|\frac{1}{n}\sum_{i=1}^{n} f'(y_i\Phi(w,x_i))y_i\nabla_1^2\Phi(w,x_i) + f''(y_i\Phi(w,x_i))\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top\right\| \\
&\le \frac{1}{n}\sum_{i=1}^{n}|f'(y_i\Phi(w,x_i))|\left\|\nabla_1^2\Phi(w,x_i)\right\| + |f''(y_i\Phi(w,x_i))|\left\|\nabla_1\Phi(w,x_i)\nabla_1\Phi(w,x_i)^\top\right\| \\
&\le \frac{1}{n}\sum_{i=1}^{n}|f'(y_i\Phi(w,x_i))|\left\|\nabla_1^2\Phi(w,x_i)\right\| + |f''(y_i\Phi(w,x_i))|\left\|\nabla_1\Phi(w,x_i)\right\|^2 \\
&\le \frac{LR^2}{\sqrt{m}}F'(w) + \ell^2R^2F''(w).
\end{aligned} \tag{57}$$

To lower-bound the minimum eigenvalue of Hessian, note that f is convex and thus $f''(\cdot) \ge 0$. Therefore the first term in (56) is positive semi-definite and the second term can be lower-bounded as follows,

$$\begin{aligned}
\lambda_{\min}(\nabla^2 \widehat{F}(w)) &\ge -\left\|\frac{1}{n}\sum_{i=1}^{n} y_i f'(y_i\Phi(w,x_i))\nabla_1^2\Phi(w,x_i)\right\| \\
&\ge -\frac{1}{n}\sum_{i=1}^{n}|y_i f'(y_i\Phi(w,x_i))|\left\|\nabla_1^2\Phi(w,x_i)\right\| \\
&\ge -\frac{LR^2}{\sqrt{m}}F'(w).
\end{aligned}$$

■

Corollary D.2.1 (Self-boundedness of Objective)**.** *Let Assumption 2 hold. If the loss satisfies Assumptions 4 (with $\beta_f = 1$) and 8, then*

1. $\|\nabla \widehat{F}(w)\| \le \ell R\, \widehat{F}(w)$.
2. $\|\nabla^2 \widehat{F}(w)\| \le \left(\ell^2R^2 + \frac{LR^2}{\sqrt{m}}\right)\widehat{F}(w)$.
3. $\lambda_{\min}\left(\nabla^2 \widehat{F}(w)\right) \ge -\frac{LR^2}{\sqrt{m}}\widehat{F}(w)$.

If in addition the loss satisfies Assumptions 3.A and 3.B with $L_f = G_f = 1$, then

6. $\|\nabla \widehat{F}(w)\| \le \ell R$.

7. $\|\nabla^2 \widehat{F}(w)\| \leq \ell^2 R^2 + \frac{LR^2}{\sqrt{m}}$.

Proof For self-bounded losses we have $\widehat{F}'(w) \leq \widehat{F}(w)$ and $\widehat{F}''(w) \leq \widehat{F}(w)$. If the loss is 1-Lipschitz and 1-smooth we have $\widehat{F}'(w) \leq 1$ and $\widehat{F}''(w) \leq 1$. Thus, the claims immediately follow from Lemma D.2. ■

Appendix E. Detailed technical comparison to most-closely related works

In terms of techniques, the most closely related works to our paper are the recent works Richards and Rabbat (2021); Richards and Kuzborskij (2021); Lei et al. (2022), which also utilize the stability-analysis framework to derive test-loss bounds of GD for shallow neural networks.

Richards and Rabbat (2021) investigates the generalization gap of weakly-convex losses for which $\lambda_{\min}(\nabla^2 \widehat{F}(w)) \geq -\epsilon$ for a constant $\epsilon > 0$. Note by Lemma D.2 that our empirical loss is weakly convex with $\epsilon = LR^2/\sqrt{m}$ since the logistic loss is 1-Lipschitz. Within the stability analysis framework, Richards and Rabbat (2021) leverage the weak-convexity property to establish an approximate expansiveness property of GD iterates that in our setting translates to

$$\left\| \left(w - \eta \nabla \widehat{F}(w) \right) - \left(w' - \eta \nabla \widehat{F}(w') \right) \right\| \lesssim \left(1 + \frac{\eta L R^2}{\sqrt{m}} \right) \|w - w'\| . \tag{58}$$

When using this inequality to bound the model stability term at iteration t, and in order to obtain non-vacuous bounds, the extra term in (58) must be chosen such that $\eta LR^2/\sqrt{m} \lesssim 1/t$. This leads to polynomial-width parameterization requirement $m \gtrsim t^2$. In this work, we reduce the requirement to logarithmic $m \gtrsim \log(t)$, by significantly tightening (58). This is achieved by introducing two crucial ideas. The first is to exploit the self-boundedness property of loss function, which yields a stronger *self-bounded* weak convexity $\lambda_{\min}(\nabla^2 \widehat{F}(w)) \geq -LR^2 \widehat{F}(w)/\sqrt{m}$. With this, we show in Corollary 5.2.1 that

$$\left\| \left(w - \eta \nabla \widehat{F}(w) \right) - \left(w' - \eta \nabla \widehat{F}(w') \right) \right\| \leq \left(1 + \frac{\eta L R^2}{\sqrt{m}} \max_{\alpha \in [0,1]} \widehat{F}(w_\alpha) \right) \|w - w'\| \tag{59}$$

for some $w_\alpha = \alpha w + (1-\alpha) w'$. Our second idea comes into bounding the term $\max_{\alpha \in [0,1]} F(w_\alpha)$ which in our bound replaces the Lipschitz constant G_f of (58). To control $\max_{\alpha \in [0,1]} F(w_\alpha)$, we identify and use the Generalized Local Quasi-convexity of Proposition 5.1. This replaces $\max_{\alpha \in [0,1]} F(w_\alpha)$ in (59) with $\tau \cdot \max\{\widehat{F}(w), \widehat{F}(w')\}$ for $\tau \approx 1 + LR^2 \|w - w'\|^2/\sqrt{m}$ and note that we can guarantee $\tau = O(1)$ provided $\sqrt{m} \gtrsim \max\{\|w - w_0\|^2, \|w' - w_0\|^2\}$. Now, in order to bound the model stability term, we apply non-expansiveness for GD iterate $w = w_t$ and its leave-one-out counterpart $w = w_t^{\neg i}$: Provided $m \gtrsim \max\{\|w_t - w_0\|^4, \|w_t^{\neg i} - w_0\|^4\} \approx \log^4(t)$,

$$\begin{aligned} \left\| \left(w_t - \eta \nabla \widehat{F}(w_t) \right) - \left(w_t^{\neg i} - \eta \nabla \widehat{F}(w_t^{\neg i}) \right) \right\| &\lesssim \left(1 + \frac{\eta L R^2}{\sqrt{m}} \max\{\widehat{F}(w), \widehat{F}(w')\} \right) \|w - w'\| \\ &\lesssim \left(1 + \frac{\eta L R^2}{t\sqrt{m}} \right) \|w - w'\| \end{aligned} \tag{60}$$

Compared to (58) note in (60) that the extra term is already of order $1/t$. Hence, the only parameterization requirement is $m \gtrsim \max\{\|w_t - w_0\|^4, \|w_t^{\neg i} - w_0\|^4\} \approx \log^4(t)$. While the above describes our main technical novelty compared to Richards and Rabbat (2021), our results surpass theirs in other aspects. Specifically, we also obtain tighter bounds on the optimization error, again thanks to leveraging self-bounded properties of the logistic loss. Overall, for the separable setting, we show a $\tilde{O}(1/n)$ test-loss bound compared to $O(T/n)$ in their paper.

In closing, we remark that our logarithmic width requirements and expansiveness bounds are also significantly tighter than those that appear in Richards and Kuzborskij (2021); Lei et al. (2022). While their results are not directly comparable to ours as they only apply to square-loss functions, we reference them here for completeness: Richards and Kuzborskij (2021) upper-bounds the expansiveness term on the left-hand side of (60) by $\lesssim (1 + \eta\sqrt{\eta t}/\sqrt{m})$ which requires $m \gtrsim t^3$ so that is of order $1 + 1/t$. More recently, Lei et al. (2022) slightly modifies their bound to $\lesssim \left(1 + \eta(\eta t)^{3/2}/(n\sqrt{m})\right)$ which requires $m \gtrsim (\eta t)^5/n^2$.