

SHARPER GUARANTEES FOR LEARNING NEURAL NETWORK CLASSIFIERS WITH GRADIENT METHODS

Hossein Taheri

Department of Computer Science and Engineering,
University of California, San Diego.
htaheri@ucsd.edu

Christos Thrampoulidis,

Department of Electrical and Computer Engineering,
University of British Columbia.
cthrampo@ece.ubc.ca

Arya Mazumdar

Department of Computer Science and Engineering,
University of California, San Diego.
arya@ucsd.edu

ABSTRACT

In this paper, we study the data-dependent convergence and generalization behavior of gradient methods for neural networks with smooth activation. Our first result is a novel bound on the excess risk of deep networks trained by the logistic loss, via an algorithmic stability analysis. Compared to previous works, our results improve upon the shortcomings of the well-established Rademacher complexity-based bounds. Importantly, the bounds we derive in this paper are tighter, hold even for neural networks of small width, do not scale unfavorably with width, are algorithm-dependent, and consequently capture the role of initialization on the sample complexity of gradient descent for deep nets. Specialized to noiseless data separable with margin γ by neural tangent kernel (NTK) features of a network of width $\Omega(\text{poly}(\log(n)))$, we show the test-error rate to be $e^{O(L)}/\gamma^2 n$, where n is the training set size and L denotes the number of hidden layers. This is an improvement in the test loss bound compared to previous works while maintaining the poly-logarithmic width conditions. We further investigate excess risk bounds for deep nets trained with noisy data, establishing that under a polynomial condition on the network width, gradient descent can achieve the optimal excess risk. Finally, we show that a large step-size significantly improves upon the NTK regime's results in classifying the XOR distribution. In particular, we show for a one-hidden layer neural network of constant width m with quadratic activation and standard Gaussian initialization that SGD with linear sample complexity and with a large step-size $\eta = m$ reaches the perfect test accuracy after only $\lceil \log(d) \rceil$ iterations, where d is the data dimension.

1 INTRODUCTION

1.1 OVERVIEW

Neural networks, with their vast capacity for capturing intricate patterns in data, have triggered a paradigm shift in machine learning. Despite the power of these networks in modeling complex relationships, the interplay between their optimization and generalization behaviors (that is the gap between training and test errors) continues to be a compelling area of research. In practice, training neural networks using gradient-based optimization methods often leads to interpolation. That is, deep networks can meticulously fit the training data, driving empirical loss to near-zero and training error to perfect classification. However, these networks also demonstrate the capability to generalize well to unseen data. Various recent research endeavors have explored the training and test error guarantees of deep networks, with a focus on the Neural Tangent Kernel (NTK) regime Jacot et al. (2018); Du et al. (2019). One prominent feature of such works is that during gradient descent iterates the network's weights are constrained to move at most a constant distance with respect to overparameterization i.e.,

	Activation	Width	Train Loss	Test Loss
Chen et al. (2021); Bartlett et al. (2017)	ReLU	$\Omega(\text{poly}(\frac{\log(n)}{\gamma}))$	$\tilde{O}(\frac{1}{\gamma^2 T})$	$\tilde{O}(\frac{e^{O(L)}}{\gamma^2} \sqrt{\frac{m}{n} + \frac{1}{\gamma^2 T}})$
Thm. 2.1, Cor. 1	Smooth	$\Omega(\text{poly}(\frac{\log(n)}{\gamma}))$	$\tilde{O}(\frac{1}{\gamma^2 T})$	$\tilde{O}(\frac{e^{O(L)}}{\gamma^2 n} + \frac{1}{\gamma^2 T})$

Table 1: Comparison of our results on learning deep nets with GD under NTK separability condition to related prior results. Here m : network width, L : network depth, γ : NTK-margin, n : number of samples and T : number of iterations.

	Activation	Width	Iteration	Sample
Telgarsky (2023); Chen et al. (2021); Taheri & Thrampoulidis (2024); Cor.1	all	$\Omega(\text{poly}(d))$	d^2	$\tilde{O}(d^2)$
Glasgow (2024)	ReLU	$\Omega(\text{poly}(\log(d)))$	$\text{poly}(\log(d))$	$\tilde{O}(d)$
Thm. 2.4	Quadratic	$\Omega(1)$	$\log(d)$	$\tilde{O}(d)$

Table 2: Comparison of our findings on learning the d -dimensional XOR distribution with SGD to relevant prior results.

$\|w^* - w_0\| = O_m(1)$, where $w^* \in \mathbb{R}^p$ denotes the vector of target weights, w_0 is the initial weight vector, and m is the network width Chen et al. (2021); Ji & Telgarsky (2020); Telgarsky (2023).

Yet, even for the relatively simple setting of learning deep nets in the kernel regime, the existing generalization bounds are still suboptimal. Moreover, the boundaries of the kernel regime are still largely unknown and an active area of research Liu et al. (2022); Banerjee et al. (2022); Telgarsky (2023). While the kernel regime can partially demonstrate the behavior of neural networks, the resulting guarantees often require large width, small step-size or large iteration and sample complexities. There is increasing evidence in recent years that for certain class of data distributions neural networks can overcome these limitations by using a large step-size which allows the network’s parameters to move a long distance from initialization, often leading to better sample and computational complexities Damian et al. (2022); Ba et al. (2022).

In this work, we study the generalization and convergence behavior of gradient-based algorithms in neural nets with smooth activation functions for a wide class of data distributions. Our first result characterizes the test and train loss rates for classification problems under the condition that deviation from initialization is bounded depending on the network’s width. In particular, for L -hidden layer networks our results hold under $\|w^* - w_0\| \lesssim m^{O(1/L)}$, allowing the network’s weights to move from initialization a distance increasing with m . This shows that the kernel regime continues to hold for a wider range of setups than previous results for which the deviations are restricted to be constant in m . The key reason for this improvement is exploiting the objective’s Hessian structure in the gradient-descent path. More importantly, using the Hessian information enables us to develop, for the first time, *algorithm-dependent* generalization bounds of deep neural networks. As will be discussed throughout the paper, the bounds we derive via algorithmic-stability are tighter than previous relevant bounds in the literature. We specialize these results to a well-known NTK separability condition tailored for noiseless data and show that our results substantially improve the prior results on the test error performance while still allowing the width to be small, specifically poly-logarithmic on sample size (Table 1). A more detailed comparison is deferred to Section 1.2. We also consider the case of noisy data distributions and show that deep models are consistent, i.e., they can achieve the optimal test loss in the presence of noise as the sample size grows.

While these results improve upon the existing bounds within the NTK regime, we show in Section 2.2 that using a large step-size can further improve both the computational and sample complexities. In particular, we show for the stylized setup of data following the XOR distribution, a two-layer neural network with quadratic activation reaches zero test error after only $\log(d)$ steps of SGD with an aggressive step-size $\eta = m$. A comparison of our findings with the guarantees of the kernel regime with both GD Taheri & Thrampoulidis (2024) and one-pass SGD Telgarsky (2023); Chen et al. (2021) and the most relevant work in the feature learning regime Glasgow (2024) is summarized in Table 2.

Below is a summary of our contributions.

- In Theorem 2.1, we develop sufficient conditions for the global convergence of gradient descent in deep and smooth networks and show that if $m = \Omega(\|w^* - w_0\|^{6L+4})$, the training loss is bounded by $O(\frac{\|w^* - w_0\|^2}{\eta T})$, where η is the step-size and w^* can be any choice of network weights that achieves small training loss. Under similar conditions on m , we show the generalization error is bounded by $O(\frac{\|w^* - w_0\|^2 G_0^2}{n})$ where G_0 is the Lipschitz parameter of the network at initialization.
- In Corollary 1, we interpret these results by specializing them to a commonly-used margin-based NTK separability condition. The results of the corollary and comparison to previous works in literature are summarized in Table 1. A promising feature of our approach is that the test loss bound does not have an unfavorable dependence on the width while still maintaining minimum poly-logarithmic width conditions, which is new in the context of deep learning. To the best of our knowledge this is the tightest test error bound for deep nets trained by GD in the NTK regime.
- We consider the more general case of noisy data with non-vanishing optimal test loss in Theorem 2.3 and show that under a polynomial growth condition on network width, GD achieves a convergence rate of $1/\sqrt{n}$ to the optimal loss after $T = \sqrt{n}$ iterations.
- In Section 2.2, we consider the d -dimensional XOR distribution and show that a one-hidden layer network of constant width after exactly $\log(d)$ iterations of SGD with step-size $\eta = m$ achieves perfect test accuracy with $n = \tilde{O}(d)$ samples, considerably improving kernel regime’s limitations.

1.2 PRIOR WORKS

Generalization of deep nets. Among prior works on the generalization capabilities of deep networks, the only initialization dependent bounds were provided in Bartlett et al. (2017) obtaining bounds of order $O(\frac{\mathcal{R}}{\sqrt{n}})$ where the Rademacher complexity is derived as $\mathcal{R} := (\prod_{i=1}^L \|W_i\|_2) (\sum_{i=1}^L \frac{\|W_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|W_i\|_2^{2/3}})^{3/2}$. Here $\|W_i\|_2$ is the spectral norm of the weight matrix of layer i (typically a constant) and M_i s are any data-independent matrices. Thus one can choose $M_i = W_{i,0}$, i.e., the initialization weight matrix. In fact, the above bound resembles the bound that we obtained via an optimization-dependent stability analysis. However, note that \mathcal{R} depends on the distance traversed by weights through the $\ell_{2,1}$ norm which is always larger than the Frobenius norm, and in the worst case, the gap can be significantly large depending on the width. To see this, note that for a matrix $V \in \mathbb{R}^{m \times m}$ it holds that $\|V\|_{2,1} \leq \sqrt{m} \|V\|_F$. We note that “initialization-independent” bounds (e.g., Neyshabur et al. (2018); Golowich et al. (2018)) that are usually proportional to $\|w_t\|$ (rather than $\|w_t - w_0\|$) are strictly looser than the bound we obtain. This is primarily due to the fact that $\|w_t - w_0\|$ can be much smaller than $\|w_t\|$ and in fact as our experiments show $\|w_t - w_0\|$ is of constant order and can even decrease with width. Whereas, $\|w_t\|$ (or $\|w_t\|/\sqrt{m}$ due to the normalization in our setup) grows by increasing m , making the initialization-independent bounds potentially grow with width at the rate $O(\sqrt{m})$, despite lacking an explicit dependence on m . Hence, for wide networks, prior generalization bounds of deep neural nets based on Rademacher complexity are larger than the bound we derive in Theorem 2.1.

Test rates under the NTK separability condition. Other works that provide generalization bounds and optimization guarantees for neural nets include Cao & Gu (2019); Nitanda et al. (2019); Ji & Telgarsky (2020); Chen et al. (2021); Richards & Rabbat (2021); Taheri & Thrampoulidis (2024); Wang et al. (2023). In particular, Ji & Telgarsky (2020) derived the width condition $m = \Omega(\frac{\text{poly}(\log(n))}{\gamma^8})$ for achieving the $\frac{1}{\gamma^2 \sqrt{n}}$ -test error rate in two-layer nets via a uniform-convergence argument Shalev-Shwartz & Ben-David (2014). This bound was extended to deep networks in Chen et al. (2021) with a generalization gap of order

$$\tilde{O}\left(\frac{4^L}{\gamma^2} \sqrt{\frac{m}{n}} \wedge \left(\frac{L^{3/2}}{\gamma^2 \sqrt{n}} + \frac{L^{11/3}}{\gamma^2 m^{1/6}}\right)\right),$$

where \wedge takes the minimum of two quantities. As discussed earlier, this bound is dependent on width since it is derived essentially by taking the minimum of two generalization bounds based on Rademacher complexity derived in Bartlett et al. (2017) and Cao & Gu (2019). Importantly, in the small width regime the bound simplifies into $\tilde{O}(\frac{4^L}{\gamma^2} \sqrt{\frac{m}{n}})$, which has an undesirable dependence on

the width. In this paper, we improve the generalization gap to $\tilde{O}(\frac{e^{O(L)}}{\gamma^2 n})$ under the width condition $m = \Omega(\text{poly}(\log(n)/\gamma))$. To the best of our knowledge, these are the smallest generalization bound and width condition in literature up to now for learning deep neural nets. The key reason behind the improvement is leveraging the Hessian structure of the objective throughout the gradient descent iterates. The improved generalization guarantees result from the algorithmic dependency of our bounds and in fact the bounds can even be expressed such that they solely depend on the cumulative training loss (c.f. Eq. 5). As the training loss captures the role of initialization and is independent of width, the resulting generalization bounds share the same favorable properties.

Feature learning and the XOR distribution. Some recent works have pointed out the limitations of the kernel regime in understanding the full power of neural networks Abbe et al. (2022). In particular, as in the kernel regime, the networks weights are bounded not to move significantly from initialization, the learned features are not considerably different from those learned at initialization. On the other hand, it is hypothesized that neural networks can learn the true underlying features of the data distribution if the network weights are allowed (by large step-sizes or avoiding early-stopping) to move a large distance from initialization. This phenomenon was first proved for specific regression tasks where the labels essentially only depend on a small number of features, such as when $y = g(Ux)$ for $U \in \mathbb{R}^{k \times d}$ where $k \ll d$ Damian et al. (2022); Ba et al. (2022); Abbe et al. (2022); Cui et al. (2024), by one large SGD step leading to superior sample complexities compared to the kernel regime. For classification tasks, some focus has been on the XOR distribution (a.k.a. parities) Wei et al. (2019). Recent works have studied the problem of learning the d -dimensional XOR distribution using neural networks in both NTK and feature learning settings Barak et al. (2022); Telgarsky (2023); Taheri & Thrampoulidis (2024); Glasgow (2024). Specifically, it has been shown that under NTK with a sufficiently small step size, a polynomially wide network requires d^2 GD steps and d^2 sample size. Some studies have achieved linear sample complexity for learning XOR Bai & Lee (2019); Glasgow (2024); Telgarsky (2023); but these methods involve more computational effort compared to our results. The work most related to ours is Glasgow (2024), which demonstrated that with a particular Gaussian initialization, a ReLU network requires $\text{poly}(\log(d))$ large SGD steps and $\text{poly}(\log(d))$ neurons to learn XOR with linear sample complexity. In contrast, we show that by using a quadratic activation, learning this distribution requires only $\log(d)$ large steps with a constant-width network, while maintaining the same linear sample complexity.

NOTATION

Probability and expectation with respect to the randomness in random variable x are denoted by $\Pr_x(\cdot)$ and $\mathbb{E}_x[\cdot]$. We use $\mathbb{E}_x[w_t]$ to represent the weights after t steps of gradient descent using the population's gradient. We use the standard complexity notation $\lesssim, o(\cdot), O(\cdot), \Theta(\cdot), \Omega(\cdot)$ and use $\tilde{o}(\cdot), \tilde{O}(\cdot), \tilde{\Theta}(\cdot), \tilde{\Omega}(\cdot)$ to hide polylogarithmic factors. We denote $a \wedge b := \min\{a, b\}$. The gradient and Hessian of the model $\Phi : \mathbb{R}^{p \times d} \rightarrow \mathbb{R}$ with respect to the first input (that is, weights) are denoted by $\nabla \Phi$ and $\nabla^2 \Phi$, respectively. The minimum eigenvalue of a symmetric matrix is denoted by $\lambda_{\min}(\cdot)$. We use $\|\cdot\|$ for the Euclidean norm of vectors and $\|\cdot\|_2$ for the spectral norm of matrices. We denote $[w_1, w_2] := \{w : w = \alpha w_1 + (1 - \alpha)w_2, \alpha \in [0, 1]\}$ the line between $w_1, w_2 \in \mathbb{R}^p$.

2 MAIN RESULTS

Throughout the paper, we consider the following unregularized objective for a neural network classifier parameterized with $w \in \mathbb{R}^p$,

$$\min_{w \in \mathbb{R}^p} \hat{F}(w) := \frac{1}{n} \sum_{i=1}^n f(y_i \Phi(w, x_i)), \quad (1)$$

with data points satisfying $\|x\| \leq 1$, the binary labels $y_i \in \{\pm 1\}$ and $f(\cdot)$ is a loss function for classification tasks such as the logistic loss, $f(t) := \log(1 + e^{-t})$ and $\Phi(\cdot, x)$ is the network's output. We also define the test loss as $F(w) := \mathbb{E}_{x,y}[f(y\Phi(w, x))]$.

2.1 TRAIN AND TEST LOSS BOUNDS IN DEEP NETS

In our first theorem, we establish conditions for the width and target weights of the network that guarantees the training loss decays to zero if the network can interpolate the training set. We consider gradient descent update rule where at any iteration $t \leq T$: $w_{t+1} = w_t - \eta \nabla \hat{F}(w_t)$. Before stating the theorem, we note that this result is valid under the standard descent-lemma condition for the step-size, as stated in Lemma 3 in the appendix. In particular, the descent lemma holds if the step-size satisfies the standard Eq. 34 in the appendix.

Theorem 2.1 (Train & Test loss of deep nets). *Consider L -layer network where $\Phi(w, x) := \frac{1}{\sqrt{m}} W_{L+1}^\top (\frac{1}{\sqrt{m}} \sigma(W_L^\top \cdots \frac{1}{\sqrt{m}} \sigma(W_1^\top x) \cdots)$ and σ is a 1-smooth and 1-Lipschitz activation function such that $\sigma(0) = 0$. Moreover, let β_L be a constant that only depends on L , let all parameters of the network be initialized as i.i.d. standard Gaussian and assume the step-size satisfies the condition of the descent lemma. Fix T and assume the target weights vector $w^* \in \mathbb{R}^p$ that obtain small training loss such that*

$$\rho^* \geq \max \left\{ \sqrt{\eta T \hat{F}(w^*)}, \sqrt{\eta \hat{F}(w_0)} \right\}. \quad (2)$$

where $\rho^* := \|w^* - w_0\|$. Moreover, assume the width m is large enough such that it satisfies,

$$m \geq 4\beta_L^2 (6\rho^*)^{6L+4} \quad (3)$$

Then, $\|w_t - w_0\| = O(\rho^*)$ and the training loss satisfies with high probability over initialization,

$$\hat{F}(w_T) \leq \frac{4\rho^{*2}}{\eta T}. \quad (4)$$

Moreover, assume for every n samples from the data distribution there exists w^* satisfying Eqs. 2-3. Then, the expected generalization gap satisfies with high probability over initialization,

$$\mathbb{E}_S \left[F(w_T) - \hat{F}(w_T) \right] \leq 2.2 \frac{\eta(G_0 + 1/4)^2}{n} \mathbb{E}_S \left[\sum_{t=0}^{T-1} \hat{F}(w_t) \right], \quad (5)$$

where the expectation is over the randomness in the training set denoted by S and G_0 is the Lipschitz parameter of network at initialization i.e., $\|\nabla \Phi(w_0, \cdot)\| \leq G_0$.

In words, the main condition of the theorem is the existence of network weights denoted by w^* that achieves small training error (Eq. 2) and its distance from initialization is at most $O(m^{1/(6L+4)})$ as implied by Eq. 3. Under these conditions, the training loss is controlled solely by $\|w^* - w_0\|$ and has no explicit dependence on the width or depth of the network. As it will be stated in Corollary 1, in the NTK regime with margin γ it holds $\|w^* - w_0\| = O(\log(n)/\gamma)$, leading to the width condition $m = \Omega(\text{poly}(\log(n)/\gamma))$. This is unlike previous results which either required polynomial width (such as (Liu et al., 2022; Cao & Gu, 2019)) or led to sub-optimal bounds (e.g., (Chen et al., 2021; Ji & Telgarsky, 2020)).

In general, the theorem is valid for any feasible minimizer $w^* \in \mathbb{R}^p$. Thus, we can choose w^* with smallest value for $\|w^* - w_0\|$ to optimize the bounds. With such choice of w^* , the distance the weights obtained by GD travel is also minimized as $\forall t \in [T] : \|w_t - w_0\| = O(\|w^* - w_0\|)$. Thus, gradient descent tends toward solutions which attain small loss and lie at minimum possible distance from initialization. This is in line with related prior observations in several other works such as (Du et al., 2019; Oymak & Soltanolkotabi, 2019).

The theorem also shows the sample complexity and iteration complexity of learning deep networks with gradient descent and demonstrates the role of initialization and weight's norms on the test error. Note that by replacing the time-averaged training loss guarantees (cf. Theorem B.1), the generalization gap simplifies into:

$$\mathbb{E}_S \left[F(w_T) - \hat{F}(w_T) \right] \leq 9 \frac{\rho^{*2} (G_0 + 1/4)^2}{n}. \quad (6)$$

Hence, the test loss after $T = \Theta(n)$ iterations takes the form of

$$\mathbb{E}_S \left[F(w_T) \right] = O \left(\frac{\|w^* - w_0\|^2 G_0^2}{n} + \frac{\|w^* - w_0\|^2}{\eta n} \right), \quad (7)$$

where the first term is the generalization error and the second term is the training loss. Remarkably, Eq. 7 shows the tight correlation between the two terms, as the generalization gap is virtually the optimization error scaled by the squared Lipschitz constant G_0^2 . This is indeed the consequence of Eq. 5 which bounds the generalization gap based on the cumulative optimization error.

At a high-level, the test loss essentially depends on two quantities: (i) the Euclidean distance between the target weights and the initialization and (ii) the Lipschitz parameter of network at initialization. Due to the algorithmic-dependent nature of our generalization bounds and unlike the prior bounds in literature, the bound in Eq. 7 captures the role of initialization on the test error: smaller deviations from initialization lead to smaller test error bounds and in particular the bound approaches zero as ρ^* goes to zero. As discussed earlier, gradient descent favors such solutions with small deviations from initialization. In addition to “distance from initialization”, the test error bounds also depend on the squared Lipschitz parameter of the network. For standard Gaussian initialization, it can be shown (cf. Section A.2) that $G_0 \lesssim e^{O(L)}$ which introduces an exponential dependence on depth to the generalization bound. We remark that this dependence also appears in the corresponding bounds derived via uniform convergence and Rademacher complexity (e.g., (Bartlett et al., 2017; Golowich et al., 2018; Chen et al., 2021)) through the term $\prod_{i=1}^L \|W_i\|_2$.

An interesting feature of our approach is the algorithmic dependent bound in Eq. 5. This bound is generally tighter than the bound in Eq. 6. With the descent lemma condition on the step-size (c.f. Lemma 3) it holds that $\eta < 1/(G_0^2 + 1/4)$ which simplifies the bound into:

$$\mathbb{E}_{\mathcal{S}} \left[F(w_T) - \hat{F}(w_T) \right] \leq \frac{2.2}{n} \mathbb{E}_{\mathcal{S}} \left[\sum_{t=0}^{T-1} \hat{F}(w_t) \right]. \quad (8)$$

Hence, we have a bound which only depends on the training performance and the number of training samples. In our experiments in Section 3, we compute this bound for real-world data and compare it with the empirical results for generalization and test loss.

Remark 2.2. Our analysis relies on the recent progress in characterising the spectral norm of the deep net’s Hessian during GD updates (Liu et al., 2020; Banerjee et al., 2022). In particular, Liu et al. (2020) proved that with standard Gaussian initialization, the model’s Hessian is bounded with high probability by $\|\nabla^2 \Phi(w, x)\| = O(\frac{R^{3L}}{\sqrt{m}})$ if $\|w - w_0\| \leq R$. These guarantees can also be used to study the convergence rate of deep networks trained by quadratic loss as done by (Liu et al., 2020; 2022) but their approach leads to excessively large width conditions. In contrast, here we consider classification tasks with an improved poly-logarithmic width requirement and also study the generalization performance.

Recently, the algorithmic stability has been employed for two-layer neural nets in (Taheri & Thrampoulidis, 2024; Richards & Rabbat, 2021) which is an improved analysis of the stability-based approach typically used for convex objectives in (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Lei & Ying, 2020). Here, we essentially extend the stability analysis to deep networks. Compared to the two-layer nets, here in every iterate of gradient descent, the Hessian’s norm guarantees depend on the network’s weights. In particular, the analysis has to take into account that both $\|w_t - w_0\|$ and $\|w_t - w^*\|$ remain small during GD updates. This is necessary in order to ensure the Hessian’s norm guarantees and the approximate quasi convexity property hold during all GD iterates. We do this by an induction based argument which bounds these terms based on the fixed quantity $\|w^* - w_0\|$, which is guaranteed to be bounded based on width by assumption.

2.1.1 SPECIALIZING TO THE NTK-SEPARABILITY CONDITION

The results in the last section can be specifically applied to a class of data distributions that includes the XOR distribution which will be discussed more through the rest of the paper. Before stating our result in the corollary, we state the neural tangent kernel (NTK) separability assumption.

Assumption 1 (NTK-separability (Nitanda et al., 2019)). Assume the tangent kernel of the model at initialization separates the data with margin γ i.e., for a unit-norm vector $w \in \mathbb{R}^p$ it holds for all $i \in [n] : y_i \langle \nabla \Phi(w_0, x_i), w \rangle \geq \gamma$.

In words, the above assumption implies that the features learned by the gradient at initialization can be linearly separated by some weights w . This assumption is commonly used in deep learning literature for studying classification tasks (Chen et al., 2021; Ji & Telgarsky, 2020).

Corollary 1 (NTK results). *Consider the same setup as Theorem 2.1 and let Assumption 1 hold. Define constant $B > 0$ that bounds the model’s output at initialization i.e., $\forall i \in [n] : |\Phi(w_0, x_i)| < B$. Assume the width is large enough such that $m \geq \beta_L^2 \left(\frac{2B + \log(1/\epsilon)}{\gamma} \right)^{6L+4}$. Then there exists $w^* \in \mathbb{R}^p$ such that $F(w^*) \leq \epsilon$ and w^* lies at a bounded Euclidean distance from initialization such that $\|w^* - w_0\| \leq \frac{1}{\gamma}(2B + \log(1/\epsilon))$.*

To interpret this result, we apply this result to Theorem 2.1 and fix $\epsilon = 1/T$. First note that the output of network at initialization is constant with high-probability over initialization (e.g., see (Liu et al., 2020, Lemma F.4), (Cao & Gu, 2019, Lemma 4.4)) which implies $B = \tilde{O}(1)$. Overall, with the stopping condition $T = \Theta(n)$ and recalling that $G_0 \leq e^{O(L)}$, Corollary 1 yields the expected test error rate of order

$$\tilde{O} \left(\frac{e^{O(L)}}{\gamma^2 n} + \frac{1}{\gamma^2 T} \right),$$

under the condition that $m = \Omega(\text{poly}(\log(n)) \cdot \beta_L / \gamma^{6L+4})$. We remark the bound does not have any explicit dependence on m . This in itself is not surprising as one intuitively expects the bound not to scale unfavorably with width. Although recent works have derived width independent generalization bounds for two-layer networks (Ji & Telgarsky, 2020; Telgarsky, 2023; Taheri & Thrampoulidis, 2024), we are not aware of any prior work proving width-independent bounds for learning multi-layer networks with GD. As discussed earlier in the introduction, the bound derived in the closely related work (Chen et al., 2021) scales unfavorably with m as it grows at the rate $\sqrt{m/n}$. In fact, the authors refer to deriving width-independent bounds as an open problem in (Chen et al., 2021, Sec 3.1).

2.1.2 CONSISTENCY OF GD WITH NOISY DATA

The results of the last section mainly apply to data settings when the network can find the optimal solution within a bounded distance from initialization that depends on the network’s width. This setting was specially tailored to the noiseless case where achieving vanishing test loss was possible. We discuss next the case of learning deep nets by noisy data and show that achieving optimal test loss might be feasible in this setting.

Theorem 2.3 (Test error for noisy data). *Consider the same setup and notation as Theorem 2.1 and assume the width of the network satisfies $m \geq \beta_L^2 n^{3L+3}$ and the step-size satisfies the conditions of the descent lemma. Then, with high probability over initialization the expected test loss at iteration $T = \sqrt{n}$ is bounded as:*

$$\mathbb{E}_{\mathcal{S}} [F(w_T)] \leq \left(1 + \frac{4}{\sqrt{n}} \right) \left(F(w^*) + \frac{\rho^{*2}}{\eta\sqrt{n}} + \frac{\rho^{*2}}{n\sqrt{n}} + \frac{1}{\sqrt{n}} \right), \quad (9)$$

where $\rho^* := \|w^* - w_0\|$ and w^* is the minimizer of the population loss i.e., $w^* = \arg \min_{w \in \mathbb{R}^p} F(w)$.

The result above shows that GD reaches optimal level given that $\sqrt{n} \gg \|w^* - w_0\|^2$; as for large n , it leads to the simplified expression

$$\mathbb{E}_{\mathcal{S}} [F(w_T)] - F(w^*) = O \left(\frac{\|w^* - w_0\|^2 + F(w^*)}{\sqrt{n}} \right).$$

Thus, even in non-interpolating regime, GD can still achieve the optimal solution of over-parameterized deep networks.

The bound in Eq. 9 is derived via a stability argument which bounds the generalization gap based on the training performance. However, contrary to the conditions of Theorem 2.1 here we do not have the interpolation condition as $F(w^*)$ is not vanishing. This comes at the expense of a larger width condition where the width is polynomial in n whereas poly-logarithmic width was sufficient in Theorem 2.1. The condition on early stopping further guarantees that the test loss reach the Bayes error given n is sufficiently large. It is worth noting that the setting above still is operating almost in the NTK regime. This can be verified by the observation that for the bound to be meaningful it should hold that $\|w^* - w_0\| \ll \sqrt{n} \lesssim m^{O(1/L)}$, as per the theorem’s width condition. Finally, we remark that the result of Thm 2.3 nicely connects to recent empirical and theoretical results (Li et al., 2020) which show that with early-stopping, GD can find a good solution for clustered data with label noise.

2.2 OVERCOMING NTK LIMITATIONS FOR XOR DATASET

Next, we show the previous NTK guarantees can be improved by using large step-sizes. We consider the stylized set-up of learning the d -dimensional XOR data distribution. Consider one-hidden layer network with quadratic activation where $x \in \mathbb{R}^d, w_i \in \mathbb{R}^d$:

$$\Phi(w, x) = \frac{1}{2m} \sum_{i=1}^m a_i (x^\top w_i)^2.$$

In the above, $a_i \in \{\pm 1\}$ are fixed during training and satisfy $\sum_i a_i = 0$ and only the first layer weights are trained. For initialization of first layer's weights we have $\forall i \leq m, j \leq d : w_{0,ij} \stackrel{\text{iid}}{\sim} N(0, \frac{1}{d})$. The data points $(x, y) \in \{\pm 1\}^d \times \{\pm 1\}$ are uniformly drawn from the resulting d -dimensional distribution of 2^d points where the labels are determined as $y = x(1) \cdot x(2)$. For the result in the next theorem, we assume the loss function is the linear loss $f(t) = -t$ and consider mini-batch SGD with the update rule $w^{t+1} = w^t - \eta \hat{F}(w^t)$, where at each step, n data points are drawn i.i.d. from the distribution to form $\hat{F}(\cdot)$. The next theorem shows the computational and sample complexities of learning this dataset for the aforementioned setup.

Theorem 2.4 (Improved guarantees for learning XOR). *Assume mini-batch SGD with batch-size $n \geq d \cdot \log^{14}(d)$ and step-size $\eta = m$. Then, after $T = \log(d)$ iterations the test accuracy satisfies with probability at least $1 - e^{\log(m) - \log^2(d)} - e^{-\frac{m}{16}} - o_d(1)$ over the randomness of initialization and data sampling,*

$$\mathbb{E}_{x,y} \left[\mathbf{1}_{\{y\Phi(w^T, x) > 0\}} \right] \geq 1 - o_d(1). \quad (10)$$

Hence, with logarithmic number of SGD steps we use $\tilde{O}(d)$ samples in total to reach almost perfect test accuracy. We remark based on the guarantees of Theorem 2.4, the network's width can be constant and at most must be polynomial in d in order for Eq. 10 to hold with high probability. This aligns with our experiments in Fig. 4, demonstrating that the network's width can be independent of data dimension as a network of constant width (where $m = 20$) suffices for learning arbitrary high-dimensional XOR data. We also note the considerable gains in iteration, sample and computational complexities by Theorem 2.4 resulting from escaping the NTK regime with our step-size selection. Contrary to Theorem 2.1 which required small step-size in accordance with the descent lemma, here the step-size is proportional to width. For a comparison with recent results for this dataset we refer to Table 2. To the best of our knowledge these are the best complexities on iteration and network width for this setup.

Remark 2.5. The proof of Theorem 2.4 (provided in Appendix E) is nuanced and involves computing expected weights and their corresponding error terms due to SGD sampling noise for each parameter of the network. It is then showed that signal strength (i.e., the strength of important features) grows as $\frac{2^t}{\sqrt{d}}$ whereas the error terms due to sampling noise and initialization grow at most at the rate $(1 + \frac{1}{\text{poly}(\log(d))})^t \frac{\sqrt{d}}{\text{poly}(\log(d))} + \frac{2^t t}{\sqrt{d} \cdot \text{poly}(\log(d))}$. Therefore, after $\log(d)$ steps the noise strength reaches $\frac{\sqrt{d}}{\text{poly}(\log(d))}$ whereas the signal's magnitude is at least \sqrt{d} , letting the signal outgrow the noise and leading to the network classifying every point correctly.

3 NUMERICAL RESULTS

Experiments on learning under NTK with small step-size. In this section, we present numerical results on the behavior of the generalization bound derived in Theorem 2.1 for real-world data (FashionMNIST and MNIST datasets) and compare it with the empirical generalization gap.

For demonstrating our theoretical results, we are interested in the algorithmic-based generalization bound (Eq. 8) derived as,

$$\mathbb{E}_{\mathcal{S}} \left[F(w_T) - \hat{F}(w_T) \right] \leq \frac{2.2}{n} \mathbb{E}_{\mathcal{S}} \left[\sum_{t=0}^{T-1} \hat{F}(w_t) \right]. \quad (11)$$

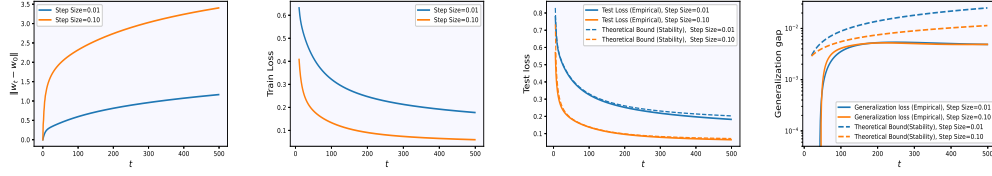


Figure 1: Iteration-based distance from initialization ($\|w_t - w_0\|$), training loss, test loss and generalization gap (i.e., test loss – train loss) for training a two hidden-layer neural network with *FashionMNIST* dataset and two choices of step-size. Here $n = 12 \times 10^3$, $m = 500$, and total number of parameters $p \approx 6 \times 10^5$.

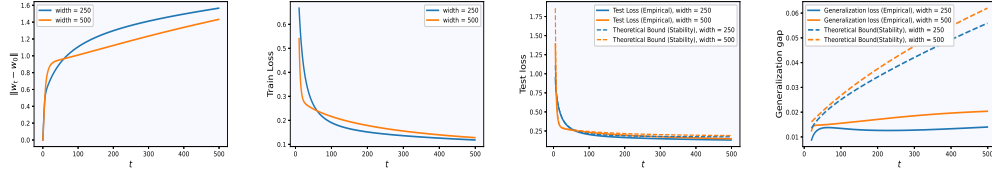


Figure 2: Iteration-based distance from initialization, training loss, test loss and generalization gap for training a two hidden-layer neural network with *FashionMNIST* dataset and $m = 250, 500$. Here $n = 4 \times 10^3$, $p \approx 2 \times 10^5$ (blue line), 6×10^5 (red line), and $\eta = 0.02$.

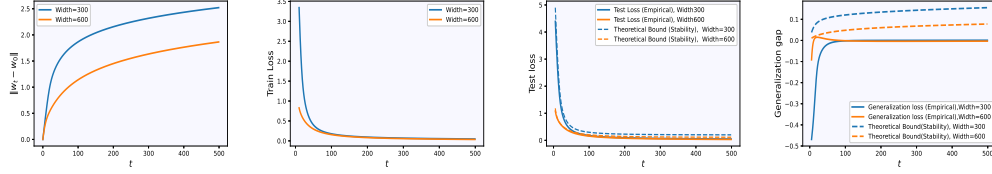


Figure 3: Iteration-based distance from initialization, training loss, test loss and generalization gap for training a two hidden-layer neural network with *MNIST* dataset and $m = 300, 600$. Here $n = 2 \times 10^3$, $p \approx 3 \times 10^5$ (blue line), 8×10^5 (red line) and $\eta = 0.02$.

We find it helpful to note that the bound above requires the width condition in the theorem (i.e., $m = \Omega(\|w^* - w_0\|^{6L+4})$) to be valid. However, verifying this condition is not feasible in general. Moreover, the bound is valid for *expected* generalization gap where the expectation is taken over data sampling. Therefore computing the exact values of both sides of the above inequality is computationally exhaustive. For our experiments we consider one realization to estimate these values. Due to both of these reasons, the theoretical test loss and generalization loss that we present in this section should only be taken as approximations of the general behavior of the bound and not as an actual verified bound on the generalization. However, in order to reduce these impacts we conduct several experiments with varying network’s width.

We consider binary classification with a 2-hidden layer network with softplus activation ($\sigma(t) = \log(1 + e^t)$) trained by the logistic loss function. Figure 1 presents train, test and generalization behavior of GD for learning a such a model with *FashionMNIST* dataset. The two lines in each figure correspond to $\eta = 0.01, 0.1$. In the two rightmost plots, the resulting theoretical generalization and test loss curves derived from Eq. 11 are compared with the empirical values. Note the good alignment between theoretical and empirical behavior for the generalization and test loss.

A similar behavior is observed in Figures 2-3. In Figure 2, we consider a similar setup but we reduce the sample size to 4000 training data in order to allow larger test-loss behavior. The resulting plots show the training and test loss performance for two choices of $m = 250$ and $m = 500$. In Figure 3, we consider the *MNIST* dataset for $m = 300$ and $m = 600$ and let the sample size be $n = 2000$. It is

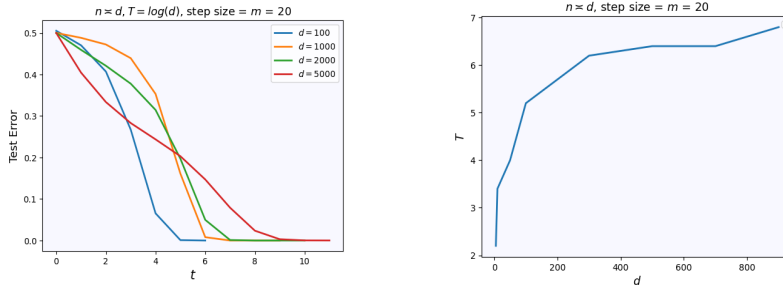


Figure 4: Left: Misclassification error based on iteration in learning the d -dimensional XOR distribution with SGD. Right: Total number of SGD steps based on data dimension to reach approximately zero test error.

noteworthy that the findings in both figures bear similarities, with the theoretical bounds providing non-vacuous and accurate approximations for the test loss and generalization gap.

Experiments on learning the XOR distribution with large step-size. Figure 4 demonstrates the test error curves associated with learning the XOR distribution according to the setting of Theorem 2.4. In particular, we fix $n = 6d$, $\eta = m = 20$ and set the total number of SGD steps as $T = \lceil \log(d) \rceil$. Note that the number of iterations required to reach perfect accuracy grows with d . The right side of Figure 4 provides further insight into the relationship between dimensionality and convergence rate. It displays the total number of SGD steps required to reach a test error below 0.01 for different values of d using $n = 3d$, $m = \eta = 20$. The results are averages over five independent experiments and highlight the logarithmic dependence of the total SGD iterations on data dimension for achieving near-zero test error.

4 CONCLUSIONS AND FUTURE DIRECTIONS

We explored the convergence and generalization of smooth neural networks trained with gradient methods. Our first goal in this paper was to derive generalization bounds through a new stability-based approach which had not been discussed in the vast literature of deep learning. These findings represent an improvement over previous results that either required substantial over-parameterization or provided suboptimal generalization rates that depended on the network width. For general noisy data distributions, we also derived generalization guarantees showing that GD can reach the optimal test loss. We also showed that orders of magnitude improvements in sample and computational complexity are possible by surpassing the NTK limitations and using SGD with large step-size. Several directions remain open to future research:

- It remains open to explore whether the minimum width conditions of Theorems 2.1-2.3 can be improved in terms of γ or n .
- It is also interesting to extend the XOR analysis to the noisy setting where a fraction of the data points have corrupted labels.
- The feature learning phenomenon in multi-index classification tasks remains largely unexplored. While we believe the XOR setup is a must-take first step, extensions to other multi-index models can shed further light on the strengths and limitations of neural networks.
- We also aim to understand the potential benefits of network depth in either the NTK regime or feature learning i.e., whether adding a single layer can improve sample complexity or reduce the total number of SGD iterations.

ACKNOWLEDGEMENTS

This work was supported by NSF grant CCF 2217058.

REFERENCES

- Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.
- Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.
- Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.
- Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Mikhail Belkin. Restricted strong convexity of deep learning models with smooth activations. *arXiv preprint arXiv:2209.15106*, 2022.
- Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.
- Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.
- Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research (JMLR)*, 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.
- Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep re{lu} networks? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fgd7we_uZa6.
- Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue M Lu, Lenka Zdeborová, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. *arXiv preprint arXiv:2402.04980*, 2024.
- Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.
- Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.
- Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. In *The Twelfth International Conference on Learning Representations*, 2024.
- Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.
- Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, volume 48, pp. 1225–1234, 2016.
- Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.
- Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *International Conference on Learning Representations*, 2020.

- Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.
- Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 5809–5819, 2020.
- Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pp. 4313–4324. PMLR, 2020.
- Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33: 15954–15964, 2020.
- 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.
- Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.
- 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. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pp. 4951–4960. PMLR, 2019.
- Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *International Conference on Artificial Intelligence and Statistics*, pp. 1990–1998. PMLR, 2021.
- Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory (COLT)*, volume 178, pp. 3380–3394, 2022.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.
- Matus Telgarsky. Feature selection and low test error in shallow low-rotation relu networks. In *The Eleventh International Conference on Learning Representations*, 2023.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Puyu Wang, Yunwen Lei, Di Wang, Yiming Ying, and Ding-Xuan Zhou. Generalization guarantees of gradient descent for multi-layer neural networks. *arXiv preprint arXiv:2305.16891*, 2023.
- 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.

APPENDIX

A PRELIMINARIES : HESSIAN AND GRADIENT NORM

A.1 HESSIAN NORM

It is essential to our analysis to obtain Hessian's spectral norm throughout the iterates of GD. With our weights' initialization in place, one can bound the Hessian's spectral norm based on the euclidean distance of the weights from initialization and network's size as in the following assumption.

Assumption 2 (Hessian's Spectral Norm). The spectral norm of the model's Hessian for any $w \in \mathbb{R}^p$, $\rho > 0$ such that $\|w - w_0\| \leq \rho$ is bounded as $\|\nabla^2 \Phi(w, \cdot)\|_2 \leq \beta_L \frac{\rho^{3L}}{\sqrt{m}}$ for some constant β_L that depends only on L .

It can be shown (Liu et al., 2020, Lemma 1) that for standard Gaussian initialization and bounded data, Assumption 2 holds with high-probability over initialization for $\beta_L = e^{O(L)}$.

For the spectral norm of the model's Hessian which is needed by our analysis, we use the following bound which bounds the spectral norm based on the distance from initialization.

Lemma 1 (Hessian's spectral norm (Liu et al., 2020)). Assume the defined L layer neural network architecture with standard Gaussian initialization i.e., $w_{0,ij}^{(\ell)} \stackrel{iid}{\sim} N(0, 1)$. Let $\|w - w_0\| \leq \rho$. Then with high probability

$$\|\nabla^2 \Phi(w, x)\|_2 \leq \beta_L \frac{\rho^{3L}}{\sqrt{m}}$$

for some constant β_L that only depends on L .

As mentioned in the main body, the result above guarantees that Assumption 2 holds with high probability.

A.2 GRADIENT NORM AT INITIALIZATION

Recall the model for Theorems 2.1-2.3,

$$\begin{aligned} \Phi(w, x) &:= \frac{1}{\sqrt{m}} \langle W^{(L+1)}, x^{(L)} \rangle \\ x_i^{(\ell+1)} &:= \frac{1}{\sqrt{m}} \sigma(\langle W_i^{(\ell+1)}, x^{(\ell)} \rangle) \quad \forall \ell \in \{1, \dots, L-1\}, \forall i \in \{1, \dots, m\} \\ x_i^{(1)} &:= \sigma(\langle x, W_i^{(1)} \rangle) \quad \forall i \in \{1, \dots, m\} \end{aligned} \tag{12}$$

Let us consider the standard Gaussian initialization. We have

$$\frac{\partial \Phi(w, x)}{\partial W_{ij}^{(\ell+1)}} = \frac{1}{\sqrt{m}} x_j^{(\ell)} \sigma'(\langle W_i^{(\ell+1)}, x^{(\ell)} \rangle) \frac{\partial \Phi(w, x)}{\partial x_i^{(\ell+1)}} \tag{13}$$

Therefore,

$$\left\| \frac{\partial \Phi(w, x)}{\partial W^{(\ell+1)}} \right\|^2 = \sum_{ij} \left\| \frac{\partial \Phi(w, x)}{\partial W_{ij}^{(\ell+1)}} \right\|^2 \leq \frac{1}{m} \|x^{(\ell)}\|^2 \left\| \frac{\partial \Phi(w, x)}{\partial x^{(\ell+1)}} \right\|^2 \tag{14}$$

by (Liu et al., 2020, Lemmas F.3 and F.5) :

$$\|x^{(\ell)}\| \leq c_0^\ell \sqrt{m}, \quad \left\| \frac{\partial \Phi}{\partial x^{(\ell)}} \right\| \leq c_0^{L-\ell+1} \tag{15}$$

Therefore $\|\frac{\partial \Phi(w_0, x)}{\partial W^{(\ell+1)}}\|^2 \leq c_0^{2L}$ and

$$\|\nabla \Phi(w_0, x)\| = \left(\sum_{\ell=0}^L \left\| \frac{\partial \Phi(w_0, x)}{\partial W^{(\ell+1)}} \right\|^2 \right)^{1/2} \leq \sqrt{L+1} c_0^L.$$

where c_0 is any constant that bounds the spectral norm of initialization weights with high probability, i.e., $\frac{\|W_0\|_2}{\sqrt{m}} \leq c_0$. Specially, well-known bounds on the spectral norm of Gaussian matrices guarantee that c_0 is constant with high probability (Vershynin, 2018).

Moreover, based on (Liu et al., 2020), if $\|w - w_0\| = \rho$ the spectral norm is bounded by,

$$\|\nabla^2 \Phi(w, x)\|_2 \leq \beta_L \frac{\rho^{3L}}{\sqrt{m}} \quad (16)$$

where β_L is a depth-only dependent variable derived to be $\beta_L = e^{O(L)}$.

Now, we use the above relation on Hessian to obtain a bound on the gradient:

$$\frac{\|\nabla \Phi(w^*, x) - \nabla \Phi(w_0, x)\|}{\|w^* - w_0\|} \leq \max_{w \in [w_0, w^*]} \|\nabla^2 \Phi(w, x)\|_2 \quad (17)$$

therefore if $w^* \in \mathcal{B}_\rho(w_0)$,

$$\|\nabla \Phi(w^*, x) - \nabla \Phi(w_0, x)\| \leq \frac{\beta_L \rho^{3L+1}}{\sqrt{m}}. \quad (18)$$

Thus,

$$\|\nabla \Phi(w^*, x)\| - \|\nabla \Phi(w_0, x)\| \leq \|\nabla \Phi(w^*, x) - \nabla \Phi(w_0, x)\| \leq \frac{\beta_L \rho^{3L+1}}{\sqrt{m}}. \quad (19)$$

Thus w.h.p.

$$\max_{i \in [n]} \|\nabla \Phi(w^*, x_i)\| \leq \max_{i \in [n]} \|\nabla \Phi(w_0, x_i)\| + \frac{\beta_L \rho^{3L+1}}{\sqrt{m}} \quad (20)$$

Let us denote the gradient and hessian bounds on Φ with C_1 and C_2 , respectively. Note that based on our derivations above it holds that $C_1 \leq \frac{\beta_L \rho^{3L+1}}{\sqrt{m}} + G_0$ and $C_2 \leq \frac{\beta_L \rho^{3L}}{\sqrt{m}}$, where G_0 is the Lipschitz parameter at initialization. With these parameters we can now obtain the Lipschitz and Smoothness parameter as well as the smallest eigenvalue of the objective along the gradient descent path. This is derived in the next lemma.

A.3 OBJECTIVE'S GRADIENT AND HESSIAN ALONG THE GD PATH

Lemma 2. *The following are true for the loss gradient and Hessian:*

1. $\|\nabla \hat{F}(w)\| \leq C_1 \hat{F}(w)$.
2. $\|\nabla^2 \hat{F}(w)\|_2 \leq (C_2 + (C_1)^2) \hat{F}(w)$.
3. $\lambda_{\min}(\nabla^2 \hat{F}(w)) \geq -C_2 \hat{F}(w)$.

where we define $C_1 := G_0 + \beta_L \|w - w_0\|^{3L+1}/\sqrt{m}$, $C_2 := \beta_L \|w - w_0\|^{3L}/\sqrt{m}$ and G_0 is the Lipschitz parameter of the network at initialization i.e., $\|\nabla \Phi(w_0, \cdot)\| \leq G_0$.

Proof. The loss gradient is derived as follows,

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

Recalling that $y_i \in \{\pm 1\}$, we can write

$$\begin{aligned}\|\nabla \hat{F}(w)\| &= \frac{1}{n} \left\| \sum_{i=1}^n f'(y_i \Phi(w, x_i)) y_i \nabla \Phi(w, x_i) \right\| \\ &\leq \frac{1}{n} \sum_{i=1}^n |f'(y_i \Phi(w, x_i))| \|\nabla \Phi(w, x_i)\| \\ &\leq C_1 \hat{F}(w).\end{aligned}\tag{21}$$

For the Hessian of loss, note that

$$\nabla^2 \hat{F}(w) = \frac{1}{n} \sum_{i=1}^n f''(y_i \Phi(w, x_i)) \nabla \Phi(w, x_i) \nabla \Phi(w, x_i)^\top + f'(y_i \Phi(w, x_i)) y_i \nabla^2 \Phi(w, x_i).\tag{22}$$

It follows that

$$\begin{aligned}\|\nabla^2 \hat{F}(w)\|_2 &= \left\| \frac{1}{n} \sum_{i=1}^n f'(y_i \Phi(w, x_i)) y_i \nabla^2 \Phi(w, x_i) + f''(y_i \Phi(w, x_i)) \nabla \Phi(w, x_i) \nabla \Phi(w, x_i)^\top \right\| \\ &\leq \frac{1}{n} \sum_{i=1}^n |f'(y_i \Phi(w, x_i))| \|\nabla^2 \Phi(w, x_i)\| + |f''(y_i \Phi(w, x_i))| \|\nabla \Phi(w, x_i) \nabla \Phi(w, x_i)^\top\| \\ &\leq \frac{1}{n} \sum_{i=1}^n |f'(y_i \Phi(w, x_i))| \|\nabla^2 \Phi(w, x_i)\| + |f''(y_i \Phi(w, x_i))| \|\nabla \Phi(w, x_i)\|^2 \\ &\leq C_2 \hat{F}(w) + (C_1)^2 \hat{F}(w).\end{aligned}\tag{23}$$

To lower-bound the minimum eigenvalue of Hessian, note that the loss function f is convex and thus $f''(\cdot) \geq 0$. Therefore the first term in Eq. 22 is positive semi-definite and the second term can be lower-bounded as follows,

$$\begin{aligned}\lambda_{\min}(\nabla^2 \hat{F}(w)) &\geq - \left\| \frac{1}{n} \sum_{i=1}^n y_i f'(y_i \Phi(w, x_i)) \nabla^2 \Phi(w, x_i) \right\| \\ &\geq - \frac{1}{n} \sum_{i=1}^n |y_i f'(y_i \Phi(w, x_i))| \|\nabla^2 \Phi(w, x_i)\| \\ &\geq -C_2 \hat{F}(w).\end{aligned}$$

□

B PROOF OF THEOREM 2.1

With the bounds on Hessian and Gradient from the last section, we are ready to prove Theorem 2.1 for obtaining the training and generalization rates of deep nets. We start with proving the training guarantees in the following theorem.

B.1 PROOF FOR TRAINING LOSS

Theorem B.1 (Convergence of GD in deep nets). *Let the descent lemma hold. Fix a T and assume the target weights vector $w^* \in \mathbb{R}^p$ that obtain small training loss such that*

$$\rho^* \geq \max \left\{ \sqrt{\eta T \hat{F}(w^*)}, \sqrt{\eta \hat{F}(w_0)} \right\}.$$

where $\rho^* := \|w^* - w_0\|$. Moreover, assume the width m is large enough such that it satisfies,

$$m \geq 4\beta_L^2 (3\rho^*)^{6L+4}.$$

The the training loss satisfies with high probability over initialization,

$$\frac{1}{T} \sum_{t=1}^T \hat{F}(w_t) \leq 2\hat{F}(w^*) + \frac{2\rho^{*2}}{\eta T}.$$

To begin the proof, we first state our descent lemma.

Lemma 3 (Descent lemma). *Consider the notation as in Lemma 2. Assume for all $t \in [T]$ we have $\eta \leq 1/((C_1)^2 + C_2)$, then*

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

Proof. By Taylor expansion and the bound on smoothness parameter of the objective, we immediately deduce:

$$\begin{aligned} \widehat{F}(w_{t+1}) &\leq \widehat{F}(w_t) - \eta \left\| \nabla \widehat{F}(w_t) \right\|^2 + \eta^2 \frac{(C_1)^2 + C_2}{2} \left\| \nabla \widehat{F}(w_t) \right\|^2 \\ &\leq \widehat{F}(w_t) - \frac{\eta}{2} \left\| \nabla \widehat{F}(w_t) \right\|^2 \end{aligned}$$

This completes the proof of the lemma. \square

Define $\rho(t) := \|w_t - w_0\|$ and $\rho^* = \|w^* - w_0\|$. Also, define $\tilde{\rho}$ as any constant such that

$$\tilde{\rho} > \max\{\rho(T-1), \rho(T-2), \dots, \rho(1), \rho^*\}.$$

Then for any $w \in [w_t, w^*]$ such that $w = \alpha w_t + (1 - \alpha)w^*$ it holds that

$$\|w - w_0\| \leq \|\alpha w_t + (1 - \alpha)w^* - w_0\| \leq \alpha \|w_t - w_0\| + (1 - \alpha) \|w^* - w_0\| \leq \tilde{\rho},$$

and thus for any $w \in [w_t, w^*]$

$$\lambda_{\min} \left(\nabla^2 \widehat{F}(w) \right) \geq -\tilde{C} \widehat{F}(w) \quad (24)$$

where $\tilde{C} := \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}}$, i.e., the smoothness parameter with respect to the distance $\tilde{\rho}$ and as computed by Lemma 2. Therefore, by Taylor remainder theorem, for any fixed w^* there exists $w'_t \in [w_t, w^*]$ such that

$$\begin{aligned} \widehat{F}(w^*) &\geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle + \frac{1}{2} \lambda_{\min} \left(\nabla^2 \widehat{F}(w'_t) \right) \|w^* - w_t\|^2 \\ &\geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle - \frac{1}{2} \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}} \widehat{F}(w'_t) \|w^* - w_t\|^2. \end{aligned}$$

As a direct result of the above inequality, we have derived that for any w^* ,

$$\widehat{F}(w^*) \geq \widehat{F}(w_t) + \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle - \frac{1}{2} \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}} \max_{w'_t \in [w_t, w^*]} \widehat{F}(w'_t) \|w^* - w_t\|^2.$$

Putting this in the relation from descent lemma (Lemma 3) followed by completion of squares using $w_{t+1} - w_t = -\eta \nabla \widehat{F}(w_t)$ gives

$$\begin{aligned} \widehat{F}(w_{t+1}) &\leq \widehat{F}(w^*) - \left\langle \nabla \widehat{F}(w_t), w^* - w_t \right\rangle - \frac{\eta}{2} \left\| \nabla \widehat{F}(w_t) \right\|^2 + \frac{1}{2} \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}} \max_{w'_t \in [w_t, w^*]} \widehat{F}(w'_t) \|w^* - w_t\|^2 \\ &= \widehat{F}(w^*) + \frac{1}{\eta} (\|w^* - w_t\|^2 - \|w^* - w_{t+1}\|^2) + \frac{1}{2} \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}} \max_{w'_t \in [w_t, w^*]} \widehat{F}(w'_t) \|w^* - w_t\|^2. \end{aligned} \quad (25)$$

Telescoping the above display for $t = 0, \dots, T-1$, we deduce that,

$$\frac{1}{T} \sum_{t=1}^T \widehat{F}(w_t) \leq \widehat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \frac{\beta_L \tilde{\rho}^{3L}}{2T} \sum_{t=0}^{T-1} \max_{w'_t \in [w_t, w^*]} \widehat{F}(w'_t) \|w^* - w_t\|^2. \quad (26)$$

In order to proceed with the above relation, it is necessary to further simplify the terms $\max_{w'_t \in [w_t, w^*]} \widehat{F}(w'_t)$. In doing so, we utilize the following Generalized Local Quasi-convexity (GLQC) property from Taheri & Thrampoulidis (2024).

Proposition 1 (Prop. 8 in Taheri & Thrampoulidis (2024)). *Let $w_1, w_2 \in \mathbb{R}^p$ be two arbitrary points. Suppose $\hat{F} : \mathbb{R}^p \rightarrow \mathbb{R}$ satisfies the self-bounded weak convexity property with parameter κ along the line $[w_1, w_2]$. Assume $\|w_1 - w_2\| \leq D$ where $D < \sqrt{2/\kappa}$. Then,*

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

where $\tau := (1 - \kappa D^2/2)^{-1}$.

For our objective we have the minimum eigenvalue of hessian being lower bounded by $\kappa = \tilde{C} := \beta_L \tilde{\rho}^{3L}$. Therefore for w_t, w^* , if $\|w^* - w_t\|^2 < D^2 < 1/2\tilde{C}$, then the GLQC property holds with $\tau = 4/3$. Recall that by the definition of $\tilde{\rho}$ it holds that $\|w^* - w_t\| \leq \tilde{\rho}$. Thus, to guarantee the conditions of the proposition above are satisfied it suffices to hold that

$$\sqrt{m} \geq 2\beta_L \tilde{\rho}^{3L+2}. \quad (28)$$

which is exactly the width condition of our theorem. With this condition on width and the resulting simplifications from the proposition, we can write:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \hat{F}(w_t) &\leq \hat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}} \frac{2}{3T} \sum_{t=0}^{T-1} \max\{\hat{F}(w^*), \hat{F}(w_t)\} \|w^* - w_t\|^2, \\ &\leq \hat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \frac{1}{3T} \sum_{t=0}^{T-1} \max\{\hat{F}(w^*), \hat{F}(w_t)\} \\ &\leq \frac{4}{3} \hat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \frac{1}{3T} \sum_{t=0}^{T-1} \hat{F}(w_t) \end{aligned}$$

Thus,

$$\frac{1}{T} \sum_{t=1}^T \hat{F}(w_t) \leq 2\hat{F}(w^*) + \frac{3\|w^* - w_0\|^2}{2\eta T} + \frac{1}{2T} \hat{F}(w_0) \quad (29)$$

Thus far, we have characterized the convergence under the condition that $\sqrt{m} \geq 2\beta_L \tilde{\rho}^{3L+2} = 2\beta_L (\max\{\|w_{T-1} - w_0\|, \dots, \|w_1 - w_0\|, \|w^* - w_0\|\})^{3L+2}$. It remains to characterize the above condition on width independent of the iteration number. In fact, in what follows we are able to show it is sufficient to reduce this condition such that it depends on only $\|w^* - w_0\|$. We do it by an iterate-wise norm bound on $\|w^* - w_t\|$ based on $\|w^* - w_0\|$.

Recall that we had,

$$\hat{F}(w_{t+1}) \leq \hat{F}(w^*) + \frac{1}{\eta} (\|w^* - w_t\|^2 - \|w^* - w_{t+1}\|^2) + \frac{1}{2} \tilde{C} \max_{w'_t \in [w_t, w^*]} \hat{F}(w'_t) \|w^* - w_t\|^2. \quad (30)$$

where recall that $\tilde{C} = \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}}$, and $\tilde{\rho} := \max\{\rho(1), \rho(2), \dots, \rho(T-1), \rho^*\}$.

Hence,

$$\|w_{t+1} - w^*\|^2 \leq \|w_t - w^*\|^2 + \eta \hat{F}(w^*) + \frac{\eta \beta_L \tilde{\rho}^{3L}}{2\sqrt{m}} \max_{w'_t \in [w_t, w^*]} \hat{F}(w'_t) \|w^* - w_t\|^2. \quad (31)$$

The following lemma proves through an induction-based argument that choosing $\tilde{\rho} = 3\rho^*$ and the corresponding width condition based on this $\tilde{\rho}$ is sufficient to ensure that $\rho(t) \leq \tilde{\rho}$ for all iterates $t \in [T]$; which in turn guarantees the above argument for convergence to be valid.

Lemma 4. *Fix any $T \geq 0$ and assume any w^* such that*

$$\|w^* - w_0\|^2 \geq \max\{\eta T \hat{F}(w^*), \eta \hat{F}(w_0)\}. \quad (32)$$

Pick $\tilde{\rho} = 3\rho^ = 3\|w^* - w_0\|$ and assume m is large enough such that $\tilde{\rho}^2 < 1/2\tilde{C}$ where $\tilde{C} = \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}}$. Then $\rho(t) := \|w_t - w_0\| \leq \tilde{\rho}$ and $\|w_t - w^*\| \leq \tilde{\rho}$ for all $t \in [T]$.*

Proof. The statements of the lemma are correct for $t = 0$. Now assume for $t \leq T - 1$ we have $\rho(t) \leq \tilde{\rho}$ and $\|w_t - w^*\| \leq \tilde{\rho}$. Recall, by 31

$$\|w_T - w^*\|^2 \leq \|w_{T-1} - w^*\|^2 + \eta \hat{F}(w^*) + \frac{\eta \beta_L \tilde{\rho}^{3L}}{2\sqrt{m}} \|w_{T-1} - w^*\|^2 \max_{w'_{T-1} \in [w_{T-1}, w^*]} \hat{F}(w'_{T-1}) \quad (33)$$

Note that by the assumption of the lemma and the induction assumption it holds that

$$\max\{\|w_{T-1} - w_0\|, \|w^* - w_0\|\} \leq \tilde{\rho}.$$

As a result, for any $w \in [w_{T-1}, w^*]$, it holds that $\|w - w^*\| \leq \tilde{\rho}$ and thus $\|\nabla \hat{F}(w)\| \geq \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}} \hat{F}(w)$.

Moreover, note that as per the lemma's assumption $\|w_{T-1} - w^*\|^2 \leq \tilde{\rho}^2 \leq 1/2\tilde{C}$. Thus, the conditions of Proposition 1 hold with $\tau = 4/3$ and we have

$$\|w_T - w^*\|^2 \leq \|w_{T-1} - w^*\|^2 + \eta \hat{F}(w^*) + \frac{2\eta \beta_L \tilde{\rho}^{3L}}{3\sqrt{m}} \|w_{T-1} - w^*\|^2 \cdot \max\{\hat{F}(w_{T-1}), \hat{F}(w^*)\}$$

Noting again that $\|w_{T-1} - w^*\|^2 \leq \tilde{\rho}^2 \leq 1/2\tilde{C}$ yields,

$$\begin{aligned} \|w_T - w^*\|^2 &\leq \|w_{T-1} - w^*\|^2 + \eta \hat{F}(w^*) + \frac{\eta}{3} \max\{\hat{F}(w_{T-1}), \hat{F}(w^*)\} \\ &\leq \|w_{T-1} - w^*\|^2 + \frac{4\eta}{3} \hat{F}(w^*) + \frac{\eta}{3} \hat{F}(w_{T-1}) \end{aligned}$$

Since the conditions of the lemma hold for all $t \in [T - 1]$, we can telescope the summation to obtain,

$$\begin{aligned} \|w_T - w^*\|^2 &\leq \|w_0 - w^*\|^2 + \frac{4\eta}{3} T \hat{F}(w^*) + \frac{\eta}{3} \sum_{t=0}^{T-1} \hat{F}(w_t) \\ &\leq \|w_0 - w^*\|^2 + \frac{4\eta}{3} T \hat{F}(w^*) + \frac{\eta}{3} \hat{F}(w_0) + \frac{\eta}{3} \sum_{t=1}^{T-1} \hat{F}(w_t) \\ &\leq \|w_0 - w^*\|^2 + \frac{4\eta}{3} T \hat{F}(w^*) + \frac{\eta}{3} \hat{F}(w_0) + \frac{\eta}{3} \left(2T \hat{F}(w^*) + \frac{3\|w^* - w_0\|^2}{2\eta} + \frac{1}{2} \hat{F}(w_0) \right) \\ &= \|w_0 - w^*\|^2 + 2\eta T \hat{F}(w^*) + \frac{\eta}{2} \hat{F}(w_0) + \frac{\|w^* - w_0\|^2}{2} \end{aligned}$$

Therefore by the conditions of the lemma and noting that $A_0 = \|w_0 - w^*\|^2 = (\rho^*)^2$:

$$\|w_T - w^*\|^2 \leq \|w_0 - w^*\|^2 + 2\|w_0 - w^*\|^2 + \frac{1}{2}\|w^* - w_0\|^2 + \frac{1}{2}\|w^* - w_0\|^2 = 4(\rho^*)^2$$

Thus, $\|w_T - w^*\|^2 \leq 9(\rho^*)^2 = \tilde{\rho}^2$. It remains to prove that $\rho(T) := \|w_T - w_0\| \leq \tilde{\rho}$. To obtain this, note that by triangle inequality,

$$\begin{aligned} \|w_T - w_0\| &\leq \|w_T - w^*\| + \|w^* - w_0\| \\ &\leq 2\rho^* + \rho^* = 3\rho^* = \tilde{\rho}. \end{aligned}$$

The proof is complete. \square

The lemma above shows that controlling $\rho^* = \|w^* - w_0\|$ is enough to guarantee that GLQC property (Proposition 1) holds throughout the gradient path. In particular, if the width is large enough such that for $\tilde{\rho} = 3\rho^*$: $\tilde{\rho}^2 \leq \frac{1}{2\tilde{C}} = \frac{\sqrt{m}}{2\beta_L \tilde{\rho}^{3L}}$ or equivalently

$$\sqrt{m} \geq 2\beta_L \tilde{\rho}^{3L+2}.$$

then the conditions of the lemma hold and the bounds on the training loss are valid. Note that we need to ensure the conditions of the descent lemma (Lem 3) hold throughout the GD iterates. This is guaranteed by taking

$$\eta \leq \frac{1}{(\tilde{C}_1)^2 + \tilde{C}} \quad (34)$$

where note that $\tilde{C}_1 = G_0 + \beta_L \frac{(3\|w^* - w_0\|)^{3L+1}}{\sqrt{m}}$, and $\tilde{C} = \beta_L \frac{(3\|w^* - w_0\|)^{3L}}{\sqrt{m}}$. As per Lemma 2, this condition ensures that step-size is smaller than the inverse of objective's smoothness parameter throughout all iterates since $\|w_t - w_0\| \leq 3\|w^* - w_0\|$ for all $t \in [T]$.

This completes the proof of Theorem B.1.

B.2 PROOF FOR GENERALIZATION LOSS

Theorem B.2 (Generalization loss of GD in deep nets). *Let the descent lemma hold. Fix a T and assume the target weights vector $w^* \in \mathbb{R}^p$ that separates the data distribution such that*

$$\rho^* \geq \max \left\{ \sqrt{\eta T \widehat{F}(w^*)}, \sqrt{\eta \widehat{F}(w_0)} \right\}. \quad (35)$$

where $\rho^* := \|w^* - w_0\|$. Moreover, assume the width m is large enough such that it satisfies,

$$\sqrt{m} \geq 4\beta_L \bar{\rho}^{3L+2} \quad (36)$$

where $\bar{\rho} = 6\rho^*$. Then the expected generalization error satisfies with high probability,

$$\mathbb{E}_{\mathcal{S}} [F(w_T) - \widehat{F}(w_T)] \leq 9 \frac{(G_0 + 1/4)^2 \rho^{*2}}{n}, \quad (37)$$

where G_0 is the Lipschitz parameter of network at initialization i.e., $\|\nabla \Phi(w_0, \cdot)\| \leq G_0$.

As per the last theorem, we know if m is large enough based on ρ^* , then the hessian during the iterates of GD satisfies $\|\nabla^2 \Phi(w, x)\| \leq C_2$ and thus as per Lemma 2 we have,

1. $\|\nabla \widehat{F}(w)\|_F \leq C_1 \widehat{F}(w)$.
2. $\|\nabla^2 \widehat{F}(w)\|_2 \leq (C_2 + (C_1)^2) \widehat{F}(w)$.
3. $\lambda_{\min}(\nabla^2 \widehat{F}(w)) \geq -C_2 \widehat{F}(w)$.

throughout this section and with a slight abuse of notation, we define $C_1 = G_0 + \beta_L(3\|w - w_0\|)^{3L+1}/\sqrt{m}$, $C_2 = \beta_L(3\|w - w_0\|)^{3L}/\sqrt{m}$ and G_0 is the Lipschitz parameter of the network at initialization i.e., $\|\nabla \Phi(w_0, \cdot)\| \leq G_0$. Recall that in obtaining these quantities we used the guarantees of Thm 2.1 that along the gradient descent path $\max\{\|w_t - w_0\|, \|w_t - w^*\|\} \leq 3\|w^* - w_0\|$

We derive the generalization bound through leave-one-out error estimates. In particular, let w_t^{-i} denote the output of t -th iteration of GD on the leave-one-out loss $F^{-i}(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}$ defined as $F^{-i}(w) := \frac{1}{n} \sum_{j \neq i} \widehat{F}_j(w)$. Then, a simple calculation based on the Lipschitz property of the objective relates the generalization loss to the leave-one-out error as stated in the following lemma.

Lemma 5 (Lem. 6 in Schliserman & Koren (2022), Thm. 2 in Lei & Ying (2020)). *Suppose each sample loss $f(\cdot, z)$ is G -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 ,*

$$\mathbb{E}_{\mathcal{S}} [F(w_T) - \widehat{F}(w_T)] \leq 2G \cdot \mathbb{E}_{\mathcal{S}} \left[\frac{1}{n} \sum_{i=1}^n \|w_T - w_T^{-i}\| \right]. \quad (38)$$

We will use the fact that under the descent lemma $\eta < 1/(C_2 + C_1^2)$, based on (Taheri & Thrampoulidis, 2024, Lem. 9):

$$\left\| (w - \eta \nabla \widehat{F}(w)) - (w' - \eta \nabla \widehat{F}(w')) \right\| \leq \max_{\tilde{w} \in [w, w']} (1 + \eta C_2 \widehat{F}(\tilde{w})) \|w - w'\|. \quad (39)$$

Then, we can bound the leave-one-out error as follows:

$$\begin{aligned} \|w_{t+1} - w_{t+1}^{-i}\| &\leq \left\| (w_t - \eta \nabla \widehat{F}^{-i}(w_t)) - (w_t^{-i} - \eta \nabla \widehat{F}^{-i}(w_t^{-i})) \right\| + \frac{\eta}{n} \left\| \nabla \widehat{F}_i(w_t) \right\| \\ &\leq \left\| (w_t - \eta \nabla \widehat{F}^{-i}(w_t)) - (w_t^{-i} - \eta \nabla \widehat{F}^{-i}(w_t^{-i})) \right\| + \frac{\eta C_1}{n} \widehat{F}_i(w_t) \\ &\leq \left(1 + \eta C_2 \max_{w'_t \in [w_t, w_t^{-i}]} \widehat{F}^{-i}(w'_t) \right) \|w_t - w_t^{-i}\| + \frac{\eta C_1}{n} \widehat{F}_i(w_t), \end{aligned}$$

By the aforementioned property from Proposition 1, we have:

$$\begin{aligned} \max_{w'_t \in [w_t, w_t^{-i}]} \widehat{F}^{\neg i}(w'_t) &\leq \frac{4}{3} \max\{\widehat{F}^{\neg i}(w_t), \widehat{F}^{\neg i}(w_t^{-i})\} \\ &\leq \frac{4}{3} \max\{\widehat{F}(w_t), \widehat{F}^{\neg i}(w_t^{-i})\} \\ &\leq \frac{4}{3} (\widehat{F}(w_t) + \widehat{F}^{\neg i}(w_t^{-i})) \end{aligned}$$

Note that for w'_t we have

$$\|w'_t - w_0\| \leq \alpha \|w_t - w_0\| + (1 - \alpha) \|w_t^{-i} - w_0\| \leq \tilde{\rho} \quad (40)$$

therefore the self-bounded weak convexity holds with $\kappa = C_2$. Moreover,

$$\|w_t - w_t^{-i}\| \leq \|w_t - w_0\| + \|w_t^{-i} - w_0\| \leq 3\rho^* + 3\rho^* = 6\rho^* = 2\tilde{\rho}$$

Denote $\bar{\rho} := 6\rho^* (= 2\tilde{\rho})$. If m is large enough such that $\bar{\rho}^2 < 1/2\bar{C}_2$ where recall $\bar{C}_2 = \frac{\beta_L \bar{\rho}^{3L}}{\sqrt{m}}$, then the GLQC property holds with $\tau = 4/3$ and we have,

$$\begin{aligned} \|w_{t+1} - w_{t+1}^{-i}\| &\leq \left(1 + \frac{4}{3}\eta\bar{C}_2(\widehat{F}(w_t) + \widehat{F}^{\neg i}(w_t^{-i}))\right) + \frac{\eta C_1}{n} \widehat{F}_i(w_t) \\ &\leq \exp\left(\frac{4}{3}\eta\bar{C}_2(\widehat{F}(w_t) + \widehat{F}^{\neg i}(w_t^{-i}))\right) + \frac{\eta C_1}{n} \widehat{F}_i(w_t) \\ &\leq \frac{\eta C_1}{n} \sum_{k=0}^t \exp\left(\sum_{\tau=1}^t \frac{4}{3}\eta\bar{C}_2(\widehat{F}(w_\tau) + \widehat{F}^{\neg i}(w_\tau^{-i}))\right) \widehat{F}_i(w_k) \end{aligned}$$

Fix a T and assume \bar{C}_2 is small enough such that

$$\frac{4}{3}\eta\bar{C}_2 \sum_{\tau=1}^T (\widehat{F}(w_\tau) + \widehat{F}^{\neg i}(w_\tau^{-i})) \leq \frac{16}{3}\bar{C}_2(\eta T \widehat{F}(w^*) + \rho^{*2}) \leq 0.05 \quad (41)$$

then,

$$\begin{aligned} \|w_T - w_T^{-i}\| &\leq \frac{\eta C_1 \exp(0.05)}{n} \sum_{t=0}^T \widehat{F}_i(w_t) \\ &\leq 1.1 \frac{\eta C_1}{n} \sum_{t=0}^T \widehat{F}_i(w_t). \end{aligned}$$

The choice of 0.05 in Eq. 41 is to ensure that the resulting constant in the generalization bound is close to 1. By averaging over $i \in [n]$ and by our training loss guarantees from Theorem 2.1:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|w_T - w_T^{-i}\| &\leq 1.1 \frac{\eta C_1}{n} \sum_{t=0}^T \widehat{F}(w_t) \\ &\leq 2.2 \frac{C_1}{n} (\eta T \widehat{F}(w^*) + \rho^{*2}) \end{aligned}$$

based on Lemma 5 and noting that $G = C_1$, we find that

$$\mathbb{E}_{\mathcal{S}} [F(w_T) - \widehat{F}(w_T)] \leq \frac{4.4}{n} \mathbb{E}_{\mathcal{S}} [C_1^2 \eta T \widehat{F}(w^*) + C_1^2 \rho^{*2}] \quad (42)$$

Recall from Lemma 2 that $C_1 \leq G_0 + \beta_L \|w_t - w_0\|^{3L+1}/\sqrt{m}$ for all $t \in [T]$. As we require $\sqrt{m} \geq 4\beta_L(6\|w^* - w_0\|^{3L+2})$ and from convergence guarantees it holds that $\|w_t - w_0\| \leq 3\|w^* - w_0\|$ we conclude that

$$C_1 \leq G_0 + \frac{1}{2^{3L+2}\|w^* - w_0\|} \leq G_0 + 1/4,$$

where in the last step we assumed without loss of generality that $\|w^* - w_0\| \geq 1$. This completes the proof for the generalization error in Theorem B.2. The proof of Theorem 2.1 is complete.

C PROOF OF COROLLARY 1

Corollary 2 (Restatement of Corollary 1). *Assume the tangent kernel of the model at initialization separates the data with margin γ i.e., for a unit-norm $w \in \mathbb{R}^p$ it holds for all $i \in [n]$: $y_i \langle \nabla \Phi(w_0, x_i), w \rangle \geq \gamma$. Define constant $B > 0$ that bounds the model's output at initialization i.e., $\forall i \in [n] : |\Phi(w_0, x_i)| < B$. Assume the width is large enough such that $m \geq \beta_L^2 (\frac{2B + \log(1/\epsilon)}{\gamma})^{6L+4}$. Then there exists $w^* \in \mathbb{R}^p$ such that $F(w^*) \leq \epsilon$ and w^* lies at a Euclidean distance of order $O(1/\gamma)$ from initialization where $\|w^* - w_0\| \leq \frac{1}{\gamma}(2B + \log(1/\epsilon))$.*

Proof of Corollary 1. By Taylor there exists $w' \in [w^*, w_0]$ such that,

$$y_i \Phi(w^*, x_i) = y_i \Phi(w_0, x_i) + y_i \langle \nabla \Phi(w_0, x_i), w^* - w_0 \rangle + \frac{1}{2} y_i \langle w^* - w_0, \nabla^2 \Phi(w', x_i) (w^* - w_0) \rangle$$

Pick $w^* = w_0 + \frac{w}{\gamma}(2B + \log(1/\epsilon))$. Since $\|w\| = 1$, we derive the desired for $\|w^* - w_0\|$. we have $\|\nabla^2 \Phi(w', x_i)\| \leq O(\frac{\rho^{3L}}{\sqrt{m}})$. Choosing $\rho = \|w^* - w_0\| = \frac{1}{\gamma}(2B + \log(1/\epsilon))$ and noting the given condition on m implies $m \geq \beta_L^2 \rho^{6L+4}$ we obtain $\frac{\beta_L \rho^{3L+2}}{\sqrt{m}} \leq 1$. Therefore with the given assumption on $m \geq \frac{\beta_L^2}{\gamma^{6L+4}}(2B + \log(1/\epsilon))^{6L+4}$ leads to $\frac{1}{2} \|\nabla^2 \Phi(w', x_i)\| \|w^* - w_0\|^2 \leq 1$. Then,

$$\begin{aligned} y_i \Phi(w^*, x_i) &\geq -|y_i \Phi(w_0, x_i)| + y_i \langle \nabla_1 \Phi(w_0, x_i), w^* - w_0 \rangle - \frac{1}{2} \|\nabla_1^2 \Phi(w', x_i)\| \|w^* - w_0\|^2 \\ &\geq -B + 2B + \log(1/\epsilon) - B \\ &= \log(1/\epsilon) \end{aligned}$$

where we assumed without loss of generality that $B \geq 1$. After applying the logistic loss function on $y_i \Phi(w^*, x_i)$, the inequality above gives $\hat{F}(w) \leq \epsilon$. This completes the proof. \square

D CONSISTENCY OF GD IN DEEP NETS: PROOF OF THEOREM 2.3

Theorem D.1 (Consistency). *Assume the width of the network satisfies $m \geq \beta_L^2 n^{3L+3}$ and the step-size satisfies $\eta < 1/L_F$. Then with high probability the expected test loss at iteration $T \geq \sqrt{n}$ is bounded as:*

$$\mathbb{E}_{\mathcal{S}} [F(w_T)] \leq \left(1 + \frac{4}{\sqrt{n}}\right) \left(F(w^*) + \frac{\rho^{*2}}{\eta\sqrt{n}} + \frac{\rho^{*2}}{n\sqrt{n}} + \frac{1}{\sqrt{n}}\right) \quad (43)$$

where w^* is the minimizer of the population loss i.e., $w^* = \arg \min_{w \in \mathbb{R}^p} F(w)$.

Recall that based on 39

$$\left\| (w - \eta \nabla \hat{F}(w)) - (w' - \eta \nabla \hat{F}(w')) \right\| \leq 1 + \eta \max_{\alpha} C_2 \|w - w'\|, \quad (44)$$

where $C_2 := O(\frac{\beta_L \rho^{3L}}{\sqrt{m}})$ and $\rho := \|w - w_0\|$

In particular, note that for GD, we can obtain for $\|w - w_0\|$:

$$\begin{aligned} \|w_T - w_0\| &\leq \eta \sum_{t=0}^{T-1} \|\nabla \hat{F}(w_t)\| \\ &\leq \eta \sum_{t=0}^{T-1} C_1 \\ &= \eta T C_1, \end{aligned} \quad (45)$$

where the first step follows by update rule of GD and in the second step we used the Lipschitz property of objective during GD iterates that ensures $\|\nabla \hat{F}(w_t)\|$ remains bounded by C_1 for all $t \in [T]$. Thus, we can thus show by induction that $\|w_T - w_0\| \leq \eta T C_1$. Then for $m = \Omega(T^2)$ and w_T^{-i}, w_T :

$$\left\| (w_T - \eta \nabla \hat{F}(w_T)) - (w_T^{-i} - \eta \nabla \hat{F}(w_T^{-i})) \right\| \leq (1 + \eta C_2) \|w_T - w_T^{-i}\|, \quad (46)$$

where $C_2 = \beta_L \tilde{\rho}^{3L} / \sqrt{m}$ is the smoothness parameter of the objective throughout the GD updates w_t for all t . Then,

$$\begin{aligned} \|w_{t+1} - w_{t+1}^{-i}\| &\leq \left\| \left(w_t - \eta \nabla \hat{F}^{-i}(w_t) \right) - \left(w_t^{-i} - \eta \nabla \hat{F}^{-i}(w_t^{-i}) \right) \right\| + \frac{\eta}{n} \left\| \nabla \hat{F}_i(w_t) \right\| \\ &\leq \left\| \left(w_t - \eta \nabla \hat{F}^{-i}(w_t) \right) - \left(w_t^{-i} - \eta \nabla \hat{F}^{-i}(w_t^{-i}) \right) \right\| + \frac{\eta C_1}{n} \hat{F}_i(w_t) \\ &\leq (1 + \eta C_2) \|w_t - w_t^{-i}\| + \frac{\eta C_1}{n} \hat{F}_i(w_t), \end{aligned}$$

Recall that in the above $\hat{F}_i(\cdot)$ refers to the objective at the i th data sample and thus has the same Lipschits constant as $\hat{F}(w) = \frac{1}{n} \sum_{i=1}^n \hat{F}_i(w)$.

Noting that $1 + x \leq e^x$ for $x \geq 0$, we can proceed with the above inequality as follows by unrolling the iterates of $\|w_t - w_t^{-i}\|$ and the fact that $w_0 = w_0^{-i}$:

$$\begin{aligned} \|w_{t+1} - w_{t+1}^{-i}\| &\leq \exp(\eta C_2) \|w_t - w_t^{-i}\| + \frac{\eta C_1}{n} \hat{F}_i(w_t) \\ &\leq \frac{\eta C_1}{n} \sum_{k=0}^t \exp(\eta C_2 k) \hat{F}_i(w_k). \end{aligned}$$

Recall that, $C_2 = \beta_L \tilde{\rho}^{3L} / \sqrt{m}$, and $\tilde{\rho} = \max\{\|w_t - w_0\|, \|w_{t-1} - w_0\|, \dots, \|w_1 - w_0\|\}$. Based on Eq. 45, $\|w_t - w_0\| \leq \eta t C_1$ which implies $\tilde{\rho} \leq \eta t C_1$. As a result, for some fixed T :

$$\|w_T - w_T^{-i}\| \leq \frac{\eta C_1}{n} \sum_{t=0}^{T-1} \exp\left(\frac{\eta \beta_L T}{\sqrt{m}} (\eta T C_1)^{3L}\right) \hat{F}_i(w_t).$$

Thus for m larger enough such that $\sqrt{m} \geq 2\beta_L(\eta C_1 T)^{3L+1}$, the exponential term is smaller than 2 and we conclude that,

$$\|w_T - w_T^{-i}\| \leq \frac{2\eta C_1}{n} \sum_{t=0}^{T-1} \hat{F}_i(w_t). \quad (47)$$

Therefore, by averaging over $i \in [n]$ we derive the on-average leave-one-out:

$$\frac{1}{n} \sum_i \|w_T - w_T^{-i}\| \leq \frac{2\eta C_1}{n} \sum_{t=0}^{T-1} \hat{F}(w_t) \quad (48)$$

The only remaining part is characterising the right hand side of the above inequality. This is done as follows.

PROOF OF OPTIMIZATION ERROR

Define $\rho(t) := \|w_t - w_0\|$ and $\rho^* = \|w^* - w_0\|$. if we define $\tilde{\rho} > \max\{\rho(t), \rho^*\}$ for all $t \in [T]$ then for any $w \in [w_t, w^*]$ it holds that $\|w - w_0\| \leq \tilde{\rho}$ and thus based on Lemma 2

$$\lambda_{\min}(\nabla^2 \hat{F}(w)) \geq -\tilde{C} \hat{F}(w), \quad (49)$$

where $\tilde{C} := \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}}$. Therefore by Eq. 26,

$$\begin{aligned} \hat{F}(w_T) &\leq \frac{1}{T} \sum_{t=1}^T \hat{F}(w_t) \leq \hat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \tilde{C} \frac{1}{2T} \sum_{t=0}^{T-1} \max_{w'_t \in [w_t, w^*]} \hat{F}(w'_t) \|w^* - w_t\|^2, \\ &\leq \hat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \tilde{C} \frac{1}{2T} \sum_{t=0}^{T-1} \|w^* - w_t\|^2 \\ &\leq \hat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \tilde{C} \|w^* - w_0\|^2 + \tilde{C} \frac{1}{T} \sum_{t=0}^{T-1} \|w_t - w_0\|^2. \end{aligned}$$

For the last term, we have a simple bound:

$$\|w_t - w_0\| = \eta \left\| \sum_{k=0}^{t-1} \nabla \hat{F}(w_k) \right\| \leq \eta \sum_{k=0}^{t-1} \|\nabla \hat{F}(w_k)\| \leq \eta t C_1$$

where C_1 denote the lipschitz parameter bound along the GD iterates and the last inequality follows by descent lemma condition.

Then,

$$\hat{F}(w_T) \leq \hat{F}(w^*) + \frac{\|w^* - w_0\|^2}{\eta T} + \tilde{C} \|w^* - w_0\|^2 + \eta^2 \tilde{C} C_1^2 T^2.$$

Pick large enough width such that $m \geq \beta_L^2 n^{3L+3} (\eta C_1)^{6L}$ and choose $T = \sqrt{n}$. Then $\tilde{\rho} < (\eta C_1 n)^{3L/2}$ and

$$\begin{aligned} \hat{F}(w_T) &\leq \hat{F}(w^*) + \frac{\rho^{*2}}{\eta \sqrt{n}} + \frac{\beta_L \tilde{\rho}^{3L}}{\sqrt{m}} (\rho^{*2} + \eta^2 C_1^2 n) \\ &\leq \hat{F}(w^*) + \frac{\rho^{*2}}{\eta \sqrt{n}} + \frac{\rho^{*2} + \eta^2 C_1^2 n}{n \sqrt{n}} \end{aligned}$$

In particular, in the above one can choose $w^* := \arg \min_{w \in \mathbb{R}^p} F(w)$. Moreover, from the generalization guarantees,

$$\begin{aligned} \mathbb{E}_{\mathcal{S}} [F(w_T)] &\leq \mathbb{E}_{\mathcal{S}} [\hat{F}(w_T)] + \frac{4\eta C_1^2}{n} \sum_{t=1}^{T-1} \mathbb{E}_{\mathcal{S}} [\hat{F}(w_t)] \\ &\leq \left(1 + \frac{4\eta C_1^2}{\sqrt{n}}\right) \left(\mathbb{E}_{\mathcal{S}} [\hat{F}(w^*)] + \frac{\rho^{*2}}{\eta \sqrt{n}} + \frac{\rho^{*2} + \eta^2 C_1^2 n}{n \sqrt{n}}\right) \\ &= \left(1 + \frac{4\eta C_1^2}{\sqrt{n}}\right) \left(F(w^*) + \frac{\rho^{*2}}{\eta \sqrt{n}} + \frac{\rho^{*2} + \eta^2 C_1^2 n}{n \sqrt{n}}\right). \end{aligned}$$

Note that $\eta \leq \frac{1}{C_1^2 + C_2}$ by the descent lemma condition (Lemma 3). Therefore $\eta \leq 1/C_1^2$. Without loss of generality we can assume $C_1 \geq 1$. This simplifies the above inequality as follows:

$$\begin{aligned} \mathbb{E}_{\mathcal{S}} [F(w_T)] &\leq \left(1 + \frac{4}{\sqrt{n}}\right) \left(F(w^*) + \frac{\rho^{*2}}{\eta \sqrt{n}} + \frac{\rho^{*2}}{n \sqrt{n}} + \frac{1}{\sqrt{n}}\right) \\ &= F(w^*) + O\left(\frac{F(w^*) + \rho^{*2}}{\sqrt{n}}\right). \end{aligned}$$

This completes the proof of Theorem 2.3.

E PROOF OF THEOREM 2.4

We recall our setup and notation for Theorem 2.4. We consider in this section the one-hidden layer network with quadratic activation where $x \in \mathbb{R}^d$, $w_i \in \mathbb{R}^d$:

$$\Phi(w, x) = \frac{1}{2m} \sum_{i=1}^m a_i (x^\top w_i)^2$$

in the above, $a_i = \pm 1$ are fixed and half of the neurons have $a_i = 1$. For initialization of first layer weights $w_i \stackrel{\text{iid}}{\sim} N(0, \frac{I_d}{d})$ that is each $w_{ij} \stackrel{\text{iid}}{\sim} N(0, \frac{1}{d})$. The data points x are uniformly drawn from the Rademacher distribution (of size 2^d) that is $x(k) = \pm 1$ i.i.d and w.p. $1/2$ and the labels $y = x(1) \cdot x(2)$. Consider the linear loss where $f(u) = -u$ and $\hat{F}(w) := \frac{1}{n} \sum_{j=1}^n f(y_j \Phi(x_j, w))$.

Consider stochastic gradient descent with the update rule as follows for each w_i :

$$\begin{aligned} w_i^{t+1} &= w_i^t - \eta \nabla \widehat{F}_i(w^t) \\ &= w_i^t - \frac{\eta a_i}{nm} \sum_{j=1}^n (x_j^\top w_i^t) x_j y_j f'(y_j \Phi(x_j, w)) \\ &= w_i^t + \frac{\eta a_i}{nm} \sum_{j=1}^n (x_j^\top w_i^t) x_j y_j \end{aligned}$$

Note that in the above the superscripts denote the iteration number and the subscript i indicates the vector $w_i \in \mathbb{R}^d$ is associated with neuron i . Fixing the initialization, define $\mathbb{E}_x[w_i^t]$ as the outcome of GD after t iterations on *population gradient*. Formally, $\mathbb{E}_x[w_i^t]$ is defined recursively as follows:

$$\mathbb{E}_x[w_i^{t+1}] := \mathbb{E}_x[w_i^t] - \eta \nabla F_i(\mathbb{E}_x[w^t]),$$

where $F(w) := \mathbb{E}_x[\widehat{F}(w)]$ is the test loss and define $\mathbb{E}_x[w^0] := w^0$. Based on previous derivations the update rule for the expected weights $\mathbb{E}_x[w_i^t]$ obeys the following:

$$\mathbb{E}_x[w_i^{t+1}] = \mathbb{E}_x[w_i^t] + \frac{\eta}{m} a_i \mathbb{E}_x[\langle x, \mathbb{E}_x[w_i^t] \rangle xy].$$

for $t = 0$:

$$\mathbb{E}_x[w_i^1] = w_i^0 + \frac{\eta}{m} a_i \mathbb{E}_x[\langle x, w_i^0 \rangle xy].$$

For $t = 1$:

$$\mathbb{E}_x[w_i^2] = \mathbb{E}_x[w_i^1] + \frac{\eta}{m} a_i \mathbb{E}_x[\langle x, \mathbb{E}_x[w_i^1] \rangle xy].$$

and similarly $\mathbb{E}_x[w_i^t]$ is defined for all iterations t .

E.1 CALCULATING $\mathbb{E}_x[w_i^1]$

For the expected gradient $g_i \in \mathbb{R}^d$ we have by noting $y = x(1)x(2)$:

$$g_i = -\mathbb{E}_x[\langle x, w_i^0 \rangle \cdot x \cdot x(1)x(2)]$$

for g_i^3 (we drop i and 0 for simplicity and denote the k 'th element of W_i^0 with W^k):

$$\begin{aligned} -g_i^3 &= \mathbb{E}_x[(x(3)w^3 + \sum_{k \neq 3} x(k)w^k) \cdot x(3) \cdot x(1)x(2)] \\ &= \mathbb{E}_x[w^3 \cdot x(1)x(2) + x(3)x(1)x(2) \sum_{k \neq 3} x(k)w^k] \\ &= \mathbb{E}_x[w^3 \cdot x(1)x(2)] + \mathbb{E}_x[x(3)x(1)x(2) \sum_{k \neq 3} x(k)w^k] \\ &= 0 + \mathbb{E}_x[x(3) \sum_{k \neq 3} x(k)w^k] \mathbb{E}_x[x(1)x(2)] \\ &= 0. \end{aligned}$$

Therefore for any $k \neq 1, 2$ we have $g_i^k = 0$.

for $k = 1$:

$$\begin{aligned}
-g_i^1 &= \mathbb{E}_x[\langle x, w \rangle \cdot x(1) \cdot x(1)x(2)] \\
&= \mathbb{E}_x[w^1 x(1)x(2) + x(1)x(2)x(1) \sum_{k \neq 1} w^k x(k)] \\
&= \mathbb{E}_x[w^1 x(1)x(2)] + \mathbb{E}_x[x(1)x(2)x(1)w^2 x(2)] + \mathbb{E}_x[x(1)x(2)x(1) \sum_{k \neq 1,2} w^k x(k)] \\
&= w^1 \mathbb{E}_x[x(1)x(2)] + w^2 \mathbb{E}_x[x(1)x(2)x(1)x(2)] + \mathbb{E}_x[x(1)x(2)x(1)] \cdot \mathbb{E}_x[\sum_{k \neq 1,2} w^k x(k)] \\
&= w^2 \mathbb{E}_x[(x(1)x(2))^2] \\
&= w^2.
\end{aligned}$$

Similarly, we derive that

$$-g_i^2 = w_i^0(1)$$

Thus, the expected weight after one step is

$$\begin{aligned}
\mathbb{E}_x[w_i^1] &= w_i^0 + \frac{\eta a_i}{m} [w_i^0(2), w_i^0(1), 0, \dots, 0] \\
&= \left[w_i^0(1) + \frac{\eta a_i}{m} w_i^0(2), w_i^0(2) + \frac{\eta a_i}{m} w_i^0(1), w_i^0(3), \dots, w_i^0(d) \right]
\end{aligned}$$

E.2 CALCULATING $\mathbb{E}_x[w_i^2]$

For $k = 1$:

$$\begin{aligned}
\mathbb{E}_x[w_i^2(1)] &= \mathbb{E}_x[w_i^1(1)] + \frac{\eta a_i}{m} \mathbb{E}_x[\langle x, \mathbb{E}_x[w_i^1] \rangle x(2)] \\
&= \mathbb{E}_x[w_i^1(1)] + \frac{\eta a_i}{m} \mathbb{E}_x[w_i^1(2)] \\
&= w_i^0(1) + \frac{\eta a_i}{m} w_i^0(2) + \frac{\eta a_i}{m} (w_i^0(2) + \frac{\eta a_i}{m} w_i^0(1))
\end{aligned}$$

Similarly we obtain for $k = 2$:

$$\begin{aligned}
\mathbb{E}_x[w_i^2(2)] &= \mathbb{E}_x[w_i^1(2)] + \frac{\eta a_i}{m} \mathbb{E}_x[\langle x, \mathbb{E}_x[w_i^1] \rangle x(1)] \\
&= \mathbb{E}_x[w_i^1(2)] + \frac{\eta a_i}{m} \mathbb{E}_x[w_i^1(1)] \\
&= w_i^0(2) + \frac{\eta a_i}{m} w_i^0(1) + \frac{\eta a_i}{m} (w_i^0(1) + \frac{\eta a_i}{m} w_i^0(2))
\end{aligned}$$

for $k \neq 1, 2$:

$$\begin{aligned}
\mathbb{E}_x[w_i^2(k)] &= \mathbb{E}_x[w_i^1(k)] + \frac{\eta a_i}{m} \mathbb{E}_x[\langle x, \mathbb{E}_x[w_i^1] \rangle x(k)x(1)x(2)] \\
&= \mathbb{E}_x[w_i^1(k)]
\end{aligned}$$

E.3 CALCULATING $\mathbb{E}_x[w_i^t]$

In general, we observe the following pattern for arbitrary t : for $k \neq 1, 2$:

$$\mathbb{E}_x[w_i^{t+1}(k)] = \mathbb{E}_x[w_i^t(k)]$$

and for $k = 1, 2$:

$$\begin{aligned}
\mathbb{E}_x[w_i^{t+1}(1)] &= \mathbb{E}_x[w_i^t(1)] + \frac{\eta a_i}{m} \mathbb{E}_x[w_i^t(2)] \\
\mathbb{E}_x[w_i^{t+1}(2)] &= \mathbb{E}_x[w_i^t(2)] + \frac{\eta a_i}{m} \mathbb{E}_x[w_i^t(1)]
\end{aligned}$$

Therefore for general t , defining $\gamma = \eta a_i / m$, $\alpha(t) := \mathbb{E}_x[w_i^t(1)]$ and $\beta(t) := \mathbb{E}_x[w_i^t(2)]$ we have the following equations:

$$\begin{aligned}\alpha(t+1) &= \alpha(t) + \gamma\beta(t) \\ \beta(t+1) &= \beta(t) + \gamma\alpha(t).\end{aligned}$$

It can be verified that the solution to the above equations take the following form:

$$\begin{aligned}\alpha(t) &= \left(\sum_{r \in [t], r: \text{even}} \binom{t}{r} \gamma^r \right) \alpha(0) + \left(\sum_{r \in [t], r: \text{odd}} \binom{t}{r} \gamma^r \right) \beta(0) \\ \beta(t) &= \left(\sum_{r \in [t], r: \text{even}} \binom{t}{r} \gamma^r \right) \beta(0) + \left(\sum_{r \in [t], r: \text{odd}} \binom{t}{r} \gamma^r \right) \alpha(0)\end{aligned}$$

Therefore with replacement:

$$\begin{aligned}\mathbb{E}_x[w_i^t(1)] &= \left(\sum_{r \in [t], r: \text{even}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(1) + \left(\sum_{r \in [t], r: \text{odd}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(2) \\ \mathbb{E}_x[w_i^t(2)] &= \left(\sum_{r \in [t], r: \text{even}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(2) + \left(\sum_{r \in [t], r: \text{odd}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(1)\end{aligned}$$

E.4 EMPIRICAL GRADIENT FOR MULTIPLE GD STEPS

The results above hold when GD is applied on the entire data distribution. Let us consider the case of SGD on n data points (on the loss $F(\cdot)$) for each iteration and bound the resulting noise.

Recall that

$$\begin{aligned}w^1 &= w^0 - \eta \nabla F(w^0), \\ \mathbb{E}_x[w^1] &= w^0 - \eta \mathbb{E}_x[\nabla F(w^0)], \\ w^1 &= \mathbb{E}_x[w^1] - \eta(\nabla F(w^0) - \mathbb{E}_x[\nabla F(w^0)]).\end{aligned}$$

Moreover, by how we defined $\mathbb{E}_x[w^t]$ for $t \geq 2$, we have

$$\begin{aligned}w^{t+1} &= w^t - \eta \nabla F(w^t), \\ \mathbb{E}_x[w^{t+1}] &= \mathbb{E}_x[w^t] - \eta \mathbb{E}_x[\nabla F(\mathbb{E}_x[w^t])], \\ w^{t+1} &= \mathbb{E}_x[w^t] + (w^t - \mathbb{E}_x[w^t]) - \eta(\nabla F(w^t) - \mathbb{E}_x[\nabla F(\mathbb{E}_x[w^t])]) - \eta \mathbb{E}_x[\nabla F(\mathbb{E}_x[w^t])].\end{aligned}$$

Overall, from the last two equations we have:

$$\begin{aligned}w^{t+1} - \mathbb{E}_x[w^{t+1}] &= -\eta \sum_{\tau=0}^t \nabla F(w^\tau) - \mathbb{E}_x[\nabla F(\mathbb{E}_x[w^\tau])] \\ &= -\eta \sum_{\tau=0}^t \underbrace{\nabla F(w^\tau) - \nabla F(\mathbb{E}_x[w^\tau])}_{\text{Term I}} + \underbrace{\nabla F(\mathbb{E}_x[w^\tau]) - \mathbb{E}_x[\nabla F(\mathbb{E}_x[w^\tau])]}_{\text{Term II}}.\end{aligned}$$

We bound each term separately.

E.5 TERM II:

Starting with the second term, recall that:

$$-\nabla F_i(\mathbb{E}_x[w^t]) = \frac{a_i}{nm} \sum_{j=1}^n \langle x_j, \mathbb{E}_x[w_i^t] \rangle x_j y_j$$

Case 1: For $k \neq 1, 2$:

$$\begin{aligned} -\nabla F_i^k(\mathbb{E}_x[w^t]) &= \frac{a_i}{nm} \sum_{j=1}^n \langle x_j, \mathbb{E}_x[w_i^t] \rangle x_j(k) x_j(1) x_j(2) \\ &= \frac{a_i}{nm} \left\langle \sum_{j=1}^n x_j \cdot x_j(k) \cdot x_j(1) \cdot x_j(2), \mathbb{E}_x[w_i^t] \right\rangle \end{aligned}$$

Denoting $v_j := x_j \cdot x_j(k) \cdot x_j(1) \cdot x_j(2)$ note that $v_j(r)$ are mutually independent for all $r \in d$ and $j \in [n]$. Moreover each $v_j(r)$ is distributed according to a Rademacher distribution and $\text{Var}(v_j(r)) = 1$.

Therefore, recalling the expression for $\mathbb{E}_x[w_i^t] = [\alpha_2^t w_i^0(1) + \alpha_1^t w_i^0(2), \alpha_1^t w_i^0(1) + \alpha_2^t w_i^0(2), w_i^0(3), \dots, w_i^0(d)]$ where $\alpha_2^t := \sum_{r \in [t], r: \text{even}} \binom{t}{r} (\frac{\eta a_i}{m})^r$ and $\alpha_1^t := \sum_{r \in [t], r: \text{odd}} \binom{t}{r} (\frac{\eta a_i}{m})^r$ by expanding the summation above:

$$\begin{aligned} -\nabla F_i^k(\mathbb{E}_x[w^t]) &= \frac{a_i}{nm} \left((\alpha_2^t \sum_j v_j(1) + \alpha_1^t \sum_j v_j(2)) w_i^0(1) \right. \\ &\quad \left. + (\alpha_1^t \sum_j v_j(1) + \alpha_2^t \sum_j v_j(2)) w_i^0(2) + \sum_{k \neq 1, 2} \sum_j v_j(k) w_i^0(k) \right). \end{aligned}$$

We use Hoeffding concentration inequality for bounding $\sum_j v_j(k) = O(\sqrt{n} \log(d))$ uniformly for every $k \neq 1, 2$. To be more concrete, by a union bound and Hoeffding it follows that uniformly over all $k \leq d$ and $t \leq T$:

$$\Pr \left(\sum_j v_j(k) \leq \sqrt{n} \log(dT) \right) \geq 1 - \exp(\log(dT) - 2 \log^2(dT)) =: 1 - p_x.$$

Moreover, recall that $w_i^0(k) \sim N(0, 1/d)$ and :

$$\Pr \left(w_i^0(k) \leq \frac{\log(dm)}{\sqrt{d}} \right) \geq 1 - \frac{\exp(-\frac{1}{2} \log^2(dm))}{\sqrt{2\pi} \log(dm)}$$

By union bound we have for all $i \leq m, k \leq d$:

$$\Pr \left(w_i^0(k) \leq \frac{\log(dm)}{\sqrt{d}} \right) \geq 1 - \frac{\exp(-\frac{1}{2} \log^2(dm) + \log(dm))}{\sqrt{2\pi} \log(dm)} =: 1 - p_w.$$

Therefore w.p. $1 - p_x - p_w$ we have for all $t \leq T, i \leq m, k \leq d$: $\sum_j v_j(k) w_i^0(k) \leq \frac{\log(dm) \log(dT) \sqrt{n}}{\sqrt{d}}$

By Hoeffding w.p. $1 - p_x - p_w$ over data sampling and initialization:

$$\Pr \left(\sum_k \sum_j v_j(k) w_i^0(k) \leq \log(dm) \log(dT) \log(d) \sqrt{n} \right) \geq 1 - \exp(-\log^2(d)) =: 1 - p_1.$$

Similarly, we obtain $\Pr \left(\sum_j v_j(1) w_i^0(1) \leq \frac{\sqrt{n}}{\sqrt{d}} \log(dT) \log(dm) \right) \geq 1 - p_x - p_w$ Overall, the above calculations show that for some constant C with probability $1 - C(p_1 - p_x - p_w)$ over initialization and data sampling it holds uniformly over all k, i, t :

$$\begin{aligned} |\nabla F_i^k(\mathbb{E}_x[w^t])| &\lesssim \frac{\log(dT) \log(dm)}{nm} \left(\frac{2\sqrt{n}(|\alpha_1^t| + |\alpha_2^t|)}{\sqrt{d}} + \sqrt{n} \log(d) \right) \\ &= \log(dT) \log(dm) \frac{|\alpha_1^t| + |\alpha_2^t| + \sqrt{d} \log(d)}{m\sqrt{nd}}. \end{aligned}$$

Case 2: Now assume $k = 1$ then in a similar way as above

$$\begin{aligned}
-\nabla F_i^1(\mathbb{E}_x[w^t]) &= \frac{a_i}{nm} \sum_{j=1}^n \langle x_j, \mathbb{E}_x[w_i^t] \rangle x_j(2) \\
&= \frac{a_i}{m} \mathbb{E}_x[w_i^t(2)] + \frac{a_i}{nm} \sum_{j=1}^n \sum_{j' \neq 2}^d \mathbb{E}_x[w_i^t(j')] x_j(j') x_j(2) \\
&= \frac{a_i}{m} \mathbb{E}_x[w_i^t(2)] + O(\log(dm) \log(dT) \frac{|\alpha_1^t| + \alpha_2^t + \sqrt{d} \log(d)}{m\sqrt{nd}}).
\end{aligned}$$

for $k = 2$:

$$-\nabla F_i^2(\mathbb{E}_x[w^t]) = \frac{a_i}{m} \mathbb{E}_x[w_i^t(1)] + O(\log(dm) \log(dT) \frac{|\alpha_1^t| + \alpha_2^t + \sqrt{d} \log(d)}{m\sqrt{nd}}). \quad (50)$$

Therefore for the weights entering the i 'th hidden neuron and for all $i \leq m, t \leq T$, the following concentration bound holds w.p. $1 - C(p_1 - p_x - p_w)$:

$$\|\nabla F_i(\mathbb{E}_x[w^t]) - \mathbb{E}_x[\nabla F_i(\mathbb{E}_x[w^t])]\| \lesssim \log(dm) \log(dT) \frac{|\alpha_1^t| + \alpha_2^t + \sqrt{d} \log(d)}{m\sqrt{n}}.$$

Therefore, as a result of Term II calculations, we deduce that with probability $1 - C(p_1 - p_x - p_w)$,

$$\forall i, t : \|\nabla F_i(\mathbb{E}_x[w^t]) - \mathbb{E}_x[\nabla F_i(\mathbb{E}_x[w^t])]\| \lesssim \log(dm) \log(dT) \frac{|\alpha_1^t| + \alpha_2^t + \sqrt{d} \log(d)}{m\sqrt{n}},$$

where x^t denotes the data chosen at iteration t of SGD and recall w^0 denotes the initialization weights.

E.6 TERM I:

Case 1: We first consider $k \neq 1, 2$. By applying Hoeffding's concentration inequality we obtain $\sum_{j=1}^n x_j(\ell) x_j(k) x_j(1) x_j(2) < \sqrt{n} \log(dT)$ uniformly for every $\ell \in [d], t \leq T$ w.p. $1 - \exp(\log(dT) - 2 \log^2(dT)) =: 1 - p_x$. Applying Hoeffding again yields,

$$\begin{aligned}
\nabla F_i^k(w^t) - \nabla F_i^k(\mathbb{E}_x[w^t]) &= \frac{a_i}{nm} \sum_{j=1}^n \langle x_j, \mathbb{E}_x[w_i^t] - w_i^t \rangle x_j(k) x_j(1) x_j(2) \\
&= \frac{a_i}{nm} \left\langle \sum_{j=1}^n x_j x_j(k) x_j(1) x_j(2), \mathbb{E}_x[w_i^t] - w_i^t \right\rangle \\
&\leq \frac{\|\mathbb{E}_x[w_i^t] - w_i^t\|}{\sqrt{nm}} \log(dT) \log(dmT),
\end{aligned}$$

w.p. $1 - \exp(\log(dmT) - 2 \log^2(dmT)) - \exp(\log(dT) - 2 \log^2(dT)) =: 1 - p_2 - p_x$ uniformly over all $k \leq d, i \leq m, t \leq T$ where in the above we assumed $\mathbb{E}_x[w_i^t] - w_i^t$ as a fixed vector.

Denote $v^t := \mathbb{E}_x[w_i^t] - w_i^t$. Concretely, we have the following bound for $k \neq 1, 2$ on Term I:

$$\Pr_{x^t} \left(\forall i, k, t : \nabla F_i^k(w^t) - \nabla F_i^k(\mathbb{E}_x[w^t]) \leq \frac{\|v^t\|}{\sqrt{nm}} \log(dT) \log(dmT) \middle| v^t \right) \geq 1 - p_2 - p_x.$$

Therefore,

$$\Pr_{x^t, v^t} \left(\forall i, k, t : \nabla F_i^k(w^t) - \nabla F_i^k(\mathbb{E}_x[w^t]) \leq \frac{\|v^t\|}{\sqrt{nm}} \log(dT) \log(dmT) \right) \geq 1 - p_2 - p_x.$$

Case 2: For $k = 1$:

$$\begin{aligned}\nabla F_i^k(w^t) - \nabla F_i^k(\mathbb{E}_x[w^t]) &= \frac{a_i}{nm} \sum_{j=1}^n \langle x_j, \mathbb{E}_x[w_i^t] - w_i^t \rangle x_j(2) \\ &= \frac{a_i}{nm} \langle \sum_{j=1}^n x_j x_j(2), \mathbb{E}_x[w_i^t] - w_i^t \rangle \\ &= \frac{a_i}{m} v^t(2) + \frac{a_i}{m\sqrt{n}} \|v^t(\ell \neq 2)\| \cdot \log(dT) \log(mT),\end{aligned}$$

w.p. $1 - p_2 - p_x$.

Similarly for $k = 2$ with the same probability it holds,

$$\begin{aligned}\nabla F_i^k(w^t) - \nabla F_i^k(\mathbb{E}_x[w^t]) &= \frac{a_i}{nm} \sum_{j=1}^n \langle x_j, \mathbb{E}_x[w_i^t] - w_i^t \rangle x_j(1) \\ &= \frac{a_i}{nm} \langle \sum_{j=1}^n x_j x_j(1), \mathbb{E}_x[w_i^t] - w_i^t \rangle \\ &= \frac{a_i}{m} v^t(1) + \frac{a_i}{m\sqrt{n}} \|v^t(\ell \neq 1)\| \cdot \log(dT) \log(mT)\end{aligned}$$

E.7 COMBINING TWO TERMS

Recall we defined $v^t = \mathbb{E}_x[w_i^t] - w_i^t$ where $v^t \in \mathbb{R}^d$. Denote its entries by $v^t(k)$ for $k \leq d$. Define $v^0 = 0$. Then, from the last two sections we obtain the following system of equations which hold w.p. $1 - C(p_1 - p_x - p_2 - p_w)$ for all $t \geq 1$:

$$\begin{aligned}|v^t(1)| &\leq \eta \sum_{\tau=0}^{t-1} \frac{|v^\tau(2)|}{m} + \frac{\|v^\tau(\ell \neq 2)\| \log(dT) \log(mT)}{m\sqrt{n}} + \frac{|\alpha_1^\tau| + \alpha_2^\tau + \sqrt{d} \log(d)}{m\sqrt{nd}} \log(dm) \log(dT) \\ |v^t(2)| &\leq \eta \sum_{\tau=0}^{t-1} \frac{|v^\tau(1)|}{m} + \frac{\|v^\tau(\ell \neq 1)\| \log(dT) \log(mT)}{m\sqrt{n}} + \frac{|\alpha_1^\tau| + \alpha_2^\tau + \sqrt{d} \log(d)}{m\sqrt{nd}} \log(dm) \log(dT) \\ |v^t(k)| &\leq \eta \sum_{\tau=0}^{t-1} \frac{\|v^\tau\| \log(dT) \log(dmT)}{m\sqrt{n}} + \frac{|\alpha_1^\tau| + \alpha_2^\tau + \sqrt{d} \log(d)}{m\sqrt{nd}} \log(dm) \log(dT), \quad \forall k \neq 1, 2\end{aligned}$$

Note that $|\alpha_1^\tau| + \alpha_2^\tau = 2^\tau$. Simplify the above by assuming $\eta = m$ and denoting $\gamma_\tau := \frac{2^\tau + \sqrt{d} \log(d)}{\sqrt{nd}} \log(dm) \log(dT)$. Then the equations above simplify to:

$$\begin{aligned}|v^t(1)| &\leq \sum_{\tau=0}^{t-1} |v^\tau(2)| + \frac{\|v^\tau\| \log^2(dmT)}{\sqrt{n}} + \sum_{\tau=0}^{t-1} \gamma_\tau \\ |v^t(2)| &\leq \sum_{\tau=0}^{t-1} |v^\tau(1)| + \frac{\|v^\tau\| \log^2(dmT)}{\sqrt{n}} + \sum_{\tau=0}^{t-1} \gamma_\tau \\ |v^t(k)| &\leq \sum_{\tau=0}^{t-1} \frac{\|v^\tau\| \log^2(dmT)}{\sqrt{n}} + \sum_{\tau=0}^{t-1} \gamma_\tau, \quad \forall k \neq 1, 2\end{aligned}$$

Define $z^t := \max_k |v^t(k)|$. Then since the RHS in the third equations is smaller than the two first equations we deduce that:

$$z^t \leq \sum_{\tau=0}^{t-1} z^\tau + \frac{z^\tau \sqrt{d} \log^2(dmT)}{\sqrt{n}} + \frac{2^\tau + \sqrt{d} \log(d)}{\sqrt{nd}} \log^2(dmT) \quad (51)$$

To proceed in simplifying the equation above, we need the following lemma which can be straightforwardly proved by induction.

Lemma 6. Let $v_k, \beta, \gamma_k \in \mathbb{R}$. If $v_0 = 0$ and for every $t < T$:

$$v_{t+1} \leq \beta \sum_{k=0}^t v_k + \sum_{k=0}^t \gamma_k \quad (52)$$

Then, it can be checked that this results in:

$$v_T \leq \sum_{k=0}^{T-1} (\beta + 1)^k \gamma_{T-k-1}. \quad (53)$$

Therefore, continuing from Eq. 51:

$$\begin{aligned} z^t &\leq \sum_{\tau=0}^{t-1} \left(2 + \frac{\sqrt{d} \log^2(dmT)}{\sqrt{n}}\right)^\tau \cdot \frac{2^{t-\tau-1} + \sqrt{d} \log(d)}{\sqrt{nd}} \log^2(dmT) \\ &\leq \frac{t \log^2(dmT)}{\sqrt{nd}} \left(2 + \frac{\sqrt{d} \log^2(dmT)}{\sqrt{n}}\right)^{t-1} + \frac{(2 + \frac{\sqrt{d} \log^2(dmT)}{\sqrt{n}})^t \log^3(dmT)}{\sqrt{n}(1 + \frac{\sqrt{d} \log^2(dmT)}{\sqrt{n}})} \end{aligned}$$

Assume $t \leq \log(d)$ and $n \geq d \cdot \log^{2c}(d)$ for some constant c . Then,

$$\begin{aligned} z^t &\leq \frac{d}{d \log^{c-3}(d)} + \frac{2d}{\sqrt{d} \log^{c-3}(d)} \\ &\leq \frac{3\sqrt{d}}{\log^{c-3}(d)} \end{aligned}$$

where in the above we used the fact that for $t \leq T \leq \log(d)$ and any $m = O(\text{poly}(d))$:

$$\left(2 + \frac{\sqrt{d} \log^2(dmT)}{\sqrt{n}}\right)^{t-1} = \left(2 + \frac{1}{\log^{c-2}(d)}\right)^{t-1} \leq \left(2 + \frac{1}{\log(d)}\right)^{t-1} \leq d$$

Therefore

$$\max_k |v^t(k)| = O\left(\frac{\sqrt{d}}{\log^{c-3}(d)}\right)$$

We next show that $\|v^t(3 : d)\|$ is much smaller. To see this, we replace our derived bound in the third equation above. define $\beta^t := \|v^t(3 : d)\|$ and note that $\beta^t \leq \sqrt{d} |v^t(k)|$ for some $k \in [3, d]$. Thus:

$$\begin{aligned} \beta^t &\leq \sum_{\tau=0}^{t-1} \frac{\|v^\tau\| \log^2(dmT) \sqrt{d}}{\sqrt{n}} + \sqrt{d} \sum_{\tau=0}^{t-1} \gamma_\tau \\ &\leq \sum_{\tau=0}^{t-1} \frac{\|v^\tau\|}{\log^{c-2}(d)} + \sqrt{d} \sum_{\tau=0}^{t-1} \gamma_\tau \\ &\leq \sum_{\tau=0}^{t-1} \frac{\sqrt{(\beta^\tau)^2 + \frac{d}{\log^{2c-6}(d)}}}{\log^{c-2}(d)} + \sqrt{d} \sum_{\tau=0}^{t-1} \gamma_\tau \\ &\leq \sum_{\tau=0}^{t-1} \frac{\beta^\tau + \frac{\sqrt{d}}{\log^{c-3}(d)}}{\log^{c-2}(d)} + \sum_{\tau=0}^{t-1} \frac{2^\tau + \sqrt{d} \log(d)}{\sqrt{n}} \log^2(dmT) \end{aligned}$$

Using the lemma for the recursive summation (Lemma 6) again:

$$\begin{aligned} \beta^t &\leq \sum_{\tau=0}^{t-1} \left(1 + \frac{1}{\log^{c-2}(d)}\right)^\tau \left(\frac{\sqrt{d}}{\log^{2c-5}(d)} + \frac{2^{t-\tau-1} + \sqrt{d} \log(d)}{\sqrt{n}} \log^2(dmT)\right) \\ &\leq \left(1 + \frac{1}{\log^{c-2}(d)}\right)^t \frac{\sqrt{d}}{\log^{c-3}(d)} + \frac{2^{t-1}t}{\sqrt{d} \log^{c-2}(d)} + \left(1 + \frac{1}{\log^{c-2}(d)}\right)^t \log^2(dmT) \end{aligned}$$

where in the above we used the conditions on m, T . By using them again and recalling $t \leq T \leq \log(d)$, the above simplifies into:

$$\begin{aligned}\beta^t &\leq \frac{2\sqrt{d}}{\log^{c-3}(d)} + \frac{\sqrt{d}}{\log^{c-3}(d)} + 2\log^2(dmT) \\ &= O\left(\frac{\sqrt{d}}{\log^{c-3}(d)}\right)\end{aligned}$$

recall that $\mathbb{E}_x[w_i^t]$ had the following form

$$\begin{aligned}\mathbb{E}_x[w_i^t(1)] &= \left(\sum_{r \in [t], r: \text{even}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(1) + \left(\sum_{r \in [t], r: \text{odd}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(2) \\ &= \frac{1}{2}((1+a_i)^t + (1-a_i)^t)w_i^0(1) + \frac{1}{2}((1+a_i)^t - (1-a_i)^t)w_i^0(2) \\ \mathbb{E}_x[w_i^t(2)] &= \left(\sum_{r \in [t], r: \text{even}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(2) + \left(\sum_{r \in [t], r: \text{odd}} \binom{t}{r} \left(\frac{\eta a_i}{m}\right)^r \right) w_i^0(1) \\ &= \frac{1}{2}((1+a_i)^t + (1-a_i)^t)w_i^0(2) + \frac{1}{2}((1+a_i)^t - (1-a_i)^t)w_i^0(1) \\ \mathbb{E}_x[w_i^t(k)] &= w_i^0(k), k > 2.\end{aligned}$$

where the last line is derived by choosing w.l.o.g. $\eta = m$:

$$\begin{aligned}\mathbb{E}_x[w_i^t(1)] &= 2^{t-1}(w_i^0(1) + a_i w_i^0(2)), \\ \mathbb{E}_x[w_i^t(2)] &= 2^{t-1}(w_i^0(2) + a_i w_i^0(1)), \\ \mathbb{E}_x[w_i^t(k)] &= w_i^0(k), \quad k > 2.\end{aligned}$$

therefore,

$$\begin{aligned}|\mathbb{E}_x[w_i^t(1)]| &= \Theta\left(\frac{2^t}{\sqrt{d}}\right), \\ |\mathbb{E}_x[w_i^t(2)]| &= \Theta\left(\frac{2^t}{\sqrt{d}}\right), \\ |\mathbb{E}_x[w_i^t(k)]| &= \Theta\left(\frac{1}{\sqrt{d}}\right).\end{aligned}$$

Select $t = \log(d)$. Then, w.h.p. it holds that

$$\begin{aligned}|\mathbb{E}_x[w_i^t(1)]| &= \Theta(\sqrt{d}), \\ |\mathbb{E}_x[w_i^t(2)]| &= \Theta(\sqrt{d}), \\ |\mathbb{E}_x[w_i^t(k)]| &= \Theta\left(\frac{1}{\sqrt{d}}\right).\end{aligned}$$

The calculations above indicate that the signal strength is already larger than the noise after $\log(d)$ SGD steps. We are now ready to compute $y\Phi(w^t, x)$. Assume for simplicity and without loss of generality that $x(1) = x(2) = y = 1$ as other cases lead to the same result. Recall $v_i^t(k) := \mathbb{E}_x[w_i^t(k)] - w_i^t(k)$. Then,

$$\begin{aligned}
y\Phi(w^t, x) &= \frac{1}{m} \sum_{i=1}^m a_i \left(w_i^t(1) + w_i^t(2) + \sum_{k=3}^d w_i^t(k) x(k) \right)^2 \\
&= \frac{1}{m} \sum_{i=1}^m a_i \left(d(w_i^0(1) + w_i^0(2))(a_i + 1) - v_i^t(1) - v_i^t(2) + \sum_{k \geq 3} (w_i^0(k) - v_i^t(k)) x(k) \right)^2 \\
&= \frac{1}{m} \sum_{i \in P} \left(2d(w_i^0(1) + w_i^0(2)) - v_i^t(1) - v_i^t(2) + \sum_{k \geq 3} (w_i^0(k) - v_i^t(k)) x(k) \right)^2 \\
&\quad + \frac{1}{m} \sum_{i \in N} \left(v_i^t(1) + v_i^t(2) + \sum_{k \geq 3} (v_i^t(k) - w_i^0(k)) x(k) \right)^2,
\end{aligned}$$

where P and N denote the set of neurons for which $a_i = 1$ and $a_i = -1$, respectively.

Note that the last term in each summation can be simplified as follows:

$$\sum_{k \geq 3} (w_i^0(k) + v_i^t(k)) x(k) = O(\|v_i^t\|) = O(\beta^t) = O\left(\frac{\sqrt{d}}{\log^{\frac{c-3}{2}}(d)}\right).$$

Concretely, by Hoeffding w.p. $1 - C(p_1 - p_2 - p_x - p_w)$ over data sampling and initialization :

$$\Pr_x \left(\sum_{k \geq 3} (w_i^0(k) + v_i^t(k)) x(k) > \frac{\sqrt{d}}{\log^{\frac{c-4}{2}}(d)} \right) \leq e^{-\log(d)} = \frac{1}{d}. \quad (54)$$

Define $\zeta_i^t := v_i^t(1) + v_i^t(2) + \sum_{k \geq 3} (v_i^0(k) - w_i^t(k)) x(k)$. Then,

$$\begin{aligned}
y\Phi(w^t, x) &= \frac{1}{m} \sum_{i \in P} (2d(w_i^0(1) + w_i^0(2)) - \zeta_i^t)^2 + \frac{1}{m} \sum_{i \in N} (\zeta_i^t)^2 \\
&= \frac{1}{m} \sum_{i \in P} 4d^2(w_i^0(1) + w_i^0(2))^2 - 4d\zeta_i^t(w_i^0(1) + w_i^0(2)) + \frac{1}{m} \sum_{i \in [m]} (\zeta_i^t)^2 \\
&\geq \frac{1}{m} \sum_{i \in P} 4d^2(w_i^0(1) + w_i^0(2))^2 - 4d|\zeta_i^t| \cdot |w_i^0(1) + w_i^0(2)|. \quad (55)
\end{aligned}$$

Moreover, we had w.p. at least $1 - C(p_1 + p_2 + p_x + p_w)$ that $|v_i^t(1)| + |v_i^t(2)| < \frac{C\sqrt{d}}{\log^{\frac{c-3}{2}}(d)}$ for some absolute constant C . Adding these two together, we deduce that w.p. $1 - o_d(1)$ it holds uniformly over i, t ,

$$|\zeta_i^t| < \frac{C\sqrt{d}}{\log^{\frac{c-4}{2}}(d)}.$$

It remains to obtain lower- and upper-bounds on $(w_i^0(1) + w_i^0(2))^2$ as required by Eq. 55. To proceed, note that $\frac{d}{2}(w_i^0(1) + w_i^0(2))^2 \sim \chi^2(1)$. Therefore with standard concentration bounds for $\chi^2(1)$ and $\chi^2(m)$ distributions (e.g., Laurent & Massart (2000)) we find for any $u \geq 0$,

$$\begin{aligned}
\Pr \left(\frac{d}{2}(w_i^0(1) + w_i^0(2))^2 - 1 \geq 2\sqrt{u} + 2u \right) &\leq e^{-u} \\
\Pr \left(m - \frac{d}{2} \sum_{i=1}^m (w_i^0(1) + w_i^0(2))^2 \geq 2\sqrt{mu} \right) &\leq e^{-u}
\end{aligned}$$

By selecting $u = \log^2(d)$ in the first inequality above, we find w.p. at most $e^{\log(m) - \log^2(d)}$ it holds uniformly for all $i \leq m$:

$$(w_i^0(1) + w_i^0(2))^2 \geq \frac{2}{d} + \frac{4}{d} \sqrt{\log^2(d)} + \frac{4}{d} \log^2(d),$$

and by noting that $|P| = |N| = m/2$ and selecting $u = m/16$ in the second inequality we have w.p. at least $1 - e^{-\frac{m}{16}}$,

$$\frac{1}{m} \sum_{i \in P} (w_i^0(1) + w_i^0(2))^2 \geq (1 - \frac{1}{\sqrt{2}}) \frac{1}{d}.$$

Hence w.p. $1 - e^{\log(m) - \log^2(d)} - e^{-\frac{m}{16}} - C(e^{-\log^2(d)} + e^{\log(dmT) - 2\log^2(dmT)} + e^{\log(dT) - 2\log^2(dT)} + e^{-\frac{1}{2}\log^2(dm) + \log(dm)})$ over data sampling and initialization,

$$\Pr_{x,y} \left(y\Phi(w^t, x) \geq d - \frac{4Cd}{\log^{\frac{c-6}{2}}(d)} \right) \geq 1 - 1/d. \quad (56)$$

For the theorem's choice of $T = \log(d)$, $m = O(\text{poly}(d))$ it is deduced that $e^{\log(m) - \log^2(d)} + C(e^{-\log^2(d)} + e^{\log(dmT) - 2\log^2(dmT)} + e^{\log(dT) - 2\log^2(dT)} + e^{-\frac{1}{2}\log^2(dm) + \log(dm)}) = o_d(1)$. Moreover, for the theorem's choice $c = 7$ and large enough d it holds that $d - 4Cd/\log^{\frac{c-6}{2}}(d) \geq 0$. This leads to the following lower bound on test accuracy,

$$\Pr_{x,y} (y\Phi(w^t, x) \geq 0) \geq 1 - \frac{1}{d}, \quad (57)$$

with probability at least $1 - e^{-\frac{m}{16}} - e^{\log(m) - \log^2(d)} - o_d(1)$ over data sampling and initialization. This completes the proof of Theorem 2.4.