LARGE LEARNING RATES FOR DEEP RELU NETWORKS: OPTIMIZATION AND GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large constant learning rates can improve the convergence rate of gradient descent in separable linear and two-layer models. We establish optimization and high-probability population risk bounds for deep ReLU networks trained by gradient descent with an arbitrary constant learning rate. For logistic loss, we assume finite-width neural tangent kernel separability with margin γ and paired Gaussian initialization, and train all hidden layers while fixing the output layer. We control the full, possibly nonmonotone trajectory and use the resulting radius to bound the complexity of a function class containing all iterates, connecting optimization to population risk analysis. The bounds hold over a finite training budget under explicit width conditions depending on depth, learning rate and budget, and yield two consequences under different learning rate choices. First, choosing a large learning rate $\eta = \Theta(\gamma^2 T)$ gives final training loss $\mathcal{O}(1/(\gamma^4 T^2))$ under the stated budget conditions, without momentum or a schedule. Second, choosing $\eta = \Theta(1)$ and $T \asymp n$ gives population risk averaged over iterates of order $\widetilde{\mathcal{O}}(1/(\gamma^2 n) + L^2/(\gamma n))$ at fixed confidence, where L is the number of hidden layers and n is the sample size. The last iterate has the same rate if the stable phase is reached with a constant fraction of the training budget remaining. At fixed depth and confidence, the bound has the same dependence on sample size and margin as the SVM-type benchmark. Extensions to a class of convex decreasing losses distinguish the conditions for trajectory control, stable-phase convergence and population risk bounds. Experiments on noisy 2-XOR and Fashion-MNIST examine the optimization and statistical effects of larger learning rates.

1 INTRODUCTION

The learning rate affects both the efficiency of optimization and the generalization of neural networks. Full batch gradient descent (GD) often exhibits nonmonotone loss near the edge of stability (EoS), beyond the regime explained by standard descent arguments (Cohen et al., 2021). Studies of learning rate annealing, sparse features and loss spikes associate large learning rates with improved generalization (Li et al., 2019; Andriushchenko et al., 2023; Zhu et al., 2024). Warmup can enable larger target learning rates and improve training and final performance (Kalra & Barkeshli, 2024). For deep architectures, suitably large initial learning rates can also improve generalization after fine-tuning with a small learning rate or weight averaging, whereas excessively large rates can impair learning (Sadrtdinov et al., 2024). These findings motivate a theoretical understanding of how large learning rates affect both optimization and generalization in deep networks. We study GD for fully connected ReLU networks with any constant learning rate and establish optimization and population risk bounds with explicit learning rate dependence.

For separable linear and two-layer models, recent theory shows that large constant learning rates can accelerate GD beyond the monotone decreasing regime. For linear models with linearly separable data and two-layer ReLU networks under neural tangent kernel (NTK) separability, Wu et al. (2024) show that GD with a suitably chosen large constant learning rate reaches a stable phase and can achieve faster convergence without momentum. Related acceleration and margin improvement results are available for smooth, nearly homogeneous two-layer networks (Cai et al., 2024). Wu et al. also suggest extending their optimization results to deep networks by combining their techniques with standard NTK arguments (Allen-Zhu et al., 2019; Chen et al., 2021). This motivates studying

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

both optimization acceleration and population risk guarantees for deep networks beyond the small learning rate regime.

Recent analyses obtain fast population risk rates for shallow and deep ReLU networks under NTK separability, but impose upper bounds on the learning rate (Lei et al., 2026; Li et al., 2026). Such learning rate restrictions leave a gap between these guarantees and empirical observations of effective training with larger learning rates in deep networks (Cohen et al., 2021; Sadrtdinov et al., 2024). It remains unclear whether the optimization benefits of large constant learning rates extend to deep ReLU networks, and what generalization guarantees can be established in this regime. This leads to the following question:

> *Can we establish optimization and population risk bounds for deep ReLU networks under arbitrary constant learning rates, and identify learning rate choices that yield accelerated optimization or fast population risk rates?*

We address this question for GD applied to all hidden layers of deep ReLU networks, using paired Gaussian initialization and a fixed output layer. Under NTK separability and width conditions that depend on the learning rate and training budget, we control the full trajectory for any constant learning rate without requiring monotone loss descent. The resulting radius also controls the complexity of a function class containing all iterates, allowing us to derive high-probability population risk bounds from the same trajectory analysis. Different learning rate choices then yield accelerated optimization or fast population risk rates. Our main contributions are as follows.

1. **Deep ReLU networks with arbitrary constant learning rates.** For any constant learning rate $\eta > 0$ and finite training budget T, we control the entire GD trajectory and establish an average training loss bound. After GD enters the stable phase, we also bound the last iterate training loss (Theorem 1). We establish a split comparator analysis for deep ReLU networks that controls the nonlinear remainder and the gradient norm squared term without a learning rate upper bound.
2. **Accelerated optimization.** Under the width and training budget conditions in Corollary 1, choosing a constant learning rate $\eta = \Theta(\gamma^2 T)$ yields a last iterate training loss of $\mathcal{O}(1/(\gamma^4 T^2))$ after T iterations. This bound has the same T^{-2} iteration dependence as classical accelerated gradient methods for smooth convex optimization (Nesterov, 2004), but is obtained here for deep ReLU networks using GD without momentum or learning rate scheduling.
3. **High-probability population risk bounds and fast rates.** To our knowledge, this is the first finite-sample analysis that establishes high-probability population risk bounds for arbitrary constant learning rates in the above deep ReLU setting (Theorem 2). Under the additional width condition in Corollary 2, choosing $\eta = \Theta(1)$ and $T \asymp n$ gives a population risk averaged over iterates of $\widetilde{\mathcal{O}}(1/(\gamma^2 n) + L^2 \log(2/\delta)/(\gamma n))$. The last iterate achieves the same rate if the stable-phase condition holds at some $s < T$ with $T - s \asymp T$. For fixed depth and confidence, the dependence on n and γ matches the optimal SVM-type benchmark up to logarithmic factors. The explicit polynomial dependence on depth in this risk bound appears only in the additive L^2 term, improving the depth dependence of the bound in Li et al. (2026) under the respective width conditions.

The accelerated optimization and fast population risk rates use different learning rate choices. See Table 1 for a comparison with existing work. Section 5 examines the optimization and generalization behavior on noisy 2-XOR and Fashion-MNIST.

2 Related Work

Optimization and training dynamics. Beyond the unregularized setting in Table 1(a), Wu et al. (2025) establish acceleration for regularized logistic regression using large learning rates, while Zhang et al. (2026) obtain minimax-optimal iteration complexity for separable logistic regression using large adaptive learning rates. For linearly separable data satisfying a nondegeneracy condition, Wu et al. (2023) characterize the implicit bias of logistic regression with arbitrary constant learning rates. At the edge of stability, Damian et al. (2023) analyze the mechanism of self-stabilization, and

Table 1: Comparison of optimization and population risk bounds for logistic loss.

(a) Optimization with large constant learning rates

Work	Model and separation	Learning rate	Final training loss
Wu et al. (2024)	Two-layer ReLU; NTK margin γ	$\Theta(\gamma^2 T)$	$\widetilde{\mathcal{O}}(\gamma^{-4}T^{-2})$
Cai et al. (2024)	Smooth two-layer network; linear margin γ	$\Theta(m\alpha^2\gamma^2 T)$	$\mathcal{O}(\alpha^{-4}\gamma^{-4}T^{-2})$
Our work	Deep ReLU; finite-width NTK margin γ	$\Theta(\gamma^2 T)$	$\mathcal{O}(\gamma^{-4}T^{-2})$

All results require their stated assumptions and sufficient training budgets. Learning rates follow each paper's parameterization. Cai et al. allow any width under smoothness and near-homogeneity conditions, with $\alpha \le |\phi'| \le 1$ and other activation constants fixed. The sufficient width in Theorem 6 of Wu et al. (2024) depends on the learning rate and training budget; our corresponding condition is (2).

(b) Population risk under NTK separability

Work	Network and sufficient width	Learning rate and budget	Population risk
Lei et al. (2026)[‡]	Shallow ReLU; $\widetilde{\Omega}(\gamma^{-8})$	$\eta \le 16/5$, $\eta T = n$	Average: $\widetilde{\mathcal{O}}(1/(\gamma^2 n))$
Li et al. (2026)	Deep ReLU; $\widetilde{\Omega}(n^3 L^{16} d/\gamma^8)$	$\eta \le 4/(5L)$, $\eta T \asymp n$	Average: $\widetilde{\mathcal{O}}(L^6/(\gamma^2 n))$
Our work	Deep ReLU; $\widetilde{\Omega}(L^{13}\gamma^{-8} + L^{11}(d+n^3L^2)\gamma^{-2})$	$\eta = \Theta(1)$, $T \asymp n$	Average / last iterate[†]: $\widetilde{\mathcal{O}}\Big(\frac{1}{\gamma^2 n} + \frac{L^2\log(2/\delta)}{\gamma n}\Big)$

Here m is the network width, d the input dimension, L the depth, n the sample size, T the number of GD updates, and δ the failure probability. $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ omit logarithmic factors. Margins are defined in the respective input or feature spaces and need not coincide. "Average" denotes population risk averaged over iterates. [†]The last iterate additionally requires the stable-phase condition at some $s < T$ with $T - s \asymp T$; see Corollary 2. [‡]The result of Lei et al. cited here assumes $\|W_*\|_{2,\infty} \le m^{-1/2}$ for the unit separating direction; see Assumption 4, Theorem 18 and Remark 19 therein. Panels (a) and (b) use different learning rate choices.

Jiang et al. (2025) study NTK eigenvector alignment through a two-layer linear model and experiments on nonlinear networks. For deep networks, Chatterji et al. (2021) establish convergence of GD with logistic loss and smoothed ReLU activations. Cai et al. (2025) characterize the asymptotic implicit bias under exponential loss for nonhomogeneous networks satisfying a near-homogeneity condition, starting from sufficiently small empirical risk.

Generalization of gradient methods. The NTK framework connects neural network training with kernel methods (Jacot et al., 2018). Under NTK separability, Ji & Telgarsky (2020) establish test error bounds for shallow ReLU networks with polylogarithmic width at fixed margin. For deep ReLU networks, the GD classification error bound of Chen et al. (2021) includes $n^{-1/2}$ statistical terms and a width-dependent approximation term. More recent analyses obtain population loss bounds of $\widetilde{\mathcal{O}}(1/(\gamma^2 n))$ for shallow ReLU networks (Lei et al., 2026) and $\widetilde{\mathcal{O}}(L^6/(\gamma^2 n))$ for deep ReLU networks with $\eta \le 4/(5L)$ (Li et al., 2026), under their respective width and training horizon conditions; see Table 1(b). For smooth activations, algorithmic stability yields expected population loss bounds for the last iterate in shallow and deep networks (Taheri & Thrampoulidis, 2024; Taheri et al., 2025). In regression, Qiao et al. (2024) study generalization at stable minima of univariate two-layer ReLU networks, Xu & Zhu (2024) analyze prediction error for streaming SGD in multi-layer networks, and Zhou et al. (2026) establish minimax rates under squared loss and source and capacity assumptions. For separable linear classification, Shamir (2021) establishes nonasymptotic

generalization bounds, while Schliserman & Koren (2023) give matching upper and lower bounds on population risk.

3 Preliminaries and Problem Setting

Notation. For a positive integer q, write $[q] = \{1, \dots, q\}$ and let I_q denote the $q \times q$ identity matrix. We denote the unit sphere in $\mathbb{R}^d$ by $\mathbb{S}^{d-1}$ and the indicator of an event A by $\mathbb{I}\{A\}$. The notation $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the spectral norm for matrices; $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\cdot\|_0$ counts nonzero entries. For collections of layer matrices, $\langle U, V\rangle_F = \sum_{l=1}^{L} \operatorname{tr}((U^l)^\top V^l)$ and $\|U\|_F^2 = \sum_{l=1}^{L} \|U^l\|_F^2$. We write $a \lesssim b$ if $a \le Cb$ for an absolute constant $C > 0$, and $a \asymp b$ if both $a \lesssim b$ and $b \lesssim a$ hold. We use $\mathcal{O}$, Ω, and Θ for asymptotic upper, lower, and two-sided bounds, respectively. The notation $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ omits logarithmic factors from the corresponding upper and lower bounds unless displayed explicitly. Subscripts on $\mathbb{P}$ and $\mathbb{E}$ indicate the sources of randomness; in particular, $\mathbb{P}_{S,W(0)}$ denotes probability over the joint distribution of the sample and initialization. Constants denoted by C may change from line to line.

Setup. We consider deep fully connected ReLU networks, as in Chen et al. (2021); Xu & Zhu (2024); Li et al. (2026). Specifically, the network has $L \ge 2$ trainable hidden layers of width m and a fixed linear output layer. Thus, L denotes the number of hidden layers throughout this paper. Let $\mathbf{x} \in \mathbb{R}^d$ be the input and $W = \{W^1, \dots, W^L\}$ be the collection of trainable weight matrices, where $W^1 \in \mathbb{R}^{m\times d}$ and $W^l \in \mathbb{R}^{m\times m}$ for $l = 2, \dots, L$. Write w_r^l for the transpose of the rth row of W^l, and let $\mathbf{a} \in \mathbb{R}^m$ be the fixed output weight vector. The network function $f_W : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$f_W(\mathbf{x}) = \mathbf{a}^\top h^L(\mathbf{x}), \qquad h^l(\mathbf{x}) = \sqrt{\frac{2}{m}}\,\sigma\big(W^l h^{l-1}(\mathbf{x})\big), \quad l \in [L],$$

where $h^0(\mathbf{x}) = \mathbf{x}$ and $\sigma(z) = \max\{0, z\}$ is the coordinatewise ReLU activation function. Appendix A gives the matrix notation used in the proofs and specifies the convention for computing gradients at zero. Let $W(0)$ denote the initialization, and define the parameter ball $\mathcal{B}_R(W(0)) := \{W : \|W - W(0)\|_F \le R\}$.

We study binary classification on the input space $\mathcal{X} = \mathbb{S}^{d-1}$. Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{-1, +1\}$, and let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ be an i.i.d. training sample of size $n \ge 2$. For the logistic loss $\ell(z) = \log(1 + e^{-z})$, define the empirical and population risks by

$$\mathcal{L}_S(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f_W(\mathbf{x}_i)), \qquad \mathcal{L}(W) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(y f_W(\mathbf{x}))\right].$$

Throughout the paper, ℓ denotes the logistic loss and ℓ' its derivative, except in Appendix D, where we consider more general losses. We also define the *gradient potential*

$$\mathcal{P}_S(W) := \frac{1}{n}\sum_{i=1}^{n} |\ell'(y_i f_W(\mathbf{x}_i))|.$$

Algorithm. We consider full-batch gradient descent (GD) with a constant learning rate $\eta > 0$:

$$W(t+1) = W(t) - \eta \nabla \mathcal{L}_S(W(t)), \tag{1}$$

where the gradient is taken with respect to W, while the output weights $\mathbf{a}$ remain fixed.

4 Main Theoretical Results

Under the assumptions below and suitable width conditions, we first control the full GD trajectory for any constant learning rate. We then use the resulting deterministic radius to bound population risk. Proofs of the main results are given in Appendix C.

4.1 Assumptions

We use a symmetric initialization, as in Nitanda & Suzuki (2021); Kuzborskij & Szepesvári (2023); Xu & Zhu (2024); Li et al. (2026), with the Gaussian pairing specified below. We also assume

normalized inputs and separability of the neural tangent features at initialization (Ji & Telgarsky, 2020; Chen et al., 2021; Li et al., 2026). The separation condition is imposed on the finite-width gradient features of the training sample.

Assumption 1 (Symmetric Initialization). The network width m is even. The fixed output weights satisfy $a_r \in \{-1, +1\}$ and $a_{r+m/2} = -a_r$ for $r \in [m/2]$. The initial hidden weights are given by

$$\begin{aligned} w_r^1(0) &\sim \mathcal{N}(0, I_d), && r \in [m], \\ w_r^l(0) &\sim \mathcal{N}(0, I_m), && l = 2, \dots, L-1, \quad r \in [m], \\ w_r^L(0) &\sim \mathcal{N}(0, I_m), && r \in [m/2], \end{aligned}$$

where all Gaussian vectors sampled above are mutually independent. Set $w_{r+m/2}^L(0) = w_r^L(0)$ for $r \in [m/2]$. The initialization $W(0)$ is independent of the training sample S.

Assumption 2 (Input Normalization). The input space satisfies $\mathcal{X} = \mathbb{S}^{d-1}$.

Assumption 3 (NTK Separability). There exist a margin $\gamma \in (0, 1]$ and a collection of matrices $W_* = \{W_*^1, \dots, W_*^L\}$ with $\|W_*\|_F = 1$ such that

$$y_i \left\langle W_*, \nabla f_{W(0)}(\mathbf{x}_i) \right\rangle_F \geq \gamma, \qquad i \in [n].$$

4.2 Optimization and Acceleration

The probability statements below concern good events under the original initialization and sample distributions. On these events, the conclusions apply to realizations satisfying Assumption 3 with the fixed margin γ.

Theorem 1 (Optimization with arbitrary constant learning rates)**.** *Suppose Assumptions 1, 2, and 3 hold with margin γ. The margin γ, learning rate $\eta > 0$, integer budget $T > 1$, and confidence parameter $\delta \in (0, 1)$ are fixed independently of S and $W(0)$. Consider GD* (1) *for T steps. Let $K_1 := \sqrt{2} + \gamma/2$, $\theta_T := 2\log(1 + 2\eta T)/\gamma$, $\rho := \eta K_1^2/\gamma$, $R_0 := 2(\theta_T + \rho) + 1$, and $R := \max\{R_0, \sqrt{2}/\gamma\}$. If*

$$m \gtrsim L^{13}(\log m)^4 \log(nL/\delta) R^8, \tag{2}$$

then, for any fixed sample S, with probability at least $1 - \delta$ over the initialization, the following conclusions hold:

1. ***Trajectory control.*** *For every $t \in \{0, \dots, T\}$, $\|W(t) - W(0)\|_F \leq R_0 \leq R$.*

2. ***Average training loss.*** *For every $t \in [T]$,*

$$\frac{1}{t} \sum_{k=0}^{t-1} \mathcal{L}_S(W(k)) \leq \frac{1}{\eta T} + \frac{(\theta_T + \rho)^2}{\eta t}. \tag{3}$$

3. ***Stable phase.*** *If there exists $s \in \{0, \dots, T-1\}$ such that*

$$\mathcal{L}_S(W(s)) \leq \min\left\{ \frac{1}{4e^2 \eta (\sqrt{2} + 5\gamma/8)^2}, \frac{\log 2}{n} \right\}, \tag{4}$$

then $\{\mathcal{L}_S(W(t))\}_{t=s}^T$ is nonincreasing and, for every $t \in \{s+1, \dots, T\}$,

$$\mathcal{L}_S(W(t)) \leq \frac{32}{\gamma^2 \eta (t - s)}. \tag{5}$$

4. ***Transition time.*** *Define*

$$\tau := \left\lceil \frac{4R}{\gamma \eta} \max\left\{ 4e^2 \eta (\sqrt{2} + 5\gamma/8)^2, \frac{n}{\log 2} \right\} \right\rceil.$$

If $\tau \leq T$, then (4) *holds for some $s \in \{0, \dots, \tau - 1\}$.*

Corollary 1 (Acceleration with a large constant learning rate)**.** *Consider the setting of Theorem 1. Let* $K_2 := \sqrt{2} + 5\gamma/8$ *and* $C_\tau := 16e^2K_2^2/\gamma$*, and choose* $\eta := \gamma T/(8C_\tau K_1^2) = \Theta(\gamma^2 T)$*. Assume that* (2) *holds with the corresponding radius* R*. There exists an absolute constant* $C > 0$ *such that, if*

$$T \geq C \max\left\{ \frac{K_1^2 n}{\gamma^2}, C_\tau \left(1 + \frac{1}{\gamma}\right), \frac{C_\tau}{\gamma} \log\left(2 + \frac{C_\tau}{\gamma}\right) \right\},$$

then, on the event in Theorem 1, we have $\tau \leq T/2$ *and*

$$\mathcal{L}_S(W(T)) \leq \frac{CC_\tau K_1^2}{\gamma^3 T^2} = \mathcal{O}\left(\frac{1}{\gamma^4 T^2}\right).$$

4.3 Population Risk

We use the trajectory control in Theorem 1 to derive population risk bounds. Conditional on initialization, the deterministic radius defines a sample-independent function class containing all GD iterates on the optimization event. A uniform generalization bound for this class can therefore be combined with the average and last iterate training loss estimates.

Let R be the radius in Theorem 1. Write $E(m, R) := CL^{13/6}(\log m)^{2/3}R^{1/3}m^{-1/6}$ for the linearization error coefficient and $\mathcal{C}_{n,m}(R) := R/\sqrt{n} + E(m, R)R$ for the corresponding Rademacher complexity bound; see Lemmas 8 and 17.

Theorem 2 (Population risk bounds)**.** *Suppose the conditions of Theorem 1 hold and, in addition,*

$$m \gtrsim L^{11} d(\log m)^5 \log(L/\delta) R^2. \tag{6}$$

Then, with probability at least $1 - \delta$ *over the sample and initialization, the following conclusions hold:*

1. ***Average population risk.***

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}(W(t)) \lesssim \frac{1}{\eta T} + \frac{\eta + \eta^{-1}\log^2(1+\eta T)}{\gamma^2 T} + \log^3 n\, \mathcal{C}_{n,m}^2(R) + \frac{\left(1 + L^2\sqrt{\log m}\, R\right)\log(2/\delta)}{n}. \tag{7}$$

2. ***Last iterate population risk.*** *If* (4) *holds at some* $s \in \{0, \ldots, T-1\}$ *with* $T - s \asymp T$*, then*

$$\mathcal{L}(W(T)) \lesssim \frac{1}{\gamma^2 \eta T} + \log^3 n\, \mathcal{C}_{n,m}^2(R) + \frac{\left(1 + L^2\sqrt{\log m}\, R\right)\log(2/\delta)}{n}. \tag{8}$$

Condition (6) controls the network outputs over the input sphere. To obtain $\mathcal{C}_{n,m}(R) \lesssim R/\sqrt{n}$, we additionally impose

$$m \gtrsim n^3 L^{13}(\log m)^4 \log(nL/\delta) R^2. \tag{9}$$

This makes the nonlinear complexity term no larger than the linear term. Substituting the radius R gives (27) and (28) in Appendix C.3. Choosing $\eta = \Theta(1)$ and $T \asymp n$ then yields the following rates. The average is taken over risks at individual iterates, not over network parameters.

Corollary 2 (Fast population risk rates)**.** *Suppose the conditions of Theorem 2 and the additional width condition* (9) *hold. Let* $T \asymp n$ *and choose an absolute constant learning rate* $\eta = \Theta(1)$*, independent of* n*,* L*, and* γ*. Then, on the same event as in Theorem 2,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}(W(t)) = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma^2 n} + \frac{L^2 \log(2/\delta)}{\gamma n}\right). \tag{10}$$

If, in addition, (4) *holds at some* $s \in \{0, \ldots, T-1\}$ *with* $T - s \asymp T$*, then*

$$\mathcal{L}(W(T)) = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma^2 n} + \frac{L^2 \log(2/\delta)}{\gamma n}\right). \tag{11}$$

Remark 1 (Population risk at the last iterate). Theorem 18 of Lei et al. (2026) and Theorem 3 of Li et al. (2026) give averaged population risk bounds for shallow and deep ReLU networks, respectively, with $\eta \leq 16/5$ and $\eta \leq 4/(5L)$ under their respective assumptions. Our uniform bound (29) and stable-phase estimate (5) also give the last iterate rate under the phase condition in Corollary 2. For $\eta = \Theta(1)$, the larger budget $T = \lceil C(n/\gamma^2)\log(2+n/\gamma^2)\rceil$, with sufficiently large C and the corresponding width conditions, ensures this phase condition with $s < T/2$ and yields the same statistical rate; see Appendix C.5. Last iterate bounds on expected population risk are also available for smooth shallow and deep networks under learning rate restrictions ensuring descent (Taheri & Thrampoulidis, 2024; Taheri et al., 2025). Our general bound instead holds with high probability for deep ReLU networks and permits arbitrary constant learning rates under the stated width conditions.

Remark 2 (Effect of the learning rate). The same analysis yields different optimization and statistical rates through the choice of learning rate. With $\eta = \Theta(1)$, Theorem 1 gives average training loss $\widetilde{\mathcal{O}}(1/(\gamma^2 T))$. Taking $T \asymp n$ also gives the fast average population risk rate in Corollary 2, whose dependence on n and γ matches the SVM-type benchmark at fixed depth and confidence, up to logarithmic factors. The same population risk rate holds at the last iterate under the stated phase condition. In contrast, the choice $\eta = \Theta(\gamma^2 T)$ in Corollary 1 yields last iterate training loss $\mathcal{O}(1/(\gamma^4 T^2))$. This larger learning rate also increases the trajectory radius: under (9), the statistical bound contains a term of order $\widetilde{\mathcal{O}}(\eta^2/(\gamma^2 n))$. For fixed γ and $T \asymp n$ satisfying the acceleration budget conditions, $\eta^2/(\gamma^2 n) = \Theta(\gamma^2 n)$, so the present bound does not ensure vanishing population risk for this choice. Determining whether the accelerated choice also admits fast population risk rates in this deep ReLU setting remains open.

4.4 Proof Technique

The main difficulty is to control the entire GD trajectory for deep ReLU networks without requiring monotone loss descent. In the two-layer analysis of Wu et al. (2024), the proof of Lemma 27 uses 1-homogeneity to control the comparator inner product. When all hidden layers are trained, our predictor is instead L-homogeneous, so the same calculation does not yield the required bound. For deep networks, the optimization analysis of Li et al. (2026) imposes learning rate upper bounds to control the squared-gradient and reference-model contributions in the distance recursion and close the trajectory induction. We therefore develop a split comparator analysis for deep ReLU networks that controls these contributions together with the nonlinear remainder, allowing the trajectory argument to accommodate any prescribed constant learning rate.

Drawing on the comparator idea of Wu et al. (2024), we construct

$$u = u_1 + u_2, \qquad u_1 = W(0) + \theta_T W_*, \qquad u_2 = \rho W_*, \qquad \rho = \frac{\eta K_1^2}{\gamma}.$$

The essential change is in the control of the comparator inner product. Instead of using 1-homogeneity, we compare the actual network losses at u_1 and the current iterate through a deep linearization estimate. We establish linearization and gradient perturbation bounds in the Frobenius norm over all trainable layers (Lemmas 8 and 9). These estimates ensure a small reference loss, preserve the separating direction, and control the loss-weighted nonlinear remainder. Writing $g_t = \nabla \mathcal{L}_S(W(t))$, we obtain

$$\begin{aligned} \langle g_t, u_1 - W(t)\rangle_F &\leq \mathcal{L}_S(u_1) - \tfrac{1}{2}\mathcal{L}_S(W(t)), \\ \eta\|g_t\|_F^2 + 2\langle g_t, u_2\rangle_F &\leq 0. \end{aligned}$$

The first inequality absorbs the deep linearization remainder into the current training loss, while the second controls the gradient norm squared term through the displacement u_2. Together with $\mathcal{L}_S(u_1) \leq e^{-\theta_T \gamma/2}$, these estimates yield the distance recursion in Lemma 10. This recursion closes the full trajectory induction without requiring monotone loss descent: the local estimates are applied at points already known to lie in the neighborhood, and the recursion then controls the next iterate. A gradient-potential estimate bounds the transition time, after which a descent argument yields the last iterate training loss bound (Appendix B.2).

Our statistical analysis uses the same deterministic radius to define a sample-independent function class conditional on initialization. At initialization, the pairing cancels the output and the gradients

with respect to the first $L-1$ hidden layers. Lemma 7 then bounds the total gradient norm without an explicit depth factor; this cancellation does not constrain later updates to the lower layers. The resulting linear Rademacher term is of order $R/\sqrt{n}$, while the nonlinear term is controlled by $E(m,R)R$ (Lemma 17). The uniform feature perturbation bound also controls the loss over the input sphere. Combining these estimates with Lemma 6 gives a high-probability population risk bound with explicit learning rate and depth dependence. The function class contains all GD iterates on the optimization event, so the uniform bound combines with the average and stable-phase training loss estimates to yield the population risk results (Appendix C.3).

Appendix D extends the analysis to a class of convex decreasing losses. The comparator displacement is enlarged to control the nonlinear remainder without requiring $|\ell'| \lesssim \ell$. Additional tail and local quadratic conditions are used for stable-phase convergence, while population risk bounds additionally require smoothness of the loss.

5 Experiments

Our primary contribution is theoretical. The experiments complement the analysis by examining the effects of the learning rate on training loss, population or test loss, and numerical stability in finite-width networks. We use the network in Section 3, with full batch GD, paired Gaussian initialization and fixed output weights, without biases, momentum, weight decay or normalization layers. We set $m = 128$ and use ten random seeds, with the same training samples and initializations across learning rates. Each learning rate remains constant throughout a run. The main figures show $L = 4, 16$; Appendix E gives results for all $L \in \{2, 4, 8, 16\}$ and the complete protocol. Lines show medians, and shaded regions span the 25th to 75th percentiles across completed runs. Runs that encounter numerical failure are reported separately and are not replaced.

Training and evaluation. We use noisy 2-XOR (Ji & Telgarsky, 2020; Li et al., 2026) with $d = 8$, $n = 128$, $T = 1024$, and binary Fashion-MNIST (Xiao et al., 2017) with classes 0 and 2, $n = 512$, $T = 512$. We compare constant learning rates $\eta \in \{1/4, 1, 4, 16\}$. Inputs are normalized to unit norm. For XOR, we compute population logistic loss by averaging over all 256 equally likely points in the distribution. For Fashion-MNIST, we report the mean logistic loss on the 2,000 images from the two selected classes in the official test split. Figure 1 shows that larger learning rates can reduce training loss faster, but lower training loss does not always give lower population or test loss. At $L = 16$ and $\eta = 16$, only $9/10$ XOR runs and $6/10$ Fashion-MNIST runs complete training. Thus, at the tested width, this combination of depth and learning rate also leads to numerical failures.

Sample size dependence. Motivated by Corollary 2, we examine how population risk varies with sample size under $T = 2n$ and a constant learning rate. On XOR, we vary $n \in \{32, 64, 128, 256\}$ with $\eta \in \{1, 16\}$, keeping the dimension, depth and width fixed within each comparison. Figure 2 plots $T^{-1}\sum_{t=0}^{T-1} \mathcal{L}(W(t))$ and $\mathcal{L}(W(T))$ against $1/n$, together with a c/n reference line. Both risks decrease substantially in several configurations, although the last iterate is not uniformly better and the curves do not consistently follow the reference. The experiment examines the empirical sample size dependence suggested by the corollary, with sample size and training budget increasing together. Appendix E additionally reports a fixed-budget control and Fashion-MNIST sample size results. It also examines how width affects training loss, completion rates and the retained separating direction, and reports when the stable-phase loss threshold is reached.

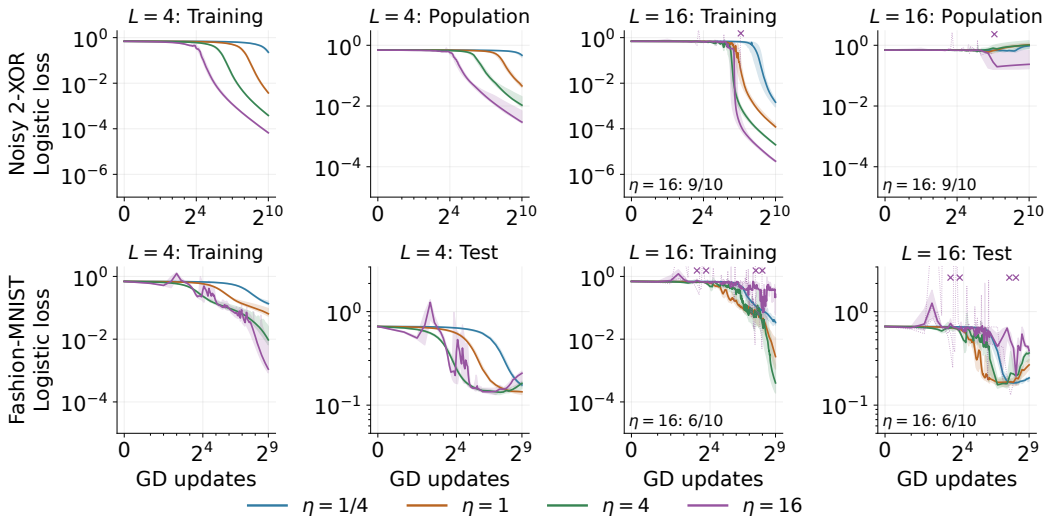

Figure 1: Training and population/test logistic loss at $L = 4, 16$. Top: noisy 2-XOR; bottom: Fashion-MNIST. Solid curves and shaded regions summarize completed runs. Dotted curves show recorded losses before numerical failure; crosses indicate the first iteration at which a nonfinite quantity is detected, with vertical positions used only for display. Completion counts are shown when fewer than ten runs finish. Values outside the axis limits are not shown. Full depth comparisons appear in Figures 3–4.

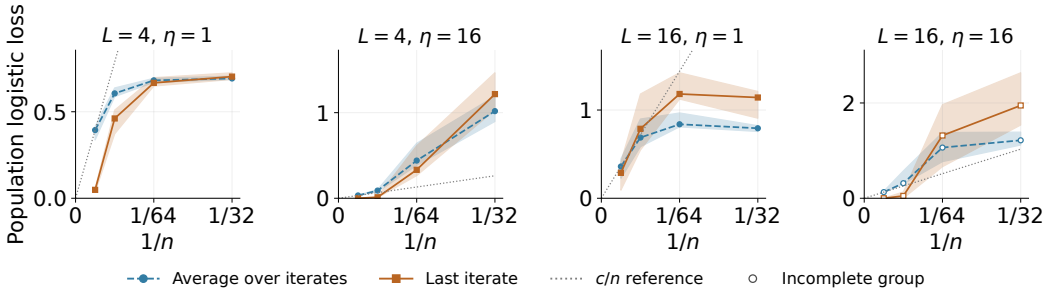

Figure 2: Average and last iterate population risks on XOR, with $m = 128$ and $T = 2n$. Lines show medians, and shaded regions span the 25th to 75th percentiles across completed runs. In each panel, the c/n reference passes through the median average risk at $n = 256$. Hollow markers indicate fewer than ten completed runs: for $L = 16, \eta = 16$, the counts are $8, 9, 9, 9$ at $n = 32, 64, 128, 256$, respectively; all other groups have ten completed runs. Full results appear in Figure 5.

6 Conclusion and Future Directions

We establish optimization and high-probability population risk bounds for deep ReLU networks with arbitrary constant learning rates under the stated initialization, separability and width conditions. Full trajectory control supports accelerated optimization and fast population risk rates under different learning rate choices; the last iterate attains the fast statistical rate when the stable-phase condition holds sufficiently early. Experiments examine training, population or test loss, and numerical stability across depths. As in the two-layer analysis of Wu et al. (2024), the sufficient width depends on the learning rate. Sharper control of the deep nonlinear remainder could weaken this dependence. Further questions concern matching statistical lower bounds under comparable separability assumptions, fast population risk bounds for the accelerated learning rate choice, and how to choose the learning rate and training budget for the best statistical rates under suitable distributional assumptions.

AI USE STATEMENT

The mathematical proofs were developed by the authors. Generative AI tools were used to polish the writing, provide feedback on the manuscript, assist with literature search, and help write the experimental code. The authors are responsible for reviewing the AI-assisted content and for the accuracy of the final text, mathematical claims, citations, code, and experimental results.

REPRODUCIBILITY STATEMENT

Section 3 states the model and algorithm, and Section 4.1 gives the initialization and separation assumptions. Appendices A–D give the auxiliary results and proofs, including the provenance of external lemmas. Appendix E specifies the datasets, preprocessing, seeds, budgets and treatment of incomplete runs, and Appendix E.7 describes implementation checks. The anonymous supplementary ZIP archive provides the experimental code, per-run records, configurations, and instructions for validating the records, regenerating the figures, and rerunning the experiments.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. arXiv:1811.03962v5, 2019.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with large step sizes learns sparse features. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 903–925, 2023.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 71306–71351, 2024.

Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6426–6504, 2025.

Niladri S. Chatterji, Philip M. Long, and Peter Bartlett. When does gradient descent with logistic loss interpolate using deep networks with smoothed ReLU activations? In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 927–1027, 2021.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep ReLU networks? In *International Conference on Learning Representations*, 2021.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Kaiqi Jiang, Jeremy Cohen, and Yuanzhi Li. Understanding the evolution of the neural tangent kernel at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Dayal Singh Kalra and Maissam Barkeshli. Why warmup the learning rate? Underlying mechanisms and improvements. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ilja Kuzborskij and Csaba Szepesvári. Learning Lipschitz functions by GD-trained shallow overparameterized ReLU neural networks. arXiv:2212.13848v2, 2023.

Yunwen Lei, Puyu Wang, Yiming Ying, and Ding-Xuan Zhou. Optimization and generalization of gradient descent for shallow ReLU networks with minimal width. *Journal of Machine Learning Research*, 27(34):1–35, 2026.

Yuanfan Li, Yunwen Lei, Zheng-Chu Guo, and Yiming Ying. Optimal rates for generalization of gradient descent for deep ReLU classification. arXiv:2510.02779v4, 2026.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2021.

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ildus Sadrtdinov, Maxim Kodryan, Eduard Pokonechny, Ekaterina Lobacheva, and Dmitry Vetrov. Where do large learning rates lead us? In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Matan Schliserman and Tomer Koren. Tight risk bounds for gradient descent on separable data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68749–68759, 2023.

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

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2199–2207, 2010.

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.

Hossein Taheri, Christos Thrampoulidis, and Arya Mazumdar. Sharper guarantees for learning neural network classifiers with gradient methods. In *International Conference on Learning Representations*, 2025.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 5019–5073, 2024.

Jingfeng Wu, Pierre Marion, and Peter L. Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.

Jiaming Xu and Hanjing Zhu. Overparametrized multi-layer neural networks: Uniform concentration of neural tangent kernel and convergence of stochastic gradient descent. *Journal of Machine Learning Research*, 25(94):1–83, 2024.

Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L. Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *Journal of Machine Learning Research*, 27(124):1–31, 2026.

Junyu Zhou, Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Optimal rates for generalization of gradient descent methods with deep neural networks. arXiv:2606.06764v1, 2026.

Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: Spikes in the training loss and their impact on generalization through feature learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62476–62509, 2024.

APPENDIX CONTENTS

A. Auxiliary results p. 13
B. Main lemmas p. 16
C. Proofs of the main results p. 25
D. Extension to other losses p. 28
E. Experimental details and additional results p. 32

A AUXILIARY RESULTS

We collect the initialization and perturbation bounds used in the proofs, followed by a generalization bound for smooth losses. The finite-input estimates hold for each fixed set of inputs, with probability over initialization; all radii and confidence parameters are deterministic. The uniform feature bound extends these estimates to the input sphere. The optimization-specific constructions are developed in Appendix B.

A.1 MATRIX NOTATION

For each $l \in [L]$, define the diagonal matrix $\Sigma^l(\mathbf{x}) \in \mathbb{R}^{m\times m}$ by

$$\Sigma^l(\mathbf{x}) = \operatorname{diag}\left(\mathbb{I}\left\{\langle w_r^l, h^{l-1}(\mathbf{x})\rangle \geq 0\right\}\right)_{r=1}^m. \tag{12}$$

When computing GD updates, we use the convention $\sigma'(z) = \mathbb{I}\{z \geq 0\}$, consistent with (12). At points where the network is not differentiable, ∇f_W denotes the backpropagation vector obtained using this convention. The network admits the representation

$$f_W(\mathbf{x}) = \mathbf{a}^\top \sqrt{\frac{2}{m}}\Sigma^L(\mathbf{x})W^L \cdots \sqrt{\frac{2}{m}}\Sigma^1(\mathbf{x})W^1\mathbf{x}.$$

Similarly,

$$h^l(\mathbf{x}) = \sqrt{\frac{2}{m}}\Sigma^l(\mathbf{x})W^l \cdots \sqrt{\frac{2}{m}}\Sigma^1(\mathbf{x})W^1\mathbf{x}, \quad l \in [L]. \tag{13}$$

Products are ordered with larger layer indices on the left, and an empty product is the identity. For $1 \leq a \leq b \leq L$, define

$$G_b^a(\mathbf{x}) = \left[\prod_{j=a+1}^{b} \sqrt{\frac{2}{m}}\Sigma^j(\mathbf{x})W^j\right] \sqrt{\frac{2}{m}}\Sigma^a(\mathbf{x}), \tag{14}$$

where $G_l^l(\mathbf{x}) = \sqrt{\frac{2}{m}}\Sigma^l(\mathbf{x})$.

A.2 DEEP PERTURBATION ESTIMATES

Let $\Sigma_0^l(\mathbf{x})$, $h_0^l(\mathbf{x})$, and $G_{b,0}^a(\mathbf{x})$ denote the quantities in (12), (13), and (14) evaluated at $W(0)$. For $l \in [L]$, let $\widehat{W}^l$ and the diagonal matrix $\widehat{\Sigma}^l(\mathbf{x})$ have the same dimensions as $W^l(0)$ and $\Sigma_0^l(\mathbf{x})$, respectively. Define

$$\begin{aligned}\widehat{G}_b^a(\mathbf{x}) = \sqrt{\frac{2}{m}}(\Sigma_0^b(\mathbf{x}) + \widehat{\Sigma}^b(\mathbf{x}))(W^b(0) + \widehat{W}^b) \cdots \sqrt{\frac{2}{m}}(\Sigma_0^{a+1}(\mathbf{x}) + \widehat{\Sigma}^{a+1}(\mathbf{x})) \\ \times (W^{a+1}(0) + \widehat{W}^{a+1})\sqrt{\frac{2}{m}}(\Sigma_0^a(\mathbf{x}) + \widehat{\Sigma}^a(\mathbf{x})), \quad 1 \leq a \leq b \leq L,\end{aligned} \tag{15}$$

and $\widehat{G}_l^l(\mathbf{x}) = \sqrt{\frac{2}{m}}(\Sigma_0^l(\mathbf{x}) + \widehat{\Sigma}^l(\mathbf{x}))$ for all $l \in [L]$.

Lemma 1 (Bound on hidden states at initialization; cf. Lemma 6 in Li et al. (2026))**.** *Suppose Assumptions 1 and 2 hold. For any $\delta \in (0,1)$, if $m \gtrsim L^2 \log(nL/\delta)$, then with probability at least $1-\delta$,*

$$\|h_0^l(\mathbf{x}_i)\|_2 \in \left[\sqrt{\frac{2}{3}}, \sqrt{\frac{4}{3}}\right], \qquad i \in [n], \quad l \in [L-1].$$

We only need this bound for $l \le L-1$, where the Gaussian rows are independent and the argument in Lemma 6 of Li et al. (2026) applies directly.

Lemma 2 (Lemma 12 in Li et al. (2026)). *Fix deterministic parameters $R > 0$ and $s \ge 0$. Suppose Assumptions 1 and 2 hold. Let $\widehat{G}_b^a(\mathbf{x})$ with $1 \le a \le b \le L$ be the matrix defined in* (15)*. Assume $\max_{l\in[L]} \|\widehat{W}^l\|_2 \le R \lesssim \sqrt{m}/(L^2\sqrt{\log m})$, $m \gtrsim L^2 \log(nL/\delta)$, and $\widehat{\Sigma}^l(\mathbf{x}_i)$, $\Sigma_0^l(\mathbf{x}_i) + \widehat{\Sigma}^l(\mathbf{x}_i) \in [-1,1]^{m\times m}$, $\|\widehat{\Sigma}^l(\mathbf{x}_i)\|_0 \le s \lesssim m/(L^2 \log m)$ for all $i \in [n], l \in [L]$. Then, with probability at least $1-\delta$ over the initialization, the following bound holds simultaneously for all perturbation matrices satisfying the above conditions, all $1 \le a \le b \le L$, and all $i \in [n]$:*

$$\|\widehat{G}_b^a(\mathbf{x}_i)\|_2 \lesssim L\sqrt{\frac{\log m}{m}}.$$

Lemma 3 (Lemma 13 in Li et al. (2026); based on Claim 11.2 and Proposition 11.3 in the full version of Allen-Zhu et al. (2019)). *For any $\mathbf{x} \in \mathcal{X}$ and $W, W' \in \mathcal{B}_R(W(0))$, there exists a collection of diagonal matrices $\{(\Sigma'')^l \in \mathbb{R}^{m\times m}\}_{l\in[L]}$ with entries in $[-1,1]$, such that $\Sigma^l + (\Sigma'')^l$ has diagonal entries in $[0,1]$ and, for every $l \in [L]$,*

(a)

$$(h')^l(\mathbf{x}) - h^l(\mathbf{x}) = \sum_{k=1}^{l}\left[\prod_{j=k+1}^{l}\sqrt{\frac{2}{m}}(\Sigma^j(\mathbf{x}) + (\Sigma'')^j)W^j\right] \times \sqrt{\frac{2}{m}}(\Sigma^k(\mathbf{x}) + (\Sigma'')^k)((W')^k - W^k)(h')^{k-1}(\mathbf{x}).$$

(b)

$$\|(\Sigma'')^l\|_0 \le \|\Sigma^l(\mathbf{x}) - (\Sigma')^l(\mathbf{x})\|_0.$$

Here, $h^l(\mathbf{x}), \Sigma^l(\mathbf{x})$ and $(h')^l(\mathbf{x}), (\Sigma')^l(\mathbf{x})$ are defined as in (13) *and* (12) *with parameters W and W', respectively. For distinct scalar inputs u and v to ReLU, the corresponding diagonal entry of $\Sigma^l + (\Sigma'')^l$ is $(\sigma(u) - \sigma(v))/(u - v) \in [0,1]$. If $u = v$, we use the corresponding entry of Σ^l.*

Following Lemma 3, define

$$\widehat{G}_L^l(\mathbf{x}_i) = \left[\prod_{j=l+1}^{L}\sqrt{\frac{2}{m}}(\Sigma^j(\mathbf{x}_i) + (\Sigma'')^j)\mathbf{W}^j\right]\sqrt{\frac{2}{m}}(\Sigma^l(\mathbf{x}_i) + (\Sigma'')^l). \tag{16}$$

Lemma 4 (Deep perturbation bounds; Lemmas 15, 17 and the proof of Lemma 19 in Li et al. (2026)). *Fix a deterministic radius $R \ge 1$. Suppose Assumptions 1 and 2 hold and $m \gtrsim L^{10}\log(nL/\delta)(\log m)^4 R^2$. Then, with probability at least $1-\delta$, for every $W, W' \in \mathcal{B}_R(W(0))$, $i \in [n]$, and $l \in [L]$, with $\widehat{G}_L^l$ constructed from W, W' as in* (16)*,*

$$\|h^l(\mathbf{x}_i) - h_0^l(\mathbf{x}_i)\|_2 \lesssim L^2\sqrt{\frac{\log m}{m}}R \quad , \quad \|\Sigma^l(\mathbf{x}_i) - \Sigma_0^l(\mathbf{x}_i)\|_0 \lesssim L^{4/3}(\log m)^{1/3}(mR)^{2/3}.$$

Moreover,

$$\|\mathbf{a}^\top(\widehat{G}_L^l(\mathbf{x}_i) - G_{L,0}^l(\mathbf{x}_i))\|_2 \lesssim \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}},$$
$$\|\mathbf{a}^\top(G_L^l(\mathbf{x}_i) - G_{L,0}^l(\mathbf{x}_i))\|_2 \lesssim \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}}.$$

These estimates apply to $\mathcal{B}_R(W(0))$ because $\|W^l - W^l(0)\|_2 \le \|W - W(0)\|_F$. The paired initialization is the same as in Li et al. (2026); scalar concentration in the last layer uses the $m/2$ independent pairs.

Lemma 5 (Adapted from Lemma 16 in Li et al. (2026)). *Suppose Assumptions 1 and 2 hold. Let $R \ge 1$ be deterministic and assume*

$$m \gtrsim L^{11}d(\log m)^5\log(L/\delta)R^2.$$

Then, with probability at least $1-\delta$, for every $W \in \mathcal{B}_R(W(0))$ and $l \in [L]$, we have

$$\sup_{\mathbf{x}\in\mathcal{X}}\|h^l(\mathbf{x}) - h_0^l(\mathbf{x})\|_2 \lesssim L^2\sqrt{\frac{\log m}{m}}R.$$

Proof. We follow the covering argument in Lemma 16 of Li et al. (2026), with a fixed deterministic radius R. Our Frobenius ball is contained in their ball defined by the spectral norm in each layer. The width condition implies $m \gtrsim d + \log(L/\delta)$ and $R \lesssim \sqrt{m}$. Gaussian matrix norm bounds therefore give, with probability at least $1 - \delta/2$, $\max_l \|W^l(0)\|_2 \le C\sqrt{m}$. For the paired final layer, its spectral norm is $\sqrt{2}$ times that of the matrix formed by its first $m/2$ rows. On this event, the 1-Lipschitz property of ReLU gives an absolute constant $A > 1$ such that

$$\|h^l(\mathbf{x}) - h^l(\mathbf{x}')\|_2 \le A^L \|\mathbf{x} - \mathbf{x}'\|_2, \qquad W \in \mathcal{B}_R(W(0)), \quad l \in [L].$$

Choose a deterministic ε-net D of $\mathbb{S}^{d-1}$ with $\varepsilon = A^{-L} m^{-1/2}$. Its cardinality satisfies $|D| \le (1 + 2/\varepsilon)^d$. Since $L \ge 2$, $d \ge 1$ and $m \ge 2$, we have

$$\log(2|D|L/\delta) \lesssim d(L + \log m) + \log(L/\delta) \lesssim Ld \log m \, \log(L/\delta).$$

Consequently, the assumed width condition implies

$$m \gtrsim L^{10} (\log m)^4 R^2 \log(2|D|L/\delta).$$

Apply Lemma 4 to the finite set D with failure probability $\delta/2$. On the intersection of the two events, choose $\mathbf{x}' \in D$ with $\|\mathbf{x} - \mathbf{x}'\|_2 \le \varepsilon$. Uniformly over $W \in \mathcal{B}_R(W(0))$ and $l \in [L]$, the triangle inequality gives

$$\|h^l(\mathbf{x}) - h_0^l(\mathbf{x})\|_2 \le \frac{2}{\sqrt{m}} + CL^2 \sqrt{\frac{\log m}{m}} R \lesssim L^2 \sqrt{\frac{\log m}{m}} R,$$

where the last inequality uses $R \ge 1$. A union bound completes the proof. The radius R may depend on the fixed learning rate, margin and training budget, but not on the realized sample or initialization, and the width condition must hold for that radius. □

A.3 A Smooth Loss Generalization Bound

Definition 1 (Rademacher Complexity). Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$ and let $S_X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Define

$$S_X^n := \{(\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n) : \widetilde{\mathbf{x}}_i \in S_X, \; i \in [n]\},$$

where repetitions are allowed. For any $\widetilde{S} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n) \in S_X^n$, define

$$\widehat{\mathfrak{R}}_{\widetilde{S}}(\mathcal{F}) := \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\widetilde{\mathbf{x}}_i) \right],$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher variables. Taking the supremum over samples drawn from S_X, we define

$$\mathfrak{R}_{S_X, n}(\mathcal{F}) := \sup_{\widetilde{S} \in S_X^n} \widehat{\mathfrak{R}}_{\widetilde{S}}(\mathcal{F}).$$

This definition allows repeated observations, as in Lei et al. (2026).

Lemma 6 (Lemma 27 in Lei et al. (2026); see also Srebro et al. (2010))**.** *Let $\mathcal{F}$ be a hypothesis class that is independent of the training sample S, and suppose ℓ is nonnegative and H-smooth. Let*

$$B_\ell := \sup_{f \in \mathcal{F}, \, (\mathbf{x}, y) \in \mathcal{X} \times \{-1, +1\}} \ell(y f(\mathbf{x})) < \infty.$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the i.i.d. sample S, simultaneously for every $f \in \mathcal{F}$,

$$\mathcal{L}(f) - \mathcal{L}_S(f) \lesssim \sqrt{\mathcal{L}_S(f)} \left(\sqrt{H \log^3 n} \, \mathfrak{R}_{S_X, n}(\mathcal{F}) + \sqrt{\frac{B_\ell \log(2/\delta)}{n}} \right) + H \log^3 n \, \mathfrak{R}_{S_X, n}^2(\mathcal{F}) + \frac{B_\ell \log(2/\delta)}{n}.$$

B MAIN LEMMAS

We first recall an initialization bound and derive the linearization and gradient perturbation estimates. We then establish trajectory control and stable-phase convergence, followed by a local Rademacher complexity bound for the population risk analysis.

B.1 INITIALIZATION AND LINEARIZATION BOUNDS

The next lemma combines the initialization identities and gradient bound in Lemmas 7–8 of Li et al. (2026). We give a direct proof, keeping track of the dependence within each neuron pair and the probability over all sample points.

Lemma 7. *Suppose Assumptions 1 and 2 hold. If the width satisfies $m \gtrsim L^2 \log(nL/\delta)$, then with probability at least $1-\delta$, for all $i \in [n]$, we have*

$$\|\nabla f_{W(0)}(\mathbf{x}_i)\|_F \le \sqrt{2}.$$

Proof. For hidden layers $l \in [L-1]$,

$$\left(\frac{\partial f_{W(0)}(\mathbf{x}_i)}{\partial W^l(0)}\right)^\top = h_0^{l-1}(\mathbf{x}_i)\mathbf{a}^\top \sqrt{\frac{2}{m}}\Sigma_0^L(\mathbf{x}_i)W^L(0)\cdots\sqrt{\frac{2}{m}}\Sigma_0^l(\mathbf{x}_i).$$

By Assumption 1, $a_{r+m/2} = -a_r$ and $w^L_{r+m/2}(0) = w^L_r(0)$. Hence the two neurons in each pair have the same activation pattern and opposite output weights, which gives

$$\mathbf{a}^\top\Sigma_0^L(\mathbf{x}_i)W^L(0) = 0 \qquad \text{and} \qquad f_{W(0)}(\mathbf{x}) = 0 \quad \text{for every } \mathbf{x}\in\mathcal{X}. \tag{17}$$

Therefore,

$$\frac{\partial f_{W(0)}(\mathbf{x}_i)}{\partial W^l(0)} = 0, \qquad l \in [L-1].$$

For the last layer,

$$\frac{\partial f_{W(0)}(\mathbf{x}_i)}{\partial W^L(0)} = \sqrt{\frac{2}{m}}\big(\Sigma_0^L(\mathbf{x}_i)\mathbf{a}\big)\big(h_0^{L-1}(\mathbf{x}_i)\big)^\top.$$

Let

$$I_r := \mathbb{I}\big\{\langle w_r^L(0), h_0^{L-1}(\mathbf{x}_i)\rangle \ge 0\big\}, \qquad r\in[m].$$

On the event in Lemma 1, $h_0^{L-1}(\mathbf{x}_i) \neq 0$ for every i. Conditioned on the first $L-1$ layers on this event, $I_1,\dots,I_{m/2}$ are independent Bernoulli$(1/2)$ variables for each fixed i, while $I_{r+m/2} = I_r$. Thus

$$\|\Sigma_0^L(\mathbf{x}_i)\mathbf{a}\|_2^2 = 2\sum_{r=1}^{m/2} I_r.$$

Hoeffding's inequality and a union bound over $i\in[n]$ imply, after replacing δ by $\delta/2$, that

$$\|\Sigma_0^L(\mathbf{x}_i)\mathbf{a}\|_2^2 \le \frac{3m}{4}, \qquad i\in[n],$$

with conditional probability at least $1-\delta/2$, provided $m \gtrsim \log(n/\delta)$. Using Lemma 1 with failure probability $\delta/2$ gives $\|h_0^{L-1}(\mathbf{x}_i)\|_2^2 \le 4/3$ simultaneously for all i. Therefore,

$$\left\|\frac{\partial f_{W(0)}(\mathbf{x}_i)}{\partial W^L(0)}\right\|_F^2 = \frac{2}{m}\|\Sigma_0^L(\mathbf{x}_i)\mathbf{a}\|_2^2\|h_0^{L-1}(\mathbf{x}_i)\|_2^2 \le 2,$$

and the claim follows. □

Lemma 8 (Linearization Error Bound)**.** *Fix a deterministic radius $R \ge 1$. Suppose Assumptions 1 and 2 hold. For $W, W' \in \mathcal{B}_R(W(0))$ and $i\in[n]$, define*

$$\xi_i(W', W) := f_{W'}(\mathbf{x}_i) - f_W(\mathbf{x}_i) - \langle \nabla f_W(\mathbf{x}_i), W'-W\rangle.$$

If $m \gtrsim L^{10}\log(nL/\delta)(\log m)^4R^2$, then, with probability at least $1-\delta$, simultaneously for all $W, W' \in \mathcal{B}_R(W(0))$ and $i\in[n]$,

$$|\xi_i(W', W)| \le C\frac{L^{13/6}(\log m)^{2/3}R^{1/3}}{m^{1/6}}\|W'-W\|_F.$$

Proof. Apply Lemmas 1, 2, and 4 with failure probability $\delta/3$ each, and work on their intersection. By Lemma 3 and (16),

$$f_{W'}(\mathbf{x}_i) - f_W(\mathbf{x}_i) = \sum_{l=1}^{L} \mathbf{a}^\top \widehat{G}_L^l(\mathbf{x}_i)((W')^l - W^l)(h')^{l-1}(\mathbf{x}_i).$$

The gradient with respect to the l-th layer is

$$\frac{\partial f_W(\mathbf{x}_i)}{\partial W^l} = (G_L^l(\mathbf{x}_i))^\top \mathbf{a}(h^{l-1}(\mathbf{x}_i))^\top.$$

Thus,

$$\begin{aligned}\langle \nabla f_W(\mathbf{x}_i), W' - W\rangle &= \sum_{l=1}^{L} \left\langle \frac{\partial f_W(\mathbf{x}_i)}{\partial W^l}, (W')^l - W^l \right\rangle \\ &= \sum_{l=1}^{L} \left\langle (G_L^l(\mathbf{x}_i))^\top \mathbf{a}(h^{l-1}(\mathbf{x}_i))^\top, (W')^l - W^l \right\rangle \\ &= \sum_{l=1}^{L} \mathbf{a}^\top G_L^l(\mathbf{x}_i)((W')^l - W^l)h^{l-1}(\mathbf{x}_i).\end{aligned}$$

Writing $\xi_i = \xi_i(W', W)$, we have

$$\begin{aligned}|\xi_i| &= \left| \sum_{l=1}^{L} \mathbf{a}^\top \widehat{G}_L^l(\mathbf{x}_i)((W')^l - W^l)(h')^{l-1}(\mathbf{x}_i) - \sum_{l=1}^{L} \mathbf{a}^\top G_L^l(\mathbf{x}_i)((W')^l - W^l)h^{l-1}(\mathbf{x}_i) \right| \\ &\le \left| \sum_{l=1}^{L} \mathbf{a}^\top (\widehat{G}_L^l(\mathbf{x}_i) - G_L^l(\mathbf{x}_i))((W')^l - W^l)(h')^{l-1}(\mathbf{x}_i) \right| \\ &\quad + \left| \sum_{l=1}^{L} \mathbf{a}^\top G_L^l(\mathbf{x}_i)((W')^l - W^l)((h')^{l-1}(\mathbf{x}_i) - h^{l-1}(\mathbf{x}_i)) \right|.\end{aligned}$$

Denote the two terms on the right-hand side by T_1 and T_2, respectively. For T_1, Lemma 4 and the triangle inequality give

$$\begin{aligned}\|\mathbf{a}^\top(\widehat{G}_L^l(\mathbf{x}_i) - G_L^l(\mathbf{x}_i))\|_2 &\le \|\mathbf{a}^\top(\widehat{G}_L^l(\mathbf{x}_i) - G_{L,0}^l(\mathbf{x}_i))\|_2 + \|\mathbf{a}^\top(G_L^l(\mathbf{x}_i) - G_{L,0}^l(\mathbf{x}_i))\|_2 \\ &\lesssim \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}}.\end{aligned}$$

For $l = 1$, $h_0^0(\mathbf{x}_i) = \mathbf{x}_i$ has unit norm by Assumption 2. For $l \ge 2$, Lemma 1 gives $\|h_0^{l-1}(\mathbf{x}_i)\|_2 \lesssim 1$. Together with the first perturbation bound in Lemma 4, this gives $\|(h')^{l-1}(\mathbf{x}_i)\|_2 \lesssim 1$ for all $l \in [L]$. Hence,

$$T_1 \lesssim \sum_{l=1}^{L} \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}} \|(W')^l - W^l\|_F.$$

For T_2, apply Lemma 2 with $\widehat{W}^j = W^j - W^j(0)$ and $\widehat{\Sigma}^j = \Sigma^j - \Sigma_0^j$. The gate sparsity bound in Lemma 4 gives the required support bound, while the width condition implies both $R \lesssim \sqrt{m}/(L^2\sqrt{\log m})$ and $s \lesssim m/(L^2 \log m)$. It follows that $\|G_L^l(\mathbf{x}_i)\|_2 \lesssim L\sqrt{\frac{\log m}{m}}$. Since $\|\mathbf{a}\|_2 = \sqrt{m}$,

$$\|\mathbf{a}^\top G_L^l(\mathbf{x}_i)\|_2 \lesssim L\sqrt{\log m}. \tag{18}$$

Since $W, W' \in \mathcal{B}_R(W(0))$, the first result of Lemma 4 and the triangle inequality give

$$\|(h')^{l-1}(\mathbf{x}_i) - h^{l-1}(\mathbf{x}_i)\|_2 \le \|(h')^{l-1}(\mathbf{x}_i) - h_0^{l-1}(\mathbf{x}_i)\|_2 + \|h^{l-1}(\mathbf{x}_i) - h_0^{l-1}(\mathbf{x}_i)\|_2 \lesssim L^2\sqrt{\frac{\log m}{m}}R.$$

Using these bounds in T_2 gives

$$T_2 \lesssim \sum_{l=1}^{L} (L\sqrt{\log m})\|(W')^l - W^l\|_2 \left(L^2\sqrt{\frac{\log m}{m}}R\right) \lesssim \sum_{l=1}^{L} \frac{L^3(\log m)R}{\sqrt{m}} \|(W')^l - W^l\|_F.$$

The width condition implies $m \gtrsim L^4(\log m)R^2$, so the $m^{-1/2}$ term in T_2 is bounded by the $m^{-1/6}$ term in T_1, up to an absolute constant. The Cauchy–Schwarz inequality $\sum_{l=1}^{L} \|(W')^l - W^l\|_F \leq \sqrt{L}\|W' - W\|_F$ now yields

$$|\xi_i| \lesssim \sqrt{L} \cdot \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}}\|W' - W\|_F = \frac{L^{13/6}(\log m)^{2/3}R^{1/3}}{m^{1/6}}\|W' - W\|_F.$$

□

Lemma 9. *Fix a deterministic radius $R \geq 1$. Suppose Assumptions 1 and 2 hold. If*

$$m \gtrsim L^{10}\log(nL/\delta)(\log m)^4 R^2,$$

then with probability at least $1-\delta$, for all $W \in \mathcal{B}_R(W(0))$ and $i \in [n]$, we have

$$\|\nabla f_W(\mathbf{x}_i) - \nabla f_{W(0)}(\mathbf{x}_i)\|_F \leq C\frac{L^{13/6}(\log m)^{2/3}R^{1/3}}{m^{1/6}}.$$

Proof. Use the same intersection of initialization and perturbation events as in the proof of Lemma 8, with total failure probability δ. For any layer $l \in [L]$, the gradient is given by

$$\frac{\partial f_W(\mathbf{x}_i)}{\partial W^l} = (G_L^l(\mathbf{x}_i))^\top \mathbf{a}(h^{l-1}(\mathbf{x}_i))^\top.$$

For every $l \in [L]$,

$$\begin{aligned}\left\|\frac{\partial f_W(\mathbf{x}_i)}{\partial W^l} - \frac{\partial f_{W(0)}(\mathbf{x}_i)}{\partial W^l(0)}\right\|_F &= \left\|(G_L^l(\mathbf{x}_i))^\top \mathbf{a}(h^{l-1}(\mathbf{x}_i))^\top - (G_{L,0}^l(\mathbf{x}_i))^\top \mathbf{a}(h_0^{l-1}(\mathbf{x}_i))^\top\right\|_F \\ &\leq \|(G_L^l(\mathbf{x}_i))^\top \mathbf{a}\|_2 \|h^{l-1}(\mathbf{x}_i) - h_0^{l-1}(\mathbf{x}_i)\|_2 \\ &\quad + \|(G_L^l(\mathbf{x}_i) - G_{L,0}^l(\mathbf{x}_i))^\top \mathbf{a}\|_2 \|h_0^{l-1}(\mathbf{x}_i)\|_2.\end{aligned}$$

The second bound in Lemma 4 gives

$$\|(G_L^l(\mathbf{x}_i) - G_{L,0}^l(\mathbf{x}_i))^\top \mathbf{a}\|_2 \lesssim \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}}.$$

The first bound in Lemma 4 gives

$$\|h^{l-1}(\mathbf{x}_i) - h_0^{l-1}(\mathbf{x}_i)\|_2 \lesssim L^2\sqrt{\frac{\log m}{m}}R.$$

For $l = 1$, $\|h_0^0(\mathbf{x}_i)\|_2 = 1$ by Assumption 2; for $l \geq 2$, Lemma 1 gives $\|h_0^{l-1}(\mathbf{x}_i)\|_2 \lesssim 1$. The remaining factor is bounded by (18). Combining these estimates,

$$\begin{aligned}\left\|\frac{\partial f_W(\mathbf{x}_i)}{\partial W^l} - \frac{\partial f_{W(0)}(\mathbf{x}_i)}{\partial W^l(0)}\right\|_F &\lesssim (L\sqrt{\log m})\left(L^2\sqrt{\frac{\log m}{m}}R\right) + \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}} \\ &\lesssim \frac{L^{5/3}(\log m)^{2/3}R^{1/3}}{m^{1/6}},\end{aligned}$$

where the last inequality follows from $m \gtrsim L^4(\log m)R^2$. Summing over the layers gives

$$\begin{aligned}\|\nabla f_W(\mathbf{x}_i) - \nabla f_{W(0)}(\mathbf{x}_i)\|_F &= \sqrt{\sum_{l=1}^{L}\left\|\frac{\partial f_W(\mathbf{x}_i)}{\partial W^l} - \frac{\partial f_{W(0)}(\mathbf{x}_i)}{\partial W^l(0)}\right\|_F^2} \\ &\lesssim \frac{L^{13/6}(\log m)^{2/3}R^{1/3}}{m^{1/6}}.\end{aligned}$$

□

B.2 Optimization Analysis

We first use the comparator recursion to control the full trajectory and the cumulative gradient potential. We then prove the loss bound conditional on entry into the stable phase, and use the potential estimate to bound the entry time.

Good Initialization Event. For any deterministic radius $R \geq 1$, let $\mathcal{E}_R$ denote the intersection of the events in Lemmas 7, 8, and 9, each with failure probability $\delta/3$ and instantiated at radius R (the event controlling the gradient at initialization itself does not depend on R). On $\mathcal{E}_R$,

$$E(m, R) := C\frac{L^{13/6}(\log m)^{2/3}R^{1/3}}{m^{1/6}}$$

controls both the gradient perturbation and the coefficient of the linearization error uniformly on $\mathcal{B}_R(W(0))$. If $m \gtrsim L^{10}\log(nL/\delta)(\log m)^4R^2$, then, for every fixed sample S, a union bound gives $\mathbb{P}_{W(0)}(\mathcal{E}_R) \geq 1 - \delta$. Whenever a larger radius is needed below, $\mathcal{E}_R$ is instantiated at that deterministic radius.

The optimization argument below adapts the comparator approach of Wu et al. (2024) to the deep linearization bound above. The comparator is $u = u_1 + u_2$: the reference point u_1 has small training loss, while the displacement u_2 supplies a negative inner product term that offsets the positive gradient norm squared term in the distance recursion. Only u_1 and the current iterate enter the local linearization bound; the full comparator u need not lie in the local ball.

Lemma 10 (Split Optimization)**.** *Fix a deterministic radius $R \geq 1$. Suppose Assumption 3 holds and the good initialization event $\mathcal{E}_R$ holds. Let $\eta > 0$ and let $T > 1$ be an integer. Define the comparator $u := u_1 + u_2$, where $u_1 = W(0) + \theta W_*$ and $u_2 = \rho W_*$, with $0 < \theta \leq R$ and $\rho := \frac{\eta(\sqrt{2}+\frac{\gamma}{2})^2}{\gamma}$. Assume that for some $t \in [T]$,*

$$\|W(k) - W(0)\|_F \leq R$$

for all $0 \leq k < t$. Furthermore, assume that the width m satisfies the overparameterization condition

$$m \gtrsim L^{13}(\log m)^4 \max\left\{\frac{R^2}{\gamma^6}, R^8\right\}.$$

Then

$$\frac{\|W(t) - u\|_F^2}{2\eta t} + \frac{1}{2t}\sum_{k=0}^{t-1}\mathcal{L}_S(W(k)) \leq e^{-\frac{\theta\gamma}{2}} + \frac{\|W(0) - u\|_F^2}{2\eta t}.$$

Proof. The width condition gives $E(m, R) \leq \gamma/2$ and $E(m, R)(R + \theta) \leq 1/2$. For $0 \leq k < t$,

$$\|W(k+1) - u\|_F^2 = \|W(k) - u\|_F^2 + \underbrace{2\eta\left\langle\nabla\mathcal{L}_S(W(k)), u_1 - W(k)\right\rangle}_{\text{Term A}}$$
$$+ \underbrace{\eta\left(2\left\langle\nabla\mathcal{L}_S(W(k)), u_2\right\rangle + \eta\|\nabla\mathcal{L}_S(W(k))\|_F^2\right)}_{\text{Term B}}. \tag{19}$$

We first consider Term B. On $\mathcal{E}_R$, for all $i \in [n]$,

$$\|\nabla f_{W(k)}(\mathbf{x}_i)\|_F \leq \|\nabla f_{W(0)}(\mathbf{x}_i)\|_F + E(m, R) \leq \sqrt{2} + \frac{\gamma}{2}.$$

Therefore,

$$\|\nabla\mathcal{L}_S(W(k))\|_F = \left\|\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i f_{W(k)}(\mathbf{x}_i))y_i\nabla f_{W(k)}(\mathbf{x}_i)\right\|_F$$
$$\leq (\sqrt{2} + \frac{\gamma}{2})\frac{1}{n}\sum_{i=1}^{n}|\ell'(y_i f_{W(k)}(\mathbf{x}_i))|.$$

Since $|\ell'(z)| < 1$, $\mathcal{P}_S(W(k)) \leq 1$, and hence

$$\eta\|\nabla\mathcal{L}_S(W(k))\|_F^2 \leq \eta(\sqrt{2} + \frac{\gamma}{2})^2\mathcal{P}_S(W(k)).$$

For the first term in Term B, the gradient perturbation bound and $E(m,R) \le \gamma/2$ imply

$$y_i \left\langle \nabla f_{W(k)}(\mathbf{x}_i), W_* \right\rangle \ge y_i \left\langle \nabla f_{W(0)}(\mathbf{x}_i), W_* \right\rangle - \|\nabla f_{W(k)}(\mathbf{x}_i) - \nabla f_{W(0)}(\mathbf{x}_i)\|_F \|W_*\|_F \ge \gamma - \frac{\gamma}{2} = \frac{\gamma}{2}.$$

Moreover,

$$2\left\langle \nabla \mathcal{L}_S(W(k)), u_2 \right\rangle = \frac{2\rho}{n} \sum_{i=1}^{n} \ell'(y_i f_{W(k)}(\mathbf{x}_i)) y_i \left\langle \nabla f_{W(k)}(\mathbf{x}_i), W_* \right\rangle.$$

Since $\ell'(z) < 0$,

$$2\left\langle \nabla \mathcal{L}_S(W(k)), u_2 \right\rangle \le \frac{2\rho}{n} \sum_{i=1}^{n} \ell'(\cdot) \frac{\gamma}{2} = -\rho\gamma \frac{1}{n} \sum_{i=1}^{n} |\ell'(\cdot)| = -\rho\gamma \mathcal{P}_{\mathcal{S}}(W(k)).$$

By the definition of ρ,

$$\text{Term B} \le \eta(-\rho\gamma \mathcal{P}_{\mathcal{S}}(W(k)) + \eta(\sqrt{2} + \frac{\gamma}{2})^2 \mathcal{P}_{\mathcal{S}}(W(k))) = 0.$$

For Term A, the convexity of ℓ gives

$$\begin{aligned} \mathcal{L}_S(u_1) - \mathcal{L}_S(W(k)) &= \frac{1}{n} \sum_{i=1}^{n} (\ell(y_i f_{u_1}(\mathbf{x}_i)) - \ell(y_i f_{W(k)}(\mathbf{x}_i))) \\ &\ge \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i f_{W(k)}(\mathbf{x}_i))(y_i f_{u_1}(\mathbf{x}_i) - y_i f_{W(k)}(\mathbf{x}_i)). \end{aligned}$$

Using the linearization of $f_{u_1}(\mathbf{x}_i)$ at $W(k)$ gives

$$\mathcal{L}_S(u_1) - \mathcal{L}_S(W(k)) \ge \langle \nabla \mathcal{L}_S(W(k)), u_1 - W(k) \rangle + \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i f_{W(k)}(\mathbf{x}_i)) y_i \xi_i(u_1, W(k)).$$

It follows that

$$\text{Term A} \le 2\eta(\mathcal{L}_S(u_1) - \mathcal{L}_S(W(k))) + 2\eta \frac{1}{n} \sum_{i=1}^{n} |\ell'(y_i f_{W(k)}(\mathbf{x}_i))||\xi_i(u_1, W(k))|.$$

Since $u_1, W(k) \in \mathcal{B}_R(W(0))$, combining the linearization bound with the triangle inequality gives $|\xi_i(u_1, W(k))| \le E(m,R)(\theta + R) \le 1/2$. Hence,

$$\begin{aligned} 2\eta \frac{1}{n} \sum |\ell'(y_i f_{W(k)}(\mathbf{x}_i))||\xi_i(u_1, W(k))| &\le 2\eta \mathcal{P}_{\mathcal{S}}(W(k)) E(m,R)(\theta + R) \\ &\le \eta \mathcal{P}_{\mathcal{S}}(W(k)) \le \eta \mathcal{L}_S(W(k)). \end{aligned}$$

The last inequality follows from $|\ell'(z)| \le \ell(z)$. Hence,

$$\text{Term A} \le 2\eta \mathcal{L}_S(u_1) - \eta \mathcal{L}_S(W(k)).$$

We next bound $\mathcal{L}_S(u_1)$. By linearizing at $W(0)$,

$$y_i f_{u_1}(\mathbf{x}_i) = y_i \left(f_{W(0)}(\mathbf{x}_i) + \langle \nabla f_{W(0)}(\mathbf{x}_i), u_1 - W(0) \rangle + \xi_i(u_1, W(0)) \right).$$

By (17), Assumption 3, and $E(m,R) \le \gamma/2$, for all $i \in [n]$,

$$y_i f_{u_1}(\mathbf{x}_i) \ge \theta\gamma - E(m,R)\theta \ge \theta\gamma - \frac{\gamma}{2}\theta = \frac{\theta\gamma}{2}.$$

Therefore,

$$\mathcal{L}_S(u_1) \le \log(1 + e^{-\theta\gamma/2}) \le e^{-\theta\gamma/2}.$$

Combining the above bounds with (19) gives

$$\|W(k+1) - u\|_F^2 \le \|W(k) - u\|_F^2 + 2\eta e^{-\frac{\theta\gamma}{2}} - \eta \mathcal{L}_S(W(k)).$$

Telescoping the sum from $k = 0$ to $t - 1$ completes the proof. □

Relation to previous analyses. In the two-layer proof of Lemma 27 in Wu et al. (2024), 1-homogeneity gives

$$\langle \nabla f_W(\mathbf{x}_i), u_1 - W \rangle = \langle \nabla f_W(\mathbf{x}_i), u_1 \rangle - f_W(\mathbf{x}_i).$$

This identity is used in the convexity comparison. For our deep predictor, the corresponding identity at differentiable parameters has $Lf_W(\mathbf{x}_i)$ in place of $f_W(\mathbf{x}_i)$, so that step cannot be used unchanged. Our proof instead compares the actual outputs at u_1 and $W(k)$ and absorbs the loss-weighted linearization remainder into the current training loss.

The restricted-step analysis of Li et al. (2026) uses two learning rate conditions. In Lemma 21, $\eta \le 4/(5L)$ makes the derivative-difference contribution nonpositive after Jensen's inequality, since its coefficient is $5\eta^2 L - 4\eta$. In the proof of their Theorem 1, the additional restriction $\eta \le 1/(20L\widetilde{F}_S(\overline{W}))$ controls the accumulated reference-model contribution, where $\widetilde{F}_S(\overline{W})$ denotes the average absolute loss derivative at the reference model. Our split comparator instead controls the gradient norm squared term through u_2, while the reference loss and the deep linearization remainder are controlled separately. This yields the trajectory recursion for arbitrary constant learning rates under the conditions of Lemma 10.

Lemma 11 (Full trajectory and average loss bounds)**.** *Fix deterministic parameters $\eta > 0$, $\gamma \in (0, 1]$, and an integer $T > 1$. Suppose Assumption 3 holds with margin γ, and define*

$$\theta_T := \frac{2\log(1 + 2\eta T)}{\gamma}, \qquad \rho := \frac{\eta(\sqrt{2} + \gamma/2)^2}{\gamma}, \qquad R_0 := 2(\theta_T + \rho) + 1.$$

Suppose $\mathcal{E}_{R_0}$ holds and

$$m \gtrsim L^{13}(\log m)^4 \max\left\{ R_0^8, \frac{R_0^2}{\gamma^6} \right\}.$$

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

$$\|W(t) - W(0)\|_F \le R_0$$

and

$$\frac{1}{t}\sum_{k=0}^{t-1} \mathcal{L}_S(W(k)) \le \frac{1}{\eta T} + \frac{(\theta_T + \rho)^2}{\eta t}. \tag{20}$$

Proof. We argue by induction. The claim is trivial at $t = 0$. Assume that $\|W(k) - W(0)\|_F \le R_0$ for all $k < t$. The displayed width condition is precisely sufficient for Lemma 10 at radius R_0. Using Lemma 10 with $u_1 = W(0) + \theta_T W_*$ and $u_2 = \rho W_*$ gives

$$\frac{\|W(t) - u\|_F^2}{2\eta t} + \frac{1}{2t}\sum_{k=0}^{t-1} \mathcal{L}_S(W(k)) \le e^{-\theta_T \gamma/2} + \frac{\|W(0) - u\|_F^2}{2\eta t}.$$

Since $e^{-\theta_T\gamma/2} = (1 + 2\eta T)^{-1} \le (2\eta T)^{-1}$ and $t \le T$,

$$\|W(t) - u\|_F^2 \le \frac{t}{T} + \|W(0) - u\|_F^2 \le 1 + (\theta_T + \rho)^2.$$

Hence $\|W(t) - u\|_F \le 1 + \theta_T + \rho$, and therefore

$$\|W(t) - W(0)\|_F \le 1 + 2(\theta_T + \rho) = R_0.$$

This closes the induction. Dropping the nonnegative distance term in the split inequality and using $2e^{-\theta_T\gamma/2} \le 1/(\eta T)$ yields (20). □

Lemma 12 (Gradient Potential Bound)**.** *Under the assumptions of Lemma 11, for every $t \in [T]$,*

$$\frac{1}{t}\sum_{k=0}^{t-1} \mathcal{P}_S(W(k)) \le \frac{2R_0}{\gamma\eta t},$$

where R_0 is defined in Lemma 11.

Proof. On $\mathcal{E}_{R_0}$, the width condition gives $\|\nabla f_{W(k)}(\mathbf{x}_i) - \nabla f_{W(0)}(\mathbf{x}_i)\|_F \le E(m, R_0) \le \gamma/2$. Together with Assumption 3, this implies

$$y_i \left\langle W_*, \nabla f_{W(k)}(\mathbf{x}_i) \right\rangle \ge y_i \left\langle W_*, \nabla f_{W(0)}(\mathbf{x}_i) \right\rangle - E(m, R_0) \ge \gamma - \frac{\gamma}{2} = \frac{\gamma}{2}.$$

Since $\ell'(z) < 0$,

$$\langle W_*, \nabla \mathcal{L}_S(W(k)) \rangle = \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i f_{W(k)}(\mathbf{x}_i)) y_i \left\langle W_*, \nabla f_{W(k)}(\mathbf{x}_i) \right\rangle \le -\frac{\gamma}{2} \mathcal{P}_{\mathcal{S}}(W(k)).$$

The GD update further gives

$$\begin{aligned} \langle W_*, W(k+1) \rangle &= \langle W_*, W(k) \rangle - \eta \langle W_*, \nabla \mathcal{L}_S(W(k)) \rangle \\ &\ge \langle W_*, W(k) \rangle + \frac{\eta\gamma}{2} \mathcal{P}_S(W(k)). \end{aligned}$$

Summing from $k = 0$ to $t - 1$ yields

$$\begin{aligned} \frac{\eta\gamma}{2} \sum_{k=0}^{t-1} \mathcal{P}_S(W(k)) &\le \langle W_*, W(t) \rangle - \langle W_*, W(0) \rangle \\ &= \langle W_*, W(t) - W(0) \rangle . \end{aligned}$$

By Cauchy–Schwarz and Lemma 11,

$$\langle W_*, W(t) - W(0) \rangle \le \|W_*\|_F \|W(t) - W(0)\|_F \le R_0.$$

$$\sum_{k=0}^{t-1} \mathcal{P}_{\mathcal{S}}(W(k)) \le \frac{2R_0}{\eta\gamma}.$$

Dividing by t proves the claim. □

Lemma 13 (Gradient Potential Properties)**.** *Suppose Assumption 3 holds and $\mathcal{E}_R$ holds for some deterministic radius $R \ge 1$. If $E(m, R) \le \gamma/2$, then for every $W \in \mathcal{B}_R(W(0))$,*

$$\mathcal{P}_S(W) \le \mathcal{L}_S(W), \qquad \frac{\gamma}{2} \mathcal{P}_S(W) \le \|\nabla \mathcal{L}_S(W)\|_F \le \left(\sqrt{2} + \frac{\gamma}{2}\right) \mathcal{P}_S(W).$$

Proof. For logistic loss, $\ell(z) = \log(1 + e^{-z}) \ge e^{-z}/(1 + e^{-z}) = |\ell'(z)|$, which proves $\mathcal{P}_S(W) \le \mathcal{L}_S(W)$. On $\mathcal{E}_R$,

$$\|\nabla f_W(\mathbf{x}_i)\|_F \le \|\nabla f_{W(0)}(\mathbf{x}_i)\|_F + E(m, R) \le \sqrt{2} + \frac{\gamma}{2}. \tag{21}$$

Therefore

$$\|\nabla \mathcal{L}_S(W)\|_F \le \left(\sqrt{2} + \frac{\gamma}{2}\right) \mathcal{P}_S(W).$$

Moreover, Assumption 3 and the gradient perturbation bound imply $y_i \langle \nabla f_W(\mathbf{x}_i), W_* \rangle \ge \gamma/2$. Since $\ell'(z) < 0$ and $\|W_*\|_F = 1$,

$$\begin{aligned} \|\nabla \mathcal{L}_S(W)\|_F &\ge |\langle \nabla \mathcal{L}_S(W), W_* \rangle| \\ &= \frac{1}{n} \sum_{i=1}^{n} |\ell'(y_i f_W(\mathbf{x}_i))| \big(y_i \langle \nabla f_W(\mathbf{x}_i), W_* \rangle\big) \ge \frac{\gamma}{2} \mathcal{P}_S(W). \end{aligned}$$

□

Lemma 14 (Stable Phase)**.** *Suppose the assumptions of Lemma 11 hold. In addition, take the absolute constant in its width condition large enough that $E(m, R_0) \le \gamma/8$. If there exists a time $s < T$ such that $\mathcal{L}_S(W(s)) \le \frac{1}{4e^2 \eta (\sqrt{2} + \frac{5\gamma}{8})^2}$, then for all $t \in [s, T-1]$, we have*

$$\mathcal{P}_{\mathcal{S}}(W(t)) \le \frac{1}{8e\eta(\sqrt{2} + \frac{5\gamma}{8})^2},$$

and

$$\mathcal{L}_S(W(t+1)) \le \mathcal{L}_S(W(t)) - \frac{\eta}{2} \|\nabla \mathcal{L}_S(W(t))\|_F^2.$$

Proof. By Lemma 11, all iterates up to time T belong to $\mathcal{B}_{R_0}(W(0))$. Hence the linearization and gradient bounds on $\mathcal{E}_{R_0}$ apply to every update considered below. We prove by induction that the displayed potential bound holds and $\mathcal{L}_S(W(t)) \le \mathcal{L}_S(W(s))$. For $t = s$, Lemma 13, instantiated at radius R_0, implies

$$\mathcal{P}_{\mathcal{S}}(W(s)) \le \mathcal{L}_S(W(s)) \le \frac{1}{4e^2\eta(\sqrt{2}+\frac{5\gamma}{8})^2} < \frac{1}{8e\eta(\sqrt{2}+\frac{5\gamma}{8})^2}.$$

Assume both bounds hold at some $s \le t < T$. Let $\Delta W := -\eta\nabla\mathcal{L}_S(W(t))$, $z_{i,t} := y_i f_{W(t)}(\mathbf{x}_i)$ and $\Delta z_{i,t} := y_i(f_{W(t+1)}(\mathbf{x}_i) - f_{W(t)}(\mathbf{x}_i))$. By the linearization bound on $\mathcal{E}_{R_0}$ and (21),

$$|\Delta z_{i,t}| \le \|\nabla f_{W(t)}(\mathbf{x}_i)\|_F \|\Delta W\|_F + |\xi_i| \le (\sqrt{2} + \frac{\gamma}{2} + E(m, R_0))\|\Delta W\|_F.$$

Since $E(m, R_0) \le \gamma/8$, Lemma 13 and the induction hypothesis give

$$\begin{aligned}|\Delta z_{i,t}| \le (\sqrt{2}+\frac{5\gamma}{8})\|\Delta W\|_F &= \eta(\sqrt{2}+\frac{5\gamma}{8})\|\nabla\mathcal{L}_S(W(t))\|_F \\ &\le \eta(\sqrt{2}+\frac{5\gamma}{8})^2\mathcal{P}_{\mathcal{S}}(W(t)) \\ &\le \eta(\sqrt{2}+\frac{5\gamma}{8})^2\frac{1}{8e\eta(\sqrt{2}+\frac{5\gamma}{8})^2} = \frac{1}{8e} < 1.\end{aligned}$$

Since $|\Delta z_{i,t}| \le 1$, for every point on the segment between $z_{i,t}$ and $z_{i,t+1}$ we have $\ell''(v) \le |\ell'(v)| \le e|\ell'(z_{i,t})|$. Taylor's theorem therefore gives, for each i,

$$\ell(z_{i,t+1}) \le \ell(z_{i,t}) + \ell'(z_{i,t})\Delta z_{i,t} + \frac{e}{2}|\ell'(z_{i,t})|(\Delta z_{i,t})^2.$$

Using $\Delta z_{i,t} = y_i\langle\nabla f_{W(t)}(\mathbf{x}_i), \Delta W\rangle + y_i\xi_i(W(t+1), W(t))$ and averaging over i gives

$$\mathcal{L}_S(W(t+1)) \le \mathcal{L}_S(W(t)) - \eta\|\nabla\mathcal{L}_S(W(t))\|_F^2 + \frac{1}{n}\sum_{i=1}^n |\ell'(z_{i,t})||\xi_i| + \frac{e}{2n}\sum_{i=1}^n |\ell'(z_{i,t})|(\Delta z_{i,t})^2.$$

The lower bound in Lemma 13 and $E(m, R_0) \le \gamma/8$ yield

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n |\ell'(z_{i,t})||\xi_i| &\le \mathcal{P}_{\mathcal{S}}(W(t))E(m, R_0)\|\Delta W\|_F \\ &\le \frac{2}{\gamma}\|\nabla\mathcal{L}_S\|_F \cdot \frac{\gamma}{8}\cdot\eta\|\nabla\mathcal{L}_S\|_F = \frac{\eta}{4}\|\nabla\mathcal{L}_S(W(t))\|_F^2.\end{aligned}$$

The induction hypothesis and the bound on $|\Delta z_{i,t}|$ also give

$$\frac{e}{2n}\sum_{i=1}^n |\ell'(z_{i,t})|(\Delta z_{i,t})^2 \le \frac{e}{2}\mathcal{P}_{\mathcal{S}}(W(t))(\eta(\sqrt{2}+\frac{5\gamma}{8})\|\nabla\mathcal{L}_S(W(t))\|_F)^2 \le \frac{\eta}{16}\|\nabla\mathcal{L}_S(W(t))\|_F^2.$$

Combining the preceding estimates,

$$\begin{aligned}\mathcal{L}_S(W(t+1)) &\le \mathcal{L}_S(W(t)) - \eta\left(1 - \frac{1}{4} - \frac{1}{16}\right)\|\nabla\mathcal{L}_S(W(t))\|_F^2 \\ &\le \mathcal{L}_S(W(t)) - \frac{\eta}{2}\|\nabla\mathcal{L}_S(W(t))\|_F^2.\end{aligned}$$

Finally, $\mathcal{P}_{\mathcal{S}}(W(t+1)) \le \mathcal{L}_S(W(t+1)) \le \mathcal{L}_S(W(s)) \le \frac{1}{4e^2\eta(\sqrt{2}+\frac{5\gamma}{8})^2}$, completing the induction. □

Lemma 15 (Convergence in Stable Phase)**.** *Suppose the assumptions of Lemma 11 hold, with the absolute constant in the width condition large enough that $E(m, R_0) \le \gamma/8$. If there exists $s < T$ such that*

$$\mathcal{L}_S(W(s)) \le \min\left\{\frac{1}{4e^2\eta(\sqrt{2}+5\gamma/8)^2}, \frac{\log 2}{n}\right\},$$

then all sample margins are nonnegative from s to T and, for $1 \le t \le T - s$,

$$\mathcal{L}_S(W(s+t)) \le \left(\mathcal{L}_S(W(s))^{-1} + \frac{\eta\gamma^2 t}{32}\right)^{-1} \le \frac{32}{\eta\gamma^2 t}.$$

All these iterates remain in $\mathcal{B}_{R_0}(W(0))$.

Proof. Lemma 14 gives monotonicity. Since individual losses are nonnegative, $\mathcal{L}_S(W(s+k)) \le \log 2/n$ implies $y_i f_{W(s+k)}(\mathbf{x}_i) \ge 0$ for every i. For $z \ge 0$, $\ell(z) \le 2|\ell'(z)|$. Thus, writing $r_k = \mathcal{L}_S(W(s+k))$, Lemma 13 gives

$$\|\nabla \mathcal{L}_S(W(s+k))\|_F \ge \frac{\gamma}{2}\mathcal{P}_S(W(s+k)) \ge \frac{\gamma}{4} r_k.$$

Combined with Lemma 14, this yields

$$r_{k+1} \le r_k - \frac{\eta\gamma^2}{32} r_k^2, \qquad \frac{1}{r_{k+1}} - \frac{1}{r_k} = \frac{r_k - r_{k+1}}{r_k r_{k+1}} \ge \frac{\eta\gamma^2}{32}.$$

Here $0 < r_{k+1} \le r_k$, since logistic loss is positive. Summing proves the loss bound. The trajectory claim follows directly from Lemma 11. □

Lemma 16 (Phase Transition Time)**.** *Suppose the assumptions of Lemma 11 hold, and let R_0 be its radius. Let $\bar{R} \ge R_0$ be deterministic and define*

$$\tau := \left\lceil \frac{4\bar{R}}{\gamma\eta} \max\left\{ 4e^2\eta \left(\sqrt{2} + \frac{5\gamma}{8}\right)^2, \frac{n}{\log 2} \right\} \right\rceil.$$

If $\tau \le T$, then there exists $s \in \{0, \dots, \tau - 1\}$ such that

$$\mathcal{L}_S(W(s)) \le \min\left\{ \frac{1}{4e^2\eta(\sqrt{2} + 5\gamma/8)^2}, \frac{\log 2}{n} \right\}.$$

Proof. Lemma 12 gives

$$\frac{1}{\tau}\sum_{k=0}^{\tau-1} \mathcal{P}_S(W(k)) \le \frac{2R_0}{\gamma\eta\tau} \le \frac{2\bar{R}}{\gamma\eta\tau}.$$

Hence some $s \in \{0, \dots, \tau - 1\}$ satisfies

$$\mathcal{P}_S(W(s)) \le \frac{1}{2}\min\left\{ \frac{1}{4e^2\eta(\sqrt{2} + 5\gamma/8)^2}, \frac{\log 2}{n} \right\}. \tag{22}$$

In particular, $n\mathcal{P}_S(W(s)) \le \log 2/2 < 1/2$. Thus every $|\ell'(y_i f_{W(s)}(\mathbf{x}_i))| < 1/2$, which, because $|\ell'(z)| = (1 + e^z)^{-1}$, implies $y_i f_{W(s)}(\mathbf{x}_i) > 0$ for all i. For $z \ge 0$, $\ell(z) \le 2|\ell'(z)|$; therefore

$$\mathcal{L}_S(W(s)) \le 2\mathcal{P}_S(W(s)).$$

Combining this with (22) proves the claim. □

B.3 Local Rademacher Complexity

We next bound the complexity of the function class containing the GD trajectory. Related local complexity bounds appear in Lei et al. (2026); Li et al. (2026). Our analysis keeps the initialization linear term and the nonlinear remainder separate. The initialization gradient bound gives a contribution of order $R/\sqrt{n}$ without an explicit depth factor, while the deep linearization estimate gives a contribution of order $E(m, R)R$. This separation makes the dependence on depth and width explicit.

Lemma 17 (Local Rademacher Complexity)**.** *Fix a deterministic radius $R \ge 1$ and define*

$$\mathcal{F}_R := \{\mathbf{x} \mapsto f_W(\mathbf{x}) : \|W - W(0)\|_F \le R\}, \qquad E(m, R) := C\frac{L^{13/6}(\log m)^{2/3}R^{1/3}}{m^{1/6}}.$$

Suppose Assumptions 1 and 2 hold and $m \gtrsim L^{10}\log(nL/\delta)(\log m)^4 R^2$. Then, for each fixed sample S, with probability at least $1 - \delta$ over the initialization,

$$\mathfrak{R}_{S_X,n}(\mathcal{F}_R) \le C\left(\frac{R}{\sqrt{n}} + E(m, R)R\right). \tag{23}$$

In particular, if

$$m \gtrsim n^3 L^{13}(\log m)^4 \log(nL/\delta)R^2, \tag{24}$$

then

$$\mathfrak{R}_{S_X,n}(\mathcal{F}_R) \le C\frac{R}{\sqrt{n}}. \tag{25}$$

The general estimate (23) uses only the local linearization regime already required by the optimization analysis. The cleaner $R/\sqrt{n}$ rate additionally requires (24); this additional condition makes the nonlinear remainder no larger than the linear Rademacher term.

Proof. Fix an ordered tuple $\widetilde{S} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n) \in S_X^n$. Use failure probability $\delta/2$ in Lemmas 7 and 8. Their intersection has probability at least $1 - \delta$, and on this event the two bounds hold simultaneously for every point of S_X, hence also for repeated points in $\widetilde{S}$. By (17), $f_{W(0)}(\mathbf{x}) = 0$ for every $\mathbf{x} \in \mathcal{X}$. For $i \in [n]$, write

$$g_i := \nabla f_{W(0)}(\widetilde{\mathbf{x}}_i), \qquad r_i(W) := f_W(\widetilde{\mathbf{x}}_i) - f_{W(0)}(\widetilde{\mathbf{x}}_i) - \langle g_i, W - W(0)\rangle.$$

Then

$$\begin{aligned}\widehat{\mathfrak{R}}_{\widetilde{S}}(\mathcal{F}_R) &\le \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\|W - W(0)\|_F \le R} \frac{1}{n} \sum_{i=1}^n \epsilon_i \langle g_i, W - W(0)\rangle \\ &\quad + \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\|W - W(0)\|_F \le R} \frac{1}{n} \sum_{i=1}^n \epsilon_i r_i(W) =: I_1 + I_2.\end{aligned}$$

For the linear term, Frobenius norm duality followed by Cauchy–Schwarz gives

$$\begin{aligned}I_1 &= R\,\mathbb{E}_{\boldsymbol{\epsilon}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i g_i \right\|_F \\ &\le \frac{R}{n} \left(\mathbb{E}_{\boldsymbol{\epsilon}} \left\| \sum_{i=1}^n \epsilon_i g_i \right\|_F^2 \right)^{1/2} = \frac{R}{n} \left(\sum_{i=1}^n \|g_i\|_F^2 \right)^{1/2} \le \frac{\sqrt{2}R}{\sqrt{n}}.\end{aligned}$$

Here, the first equality follows from Frobenius norm duality, the first inequality follows from Cauchy–Schwarz, and the second equality uses $\mathbb{E}[\epsilon_i \epsilon_j] = \mathbb{I}\{i = j\}$. The last inequality follows from Lemma 7.

For the nonlinear remainder, Lemma 8 with base point $W(0)$ yields, uniformly over $\|W - W(0)\|_F \le R$,

$$|r_i(W)| \le E(m, R)\|W - W(0)\|_F \le E(m, R)R.$$

Consequently,

$$\begin{aligned}I_2 &\le \frac{1}{n} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\|W - W(0)\|_F \le R} \left| \sum_{i=1}^n \epsilon_i r_i(W) \right| \\ &\le \frac{1}{n} \sup_{\|W - W(0)\|_F \le R} \sum_{i=1}^n |r_i(W)| \le E(m, R)R.\end{aligned}$$

Combining the bounds for I_1 and I_2, and then taking the supremum over $\widetilde{S} \in S_X^n$, proves (23).

Finally, under (24), after increasing the absolute constant if necessary,

$$E(m, R) = C \left(\frac{L^{13} (\log m)^4 R^2}{m} \right)^{1/6} \le \frac{C}{\sqrt{n}}.$$

Plugging this bound into (23) gives (25). □

C Proofs of the Main Results

All parameters defining a training budget or a parameter radius are fixed before drawing the sample and initialization. The optimization statements below hold on the corresponding initialization events whenever the finite-width separation assumption holds with the fixed margin γ.

C.1 Proof of Theorem 1

Proof. Fix the sample S. Apply the initialization and perturbation lemmas at radii R_0 and R, assigning suitable fractions of δ to their failure probabilities. Since $R \geq R_0 \geq 1$ and (2) implies the required width conditions, a union bound gives an event $\mathcal{G}_S$ with $\mathbb{P}_{W(0)}(\mathcal{G}_S) \geq 1-\delta$ on which both $\mathcal{E}_{R_0}$ and $\mathcal{E}_R$ hold. On this event, suppose the sample is separable with margin γ. Since $R \geq \sqrt{2}/\gamma$ and $R \geq R_0$,

$$R^8 = R^2 R^6 \geq 8\frac{R_0^2}{\gamma^6}.$$

Thus (2) also implies the width condition in Lemma 11, which proves the first two claims. If (4) holds, Lemmas 14 and 15 give monotonicity and (5), respectively. Applying Lemma 16 with radius R proves the final claim. □

C.2 Proof of Corollary 1

Proof. From Theorem 1,

$$R \leq \frac{2\eta K_1^2}{\gamma} + \frac{4\log(1+2\eta T)}{\gamma} + 1 + \frac{\sqrt{2}}{\gamma}.$$

The chosen learning rate gives $2\eta K_1^2/\gamma = T/(4C_\tau)$ and $\eta T \lesssim T^2$, hence

$$R \leq \frac{T}{4C_\tau} + \frac{C\log(2T)}{\gamma} + C\left(1+\frac{1}{\gamma}\right). \tag{26}$$

The condition $T \gtrsim K_1^2 n/\gamma^2$ ensures $4e^2\eta K_2^2 \geq n/\log 2$. Therefore,

$$\tau \leq \frac{4R}{\gamma\eta}(4e^2\eta K_2^2) + 1 = C_\tau R + 1.$$

Plugging (26) into the preceding bound gives

$$\tau \leq \frac{T}{4} + \frac{CC_\tau}{\gamma}\log(2T) + CC_\tau\left(1+\frac{1}{\gamma}\right) + 1.$$

The last two lower bounds on T, after enlarging the absolute constant C, ensure that $\frac{CC_\tau}{\gamma}\log(2T) \leq T/8$ and $CC_\tau(1+1/\gamma)+1 \leq T/8$; here we use the elementary implication $T \gtrsim a\log(2+a) \Rightarrow a\log(2T) \lesssim T$. Hence $\tau \leq T/2$. Therefore there is $s \leq T/2$ satisfying the condition for the stable phase, so $T - s \geq T/2$. Theorem 1 then gives

$$\begin{aligned}\mathcal{L}_S(W(T)) &\leq \frac{32}{\gamma^2\eta(T-s)} \\ &\leq \frac{C}{\gamma^2\eta T} = \frac{CC_\tau K_1^2}{\gamma^3 T^2}.\end{aligned}$$

□

C.3 Proof of Theorem 2

We first record the expanded bounds from Theorem 2. Under the additional width condition (9), equivalently (24), we have $\mathcal{C}_{n,m}(R) \lesssim R/\sqrt{n}$, and the average bound becomes

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{L}(W(t)) = \widetilde{\mathcal{O}}\Bigg(&\frac{1}{\eta T} + \frac{\eta + \eta^{-1}\log^2(1+\eta T)}{\gamma^2 T} \\ &+ \frac{1+\eta^2+\log^2(1+\eta T)}{\gamma^2 n} + \frac{L^2\big(1+\eta+\log(1+\eta T)\big)\log(2/\delta)}{\gamma n}\Bigg),\end{aligned} \tag{27}$$

and, if (4) holds at some $s < T$ with $T - s \asymp T$,

$$\begin{aligned}\mathcal{L}(W(T)) = \widetilde{\mathcal{O}}\Bigg(&\frac{1}{\gamma^2\eta T} + \frac{1+\eta^2+\log^2(1+\eta T)}{\gamma^2 n} \\ &+ \frac{L^2\big(1+\eta+\log(1+\eta T)\big)\log(2/\delta)}{\gamma n}\Bigg).\end{aligned} \tag{28}$$

Proof. Using Theorem 1 and Lemmas 17 and 5, each with failure probability $\delta/4$, gives an initialization event $\mathcal{G}_{\text{init}}$. Replacing δ by $\delta/4$ only changes the absolute constants in the logarithmic width conditions. For every fixed S,

$$\mathbb{P}_{W(0)}(\mathcal{G}_{\text{init}}^c \mid S) \le \frac{3\delta}{4},$$

and therefore

$$\mathbb{P}_{S,W(0)}(\mathcal{G}_{\text{init}}^c) = \mathbb{E}_S\big[\mathbb{P}_{W(0)}(\mathcal{G}_{\text{init}}^c \mid S)\big] \le \frac{3\delta}{4}.$$

For every fixed $W(0)$, the radius $R = R(\eta, T, \gamma)$ is deterministic and $\mathcal{F}_R$ does not depend on S. Continuity of $f_W(\mathbf{x})$ on the compact set $\mathcal{B}_R(W(0)) \times \mathcal{X}$ gives $B_\ell < \infty$ for each such initialization. Lemma 6, with failure probability $\delta/4$, then gives an event $\mathcal{G}_{\text{gen}}$ satisfying

$$\mathbb{P}_S(\mathcal{G}_{\text{gen}}^c \mid W(0)) \le \frac{\delta}{4},$$

and hence

$$\mathbb{P}_{S,W(0)}(\mathcal{G}_{\text{gen}}^c) = \mathbb{E}_{W(0)}\big[\mathbb{P}_S(\mathcal{G}_{\text{gen}}^c \mid W(0))\big] \le \frac{\delta}{4}.$$

By a union bound, $\mathcal{G} := \mathcal{G}_{\text{init}} \cap \mathcal{G}_{\text{gen}}$ has probability at least $1 - \delta$. Whenever Assumption 3 holds with the fixed margin γ, the optimization conclusions of Theorem 1 also hold on $\mathcal{G}$.

We first bound B_ℓ. By (17), $f_{W(0)}(\mathbf{x}) = 0$ for every $\mathbf{x} \in \mathcal{X}$. Under (6), Lemma 5 gives, uniformly over $W \in \mathcal{B}_R(W(0))$ and $\mathbf{x} \in \mathcal{X}$,

$$|f_W(\mathbf{x})| \le \|\mathbf{a}\|_2 \|h^L(\mathbf{x}) - h_0^L(\mathbf{x})\|_2 \le CL^2\sqrt{\log m}\, R.$$

Since $\ell(z) = \log(1 + e^{-z}) \le \log 2 + |z|$,

$$B_\ell \le \log 2 + CL^2\sqrt{\log m}\, R.$$

For the logistic loss, $H = 1/4$. Combining this bound with Lemmas 6 and 17, and applying Young's inequality to the terms containing $\sqrt{\mathcal{L}_S(W)}$, we have, uniformly for $W \in \mathcal{B}_R(W(0))$,

$$\mathcal{L}(W) \le 2\mathcal{L}_S(W) + C\log^3 n\, \mathcal{C}_{n,m}^2(R) + C\frac{\big(1 + L^2\sqrt{\log m}\, R\big)\log(2/\delta)}{n}. \tag{29}$$

Theorem 1 ensures that $W(t) \in \mathcal{B}_R(W(0))$ for every $t \le T$. Averaging (29) over $t = 0, \dots, T-1$ and using (3) proves (7). If (4) holds and $T - s \asymp T$, applying (29) to $W(T)$ and using (5) proves (8).

Under (24), Lemma 17 gives $\mathcal{C}_{n,m}(R) \lesssim R/\sqrt{n}$. Since $\gamma \in (0, 1]$, the definition of R gives

$$R \lesssim \frac{1 + \eta + \log(1 + \eta T)}{\gamma}.$$

Therefore,

$$\frac{R^2}{n} \lesssim \frac{1 + \eta^2 + \log^2(1 + \eta T)}{\gamma^2 n}, \qquad \frac{1 + L^2\sqrt{\log m}\, R}{n} = \widetilde{\mathcal{O}}\left(\frac{L^2(1 + \eta + \log(1 + \eta T))}{\gamma n}\right).$$

Plugging these estimates into (7) and (8), and absorbing logarithmic factors into $\widetilde{\mathcal{O}}(\cdot)$, proves (27) and (28). □

C.4 Proof of Corollary 2

Proof. Plugging $T \asymp n$ and $\eta = \Theta(1)$ into (27) proves (10). Under the additional phase condition, the same choice in (28) proves (11). □

C.5 A Sufficient Budget for Early Entry

We justify the sufficient budget in Remark 1, without changing the conditional last iterate statement of Corollary 2. Fix an absolute constant learning rate $\eta = \Theta(1)$, independent of n, L, γ, and assume the width conditions of Theorem 2 and (9) for the budget chosen below. Constants in this argument may depend on this fixed absolute choice of η.

Since $0 < \gamma \le 1$ and $n \ge 2$, the definitions in Theorem 1 give

$$R \lesssim \frac{\log(2+T)}{\gamma}, \qquad \max\left\{4e^2\eta(\sqrt{2}+5\gamma/8)^2, \frac{n}{\log 2}\right\} \lesssim n.$$

Thus $\tau \le C_0(n/\gamma^2)\log(2+T)+1$ for an absolute constant C_0. Set $A = n/\gamma^2$, $H = \log(2+A)$, and $T = \lceil CAH \rceil$. Here $A \ge 2$ and $H \ge \log 4$. For $C \ge 1$,

$$2 + T \le (C+3)AH, \qquad \log(2+T) \le \log(C+3) + \log A + \log H \le \left(2 + \frac{\log(C+3)}{\log 4}\right) H.$$

Choose C large enough that

$$C_0\left(2 + \frac{\log(C+3)}{\log 4}\right) + \frac{1}{2\log 4} \le \frac{C}{2}.$$

Such a choice exists because the left-hand side grows only logarithmically in C. Since $AH \ge 2\log 4$, it follows that $\tau \le CAH/2 \le T/2$. Theorem 1 therefore guarantees an index $s < \tau$ satisfying (4), and hence $T - s \ge T/2$. Finally, $T \ge n$ and $\log(2+T) \lesssim \log(2+n/\gamma^2)$, so (28) gives the rate in (11), with the additional logarithmic factors absorbed into $\widetilde{\mathcal{O}}$. This is a sufficient budget, not a necessary lower bound, and no new probability event is required.

D Extension to Other Losses

We extend the trajectory and stable-phase arguments to a class of convex decreasing losses, keeping the network, paired Gaussian initialization, fixed output weights and GD update unchanged. Population risk bounds additionally require smoothness of the loss. The notation in this appendix is local: θ, ρ, D and R_ℓ are defined separately for the chosen loss. The empirical risk, population risk and gradient potential use this loss in the definitions of Section 3. Write $z_{i,k} := y_i f_{W(k)}(\mathbf{x}_i)$ for the prediction margin of sample i at iteration k; it is distinct from the NTK separation margin γ in Assumption 3.

Assumption 4 (Regularity of the Loss). The loss $\ell : \mathbb{R} \to (0,\infty)$ is continuously differentiable, convex, and nonincreasing, with $\lim_{z\to\infty} \ell(z) = 0$. There is a constant $M_\ell > 0$ such that $|\ell'(z)| \le M_\ell$ for every $z \in \mathbb{R}$. For the population risk bound, assume in addition that ℓ' is H_ℓ-Lipschitz for a constant $H_\ell > 0$.

Assumption 5 (Conditions for the Stable Phase). There are constants $Q_\ell, C_\ell > 0$ such that

$$\ell(z) \le C_\ell |\ell'(z)|, \qquad z \ge 0, \tag{30}$$

and, for $z, z+v \ge 0$ with $|v| \le 1$,

$$\ell(z+v) \le \ell(z) + \ell'(z)v + Q_\ell |\ell'(z)| v^2. \tag{31}$$

The trajectory bound uses Assumption 4 without smoothness; the uniform population risk bound also requires smoothness. Assumption 5 supplies the additional tail and local quadratic conditions for the stable phase and acceleration, following Wu et al. (2024) but imposing them only on non-negative prediction margins. These conditions are not needed for Corollary 3; they also yield the comparator loss decay used for the fast average population risk rate in Corollary 4.

Examples. Logistic loss satisfies both assumptions with $M_\ell = 1$, $H_\ell = 1/4$, $Q_\ell = e/2$, and $C_\ell = 2$. The exponential loss with a linear continuation, $\ell(z) = e^{-z}$ for $z \ge 0$ and $\ell(z) = 1 - z$ for $z < 0$, also satisfies both, with $M_\ell = H_\ell = C_\ell = 1$ and $Q_\ell = e/2$. For $p > 0$, the loss $\ell(z) = (1+z)^{-p}$ for $z \ge 0$ and $\ell(z) = 1 - pz$ for $z < 0$ satisfies Assumption 4 with $M_\ell = p$ and $H_\ell = p(p+1)$, but not (30), since $\ell(z)/|\ell'(z)| = (1+z)/p$. These examples are also considered by

Wu et al. (2024). Ordinary exponential loss has an unbounded derivative on negative margins and is not covered here.

The regularity assumption does not imply a bound of the form $|\ell'| \le A\ell$. For example, let

$$\ell(z) = \begin{cases} \displaystyle\int_z^\infty e^{-u^2}\,du, & z \ge 0, \\ \sqrt{\pi}/2 - z, & z < 0. \end{cases}$$

This loss satisfies Assumption 4 with $M_\ell = 1$ and $H_\ell = \sqrt{2/e}$. For $z > 0$, $\ell(z) \le e^{-z^2}/(2z)$, so $|\ell'(z)|/\ell(z) \ge 2z$. Thus the first corollary below applies to losses that do not satisfy the self-boundedness condition used for the stable phase in Wu et al. (2024). We do not claim the accelerated rate for this example.

Corollary 3 (Trajectory and Population Risk for General Losses)**.** *Suppose Assumptions 1, 2, and 4 hold. Fix $\eta > 0$, an integer $T \ge 1$, $\gamma \in (0, 1]$, $\delta \in (0, 1)$, and $z_T \ge 0$, independently of S and $W(0)$. Define*

$$K = \sqrt{2} + \gamma/2, \qquad \theta = 2z_T/\gamma, \qquad \rho = \frac{\eta M_\ell K^2 + 1}{\gamma}, \qquad D = \theta + \rho,$$

$$R_\ell = 2D + \sqrt{2\eta T \ell(z_T)} + 1 + \gamma^{-1}.$$

Assume (2) *with $R = R_\ell$. For every fixed sample S, there is an initialization event of probability at least $1 - \delta$ on which, whenever Assumption 3 holds with margin γ, all iterates up to time T belong to $\mathcal{B}_{R_\ell}(W(0))$ and, for $1 \le t \le T$,*

$$\frac{1}{t}\sum_{k=0}^{t-1} \mathcal{L}_S(W(k)) \le \ell(z_T) + \frac{D^2}{2\eta t}, \qquad \sum_{k=0}^{t-1} \mathcal{P}_S(W(k)) \le \frac{2R_\ell}{\eta\gamma}. \tag{32}$$

If the smoothness condition in Assumption 4 and the width conditions (6) *and* (24) *also hold with $R = R_\ell$, there is a joint event over $S \sim \mathcal{D}^n$ and the independent initialization, of probability at least $1 - \delta$, on which the same implication holds and, for all $0 \le t \le T$,*

$$\begin{aligned} \mathcal{L}(W(t)) \le 2\mathcal{L}_S(W(t)) &+ CH_\ell \log^3 n\, \frac{R_\ell^2}{n} \\ &+ C\frac{(\ell(0) + M_\ell L^2 \sqrt{\log m}\, R_\ell)\log(2/\delta)}{n}. \end{aligned} \tag{33}$$

Here and below the loss is fixed; constants may depend on its parameters, but not on n, m, L, γ, η, T, or δ.

Proof. Write $E = E(m, R_\ell)$. As in Theorem 1, the width condition gives an event on which $\mathcal{E}_{R_\ell}$ holds, $E \le \gamma/8$, and $2ER_\ell \le 1/2$. The proof of Lemma 13, using only $\ell' < 0$, gives

$$\frac{\gamma}{2}\mathcal{P}_S(W) \le \|\nabla \mathcal{L}_S(W)\|_F \le K\mathcal{P}_S(W), \qquad \mathcal{P}_S(W) \le M_\ell$$

throughout the ball. To see that $\ell'(z) < 0$, choose $b > z$ with $\ell(b) < \ell(z)$, which is possible since $\ell(z) > 0$ and $\ell(+\infty) = 0$. Convexity gives $\ell'(z) \le (\ell(b) - \ell(z))/(b - z) < 0$. Moreover, ℓ' is nondecreasing, so $|\ell'| = -\ell'$ is nonincreasing.

Repeat the argument in the proof of Lemma 10 with $u_1 = W(0) + \theta W_*$ and $u = u_1 + \rho W_*$. Linearization gives $\mathcal{L}_S(u_1) \le \ell(z_T)$. Writing $P_k = \mathcal{P}_S(W(k))$, the same distance expansion gives, whenever $W(k)$ is in the ball,

$$\begin{aligned} \|W(k+1) - u\|_F^2 \le \|W(k) - u\|_F^2 &+ 2\eta\{\ell(z_T) - \mathcal{L}_S(W(k))\} \\ &+ \eta\{2E(\theta + R_\ell) - \rho\gamma + \eta M_\ell K^2\} P_k. \end{aligned}$$

Since $\theta \le R_\ell$ and $4ER_\ell \le 1$, we have $2E(\theta + R_\ell) - \rho\gamma + \eta M_\ell K^2 \le 1 - 1 = 0$. Thus

$$\|W(k+1) - u\|_F^2 \le \|W(k) - u\|_F^2 + 2\eta\ell(z_T) - 2\eta\mathcal{L}_S(W(k)). \tag{34}$$

The term involving ρ controls the linearization error, so no bound of the form $|\ell'| \lesssim \ell$ is needed. Assuming $W(0), \dots, W(k)$ belong to the ball, summing (34) and using nonnegativity of the loss gives

$$\|W(k+1) - W(0)\|_F \le D + \sqrt{D^2 + 2\eta T \ell(z_T)} < R_\ell.$$

Summing (34) proves the average loss bound. Summing the updates in direction W_*, as in Lemma 12, proves the potential bound.

For generalization, the uniform feature bound and $|\ell'| \le M_\ell$ give $B_\ell \le \ell(0) + CM_\ell L^2 \sqrt{\log m}\, R_\ell$. Apply the proof of Theorem 2 with this bound, smoothness constant H_ℓ, and deterministic radius R_ℓ. Lemmas 6 and 17 give (33). The same allocation of failure probabilities gives the stated joint event. The separating direction may depend on the sample; the function class and its radius do not, once $W(0)$ is fixed. □

Corollary 4 (Acceleration for General Losses)**.** *In Corollary 3, assume also Assumption 5 and choose $z_T = C_\ell \log(1 + 2\ell(0)\eta T)$. Set*

$$K' = \sqrt{2} + 5\gamma/8, \qquad A_\ell = 2\max\{1, 2Q_\ell\}, \qquad r_* = \min\left\{\frac{1}{8\eta A_\ell (1+Q_\ell)(K')^2}, \frac{\ell(1)}{2n}\right\},$$

$$\tau_\ell = \left\lceil \frac{2R_\ell}{\eta\gamma} \max\left\{\frac{2n}{|\ell'(1)|}, \frac{C_\ell}{r_*}\right\}\right\rceil.$$

On the optimization event in Corollary 3, whenever separability holds and $\tau_\ell \le T$, there is $s < \tau_\ell$ such that the loss is nonincreasing from s to T and, for $1 \le t \le T - s$,

$$\mathcal{L}_S(W(s+t)) \le \left(\mathcal{L}_S(W(s))^{-1} + \frac{\eta\gamma^2 t}{8C_\ell^2}\right)^{-1} \le \frac{8C_\ell^2}{\eta\gamma^2 t}. \tag{35}$$

There are constants $c_*, C_* > 0$, depending only on the loss, such that $\eta = c_*\gamma^2 T$ and $T \ge C_*\gamma^{-2}(n + \log(2/\gamma))$ imply $\tau_\ell \le T/2$ and $\mathcal{L}_S(W(T)) = \mathcal{O}(1/(\gamma^4 T^2))$. Under the additional conditions for (33)*, choosing $\eta = \Theta(1)$ and $T \asymp n$ instead gives average population risk $\widetilde{\mathcal{O}}(1/(\gamma^2 n) + L^2 \log(2/\delta)/(\gamma n))$.*

Proof. First, (30) and $\ell' < 0$ give

$$\frac{d}{dz}\left(e^{z/C_\ell}\ell(z)\right) = e^{z/C_\ell}\left(\ell'(z) + \frac{\ell(z)}{C_\ell}\right) \le 0, \qquad z \ge 0.$$

Thus $\ell(z) \le \ell(0)e^{-z/C_\ell}$, and the choice of z_T yields $\ell(z_T) \le 1/(2\eta T)$ and $R_\ell \lesssim (1 + \eta + \log(1 + \eta T))/\gamma$. Next, take $v = \min\{1, (2Q_\ell)^{-1}\}$ in (31). Since $Q_\ell v \le 1/2$, we have $0 < \ell(z+v) \le \ell(z) - v|\ell'(z)|/2$, and hence $|\ell'(z)| \le 2\ell(z)/v = A_\ell \ell(z)$ for $z \ge 0$.

The potential bound in (32) gives some $s < \tau_\ell$ with

$$\mathcal{P}_S(W(s)) \le \min\{|\ell'(1)|/(2n), r_*/C_\ell\}.$$

For every $i \in [n]$, $|\ell'(z_{i,s})| \le n\mathcal{P}_S(W(s)) \le |\ell'(1)|/2 < |\ell'(1)|$. Since $|\ell'|$ is nonincreasing, $z_{i,s} > 1$. We can therefore apply (30) to obtain $\mathcal{L}_S(W(s)) \le C_\ell \mathcal{P}_S(W(s)) \le r_*$.

We argue by induction, following Lemma 14. Write $r_k = \mathcal{L}_S(W(k))$, $P_k = \mathcal{P}_S(W(k))$, and $V_k = \|\nabla \mathcal{L}_S(W(k))\|_F$. If $r_k \le r_*$, then for every i,

$$\ell(z_{i,k}) \le n r_k \le n r_* \le \ell(1)/2 < \ell(1).$$

As shown in the proof of Corollary 3, $\ell' < 0$. Thus ℓ is strictly decreasing, so $z_{i,k} > 1$ and $P_k \le A_\ell r_k$. Both iterates belong to $\mathcal{B}_{R_\ell}(W(0))$ by Corollary 3. The linearization bound therefore gives

$$|z_{i,k+1} - z_{i,k}| \le \eta K' V_k \le \eta (K')^2 P_k \le \frac{1}{8(1+Q_\ell)} < 1.$$

Hence $z_{i,k+1} > 0$, and (31) applies to $z_{i,k}$ and $z_{i,k+1}$. Averaging this inequality and using the GD update, as in Lemma 14, gives

$$\begin{aligned} r_{k+1} &\le r_k - \eta V_k^2 + \eta E(m, R_\ell) P_k V_k + Q_\ell \eta^2 (K')^2 P_k V_k^2 \\ &\le r_k - \eta(1 - 1/4 - 1/8) V_k^2 \le r_k - \tfrac{\eta}{2} V_k^2. \end{aligned}$$

Here $EP_k \le V_k/4$ follows from $E \le \gamma/8$ and $V_k \ge \gamma P_k/2$, while $Q_\ell \eta (K')^2 P_k \le 1/8$ follows from the definition of r_*. Thus $r_{k+1} \le r_k \le r_*$, which closes the induction. The same individual loss bound gives $z_{i,k+1} > 1$ for every i; no monotonicity of the individual prediction margins is required. By (30), $P_k \ge r_k / C_\ell$, and hence

$$r_{k+1} \le r_k - \frac{\eta\gamma^2}{8C_\ell^2} r_k^2, \qquad \frac{1}{r_{k+1}} \ge \frac{1}{r_k} + \frac{\eta\gamma^2}{8C_\ell^2}.$$

Summing proves (35).

Finally, $\tau_\ell \le 1 + C\gamma^{-2}(1 + \eta + \log(1+\eta T)) \max\{1, n/\eta\}$. Put $u = \gamma^2 T$ and choose $0 < c_* \le 1$ so that $C c_* \le 1/8$. For $\eta = c_* u$, we have

$$\log(1+\eta T) = \log\left(1 + \frac{c_* u^2}{\gamma^2}\right) \le 2\log(1+u) + 2\log(2/\gamma).$$

Choose C_* sufficiently large. Since $u \ge C_*(n + \log(2/\gamma))$ and $\log(1+u)/u \to 0$, this ensures $c_* u \ge n$, $T \ge 4$, and $C\{1 + 2\log(1+u) + 2\log(2/\gamma)\} \le u/8$. Therefore $\tau_\ell \le 1 + T/4 \le T/2$. Combining this with (35) gives the accelerated rate. For constant η and $T \asymp n$, we instead have $D, R_\ell = \widetilde{\mathcal{O}}(\gamma^{-1})$; averaging (33) and using (32) proves the population risk bound. □

E Experimental Details and Additional Results

E.1 Experimental Setup

Data. We use noisy 2-XOR, following the construction studied by Ji & Telgarsky (2020) and the numerical setting of Li et al. (2026). Let j be uniform on $\{1, 2\}$, and let $s, \xi_3, \ldots, \xi_d$ be independent uniform signs. With e_k denoting the kth coordinate vector, set

$$\mathbf{x} = \frac{s e_j + \sum_{k=3}^{d} \xi_k e_k}{\sqrt{d-1}}, \qquad y = \begin{cases} +1, & j = 1, \\ -1, & j = 2. \end{cases}$$

The label is determined by which of the first two coordinates is nonzero; the remaining coordinates add nuisance variation, not label noise. We fix $d = 8$ in the main experiments and draw training observations independently with replacement. The distribution has $2^d = 256$ equally likely points, so its population logistic loss can be evaluated exactly by enumeration.

For real data, we use the official Fashion-MNIST training and test splits (Xiao et al., 2017). We select T-shirt/top (class 0, labeled $+1$) and Pullover (class 2, labeled -1) to obtain a binary classification problem. Each image is flattened into $\mathbb{R}^{784}$ and normalized to unit Euclidean norm. Training subsets contain equal numbers from the two classes, sampled without replacement within each class. Evaluation uses all 2,000 corresponding test images. We report their mean logistic loss as *test loss*; unlike the XOR evaluation, this is not the exact population risk.

Network and optimization. We use the network and paired initialization of Sections 3 and 4.1, with width $m = 128$ and $L \in \{2, 4, 8, 16\}$ hidden layers. All hidden layers are trained by full batch GD on logistic loss; the output weights remain fixed. We use no biases, momentum, weight decay, normalization layers or early stopping. Learning rates are chosen on the geometric grid $\eta \in \{2^{-2}, 2^0, 2^2, 2^4\}$. An additional value 2^6 is included in the stability experiment. Table 2 gives the sample sizes and budgets. The grid was fixed for the unified runs after exploratory experiments, without selection by test performance.

Table 2: Experimental protocol. Each configuration uses ten seeds, width 128 and all four depths. The Fashion-MNIST sample size experiment is reported in Appendix E.5.

Experiment	Sample size n	Learning rate η	Updates T
XOR trajectories	128	$1/4, 1, 4, 16$	1024
Fashion-MNIST trajectories	512	$1/4, 1, 4, 16$	512
XOR sample size	$32, 64, 128, 256$	$1, 16$	$2n$
Fashion-MNIST sample size	$64, 128, 256, 512$	$1, 16$	$2n$

Repeated runs and evaluation. We use ten seeds, indexed by $0, \ldots, 9$, for each configuration. Data and initialization are paired across learning rates; samples are nested within each sample size experiment. Lines show the median, and shaded regions span the 25th to 75th percentiles across completed runs. These are measures of variation across seeds, not confidence intervals. We retain one fixed cohort of completed runs along each trajectory and report nonfinite runs separately. Dotted curves show the recorded finite losses before numerical failure. Crosses mark the first iteration at which a nonfinite quantity is detected; their vertical positions are for display only. Points with incomplete runs are hollow in the sample size plots. Appendix E.7 gives implementation and reporting details.

These finite-width experiments examine training and prediction behavior in the stated model. They do not certify the theorem assumptions or estimate the accelerated budget-dependent rate in Corollary 1.

E.2 Training Loss and Learning Rate

Figure 3 shows the training loss at each update. Larger learning rates substantially reduce the number of updates required to reach a small loss in several depth configurations. For XOR with $L = 4$,

the median numbers of updates to reach training loss 10^{-2} are 675, 169 and 46 for $\eta = 1, 4, 16$, respectively. Their median final losses are 3.75×10^{-3}, 3.85×10^{-4} and 6.64×10^{-5}. On Fashion-MNIST at the same depth, the corresponding final losses are 6.43×10^{-2}, 9.51×10^{-3} and 1.11×10^{-3}.

The products of learning rate and median threshold time are 675, 676 and 736, respectively, indicating an approximately inverse dependence of the threshold hitting time on the learning rate over the tested range.

The benefit depends on depth and initialization. At $L = 16$ and $\eta = 16$, nine of ten XOR runs and six of ten Fashion-MNIST runs complete the budget. Thus, increasing the learning rate at fixed width can also lead to numerical instability.

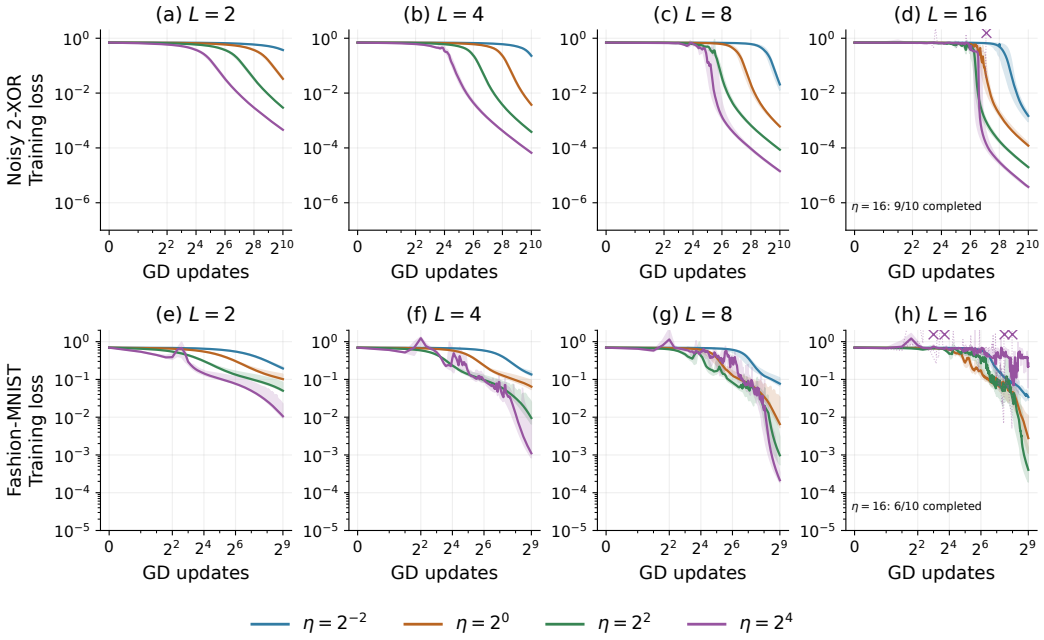

Figure 3: Training logistic loss on noisy 2-XOR (top) and binary Fashion-MNIST (bottom). The sample sizes are 128 and 512, respectively, and $m = 128$. Lines and bands show medians and interquartile intervals across completed runs. At $L = 16$, $\eta = 16$, the completion counts are $9/10$ and $6/10$; all other displayed configurations complete all ten runs. Dotted finite prefixes and crosses retain information about failed runs. Values above the displayed range are clipped. The horizontal scale is logarithmic in base two away from the origin and linear near zero.

E.3 Population and Test Loss during Training

Figure 4 evaluates the same trajectories on the full XOR distribution and the held-out Fashion-MNIST test set. This comparison distinguishes faster fitting from lower population or test loss. On XOR, larger learning rates can bring the reduction in population loss forward in training. On Fashion-MNIST, the ordering of test losses need not follow that of training losses: a configuration that fits the training set more accurately can have higher test logistic loss. This distinction motivates the explicit learning rate dependence in Theorem 2, rather than an inference about generalization from optimization alone.

E.4 Sample Size Dependence on XOR

Corollary 2 suggests examining population risk with $T \asymp n$ and a constant learning rate. Both $\eta = 1$ and $\eta = 16$ are absolute constants; neither is scaled with the budget in this experiment. We therefore set $T = 2n$, keep d, L, m, η fixed within each panel, and plot risk against $1/n$. For a trajectory, we

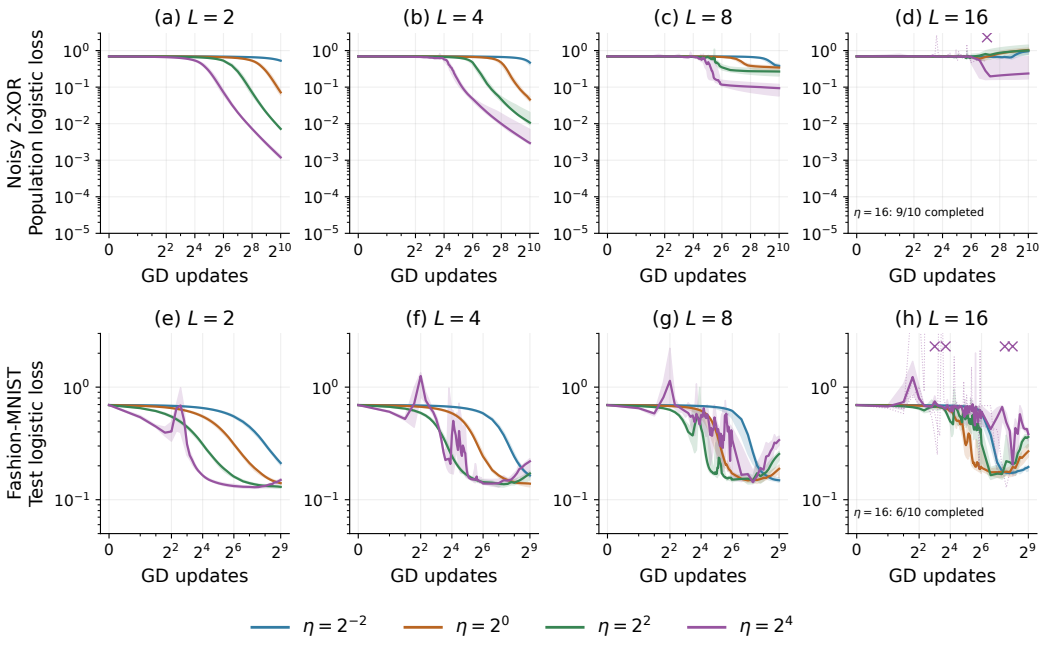

Figure 4: Exact population logistic loss on XOR (top) and test logistic loss on Fashion-MNIST (bottom), for the same runs as Figure 3. Evaluation is performed at every update through $t = 64$, then every 32 updates and at the final iterate. Curves connect the recorded evaluations. Bands, failure markers and completion counts have the same meaning as in Figure 3; values outside the displayed range are clipped. The test set is used only for evaluation.

record $\overline{\mathcal{L}}_T = T^{-1} \sum_{t=0}^{T-1} \mathcal{L}(W(t))$ and $\mathcal{L}(W(T))$. Both quantities are evaluated on the full XOR distribution. For the average, we use every iterate $t = 0, \ldots, T-1$; the sparse evaluation schedule in Figure 4 applies only to the trajectory experiment.

Figure 5 shows a substantial reduction in risk in several configurations as the sample size and training budget increase. For $L = 4$ and $\eta = 16$, the median average risk decreases from 1.018 at $n = 32$ to 0.0329 at $n = 256$, while the median last iterate risk decreases from 1.217 to 4.43×10^{-4}. The last iterate can therefore improve considerably on an average that includes the early training phase, but it is not uniformly better at every sample size.

Each panel includes a reference line c/n, anchored at the median average risk for the largest sample size. Specifically, $c = n_{\max} \operatorname{median}(\overline{\mathcal{L}}_{2n_{\max}})$, where $n_{\max} = 256$. This constant is fixed within the panel and shared by the average and last iterate curves; it is not fitted across sample sizes. The reference provides a visual comparison with $1/n$ scaling over the tested range, and the observed curves do not follow it uniformly. Since $T = 2n$ increases optimization time together with sample size, Appendix E.5.2 also reports a fixed-budget control. The choice $T = 2n$ does not by itself ensure the early entry condition for the last iterate in Corollary 2.

The dimension dependence also requires care. A margin bound $\gamma^{-1} = \mathcal{O}(d)$, when valid for the relevant NTK, gives $\widetilde{\mathcal{O}}((d^2 + L^2 d \log(2/\delta))/n)$ after plugging it into our corollary. The shallow-network construction of Ji & Telgarsky (2020) motivates the data choice, but does not by itself supply that margin for every depth and finite initialization used here. We therefore fix d in the main comparison and do not infer a common d^2/n law across depths. Throughout this section, the reported quantity is population logistic risk. For this deterministic-label distribution, the infimum over all measurable predictors is zero, but no such zero-risk baseline is assumed for Fashion-MNIST.

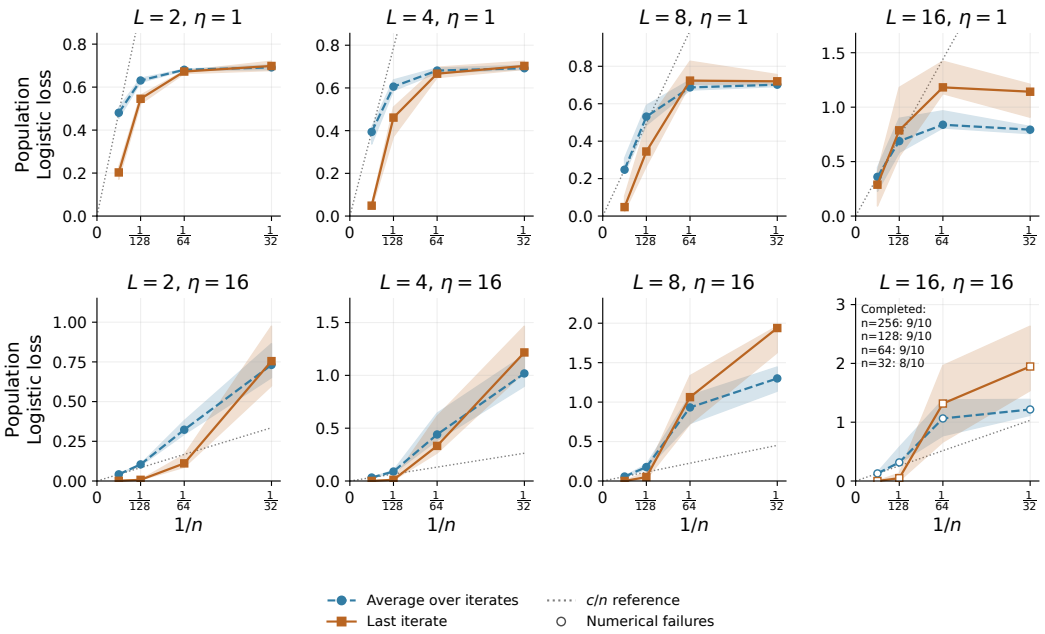

Figure 5: Average and last iterate population logistic risks on noisy 2-XOR with $d = 8$, $m = 128$ and $T = 2n$. Both axes are linear; moving left increases n. The dotted c/n line is anchored at the median average risk for $n = 256$, separately in each panel. Bands show interquartile intervals. All ten runs complete except at $L = 16$, $\eta = 16$, where the counts for $n = 32, 64, 128, 256$ are $8, 9, 9, 9$. Hollow points denote these incomplete groups.

E.5 Additional Experiments

E.5.1 Sample Size Dependence on Fashion-MNIST

Figure 6 uses balanced training subsets of sizes $n \in \{64, 128, 256, 512\}$, with $T = 2n$, $m = 128$ and the same ten seeds for each configuration. The test set is fixed throughout. Test loss is evaluated at every iterate, and the average uses $t = 0, \ldots, T - 1$, as in the XOR sample size experiment. For $L = 4$ and $\eta = 1$, the median average test loss decreases from 0.325 to 0.163, whereas the last iterate loss decreases from 0.171 to 0.141. With $\eta = 16$, the final loss at $n = 512$ is approximately 0.257, despite the smaller training losses attainable with this learning rate.

The test loss does not exhibit a uniform $1/n$ dependence over the displayed range.

E.5.2 A Fixed Training Budget

To distinguish the effect of sample size from the increase in training time, Figure 7 compares the last iterate population risk on XOR under $T = 2n$ and $T = 512$. Data, initialization, width and learning rate are paired between the two budgets. At the smaller sample sizes, additional optimization changes the population risk, confirming that Figure 5 should be read as an evaluation of the combined sample and iteration budget. Neither experiment alone identifies an asymptotic statistical rate.

E.5.3 Width and a Separating Direction

We vary $m \in \{64, 128, 256, 512\}$ on XOR, keeping $d = 8$, $n = 128$, $\eta = 16$ and $T = 1024$. Figure 8 reports training loss for $L = 4, 16$. All ten runs complete at $L = 4$. At $L = 16$, the completion counts are $7, 9, 10, 10$ as width increases. Thus width can affect numerical stability even at a fixed learning rate; these four widths do not determine the polynomial width dependence of the theory.

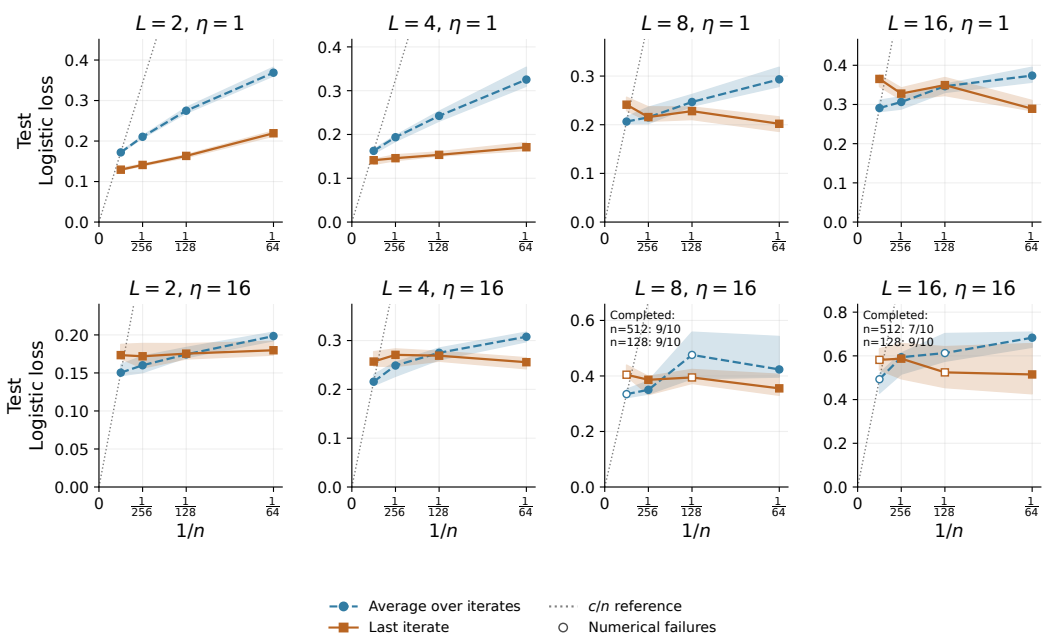

Figure 6: Average and last iterate test logistic losses on binary Fashion-MNIST against $1/n$, with $T = 2n$. The dotted reference is anchored at the largest sample size, as in Figure 5. All groups complete ten runs except those with $\eta = 16$: $L = 8$, $n = 128, 512$ have $9, 9$ completed runs, and $L = 16$, $n = 128, 512$ have $9, 7$. Hollow markers identify these groups.

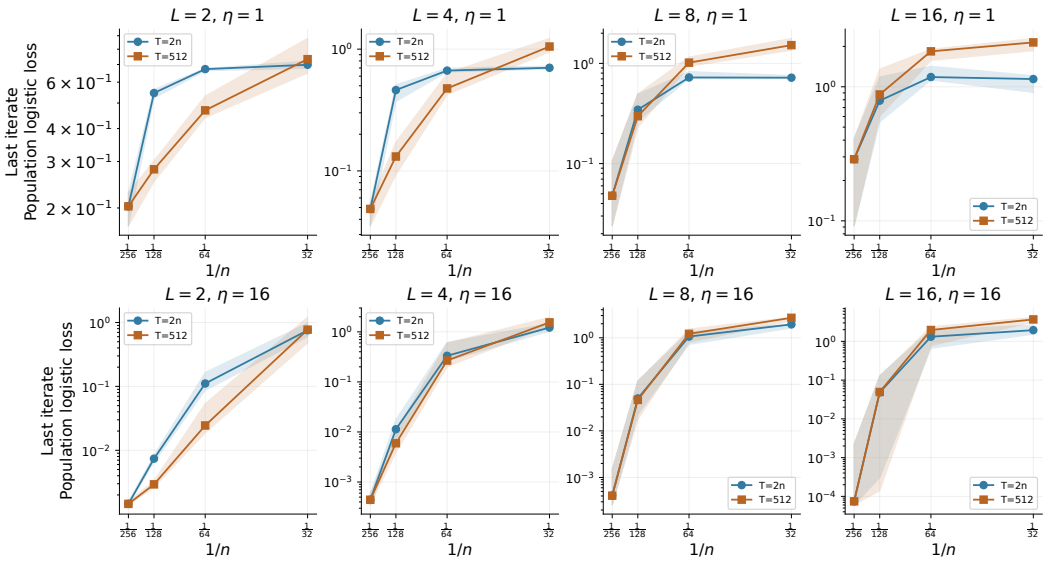

Figure 7: Last iterate population logistic loss on XOR with $d = 8$ under two training budgets. The horizontal axis is linear in $1/n$ and the vertical axis is logarithmic. Lines and bands show medians and interquartile intervals. All configurations complete ten runs except $L = 16$, $\eta = 16$: for $n = 32, 64, 128, 256$, the counts are $8, 9, 9, 9$ with $T = 2n$ and $7, 6, 9, 9$ with $T = 512$. The statistics are conditional on completion under each budget.

We also examine the persistence of one separating direction on the training sample. Let Φ_0 have rows $\mathrm{vec}(\nabla f_{W(0)}(\mathbf{x}_i))^\top$, let $\mathbf{y} = (y_1, \ldots, y_n)^\top$, and set $u = \Phi_0^\top(\Phi_0\Phi_0^\top)^\dagger\mathbf{y}$. If $u \neq 0$, define $v = u/\|u\|_2$ and $\widehat{\gamma}_0 = \min_i y_i(\Phi_0 v)_i$. We retain the direction diagnostic only when $\widehat{\gamma}_0 > 0$ and

record

$$\frac{\widehat{\gamma}_t}{\widehat{\gamma}_0}, \qquad \widehat{\gamma}_t = \min_{i\in[n]} y_i \langle \mathrm{vec}(\nabla f_{W(t)}(\mathbf{x}_i)), v\rangle.$$

The pseudoinverse is computed using NumPy with `rcond=1e-12`. The direction is constructed from the initial features and labels, not drawn randomly. It is then held fixed during training. A value above $1/2$ means that this direction retains at least half its measured initial margin on the training sample at that recorded iterate. The ratio measures the persistence of this fixed separating direction on the training sample; it does not characterize separation by other directions or over the population.

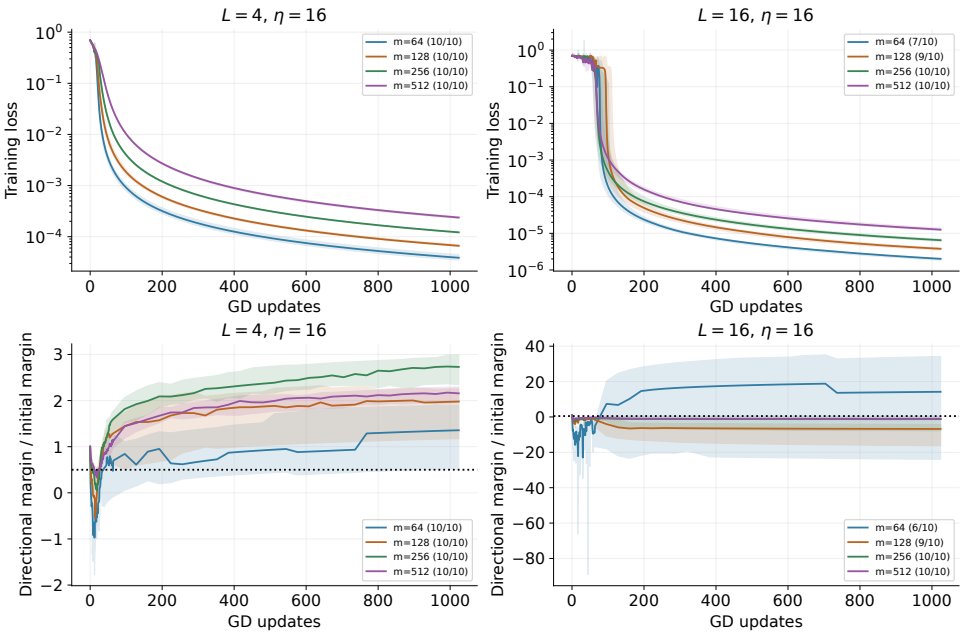

Figure 8: Width dependence of training loss (top) and the retained directional margin (bottom) on XOR at $\eta = 16$. Legends report the number of included runs out of ten. The lower panels additionally require a positive initial directional margin; at $L = 16, m = 64$, six completed runs meet this condition. The dotted horizontal line is $1/2$. Directions are fixed within runs but differ across seeds and widths.

E.5.4 Numerical Failures and Classification Error

Figure 9 reports all numerical failures in the trajectory grid, including $\eta = 64$. A run fails when the loss, gradient or a required evaluation becomes nonfinite in double precision. No failed run is replaced by a zero loss or a successful seed. A larger learning rate can therefore produce faster fitting among completed runs while reducing the completion frequency. Nonfinite arithmetic is not, by itself, a proof of divergence of the exact real-valued GD dynamics.

For completeness, Figures 10 and 11 report classification error for the sample size experiments. We predict $+1$ when $f_W(\mathbf{x}) \geq 0$ and -1 otherwise. The average curve averages these errors over $t = 0, \ldots, T-1$. Classification error is bounded above by logistic risk divided by $\log 2$, but this upper bound does not imply that the two quantities have the same observed dependence on n. The main comparisons therefore use the logistic loss appearing in our theorems.

E.6 A Stable-Phase Threshold Diagnostic

We check the loss threshold in (4) without estimating γ. Since $0 < \gamma \leq 1$, the following threshold is no larger than the threshold in (4):

$$b_{\eta,n} := \min\left\{ \frac{1}{4e^2\eta(\sqrt{2}+5/8)^2}, \frac{\log 2}{n} \right\}.$$

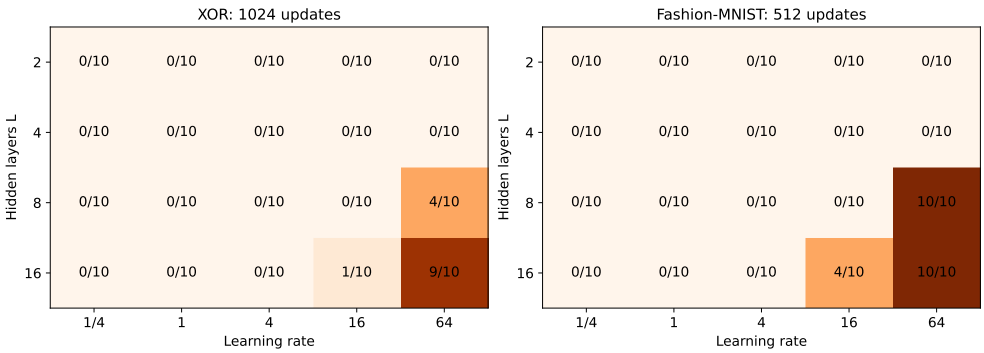

Figure 9: Numbers of nonfinite runs out of ten, at $m = 128$. XOR uses $n = 128, T = 1024$; Fashion-MNIST uses $n = 512, T = 512$. The additional learning rate 64 shows that the range of reliable training depends on depth and dataset.

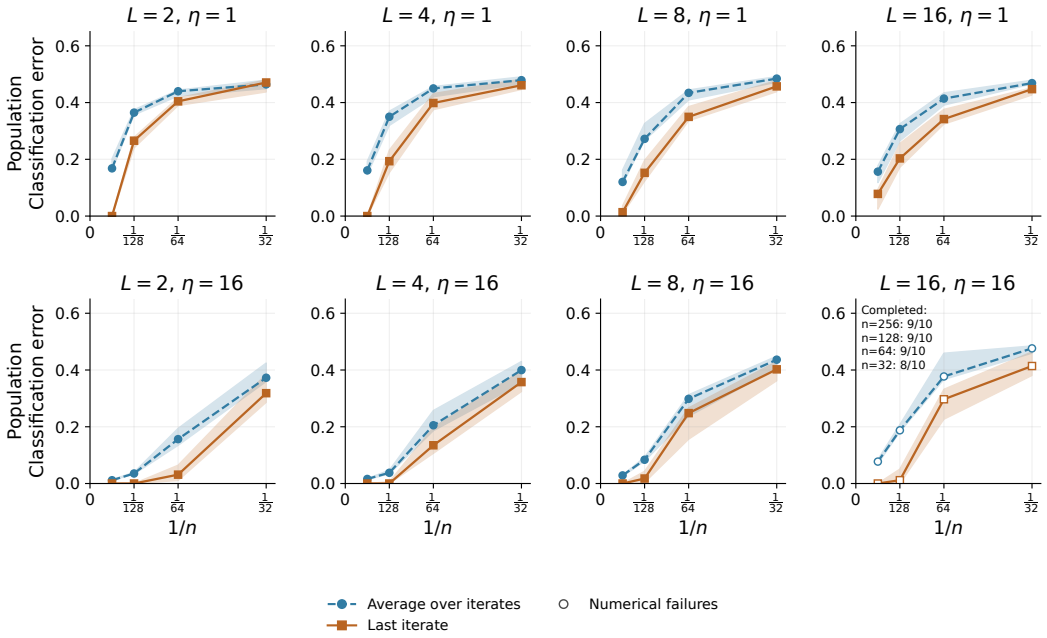

Figure 10: Average and last iterate classification errors on the exact XOR distribution, for the runs in Figure 5. Shading and hollow markers use the same reporting convention.

For each archived trajectory, let s be the first recorded update with $\mathcal{L}_S(W(s)) \le b_{\eta,n}$ and $s < T$, if one exists. Table 3 reports how many of the ten seeds reach this threshold by $T/2$ and before T. Every update is recorded, and finite prefixes of failed runs are checked as well; no seed is removed from the denominator.

The threshold is reached early in several XOR configurations but rarely by half the budget on Fashion-MNIST. Because $b_{\eta,n}$ is conservative, not crossing it does not rule out satisfying (4) with the actual margin. Crossing it verifies the loss threshold only, not the remaining assumptions of Theorem 1.

E.7 Implementation and Reproducibility

The implementation uses NumPy in double precision, stable evaluation of $\log(1 + \exp(-z))$, and the convention $\sigma'(0) = 1$ from Appendix A. All hidden layers are updated from the same iterate.

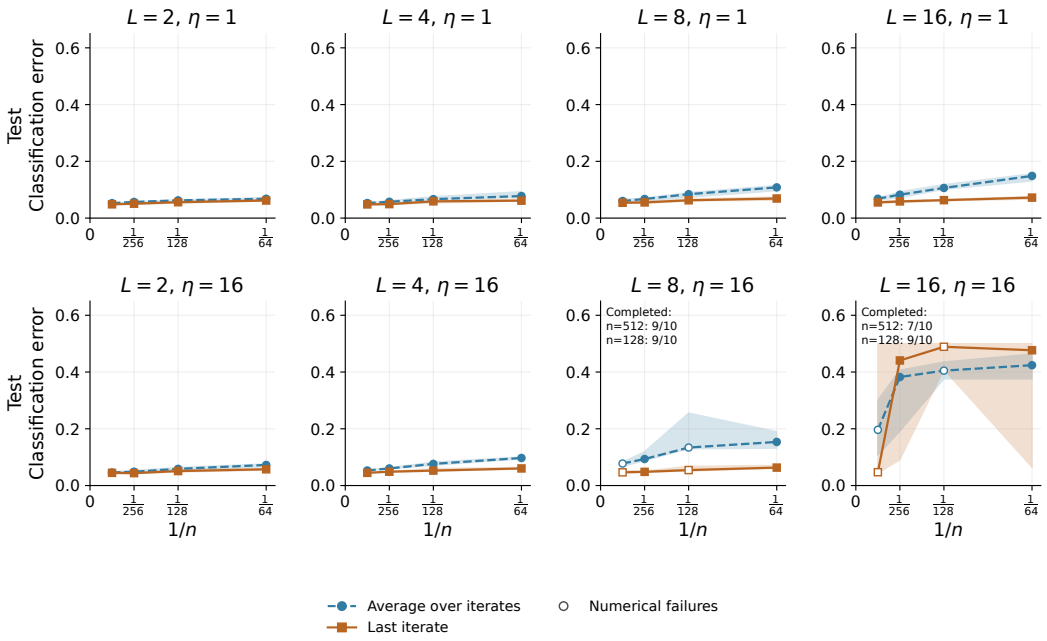

Figure 11: Average and last iterate classification errors on the official binary Fashion-MNIST test set, for the runs in Figure 6. Shading and hollow markers use the same reporting convention.

Table 3: Conservative loss-threshold diagnostic at $m = 128$. Each entry (a, b) reports the number of runs reaching the threshold by $T/2$ and before T, respectively, out of ten seeds. XOR uses $n = 128, T = 1024$; Fashion-MNIST uses $n = 512, T = 512$.

Dataset	L	$\eta = 1/4$	$\eta = 1$	$\eta = 4$	$\eta = 16$
XOR	2	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 10)$
	4	$(0, 0)$	$(0, 10)$	$(10, 10)$	$(10, 10)$
	8	$(0, 0)$	$(10, 10)$	$(10, 10)$	$(10, 10)$
	16	$(3, 10)$	$(10, 10)$	$(10, 10)$	$(9, 9)$
Fashion-MNIST	2	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
	4	$(0, 0)$	$(0, 0)$	$(0, 0)$	$(0, 0)$
	8	$(0, 0)$	$(0, 0)$	$(0, 6)$	$(0, 8)$
	16	$(0, 0)$	$(0, 3)$	$(2, 7)$	$(0, 0)$

Seeds and separate data and initialization offsets are recorded. Quartiles use linear interpolation; summaries exclude failed runs and report completion counts, so incomplete groups describe performance conditional on completion.

Saved records contain training curves, evaluation checkpoints, sample hashes, failure times and configuration metadata. Validation checks dataset checksums, input norms, nested samples, risk averages and record lengths. Finite-difference gradient checks and independent replays of selected trajectories agree up to numerical precision.