Large Stepsize Gradient Descent for Logistic Loss: Non-Monotonicity of the Loss Improves Optimization Efficiency

Jingfeng Wu UUUJF@BERKELEY.EDU
University of California, Berkeley

Peter L. Bartlett* PETER@BERKELEY.EDU
University of California, Berkeley and Google DeepMind

Matus Telgarsky* MJT10041@NYU.EDU
New York University

Bin Yu* BINYU@BERKELEY.EDU
University of California, Berkeley

Editors: Shipra Agrawal and Aaron Roth

Abstract

We consider *gradient descent* (GD) with a constant stepsize applied to logistic regression with linearly separable data, where the constant stepsize η is so large that the loss initially oscillates. We show that GD exits this initial oscillatory phase rapidly — in $\mathcal{O}(\eta)$ steps, and subsequently achieves an $\tilde{\mathcal{O}}(1/(\eta t))$ convergence rate after t additional steps. Our results imply that, given a budget of T steps, GD can achieve an *accelerated* loss of $\tilde{\mathcal{O}}(1/T^2)$ with an aggressive stepsize $\eta := \Theta(T)$, without any use of momentum or variable stepsize schedulers. Our proof technique is versatile and also handles general classification loss functions (where exponential tails are needed for the $\tilde{\mathcal{O}}(1/T^2)$ acceleration), nonlinear predictors in the *neural tangent kernel* regime, and online *stochastic gradient descent* (SGD) with a large stepsize, under suitable separability conditions.

Keywords: optimization; logistic regression; gradient descent; edge of stability; acceleration

1. Introduction

Gradient-based methods are arguably the most popular optimization methods in modern machine learning. The simplest version, constant-stepsize *gradient descent* (GD), is defined as

$$\mathbf{w}_t := \mathbf{w}_{t-1} - \eta \nabla L(\mathbf{w}_{t-1}), \quad t = 1, 2, \ldots, T, \tag{GD}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the trainable parameter, $L(\cdot)$ is the loss function, $\eta > 0$ is a stepsize fixed across the iterates, and $\mathbf{w}_0$ is the initialization. We focus on understanding GD with a *large* stepsize, which means that the induced loss does not decrease monotonically.

The classical theory for constant-stepsize GD assumes a small stepsize such that the induced loss decreases monotonically (Nesterov, 2018). For example, if $L(\mathbf{w})$ is β-smooth, then choosing $\eta < 2/\beta$ guarantees $L(\mathbf{w}_t)$ to decrease monotonically, a result known as the *descent lemma* (Nesterov, 2018, Section 1.2.3). Fruitful convergence results have been developed for small stepsize GD based on the descent lemma or its variants. However, little theory is known when GD is run with a large stepsize (for instance, $\eta > 2/\beta$ for a β-smooth function), which induces non-monotonic loss values (exceptions will be discussed later in Section 1.1). Cohen et al. (2020) refer to this as the *edge of*

* Alphabetical order

stability (EoS) regime, and they further point out that GD usually operates in the EoS regime for trained neural networks to exhibit reasonable optimization or generalization performance (Wu and Ma, 2018; Cohen et al., 2020).

This work considers large stepsize GD in minimal yet natural machine learning settings, such as *logistic regression*:

$$L(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ln\left(1+\exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right), \quad (\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}. \tag{1}$$

Our first main result (Theorem 1) characterizes the dynamics of large stepsize GD for logistic regression with *linearly separable* data. We show that GD initially induces a non-monotonic loss (the EoS phase) and then exits this regime in finite time (phase transition), and afterwards the loss decreases monotonically (the stable phase). Specifically:

1. **The EoS phase.** First, we show the loss *averaged* over t steps decreases at a rate of $\tilde{\mathcal{O}}\big((1+\eta^2)/(\eta t)\big)$ for GD with an arbitrarily large stepsize η. In particular, this applies to GD in the EoS phase where the loss oscillates locally.

2. **Phase transition.** Second, we show that GD exits the initial EoS phase (if it ever enters) in $\mathcal{O}(\eta)$ steps. Then GD undergoes a phase transition and enters a stable phase, where the loss decreases monotonically, and stays in this phase.

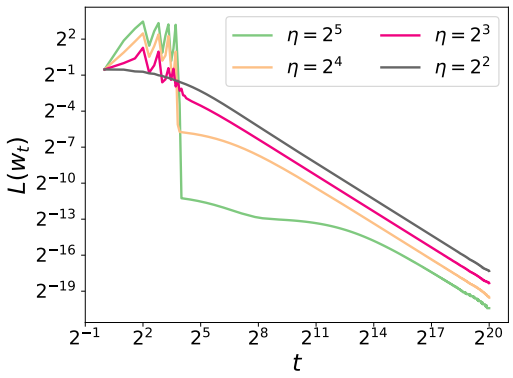
3. **The Stable phase.** Finally, we show in the stable phase, the loss decreases monotonically at an $\tilde{\mathcal{O}}\big(1/(\eta t)\big)$ rate after t steps. Figures 1(*b*) and 1(*d*) empirically verify that this rate is sharp asymptotically (as $t \to \infty$) ignoring logarithmic factors.

The above result immediately justifies the **benefits of a large stepsize** for improving optimization efficiency. First, a larger η yields a smaller constant factor in the *asymptotic* bound, $\tilde{\mathcal{O}}(1/(\eta t))$. We highlight that η can be an arbitrarily large constant, beyond the classical wisdom that constrains η by the inverse smoothness (Nesterov, 2018). Second, given a budget of T steps, GD attains an $\tilde{\mathcal{O}}(1/T^2)$ loss when equipped with an aggressive stepsize $\eta := \Theta(T)$ that optimally balances the steps spent in the EoS and stable phases (see Corollary 2). In contrast, we also provide an $\Omega(1/T)$ loss lower bound for constant-stepsize GD that does not enter the EoS phase (see Theorem 3). Our theory explains the empirical observations in Figure 1.

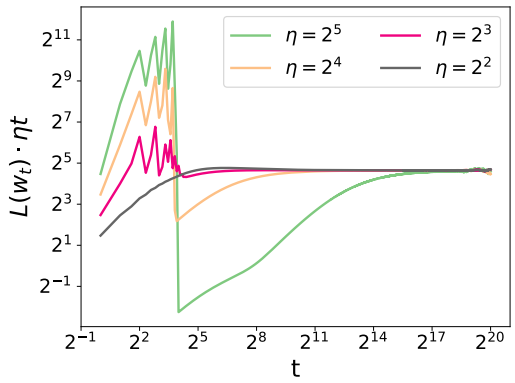
We extend the above result in three notable aspects.

- **General losses.** We relax the logistic loss in (1) to general classification loss functions under a set of conditions (see Assumption 3 and Theorem 6). Specifically, the EoS bound relies on the *Lipschitzness* of the loss, the stable phase bound relies on an additional *self-boundedness* condition, and the $\tilde{\mathcal{O}}(1/T^2)$ acceleration relies on the loss having an *exponential tail*.

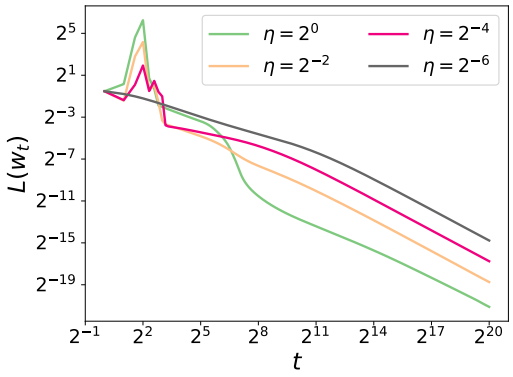
- **Nonlinear models.** We extend the linear predictor in (1) to a two-layer network in the *neural tangent kernel* (NTK, Jacot et al., 2018) regime (Theorem 6). To our knowledge, this is the first NTK result that accommodates an arbitrarily large stepsize (the minimum network width depends on the stepsize). For a constant stepsize, the minimum network width is *polylogarithmic* in the number of steps and the number of samples, recovering the main result of Ji and Telgarsky (2019) as a special case.



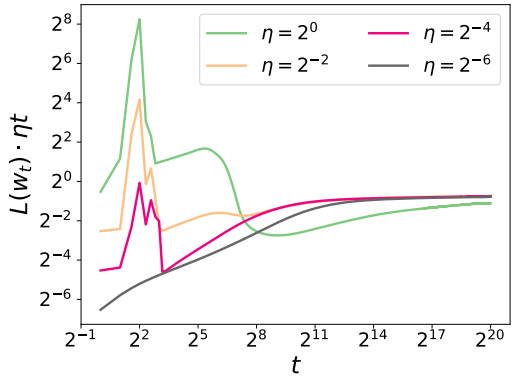
(a) Training loss, toy dataset.



(b) Asymptotic rate, toy dataset.



(c) Training loss, subset of MNIST.



(d) Asymptotic rate, subset of MNIST.

Figure 1: Behaviors of (GD) for logistic regression (1) with initialization $\mathbf{w}_0 := 0$. Figures 1(a) and 1(b) are results on a toy dataset consisting of four samples, $\mathbf{x}_1 = (1, 0.2)$, $y_1 = 1$, $\mathbf{x}_2 = (-2, 0.2)$, $y_2 = 1$, $\mathbf{x}_3 = (-1, -0.2)$, $y_3 = -1$, and $\mathbf{x}_4 = (2, -0.2)$, $y_4 = -1$. Here, the stepsize $\eta = 2^2$ is small, and $\eta \in \{2^3, 2^4, 2^5\}$ is large (depending on whether the loss decreases monotonically). Figures 1(c) and 1(d) are results on 1000 samples from the MNIST dataset with labels "0" or "8". Here, the stepsize $\eta = 2^{-6}$ is small, and $\eta \in \{2^{-4}, 2^{-2}, 2^0\}$ is large. Figures 1(a) and 1(c) suggest that (GD) with a larger stepsize incurs a smaller loss *when stabilized*. Figures 1(b) and 1(d) suggest that $L(\mathbf{w}_t) = \Theta(1/(\eta t))$ asymptotically as $t \to \infty$.

- **SGD.** Finally, we consider constant-stepsize online *stochastic gradient descent* (SGD) for logistic regression with separable data. We provide upper bounds on the population logistic loss and the population zero-one error (Theorem 4), allowing the stepsize to be $o(T)$ (where T is the maximum number of steps) and arbitrarily large, respectively. These are in contrast to the typical SGD results in online smooth convex optimization that only allow small (even vanishing) stepsizes (see, for example, Hazan, 2022, Chapter 3).

Implication and limitations of our theory in practice. The work by Cohen et al. (2020) empirically identifies two phenomena when training neural networks using GD with a large stepsize (η):

progressive sharpening, where the sharpness increases until around $2/\eta$, and EoS, where the sharpness oscillates around $2/\eta$. Their definition of EoS and ours are equivalent for quadratic functions, and our definition highlights the unstable loss behaviors more explicitly for non-quadratic functions. Additionally, a transition from the EoS phase to the stable phase is observed in neural networks trained with cross-entropy loss (see Cohen et al., 2020, Appendix A). Our theory offers valuable insights into the EoS and stable phases but remains inconclusive regarding the progressive sharpening phenomenon.

1.1. Related Works

Large stepsize and EoS. Empirically, training neural networks with GD (or large batch SGD) often requires a large stepsize to achieve reasonable optimization and generalization performance (see, for example, Hoffer et al., 2017; Goyal et al., 2017; Wu and Ma, 2018; Cohen et al., 2020), whereby the training loss exhibits non-monotonic behavior. Cohen et al. (2020) introduce the term *edge of stability* (EoS) to refer to this. The theory of EoS and large stepsize has been studied in various cases, such as one- or two-dimensional (nearly analytic) functions (Zhu et al., 2022; Ahn et al., 2023; Kreisler et al., 2023; Chen et al., 2023; Wang et al., 2023), scale-invariant networks (Lyu et al., 2022), diagonal linear networks (Even et al., 2023; Andriushchenko et al., 2023), and matrix factorization (Wang et al., 2022b; Chen and Bruna, 2023). Besides, there are some general theories for EoS (Kong and Tao, 2020; Ahn et al., 2022; Damian et al., 2022; Ma et al., 2022; Wang et al., 2022c; Lu et al., 2023), but they are built up on subtle assumptions. In comparison, we work in a natural classification setting under standard assumptions.

Among the literature of EoS, the work by Wu et al. (2023) is most relevant to us. They show the convergence of GD with an arbitrarily large stepsize for logistic regression with linearly separable and *non-degenerate* data. Our work significantly improves theirs. First, the constant factors in their bounds are *exponential* in the stepsize η. In comparison, we nail down the dependence on η in each phase (see Theorem 1), establishing the $\tilde{\mathcal{O}}(1/(\eta t))$ asymptotic rate and the $\tilde{\mathcal{O}}(1/T^2)$ acceleration with stepsize $\eta = \Theta(T)$ (see Corollary 2). So we explain the benefits of large stepsizes while they cannot. Second, they require the dataset to be *non-degenerate* such that the support vectors span the dataset (see their Assumption 3), while we do not need this condition. Finally, their results are specialized to GD and logistic regression, while our approach is versatile and also handles general classification losses, nonlinear predictors in the NTK regime, and SGD with a large stepsize.

Variable stepsize schedulers. AdaBoost (Freund and Schapire, 1997), in the language of convex optimization, is *coordinate descent* with an exponentially increasing stepsize scheduler, achieving an $\exp(-\Omega(t))$ rate (see Telgarsky, 2013, for example). Following this idea, a line of work (Nacson et al., 2019; Ji and Telgarsky, 2021; Ji et al., 2021) considers GD with an increasing stepsize scheduler (that adapts to the loss value) and obtains faster rates for linear classification under exponentially tailed losses. Despite the increase in stepsize, the loss in these works *always* decreases monotonically as the gradient dynamics satisfy a local version of the descent lemma. In contrast, we focus on a large stepsize that allows GD to enter the EoS phase where the loss oscillates. We remark that our results for constant stepsize can be applied recursively for analyzing variable stepsize schedulers, but this is beyond the scope of the current paper.

A recent line of work uses variable stepsize schedulers to accelerate the *worst-case* rate of GD for a class of problems (Kelner et al., 2022; Altschuler and Parrilo, 2023, and references therein). In particular, Altschuler and Parrilo (2023) show that GD with the *Silver* stepsize scheduler achieves an

$\mathcal{O}(1/t^{1.2716})$ rate for smooth convex optimization. Similarly to our work, they achieve acceleration by using aggressive stepsizes that result in oscillatory losses. However, they design a delicate variable stepsize scheduler that improves the worst-case rate of GD, while we focus on constant-stepsize GD for linear classification problems.

Malitsky and Mishchenko (2020) design an adaptive stepsize scheduler for GD for smooth convex optimization, where they show a convergence rate without an explicit dependence on the global smoothness parameter. Their analysis involves two consecutive GD steps, instead of using the classical descent lemma. Thus their analysis also allows non-monotonic loss. In comparison, we focus on GD with a large but constant stepsize, and our analysis relies on the separability condition of the dataset instead of variable stepsizes.

Neural tangent kernel. The *neural tangent kernel* (NTK) is most prominently presented and named by Jacot et al. (2018), and similar concepts are used by subsequent works to analyze the early training dynamics of GD. In the majority of these works (see Du et al., 2018; Zou et al., 2018; Zou and Gu, 2019; Allen-Zhu et al., 2019, for example), the stepsize of GD is $o(1)$ and vanishes as a function of other parameters such as the target error or the level of overparameterization. The only two exceptions are by Ji and Telgarsky (2019) and Chen et al. (2020), allowing a constant stepsize smaller than $2/\beta$, where β is the smoothness of the objective near initialization. In comparison, our NTK result (see Theorem 6) handles arbitrarily large stepsizes, in particular, it recovers the related result of Ji and Telgarsky (2019) when specializing the stepsize to a constant.

2. GD for Logistic Regression

We now present our results on GD for logistic regression. We assume the dataset is bounded and linearly separable.

Assumption 1 (Bounded and separable data) *Assume the training data $(\mathbf{x}_i, y_i)_{i=1}^n$ satisfies*

A. for every $i = 1, \dots, n$, $\|\mathbf{x}_i\| \le 1$ and $y_i \in \{\pm 1\}$;

B. there is a margin *$\gamma > 0$ and a unit vector $\mathbf{w}_*$ such that $\langle y_i \mathbf{x}_i, \mathbf{w}_* \rangle \ge \gamma$ for every $i = 1, \dots, n$.*

The next theorem characterizes the behaviors of GD for logistic regression.

Theorem 1 (Bounds on risk and phase transition time) *Consider (GD) with stepsize $\eta > 0$ for logistic regression (1) under Assumption 1. Without loss of generality, assume $\mathbf{w}_0 = 0$. We say that GD is* in the stable phase *at step t when $L(w_t)$ decreases monotonically from t onwards, and* in the EoS phase *when it does not. Then we have the following:*

- ***The EoS phase.*** *For every $t > 0$ (and in particular in the EoS phase), we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le \frac{1 + \ln^2(\gamma^2 \eta t) + \eta^2/4}{\gamma^2 \eta t}.$$

- ***The stable phase.*** *If s is such that*

$$L(\mathbf{w}_s) \le \frac{1}{\eta}, \tag{2}$$

then (GD) *is in the stable phase, that is,* $\big(L(\mathbf{w}_t)\big)_{t\geq s}$ *decreases monotonically, and moreover,*

$$L(\mathbf{w}_t) \leq \frac{2F(\mathbf{w}_s) + \ln^2(\gamma^2\eta(t-s))}{\gamma^2\eta(t-s)}, \quad \text{where } F(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n}\exp(-y_i\mathbf{x}_i^\top\mathbf{w}).$$

- ***Phase transition time.*** *There exists* $s \leq \tau$ *such that* (2) *holds and* $F(\mathbf{w}_s) \leq 1$*, where*

$$\tau := \frac{60}{\gamma^2}\max\left\{\eta,\ n,\ e,\ \frac{\eta+n}{\eta}\ln\frac{\eta+n}{\eta}\right\}.$$

The proof of Theorem 1 is deferred to Appendix B. Theorem 1 provides control over the average loss in the EoS phase, the last iterate loss in the stable phase, and the maximum length of the EoS phase. Note that GD might never enter the EoS phase. For instance, this happens when the stepsize is sufficiently small such that (2) holds for $\mathbf{w}_s = \mathbf{w}_0 = 0$. However, if GD enters the EoS phase at the beginning of the optimization, the algorithm must undergo a phase transition and switch to the stable phase after a finite time. We emphasize that Theorem 1 allows an arbitrarily large η.

Benefits of large stepsizes. Theorem 1 shows that a large stepsize improves optimization efficiency in two ways. First, when $\eta = \mathcal{O}(1)$, the EoS phase ends in finite time, so asymptotically GD stays in the stable phase and attains an $\tilde{\mathcal{O}}\big(1/(\eta t)\big)$ rate. In this case, a larger stepsize leads to a smaller constant factor. We also note that this rate can be sharp ignoring logarithmic factors according to the experimental results in Figures 1(*b*) and 1(*d*). Interestingly, recall the classical rate of GD for smooth convex optimization is $\mathcal{O}\big(1/(\eta t)\big)$ for $\eta < 2/\beta$, where β is the smoothness parameter (Nesterov, 2018, Theorem 2.1.14). Our results (nearly) match this rate but remove the $\eta < 2/\beta$ condition for logistic regression, a special class of smooth convex optimization problems.

Second, Theorem 1 suggests that GD with a larger stepsize converges faster in the stable phase, but it takes more steps to exit the initial EoS phase. Picking a stepsize proportional to the total number of steps will balance these two effects and lead to an acceleration effect. This is formalized by the following corollary (the proof is deferred to Appendix C).

Corollary 2 (Acceleration of a large stepsize) *Under the same setup as in Theorem 1, consider* (GD) *with a given budget of* T *steps, where* $T \geq 120\max\{e, n\}/\gamma^2$*. Then for*

$$\eta := \frac{\gamma^2}{120}T,$$

we have $\tau \leq T/2$ *and*

$$L(\mathbf{w}_T) \leq 480\frac{\ln^2(\gamma^4T^2)}{\gamma^4T^2} = \mathcal{O}\left(\frac{\ln^2(T)}{T^2}\right).$$

Corollary 2 shows that GD attains an $\tilde{\mathcal{O}}(1/T^2)$ loss after T steps with a large stepsize $\eta = \Theta(T)$. We point out that the final low loss is a consequence of the initial EoS phase. In particular, the next theorem (the proof is deferred to Appendix D) shows that in general, GD with T steps of fixed stepsize must suffer an $\Omega(1/T)$ loss if it never enters the EoS phase.

Theorem 3 (Lower bound in the classical regime) *Consider (GD) with $\mathbf{w}_0 = 0$ and stepsize $\eta > 0$ for logistic regression (1) on the following dataset:*

$$\mathbf{x}_1 = \big(\gamma,\ \sqrt{1-\gamma^2}\big), \quad \mathbf{x}_2 = \big(\gamma,\ -\sqrt{1-\gamma^2}/2\big), \quad y_1 = y_2 = 1, \quad 0 < \gamma < 0.1.$$

It is clear that $(\mathbf{x}_i, y_i)_{i=1,2}$ satisfy Assumption 1. If $\big(L(\mathbf{w}_t)\big)_{t\geq 0}$ is non-increasing, then

$$L(\mathbf{w}_t) \geq c_0/t, \quad t \geq 1,$$

where $c_0 > 0$ is a function of γ but is independent of t and η.

We remark that the acceleration in Corollary 2 requires the budget to be at least linear in the sample size, that is, $T \geq \Omega(n)$. This requirement is rooted in the phase transition time bound in Theorem 1. We conjecture our phase transition time bound, especially its dependence on the sample size, is improvable, and thus the $T \geq \Omega(n)$ condition in Corollary 2 can be relaxed. Some empirical evidence is provided in Figure 2 in Appendix A.

A proof technique. We now showcase a powerful yet surprisingly simple technique for handling large stepsizes, which lies at the heart of the proof of Theorem 1. To begin with, consider a comparator $\mathbf{u} := \mathbf{u}_1 + \mathbf{u}_2$. Then from the definition of (GD), we have

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L(\mathbf{w}_t), \mathbf{u} - \mathbf{w}_t\rangle + \eta^2 \big\|\nabla L(\mathbf{w}_t)\big\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L(\mathbf{w}_t), \mathbf{u}_1 - \mathbf{w}_t\rangle + \eta^2\Big(\langle \nabla L(\mathbf{w}_t), 2\mathbf{u}_2/\eta\rangle + \big\|\nabla L(\mathbf{w}_t)\big\|_2^2\Big).\end{aligned}$$

The last term is *non-positive* when we choose $\mathbf{u}_2 := \mathbf{w}_* \cdot \eta/(2\gamma)$, because

$$\langle \nabla L(\mathbf{w}_t), 2\mathbf{u}_2/\eta\rangle = \frac{1}{n}\sum_{i=1}^n \ell'(y_i\mathbf{x}_i^\top \mathbf{w}_t)\langle y_i\mathbf{x}_i, 2\mathbf{u}_2/\eta\rangle \leq \frac{1}{n}\sum_{i=1}^n \ell'(y_i\mathbf{x}_i^\top \mathbf{w}_t),$$

where we use $\langle y_i\mathbf{x}_i, \mathbf{w}_*\rangle \geq \gamma$ and $\ell'(\cdot) \leq 0$, and

$$\big\|\nabla L(\mathbf{w}_t)\big\|_2^2 = \left\|\frac{1}{n}\sum_{i=1}^n \ell'(y_i\mathbf{x}_i^\top \mathbf{w}_t) y_i\mathbf{x}_i\right\|_2^2 \leq \left(\frac{1}{n}\sum_{i=1}^n \ell'(y_i\mathbf{x}_i^\top \mathbf{w}_t)\right)^2 \leq \frac{1}{n}\sum_{i=1}^n |\ell'(y_i\mathbf{x}_i^\top \mathbf{w}_t)|,$$

where we use $\|y_i\mathbf{x}_i\| \leq 1$ and $|\ell'(\cdot)| \leq 1$. Dropping the last term, we have

$$\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 \leq \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L(\mathbf{w}_t), \mathbf{u}_1 - \mathbf{w}_t\rangle \leq \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\big(L(\mathbf{u}_1) - L(\mathbf{w}_t)\big),$$

where we use the convexity of $L(\cdot)$. Telescoping the sum from 0 to t and rearranging, we obtain

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \leq L(\mathbf{u}_1) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t}. \tag{3}$$

Let us choose $\mathbf{u}_1 := \mathbf{w}_* \cdot \ln(\gamma^2\eta t)/\gamma$, then we have $L(\mathbf{u}_1) \leq 1/(\gamma^2\eta t)$ and $\|\mathbf{w}_0 - \mathbf{u}\| = \|\mathbf{u}\| = \ln(\gamma^2\eta t)/\gamma + \eta/(2\gamma)$. Then (3) allows us to control the average loss and the norm of $\mathbf{w}_t$ under any stepsize $\eta > 0$. We refer the reader to Appendix B for a complete proof of Theorem 1.

Other acceleration techniques. Our focus in this paper is to understand the benefits of using a large but constant stepsize in GD. Based on a simple yet powerful technique, we show that GD with a suitably large stepsize leads to non-monotonic loss and achieves improved optimization efficiency. Our techniques are orthogonal to standard acceleration methods such as momentum and variable stepsize schedulers. We conjecture these techniques can be combined to obtain an even faster optimization. This is left for future work.

Implicit bias and generalization. We conclude this part by discussing the implicit bias and generalization of the output of GD with a large stepsize, η. When $\eta = \mathcal{O}(1)$ compared to the number of steps, we can apply existing implicit bias results (Soudry et al., 2018; Ji and Telgarsky, 2018) to GD in the stable phase (replacing the initialization by $\mathbf{w}_s$) to show the direction of the GD iterate asymptotically converges to the max-margin direction. In this situation, the generalization of the GD output is guaranteed by standard margin-based generalization bounds (Bartlett and Shawe-Taylor, 1999, for example).

When $\eta \geq \omega(1)$ compared to the total number of steps T (for example, $\eta = \Theta(T)$ as in Corollary 2), existing implicit bias results (Soudry et al., 2018; Ji and Telgarsky, 2018) can no longer be applied. In particular, our bound on the norm of the GD iterates is as large as $\mathcal{O}(\eta)$ (see (3) or Lemma 8 in Appendix B), which prevents us from proving a good margin for the GD iterates. Nonetheless, as the final GD iterate attains low training error, the standard uniform-convergence theory is applicable to guarantee its generalization. We refer the reader to Proposition 15 in Appendix E for one such bound.

We leave for future work the study of the implicit bias and generalization of GD with a very large stepsize, that is, $\eta \geq \omega(1)$ compared to the total number of steps T.

3. SGD for Logistic Regression

We next consider constant-stepsize online *stochastic gradient descent* (SGD) for logistic regression, defined as

$$\mathbf{w}_{t+1} := \mathbf{w}_t - \eta \nabla L_t(\mathbf{w}_t), \quad \text{where} \quad L_t(\mathbf{w}) := \ln\left(1 + \exp(-y_t \mathbf{x}_t^\top \mathbf{w})\right), \quad t \geq 0. \tag{SGD}$$

Here, $(\mathbf{x}_t, y_t)_{t\geq 0}$ are independent and identically distributed according to the following assumption.

Assumption 2 (Bounded and separable distribution) *Assume that $(\mathbf{x}_t, y_t)_{t\geq 0}$ are independent copies of $(\mathbf{x}, y)$ that follows a distribution such that*

A. the label is binary, $y \in \{\pm 1\}$, and $\|\mathbf{x}\| \leq 1$, almost surely;

B. there exist a margin $\gamma > 0$ and a unit vector $\mathbf{w}_*$ such that $\langle y\mathbf{x}, \mathbf{w}_* \rangle \geq \gamma$, almost surely.

The following theorem controls the population logistic loss and the population zero-one loss of SGD with a large stepsize.

Theorem 4 (Risk and error bounds) *Consider (SGD) with stepsize $\eta > 0$ for logistic regression under Assumption 2. Without loss of generality, assume $\mathbf{w}_0 = 0$. Denote*

$$L(\mathbf{w}) := \mathbb{E} \ln\left(1 + \exp(-y\mathbf{x}^\top \mathbf{w})\right),$$

where the expectation is over $(\mathbf{x}, y)$ *as in Assumption 2. Then with probability at least* $1-\delta$ *over the randomness of* $(\mathbf{x}_k, y_k)_{k=0}^{t-1}$*, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le \frac{2+2\ln^2(\gamma^2\eta t)+\eta^2/2}{\gamma^2\eta t} + \frac{3+2\ln(\gamma^2\eta t)+\eta}{\gamma}\cdot\frac{18\ln(1/\delta)}{t},$$

and

$$\frac{1}{t}\sum_{k=0}^{t-1} \Pr(y\mathbf{x}^\top\mathbf{w}_k \le 0) \le \frac{4\big(\sqrt{2}+2\ln(\gamma^2\eta t)+\eta\big)}{\gamma^2\eta t} + \frac{36\ln(1/\delta)}{t}.$$

The proof of Theorem 4 is deferred to Appendix F. The population logistic loss bound in Theorem 4 allows SGD with a stepsize as large as $o(t)$. For instance, in t steps, SGD achieves an $\mathcal{O}(1/\sqrt{t})$ population logistic loss on average with $\eta := \sqrt{t}$. This is in contrast to the typical online smooth convex optimization results, where the stepsize for SGD is a small constant or even vanishing (see Hazan, 2022, Chapter 3, for example).

The population zero-one error bound in Theorem 4 allows SGD with an arbitrarily large stepsize. In particular, for a large stepsize $\eta \ge \ln(\gamma^2 t)$, there exists $\hat{\mathbf{w}} \in (\mathbf{w}_k)_{k=0}^{t-1}$ such that

$$\Pr\left(y\mathbf{x}^\top\hat{\mathbf{w}} \le 0\right) \le C\left(\frac{1}{\gamma^2 t} + \frac{\ln(1/\delta)}{t}\right),$$

where $C \ge 1$ is an absolute constant. This bound for SGD matches that achieved by the Perceptron algorithm (Novikoff, 1962; Hanneke and Kontorovich, 2021), and matches (or improves by logarithmic factors) that of large margin classifiers (Bartlett and Shawe-Taylor, 1999; Grønlund et al., 2020; Hanneke and Kontorovich, 2021). Note that the parameter $\hat{\mathbf{w}}$ might not exhibit a large margin over the training data.

Unlike the result in Section 2 for batch GD, Theorem 4 does not show the benefits of a large stepsize for SGD (except on logarithmic factors). Specifically, the logistic loss bound in Theorem 4 is minimized when $\eta = \Theta\big(\ln(t)\big)$ (and larger stepsize such as $\eta \ge \omega\big(\ln(t)\big)$ hurts the bound), and the zero-one error bound in Theorem 4 is minimized when $\eta \ge \Omega\big(\ln(t)\big)$ (so larger stepsize is as good). However, both optimized bounds do not improve the bounds for $\eta = \Theta(1)$ ignoring logarithmic factors. We leave it as an open problem to investigate whether or not a large stepsize benefits SGD.

Finally, the above discussion applies to multi-pass SGD, empirical logistic loss, and empirical zero-one error by setting the distribution in Assumption 2 to an empirical distribution.

4. General Loss Functions and Nonlinear Predictors

We now extend our results in Section 2 to general classification loss functions and a two-layer network in the *neural tangent kernel* (NTK, Jacot et al., 2018) regime.

General classification losses. We first introduce a set of conditions for the loss functions.

Assumption 3 (Loss conditions) *Consider a loss function* $\ell : \mathbb{R} \to \mathbb{R}_+$.

A. ***Regularity.*** *Assume $\ell(\cdot)$ is continuously differentiable, convex, non-increasing, and $\ell(+\infty) = 0$. Then the following function $\rho : [1, \infty) \to \mathbb{R}_+$ is well-defined:*

$$\rho(\lambda) := \min_{z \in \mathbb{R}} \lambda \ell(z) + z^2, \quad \lambda \geq 1.$$

B. ***Lipschitzness.*** *Denote $g(\cdot) := |\ell'(\cdot)|$. Assume there is a constant $C_g > 0$ such that $g(\cdot) \leq C_g$.*

C. ***Self-boundedness.*** *Assume there is a constant $C_\beta > 0$ such that $g(\cdot) \leq C_\beta \ell(\cdot)$ and*

$$\ell(z) \leq \ell(x) + \ell'(x)(z - x) + C_\beta g(x)(z - x)^2 \text{ for } z \text{ and } x \text{ such that } |z - x| < 1.$$

D. ***An exponential tail.*** *Assume there is a constant $C_e > 0$ such that $\ell(z) \leq C_e g(z)$ for $z \geq 0$.*

The function ρ in Assumption 3A measures the squared length of the regularization path, which plays a central role in our bounds. The second condition in Assumption 3C is satisfied (with a possibly different constant) if ℓ is sufficiently differentiable and $\ell''(\cdot) \leq C g(\cdot)$ for a constant $C > 0$, hence it reflects the self-boundedness of ℓ. Finally, one can verify that Assumptions 3A and 3D imply that $\ell(z) \leq \ell(0) \exp(-C_e^{-1} z)$ for $z \geq 0$, that is, $\ell(\cdot)$ is exponentially tailed.

Proposition 5 provides three examples of loss functions. The proof is included in Appendix G.

Proposition 5 (Examples) *The following loss functions satisfy (parts of) Assumption 3.*

1. *The logistic loss, $\ell_{\log}(z) := \ln(1 + \exp(-z))$, satisfies Assumption 3 with $C_g = 1$, $C_\beta = e/2$, $C_e = 2$, and*

$$\rho(\lambda) \leq \rho_{\log}(\lambda) := 1 + \ln^2(\lambda).$$

2. *The "flattened" exponential loss with temperature $a > 0$,*

$$\ell_{\exp}(z) := \begin{cases} e^{-az} & z > 0, \\ 1 - az & z \leq 0, \end{cases}$$

satisfies Assumption 3 with $C_g = a$, $C_\beta = \max\{a, ae^a/2, 1\}$, $C_e = 1/a$, and

$$\rho(\lambda) \leq \rho_{\exp}(\lambda) := 1 + \ln^2(\lambda)/a^2.$$

3. *The "flattened" polynomial loss of degree $a > 0$,*

$$\ell_{\text{poly}}(z) := \begin{cases} (1 + z)^{-a} & z > 0, \\ 1 - az & z \leq 0, \end{cases}$$

satisfies Assumptions 3A to 3C with $C_g = a$, $C_\beta = \max\{a, (a + 1)2^a\}$, and

$$\rho(\lambda) \leq \rho_{\text{poly}}(\lambda) := 2\lambda^{2/(a+2)}.$$

The (flattened) polynomial loss is introduced in (Ji et al., 2020; Ji and Telgarsky, 2021) and is later used by Wang et al. (2022a) to improve importance weighting in distribution shift problems. In the second and third examples, we flatten the negative parts of the exponential and polynomial losses to satisfy Assumption 3B. This is necessary in our setting because, under (unflattened) exponential loss, large stepsize GD may oscillate catastrophically as shown by Wu et al. (2023).

A two-layer network. We consider a two-layer network[1] under a loss function ℓ, defined as

$$L(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i f(\mathbf{x}_i;\mathbf{w})\big), \quad f(\mathbf{x};\mathbf{w}) := \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \phi(\mathbf{x}^\top \mathbf{w}^{(s)}), \quad \phi(\cdot) := \max\{\cdot, 0\}, \quad (4)$$

where the trainable parameters are denoted by $\mathbf{w} \in \mathbb{R}^{md}$, a stack of $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)} \in \mathbb{R}^d$. We define $\phi'(0) := 0$ for concreteness but our analysis can be easily generalized to other subderivatives, that is, $\phi'(0) = c$ for any $c \in [0, 1]$. Here, we follow the standard NTK setup (Du et al., 2018; Ji and Telgarsky, 2019) and assume that $(a_s)_{s=1}^m$ are fixed parameters satisfying[2]

$$a_1, \ldots, a_m \in \{\pm 1\}, \quad \left|\sum_{s=1}^{m} a_s\right| \le C_a\sqrt{m} \text{ for some constant } C_a > 0. \quad (5)$$

It is clear that the optimization problem in (4) is non-convex.

We analyze the training of a nonlinear predictor (see f in (4)) in the NTK regime (Jacot et al., 2018). The key idea is that, when the network width m is large, the predictors induced by (GD) iterates (in finite steps) are close to the gradient dynamics in a *reproducing kernel Hilbert space* (RKHS), under the so-called *neural tangent kernel* (NTK). Therefore, the nonlinear predictors induced by the GD iterates are approximately linear in the NTK RKHS, allowing us to recycle our techniques for analyzing linear predictors in Section 2.

We reformulate our separability assumption within the NTK RKHS (Ji and Telgarsky, 2019).

Assumption 4 (NTK separability) *Assume the training data $(\mathbf{x}_i, y_i)_{i=1}^n$ satisfies*

A. for every $i = 1, \ldots, n$, $\|\mathbf{x}_i\| \le 1$ and $y_i \in \{\pm 1\}$;

B. there is a margin $\gamma > 0$ and a map $\chi : \mathbb{R}^d \to \mathbb{R}^d$ such that $\|\chi(\cdot)\| \le 1$ and

$$\min_{i=1,\ldots,n} \mathbb{E}_{\mathbf{u}\sim\mathcal{N}(0,\mathbf{I}_d)} \big\langle y_i \phi'(\mathbf{x}_i^\top \mathbf{u})\mathbf{x}_i,\, \chi(\mathbf{u})\big\rangle \ge \gamma,$$

that is, the dataset is separable in the NTK RKHS space.

Assumption 4 corresponds to Assumption 2.1 in (Ji and Telgarsky, 2019). Assumption 4 requires the dataset being linearly separable under the infinite-dimensional NTK feature, that is, $\phi'(\mathbf{x}^\top \mathbf{u})\mathbf{x}$ indexed by $\mathbf{u}$, which corresponds to the gradient of $f(\mathbf{x};\cdot)$ at the Gaussian random initialization as the network width $m \to \infty$ (ignoring $\frac{1}{\sqrt{m}} a_s$'s for simplicity). Assumption 4 is satisfied when the dataset is linearly separable (as in Assumption 2, with a possibly different margin parameter). In addition, Assumption 4 can capture non-linearly separable datasets. We refer the reader to Section 5 in (Ji and Telgarsky, 2019) for more discussion.

We are ready to state our theorem for general loss functions and two-layer networks.

1. We present our NTK results in a two-layer network for conciseness. Our results are ready to be extended to deep networks by combining our techniques with standard NTK arguments (Allen-Zhu et al., 2019; Chen et al., 2020).
2. Some papers assume that $(a_s)_{s=1}^m$ are uniformly sampled from $\{\pm 1\}$. In this case, $\left|\sum_{s=1}^m a_s\right| \le \sqrt{2m\ln(2/\delta)}$ with probability at least $1-\delta$. So we recover theirs by setting $C_a := \sqrt{2\ln(2/\delta)}$ in (5) and applying a union bound.

Theorem 6 (General losses and NTK) *Consider (GD) with stepsize $\eta > 0$ for learning a two-layer network (4) under a loss function ℓ that satisfies Assumptions 3A and 3B. Suppose (5) and Assumption 4 hold, and the network is initialized by*

$$\mathbf{w}_0 \sim \mathcal{N}(0, \mathbf{I}_{md}).$$

Let T be the maximum number of steps. Suppose the network width m is at least

$$m \geq \left(\frac{30R^{1/3} + 10\ln^{1/4}(n/\delta)}{\gamma}\right)^6, \quad \text{where } R := 6\frac{\sqrt{\rho(\gamma^2\eta T)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma}.$$

Then with probability at least $1 - 3\delta$ over the randomness of initialization, the following holds:

- ***Lazy training.*** *For every $t \leq T$, we have*

$$\|\mathbf{w}_t - \mathbf{w}_0\| \leq R.$$

- ***The EoS phase.*** *For every $t \leq T$ (and in particular in the EoS phase), we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \leq 9\frac{\rho(\gamma^2\eta t) + \left(C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g\right)^2}{\gamma^2\eta t}.$$

- ***The stable phase.*** *Assume the loss ℓ also satisfies Assumption 3C. If $s < T$ is such that*

$$L(\mathbf{w}_s) \leq \min\left\{\frac{1}{12C_\beta^2\eta}, \frac{\ell(0)}{n}\right\}, \tag{6}$$

  then (GD) is in the stable phase, that is, $\big(L(\mathbf{w}_t)\big)_{t=s}^{T}$ decreases monotonically, and moreover,

$$L(\mathbf{w}_t) \leq 15\frac{\rho(\gamma^2\eta(t-s))}{\gamma^2\eta(t-s)}, \quad t \in (s, T].$$

- ***Phase transition time.*** *There exists a constant $C_1 > 0$ that only depends on C_g, C_β, C_a, $\ell(0)$, and $\ln(1/\delta)$ such that the following holds. Let*

$$\tau := \frac{1}{\gamma^2}\max\left\{\frac{\psi^{-1}\big(C_1(\eta+n)\big)}{\eta}, C_1(\eta+n)\eta\right\}, \quad \text{where } \psi(\lambda) := \frac{\lambda}{\rho(\lambda)}.$$

  If $\tau \leq T$, then (6) holds for some $s \leq \tau$.

- ***Phase transition time under an exponential tail.*** *Assume the loss ℓ further satisfies Assumption 3D, then there exists a constant $C_2 > 0$ that only depends on C_e, C_g, C_β, C_a, $\ell(0)$, and $\ln(1/\delta)$ such that the following holds. Let*

$$\tau := \frac{C_2}{\gamma^2}\max\big\{\eta, n\ln(n)\big\}.$$

  If $\tau \leq T$, then (6) holds for some $s \leq \tau$.

The proof of Theorem 6 is deferred to Appendix H. Theorem 6 characterizes the behaviors of large stepsize GD for a wide two-layer network under general classification losses. We remark that the proof of Theorem 6 directly adapts when replacing the two-layer network with a linear predictor. For completeness, we state a variant of Theorem 6 for GD in linear classification under general losses as Theorem 35 in Appendix I.

Table 1: Effects of the loss functions and stepsizes in Theorem 6.

loss function	logistic / flattened exponential		flattened polynomial of degree a		
degree condition	N/A		$a > 0$	$0 < a \le 1$	$a > 1$
$\rho(\lambda)$	$\Theta\big(\ln^2(\lambda)\big)$		$\Theta\big(\lambda^{\frac{2}{a+2}}\big)$		
stepsize η	1	$\Theta(T)$	1	$\Theta(T^{\frac{a}{2}})$	$\Theta(T^{\frac{1}{2}})$
width m	$\Omega\big(\ln^2(T)\big)$	$\Omega(T^2)$	$\Omega\big(T^{\frac{2}{a+2}}\big)$	$\Omega(T)$	$\Omega(T)$
phase transition time s	N/A	$\le T/2$	N/A	$\le T/2$	$\le T/2$
loss $L(\mathbf{w}_T)$	$\mathcal{O}\big(\ln^2(T)/T\big)$	$\mathcal{O}\big(\ln^2(T)/T^2\big)$	$\mathcal{O}\big(T^{\frac{-a}{a+2}}\big)$	$\mathcal{O}(T^{-\frac{a}{2}})$	$\mathcal{O}\big(T^{\frac{-3a}{2a+4}}\big)$

Convergence rate. The final convergence rate of GD in Theorem 6 depends on the loss functions and the stepsizes. We summarize and compare several key examples in Table 1, where we assume the total number of steps T is large and treat all other instance specific quantities (such as n and γ) as constants. Results in Table 1 are direct consequences of Theorem 6 and Proposition 5. In particular, we see that GD attains an improved loss by using a large stepsize that balances the length of the EoS and the stable phases.

Loss conditions. Notably, Theorem 6 (also Theorem 35 in Appendix I) reflects the role of each loss function condition in Assumption 3. Specifically, a general classification loss is specified by Assumption 3A. Then the EoS phase bound in Theorem 6 holds when the loss satisfies an additional Lipschitz condition (Assumption 3B), and the stable phase bound holds if the loss further satisfies the self-boundedness condition (Assumption 3C). Finally, an exponential tail condition (Assumption 3D) implies a better phase transition bound.

The width condition. The width condition in Theorem 6 depends on the loss function ℓ and the stepsize η. When we specialize ℓ to the logistic loss (so $\rho(\lambda) = \mathcal{O}(\ln^2(\lambda))$ by Proposition 5) and the stepsize to a constant ($\eta = \Theta(1)$), the width condition is *polylogarithmic* in the number of samples and the number of iterates, and the achieved convergence rate is $\tilde{\mathcal{O}}(1/t)$. This recovers the main result of Ji and Telgarsky (2019).

On the other hand, the width condition in Theorem 6 is at least $\Omega(\eta^2)$, which will become *polynomial* in the number of steps when GD is equipped with a polynomially large stepsize $\eta = \mathrm{poly}(T)$. This is in stark contrast to the previous example where $\eta = \Theta(1)$ and the width only needs to be polylogarithmic in T, suggesting that a larger stepsize helps GD escape the NTK regime in the early optimization phase. We leave it as a future direction to study the early-phase feature learning of GD caused by a large stepsize.

The loss function affects the width condition in Theorem 6 through $\rho(\gamma^2\eta T)$. For instance, the width condition becomes $m \ge \Omega(T^{2/(a+2)})$ for the constant stepsize and the flattened polynomial loss of degree $a > 0$ (see Table 1). Such results are new in the NTK literature to our knowledge.

5. Conclusion

We study constant-stepsize GD for training linear and nonlinear predictors (in the NTK regime) under general classification loss functions and constant-stepsize SGD for logistic regression, assuming suitable separability conditions on the dataset. We show GD and SGD converge even with a large stepsize that leads to a locally oscillatory loss. Moreover, we show that a large stepsize allows GD to attain an accelerated loss by undergoing an unstable initial phase, which cannot be attained if the stepsize is small such that GD converges monotonically. We leave for future work to relax the separability conditions.

Acknowledgments

We thank the anonymous reviewers and area chairs for their helpful comments. We thank Fabian Pedregosa for his suggestions on an early draft. We gratefully acknowledge the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639 respectively, NSF Grant DMS 2015341 and NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI).

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning*, pages 247–257. PMLR, 2022.

Kwangjun Ahn, Sebastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging II: Silver stepsize schedule for smooth convex optimization. *arXiv preprint arXiv:2309.16530*, 2023.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel methods—support vector learning*, pages 43–54, 1999.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *International Conference on Machine Learning*, pages 4330–4391. PMLR, 2023.

Xuxing Chen, Krishnakumar Balasubramanian, Promit Ghosal, and Bhavya Agrawalla. From stability to chaos: Analyzing gradient descent dynamics in quadratic regression. *arXiv preprint arXiv:2310.01687*, 2023.

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

Jeremy 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*, 2020.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2022.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6(6):899–929, 1978.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Allan Grønlund, Lior Kamma, and Kasper Green Larsen. Near-tight margin-based generalization bounds for support vector machines. In *International Conference on Machine Learning*, pages 3779–3788. PMLR, 2020.

Steve Hanneke and Aryeh Kontorovich. Stable sample compression schemes: New applications and an optimal SVM margin bound. In *Algorithmic Learning Theory*, pages 697–721. PMLR, 2021.

Elad Hazan. *Introduction to online convex optimization*. MIT Press, 2022.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations*, 2019.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

Jonathan Kelner, Annie Marsden, Vatsal Sharan, Aaron Sidford, Gregory Valiant, and Honglin Yuan. Big-step-little-step: Efficient gradient methods for objectives with multiple scales. In *Conference on Learning Theory*, pages 2431–2540. PMLR, 2022.

Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. *Advances in Neural Information Processing Systems*, 33:2625–2638, 2020.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17684–17744. PMLR, 23–29 Jul 2023.

Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rate. In *The Twelfth International Conference on Learning Representations*, 2023.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35: 34689–34708, 2022.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *arXiv preprint arXiv:2204.11326*, 2022.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pages 6702–6712. PMLR, 2020.

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

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

Albert BJ Novikoff. On convergence proofs for perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*. New York, NY, 1962.

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

Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

Matus Telgarsky. Deep learning theory lecture notes, 2021. URL https://mjt.cs.illinois.edu/dlt/.

Ke Alexander Wang, Niladri Shekhar Chatterji, Saminul Haque, and Tatsunori Hashimoto. Is importance weighting incompatible with interpolating classifiers? In *International Conference on Learning Representations*, 2022a.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022b.

Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. *arXiv preprint arXiv:2310.17087*, 2023.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 35:9983–9994, 2022c.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lei Wu and Chao Ma. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2022.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018.

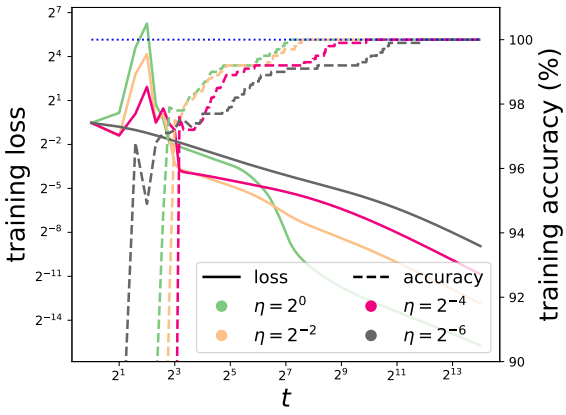


Figure 2: Training loss and training accuracy of (GD) for logistic regression on a subset of MNIST. This is a replication of Figure 1 with training accuracy curves. The plots show that the stable phase is entered before reaching 100% training accuracy. Moreover, GD with larger stepsizes reaches perfect training accuracy faster.

Appendix A. Additional Simulation

In Figure 2, we recreate Figure 1 with training accuracy curves. The plots show that the stable phase is entered before reaching 100% training accuracy. Moreover, GD with larger stepsizes reaches perfect training accuracy faster. The simulations suggest that the phase transition time bound given in Theorem 1, especially the dependence on the sample size n, might be improvable.

Appendix B. Proof of Theorem 1

Throughout this part, we assume Assumption 1 holds. In addition, without loss of generality, we assume $y = 1$. We define

$$L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i^\top \mathbf{w}), \quad \ell(t) = \ln(1 + e^{-t}), \quad G(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} |\ell'(\mathbf{x}_i^\top \mathbf{w})|.$$

The following lemma presents a new split optimization method for handling large stepsizes.

Lemma 7 (A split optimization bound) *For every $\eta > 0$, $\mathbf{w}_0$, and $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ such that*

$$\mathbf{u}_2 = \frac{\eta}{2\gamma}\mathbf{w}_*,$$

we have

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le L(\mathbf{u}_1) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t}, \quad t \ge 1.$$

Proof (of Lemma 7) Using the definition of GD, we have

$$\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 = \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L(\mathbf{w}_t), \mathbf{u} - \mathbf{w}_t\rangle + \eta^2 \left\|\nabla L(\mathbf{w}_t)\right\|_2^2$$

$$= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L(\mathbf{w}_t), \mathbf{u}_1 - \mathbf{w}_t\rangle + \eta\Big(2\langle \nabla L(\mathbf{w}_t), \mathbf{u}_2\rangle + \eta\big\|\nabla L(\mathbf{w}_t)\big\|_2^2\Big).$$

Plugging $\mathbf{u}_2$ and ∇L, we can show the second term is always non-positive:

$$\begin{aligned}
&2\langle \nabla L(\mathbf{w}_t), \mathbf{u}_2\rangle + \eta\big\|\nabla L(\mathbf{w}_t)\big\|_2^2 \\
&= \frac{2}{n}\sum_{i=1}^n \ell'(\mathbf{w}_t^\top \mathbf{x}_i)\langle \mathbf{x}_i, \mathbf{u}_2\rangle + \eta\left\|\frac{1}{n}\sum_{i=1}^n \ell'(\mathbf{w}_t^\top \mathbf{x}_i)\mathbf{x}_i\right\|_2^2 \\
&\le \frac{2\gamma\|\mathbf{u}_2\|}{n}\sum_{i=1}^n \ell'(\mathbf{w}_t^\top \mathbf{x}_i) + \eta\left(\frac{1}{n}\sum_{i=1}^n \ell'(\mathbf{w}_t^\top \mathbf{x}_i)\right)^2 && \text{by } \mathbf{x}_i^\top \mathbf{w}_* \ge \gamma \text{ and } \|\mathbf{x}_i\| \le 1 \\
&\le (-2\gamma\|\mathbf{u}_2\| + \eta)\cdot \frac{1}{n}\sum_{i=1}^n |\ell'(\mathbf{w}_t^\top \mathbf{x}_i)| && \text{since } |\ell'| \le 1 \\
&\le 0. && \text{by the choice of } \mathbf{u}_2
\end{aligned}$$

So we have

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &\le \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L(\mathbf{w}_t), \mathbf{u}_1 - \mathbf{w}_t\rangle \\
&\le \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\big(L(\mathbf{u}_1) - L(\mathbf{w}_t)\big). && \text{by convexity}
\end{aligned}$$

Telescoping the sum from 0 to t, we obtain

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta} + \sum_{k=0}^{t-1} L(\mathbf{w}_k) \le tL(\mathbf{u}_1) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta},$$

which completes the proof. ■

Applying the above lemma with an appropriate comparator, we can get the following bounds on the risk and parameter norm.

Lemma 8 (Parameter and risk bounds in the EoS phase) *Let $\mathbf{w}_0 = 0$. For every $\eta > 0$, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le \frac{1 + \ln^2(\gamma^2\eta t) + \eta^2/4}{\gamma^2\eta t}, \quad \|\mathbf{w}_t\| \le \frac{\sqrt{2} + 2\ln(\gamma^2\eta t) + \eta}{\gamma}, \quad t \ge 1.$$

Proof (of Lemma 8) In Lemma 7, choose

$$\mathbf{w}_0 = 0, \quad \mathbf{u}_1 = \frac{\ln(\gamma^2\eta t)}{\gamma}\mathbf{w}_*,$$

and check that

$$\begin{aligned}
L(\mathbf{u}_1) &\le \frac{1}{n}\sum_{i=1}^n \exp(-\mathbf{x}_i^\top \mathbf{u}_1) \\
&\le \exp(-\gamma\|\mathbf{u}_1\|)
\end{aligned}$$

$$\leq \frac{1}{\gamma^2 \eta t},$$

and that

$$\|\mathbf{u}\| = \|\mathbf{u}_1 + \mathbf{u}_2\| = \frac{\ln(\gamma^2 \eta t) + \eta/2}{\gamma}.$$

We then apply Lemma 7 to get

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \leq L(\mathbf{u}_1) + \frac{\|\mathbf{u}\|^2}{2\eta t} \leq \frac{1 + \ln^2(\gamma^2 \eta t) + \eta^2/4}{\gamma^2 \eta t},$$

and

$$\|\mathbf{w}_t\| \leq \|\mathbf{w}_t - \mathbf{u}\| + \|\mathbf{u}\| \leq \sqrt{2\eta t L(\mathbf{u}_1)} + 2\|\mathbf{u}\| \leq \frac{\sqrt{2} + 2\ln(\gamma^2 \eta t) + \eta}{\gamma}.$$

These complete the proof. ■

The following lemma controls the gradient potential. We remark that the bound on gradient potential in Lemma 9 does not scale with η, while the bound on risk in Lemma 8 linearly scales with η. This difference will be crucial for getting a sharp bound on the phase transition time.

Lemma 9 (Gradient potential bound in the EoS phase) *Let $\mathbf{w}_0 = 0$. For every $\eta > 0$, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} G(\mathbf{w}_k) \leq \frac{\sqrt{2} + 2\ln(\gamma^2 \eta t) + \eta}{\gamma^2 \eta t}, \quad t \geq 1.$$

Proof (of Lemma 9) This is from the perceptron argument (Novikoff, 1962). Specifically,

$$\begin{aligned}
\langle \mathbf{w}_{t+1}, \mathbf{w}_* \rangle &= \langle \mathbf{w}_t, \mathbf{w}_* \rangle - \eta \langle \nabla L(\mathbf{w}_t), \mathbf{w}_* \rangle \\
&= \langle \mathbf{w}_t, \mathbf{w}_* \rangle - \frac{\eta}{n}\sum_{i=1}^{n} \ell'(\mathbf{x}_i^\top \mathbf{w}_t) \langle \mathbf{x}_i, \mathbf{w}_* \rangle \\
&\geq \langle \mathbf{w}_t, \mathbf{w}_* \rangle - \frac{\gamma\eta}{n}\sum_{i=1}^{n} \ell'(\mathbf{x}_i^\top \mathbf{w}_t) \\
&= \langle \mathbf{w}_t, \mathbf{w}_* \rangle + \gamma\eta G(\mathbf{w}_t).
\end{aligned}$$

So we have

$$\frac{1}{t}\sum_{k=0}^{t-1} G(\mathbf{w}_k) \leq \frac{\langle \mathbf{w}_t, \mathbf{w}_* \rangle - \langle \mathbf{w}_0, \mathbf{w}_* \rangle}{\gamma\eta t} \leq \frac{\|\mathbf{w}_t - \mathbf{w}_0\|}{\gamma\eta t}.$$

We complete the proof by applying the parameter bound in Lemma 8. ■

For logistic regression, the local sharpness is related to the loss value. So the landscape will become sufficiently flat when the loss is sufficiently small, allowing GD to fall into the stable convergence phase. Our next lemma formally justifies this discussion.

Lemma 10 (Stable phase) *If there exists $s \geq 0$ such that*

$$L(\mathbf{w}_s) \leq \frac{2}{\eta},$$

then for every $t \geq s$, $L(\mathbf{w}_t)$ is non-increasing.

Proof (of Lemma 10) We verify that $\ln L(\mathbf{w})$ is 1-smooth (this is used by Ji and Telgarsky (2018)) by direct computation:

$$\begin{aligned}
\frac{\mathrm{d}^2}{(\mathrm{d}\mathbf{w})^2} \ln L(\mathbf{w}) &= \frac{1}{L(\mathbf{w})^2}\Big(\nabla^2 L(\mathbf{w}) \cdot L(\mathbf{w}) - \big(\nabla L(\mathbf{w})\big)^{\otimes 2}\Big) \\
&\preceq \frac{1}{L(\mathbf{w})}\nabla^2 L(\mathbf{w}) \\
&\preceq \frac{1}{L(\mathbf{w})} L(\mathbf{w}) \cdot \mathbf{I} = \mathbf{I}. && \text{since } |\ell''| \leq \ell \text{ and } \|\mathbf{x}_i\| \leq 1
\end{aligned}$$

Using the 1-smoothness of $\ln L(\cdot)$, we get

$$\begin{aligned}
\ln L(\mathbf{w}_{t+1}) &\leq \ln L(\mathbf{w}_t) + \left\langle \frac{\nabla L(\mathbf{w}_t)}{L(\mathbf{w}_t)}, \mathbf{w}_{t+1} - \mathbf{w}_t \right\rangle + \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&= \ln L(\mathbf{w}_t) - \frac{\eta}{L(\mathbf{w}_t)}\|\nabla L(\mathbf{w}_t)\|^2 + \frac{\eta^2}{2}\|\nabla L(\mathbf{w}_t)\|^2 \\
&= \ln L(\mathbf{w}_t) - \eta\left(\frac{1}{L(\mathbf{w}_t)} - \frac{\eta}{2}\right)\|\nabla L(\mathbf{w}_t)\|^2.
\end{aligned}$$

The above implies

$$L(\mathbf{w}_{t+1}) \leq L(\mathbf{w}_t) \;\text{ if }\; L(\mathbf{w}_t) \leq \frac{2}{\eta}.$$

We complete the proof by induction and the condition that $L(\mathbf{w}_s) \leq 2/\eta$. ■

The next lemma provides a risk bound for GD in the stable phase. Notably, the bound depends on the initial condition ($\mathbf{w}_s$) through an exponential potential (see the definition of $F(\cdot)$) instead of a quadratic potential. We will see later that $F(\mathbf{w}_s)$ can be made a constant while $\|\mathbf{w}_s\|^2$ might be as large as $\Theta(\eta^2)$.

Lemma 11 (A risk bound in the stable phase) *Suppose there exists a time s such that*

$$L(\mathbf{w}_s) \leq \frac{1}{\eta}.$$

Then for every $t \geq s$, we have

$$L(\mathbf{w}_t) \leq \frac{2F(\mathbf{w}_s) + \ln^2(\gamma^2\eta(t-s))}{\gamma^2\eta(t-s)}, \;\; \textit{where} \;\; F(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \exp(-\mathbf{x}_i^\top \mathbf{w}).$$

Proof (of Lemma 11) The assumption enables Lemma 10 for all $t \geq s$. Therefore we have

$$\text{for every } t \geq s, \quad L(\mathbf{w}_t) \leq L(\mathbf{w}_{t-1}) \leq \cdots \leq L(\mathbf{w}_s) \leq \frac{1}{\eta}.$$

Because the logistic loss with $\|\mathbf{x}\| \leq 1$ satisfies $\|\nabla L(\cdot)\| \leq L(\cdot)$, for a comparator $\mathbf{u}$ and $t \geq s$, we have

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L(\mathbf{w}_t), \mathbf{u} - \mathbf{w}_t\rangle + \eta^2 \|\nabla L(\mathbf{w}_t)\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\big(L(\mathbf{u}) - L(\mathbf{w}_t)\big) + \eta^2 L(\mathbf{w}_t)^2 && \text{since } |\ell'| \leq \ell \\
&\leq \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\big(L(\mathbf{u}) - L(\mathbf{w}_t)\big) + \eta L(\mathbf{w}_t) && \text{since } L(\mathbf{w}_t) \leq 1/\eta \\
&= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta L(\mathbf{u}) - \eta L(\mathbf{w}_t).
\end{aligned}$$

Telescoping the sum from s to t, we get

$$\|\mathbf{w}_t - \mathbf{u}\|^2 + \eta \sum_{k=s}^{t-1} L(\mathbf{w}_k) \leq \|\mathbf{w}_s - \mathbf{u}\|^2 + 2\eta(t-s)L(\mathbf{u}),$$

that is,

$$\begin{aligned}
\frac{1}{t-s}\sum_{k=s}^{t-1} L(\mathbf{w}_k) &\leq 2L(\mathbf{u}) + \frac{\|\mathbf{w}_s - \mathbf{u}\|^2 - \|\mathbf{w}_t - \mathbf{u}\|^2}{\eta(t-s)} \\
&\leq 2L(\mathbf{u}) + \frac{\|\mathbf{w}_s - \mathbf{u}\|^2}{\eta(t-s)}.
\end{aligned}$$

Choose

$$\mathbf{u} = \mathbf{w}_s + \mathbf{u}_1, \quad \mathbf{u}_1 = \frac{\ln\big(\gamma^2\eta(t-s)\big)}{\gamma}\mathbf{w}_*,$$

and verify that

$$\|\mathbf{w}_s - \mathbf{u}\| = \frac{\ln\big(\gamma^2\eta(t-s)\big)}{\gamma},$$

and

$$\begin{aligned}
L(\mathbf{u}) \leq F(\mathbf{u}) &= \frac{1}{n}\sum_{i=1}^{n} \exp(-\langle \mathbf{u}, \mathbf{x}\rangle) \\
&= \frac{1}{n}\sum_{i=1}^{n} \exp(-\langle \mathbf{w}_s, \mathbf{x}\rangle - \langle \mathbf{u}_1, \mathbf{x}\rangle) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \exp(-\langle \mathbf{w}_s, \mathbf{x}\rangle) \cdot \frac{1}{\gamma^2\eta(t-s)} \\
&= \frac{F(\mathbf{w}_s)}{\gamma^2\eta(t-s)}.
\end{aligned}$$

Bringing these back, we obtain

$$\frac{1}{t-s}\sum_{k=1}^{t-s} L(\mathbf{w}_{s+k}) \le 2L(\mathbf{u}) + \frac{\|\mathbf{w}_s - \mathbf{u}\|^2}{\eta(t-s)} \le \frac{2F(\mathbf{w}_s) + \ln^2\left(\gamma^2\eta(t-s)\right)}{\gamma^2\eta(t-s)}.$$

We complete the proof using the monotonicity of $L(\mathbf{w}_t)$ for $t \ge s$ by Lemma 10. ■

The following lemma provides a bound on the phase transition time. The bound is linear in η because we use the gradient potential bound in Lemma 9 instead of the risk bound in Lemma 8.

Lemma 12 (Phase transition time) *Let $\mathbf{w}_0 = 0$. For every $\eta > 0$, let*

$$\tau := \frac{60}{\gamma^2}\max\left\{\eta,\ n,\ e,\ \frac{\eta+n}{\eta}\ln\frac{\eta+n}{\eta}\right\}.$$

Then there exists $0 \le s \le \tau$ such that

$$L(\mathbf{w}_s) \le \frac{1}{\eta}, \quad F(\mathbf{w}_s) \le 1.$$

Proof (of Lemma 12) Applying Lemma 9 with $t = \tau$, we have

$$\begin{aligned}
\frac{1}{\tau}\sum_{k=0}^{\tau-1} G(\mathbf{w}_k) &\le \frac{\sqrt{2} + 2\ln(\gamma^2\eta\tau) + \eta}{\gamma^2\eta\tau} \\
&\le \frac{\sqrt{2} + 2\ln(\gamma^2\tau) + 3\eta}{\eta\gamma^2\tau} && \text{since } \ln(\eta) \le \eta \\
&\le \frac{1}{2(\eta+n)} \le \min\left\{\frac{1}{2\eta}, \frac{1}{2n}\right\},
\end{aligned}$$

where the third inequality is by verifying that

$$\begin{aligned}
\frac{\sqrt{2} + 2\ln(\gamma^2\tau)}{\gamma^2\tau} \le \frac{\eta}{4(\eta+n)} \quad &\text{if} \quad \gamma^2\tau \ge 60\frac{\eta+n}{\eta}\max\left\{\ln\frac{\eta+n}{\eta},\ 1\right\}, \\
\frac{3}{\gamma^2\tau} \le \frac{1}{4(\eta+n)} \quad &\text{if} \quad \gamma^2\tau \ge 12(\eta+n),
\end{aligned}$$

and that the two conditions are satisfied because

$$\gamma^2\tau := 60\max\left\{\eta,\ n,\ e, \frac{\eta+n}{\eta}\ln\frac{\eta+n}{\eta}\right\} \ge \max\left\{12(\eta+n),\ 60\frac{\eta+n}{\eta}\max\left\{\ln\frac{\eta+n}{\eta}, 1\right\}\right\}.$$

So there exists $s \le \tau$ such that

$$G(\mathbf{w}_s) \le \min\left\{\frac{1}{2\eta}, \frac{1}{2n}\right\}.$$

The second bound ensures that

$$\text{for every } i, \quad \frac{1}{n}\cdot\frac{1}{1+\exp(\mathbf{x}_i^\top\mathbf{w}_s)} \le G(\mathbf{w}_s) \le \frac{1}{2n},$$

that is,

$$\text{for every } i, \quad \mathbf{x}_i^\top \mathbf{w}_s \geq 0.$$

So we have

$$F(\mathbf{w}_s) = \frac{1}{n}\sum_{i=1}^n \frac{2}{2\exp(\mathbf{x}_i^\top \mathbf{w}_s)} \leq \frac{1}{n}\sum_{i=1}^n \frac{2}{1+\exp(\mathbf{x}_i^\top \mathbf{w}_s)} = 2G(\mathbf{w}_s) \leq \min\left\{\frac{1}{\eta},\ \frac{1}{n}\right\}.$$

Furthermore, by definitions of L and F we have

$$L(\mathbf{w}_s) \leq F(\mathbf{w}_s) \leq \min\left\{\frac{1}{\eta},\ \frac{1}{n}\right\} \leq 1,$$

which completes the proof. ■

The above lemmas imply Theorem 1.

Proof (of Theorem 1) It follows from Lemmas 8, 11 and 12. ■

Appendix C. Proof of Corollary 2

Proof (of Corollary 2) We first verify $\tau \leq T/2$ for

$$\tau := \frac{60}{\gamma^2}\max\left\{\eta,\ n,\ e,\ \frac{\eta+n}{\eta}\ln\frac{\eta+n}{\eta}\right\}.$$

The first term is

$$\frac{60}{\gamma^2}\eta = \frac{T}{2}, \qquad \text{since } \eta := \frac{\gamma^2}{120}T$$

the second term is

$$\frac{60n}{\gamma^2} \leq \frac{T}{2}, \qquad \text{since } T \geq \frac{120\max\{e,n\}}{\gamma^2},$$

the last term is

$$\frac{60e}{\gamma^2} \leq \frac{T}{2}, \qquad \text{since } T \geq \frac{120\max\{e,n\}}{\gamma^2},$$

and the third term is less than $2\ln(2) \leq T/2$ (note that $T \geq 120$ since $\gamma < 1$ and $n \geq 1$) because

$$\begin{aligned}
\frac{\eta+n}{\eta} &= \frac{T\gamma^2/120 + n}{T\gamma^2/120} && \text{since } \eta := \frac{\gamma^2}{120}T \\
&\leq \frac{T\gamma^2/120 + T\gamma^2/120}{T\gamma^2/120} && \text{since } T \geq \frac{120n}{\gamma^2} \\
&= 2.
\end{aligned}$$

We have verified that $\tau \leq T/2$.

Next, we apply Theorem 1 with $s = \tau \le T/2$ and get

$$\begin{aligned} L(\mathbf{w}_T) &\le \frac{2F(\mathbf{w}_s) + \ln^2(\gamma^2\eta(T-s))}{\gamma^2\eta(T-s)} \\ &\le \frac{2 + \ln^2(\gamma^2\eta T/2)}{\gamma^2\eta T/2} && \text{since } F(\mathbf{w}_s) \le 1 \text{ and } s \le T/2 \\ &= \frac{2 + \ln^2(\gamma^4 T^2/240)}{\gamma^4 T^2/240} && \text{since } \eta := \frac{\gamma^2}{120}T \\ &\le \frac{480\ln^2(\gamma^4 T^2)}{\gamma^4 T^2}, \end{aligned}$$

which completes the proof. ■

Appendix D. Proof of Theorem 3

The following lemma suggests an $\Omega(1/(\eta t))$ lower bound on the risk for GD with a fixed stepsize.

Lemma 13 (A lower bound) *Suppose that $(\mathbf{x}_i, y_i)_{i=1}^n$ are linearly separable with max-margin direction $\mathbf{w}_*$ and margin γ. Denote*

$$\bar{\mathbf{x}}_i = \mathbf{x}_i - \langle \mathbf{x}_i, \mathbf{w}_* \rangle \mathbf{w}_*, \quad i = 1, \dots, n.$$

Assume that $(\bar{\mathbf{x}}_i, y_i)_{i=1}^n$ are non-separable, that is, there exists $b > 0$ such that

$$\sup_{\bar{\mathbf{v}} \in \mathbb{R}^{d-1}, \|\bar{\mathbf{v}}\|=1} \min_i \langle y_i \bar{\mathbf{x}}_i, \bar{\mathbf{v}} \rangle \le -b.$$

Then for every stepsize $\eta > 0$, we have

$$L(\mathbf{w}_t) \ge \frac{1}{c_0 \eta \exp(c_0 \eta) t}, \quad t \ge 1,$$

where $c_0 > 0$ is a function of γ, b and n but is independent of η and t.

Proof (of Lemma 13) Assume $y_i = 1$ without loss of generality. Define

$$w_t = \langle \mathbf{w}_t, \mathbf{w}_* \rangle, \quad \bar{\mathbf{w}}_t = \mathbf{w}_t - \langle \mathbf{w}_t, \mathbf{w}_* \rangle \mathbf{w}_*.$$

We call some results in (Wu et al., 2023). By Lemma B.3 in (Wu et al., 2023) (replacing their ηn by η since they do not take average in their loss definition), we get

$$\|\bar{\mathbf{w}}_t\| \le \max\{4n/b, \eta n/b\} + \eta \le c_1 \eta,$$

where c_1 is a positive constant independent of η. By Lemma B.7 in (Wu et al., 2023), we get

$$w_t \le \frac{1}{\gamma} \ln\left(\left(e\frac{\eta}{n}\gamma^2 G_{\max} + e\frac{\eta}{n}\gamma H_{\max} \right)(t+1) \right), \quad t \ge 0,$$

where $G_{\max}$ and $H_{\max}$ can be upper bounded by (see their Definition 3 in Appendix B.3)

$$G_{\max} + H_{\max} = \sup_{\|\bar{\mathbf{w}}_t\| \le c_1\eta} \sum_{i=1}^{n} \exp(-\bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}) \le n \exp(c_1\eta).$$

So we have

$$w_t \le \frac{1}{\gamma} \ln\big(e\gamma\eta \exp(c_1\eta)(t+1)\big), \quad t \ge 0.$$

The above implies that

$$\begin{aligned}
L(\mathbf{w}_t) &= \frac{1}{n}\sum_{i=1}^{n} \ln(1+\exp(-\mathbf{x}_i^\top \mathbf{w}_t)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \ln\big(1+\exp(-\gamma w_t - \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_t)\big) \\
&\ge \ln\big(1+\exp(-\gamma w_t - c_1\eta)\big) \\
&\ge \exp(-\gamma w_t - c_1\eta) \\
&= \exp\big(-\ln\big(e\gamma\eta\exp(2c_1\eta)(t+1)\big)\big) \\
&\ge \frac{1}{2e\gamma\eta\exp(2c_1\eta)(t+1)} \\
&\ge \frac{1}{c_0\eta\exp(c_0\eta)t},
\end{aligned}$$

where $c_0 > 0$ is independent of η and t. ■

The next lemma provides a set of examples where the stepsize has to be small for GD to induce a monotonically decreasing risk.

Lemma 14 *Let* $\mathbf{w}_0 = 0$ *and*

$$\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.$$

Assume there exist constants r *and* q *such that*

$$\frac{|\{i \in [n] : \mathbf{x}_i^\top \bar{\mathbf{x}} < -r\}|}{n} \ge q, \quad r > 0,\ 0 < q < 1.$$

Then $L(\mathbf{w}_1) \le L(\mathbf{w}_0)$ *implies that*

$$\eta \le \frac{2\ln(2)}{rq}.$$

Proof (of Lemma 14) Since $\mathbf{w}_0 = 0$, we have $L(0) = \ln(2)$ and

$$\nabla L(\mathbf{w}_0) = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\exp(\mathbf{x}_i^\top \mathbf{w}_0)}\mathbf{x}_i = -\frac{1}{2}\bar{\mathbf{x}}.$$

So we have

$$\mathbf{w}_1 = \mathbf{w}_0 - \eta \nabla L(\mathbf{w}_0) = \frac{\eta}{2}\bar{\mathbf{x}},$$

and

$$\begin{aligned} L(\mathbf{w}_1) &= \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-\mathbf{x}_i^\top \mathbf{w}_1)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-\mathbf{x}_i^\top \bar{\mathbf{x}}\eta/2)\right). \end{aligned}$$

If $\eta > 2\ln(2)/(rq)$, then

$$\begin{aligned} L(\mathbf{w}_1) &\geq \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-\mathbf{x}_i^\top \bar{\mathbf{x}}\eta/2)\right)\mathbb{1}\left[\mathbf{x}_i^\top \bar{\mathbf{x}} < -r\right] \\ &\geq \ln\left(1 + \exp(r\eta/2)\right)q && \text{by assumption} \\ &\geq \frac{rq\eta}{2} \\ &\geq \ln(2) = L(\mathbf{w}_0). \end{aligned}$$

This proves the original claim by contradiction. ■

The proof of Theorem 3 is a combination of Lemmas 13 and 14.

Proof (of Theorem 3) We verify that our dataset satisfies the conditions in Lemma 14 with

$$r = \frac{1}{10}, \quad q = \frac{1}{2}.$$

So by Lemma 14 and the monotonic loss, we have

$$\eta \leq \frac{2\ln(2)}{rq} \leq 40\ln(2).$$

Then we verify that our dataset satisfies the conditions in Lemma 13 with $b = 0.5$. So by Lemma 13 and the upper bound on η, we have

$$L(\mathbf{w}_t) \geq \frac{1}{c_0\eta\exp(c_0\eta)t} \geq \frac{c_1}{t},$$

where $c_1 > 0$ is independent of η and t since η is upper bounded and $c_0 > 0$ is independent of η and t. ■

Appendix E. A Uniform-Convergence Generalization Bound

Proposition 15 (Uniform convergence) *Suppose Assumption 1 holds. Assume that the training samples $(\mathbf{x}_i, y_i)_{i=1}^n$ are i.i.d. copies of $(\mathbf{x}, y)$. Consider a parameter $\hat{\mathbf{w}}$ that classifies $(\mathbf{x}_i, y_i)_{i=1}^n$*

correctly, that is, $y_i \mathbf{x}_i^\top \hat{\mathbf{w}} > 0$ *for every* i*. Then with probability at least* $1 - \delta$ *over the randomness of* $(\mathbf{x}_i, y_i)_{i=1}^n$*, we have*

$$\Pr(y\mathbf{x}^\top \hat{\mathbf{w}} < 0) \leq 4\frac{d\ln(n+1) + \ln(4/\delta)}{n}.$$

In particular, every (GD) *output with* $L(\mathbf{w}_t) \leq 1/n$ *is such a parameter.*

Proof (of Proposition 15) Consider a hypothesis class of hyperplanes

$$\mathcal{H} := \left\{\mathbf{x} \mapsto \operatorname{sgn}(\mathbf{x}^\top \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d\right\}.$$

The Vapnik-Chervonenkis (VC) dimension of $\mathcal{H}$ is $\texttt{VC}(\mathcal{H}) = d$ (Dudley, 1978). Denote the empirical (zero-one) error by

$$R_n(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}[y_i \neq \operatorname{sgn}(\mathbf{x}_i^\top \mathbf{w})] = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[y_i \mathbf{x}_i^\top \mathbf{w} < 0],$$

and the population (zero-one) error by

$$R(\mathbf{w}) := \mathbb{E}\mathbb{1}[y \neq \operatorname{sgn}(\mathbf{x}^\top \mathbf{w})] = \mathbb{E}\mathbb{1}[y\mathbf{x}^\top \mathbf{w} < 0] = \Pr(y\mathbf{x}^\top \mathbf{w} < 0).$$

Then by the classical VC-theory for empirical risk minimizer (ERM) (see Boucheron et al., 2005, Corollary 5.2, for example), for every ERM,

$$\hat{\mathbf{w}} \in \arg\min_{\mathbf{w}} R_n(\mathbf{w}),$$

we have the following with probability at least $1 - \delta$,

$$R(\hat{\mathbf{w}}) \leq R_n(\hat{\mathbf{w}}) + 2\sqrt{R_n(\hat{\mathbf{w}})\frac{2\texttt{VC}(\mathcal{H})\ln(n+1) + \ln(4/\delta)}{n}} + 4\frac{\texttt{VC}(\mathcal{H})\ln(n+1) + \ln(4/\delta)}{n}.$$

In our case, $R_n(\hat{\mathbf{w}}) = 0$, so with probability at least $1 - \delta$ we have

$$R(\hat{\mathbf{w}}) \leq 4\frac{d\ln(n+1) + \ln(4/\delta)}{n}.$$

We complete the proof by noting that $\hat{\mathbf{w}}$ can be any EMR. ■

Appendix F. Proof of Theorem 4

Throughout this part, we assume Assumption 2 holds. Without loss of generality, we also assume $y = 1$. For simplicity, we define

$$L_{\mathbf{x}}(\mathbf{w}) := \ell(\mathbf{x}^\top \mathbf{w}), \quad G_{\mathbf{x}}(\mathbf{w}) := \frac{1}{1 + \exp(\mathbf{x}^\top \mathbf{w})}.$$

The following lemma for SGD is an analogy of Lemma 7 for GD.

Lemma 16 *For every $\eta > 0$, $\mathbf{w}_0$, and $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ such that*

$$\mathbf{u}_2 = \frac{\eta}{2\gamma}\mathbf{w}_*,$$

the following holds for every t:

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1} L_k(\mathbf{w}_k) \le \frac{1}{t}\sum_{k=0}^{t-1} L_k(\mathbf{u}_1) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t}, \quad \textit{almost surely}.$$

Proof (of Lemma 16) The proof repeats the arguments in Lemma 7. By the SGD update rule, we have

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L_t(\mathbf{w}_t), \mathbf{u} - \mathbf{w}_t\rangle + \eta^2\big\|\nabla L_t(\mathbf{w}_t)\big\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L_t(\mathbf{w}_t), \mathbf{u}_1 - \mathbf{w}_t\rangle + \eta\Big(2\langle \nabla L_t(\mathbf{w}_t), \mathbf{u}_2\rangle + \eta\big\|\nabla L_t(\mathbf{w}_t)\big\|_2^2\Big).\end{aligned}$$

Plugging $\mathbf{u}_2 = \mathbf{w}_* \cdot \eta/(2\gamma)$, we can show the second term is non-positive:

$$\begin{aligned}&2\langle \nabla L_t(\mathbf{w}_t), \mathbf{u}_2\rangle + \eta\big\|\nabla L_t(\mathbf{w}_t)\big\|_2^2 \\ &= 2\ell'(\mathbf{x}_t^\top \mathbf{w}_t)\langle \mathbf{x}_t, \mathbf{u}_2\rangle + \eta\big\|\ell'(\mathbf{x}_t^\top \mathbf{w}_t)\mathbf{x}_t\big\|_2^2 \\ &\le \eta\ell'(\mathbf{x}_t^\top \mathbf{w}_t) + \eta\big(\ell'(\mathbf{x}_t^\top \mathbf{w}_t)\big)^2 && \text{since } \mathbf{u}_2 = \mathbf{w}_* \cdot \eta/(2\gamma) \\ &\le 0, \quad \text{almost surely.}\end{aligned}$$

Therefore, we have

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &\le \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\langle \nabla L_t(\mathbf{w}_t), \mathbf{u}_1 - \mathbf{w}_t\rangle \\ &\le \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\big(L_t(\mathbf{u}_1) - L_t(\mathbf{w}_t)\big), \quad \text{almost surely.} \qquad \text{by convexity}\end{aligned}$$

Summing the above from 0 to $t-1$ and rearranging, we obtain

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta} + \sum_{k=0}^{t-1} L_k(\mathbf{w}_k) \le \sum_{k=0}^{t-1} L_k(\mathbf{u}_1) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta}, \quad \text{almost surely,}$$

dividing both sides by t completes our proof. ∎

Similarly, the following lemma for SGD is an analogy of Lemma 8 for GD.

Lemma 17 (Parameter and risk bounds) *Let $\mathbf{w}_0 = 0$. For every $t \ge 0$, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} L_k(\mathbf{w}_k) \le \frac{1 + \ln^2(\gamma^2\eta t) + \eta^2/4}{\gamma^2\eta t}, \quad \textit{almost surely},$$

and for every $k \le t$,

$$\|\mathbf{w}_k\| \le \frac{\sqrt{2} + 2\ln(\gamma^2\eta t) + \eta}{\gamma}, \quad \textit{almost surely}.$$

Proof (of Lemma 17) In Lemma 16, choose

$$\mathbf{w}_0 = 0, \quad \mathbf{u}_1 = \frac{\ln(\gamma^2\eta t)}{\gamma}\mathbf{w}_*,$$

and check that

$$\|\mathbf{u}\| = \frac{\ln(\gamma^2\eta t) + \eta/2}{\gamma},$$

and that for every $\mathbf{x}$,

$$\begin{aligned} L_{\mathbf{x}}(\mathbf{u}_1) &= \ln\left(1 + \exp(-\mathbf{x}^\top \mathbf{u}_1)\right) \\ &\le \exp(-\mathbf{x}^\top \mathbf{u}_1) \\ &\le \frac{1}{\gamma^2\eta t}, \quad \text{almost surely.} \end{aligned}$$

Therefore, we have

$$\frac{1}{t}\sum_{k=0}^{t-1} L_k(\mathbf{u}_1) \le \frac{1}{\gamma^2\eta t}, \quad \text{almost surely.}$$

Using Lemma 16, we get

$$\begin{aligned} \frac{1}{t}\sum_{k=0}^{t-1} L_k(\mathbf{w}_k) &\le \frac{1}{t}\sum_{k=0}^{t-1} L_k(\mathbf{u}_1) + \frac{\|\mathbf{u}\|^2}{2\eta t} \\ &\le \frac{1}{\gamma^2\eta t} + \frac{\ln^2(\gamma^2\eta t) + \eta^2/4}{\gamma^2\eta t}, \quad \text{almost surely,} \end{aligned}$$

which verifies the first claim. For the second claim, we use Lemma 16 for $s \le t$ and get

$$\begin{aligned} \text{for every } s \le t, \quad \|\mathbf{w}_s - \mathbf{u}\|^2 &\le 2\eta \sum_{k=0}^{s-1} \ell_k(\mathbf{u}_1) + \|\mathbf{u}\|^2 \\ &\le 2\eta s \cdot \frac{1}{\gamma^2\eta t} + \|\mathbf{u}\|^2 \\ &\le \frac{2}{\gamma^2} + \|\mathbf{u}\|^2, \quad \text{almost surely,} \end{aligned}$$

which implies that

$$\begin{aligned} \text{for every } s \le t, \quad \|\mathbf{w}_s\| &\le \|\mathbf{u}\| + \|\mathbf{w}_s - \mathbf{u}\| \\ &\le \frac{\sqrt{2}}{\gamma} + 2\|\mathbf{u}\| \\ &\le \frac{\sqrt{2} + 2\ln(\gamma^2\eta t) + \eta}{\gamma}, \quad \text{almost surely.} \end{aligned}$$

We complete the proof. ■

The next lemma provides a population risk bound. The main idea is to use martingale concentration tools to convert the "regret" bound in Lemma 17 to a population risk bound.

Lemma 18 (Population risk bound) *Let $L(\mathbf{w}) := \mathbb{E}\ell(\mathbf{x}^\top \mathbf{w})$. Let $\mathbf{w}_0 = 0$. For every $\eta > 0$, with probability at least $1 - \delta$ we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le \frac{2 + 2\ln^2(\gamma^2\eta t) + \eta^2 B^4/2}{\gamma^2 \eta t} + \frac{3 + 2\ln(\gamma^2\eta t) + \eta}{\gamma} \cdot \frac{18\ln(1/\delta)}{t}.$$

Proof (of Lemma 18) Based on the uniform upper bound on the norm of $(\mathbf{w}_k)_{k\le t}$ in Lemma 17, we can show the following uniform upper bound on risk:

$$\begin{aligned}
\text{for every } k \le t \text{ and every } \mathbf{x}, \quad L_{\mathbf{x}}(\mathbf{w}_k) &= \ln(1 + \exp(-\mathbf{w}_k^\top \mathbf{x})) \\
&\le \ln(1 + \exp(\|\mathbf{w}_k\|)) \\
&\le \ln(2) + \|\mathbf{w}_k\| \\
&\le \ln(2) + \frac{\sqrt{2} + 2\ln(\gamma^2\eta t) + \eta}{\gamma} \quad \text{almost surely} \\
&\le M := \frac{3 + 2\ln(\gamma^2\eta t) + \eta}{\gamma}.
\end{aligned}$$

Note that $L_{\mathbf{x}}(\cdot)$ is non-negative and that $L(\cdot) = \mathbb{E}L_{\mathbf{x}}(\cdot)$. We then have

$$\text{for every } k \le t \text{ and every } \mathbf{x}, \quad |L(\mathbf{w}_k) - L_{\mathbf{x}}(\mathbf{w}_k)| \le M, \quad \text{almost surely.}$$

As a direct consequence, we have

$$\begin{aligned}
\mathbb{E}\big(L(\mathbf{w}_k) - L_{\mathbf{x}}(\mathbf{w}_k)\big)^2 &\le M \cdot \mathbb{E}\big(L(\mathbf{w}_k) + \ell_{\mathbf{x}}(\mathbf{w}_k)\big) \\
&= 2M \cdot L(\mathbf{w}_k), \quad k = 0, 1, \ldots, t.
\end{aligned}$$

Next, we use (one-side) martingale Bernstein inequality (a.k.a. Freedman inequality). With probability at least $1 - \delta$, it holds that

$$\begin{aligned}
\sum_{k=0}^{t-1} L(\mathbf{w}_k) - L_k(\mathbf{w}_k) &\le \sqrt{2\sum_{k=0}^{k-1} \mathbb{E}\big(L(\mathbf{w}_k) - L_k(\mathbf{w}_k)\big)^2 \ln\frac{1}{\delta}} + \frac{2M}{3}\ln\frac{1}{\delta} \\
&\le \sqrt{\sum_{k=0}^{t-1} L(\mathbf{w}_k) \cdot 4M \ln\frac{1}{\delta}} + \frac{2M}{3}\ln\frac{1}{\delta} \\
&\le \frac{1}{2}\sum_{k=0}^{t-1} L(\mathbf{w}_t) + 9M\ln\frac{1}{\delta},
\end{aligned}$$

which implies that

$$\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le 2\sum_{k=0}^{t-1} L_k(\mathbf{w}_k) + 18M\ln\frac{1}{\delta}.$$

Using Lemma 17 and the definition of M, we get

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le 2 \cdot \frac{1 + \ln^2(\gamma^2\eta t) + \eta^2/4}{\gamma^2\eta t} + \frac{18M}{t}\ln\frac{1}{\delta}$$

$$\leq \frac{2 + 2\ln^2(\gamma^2\eta t) + \eta^2/2}{\gamma^2\eta t} + \frac{18\ln(1/\delta)}{t} \cdot \frac{3 + 2\ln(\gamma^2\eta t) + \eta}{\gamma},$$

which concludes our analysis. ■

The next lemma gives a population zero-one error bound. The proof combines a gradient potential bound and a martingale concentration argument.

Lemma 19 (Population error bound) *Consider one-pass SGD initialized from* $\mathbf{w}_0 = 0$*. For every* $\eta > 0$*, with probability at least* $1 - \delta$*, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} \Pr(y\mathbf{x}^\top \mathbf{w}_k) \leq \frac{4\big(\sqrt{2} + 2\ln(\gamma^2\eta t) + \eta\big)}{\gamma^2\eta t} + \frac{36\ln(1/\delta)}{t}.$$

Proof (of Lemma 19) Repeating the perceptron argument as in Lemma 9 , we get

$$\frac{1}{t}\sum_{k=0}^{t-1} G_k(\mathbf{w}_k) \leq \frac{\langle \mathbf{w}_t - \mathbf{w}_0, \mathbf{w}_* \rangle}{\gamma\eta t} \leq \frac{\|\mathbf{w}_t - \mathbf{w}_0\|}{\gamma\eta t}, \quad \text{almost surely.}$$

Using the parameter norm upper bound in Lemma 17, we have

$$\frac{1}{t}\sum_{k=0}^{t-1} G_k(\mathbf{w}_k) \leq \frac{\|\mathbf{w}_t - \mathbf{w}_0\|}{\gamma\eta t} \leq \frac{\sqrt{2} + 2\ln(\gamma^2\eta t) + \eta}{\gamma^2\eta t}, \quad \text{almost surely.}$$

Note that $0 \leq G_{\mathbf{x}}(\cdot) \leq 1$. Applying Freedman's inequality, with probability at least $1 - \delta$ we have

$$\begin{aligned}
\sum_{k=0}^{t-1} G(\mathbf{w}_k) - G_k(\mathbf{w}_k) &\leq \sqrt{2\sum_{k=0}^{k-1} \mathbb{E}\big(G(\mathbf{w}_k) - G_k(\mathbf{w}_k)\big)^2 \ln\frac{1}{\delta}} + \frac{2}{3}\ln\frac{1}{\delta} \\
&\leq \sqrt{2\sum_{k=0}^{t-1} \mathbb{E}\big(G(\mathbf{w}_k) + G_k(\mathbf{w}_k)\big) \ln\frac{1}{\delta}} + \frac{2}{3}\ln\frac{1}{\delta} \\
&= \sqrt{\sum_{k=0}^{t-1} G(\mathbf{w}_k) \cdot 4\ln\frac{1}{\delta}} + \frac{2}{3}\ln\frac{1}{\delta} \\
&\leq \frac{1}{2}\sum_{k=0}^{t-1} G(\mathbf{w}_t) + 9\ln\frac{1}{\delta},
\end{aligned}$$

which implies that

$$\sum_{k=0}^{t-1} G(\mathbf{w}_k) \leq 2\sum_{k=0}^{t-1} G_k(\mathbf{w}_k) + 18\ln\frac{1}{\delta}.$$

So with probability at least $1 - \delta$ we have

$$\frac{1}{t}\sum_{k=0}^{t-1} G(\mathbf{w}_k) \leq \frac{2\big(\sqrt{2} + 2\ln(\gamma^2\eta t) + \eta\big)}{\gamma^2\eta t} + \frac{18\ln(1/\delta)}{t}.$$

We conclude our proof by noting that zero-one loss is bounded by $2G(\cdot)$. ■

The above lemmas imply Theorem 4.

Proof (of Theorem 4) It follows from Lemmas 18 and 19. ■

Appendix G. Proof of Proposition 5

Proof (of Proposition 5) We now verify Proposition 5.

Logisitic loss. For the logistic loss, we have

$$\ell(z) := \ln(1+e^{-z}), \quad g(z) := \frac{1}{1+e^{z}}.$$

By direct computation, we can verify the first part of Assumption 3A and Assumptions 3B and 3D for $C_g = 1$ and $C_e = 2$. We check the bound on $\rho(\lambda)$ in Assumption 3A by

$$\begin{aligned}\rho(\lambda) &= \min_z \lambda\ell(z) + z^2 \\ &\le \lambda\ell(\ln(\lambda)) + \ln^2(\lambda) && \text{by setting } z := \ln(\lambda) \\ &\le 1 + \ln^2(\lambda).\end{aligned}$$

We next verify Assumption 3C for $C_\beta = e/2$. Note that $g(z) \le \ell(z) \le C_\beta \ell(z)$. For the second part of Assumption 3C, assume $z < x$ without loss of generality (if $x < z$, we apply $g(x) \le g(z)$ and exchange x and z in Assumption 3C). So we have $x - 1 \le z < x$ since $|z - x| \le 1$. By the mean-value form of the Taylor's theorem, we have

$$\text{there exists } y \in (z, x) \text{ such that } \ell(z) = \ell(x) + \ell'(x)(z-x) + \frac{\ell''(y)}{2}(z-x)^2. \tag{7}$$

Note that $\ell''(\cdot) \le g(\cdot)$, $g(\cdot)$ is decreasing, and $|z - x| \le 1$, so we have

$$\ell''(y) \le g(y) \le g(z) \le g(x-1) = \frac{e}{e+e^{x}} \le e \cdot g(x).$$

Plugging the above into (7) verifies Assumption 3C for $C_\beta = e/2$.

Flattened exponential loss. For the flattened exponential loss with temperature $a > 0$, we have

$$\ell(z) := \begin{cases} e^{-az} & z > 0, \\ 1 - az & z \le 0, \end{cases} \qquad g(z) := \begin{cases} ae^{-az} & z > 0, \\ a & z \le 0. \end{cases}$$

By direct computation, we can verify the first part of Assumption 3A and Assumptions 3B and 3D for $C_g = a$ and $C_e = 1/a$. We check the bound on $\rho(\lambda)$ in Assumption 3A by

$$\begin{aligned}\rho(\lambda) &= \min_z \lambda\ell(z) + z^2 \\ &\le \lambda\ell\left(\frac{\ln(\lambda)}{a}\right) + \frac{\ln^2(\lambda)}{a^2} && \text{by setting } z := \ln(\lambda)/a \ge 0\end{aligned}$$

$$= 1 + \frac{\ln^2(\lambda)}{a^2}.$$

We next verify Assumption 3C for $C_\beta = \max\{a, ae^a/2\}$. The first part of Assumption 3C is because

$$g(z) \le a\ell(z) \le C_\beta \ell(z).$$

For the second part of Assumption 3C, assume $z < x$ without loss of generality. So we have $x - 1 \le z < x$. We discuss two cases.

- If $z > 0$, then $0 < z < x$, so ℓ'' is continuous in $[z, x]$. Then (7) holds. Using $\ell''(z) = ag(z)$ for $z > 0$, $g(\cdot)$ is non-increasing, and $0 < z < y < x$, we have

  $$\ell''(y) \le ag(y) \le ag(z) = ae^{-a(z-x)}g(x) \le ae^a g(x).$$

  Plugging the above into (7) verifies Assumption 3C.

- If $z \le 0$, then $x \le 1$. So

  $$g(x) \ge g(1) = ae^{-a}.$$

  Note that ℓ is a^2-smooth, so we have

  $$\begin{aligned}\ell(z) &\le \ell(x) + \ell'(x)(z-x) + \frac{a^2}{2}(z-x)^2 \\ &\le \ell(x) + \ell'(x)(z-x) + \frac{ae^a}{2}g(x)(z-x)^2,\end{aligned}$$

  which verifies Assumption 3C.

Flattend polynomial loss. For the flattened polynomial loss of degree $a > 0$, we have

$$\ell(z) := \begin{cases}(1+z)^{-a} & z > 0, \\ 1 - az & z \le 0,\end{cases} \qquad g(z) := \begin{cases}a(1+z)^{-(a+1)} & z > 0, \\ a & z \le 0.\end{cases}$$

By direct computation, we can verify the first part of Assumption 3A and Assumption 3B for $C_g = a$. We check the bound on $\rho(\lambda)$ in Assumption 3A by

$$\begin{aligned}\rho(\lambda) &= \min_z \lambda\ell(z) + z^2 \\ &\le \lambda\ell\big(\lambda^{1/(a+2)}\big) + \lambda^{2/(a+2)} && \text{by setting } z := \lambda^{1/(a+2)} \ge 0 \\ &\le 2\lambda^{2/(a+2)}.\end{aligned}$$

We next verify Assumption 3C for $C_\beta = \max\{a, (a+1)2^a)\}$. The first part of Assumption 3C is because

$$g(z) \le a\ell(z) \le C_\beta \ell(z).$$

For the second part of Assumption 3C, assume $z < x$ without loss of generality. So we have $x - 1 \le z < x$. We discuss two cases.

- If $z > 0$, then $0 < z < x$, so ℓ'' is continuous in $[z, x]$. Then (7) holds. Using $\ell''(z) \leq (a+1)g(z)$ for $z > 0$, $g(\cdot)$ is non-increasing, and $0 < z < y < x$, we have

$$\ell''(y) \leq (a+1)g(y) \leq (a+1)g(z) = (a+1)\left(\frac{1+x}{1+z}\right)^{a+1} g(x) \leq (a+1)2^{a+1}g(x).$$

  Plugging the above into (7) verifies Assumption 3C.

- If $z \leq 0$, then $x \leq 1$. So

$$g(x) \geq g(1) = a2^{-(a+1)}.$$

  Note that ℓ is $a(a+1)$-smooth, so we have

$$\begin{aligned}\ell(z) &\leq \ell(x) + \ell'(x)(z-x) + \frac{a(a+1)}{2}(z-x)^2 \\ &\leq \ell(x) + \ell'(x)(z-x) + (a+1)2^a g(x)(z-x)^2,\end{aligned}$$

  which verifies Assumption 3C.

We have verified all examples. ■

The following lemma is useful in our analysis of general losses.

Lemma 20 *Under Assumption 3A, we have*

$$\ell\big(\sqrt{\rho(\lambda)}\big) \leq \frac{\rho(\lambda)}{\lambda}.$$

Proof (of Lemma 20) Recall that

$$\rho(\lambda) = \min_z \lambda\ell(z) + z^2.$$

Let z_* be the optimal value. Then

$$\rho(\lambda) = \lambda\ell(z_*) + z_*^2.$$

Since $\ell(\cdot)$ is positive by Assumption 3A, we have

$$z_* \leq \sqrt{\rho(\lambda)}, \quad \ell(z_*) \leq \frac{\rho(\lambda)}{\lambda}.$$

Since $\ell(\cdot)$ is non-increasing by Assumption 3A, we have

$$\ell\big(\sqrt{\rho(\lambda)}\big) \leq \ell(z_*) \leq \frac{\rho(\lambda)}{\lambda},$$

which completes the proof. ■

Appendix H. Proof of Theorem 6

We first introduce some notation that will be used extensively in this section. For a loss function ℓ satisfying Assumptions 3A and 3B, define a radius

$$R := 6\frac{\sqrt{\rho(\gamma^2\eta T)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma}. \tag{8}$$

Define a ball centered at the initialization,

$$\mathcal{B} := \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}_0\| \le R\}.$$

Define a point-wise linearization error of the network within the ball as

$$\xi_i(\mathbf{w}, \mathbf{v}) := f_i(\mathbf{w}) - f_i(\mathbf{v}) - \langle \nabla f_i(\mathbf{v}), \mathbf{w} - \mathbf{v}\rangle, \quad \mathbf{w}, \mathbf{v} \in \mathcal{B}.$$

Define the maximum linearization error as

$$\xi = \sup_{\mathbf{w},\mathbf{v}\in\mathcal{B}} \sup_i \left|\xi_i(\mathbf{w}, \mathbf{v})\right|$$

Under Assumption 4, we verify that the gradient is bounded by

$$\|\nabla f(\mathbf{x}; \mathbf{w})\| = \sqrt{\frac{1}{m}\sum_{s=1}^{m} a_s^2\big(\phi'(\mathbf{x}^\top \mathbf{w}^{(s)})\big)^2\|\mathbf{x}\|^2} \le 1. \tag{9}$$

The proof of Theorem 6 breaks down into three parts. In Appendix H.1, we control the network properties at the initialization; then in Appendix H.2, we bound the linearization error. These two parts use standard techniques in the NTK literature. Results in Appendices H.1 and H.2 allow us to show that a good event, under which the network is well-behaved at initialization and the linearization error is small, happens with high probability provided there is a large enough network width. Finally, in Appendix H.3, we analyze the GD dynamics under general losses assuming that the good event holds.

It is worth noting that our analysis in Appendix H.3 is decoupled from Appendices H.1 and H.2 except assuming the good event holds. Therefore, our results for GD under general losses in Appendix H.3 directly apply to a simpler case of training linear models with zero initialization, since the good event automatically holds in this case; we include this consequence explicitly as Theorem 35 in Appendix I.

H.1. Initilization Bounds

The following lemma is from (Ji and Telgarsky, 2019), showing that the NTK features lead to a large margin under Assumption 4.

Lemma 21 (NTK feature margin) *Under Assumption 4, if $m \ge 50\ln(n/\delta)/\gamma^2$, then there exists $\mathbf{w}_* \in \mathbb{R}^{md}$ with $\|\mathbf{w}_*\| \le 1$ such that with probability at least $1-\delta$, we have*

$$\textit{for } i = 1, \dots, n, \quad \langle y_i \nabla f_i(\mathbf{w}_0), \mathbf{w}_*\rangle \ge \frac{9\gamma}{10} > 0.$$

Proof (of Lemma 21) This is Lemma 2.3 in (Ji and Telgarsky, 2019). Choose

$$\mathbf{w}_* := \left(\tfrac{a_1}{\sqrt{m}}\chi(\mathbf{w}_0^{(1)}), \quad \cdots, \quad \tfrac{a_m}{\sqrt{m}}\chi(\mathbf{w}_0^{(m)})\right).$$

By the condition on χ in Assumption 4, we have $\|\mathbf{w}_*\| \le 1$. In addition, note that

$$\langle y_i \nabla f_i(\mathbf{w}_0), \mathbf{w}_* \rangle = \frac{1}{m}\sum_{s=1}^{m} y_i \phi'(\mathbf{x}_i^\top \mathbf{w}_0^{(s)}) \mathbf{x}_i^\top \chi(\mathbf{w}_0^{(s)}).$$

By Assumption 4, we have

$$\mathbb{E}\langle y_i \nabla f_i(\mathbf{w}_0), \mathbf{w}_* \rangle = \mathbb{E} y_i \phi'(\mathbf{x}_i^\top \mathbf{w}_0^{(s)}) \mathbf{x}_i^\top \chi(\mathbf{w}_0^{(s)}) \ge \gamma.$$

By Hoeffding inequality, we have with probability at least $1-\delta$,

$$\langle y_i \nabla f_i(\mathbf{w}_0), \mathbf{w}_* \rangle \ge \mathbb{E}\langle y_i \nabla f_i(\mathbf{w}_0), \mathbf{w}_* \rangle - \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}} \ge \gamma - \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}.$$

By a union bound over $i = 1, \ldots, n$, we have with probability at least $1-\delta$,

$$\min_i \langle y_i \nabla f_i(\mathbf{w}_0), \mathbf{w}_* \rangle \ge \gamma - \sqrt{\frac{1}{2m}\ln\frac{n}{\delta}} \ge \frac{9\gamma}{10},$$

where the last inequality is because $m \ge 50\ln(n/\delta)/\gamma^2$ ■

The next lemma is also from (Ji and Telgarsky, 2019), showing that the model output is bounded at initialization.

Lemma 22 (Initial function value) *At initialization, with probability at least $1-\delta$, we have*

$$\max_i |f_i(\mathbf{w}_0)| \le C_a + \sqrt{2\ln(2n/\delta)}.$$

Proof (of Lemma 22) This is a variant of Lemma 2.4 in (Ji and Telgarsky, 2019). Note that $\mathbf{x}^\top \mathbf{w}_0^{(s)} \sim \mathcal{N}(0, \|\mathbf{x}\|^2)$, $\|\mathbf{x}\| \le 1$, and $\phi(\cdot)$ is 1-Lipschitz, so $\phi(\mathbf{x}^\top \mathbf{w}_0^{(s)})$ is 1-subGaussian. Therefore,

$$\left(\phi(\mathbf{x}^\top \mathbf{w}_0^{(1)}), \quad \cdots, \quad \phi(\mathbf{x}^\top \mathbf{w}_0^{(m)})\right)^\top$$

is an m-dimensional, 1-subGaussian random vector. Note that

$$\sqrt{\sum_{s=1}^{m}\left(\frac{a_s}{\sqrt{m}}\right)^2} \le 1, \qquad \text{by (5)}$$

so as an inner product of a unit vector and a 1-subGaussian random vector,

$$f(\mathbf{x}; \mathbf{w}_0) = \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \phi(\mathbf{x}^\top \mathbf{w}_0^{(s)})$$

is 1-subGaussian. So with probability at least $1-\delta$, we have

$$|f(\mathbf{x};\mathbf{w}_0) - \mathbb{E}f(\mathbf{x};\mathbf{w}_0)| \le \sqrt{2\ln(2/\delta)} \le \sqrt{2\ln(2/\delta)}.$$

On the other hand, recall that $\mathbf{x}^\top \mathbf{w}_0^{(s)} \sim \mathcal{N}(0, \|\mathbf{x}\|^2)$ and that

$$Z \sim \mathcal{N}(0,\sigma^2), \quad \mathbb{E}\max\{Z,0\} = \frac{\sigma}{\sqrt{2\pi}} \le \sigma.$$

So we have

$$\begin{aligned}
\mathbb{E}f(\mathbf{x};\mathbf{w}_0) &= \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \mathbb{E}\phi(\mathbf{x}^\top \mathbf{w}_0^{(s)}) \\
&\le \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \|\mathbf{x}\| \\
&\le C_a. && \text{since } \sum_s a_s \le C_a\sqrt{m} \text{ by (5)}
\end{aligned}$$

Applying the triangle inequality and a union bound over $\mathbf{x}_1,\dots,\mathbf{x}_n$, we have with probability at least $1-\delta$,

$$\begin{aligned}
\max_i |f_i(\mathbf{w}_0)| &\le \max_i \left(\mathbb{E}f_i(\mathbf{w}_0) + |f_i(\mathbf{w}_0) - \mathbb{E}f_i(\mathbf{w}_0)|\right) \\
&\le C_a + \sqrt{2\ln(2n/\delta)}.
\end{aligned}$$

This completes the proof. ■

H.2. Linearization Error Bounds

The next lemma is a variant of Lemma 4.1 in (Telgarsky, 2021) (which comes from (Ji and Telgarsky, 2019)), connecting the linearization error with the network width m. The original version in (Telgarsky, 2021) only provides a bound on the maximum linearization error. Here, we need to use a slightly stronger version that tracks the point-wise linearization error.

Lemma 23 (Linearization error) *With probability at least $1-\delta$, we have the following for every $\mathbf{w},\mathbf{v} \in \mathcal{B}$:*

$$\sup_i |\xi_i(\mathbf{w},\mathbf{v})| \le \frac{3R^{1/3} + \ln^{1/4}(n/\delta)}{m^{1/6}} \cdot \|\mathbf{w}-\mathbf{v}\|,$$

and

$$\sup_i \|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{v})\| \le \frac{3R^{1/3} + \ln^{1/4}(n/\delta)}{m^{1/6}}.$$

Proof (of Lemma 23) This is a variant of Lemma 4.1 in (Telgarsky, 2021). The proof is included for completeness.

Fix $\mathbf{x}$. Using homogeneity of ϕ, we have

$$f(\mathbf{x};\mathbf{w}) = \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \phi(\mathbf{x}^\top \mathbf{w}^{(s)}) = \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \mathbb{1}\left[\mathbf{x}^\top \mathbf{w}^{(s)} > 0\right] \mathbf{x}^\top \mathbf{w}^{(s)},$$

and

$$\begin{aligned}
&f(\mathbf{x};\mathbf{v}) + \langle \nabla f(\mathbf{x};\mathbf{v}), \mathbf{w}-\mathbf{v}\rangle \\
&= \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \phi(\mathbf{x}^\top \mathbf{v}^{(s)}) + \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \mathbf{x}^\top \big(\mathbf{w}^{(s)} - \mathbf{v}^{(s)}\big) \\
&= \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \mathbf{x}^\top \mathbf{v}^{(s)} + \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \mathbf{x}^\top \big(\mathbf{w}^{(s)} - \mathbf{v}^{(s)}\big) \\
&= \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \mathbf{x}^\top \mathbf{w}^{(s)},
\end{aligned}$$

then we have

$$\begin{aligned}
\xi_{\mathbf{x}}(\mathbf{w},\mathbf{v}) &= f(\mathbf{x};\mathbf{w}) - f(\mathbf{x};\mathbf{v}) - \langle \nabla f(\mathbf{x};\mathbf{v}), \mathbf{w}-\mathbf{v}\rangle \\
&= \frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s \left(\mathbb{1}\left[\mathbf{x}^\top \mathbf{w}^{(s)} > 0\right] - \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \right) \mathbf{x}^\top \mathbf{w}^{(s)}.
\end{aligned}$$

Control linearization error. For r to be determined, define

$$\begin{aligned}
\mathcal{S}_0 &:= \big\{ s \in [m] : |\mathbf{x}^\top \mathbf{w}_0^{(s)}| \le r\|\mathbf{x}\| \big\} \\
\mathcal{S}_1 &:= \big\{ s \in [m] : \|\mathbf{w}^{(s)} - \mathbf{w}_0^{(s)}\| \ge r \big\} \\
\mathcal{S}_2 &:= \big\{ s \in [m] : \|\mathbf{v}^{(s)} - \mathbf{w}_0^{(s)}\| \ge r \big\}.
\end{aligned}$$

One can verify that for every $s \notin \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2$,

$$\operatorname{sgn}\big(\mathbf{x}^\top \mathbf{w}^{(s)}\big) = \operatorname{sgn}\big(\mathbf{x}^\top \mathbf{w}_0^{(s)}\big), \quad \operatorname{sgn}\big(\mathbf{x}^\top \mathbf{v}^{(s)}\big) = \operatorname{sgn}\big(\mathbf{x}^\top \mathbf{w}_0^{(s)}\big).$$

So for every $s \notin \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2$,

$$\mathbb{1}\left[\mathbf{x}^\top \mathbf{w}^{(s)} > 0\right] = \mathbb{1}\left[\mathbf{x}^\top \mathbf{w}_0^{(s)} > 0\right] = \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right].$$

Then we have

$$\begin{aligned}
|\xi_{\mathbf{x}}(\mathbf{w},\mathbf{v})| &\le \frac{1}{\sqrt{m}} \sum_{s\in[m]} \left| \mathbb{1}\left[\mathbf{x}^\top \mathbf{w}^{(s)} > 0\right] - \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \right| \cdot \left|\mathbf{x}^\top \mathbf{w}^{(s)}\right| \\
&\le \frac{1}{\sqrt{m}} \sum_{s\in[m]} \left| \mathbb{1}\left[\mathbf{x}^\top \mathbf{w}^{(s)} > 0\right] - \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \right| \cdot \left|\mathbf{x}^\top \mathbf{w}^{(s)} - \mathbf{x}^\top \mathbf{v}^{(s)}\right| \\
&\qquad \text{since } \left|\mathbf{x}^\top \mathbf{w}^{(s)} - \mathbf{x}^\top \mathbf{v}^{(s)}\right| \ge \left|\mathbf{x}^\top \mathbf{w}^{(s)}\right| \text{ if the indicator factor is non-zero}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\sqrt{m}} \sum_{s\in\mathcal{S}_0\cup\mathcal{S}_1\cup\mathcal{S}_2} \left| \mathbb{1}\left[\mathbf{x}^\top \mathbf{w}^{(s)} > 0\right] - \mathbb{1}\left[\mathbf{x}^\top \mathbf{v}^{(s)} > 0\right] \right| \cdot \left|\mathbf{x}^\top \mathbf{w}^{(s)} - \mathbf{x}^\top \mathbf{v}^{(s)}\right| \\
&\text{the indicator factor is zero if } s \notin \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2 \\
&\leq \frac{1}{\sqrt{m}} \sum_{s\in\mathcal{S}_0\cup\mathcal{S}_1\cup\mathcal{S}_2} \left|\mathbf{x}^\top \mathbf{w}^{(s)} - \mathbf{x}^\top \mathbf{v}^{(s)}\right| \\
&\leq \frac{1}{\sqrt{m}} \sqrt{|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2|} \cdot \sqrt{\sum_{s\in\mathcal{S}_1\cup\mathcal{S}_2\cup\mathcal{S}_3} \left|\mathbf{x}^\top \mathbf{w}^{(s)} - \mathbf{x}^\top \mathbf{v}^{(s)}\right|^2} \\
&\leq \frac{1}{\sqrt{m}} \sqrt{|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2|} \cdot \|\mathbf{w} - \mathbf{v}\|.
\end{aligned}
$$

Count undesirable indexes. Then by Hoeffding inequality, with probability at least $1-\delta$, we have

$$
\begin{aligned}
|\mathcal{S}_0| &= \sum_{s=1}^{m} \mathbb{1}\left[|\mathbf{x}^\top \mathbf{w}_0^{(s)}| \leq r\|\mathbf{x}\|\right] \\
&\leq m \Pr\left\{|\mathbf{x}^\top \mathbf{w}_0^{(s)}| \leq r\|\mathbf{x}\|\right\} + \sqrt{\frac{m}{2}\ln(1/\delta)}.
\end{aligned}
$$

Since $\mathbf{x}^\top \mathbf{w}_0^{(s)}/\|\mathbf{x}\| \sim \mathcal{N}(0,1)$, we have

$$
\begin{aligned}
\Pr\left\{|\mathbf{x}^\top \mathbf{w}_0^{(s)}| \leq r\|\mathbf{x}\|\right\} &= \Pr\left\{|\mathbf{x}^\top \mathbf{w}_0^{(s)}|/\|\mathbf{x}\| \leq r\right\} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-r}^{r} \exp\big(-z^2/2\big)\mathrm{d}z \\
&\leq \frac{2r}{\sqrt{2\pi}} \leq r.
\end{aligned}
$$

So with probability at least $1-\delta$,

$$
|\mathcal{S}_0| \leq mr + \sqrt{\frac{m}{2}\ln(1/\delta)}.
$$

On the other hand,

$$
R^2 \geq \sum_{s=1}^{m} \|\mathbf{w}^{(s)} - \mathbf{w}_0^{(s)}\|^2 \geq r^2 |\mathcal{S}_1|,
$$

which implies

$$
|\mathcal{S}_1| \leq \frac{R^2}{r^2}.
$$

Similarly, we have

$$
|\mathcal{S}_2| \leq \frac{R^2}{r^2}.
$$

So

$$|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2| \le mr + \sqrt{\frac{m}{2}\ln(1/\delta)} + \frac{2R^2}{r^2}.$$

Picking $r = R^{2/3}m^{-1/3}$, then we have

$$\begin{aligned}|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2| &\le mr + \sqrt{\frac{m}{2}\ln(1/\delta)} + \frac{2R^2}{r^2} \\ &\le m^{2/3}\big(3R^{2/3} + \sqrt{\ln(1/\delta)}\big).\end{aligned}$$

Union bound. Using all the above, we have: with probability at least $1-\delta$,

$$\begin{aligned}|\xi_{\mathbf{x}}(\mathbf{w},\mathbf{v})| &\le \frac{1}{\sqrt{m}}\sqrt{|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2|}\cdot\|\mathbf{w}-\mathbf{v}\| \\ &\le \frac{1}{\sqrt{m}}\sqrt{m^{2/3}\big(3R^{2/3} + \sqrt{\ln(1/\delta)}\big)}\cdot\|\mathbf{w}-\mathbf{v}\| \\ &\le \frac{3R^{1/3} + \ln^{1/4}(1/\delta)}{m^{1/6}}\cdot\|\mathbf{w}-\mathbf{v}\|, \quad \text{for every } \mathbf{w},\mathbf{v}\in\mathcal{B}.\end{aligned}$$

Applying a union bound for the above to hold over $\mathbf{x}_1,\dots,\mathbf{x}_n$, we get with probability at least $1-\delta$,

$$\sup_i |\xi_i(\mathbf{w},\mathbf{v})| \le \frac{3R^{1/3} + \ln^{1/4}(n/\delta)}{m^{1/6}}\cdot\|\mathbf{w}-\mathbf{v}\|, \quad \text{for every } \mathbf{w},\mathbf{v}\in\mathcal{B}.$$

Gradient bound. This part is similar to the previous argument. Fix $\mathbf{x}$, we have

$$\begin{aligned}\|\nabla f(\mathbf{x};\mathbf{w}) - \nabla f(\mathbf{x};\mathbf{v})\|^2 &= \frac{1}{m}\sum_{s=1}^m \left\| \mathbb{1}\left[\mathbf{x}^\top\mathbf{w}^{(s)} > 0\right]\mathbf{x} - \mathbb{1}\left[\mathbf{x}^\top\mathbf{v}^{(s)} > 0\right]\mathbf{x}\right\|^2 \\ &\le \frac{1}{m}\sum_{s=1}^m \left| \mathbb{1}\left[\mathbf{x}^\top\mathbf{w}^{(s)} > 0\right] - \mathbb{1}\left[\mathbf{x}^\top\mathbf{v}^{(s)} > 0\right]\right|^2 \\ &\le \frac{1}{m}\cdot|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2|.\end{aligned}$$

Applying our bound on $|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2|$ and a union bound over $\mathbf{x}$, we get

$$\max_i \|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w})\| \le \frac{1}{\sqrt{m}}\cdot\sqrt{|\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2|} \le \frac{3R^{1/3} + \ln^{1/4}(n/\delta)}{m^{1/6}}$$

holds with probability at least $1-\delta$. ■

Using Lemma 23, we can show that the linearization error is well-bounded when the network width is sufficiently large.

Lemma 24 (Linearization error conditions) *Recall that*

$$R := 6\frac{\sqrt{\rho(\gamma^2\eta T)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma}.$$

Assume that

$$m \geq \left(\frac{10\big(3R^{1/3} + \ln^{1/4}(n/\delta)\big)}{\gamma}\right)^6,$$

then with probability at least $1-\delta$*, we have the following for every* $\mathbf{w}, \mathbf{v} \in \mathcal{B}$*:*

$$\sup_i |\xi_i(\mathbf{w}, \mathbf{v})| \leq \frac{\gamma}{10}\|\mathbf{w} - \mathbf{v}\|, \quad \sup_i \|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{v})\| \leq \frac{\gamma}{10}.$$

Proof (of Lemma 24) These follow from our choice of m and Lemma 23. ■

H.3. Optimization Analysis under General Losses

The good event and the gradient potential. Let $\mathcal{E}$ be the intersection of the events in Lemmas 21, 22 and 24. Then we know $\mathcal{E}$ holds with probability at least $1-3\delta$ if the width m is large enough as specified in Theorem 6. In what follows, we always operate under the good event $\mathcal{E}$. We denote the gradient potential by

$$G(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n g\big(y_i f(\mathbf{x}_i; \mathbf{w})\big).$$

Our next lemma establishes both a uniform margin bound and a relative margin bound for parameters in a ball. The relative margin bound is crucial to get the right condition on width m and the ball radius R.

Lemma 25 (Margin in a parameter ball) *Under event* $\mathcal{E}$*, we have*

$$\inf_{\mathbf{w}\in\mathcal{B}} \min_i \langle y_i \nabla f_i(\mathbf{w}), \mathbf{w}_*\rangle \geq \frac{4\gamma}{5}.$$

As a consequence, for

$$\mathbf{u} := \mathbf{v} + \theta \mathbf{w}_*, \quad \mathbf{v} \in \mathcal{B},$$

we have

$$\textit{for every } i \textit{ and every } \mathbf{w} \in \mathcal{B}, \quad \langle y_i \nabla f_i(\mathbf{w}), \mathbf{u}\rangle \geq \frac{\gamma}{10}\big(8\theta - \|\mathbf{w} - \mathbf{v}\|\big) + y_i f_i(\mathbf{v}).$$

Proof (of Lemma 25) The first claim is because

$$\begin{aligned}
\langle y_i \nabla f_i(\mathbf{w}), \mathbf{w}_*\rangle &= \langle y_i \nabla f_i(\mathbf{w}_0), \mathbf{w}_*\rangle + \langle y_i \nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}_0), \mathbf{w}_*\rangle \\
&\geq \frac{9\gamma}{10} - \|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}_0)\| && \text{by Lemma 21} \\
&\geq \frac{4\gamma}{5}. && \text{by Lemma 24}
\end{aligned}$$

To prove the second claim, for $\mathbf{w} \in \mathcal{B}$, we have

$$\begin{aligned}
y_i f_i(\mathbf{v}) &= y_i\big(f_i(\mathbf{w}) + \langle \nabla f_i(\mathbf{w}), \mathbf{v} - \mathbf{w}\rangle + \xi_i(\mathbf{v}, \mathbf{w})\big) \\
&= y_i\big(\langle \nabla f_i(\mathbf{w}), \mathbf{v}\rangle + \xi_i(\mathbf{v}, \mathbf{w})\big) && \text{by the homogeneity of } f
\end{aligned}$$

$$\leq y_i\langle \nabla f_i(\mathbf{w}), \mathbf{v}\rangle + \frac{\gamma}{10}\|\mathbf{w}-\mathbf{v}\|. \qquad \text{by Lemma 24}$$

Therefore, we have

$$\begin{aligned}\langle y_i\nabla f_i(\mathbf{w}), \mathbf{u}\rangle &= \langle y_i\nabla f_i(\mathbf{w}), \theta\mathbf{w}_*\rangle + \langle y_i\nabla f_i(\mathbf{w}), \mathbf{v}\rangle \\ &\geq \frac{\gamma}{10}(8\theta - \|\mathbf{w}-\mathbf{v}\|) + y_i f_i(\mathbf{v}).\end{aligned}$$

These complete the proof. ■

The next lemma establishes useful relationships between the gradient potential and the loss.

Lemma 26 (Gradient potential properties) *Consider the loss and the gradient potential.*

1. *For ℓ satisfying Assumption 3C, we have*

$$G(\mathbf{w}) \leq C_\beta \cdot L(\mathbf{w}).$$

2. *Under event $\mathcal{E}$, we have*

$$\frac{4\gamma}{5}\cdot G(\mathbf{w}) \leq \|\nabla L(\mathbf{w})\| \leq G(\mathbf{w}).$$

Proof (of Lemma 26) The first claim is by

$$\begin{aligned}G(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^n g\big(y_i f_i(\mathbf{w})\big) \\ &\leq \frac{C_\beta}{n}\sum_{i=1}^n \ell\big(y_i f_i(\mathbf{w})\big) \qquad \text{by Assumption 3C} \\ &= C_\beta L(\mathbf{w}).\end{aligned}$$

The upper bound in the second claim is by

$$\begin{aligned}\|\nabla L(\mathbf{w})\| &= \left\|\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w})\big) y_i \nabla f_i(\mathbf{w})\right\| \\ &\leq \frac{1}{n}\sum_{i=1}^n \big|\ell'\big(y_i f_i(\mathbf{w})\big)\big| \cdot \|\nabla f_i(\mathbf{w})\| \\ &\leq G(\mathbf{w}). \qquad \text{by (9)}\end{aligned}$$

The lower bound in the second claim is by

$$\begin{aligned}\|\nabla L(\mathbf{w})\| &\geq \langle \nabla L(\mathbf{w}), -\mathbf{w}_*\rangle \\ &= -\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w})\big)\langle y_i\nabla f_i(\mathbf{w}), \mathbf{w}_*\rangle \\ &\geq -\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w})\big)\cdot\frac{4\gamma}{5} \qquad \text{by Lemma 25}\end{aligned}$$

$$= \frac{4\gamma}{5} \cdot G(\mathbf{w}).$$

We have completed the proof. ■

In the next lemma, we redo Lemma 7 for nonlinear models with controlled linearization errors and general loss functions.

Lemma 27 (Split optimization) *Consider a loss ℓ satisfying Assumptions 3A and 3B. Assume event $\mathcal{E}$ holds. Let $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$ such that*

$$\mathbf{u}_1 = \theta \mathbf{w}_* + \mathbf{w}_0, \quad \mathbf{u}_2 = \frac{2\eta C_g}{\gamma} \mathbf{w}_*.$$

For every $t > 0$ and $R_t \in (0, R]$ such that

$$\|\mathbf{w}_k - \mathbf{w}_0\| \le R_t, \quad k = 0, 1, \ldots, t-1,$$

we have

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \frac{1}{t} \sum_{k=0}^{t-1} L(\mathbf{w}_k) \le \ell\Big(\gamma(8\theta - R_t)/10 - \big(C_a + \sqrt{2\ln(2n/\delta)}\big)\Big) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t}.$$

Proof (of Lemma 27) For $k < t$, we have

$$\begin{aligned}
\|\mathbf{w}_{k+1} - \mathbf{u}\|^2 &= \|\mathbf{w}_k - \mathbf{u}\|^2 + 2\eta \langle \nabla L(\mathbf{w}_k), \mathbf{u} - \mathbf{w}_k \rangle + \eta^2 \|\nabla L(\mathbf{w}_k)\|^2 \\
&= \|\mathbf{w}_k - \mathbf{u}\|^2 + 2\eta \langle \nabla L(\mathbf{w}_k), \mathbf{u}_1 - \mathbf{w}_k \rangle + \eta\big(2\langle \nabla L(\mathbf{w}_k), \mathbf{u}_2 \rangle + \eta \|\nabla L(\mathbf{w}_k)\|^2\big).
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\langle \nabla L(\mathbf{w}_k), \mathbf{u}_1 - \mathbf{w}_k \rangle \\
&= \frac{1}{n} \sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_t)\big) y_i \langle \nabla f_i(\mathbf{w}_k), \mathbf{u}_1 - \mathbf{w}_k \rangle \\
&= \frac{1}{n} \sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_k)\big) \big(\langle y_i \nabla f_i(\mathbf{w}_k), \mathbf{u}_1 \rangle - y_i f_i(\mathbf{w}_k)\big) && \text{by the homogeneity of } f \\
&\le \frac{1}{n} \sum_{i=1}^n \Big(\ell\big(\langle y_i \nabla f_i(\mathbf{w}_k), \mathbf{u}_1 \rangle\big) - \ell\big(y_i f_i(\mathbf{w}_k)\big)\Big) && \text{by Assumption 3A} \\
&\le \frac{1}{n} \sum_{i=1}^n \ell\big(\gamma(8\theta - \|\mathbf{w}_k - \mathbf{w}_0\|)/10 + y_i f_i(\mathbf{w}_0)\big) - L(\mathbf{w}_t) && \text{by Lemma 25} \\
&\le \frac{1}{n} \sum_{i=1}^n \ell\big(\gamma(8\theta - R_t)/10 + y_i f_i(\mathbf{w}_0)\big) - L(\mathbf{w}_t). && \text{since } \|\mathbf{w}_k - \mathbf{w}_0\| \le R_t \qquad (10)
\end{aligned}$$

From Lemma 22 we have

$$|y_i f_i(\mathbf{w}_0)| \le C_a + \sqrt{2\ln(2n/\delta)},$$

then by the monotinicity of ℓ from Assumption 3A, we have

$$\ell\big(\gamma(8\theta - R_t)/10 + y_i f_i(\mathbf{w}_0)\big) \le \ell\Big(\gamma(8\theta - R_t)/10 - \big(C_a + \sqrt{2\ln(2n/\delta)}\big)\Big).$$

Therefore, the second term can be bounded by

$$\langle \nabla L(\mathbf{w}_k), \mathbf{u}_1 - \mathbf{w}_k\rangle \le \ell\Big(\gamma(8\theta - R_t)/10 - \big(C_a + \sqrt{2\ln(2n/\delta)}\big)\Big) - L(\mathbf{w}_k).$$

For the third term, we have

$$\begin{aligned}
&2\langle \nabla L(\mathbf{w}_k), \mathbf{u}_2\rangle + \eta\|\nabla L(\mathbf{w}_k)\|^2 \\
&= \frac{2}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_k)\big) y_i \langle \nabla f_i(\mathbf{w}_k), \mathbf{u}_2\rangle + \eta\left\|\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_k)\big) y_i \nabla f_i(\mathbf{w}_k)\right\|^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_k)\big) y_i \langle \nabla f_i(\mathbf{w}_k), \mathbf{u}_2\rangle + \eta\left(\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_k)\big)\right)^2 \qquad \text{by (9)} \\
&\le \frac{2\|\mathbf{u}_2\|}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_k)\big) y_i \langle \nabla f_i(\mathbf{w}_k), \mathbf{w}_*\rangle + \eta C_g \cdot G(\mathbf{w}_k) \qquad \text{by Assumption 3B} \\
&\le -\frac{\gamma\|\mathbf{u}_2\|}{2} G(\mathbf{w}_k) + \eta C_g \cdot G(\mathbf{w}_k) \qquad \text{by Lemma 25} \\
&\le 0,
\end{aligned}$$

where the last inequality is by the choice of $\mathbf{u}_2$.

Putting these together, we have for $k < t$

$$\|\mathbf{w}_{k+1} - \mathbf{u}\|^2 \le \|\mathbf{w}_k - \mathbf{u}\|^2 + 2\eta\left(\ell\Big(\gamma(8\theta - R_t)/10 - \big(C_a + \sqrt{2\ln(2n/\delta)}\big)\Big) - L(\mathbf{w}_k)\right).$$

Telescoping the sum from 0 to $t-1$ and rearranging, we get

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le \ell\Big(\gamma(8\theta - R_t)/10 - \big(C_a + \sqrt{2\ln(2n/\delta)}\big)\Big) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t},$$

which completes the proof. ∎

The following lemma combines an induction argument with the arguments in Lemma 8.

Lemma 28 (Risk and parameter bounds in the EoS phase) *Consider a loss ℓ that satisfies Assumptions 3A and 3B. Under event $\mathcal{E}$, for every $t \le T$,*

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le 9\frac{\rho(\gamma^2\eta t) + \big(C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g\big)^2}{\gamma^2\eta t},$$

and

$$\|\mathbf{w}_t - \mathbf{w}_0\| \le 6\frac{\sqrt{\rho(\gamma^2\eta t)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma} =: R_t \le R.$$

Proof (of Lemma 28) We prove the claims by induction. For $t = 0$, the two claims hold.

Now suppose the claims hold for $0, \dots, t-1$, and consider t. By the inductive hypothesis, we have

$$\text{for } k = 0, 1 \dots, t-1, \quad \|\mathbf{w}_k - \mathbf{w}_0\| \le R_k \le R_t := 6\frac{\sqrt{\rho(\gamma^2\eta t)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma}.$$

So we can invoke Lemma 27. Recall that

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2, \quad \mathbf{u}_1 = \theta \mathbf{w}_* + \mathbf{w}_0, \quad \mathbf{u}_2 = \frac{2\eta C_g}{\gamma}\mathbf{w}_*.$$

Let us choose

$$\theta = \frac{1}{8}R_t + \frac{5}{4}\frac{\sqrt{\rho(\gamma^2\eta t)} + C_a + \sqrt{2\ln(2n/\delta)}}{\gamma}.$$

Then we can check that

$$\begin{aligned}\ell\Big(\gamma(8\theta - R_t)/10 - \big(C_a + \sqrt{2\ln(2n/\delta)}\big)\Big) &= \ell\big(\sqrt{\rho(\gamma^2\eta t)}\big) \\ &\le \frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t}, \qquad \text{by Lemma 20}\end{aligned}$$

and that

$$\|\mathbf{u} - \mathbf{w}_0\| = \theta + \frac{2\eta C_g}{\gamma} = \frac{1}{8}R_t + \frac{5}{4}\frac{\sqrt{\rho(\gamma^2\eta t)} + C_a + \sqrt{2\ln(2n/\delta)}}{\gamma} + \frac{2\eta C_g}{\gamma}.$$

Now using Lemma 27, we get

$$\begin{aligned}\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) &\le \ell\big(\gamma(8\theta - R_t)/10 - (C_a + \sqrt{2\ln(2n/\delta)})\big) + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t} \\ &\le \frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t} + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t},\end{aligned}$$

which implies that

$$\begin{aligned}\|\mathbf{w}_t - \mathbf{w}_0\| &\le \|\mathbf{w}_t - \mathbf{u}\| + \|\mathbf{u} - \mathbf{w}_0\| \\ &\le \sqrt{\frac{2\rho(\gamma^2\eta t)}{\gamma^2} + \|\mathbf{w}_0 - \mathbf{u}\|^2} + \|\mathbf{u} - \mathbf{w}_0\| \\ &\le \frac{\sqrt{2\rho(\gamma^2\eta t)}}{\gamma} + 2\|\mathbf{w}_0 - \mathbf{u}\| \\ &\le \frac{\sqrt{2\rho(\gamma^2\eta t)}}{\gamma} + \frac{1}{4}R_t + \frac{5}{2}\frac{\sqrt{\rho(\gamma^2\eta t)} + C_a + \sqrt{2\ln(2n/\delta)}}{\gamma} + \frac{4\eta C_g}{\gamma} \\ &\le R_t := 6\frac{\sqrt{\rho(\gamma^2\eta t)} + \eta C_g + C_a + \sqrt{2\ln(2n/\delta)}}{\gamma},\end{aligned}$$

and

$$\begin{aligned}\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) &\le \frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t} + \frac{\|\mathbf{w}_0 - \mathbf{u}\|^2}{2\eta t} \\ &\le \frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t} + \frac{1}{2\eta t}\left(\frac{1}{8}R_t + \frac{5}{4}\frac{\sqrt{\rho(\gamma^2\eta t)} + C_a + \sqrt{2\ln(2n/\delta)}}{\gamma} + \frac{2\eta C_g}{\gamma}\right)^2 \\ &\le 9\frac{\rho(\gamma^2\eta t) + \left(\eta C_g + C_a + \sqrt{2\ln(2n/\delta)}\right)^2}{\gamma^2\eta t}.\end{aligned}$$

These verify the claim for t. We have completed the proof. ■

The following lemma is analogous to Lemma 9.

Lemma 29 (Gradient potential bound in the EoS phase) *Consider a loss ℓ satisfying Assumptions 3A and 3B. Under event $\mathcal{E}$, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} G(\mathbf{w}_k) \le 12\frac{\sqrt{\rho(\gamma^2\eta t)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma^2\eta t}, \quad t \le T.$$

Proof (of Lemma 29) By Lemma 28, we know $\mathbf{w}_t \in \mathcal{B}$ for $t \le T$. Then we use the perceptron argument (Novikoff, 1962),

$$\begin{aligned}\langle \mathbf{w}_{t+1}, \mathbf{w}_* \rangle &= \langle \mathbf{w}_t, \mathbf{w}_* \rangle - \eta\langle \nabla L(\mathbf{w}_t), \mathbf{w}_* \rangle \\ &= \langle \mathbf{w}_t, \mathbf{w}_* \rangle - \frac{\eta}{n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_t)\big)\langle y_i \nabla f_i(\mathbf{w}_t), \mathbf{w}_* \rangle \\ &\ge \langle \mathbf{w}_t, \mathbf{w}_* \rangle - \frac{4\gamma\eta}{5n}\sum_{i=1}^n \ell'\big(y_i f_i(\mathbf{w}_t)\big) \qquad \text{by Lemma 25} \\ &= \langle \mathbf{w}_t, \mathbf{w}_* \rangle + \frac{4\gamma\eta}{5} G(\mathbf{w}_t).\end{aligned}$$

Telescoping the sum, we have

$$\frac{1}{t}\sum_{k=0}^{t-1} G(\mathbf{w}_k) \le \frac{5\big(\langle \mathbf{w}_t, \mathbf{w}_* \rangle - \langle \mathbf{w}_0, \mathbf{w}_* \rangle\big)}{4\gamma\eta t} \le \frac{5\|\mathbf{w}_t - \mathbf{w}_0\|}{4\gamma\eta t}.$$

Plugging in the parameter bound in Lemma 28 completes the proof. ■

The following lemma is analogous to Lemma 10. However, our focus here is a nonlinear predictor that might not be twice differentiable. We address this issue using the self-boundedness of the loss in Assumption 3C and a slightly longer induction argument.

Lemma 30 (Stable phase) *Consider a loss ℓ satisfying Assumptions 3A to 3C. Suppose event $\mathcal{E}$ holds. Suppose there exists $s < T$ such that*

$$L(\mathbf{w}_s) \le \frac{1}{12C_\beta^2\eta},$$

then for every $t \in [s, T]$ we have,

1. $G(\mathbf{w}_t) \le 1/(12C_\beta\eta)$.
2. $L(\mathbf{w}_{t+1}) \le L(\mathbf{w}_t) - \frac{5\eta}{8}\|\nabla L(\mathbf{w}_t)\|^2 \le L(\mathbf{w}_t)$.

Proof (of Lemma 30) Note that $\mathbf{w}_1, \dots, \mathbf{w}_T$ all belong to $\mathcal{B}$ by Lemma 28.

We first show that Claim 1 implies Claim 2. Note that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_t\| &= \eta\|\nabla L(\mathbf{w}_t)\| \\ &\le \eta \cdot G(\mathbf{w}_t) && \text{by Lemma 26} \\ &\le \frac{1}{2}. && \text{by Claim 1}\end{aligned}$$

So we have

$$\begin{aligned}|y_i f_i(\mathbf{w}_{t+1}) - y_i f_i(\mathbf{w}_t)| &= \left|\langle y_i \nabla f_i(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \xi_i(\mathbf{w}_{t+1}, \mathbf{w}_t)\right| \\ &\le \|\mathbf{w}_{t+1} - \mathbf{w}_t\| + \frac{\gamma}{10}\|\mathbf{w}_{t+1} - \mathbf{w}_t\| && \text{by (9)} \\ &\le \frac{1}{2} + \frac{\gamma}{20} \le 1. && \text{since } \gamma \le 1\end{aligned}$$

Denote $z_{i,t} = y_i f_i(\mathbf{w}_t)$, then $|z_{i,t} - z_{i,t+1}| \le 1$. By Assumption 3C we have

$$\begin{aligned}\ell(z_{i,t+1}) &\le \ell(z_{i,t}) + \ell'(z_{i,t})(z_{i,t+1} - z_{i,t}) + C_\beta g(z_{i,t})(z_{i,t+1} - z_{i,t})^2 \\ &\le \ell(z_{i,t}) + \ell'(z_{i,t})\langle y_i \nabla f_i(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + |\ell'(z_{i,t})| \cdot |\xi_i(\mathbf{w}_{t+1}, \mathbf{w}_t)| \\ &\quad + 2C_\beta g(z_{i,t})\Big(|\langle y_i \nabla f_i(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle|^2 + |\xi_i(\mathbf{w}_{t+1}, \mathbf{w}_t)|^2\Big).\end{aligned}$$

Applying the gradient bound (9) and the local linearization error bound in Lemma 24,

$$|\xi_i(\mathbf{w}, \mathbf{v})| \le \frac{\gamma}{10}\|\mathbf{w} - \mathbf{v}\|, \quad \mathbf{w}, \mathbf{v} \in \mathcal{B},$$

we get

$$\begin{aligned}\ell(z_{i,t+1}) &\le \ell(z_{i,t}) + \ell'(z_{i,t})\langle y_i \nabla f_i(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + g(z_{i,t}) \cdot \frac{\gamma}{10}\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \\ &\quad + 2C_\beta g(z_{i,t})\Big(\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + \frac{\gamma^2}{100}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2\Big) \\ &\le \ell(z_{i,t}) + \ell'(z_{i,t})\langle y_i \nabla f_i(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + g(z_{i,t}) \cdot \frac{\gamma}{10}\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \\ &\quad + 3C_\beta g(z_{i,t})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2,\end{aligned}$$

where the last inequality is because $\gamma \le 1$. Taking average over $i = 1, \dots, n$, we get

$$\begin{aligned}L(\mathbf{w}_{t+1}) &\le L(\mathbf{w}_t) + \langle \nabla L(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + G(\mathbf{w}_t) \cdot \frac{\gamma}{10}\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \\ &\quad + 3C_\beta G(\mathbf{w}_{t+1}) \cdot \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\ &= L(\mathbf{w}_t) - \eta\|\nabla L(\mathbf{w}_t)\|^2 + G(\mathbf{w}_t) \cdot \frac{\gamma\eta}{10}\|\nabla L(\mathbf{w}_t)\| \\ &\quad + 3C_\beta \eta^2 G(\mathbf{w}_t) \cdot \|\nabla L(\mathbf{w}_t)\|^2\end{aligned}$$

$$\begin{aligned}
&\le L(\mathbf{w}_t) - \eta\|\nabla L(\mathbf{w}_t)\|^2 + \frac{\eta}{8}\|\nabla L(\mathbf{w}_t)\|^2 \qquad \text{by Lemma 26} \\
&\quad + \frac{\eta}{4}\|\nabla L(\mathbf{w}_t)\|^2 \qquad\qquad\qquad\qquad \text{by } G(\mathbf{w}_t) \le 1/(12C_\beta\eta) \\
&= L(\mathbf{w}_t) - \frac{5\eta}{8}\|\nabla L(\mathbf{w}_t)\|^2.
\end{aligned}$$

which verifies Claim 2. We have shown that Claim 1 implies Claim 2.

Next, we prove a stronger version of Claim 1 by induction, that is,

$$\text{for every } t \in [s, T], \quad L(\mathbf{w}_t) \le \frac{1}{12C_\beta^2\eta}, \quad G(\mathbf{w}_t) \le \frac{1}{12C_\beta\eta}. \tag{11}$$

- For $t = s$, using our assumption on $L(\mathbf{w}_s)$ and Lemma 26, we have

  $$G(\mathbf{w}_t) \le C_\beta L(\mathbf{w}_t) \le \frac{1}{12C_\beta\eta},$$

  which verifies (11) for s.

- Now suppose that (11) holds for $s, s+1, \ldots, t$. Since Claim 1 implies Claim 2, we have

  $$L(\mathbf{w}_{t+1}) \le L(\mathbf{w}_t) \le \cdots \le L(\mathbf{w}_s) \le \frac{1}{12C_\beta^2\eta}.$$

  From Lemma 26, we have

  $$G(\mathbf{w}_{t+1}) \le C_\beta L(\mathbf{w}_{t+1}) \le \frac{1}{12C_\beta\eta}.$$

  These together verify (11) for $t+1$ and thus complete our induction.

We have proved all claims. ■

The following lemma shows that all data is classified correctly when the training loss is low. This lemma is only used in the convergence analysis in the stable phase.

Lemma 31 *Consider a loss ℓ satisfying Assumption 3A. For every $\mathbf{w}$ such that*

$$L(\mathbf{w}) \le \frac{\ell(0)}{n},$$

then $y_i f_i(\mathbf{w}) \ge 0$ for $i = 1, \ldots, n$,

Proof (of Lemma 31) Since $\ell(\cdot) \ge 0$ by Assumption 3A, we have

$$\frac{1}{n}\ell\big(y_i f_i(\mathbf{w}_s)\big) \le L(\mathbf{w}_s) = \frac{1}{n}\sum_{i=1}^{n}\ell\big(y_i f_i(\mathbf{w}_s)\big) \le \frac{\ell(0)}{n},$$

which implies

$$\ell\big(y_i f_i(\mathbf{w})\big) \le \ell(0).$$

Then the monotonicity of $\ell(\cdot)$ by Assumption 3A implies $y_i f_i(\mathbf{w}) \ge 0$. ■

The following lemma is analogous to Lemma 11. Again, we use an additional induction argument to get a sharp width condition.

Lemma 32 (Convergence in the stable phase) *Consider a loss ℓ satisfying Assumptions 3A to 3C. Suppose event $\mathcal{E}$ holds. Suppose there exists a time $s < T$ such that*

$$L(\mathbf{w}_s) \le \min\left\{\frac{1}{12C_\beta^2\eta}, \frac{\ell(0)}{n}\right\}.$$

Then for every $0 \le t \le T - s$, we have

$$L(\mathbf{w}_{s+t}) \le 16\frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t}, \quad \|\mathbf{w}_{s+t} - \mathbf{w}_s\| \le 6\frac{\sqrt{\rho(\gamma^2\eta t)}}{\gamma}.$$

Proof (of Lemma 32) The first upper bound on $L(\mathbf{w}_s)$ enables Lemma 30 for s onwards. Therefore we have for $k \ge 0$,

$$\eta\|\nabla L(\mathbf{w}_{s+k})\|^2 \le \frac{8}{5}\big(L(\mathbf{w}_{s+k}) - L(\mathbf{w}_{s+k+1})\big) \le \frac{8}{5}L(\mathbf{w}_{s+k}). \tag{12}$$

Next, we use induction to prove the claims. For $t = 0$, the claims hold. Consider t. By the inductive hypothesis, we have

$$\text{for } k = 0, \dots, t-1, \quad \|\mathbf{w}_{s+k} - \mathbf{w}_s\| \le 6\frac{\sqrt{\rho(\gamma^2\eta k)}}{\gamma} \le 6\frac{\sqrt{\rho(\gamma^2\eta t)}}{\gamma} =: R'.$$

Choose a comparator centered at $\mathbf{w}_s$,

$$\mathbf{u} := \mathbf{w}_s + \theta\mathbf{w}_*, \quad \theta := \frac{1}{8}R' + \frac{5}{4}\frac{\sqrt{\rho\big(\gamma^2\eta t\big)}}{\gamma}.$$

For $k \le t - 1$, we have

$$\begin{aligned}
\|\mathbf{w}_{s+k+1} - \mathbf{u}\|^2 &= \|\mathbf{w}_{s+k} - \mathbf{u}\|^2 + 2\eta\langle\nabla L(\mathbf{w}_{s+k}), \mathbf{u} - \mathbf{w}_{s+k}\rangle + \eta^2\|\nabla L(\mathbf{w}_{s+k})\|^2 \\
&\le \|\mathbf{w}_{s+k} - \mathbf{u}\|^2 + 2\eta\langle\nabla L(\mathbf{w}_{s+k}), \mathbf{u} - \mathbf{w}_{s+k}\rangle + \frac{8\eta}{5}L(\mathbf{w}_{s+k}). \qquad \text{by (12)}
\end{aligned}$$

Repeating the argument in (10), we can bound the second term by

$$\langle\nabla L(\mathbf{w}_{s+k}), \mathbf{u} - \mathbf{w}_{s+k}\rangle \le \frac{1}{n}\sum_{i=1}^n \ell\big(\gamma(8\theta - \|\mathbf{w}_{s+k} - \mathbf{w}_s\|)/10 + y_i f_i(\mathbf{w}_s)\big) - L(\mathbf{w}_{s+k}).$$

The assumption that $L(\mathbf{w}_s) \le \ell(0)/n$ allows us to apply Lemma 31, so $y_i f_i(\mathbf{w}_s) \ge 0$ and thus

$$\begin{aligned}
\gamma(8\theta - \|\mathbf{w}_{s+k} - \mathbf{w}_s\|)/10 + y_i f_i(\mathbf{w}_s) &\ge \gamma(8\theta - \|\mathbf{w}_{s+k} - \mathbf{w}_s\|)/10 \\
&\ge \gamma(8\theta - R')/10 \qquad \text{by the inductive hypothesis} \\
&= \sqrt{\rho(\gamma^2\eta t)}. \qquad \text{by the choice of } \theta
\end{aligned}$$

Using the monotonicity of ℓ by Assumption 3A, we get

$$\ell\big(\gamma(8\theta - \|\mathbf{w}_{s+k} - \mathbf{w}_s\|)/10 + y_i f_i(\mathbf{w}_s)\big) \le \ell\big(\sqrt{\rho(\gamma^2\eta t)}\big).$$

So we can further control the second term by

$$\langle \nabla L(\mathbf{w}_{s+k}), \mathbf{u} - \mathbf{w}_{s+k} \rangle \le \ell\big(\gamma(8\theta - R')/10\big) - L(\mathbf{w}_{s+k}).$$

Bringing this back, we get

$$\begin{aligned}\|\mathbf{w}_{s+k+1} - \mathbf{u}\|^2 &\le \|\mathbf{w}_{s+k} - \mathbf{u}\|^2 + 2\eta\Big(\ell\big(\sqrt{\rho(\gamma^2\eta t)}\big) - L(\mathbf{w}_t)\Big) + \frac{8\eta}{5} L(\mathbf{w}_{s+k}) \\ &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta\ell\big(\sqrt{\rho(\gamma^2\eta t)}\big) - \frac{2\eta}{5} L(\mathbf{w}_{s+k}).\end{aligned}$$

Telescoping the sum from 0 to $t-1$ and rearranging, we get

$$\begin{aligned}\frac{5\|\mathbf{w}_{s+t} - \mathbf{u}\|^2}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_{s+k}) &\le 5\ell\big(\sqrt{\rho(\gamma^2\eta t)}\big) + \frac{5\|\mathbf{w}_s - \mathbf{u}\|^2}{2\eta t} \\ &\le 5\frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t} + \frac{5\|\mathbf{w}_s - \mathbf{u}\|^2}{2\eta t}. \qquad \text{by Lemma 20.} \qquad (13)\end{aligned}$$

Back to our induction. For case t, we have

$$\begin{aligned}\|\mathbf{w}_{s+t} - \mathbf{w}_s\| &\le \|\mathbf{w}_{s+t} - \mathbf{u}\| + \|\mathbf{w}_s - \mathbf{u}\| \\ &\le \sqrt{2\rho(\gamma^2\eta t)/\gamma^2} + 2\|\mathbf{w}_s - \mathbf{u}\| \qquad \text{by (13)} \\ &= \frac{\sqrt{2\rho(\gamma^2\eta t)}}{\gamma} + \frac{1}{4}R' + \frac{5}{2}\frac{\sqrt{\rho(\gamma^2\eta t)}}{\gamma} \qquad \text{by the choice of } \mathbf{u} \\ &\le R' := 6\frac{\sqrt{\rho(\gamma^2\eta t)}}{\gamma},\end{aligned}$$

and

$$\begin{aligned}\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_{s+k}) &\le 5\frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t} + \frac{5}{2\eta t}\left(\frac{1}{8}R' + \frac{5}{4}\frac{\sqrt{\rho(\gamma^2\eta t)}}{\gamma}\right)^2 \qquad \text{by (13)} \\ &= 15\frac{\rho(\gamma^2\eta t)}{\gamma^2\eta t}.\end{aligned}$$

We complete our induction by using the monotonicity of $L(\mathbf{w}_t)$ for $t \ge s$ by Lemma 30. ■

The next lemma shows a (weak) phase transition time bound without using the exponential tail condition of the loss.

Lemma 33 (Phase transition time) *Consider a loss ℓ satisfying Assumptions 3A and 3B. Suppose event $\mathcal{E}$ holds. Define*

$$\psi(\lambda) = \frac{\lambda}{\rho(\lambda)}, \quad \lambda > 0.$$

Then there is $C > 0$ as a function of C_g, C_a, C_β, $\ell(0)$, $\ln(1/\delta)$ for the following to hold. Let

$$\tau := \frac{1}{\gamma^2}\max\left\{\frac{\psi^{-1}\big(C(\eta+n)\big)}{\eta},\ C(\eta+n)\eta\right\}.$$

If $\tau < T$, then there exists $0 \le s \le \tau$ such that

$$L(\mathbf{w}_s) \le \min\left\{\frac{1}{12C_\beta^2\eta}, \frac{\ell(0)}{n}\right\}.$$

Proof (of Lemma 33) Applying Lemma 28 with $t = \tau$, we have

$$\frac{1}{\tau}\sum_{k=0}^{\tau-1} L(\mathbf{w}_k) \le 9\frac{\rho(\gamma^2\eta\tau) + \left(C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g\right)^2}{\gamma^2\eta\tau}.$$

Choose τ such that

$$\eta\gamma^2\tau \ge \max\left\{\psi^{-1}\left(18\left(12C_\beta^2\eta + \frac{n}{\ell(0)}\right)\right), \right.$$
$$\left. 18\left(12C_\beta^2\eta + \frac{n}{\ell(0)}\right)\left(C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g\right)^2\right\}.$$

It is clear that

$$\frac{1}{\psi(\lambda)} = \frac{\rho(\lambda)}{\lambda} = \min_z \ell(z) + \frac{z^2}{\lambda}$$

is a decreasing function. So for $\eta\gamma^2\tau$ we have

$$\begin{aligned} 9\frac{\rho(\gamma^2\eta\tau)}{\gamma^2\eta\tau} &= 9\frac{1}{\psi(\eta\gamma^2\tau)} \\ &\le 9\frac{1}{18\left(12C_\beta^2\eta + n/\ell(0)\right)} \\ &= \frac{1}{2}\cdot\frac{1}{12C_\beta^2\eta + n/\ell(0)}, \end{aligned}$$

and

$$9\frac{\left(C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g\right)^2}{\gamma^2\eta\tau} \le \frac{1}{2}\cdot\frac{1}{12C_\beta^2\eta + n/\ell(0)}.$$

These two inequalities together imply that

$$\begin{aligned} \frac{1}{\tau}\sum_{k=0}^{\tau-1} L(\mathbf{w}_k) &\le 9\frac{\rho(\gamma^2\eta\tau) + \left(C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g\right)^2}{\gamma^2\eta\tau} \\ &\le \frac{1}{12C_\beta^2\eta + n/\ell(0)} \\ &\le \min\left\{\frac{1}{12C_\beta^2\eta}, \frac{\ell(0)}{n}\right\}, \end{aligned}$$

which implies that there exists $s \le \tau$ for $L(\mathbf{w}_s)$ satisfies the right hand side bound. ■

The next lemma shows a stronger phase transition bound assuming an exponentially tailed loss.

Lemma 34 (Phase transition time under exponential tail) *Consider a loss ℓ satisfying Assumptions 3A, 3B and 3D. Suppose event $\mathcal{E}$ holds. Then there is $C > 0$ as a function of C_e, C_g, C_a, C_β, $\ell(0)$, $\ln(1/\delta)$ for the following to hold. Let*

$$\tau := \frac{C}{\gamma^2} \max\{\eta,\ n\ln(n)\},$$

if $\tau \le T$, then there exists $0 \le s \le \tau$ such that

$$L(\mathbf{w}_s) \le \min\left\{\frac{1}{12C_\beta^2\eta}, \frac{\ell(0)}{n}\right\}.$$

Proof (of Lemma 34) Under Assumption 3D, we have,

$$\ell(z) \le C_e g(z) = -C_e \ell'(z), \quad z \ge 0,$$

which implies

$$\frac{\ell'(z)}{\ell(z)} \le -C_e^{-1}, \quad z \ge 0.$$

Integrating both sides, we get

$$\ln \ell(x) \le \ln \ell(0) + \int_{z=0}^{x} \frac{\ell'(z)}{\ell(z)} \mathrm{d}x \le \ln \ell(0) - C_e^{-1} x, \quad x \ge 0,$$

which implies

$$\ell(x) \le \ell(0) \exp\big(-C_e^{-1}x\big), \quad x \ge 0.$$

Using the exponential tail property, we have

$$\begin{aligned}
\rho(\lambda) &= \min \lambda\ell(z) + z^2 \\
&\le \lambda\ell\big(C_e \ln(\lambda)\big) + C_e^2 \ln^2(\lambda) \\
&\le \ell(0) + C_e^2 \ln^2(\lambda). \qquad (14)
\end{aligned}$$

Applying Lemma 29 for τ, we have

$$\begin{aligned}
\frac{1}{\tau}\sum_{k=0}^{\tau-1} G(\mathbf{w}_k) &\le 12\frac{\sqrt{\rho(\gamma^2\eta\tau)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma^2\eta\tau} \\
&\le 12\frac{\sqrt{\ell(0) + C_e^2\ln^2(\gamma^2\eta\tau)} + C_a + \sqrt{2\ln(2n/\delta)} + \eta C_g}{\gamma^2\eta\tau} \qquad \text{by (14)} \\
&\le 12\frac{C_e\ln(\gamma^2\tau) + \eta(C_g + C_e) + \sqrt{\ell(0)} + C_a + \sqrt{2\ln(2n/\delta)}}{\gamma^2\eta\tau} \\
&\le 12\left(\frac{C_e}{\eta}\frac{\ln(\gamma^2\tau)}{\gamma^2\tau} + \frac{C_g + C_e}{\gamma^2\tau} + \frac{\sqrt{\ell(0)} + C_a + \sqrt{2\ln(2n/\delta)}}{\eta}\cdot\frac{1}{\gamma^2\tau}\right).
\end{aligned}$$

So there exists $C > 0$ as a function of C_e, C_g, C_a, C_β, $\ell(0)$, $\ln(1/\delta)$, such that

$$\gamma^2 \tau \geq C \max\left\{\eta,\ n\ln(n)\right\}$$

implies that

$$\frac{1}{\tau}\sum_{k=0}^{\tau-1} G(\mathbf{w}_k) \leq \min\left\{\frac{1}{12C_eC_\beta^2\eta},\ \frac{\ell(0)}{C_e n}\right\}.$$

So there exists $s \leq \tau$ such that

$$G(\mathbf{w}_s) \leq \min\left\{\frac{1}{12C_eC_\beta^2\eta},\ \frac{\ell(0)}{C_e n}\right\}.$$

The last upper bound ensures that for every i,

$$\frac{1}{n}g\big(y_i f_i(\mathbf{w}_s)\big) \leq G(\mathbf{w}_s) = \frac{1}{n}\sum_{i=1}^{n} g\big(y_i f_i(\mathbf{w}_s)\big) \leq \frac{\ell(0)}{C_e n} \leq \frac{g(0)}{n},$$

where the last inequality is due to Assumption 3D. The above implies

$$y_i f_i(\mathbf{w}_s) \geq 0$$

since $g(\cdot)$ is non-increasing by Assumption 3A. So we can apply Assumption 3D for $y_i f_i(\mathbf{w}_s)$ and get

$$\ell(y_i f_i(\mathbf{w}_s)) \leq C_e g(y_i f_i(\mathbf{w}_s)).$$

Taking an average over $i = 1, \ldots, n$, we get

$$L(\mathbf{w}_s) \leq C_e G(\mathbf{w}_s).$$

We complete the proof by plugging in the upper bound on $G(\mathbf{w}_s)$. ■

The proof of Theorem 6 follows from the above lemmas.

Proof (of Theorem 6) It follows from Lemmas 28 and 32 to 34. ■

Appendix I. General Losses and Linear Model

We restate our Theorem 6 for a linear model under general loss functions.

Theorem 35 (General losses and linear model) *Consider (GD) with stepsize $\eta > 0$ for linear classification*

$$L(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \mathbf{x}_i^\top \mathbf{w}), \quad \mathbf{w} \in \mathbb{R}^d,$$

where the dataset $(\mathbf{x}_i, y_i)_{i=1}^n$ satisfies Assumption 1 and the loss function ℓ satisfies Assumptions 3A and 3B. Then we have the following.

- ***The EoS phase.*** *For every $t > 0$ (and in particular in the EoS phase), we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\mathbf{w}_k) \le 9\frac{\rho(\gamma^2\eta t) + \big(\eta C_g\big)^2}{\gamma^2\eta t}.$$

- ***The stable phase.*** *Assume that the loss ℓ also satisfies Assumption 3C. If s is such that*

$$L(\mathbf{w}_s) \le \min\left\{\frac{1}{12C_\beta^2\eta},\ \frac{\ell(0)}{n}\right\}, \tag{15}$$

then (GD) is in the stable phase, that is, for every $t \in [s, T]$, $L(\mathbf{w}_t)$ decreases monotonically, and moreover,

$$L(\mathbf{w}_t) \le 15\frac{\rho(\gamma^2\eta(t-s))}{\gamma^2\eta(t-s)}.$$

- ***Phase transition time.*** *There exists a constant $C_1 > 0$ that only depends on C_g, C_β, and $\ell(0)$ such that the following holds. Let*

$$\tau := \frac{1}{\gamma^2}\max\left\{\frac{\psi^{-1}\big(C_1(\eta+n)\big)}{\eta},\ C_1(\eta+n)\eta\right\},\ \ \textit{where}\ \ \psi(\lambda) := \frac{\lambda}{\rho(\lambda)}.$$

If $\tau \le T$, then (15) holds for some $s \le \tau$.

- ***Phase transition time under an exponential tail.*** *Assume that the loss ℓ further satisfies Assumption 3D, then there exists a constant $C_2 > 0$ that only depends on C_e, C_g, C_β, $\ell(0)$, and $\ln(1/\delta)$ such that the following holds. Let*

$$\tau := \frac{C_2}{\gamma^2}\max\big\{\eta,\ n\big\}.$$

If $\tau \le T$, then (15) holds for some $s \le \tau$.

Proof (of Theorem 35) This is proved by reusing results in Appendix H.3 under a good event where the margin is $\gamma \ge 0.9\gamma$, the initial predictor is zero,

$$y_i\mathbf{x}_i^\top\mathbf{w}_0 = 0,$$

the linearization error is zero, and the predictor gradient difference is zero (that is, for $f(\mathbf{x};\mathbf{w}) = \mathbf{x}^\top\mathbf{w}$, we have $\|\nabla f(\mathbf{x};\mathbf{w}) - \nabla f(\mathbf{x};\mathbf{v})\| = 0$).

We remark that different from Theorem 6, here the phase transition time bound under an exponential tail only depends on n instead of $n\ln(n)$. This difference is because the initial predictor in the NTK case depends on $\ln(n)$, while our initial predictor is zero. See more details in the proof of Lemma 34. ■