Tight Bounds for Logistic Regression with Large Stepsize Gradient Descent in Low Dimension

Michael Crawshaw MCRAWSHA@GMU.EDU
Mingrui Liu MINGRUIL@GMU.EDU
Department of Computer Science, George Mason University, Fairfax, VA 22030

Editors: Steve Hanneke and Tor Lattimore

Abstract

We consider the optimization problem of minimizing the logistic loss with gradient descent to train a linear model for binary classification with separable data. With a budget of T iterations, it was recently shown that an accelerated $1/T^2$ rate is possible by choosing a large stepsize $\eta = \Theta(\gamma^2 T)$ (where γ is the dataset's margin) despite the resulting non-monotonicity of the loss. In this paper, we provide a tighter analysis of gradient descent for this problem when the data is two-dimensional: we show that GD with a sufficiently large learning rate η finds a point with loss smaller than $\mathcal{O}(1/(\eta\gamma^2 T))$, as long as $T \geq \Omega(n/\gamma + 1/\gamma^2)$, where n is the dataset size. Our improved rate comes from a tighter bound on the time τ that it takes for GD to transition from unstable (non-monotonic loss) to stable (monotonic loss), via a fine-grained analysis of the oscillatory dynamics of GD in the subspace orthogonal to the max-margin classifier. We also provide a lower bound of τ matching our upper bound up to logarithmic factors, showing that our analysis is tight.

1. Introduction

In modern machine learning, optimization algorithms tend to operate in "unstable" regimes, where the loss does not monotonically decrease over time, even with full-batch gradients (Cohen et al., 2021). However, the theory of optimization for machine learning largely considers only stable regimes, where sufficiently small stepsizes for gradient descent (GD) and its variants will safely ensure monotonic loss decrease (Nesterov, 2013). Our limited understanding of unstable optimization has created a significant gap between optimization algorithms that work, and optimization algorithms that we understand theoretically.

In this paper, we consider the problem of training linear models for binary classification by minimizing the logistic loss using gradient descent with large stepsizes. Despite this problem's simplicity and its fundamental role in machine learning, the optimization dynamics of gradient descent with large step sizes in this setting is still not entirely understood. We aim to provide upper and lower bounds on the iteration complexity required for gradient descent to find an approximate solution to the optimization problem in question.

Given a dataset of n samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where each sample consists of an input $\boldsymbol{x}_i \in \mathbb{R}^d$ and a label $y_i \in \{-1, 1\}$, we denote $\ell(z) = \log(1 + \exp(-z))$ and consider the optimization problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ F(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^n \ell(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \right\}. \tag{1}$$

We make the following assumptions on the dataset.

Assumption 1.1

(a) Linear separability: There exists $\boldsymbol{w} \in \mathbb{R}^d$ *such that* $y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0$ *for all* $i \in [n]$.

(b) Bounded data norm: $\|\boldsymbol{x}_i\| \leq 1$ *for all* $i \in [n]$.

(c) Identical label: $y_i = 1$ *for all* $i \in [n]$.

Assumption 1.1(b) can be satisfied in practice by replacing all data $\boldsymbol{x}_i$ with $\boldsymbol{x}_i / \max_{j \in [n]} \|\boldsymbol{x}_j\|$, and Assumption 1.1(c) can be made without loss of generality: the objective F only depends on the label y_i through the product $y_i \boldsymbol{x}_i$, so that any sample $(\boldsymbol{x}_i, -1)$ can be replaced by $(-\boldsymbol{x}_i, 1)$ while preserving the objective. The same assumptions are made in previous work (Wu et al., 2023, 2024).

For a given dataset, we define the maximum margin and the associated classifier by

$$\gamma = \max_{\|\boldsymbol{w}\|=1} \min_{i \in [n]} y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle, \quad \boldsymbol{w}_* = \operatorname*{argmax}_{\|\boldsymbol{w}\|=1} \min_{i \in [n]} y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle. \tag{2}$$

Note $0 < \gamma < 1$, since separable data implies $\gamma > 0$ and bounded data implies $\gamma \leq \|\boldsymbol{x}_i\| \leq 1$.

We consider Gradient Descent (GD) with a constant stepsize $\eta > 0$ for minimizing Equation (1):

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla F(\boldsymbol{w}_t). \tag{3}$$

Similar to previous work (Wu et al., 2023), we fix the initialization $\boldsymbol{w}_0 = \boldsymbol{0}$, though our results can be extended for general initialization.

This optimization problem is convex, smooth, and Lipschitz, so classical theory (Nesterov, 2013) provides myriad guarantees for GD with sufficiently small stepsize. In particular, since this problem is L-smooth with $L = 1/4$, we can show that $F(\boldsymbol{w}_t) \leq \widetilde{\mathcal{O}}(1/(\gamma^2 t))$ when $\eta = 1/L = 4$.

However, the classical rate is not the end of the story. Recently, Wu et al. (2024) showed that GD for T iterations satisfies $F(\boldsymbol{w}_T) \leq \mathcal{O}(1/(\gamma^4 T^2))$ if $\eta = \Theta(\gamma^2 T)$, so a large stepsize accelerates optimization for this problem, despite the resulting non-monotonicity of the loss. In this work, we show that for the low-dimensional case $d = 2$, this rate can be further improved to $\mathcal{O}(1/(\eta \gamma^2 T))$ for any $\eta \geq \widetilde{\Omega}(n + 1/\gamma^2)$, as long as $T \geq \Omega(n/\gamma + \log(1/\gamma)/\gamma^2)$. We achieve this improved rate via a sharper bound of the time it takes to transition into a stable phase, based on a fine-grained analysis of the *oscillatory dynamics* of $\boldsymbol{w}_t$ in the subspace orthogonal to $\boldsymbol{w}_*$. We also provide a lower bound on the transition time that matches our upper bound up to logarithmic factors.

From a higher level, we should point out that our goal here is not to achieve the fastest optimization guarantees by any means necessary. Rather, our primary motivation is to develop a fine-grained understanding of the unstable dynamics of GD with large stepsizes. Given that optimization in machine learning tends to operate in unstable regimes in practice (Cohen et al., 2021), we believe that it is important to develop a rigorous mathematical understanding of unstable optimization in machine learning with fundamental algorithms like GD, and our results here are a step in this direction.

Notation After the abstract, we use $\mathcal{O}, \Omega, \Theta$ to omit only universal constants, and $\widetilde{\mathcal{O}}, \widetilde{\Omega}$, and $\widetilde{\Theta}$ to omit only universal constants and factors logarithmic in $n, 1/\gamma, t, 1/\epsilon$. We denote $[n] = \{1, \ldots, n\}$.

1.1. Technical Overview

Characterizing the instability induced by large stepsizes is a fundamental difficulty of analyzing GD in our setting. We know from previous work (Wu et al., 2024) that if $F(\boldsymbol{w}_t) \leq 2/\eta$ for some t, then

	Complexity	Stepsize	Setting
Gradient Descent (Nesterov, 2013)	$\mathcal{O}\left(LB^2/\epsilon\right)$	$\eta = 1/L$	Convex, L-smooth
First-Order Algorithms (Nesterov, 2013)	$\Omega\left(B\sqrt{L/\epsilon}\right)$	-	Convex, L-smooth
Gradient Descent (Wu et al., 2024)	$\widetilde{\mathcal{O}}\left(\frac{1}{\gamma^2\sqrt{\epsilon}}\right)$	$\eta = \Theta\left(1/\sqrt{\epsilon}\right)$	Logistic regression
Adaptive Gradient Descent (Zhang et al., 2025a)	$\mathcal{O}\left(1/\gamma^2\right)$	-	Logistic regression
First-Order Algorithms (Zhang et al., 2025a)	$\Omega\left(\min\left(\log n, 1/\gamma^2\right)\right)^{(a)}$	-	Logistic regression
Gradient Descent (Theorem 2.1)	$\mathcal{O}\left(\frac{n}{\gamma} + \frac{\log(1/\gamma)}{\gamma^2}\right)$	$\eta \geq \Omega\left(\frac{1}{\epsilon(\gamma n+1)}\right)$	Logistic regression $d = 2$
Gradient Descent (Theorem 3.1)	$\Omega\left(\frac{n}{\gamma} + \frac{1}{\gamma^2}\right)^{(b)}$	$\eta \geq \widetilde{\Omega}\left(n + \frac{1}{\gamma^2}\right)$	Logistic regression

Table 1: Iteration complexity to find an ϵ-approximate solution of Equation (1). Note that (a) and (b) show the time to find a linear separator and time to reach $\mathcal{O}(1/\eta)$ loss, respectively, however both of these conditions are necessary for finding an ϵ-approximate solution for sufficiently small ϵ and sufficiently large η.

the loss $F(\boldsymbol{w}_s)$ decreases monotonically for $s \geq t$. So defining $\tau = \min\{t \geq 0 : F(\boldsymbol{w}_t) \leq 1/8\eta\}$[1], we divide the trajectory into two phases: the unstable phase $t \leq \tau$, where the loss may be non-monotonic, and the stable phase $t \geq \tau$, where it is known that $F(\boldsymbol{w}_t) \leq \widetilde{\mathcal{O}}(1/(\gamma^2\eta t))$. The challenge is proving that GD must enter the stable phase, and bounding the time when this happens.

Wu et al. (2024) used a "split comparator" technique to prove that $\tau \leq \widetilde{\mathcal{O}}(\eta/\gamma^2)$ (for sufficiently large η), which aligns with the intuitive feeling that the length of the unstable phase might increase under larger stepsizes. A key part of their argument is the connection between $\hat{w}_t := \langle \boldsymbol{w}_t, \boldsymbol{w}_* \rangle$ and a quantity called the gradient potential:

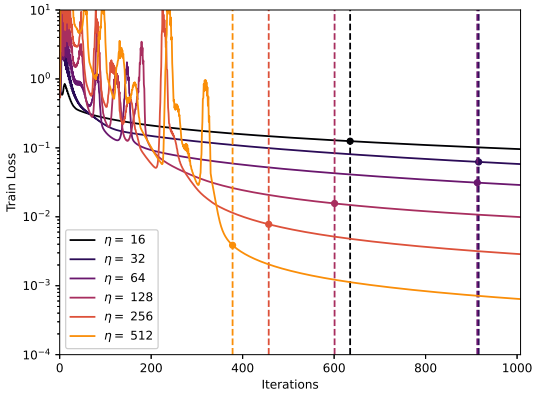
$$G(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} |\ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)| = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\exp(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) + 1}. \tag{4}$$

To rephrase their argument, the component $\hat{w}_t$ of $\boldsymbol{w}_t$ in the direction of the max-margin classifier increases at least proportionally to the gradient potential:

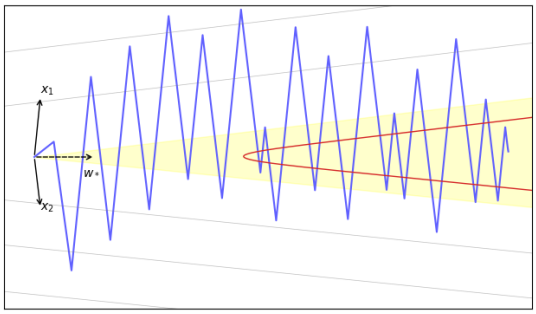
$$\hat{w}_{t+1} - \hat{w}_t = \langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \boldsymbol{w}_* \rangle = \frac{\eta}{n}\sum_{i=1}^{n} \frac{\langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) + 1} \geq \eta\gamma G(\boldsymbol{w}_t), \tag{5}$$

where the last inequality uses that $\langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle \geq \gamma$. Now, to bound τ, we note that $t < \tau$ means $F(\boldsymbol{w}_t) \geq \Omega(1/\eta)$, so that $G(\boldsymbol{w}_t) \geq \Omega(1/\eta)$ (by Lemma C.1); this means $\hat{w}_{t+1} - \hat{w}_t \geq \Omega(\gamma)$, so $\hat{w}_t$ must increase at least *linearly* during the unstable phase. Combining with $\hat{w}_t \leq \|\boldsymbol{w}_t\| \leq \widetilde{\mathcal{O}}(\eta/\gamma)$ (the second inequality is proven by Wu et al. (2024) by other means), we conclude that the unstable phase cannot last more than $\widetilde{\mathcal{O}}(\eta/\gamma^2)$ iterations.

1. Wu et al. (2024) defined τ as the first time at which $F(\boldsymbol{w}_t) \leq 1/\eta$, compared to our $1/8\eta$, though this difference only affects universal constants in the analysis.



(*a*) MNIST loss and transition times.



(*b*) Example GD trajectory in two dimensions.

Figure 1: **(a)** For GD on a subset of MNIST ($n = 256$, binarized labels), larger learning rates create instability and faster optimization. As the stepsize increases exponentially, the stable transition time (i.e. the first timestep at which $F(\boldsymbol{w}_t) \leq 2/\eta$) does not increase. **(b)** Data $\boldsymbol{x}_i$ and max-margin direction $\boldsymbol{w}_*$ are shown in black, GD trajectory in blue, the contour line $F(\boldsymbol{w}) = 2/\eta$ in red, and the region where $\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 0$ for both $\boldsymbol{x}_i$ in yellow. Our proof uses the idea that avoiding the region where $F(\boldsymbol{w}) \leq \mathcal{O}(1/\eta)$ requires large oscillations in the subspace orthogonal to $\boldsymbol{w}_*$, and every such oscillation necessarily yields fast progress in the direction of $\boldsymbol{w}_*$.

However, it is not known if the bound $\tau \leq \widetilde{\mathcal{O}}(\eta/\gamma^2)$ is tight, and experimental observations suggest that it is not: Figure 1(*a*) shows that for MNIST data, τ does not seem to increase with η.

Our Upper Bound The argument above essentially takes a *static* view of the gradient potential: we pessimistically allowed the possibility that $G(\boldsymbol{w}_t) \approx 1/\eta$ for every $t < \tau$. It might hold that $G(\boldsymbol{w}_t) \approx 1/\eta$ for a single iteration, but can this really occur at every iteration before τ along the trajectory of GD? To get a tighter bound of τ, we will take a *dynamical* viewpoint by considering the gradient potential along a trajectory that avoids the sublevel set where $F(\boldsymbol{w}_t) \leq 1/8\eta$.

The intuition behind our proof is demonstrated in Figure 1(*b*). By convexity of F, the gradient points into the sublevel set where $F(\boldsymbol{w}) \leq 1/8\eta$, so that after a certain point, the only way for GD to avoid the sublevel set is to "jump" over it and land on the other side. Notice that the "height" of the sublevel set (that is, the length of each cross-section orthogonal to $\boldsymbol{w}_*$) tends to increase linearly along the $\boldsymbol{w}_*$ axis, so a jump over the sublevel set requires a parameter update with magnitude proportional to $\hat{w}_t$. More precisely, if t is a step where such a jump occurs, we show that $\hat{w}_{t+1} - \hat{w}_t \geq \gamma^2 \hat{w}_t$, that is, $\hat{w}_t$ increases *exponentially*! We then show that these oscillations across the sublevel set must happen with a certain frequency. Compared to the linear rate of $\hat{w}_t$ from the argument of Wu et al. (2024), our exponential rate of $\hat{w}_t$ implies the tighter bound $\tau \leq \widetilde{\mathcal{O}}(n/\gamma + 1/\gamma^2)$. This shows that the transition time can be bounded independently of η, so that the stable, accelerated second phase can be quickly reached even with arbitrarily large stepsize.

Our Lower Bound We further show that our bound $\tau \leq \widetilde{\mathcal{O}}(n/\gamma + 1/\gamma^2)$ is tight up to logarithmic factors in the worst-case (for GD) over datasets satisfying Assumption 1.1. We provide two hard instances. The first instance requires $\Omega(n/\gamma)$ iterations until all n data points are correctly classified (which is a necessary condition for $F(\boldsymbol{w}_t) \leq 1/8\eta$ when η is sufficiently large), which is a slight

generalization of a lower bound construction from Tyurin (2025). On the second instance, GD correctly classifies all points quickly, but requires $\Omega(1/\gamma^2)$ iterations until $F(\boldsymbol{w}_t) \leq 2/\eta$.

1.2. Related Work

Logistic Regression Many works in recent years have studied logistic regression as a fundamental testbed for optimization in machine learning. The role played by gradient-based optimization methods in generalization was studied through implicit bias, first for GD under separable data (Soudry et al., 2018) and non-separable data (Ji and Telgarsky, 2019), and later for SGD (Nacson et al., 2019) and steepest descent with momentum (Gunasekar et al., 2018).

Recently, it was shown that GD for logistic regression with separable data can converge with any stepsize (Wu et al., 2023), and it was later shown that a large stepsize could induce an accelerated convergence rate, despite the resulting instability (Wu et al., 2024). The techniques used by Wu et al. (2024) were subsequently applied to achieve accelerated rates in various settings, such as for two-layer networks (Cai et al., 2024), regularized logistic regression (Wu et al., 2025), and GD with adaptive stepsizes (Zhang et al., 2025a). For non-separable data, the behavior of GD with large stepsizes was explored by Meng et al. (2024, 2025), who provided negative results showing that global convergence is not guaranteed when the stepsize is larger than a critical threshold.

Zhang et al. (2025a) also provided lower bounds, showing that any first-order optimization algorithm that minimizes the logistic loss requires $\Omega(\min(\log n, 1/\gamma^2))$ iterations to correctly classify all data. Also, Kornowski and Shamir (2024) use a game-theoretic formulation to provide a lower bound of $\Omega(1/\gamma^2)$ for finding a linear separator for algorithms that use a "one-sided" oracle (which includes first-order optimization algorithms that minimize the logistic loss), and another lower bound of $\Omega(1/\gamma^{2/3})$ for a broader class of algorithms. Note that the lower bounds of Kornowski and Shamir (2024) are formulated as the worst-case over all dataset sizes n, whereas our formulation (and that of Zhang et al. (2025a)) considers a fixed n.

Edge of Stability Our study is motivated by the ubiquity of unstable optimization in practical machine learning. Cohen et al. (2021) discovered the Edge of Stability (EoS) phenomenon, where GD in deep learning operates in unstable regimes, with loss not decreasing monotonically but still tending to decrease in the long term. EoS was also observed for adaptive optimization algorithms (Cohen et al., 2022), and was later elaborated by the central flows framework (Cohen et al., 2025).

Many follow up works have studied EoS theoretically. Arora et al. (2022) showed that, under certain general conditions on the loss, GD at the edge of stability follows a deterministic flow on the manifold of global minimizers. Several works have studied surrogate models of EoS dynamics, such as 4 layer scalar networks (Zhu et al., 2023), two-layer, one-neuron neural networks (Chen and Bruna, 2023), diagonal linear networks (Even et al., 2023), and a two-parameter model of two-layer ReLU networks (Ahn et al., 2023). Damian et al. (2023) proved that GD at EoS has a self-stabilization property for objective functions with a progressive sharpening property.

In this work, our goal is not to study EoS in deep learning, but rather to provide a tight, mathematically rigorous characterization of gradient descent under unstable regimes of a natural learning problem. See Cohen et al. (2025) for a comprehensive review of the literature around EoS.

Large Stepsizes in Convex Optimization GD for smooth, convex optimization can also be accelerated by allowing large steps/non-monotonic loss. Classical methods such as mirror descent (Bubeck, 2015) and Nesterov acceleration (Nesterov, 2013) do not require monotonic decrease of

the loss. Malitsky and Mishchenko (2020) proposed an adaptive stepsize for gradient descent based on local smoothness rather than global smoothness, and which does not enforce monotonic loss decrease. Altschuler (2018) showed that the classical convergence rate of gradient descent can be improved by constant factors, at least for a couple of iterations, with a particular stepsize schedule that occasionally includes very big steps. Grimmer (2024) showed that a similar improvement by constant factors can be achieved for longer horizons (up to 127 steps). Concurrently, Altschuler and Parrilo (2024) and Grimmer et al. (2025) showed accelerated convergence rates with stepsize schedules that include occasional large steps. Zhang et al. (2025b) achieved similar acceleration with an "anytime" guarantee, where the stepsize schedule is not defined in terms of a prior stopping time, and the convergence guarantee holds for any stopping time. See Altschuler and Parrilo (2024) for a thorough discussion of this line of work.

2. Upper Bounding the Stable Transition Time

In this section, we present our improved convergence analysis of GD for Equation (1) in two dimensions, which is based on a sharper analysis of the time required for GD to transition from unstable to stable. For the entirety of this section, we fix $d = 2$.

2.1. Statement of Results

Theorem 2.1 *If $d = 2$ and $\eta \geq \eta_0 := \max(n, 32/\gamma^2 \log(256/\gamma^2))$, then the transition time $\tau := \min\{t \geq 0 : F(\boldsymbol{w}_t) \leq 1/8\eta\}$ of GD for Equation (1) satisfies*

$$\tau \leq \mathcal{O}\left(\frac{n}{\gamma} + \frac{\log(1/\gamma)}{\gamma^2}\right). \tag{6}$$

Further, $F(\boldsymbol{w}_t) \leq \mathcal{O}\left(1/(\eta\gamma^2(t-\tau))\right)$ for all $t > \tau$.

The key feature of the above theorem is the fact that the transition time τ is bounded by a quantity independent of η, whereas the previously best known bound was $\tau \leq \widetilde{\mathcal{O}}(\eta/\gamma^2)$ (Wu et al., 2024), and this dependence is a crucial bottleneck for the overall convergence rate. Indeed, if we only know $\tau \leq \widetilde{\mathcal{O}}(\eta/\gamma^2)$, then for a budget of T iterations the largest acceptable stepsize which ensures that GD enters the stable phase is $\eta = \widetilde{\Theta}(\gamma^2 T)$, which leads to the $\widetilde{\mathcal{O}}(1/\gamma^4 T^2)$ rate of Wu et al. (2024). With our improved transition time $\tau \leq \widetilde{\mathcal{O}}(n/\gamma + 1/\gamma^2)$, GD will definitely enter the stable phase as long as $T \geq \widetilde{\Omega}(n/\gamma + 1/\gamma^2)$, even with arbitrary large η, and such a large η accelerates convergence during the stable phase. This difference is shown in the complexities of Table 1; Theorem 2.1 implies that GD finds an ϵ-approximate solution in time independent of ϵ!

Note from Table 1 that the complexity of GD with a constant stepsize matches that of GD with the adaptive stepsize of Zhang et al. (2025a) in terms of the dependence on ϵ and γ, and is worse in terms of n. This establishes exactly when Adaptive GD outperforms GD for this problem: if $n \leq \mathcal{O}(1/\gamma)$, then the two algorithms have the same worst-case complexity, and otherwise Adaptive GD is provably faster by a factor of $n\gamma$.

We can also compare GD against the lower bounds on all first-order algorithms from Zhang et al. (2025a), who showed two separate lower bounds on the time to find a linear separator: $\Omega(\min(1/\gamma^2, \log n))$ and $\Omega(\min(1/\gamma^{2/3}, n))$. In the regime of a large dataset $n \geq \Omega(\exp(1/\gamma^2))$, their combined lower bounds simplify to $\Omega(1/\gamma^2)$, so that GD is suboptimal by a factor of $n\gamma$, while Adaptive GD is optimal. For a small dataset $n \leq \mathcal{O}(1/\gamma^{2/3})$, the combined lower bounds

simplify to $\Omega(n)$, which is not met by any first-order algorithm. Finally, in the intermediate regime $\Omega(1/\gamma^{2/3}) \leq n \leq \mathcal{O}(\exp(1/\gamma^2))$, the combined lower bounds simplify to $\Omega(1/\gamma^{2/3} + \log n)$, for which GD is suboptimal in both n and γ, and Adaptive GD is suboptimal in γ. In all regimes, for the problem of making the loss smaller than ϵ, both GD and Adaptive GD match the optimal complexity in terms of the dependence on ϵ alone, namely both algorithms can do so in time independent of ϵ.

2.2. Previous Bottleneck: Average-Iterate vs Last-Iterate

In Section 1.1, we rephrased the argument of Wu et al. (2024) that $\tau \leq \tilde{\mathcal{O}}(\eta/\gamma^2)$, which pessimistically allows for the possibility that $G(\boldsymbol{w}_t) \approx 1/\eta$ for every $t < \tau$. A related bottleneck of their proof is the analysis of the average gradient potential $\frac{1}{t}\sum_{s=0}^{t-1} G(\boldsymbol{w}_s)$. Specifically, denoting $\tilde{\tau}$ as the first iteration where $\frac{2}{t}\sum_{s=0}^{t-1} G(\boldsymbol{w}_s) \leq 1/8\eta$, the analysis of Wu et al. (2024) uses

$$\min_{0\leq s<t} F(\boldsymbol{w}_s) \leq 2 \min_{0\leq s<t} G(\boldsymbol{w}_s) \leq \frac{2}{t}\sum_{s=0}^{t-1} G(\boldsymbol{w}_s), \tag{7}$$

(where the first inequality uses that $F(\boldsymbol{w}) \leq 2G(\boldsymbol{w})$ for all $\boldsymbol{w}$ when $G(\boldsymbol{w})$ is small enough, see Lemma C.6), and concludes that $\tau \leq \tilde{\tau}$, then proceeds to bound $\tilde{\tau}$. However, a quick argument (which we give below) shows that $\tilde{\tau}$ is necessarily linear in η in the worst-case. From our lower bound in Section 3, we know that there is a dataset for which at least one point $\boldsymbol{x}_j$ is misclassified (i.e. $\langle \boldsymbol{w}_t, \boldsymbol{x}_j\rangle \leq 0$) for the first $\Theta(n/\gamma)$ iterations. So

$$G(\boldsymbol{w}_t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i\rangle) + 1} \geq \frac{1}{n}\frac{1}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_j\rangle) + 1} \geq \frac{1}{2n}. \tag{8}$$

for the first $\Theta(n/\gamma)$ iterations. Therefore, for $t \geq \Theta(n/\gamma)$,

$$\frac{2}{t}\sum_{s=0}^{t-1} G(\boldsymbol{w}_s) \geq \frac{2}{t}\cdot\frac{1}{2n}\Theta\left(\frac{n}{\gamma}\right) = \Theta\left(\frac{1}{\gamma t}\right). \tag{9}$$

Therefore, $\frac{2}{t}\sum_{s=0}^{t-1} G(\boldsymbol{w}_s) \geq 1/8\eta$ whenever $t \leq \mathcal{O}(\eta/\gamma)$, so $\tilde{\tau} \geq \Omega(\eta/\gamma)$. Therefore, using $\tau \leq \tilde{\tau}$ and bounding the time-averaged potential is insufficient to upper bound τ independently of η. To achieve such a bound, we need a more fine-grained analysis that considers $G(\boldsymbol{w}_t)$ for individual t.

2.3. Proof of Theorem 2.1

Notation Recall that $\boldsymbol{w}_* = \operatorname{argmax}_{\|\boldsymbol{w}\|=1} \min_{i\in[n]} \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle$. Choose $\boldsymbol{v}_*$ with $\langle \boldsymbol{v}_*, \boldsymbol{w}_*\rangle = 0$ with $\|\boldsymbol{v}_*\| = 1$. We define $\hat{w}_t = \langle \boldsymbol{w}_t, \boldsymbol{w}_*\rangle$ and $\tilde{w}_t = \langle \boldsymbol{w}_t, \boldsymbol{v}_*\rangle$. For each data $i \in [n]$, we define $\tilde{x}_i = \langle \boldsymbol{x}_i, \boldsymbol{v}_*\rangle$ and $a_t^i = \langle \boldsymbol{w}_t, \boldsymbol{x}_i\rangle$. Note that the loss for each data point $\ell(a_t^i)$ is decreasing in a_t^i, and

$$a_t^i = \langle \boldsymbol{w}_t, \boldsymbol{x}_i\rangle = \hat{w}_t\langle \boldsymbol{w}_*, \boldsymbol{x}_i\rangle + \tilde{w}_t\tilde{x}_i \geq \gamma\hat{w}_t + \tilde{w}_t\tilde{x}_i, \tag{10}$$

which we will use repeatedly.

Given $\eta > 0$, we define $\lambda = \log(1/(\exp(1/8\eta) - 1))/\gamma$, so that $F(\lambda\boldsymbol{w}_*) \leq \ell(\gamma\lambda) = 1/8\eta$. We will sometimes use the slightly larger but more convenient $\tilde{\lambda} = \log(8\eta)/\gamma$, and $\lambda \leq \tilde{\lambda}$ can be seen by applying $\exp(1/8\eta) \geq 1 + 1/8\eta$ in the definition of λ. Similarly, we have $F(\tilde{\lambda}\boldsymbol{w}_*) \leq 1/8\eta$.

Denoting $\eta_0 = \max(n, 32/\gamma^2 \log(256/\gamma^2))$, we will often require $\eta \geq \eta_0$.

At each iteration, we will split the dataset into two subsets depending on whether the current iterate $\boldsymbol{w}_t$ has positive or negative alignment with each $\boldsymbol{x}_i$ in the subspace orthogonal to $\boldsymbol{w}_*$:

$$D_t^+ = \{i \in [n] \mid \tilde{x}_i \tilde{w}_t \geq 0\}, \quad D_t^- = \{i \in [n] \mid \tilde{x}_i \tilde{w}_t < 0\} \tag{11}$$

So for $i \in D_t^+$, we have $a_t^i \geq \gamma \hat{w}_t + \tilde{w}_t \tilde{x}_i \geq \gamma \hat{w}_t$, which we will show is large for all $t \geq 1$. Essentially, the loss for each data point $i \in D_t^+$ is negligible for $t \geq 1$.

We start by establishing the linear growth of $\hat{w}_t$ as discussed in Section 1.1 (proof in Appendix A).

Lemma 2.1 *$\hat{w}_t$ is strictly increasing. Also, if $\eta \geq \eta_0$, then $\hat{w}_t \geq \gamma\eta/2 + \gamma(t-1)/16$ for all $1 \leq t \leq \tau$. In particular, $\hat{w}_t \geq 8\tilde{\lambda}$ for all $t \geq 1$.*

We will say that an *oscillation* occurs at iteration $t \geq 0$ if all of the following hold: **(1)** $\hat{w}_t \geq \lambda$. **(2)** $F(\boldsymbol{w}_t) > 1/8\eta$ and $F(\boldsymbol{w}_{t+1}) > 1/8\eta$. **(3)** $\tilde{w}_{t+1}\tilde{w}_t < 0$, that is, $\tilde{w}_t$ changes sign from t to $t+1$.

Note that condition 2 means $t < \tau - 1$. Essentially, an oscillation happens when the trajectory jumps over the sublevel set where $F(\boldsymbol{w}) \leq 1/8\eta$ (see Figure 1(*b*)). First, we show that $\hat{w}_t$ increases geometrically whenever an oscillation occurs, but an oscillation can only happen when $\hat{w}_t \leq \eta/\gamma$.

Lemma 2.2 *If an oscillation happens at iteration t and $\eta \geq \eta_0$, then $\hat{w}_{t+1} \geq (1+\gamma^2)\hat{w}_t$.*

Proof

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| = \eta \left\| \frac{1}{n} \sum_{i=1}^n \frac{\boldsymbol{x}_i}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) + 1} \right\| \leq \frac{\eta}{n} \sum_{i=1}^n \frac{\|\boldsymbol{x}_i\|}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) + 1} \tag{12}$$

$$\leq \frac{\eta}{n} \sum_{i=1}^n \frac{1}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) + 1} = \eta G(\boldsymbol{w}_t) \leq \frac{1}{\gamma}(\hat{w}_{t+1} - \hat{w}_t), \tag{13}$$

where the last inequality uses $\hat{w}_{t+1} - \hat{w}_t \geq \eta\gamma G(\boldsymbol{w}_t)$ from Equation (5). Therefore

$$\hat{w}_{t+1} - \hat{w}_t \geq \gamma \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| \geq \gamma |\tilde{w}_{t+1} - \tilde{w}_t| \overset{(i)}{=} \gamma(|\tilde{w}_{t+1}| + |\tilde{w}_t|) \tag{14}$$

$$\overset{(ii)}{\geq} \gamma^2 \hat{w}_{t+1}/2 + \gamma^2 \hat{w}_t/2 \overset{(iii)}{\geq} \gamma^2 \hat{w}_t, \tag{15}$$

where (i) uses that $\tilde{w}_t$ and $\tilde{w}_{t+1}$ have different signs (from the definition of oscillation), (ii) uses Lemma A.1, and (iii) uses $\hat{w}_{t+1} \geq \hat{w}_t$ from Lemma 2.1. ■

Lemma 2.3 *If an oscillation happens at iteration t and $\eta \geq \eta_0$, then $\hat{w}_t \leq \eta/\gamma$.*

Proof Since $t+1 < \tau$, we know from Lemma A.1 that $|\tilde{w}_t| \geq \gamma \hat{w}_t/2$ and $|\tilde{w}_{t+1}| \geq \gamma \hat{w}_{t+1}/2 \geq \gamma \hat{w}_t/2$, where the last inequality uses Lemma 2.1. From the definition of oscillation, we know that $\tilde{w}_t$ and $\tilde{w}_{t+1}$ have different signs, so $|\tilde{w}_{t+1} - \tilde{w}_t| = |\tilde{w}_{t+1}| + |\tilde{w}_t| \geq \gamma \hat{w}_t$. Therefore

$$\gamma \hat{w}_t \leq |\tilde{w}_{t+1} - \tilde{w}_t| \leq \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| = \eta \|\nabla F(\boldsymbol{w}_t)\| \leq \eta, \tag{16}$$

where the last line uses that F is 1-Lipschitz. So $\hat{w}_t \leq \eta/\gamma$. ■

Now we want to show that oscillations must happen at a certain frequency. To do so, we use the following two lemmas (proofs in Appendix A), which give us rates of progress on the loss for individual data points between oscillations (Lemma 2.4) and a global bound for $|\tilde{w}_t|$ (Lemma 2.5).

Lemma 2.4 *For every t with $1 \leq t < \tau$ and $i \in D_t^-$, if $\eta \geq \eta_0$ then*

$$a_{t+1}^i - a_t^i \geq \max\left(\frac{\eta\|\boldsymbol{x}_i\|^2}{2n(\exp(a_t^i)+1)}, \frac{1}{2}\eta\gamma^2 G(\boldsymbol{w}_t)\right) \tag{17}$$

Lemma 2.5 *If $\eta \geq \eta_0$, then $|\tilde{w}_t| \leq \eta$ for all $t < \tau$.*

Lemma 2.4 tells us that between oscillations, if a data point $\boldsymbol{x}_i$ has non-negligible loss (i.e. $i \in D_t^-$), then its loss must decrease. This recurrence on a_t^i can be combined with the initial condition $a_t^i \geq -\eta$ implied by Lemma 2.5:

$$a_t^i = \langle \boldsymbol{w}_t, \boldsymbol{x}_i\rangle = \hat{w}_t\langle \boldsymbol{w}_*, \boldsymbol{x}_i\rangle + \tilde{w}_t\langle \boldsymbol{v}_*, \boldsymbol{x}_i\rangle \geq -|\tilde{w}_t|\|\boldsymbol{v}_*\|\|\boldsymbol{x}_i\| \geq -\eta, \tag{18}$$

in order to bound the time s until $\ell(a_s^i) \leq 1/8\eta$ for $i \in D_t^-$. At this point, there are two cases: either all data have low loss, implying $F(\boldsymbol{w}_s) \leq 1/8\eta$ and we have transitioned to stability, or some points with negligible loss at step t experienced an increase in loss at step s, implying an oscillation has occurred. The resulting conclusion is stated formally below and proved in Appendix A.

Lemma 2.6 *Suppose $\eta \geq \eta_0$ and $t < \tau$. Then for some $s \leq t + 1 + 4n/\gamma + 192/\gamma^2$, either $s = \tau$ or an oscillation happens at iteration s.*

The last piece of the puzzle before bounding τ is to handle the steps between oscillations; the following lemma shows that $\hat{w}_t$ will grow exponentially not only on iterations where oscillations occur, but at any iteration before an oscillation (proof in Appendix A).

Lemma 2.7 *For any t, if there exists an oscillation $t_k > t$ and $\eta \geq \eta_0$, then $\hat{w}_{t+1} \geq (1+\gamma^2/2)\hat{w}_t$.*

Finally, we can prove our key lemma: a tight upper bound on the transition time τ.

Lemma 2.8 *If $\eta \geq \eta_0$, then $\tau \leq 2 + 4n/\gamma + 280\log(2/\gamma^2)/\gamma^2$.*

Proof First, we sketch the argument from a high-level. We know from Lemma 2.7 that $\hat{w}_t$ grows exponentially until the last oscillation occurs. This means that no oscillations can occur after the first $\widetilde{\mathcal{O}}(1/\gamma^2)$ iterations, since, if an oscillation were to occur after step $t = \widetilde{\Theta}(1/\gamma^2))$ then $\hat{w}_t \geq \eta/\gamma$ (by repeated applications of Lemma 2.7) at which point Lemma 2.3 implies that no further oscillations can occur. Then by Lemma 2.6, it will be no more than $\widetilde{\mathcal{O}}(n/\gamma + 1/\gamma^2)$ iterations until the stable transition happens. We execute this argument below.

Let $t_0, t_1, \ldots$ be the iterations where oscillations occur. We want to show that this list is finite, and to bound the last iteration t_N where an oscillation happens. For any oscillation t_k, and iteration $t < t_k$ we know from Lemma 2.7 that $\hat{w}_{t+1} \geq (1+\gamma^2/2)\hat{w}_t$, so

$$\hat{w}_{t_k} \geq (1+\gamma^2/2)^{t_k-1}\hat{w}_1 \overset{(i)}{\geq} \frac{1}{2}(1+\gamma^2/2)^{t_k-1}\eta\gamma, \tag{19}$$

where (i) uses Lemma 2.1. We also know from Lemma 2.3 that $\hat{w}_{t_k} \leq \eta/\gamma$, so

$$\frac{1}{2}(1+\gamma^2/2)^{t_k-1}\eta\gamma \leq \frac{\eta}{\gamma}, \tag{20}$$

or

$$t_k \leq 1 + \frac{\log(2/\gamma^2)}{\log\left(1+\gamma^2/2\right)} \overset{(i)}{\leq} 1 + \frac{1+\gamma^2/2}{\gamma^2/2}\log(2/\gamma^2) \leq 1 + \frac{3\log(2/\gamma^2)}{\gamma^2}, \tag{21}$$

where (i) uses $\log(1+z) \geq z/(1+z)$ for all $z \geq 0$. Therefore, there are a finite number N of oscillations, and $t_N \leq 1 + 3\log(2/\gamma^2)/\gamma^2$.

We can now apply Lemma 2.6 with $t = t_N + 1$, which implies that, for some s with $t_N < s \leq t_N + 1 + 4n/\gamma + 192/\gamma^2$, either s is an oscillation or $s = \tau$. However, s cannot be an oscillation, since t_N is the last oscillation and $s > t_N$, so $s = \tau$, and

$$\tau \leq t_N + 1 + \frac{4n}{\gamma} + \frac{192}{\gamma^2} \leq 2 + \frac{4n}{\gamma} + \frac{192 + 3\log(2/\gamma^2)}{\gamma^2} \leq 2 + \frac{4n}{\gamma} + \frac{280\log(2/\gamma^2)}{\gamma^2}. \tag{22}$$

■

To prove Theorem 2.1, it only remains to upper bound $F(\boldsymbol{w}_t)$ for $t > \tau$. Rather than using the split comparator technique of Wu et al. (2024), we follow the analysis of Crawshaw et al. (2025a), which explicitly bounds $\|\nabla^2 F(\boldsymbol{w}_t)\|$ along the trajectory, then uses essentially classical descent arguments for smooth objectives. This allows us to eliminate $\log$ terms from the final rate.

3. Lower Bounding the Stable Transition Time

In this section, we present a lower bound of τ that matches our upper bound up to factors logarithmic in $1/\gamma$, implying that our analysis of GD's stable transition time from Section 2 is tight.

Theorem 3.1 *If $\gamma \leq 1/6, n \geq 2$, and $\eta \geq \eta_1 := \max\left\{n, 32/\gamma^2 \log(3/\gamma)\right\}$, then there exists a dataset satisfying Assumption 1.1 such that the transition time $\tau := \min\{t \geq 0 : F(\boldsymbol{w}_t) \leq 1/8\eta\}$ of GD for Equation (1) satisfies $\tau \geq \Omega\left(n/\gamma + 1/\gamma^2\right)$.*

The lower bound of Zhang et al. (2025a) shows that any first-order optimization algorithm for minimizing the logistic loss requires $\Omega(\min(\log n, 1/\gamma^2))$ iterations to find a linear separator. Theorem 3.1 implies that GD requires $\Omega(n/\gamma)$ iterations for the same task[2], so GD is suboptimal among first-order algorithms by a factor of $n\gamma$ for large n.

A similar suboptimality conclusion was reached by Tyurin (2025) for the perceptron algorithm, although they did not consider the dependence on γ. Specifically, the lower bound of Tyurin shows that the number of steps required by the perceptron algorithm (or GD with $\eta \to \infty$) to find a linear separator is $\Omega(n)$. This is worse than $\mathcal{O}(1/\gamma^2)$ required by adaptive GD to find a separator (Zhang et al., 2025a) when $n \gg 1/\gamma^2$. We show an improved lower bound of $\Omega((1+n\gamma)/\gamma^2)$ for large stepsize GD, implying that GD is suboptimal when $n \gg 1/\gamma$.

Theorem 3.1 applies for sufficiently large stepsizes, and the threshold η_1 matches that of our upper bound up to constant factors. Both hard datasets have $d = 2$, however they can be trivially generalized to any $d \geq 2$ by embedding them into a 2-dimensional subspace of $\mathbb{R}^d$. The conditions $n \geq 2$ and $\gamma \leq 1/6$ are to some extent unavoidable: If $n = 1$, then a sufficiently large η will make

2. Although Theorem 3.1 is stated in terms of finding a point with small loss, the hard dataset in Lemma 3.1 provides a lower bound for finding a linear separator.

the loss arbitrarily small in one GD step. Similarly, if $\gamma \geq 1/\sqrt{2}$ then for every pair of data points $\boldsymbol{x}_i, \boldsymbol{x}_j$,

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle \langle \boldsymbol{x}_j, \boldsymbol{w}_* \rangle + \langle (\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top) \boldsymbol{x}_i, (\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top) \boldsymbol{x}_j \rangle \tag{23}$$

$$\geq \gamma^2 - \left\| (\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top) \boldsymbol{x}_i \right\| \left\| (\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top) \boldsymbol{x}_j \right\| \tag{24}$$

$$\geq \gamma^2 - (1 - \gamma^2) = 2\gamma^2 - 1 \geq 0, \tag{25}$$

and this pairwise alignment among the dataset again implies that the loss can be made arbitrarily small in a single GD step.

Proof Sketch for Theorem 3.1 Here we informally describe the construction of two hard datasets corresponding to the two terms of the lower bound from Theorem 3.1. The two datasets yield $\tau \geq \Omega(n/\gamma)$ and $\tau \geq \Omega(1/\gamma^2)$, respectively, and Theorem 3.1 follows from $\tau \geq \Omega(\max(n/\gamma, 1/\gamma^2)) \geq \Omega(n/\gamma + 1/\gamma^2)$. The complete proof can be found in Appendix B.

Lemma 3.1 *[Time until Classification] Suppose that $\gamma \leq 1/6$, $n \geq 6$, and $\eta \geq \eta_1$. Then there exists some dataset satisfying Assumption 1.1, such that for every $t \leq n/(16\gamma)$, there exists $i \in [n]$ with $\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle < 0$.*

We define the hard dataset as

$$\boldsymbol{x}_i = \begin{cases} (\gamma, -\gamma) & i = 1 \\ (\gamma, \sqrt{1-\gamma^2}) & i \in \{2, \ldots, n\} \end{cases}. \tag{26}$$

Recall that the GD update can be decomposed into contributions in the direction of each $\boldsymbol{x}_i$:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \frac{\eta}{n} \sum_{i=1}^{n} \frac{1}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) + 1} \boldsymbol{x}_i. \tag{27}$$

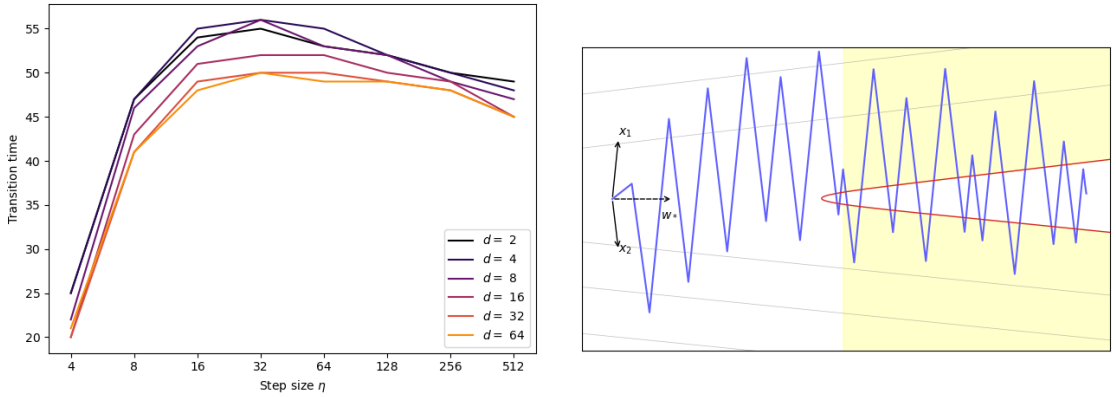
The trajectory of GD on this dataset is easy to imagine: $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are in agreement and have norm 1, so together their contribution to the first update overpowers that of $\boldsymbol{x}_1$ (whose norm is only $\mathcal{O}(\gamma)$). After the first step, all data points except $\boldsymbol{x}_1$ have very low loss, so the gradient is dominated by $\boldsymbol{x}_1$; until $\boldsymbol{x}_1$ is correctly classified, the GD trajectory approximately moves on a line from $\boldsymbol{w}_1$ in the direction of $\boldsymbol{x}_1$. Using this, we get a recurrent upper bound on $\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle$, and we can lower bound the time until $\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle \geq 0$. Note that this construction is a slight generalization of a lower bound for the perceptron algorithm from Tyurin (2025).

Lemma 3.2 *[Time until Stability] Suppose that $\gamma \leq 1/6, n \geq 2$, and $\eta \geq \eta_1$. Then there exists some dataset satisfying Assumption 1.1 such that $F(\boldsymbol{w}_t) > 2/\eta$ for all $t \leq 1 + 1/(59\gamma^2)$.*

Denoting $k = \lceil n/2 \rceil$, we define the hard dataset as

$$\boldsymbol{x}_i = \begin{cases} (\gamma, -\delta) & i \leq k \\ (\gamma, \sqrt{1-\gamma^2}) & i > k \end{cases}, \tag{28}$$

where $\delta \in [0, \sqrt{1-\gamma^2}]$. The idea is to choose δ small enough that after the first step, the loss for $\boldsymbol{x}_i$ is negligibly small for $i > k$, but large enough that the loss for $\boldsymbol{x}_i$ with $i \leq k$ is only slightly



(*a*) Maximum transition time as a function of η. (*b*) A disconnected set of $\boldsymbol{w}$.

Figure 2: **Left:** Numerical evidence suggesting that our results might hold in higher dimensions. Over a random search of datasets, the worst-case transition time τ does not increase past 60 even as the learning rate η increases exponentially. **Right:** An illustration of a property helping our two-dimensional analysis: the set $\{\boldsymbol{w} \in \mathbb{R}^d : \langle \boldsymbol{w}, \boldsymbol{w}_* \rangle \geq \lambda \text{ and } \|(\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top)\boldsymbol{w}\| \geq \gamma \langle \boldsymbol{w}, \boldsymbol{w}_* \rangle / 2\}$ is disconnected when $d = 2$ (shown in yellow), which implies that the GD trajectory must make large jumps to traverse between connected components. This set is connected for $d > 2$.

larger than $2n/k\eta$. This would imply $F(\boldsymbol{w}_1)$ is slightly larger than $2/\eta$; so after one step, the GD trajectory is close to the sublevel set $F(\boldsymbol{w}) \leq 2/\eta$, but the loss (and the gradient norm) are small enough that it takes time to actually enter that sublevel set. Indeed, for $i > k$, $\boldsymbol{x}_i$ contributes negligibly to each update from $t = 1$ to $t = \tau$, so for these steps the updates to $\boldsymbol{w}_t$ are dominated by $\boldsymbol{x}_i$ with $i \leq k$. And, since the loss for $\boldsymbol{x}_i$ with $i \leq k$ is quite small (only slightly larger than $2n/k\eta$), the update size $\eta \|\nabla F(\boldsymbol{w}_t)\|$ can be bounded. It then requires $1/\gamma^2$ iterations until the loss for $\boldsymbol{x}_i$ with $i \leq k$ is smaller than $2n/k\eta$.

4. Possible Extension to Higher Dimensions

The most important limitation of our work is the restriction to two dimensions for the upper bound. In this section, we discuss whether our results could be extended to higher dimensions, and some challenges of generalizing our oscillation-based analysis for $d \geq 2$.

First, we present numerical evidence suggesting that our conclusions when $d = 2$ might also hold for $d > 2$. We can see from Figure 2(*a*) that even as η grows exponentially, the stable transition time τ appears bounded over a random search of many datasets with $d \geq 2$. For a given d, we generate a dataset of $n = d$ samples as follows: for each $i \in [n]$ we set the max-margin component $\langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle$ equal to γ, then sample the orthogonal complement $\tilde{\boldsymbol{x}}_i = (\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top)\boldsymbol{x}_i$ uniformly over a $d - 1$ dimensional ball centered at $\boldsymbol{0}$ with radius $\sqrt{1 - \gamma^2}$. This procedure enforces separability with a margin of γ and $\|\boldsymbol{x}_i\| \leq 1$. For every $d \in \{2^1, \ldots, 2^7\}$ and every $\eta \in \{2^2, \ldots, 2^9\}$, we generate 4096 datasets, and compute the maximum time (over datasets) that it takes to achieve $F(\boldsymbol{w}_t) \leq 2/\eta$. In Figure 2(*a*), for each d we plot the worst-case τ as a function of η. While this random search is certainly not exhaustive (especially in very high dimensions), these results suggest that our conclusion that τ is bounded independently of η might extend to higher dimensions.

Indeed, some parts of our analysis (e.g. Lemma 2.1) do not rely on low dimension. However, there are some difficulties in characterizing the oscillatory behavior of GD's trajectory when the orthogonal component to $\boldsymbol{w}_*$ is not a scalar, which we discuss below. For datasets with $d \geq 2$, we denote the orthogonal component as $\tilde{\boldsymbol{w}}_t = (\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top)\boldsymbol{w}_t$. We keep the definition $\hat{w}_t = \langle \boldsymbol{w}_t, \boldsymbol{w}_* \rangle$.

First, we need a definition of oscillation that does not rely on d. One possibility is to replace the condition $\tilde{w}_{t+1}\tilde{w}_t < 0$ with $\langle \tilde{\boldsymbol{w}}_{t+1}, \tilde{\boldsymbol{w}}_t \rangle < 0$. With this generalization, many of our key lemmas extend for $d \geq 2$, such as Lemma 2.2, which says $\hat{w}_t$ increases by a factor of $1 + \Omega(\gamma^2)$ when an oscillation happens, and Lemma 2.3, which says oscillations can only happen when $\hat{w}_t \leq \eta/\gamma$.

The main obstacle is extending Lemma 2.6, that is, to show that $\langle \tilde{\boldsymbol{w}}_{t+1}, \tilde{\boldsymbol{w}}_t \rangle < 0$ must happen frequently along GD trajectories that avoid the sublevel set where $F(\boldsymbol{w}) \leq 2/\eta$. In our two-dimensional analysis, we showed this by implicitly leveraging a nice topological property: for $1 \leq t < \tau$, we know that $\hat{w}_t \geq \lambda$ and $|\tilde{w}_t| \geq \frac{1}{2}\gamma \hat{w}_t$, and the set of such $\boldsymbol{w}$ is *disconnected* when $d = 2$ (see Figure 2(*b*)). By showing that the trajectory cannot stay in one connected component for too long, we know that the trajectory must "jump" between components, which we call an oscillation. However, in higher dimensions, the set of $\boldsymbol{w}$ satisfying $\langle \boldsymbol{w}, \boldsymbol{w}_* \rangle \geq \lambda$ and $\|(\boldsymbol{I} - \boldsymbol{w}_* \boldsymbol{w}_*^\top)\boldsymbol{w}\| \geq \frac{1}{2}\gamma \langle \boldsymbol{w}, \boldsymbol{w}_* \rangle$ is *connected* (it is a half-space minus a cone), so it is a priori possible for the trajectory to avoid the sublevel set without making any large jumps. Due to this difficulty, it is unclear whether our oscillation-based analysis can directly generalize for $d > 2$, but based on the numerical evidence in Figure 2(*a*), we conjecture that $\tau \leq \widetilde{\mathcal{O}}(n/\gamma + 1/\gamma^2)$ still holds for any $d \geq 2$. We leave this problem of tightly characterizing the transition time in general dimension open for future work.

Acknowledgments

We thank Blake Woodworth for many helpful conversations in the early stages of this project. Michael Crawshaw is supported by the Doctoral Research Scholarship of George Mason University. Mingrui Liu is supported by NSF grants #2436217, #2425687.

References

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36:19540–19569, 2023.

Jason Altschuler. *Greed, hedging, and acceleration in convex optimization*. PhD thesis, Massachusetts Institute of Technology, 2018.

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, 213(1–2):1105–1118, November 2024. ISSN 1436-4646. doi: 10.1007/s10107-024-02164-2. URL http://dx.doi.org/10.1007/s10107-024-02164-2.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):231–357, 2015.

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

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.

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*, 2021.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Michael Crawshaw, Blake Woodworth, and Mingrui Liu. Constant stepsize local gd for logistic regression: Acceleration by instability. In *Forty-second International Conference on Machine Learning*, 2025a.

Michael Crawshaw, Blake Woodworth, and Mingrui Liu. Local steps speed up local gd for heterogeneous distributed logistic regression. In *The Thirteenth International Conference on Learning Representations*, 2025b.

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*, 2023.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s) gd over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. *Advances in Neural Information Processing Systems*, 36:29406–29448, 2023.

Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34(3):2588–2608, 2024.

Benjamin Grimmer, Kevin Shu, and Alex L Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *INFORMS Journal on Optimization*, 7(2):156–169, 2025.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pages 1772–1798. PMLR, 2019.

Zoran Kadelburg, Dusan Dukic, Milivoje Lukic, and Ivan Matic. Inequalities of karamata, schur and muirhead, and some applications. *The Teaching of Mathematics*, 8(1):31–45, 2005.

Guy Kornowski and Ohad Shamir. The oracle complexity of simplex-based matrix games: Linear separability and nash equilibria. *arXiv preprint arXiv:2412.06990*, 2024.

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

Si Yi Meng, Antonio Orvieto, Daniel Yiming Cao, and Christopher De Sa. Gradient descent on logistic regression with non-separable data and large step sizes. *arXiv preprint arXiv:2406.05033*, 2024.

Si Yi Meng, Baptiste Goujaud, Antonio Orvieto, and Christopher De Sa. Gradient descent on logistic regression: Do large step-sizes work with data on the sphere? *arXiv preprint arXiv:2507.11228*, 2025.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

Alexander Tyurin. From logistic regression to the perceptron algorithm: Exploring gradient descent with large step sizes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20938–20946, 2025.

Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36: 74229–74256, 2023.

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

Jingfeng Wu, Pierre Marion, and Peter Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. *arXiv preprint arXiv:2506.02336*, 2025.

Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. Gradient descent converges arbitrarily fast for logistic regression via large and adaptive stepsizes. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=cufJbug7du.

Zihan Zhang, Jason Lee, Simon Du, and Yuxin Chen. Anytime acceleration of gradient descent. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 5991–6013. PMLR,

30 Jun–04 Jul 2025b. URL https://proceedings.mlr.press/v291/zhang25a.html.

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*, 2023. URL https://openreview.net/forum?id=p7EagBsMAEO.

Appendix A. Deferred Proofs from Section 2

Lemma 2.1 *$\hat{w}_t$ is strictly increasing. Also, if $\eta \geq \eta_0$, then $\hat{w}_t \geq \gamma\eta/2 + \gamma(t-1)/16$ for all $1 \leq t \leq \tau$. In particular, $\hat{w}_t \geq 8\tilde{\lambda}$ for all $t \geq 1$.*

Proof For any $t \geq 0$,

$$\begin{aligned}
\hat{w}_{t+1} - \hat{w}_t &= \langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \boldsymbol{w}_* \rangle && (29)\\
&= -\eta \langle \nabla F(\boldsymbol{w}_t), \boldsymbol{w}_* \rangle && (30)\\
&= \frac{\eta}{n} \sum_{i=1}^{n} \frac{1}{\exp(a_t^i) + 1} \langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle && (31)\\
&\geq \frac{\eta\gamma}{n} \sum_{i=1}^{n} \frac{1}{\exp(a_t^i) + 1} && (32)\\
&> 0, && (33)
\end{aligned}$$

so $\hat{w}_t$ is increasing.

For $t = 1$, we can bound $\hat{w}_t$ directly as

$$\begin{aligned}
\hat{w}_1 &= \langle \boldsymbol{w}_1, \boldsymbol{w}_* \rangle && (34)\\
&= \eta \langle -\nabla F(\boldsymbol{w}_0), \boldsymbol{w}_* \rangle && (35)\\
&= \frac{\eta}{n} \sum_{i=1}^{n} \frac{\langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle}{\exp(a_0^i) + 1} && (36)\\
&= \frac{\eta}{2n} \sum_{i=1}^{n} \langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle && (37)\\
&\geq \frac{\eta\gamma}{2}. && (38)
\end{aligned}$$

For $1 < t < \tau$, we know $F(\boldsymbol{w}_t) \geq 1/8\eta$. Therefore, by Equation (32) and Lemma C.1 (together with the condition $\eta \geq \eta_0 \geq n$),

$$\hat{w}_{t+1} - \hat{w}_t \geq \eta\gamma G(\boldsymbol{w}_t) \geq \eta\gamma \frac{1}{16\eta} = \frac{\gamma}{16}. \tag{39}$$

Unrolling back to $t = 1$ yields the result.

To get $\hat{w}_t \geq 8\tilde{\lambda}$, it suffices that $\eta\gamma/2 \geq 8\tilde{\lambda}$, since $\hat{w}_t \geq \hat{w}_1 \geq \eta\gamma/2$. So we want

$$\frac{8\log(8\eta)}{\gamma} \leq \frac{1}{2}\eta\gamma, \tag{40}$$

or equivalently

$$\log(8\eta) \leq \frac{\eta\gamma^2}{16}. \tag{41}$$

By concavity of $\log$,

$$\log(8\eta) \leq \log\left(\frac{256}{\gamma^2}\right) + \frac{\gamma^2}{256}\left(8\eta - \frac{256}{\gamma^2}\right) \leq \log\left(\frac{256}{\gamma^2}\right) + \frac{\eta\gamma^2}{32} \leq \frac{\eta\gamma^2}{16}, \tag{42}$$

where the last inequality uses $\eta \geq \eta_0 \geq (32/\gamma^2)\log(128/\gamma^2)$. ■

Lemma A.1 *For $t \geq 1$, if $\eta \geq \eta_0$ and*

$$|\tilde{w}_t| \leq \frac{1}{2}\gamma\hat{w}_t, \tag{43}$$

then $F(\boldsymbol{w}_t) \leq 1/8\eta$.

Proof Recall that $\boldsymbol{w}_t = \hat{w}_t\boldsymbol{w}_* + \tilde{w}_t\boldsymbol{v}_*$. So for each $i \in [n]$,

$$\begin{aligned}
a_t^i &= \langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle && (44)\\
&= \hat{w}_t\langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle + \tilde{w}_t\langle \boldsymbol{v}_*, \boldsymbol{x}_i \rangle && (45)\\
&\geq \gamma\hat{w}_t - |\tilde{w}_t|\|\boldsymbol{x}_i\| && (46)\\
&\overset{(i)}{\geq} \gamma\hat{w}_t - \frac{1}{2}\gamma\hat{w}_t && (47)\\
&= \frac{1}{2}\gamma\hat{w}_t && (48)\\
&\overset{(ii)}{\geq} \gamma\tilde{\lambda} && (49)\\
&\geq \gamma\lambda, && (50)
\end{aligned}$$

where (i) uses $\|\boldsymbol{x}_i\| \leq 1$ and $|\tilde{w}_t| \leq \gamma\hat{w}_t/2$ and (ii) uses $\hat{w}_t \geq 8\tilde{\lambda}$ from Lemma 2.1. Therefore

$$F(\boldsymbol{w}_t) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-a_t^i)) \leq \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-\gamma\lambda)) = 1/8\eta, \tag{51}$$

where the last equality follows from the definition of λ. ■

Lemma A.2 *Define $B_t = \{i \in [n] : \tilde{x}_i\tilde{w}_t \leq -\frac{1}{2}\gamma\hat{w}_t\}$. If $\eta \geq \eta_0$, then for $t < \tau$,*

$$\left(1 - \frac{16}{\eta^3}\right) G(\boldsymbol{w}_t) \leq \frac{1}{n}\sum_{i\in B_t}\frac{1}{\exp(a_t^i)+1} \leq G(\boldsymbol{w}_t). \tag{52}$$

Proof The second desired inequality

$$\frac{1}{n}\sum_{i\in B_t}\frac{1}{\exp(a_t^i)+1} \leq G(\boldsymbol{w}_t) \tag{53}$$

is obvious from the definition of $G(\boldsymbol{w}_t)$, so we focus on the first desired inequality.

For $i \notin B_t$,

$$\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle = \hat{w}_t\langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle + \langle \tilde{\boldsymbol{w}}_t, \tilde{\boldsymbol{x}}_i \rangle \geq \gamma\hat{w}_t - \frac{1}{2}\gamma\hat{w}_t = \frac{1}{2}\gamma\hat{w}_t \geq \gamma\lambda, \tag{54}$$

(where the last inequality uses $\hat{w}_t \geq 8\tilde{\lambda}$ from Lemma 2.1), so

$$\ell(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) \leq \ell(\gamma\lambda) = 1/8\eta, \tag{55}$$

and therefore

$$\frac{1}{n-|B_t|}\sum_{i\notin B_t}\ell(\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)\le 1/8\eta. \tag{56}$$

Combining with the fact that $F(\boldsymbol{w}_t)>1/8\eta$ (since $t<\tau$), this means B_t is not empty. Essentially, B_t consists of the data points that have non-negligible loss. Now, note that for $i\notin B_t$,

$$\frac{1}{\exp(\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)+1}\le \exp(-\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)\overset{(i)}{\le}\exp\left(-\frac{1}{2}\gamma\hat{w}_t\right) \tag{57}$$

$$\overset{(ii)}{\le}\exp(-4\gamma\tilde{\lambda})=1/\eta^4\overset{(iii)}{\le}16G(\boldsymbol{w}_t)/\eta^3, \tag{58}$$

where (i) uses Equation (54), (ii) uses $\hat{w}_t\ge 8\tilde{\lambda}$ from Lemma 2.1, and (iii) uses $F(\boldsymbol{w}_t)\ge 1/8\eta \implies G(\boldsymbol{w}_t)\ge 1/(16\eta)$ from Lemma C.1. Therefore

$$G(\boldsymbol{w}_t)=\frac{1}{n}\sum_{i\in B_t}\frac{1}{\exp(\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)+1}+\frac{1}{n}\sum_{i\notin B_t}\frac{1}{\exp(\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)+1} \tag{59}$$

$$G(\boldsymbol{w}_t)\le\frac{1}{n}\sum_{i\in B_t}\frac{1}{\exp(\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)+1}+\frac{n-|B_t|}{n}\frac{16G(\boldsymbol{w}_t)}{\eta^3} \tag{60}$$

$$G(\boldsymbol{w}_t)\le\frac{1}{n}\sum_{i\in B_t}\frac{1}{\exp(\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)+1}+\frac{16G(\boldsymbol{w}_t)}{\eta^3} \tag{61}$$

$$\left(1-\frac{16}{\eta^3}\right)G(\boldsymbol{w}_t)\le\frac{1}{n}\sum_{i\in B_t}\frac{1}{\exp(\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle)+1} \tag{62}$$

which proves the first inequality in Equation (52). ■

Lemma 2.4 *For every t with $1\le t<\tau$ and $i\in D_t^-$, if $\eta\ge\eta_0$ then*

$$a_{t+1}^i-a_t^i\ge\max\left(\frac{\eta\|\boldsymbol{x}_i\|^2}{2n(\exp(a_t^i)+1)},\frac{1}{2}\eta\gamma^2G(\boldsymbol{w}_t)\right) \tag{17}$$

Proof The idea is that the data points in D_t^+ all have low loss, so they don't contribute much to each gradient update, while the contributions of each point in D_t^- are all "aligned" in that each pair has positive dot product.

To make this concrete, denote $m=\operatorname{argmin}_{i\in[n]}a_t^i$, and notice that $m\in D_t^-$, since for every $i\in D_t^+$:

$$a_t^i=\langle \boldsymbol{w}_t,\boldsymbol{x}_i\rangle=\hat{w}_t\langle \boldsymbol{w}_*,\boldsymbol{x}_i\rangle+\tilde{w}_t\langle \boldsymbol{v}_*,\boldsymbol{x}_i\rangle\overset{(i)}{\ge}\hat{w}_t\langle \boldsymbol{w}_*,\boldsymbol{x}_i\rangle\ge\gamma\hat{w}_t\ge\gamma\tilde{\lambda}, \tag{63}$$

(where (i) uses $\tilde{w}_t\langle \boldsymbol{v}_*,\boldsymbol{x}_i\rangle\ge 0$ from the definition of D_t^+) and we know that $\min_{i\in[n]}a_t^i<\gamma\tilde{\lambda}$, since otherwise $F(\boldsymbol{w}_t)\le\ell(\gamma\tilde{\lambda})\le 1/8\eta$. For each $i\in D_t^-$, we can now bound the change in a_t^i as

follows:

$$a_{t+1}^i - a_t^i = \langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \boldsymbol{x}_i \rangle \tag{64}$$

$$= \frac{\eta}{n} \sum_{j=1}^{n} \frac{\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle}{\exp(a_t^j) + 1} \tag{65}$$

$$= \eta \Bigg(\underbrace{\frac{1}{n} \sum_{j \in D_t^-} \frac{\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle}{\exp(a_t^j) + 1}}_{A_1} + \underbrace{\frac{1}{n} \sum_{j \in D_t^+} \frac{\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle}{\exp(a_t^j) + 1}}_{A_2} \Bigg). \tag{66}$$

We will show that A_1 dominates A_2. Note that

$$A_1 \overset{(i)}{\geq} \frac{\gamma^2}{n} \sum_{j \in D_t^-} \frac{1}{\exp(a_t^j) + 1} \overset{(ii)}{\geq} \frac{\gamma^2}{n} \sum_{j \in B_t} \frac{1}{\exp(a_t^j) + 1} \overset{(iii)}{\geq} \gamma^2 \left(1 - \frac{16}{\eta^3}\right) G(\boldsymbol{w}_t) \overset{(iv)}{\geq} \frac{9}{10} \gamma^2 G(\boldsymbol{w}_t), \tag{67}$$

where (i) uses that for $i, j \in D_t^-$,

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle \langle \boldsymbol{x}_j, \boldsymbol{w}_* \rangle + \langle \boldsymbol{x}_i, \boldsymbol{v}_* \rangle \langle \boldsymbol{x}_j, \boldsymbol{v}_* \rangle \geq \langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle \langle \boldsymbol{x}_j, \boldsymbol{w}_* \rangle \geq \gamma^2, \tag{68}$$

(ii) uses that $B_t \subset D_t^-$, (iii) uses Lemma A.2, and (iv) uses $\eta \geq \eta_0 \geq 32 \log(256)$. For A_2,

$$|A_2| \leq \frac{1}{n} \sum_{j \in D_t^+} \frac{1}{\exp(a_t^j) + 1} \overset{(i)}{\leq} \frac{1}{n} \sum_{j \notin B_t} \frac{1}{\exp(a_t^j) + 1} \overset{(ii)}{\leq} \frac{16}{\eta^3} G(\boldsymbol{w}_t), \tag{69}$$

where (i) uses $B_t \subset D_t^- \implies D_t^+ \subset [n] \backslash B_t$ and (ii) uses Lemma A.2. Since $\eta \geq \eta_0 \geq 32 \log(256)/\gamma^2 \geq 128/\gamma^2$, this means

$$|A_2| \leq \frac{1}{2^{17}} \gamma^2 G(\boldsymbol{w}_t) \leq \frac{1}{2^{16}} A_1, \tag{70}$$

so

$$a_{t+1}^i - a_t^i = \eta(A_1 + A_2) \geq \frac{2}{3} \eta A_1 = \frac{2\eta}{3n} \sum_{j \in D_t^-} \frac{\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle}{\exp(a_t^j) + 1}. \tag{71}$$

Now we can derive the two desired bounds. First, recall that $i \in D_t^-$, so

$$a_{t+1}^i - a_t^i \geq \frac{2\eta}{3n} \frac{\|\boldsymbol{x}_i\|^2}{\exp(a_t^i) + 1} \geq \frac{\eta \|\boldsymbol{x}_i\|^2}{2n(\exp(a_t^i) + 1)}. \tag{72}$$

Second,

$$a_{t+1}^i - a_t^i \geq \frac{2\eta}{3n} \sum_{j \in B_t} \frac{\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle}{\exp(a_t^j) + 1} \overset{(i)}{\geq} \frac{2\eta\gamma^2}{3n} \sum_{j \in B_t} \frac{1}{\exp(a_t^j) + 1} \tag{73}$$

$$\overset{(ii)}{\geq} \frac{2}{3} \left(1 - \frac{16}{\eta^3}\right) \eta \gamma^2 G(\boldsymbol{w}_t) \overset{(iii)}{\geq} \frac{1}{2} \eta \gamma^2 G(\boldsymbol{w}_t). \tag{74}$$

where (i) uses $\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle \geq \gamma^2$ as in Equation (68), (ii) uses Lemma A.2, and (iii) uses $\eta \geq \eta_0 \geq 32 \log(256)$. ■

Lemma 2.5 *If $\eta \geq \eta_0$, then $|\tilde{w}_t| \leq \eta$ for all $t < \tau$.*

Proof Since $\boldsymbol{w}_0 = \mathbf{0}$, it clearly holds for $t = 0$, and it also holds for $t = 1$ since

$$|\tilde{w}_1| \leq \|\boldsymbol{w}_1\| = \|\boldsymbol{w}_0 - \eta \nabla F(\boldsymbol{w}_0)\| = \eta \|\nabla F(\mathbf{0})\| \leq \eta, \tag{75}$$

using that F is 1-Lipschitz.

Now suppose $|\tilde{w}_t| \leq \eta$ for some $t \geq 1$. We consider two cases. If an oscillation happens at step t, then

$$|\tilde{w}_{t+1}| \leq |\tilde{w}_{t+1}| + |\tilde{w}_t| \overset{(i)}{=} |\tilde{w}_{t+1} - \tilde{w}_t| = \eta |\langle \nabla F(\boldsymbol{w}_t), \boldsymbol{v}_* \rangle| \leq \eta \|\nabla F(\boldsymbol{w}_t)\| \overset{(ii)}{\leq} \eta, \tag{76}$$

where (i) uses that $\tilde{w}_{t+1}$ and $\tilde{w}_t$ have opposite sign, and (ii) again uses that F is 1-Lipschitz.

Now suppose an oscillation does not happen at step t, so that $\tilde{w}_{t+1}$ and $\tilde{w}_t$ have the same sign. First, we claim that $\tilde{w}_{t+1} - \tilde{w}_t$ has opposite sign from $\tilde{w}_t$. To see why, recall the definition of $B_t = \{i \in [n] : \tilde{x}_i \tilde{w}_t \leq -\frac{1}{2}\gamma \hat{w}_t\}$ from Lemma A.2, and notice

$$\tilde{w}_t(\tilde{w}_{t+1} - \tilde{w}_t) = \frac{\eta}{n} \sum_{i=1}^{n} \frac{\tilde{w}_t \tilde{x}_i}{\exp(a_t^i) + 1} \tag{77}$$

$$= \eta \Bigg(\underbrace{\frac{1}{n} \sum_{i \in B_t} \frac{\tilde{w}_t \tilde{x}_i}{\exp(a_t^i) + 1}}_{A_1} + \underbrace{\frac{1}{n} \sum_{i \notin B_t} \frac{\tilde{w}_t \tilde{x}_i}{\exp(a_t^i) + 1}}_{A_2} \Bigg). \tag{78}$$

We can bound A_1 as

$$A_1 = \frac{1}{n} \sum_{i \in B_t} \frac{\tilde{w}_t \tilde{x}_i}{\exp(a_t^i) + 1} \leq -\frac{\gamma \hat{w}_t}{2n} \sum_{i \in B_t} \frac{1}{\exp(a_t^i) + 1} \overset{(i)}{\leq} -\frac{1}{2} \gamma \hat{w}_t \left(1 - \frac{16}{\eta^3}\right) G(\boldsymbol{w}_t) \tag{79}$$

$$\overset{(ii)}{\leq} -\frac{1}{4} \eta \gamma^2 \left(1 - \frac{16}{\eta^3}\right) G(\boldsymbol{w}_t) \overset{(iii)}{\leq} -38 G(\boldsymbol{w}_t), \tag{80}$$

where (i) uses Lemma A.2, (ii) uses $\hat{w}_t \geq \hat{w}_1 \geq \eta\gamma/2$ from Lemma 2.1, and (iii) uses $\eta \geq \eta_0 \geq 32 \log(256)/\gamma^2$. Similarly, we can bound A_2 as

$$A_2 = \frac{1}{n} \sum_{i \notin B_t} \frac{\tilde{w}_t \tilde{x}_i}{\exp(a_t^i) + 1} \overset{(i)}{\leq} \frac{\eta}{n} \sum_{i \notin B_t} \frac{1}{\exp(a_t^i) + 1} \overset{(ii)}{\leq} \frac{16}{\eta^2} G(\boldsymbol{w}_t) \overset{(iii)}{\leq} G(\boldsymbol{w}_t), \tag{81}$$

where (i) uses $\tilde{w}_t \tilde{x}_i \leq |\tilde{w}_t| \leq \eta$, (ii) uses Lemma A.2, and (iii) uses $\eta \geq \eta_0 \geq 16$. So

$$\tilde{w}_t(\tilde{w}_{t+1} - \tilde{w}_t) = \eta(A_1 + A_2) < 0, \tag{82}$$

which proves the claim. So $\text{sign}(\tilde{w}_t) = \text{sign}(\tilde{w}_{t+1}) = -\text{sign}(\tilde{w}_{t+1} - \tilde{w}_t)$, and

$$|\tilde{w}_{t+1} - (\tilde{w}_{t+1} - \tilde{w}_t)| = |\tilde{w}_{t+1}| + |\tilde{w}_{t+1} - \tilde{w}_t| \tag{83}$$

$$|\tilde{w}_t| = |\tilde{w}_{t+1}| + |\tilde{w}_{t+1} - \tilde{w}_t| \tag{84}$$

$$|\tilde{w}_{t+1}| = |\tilde{w}_t| - |\tilde{w}_{t+1} - \tilde{w}_t| \tag{85}$$

$$|\tilde{w}_{t+1}| \leq |\tilde{w}_t| \leq \eta. \tag{86}$$

■

Lemma 2.6 *Suppose $\eta \geq \eta_0$ and $t < \tau$. Then for some $s \leq t + 1 + 4n/\gamma + 192/\gamma^2$, either $s = \tau$ or an oscillation happens at iteration s.*

Proof The idea is to analyze a_t^i in two phases: first we show that $a_s^i \geq 0$ for all i after $\mathcal{O}(n/\gamma)$ iterations, then we derive a recurrence relation on $G(\boldsymbol{w}_s)$ showing that $G(\boldsymbol{w}_s) \leq 1/\eta$ after $1/\gamma^2$ iterations, unless an oscillation happens first.

Let t_{osc} be the first iteration after t such that either an oscillation happens at step t_{osc} or $\tau = t_{\text{osc}}$. From the definition of an oscillation, we know that $\langle \boldsymbol{w}_s, \boldsymbol{v}_* \rangle$ does not change sign for $s \in \{t, \dots, t_{\text{osc}}\}$, so $D_t^+ = D_{t+1}^+ = \dots = D_{t_{\text{osc}}}^+$ and $D_t^- = D_{t+1}^- = \dots = D_{t_{\text{osc}}}^-$. Accordingly, we denote $D_+ = D_t^+$ and $D_- = D_t^-$.

First, we bound the number of steps until $a_s^i \geq 0$ for every i. Note that $a_s^i \geq 0$ for every $i \in D_+$, since

$$a_s^i = \langle \boldsymbol{w}_s, \boldsymbol{x}_i \rangle = \hat{w}_s \langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle + \tilde{w}_s \langle \boldsymbol{v}_*, \boldsymbol{x}_i \rangle \overset{(i)}{\geq} \gamma \hat{w}_s \overset{(ii)}{\geq} 0, \tag{87}$$

where (i) uses $\tilde{w}_s \langle \boldsymbol{v}_*, \boldsymbol{x}_i \rangle \geq 0$ from the definition of D_s^+ and (ii) uses $\hat{w}_s \geq \hat{w}_0 = 0$ from Lemma 2.1. So we only need to worry about a_s^i for $i \in D_-$.

Suppose for some $i \in D_-$ that $a_t^i < 0$, and let $s \geq t$ such that $a_r^i < 0$ for all r with $t \leq r \leq s$. Then according to Lemma 2.4,

$$a_{s+1}^i - a_s^i \geq \frac{\eta \|\boldsymbol{x}_i\|^2}{2n(\exp(a_s^i) + 1)} \geq \frac{\eta \|\boldsymbol{x}_i\|^2}{4n}, \tag{88}$$

so that

$$a_s^i - a_t^i \geq \frac{\eta \|\boldsymbol{x}_i\|^2 (s - t)}{4n}. \tag{89}$$

This means that $a_s^i \geq 0$ when

$$s - t \geq \frac{4n \max(0, -a_t^i)}{\eta \|\boldsymbol{x}_i\|^2}. \tag{90}$$

Note that

$$a_t^i = \langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle = \hat{w}_t \langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle + \tilde{w}_t \langle \boldsymbol{v}_*, \boldsymbol{x}_i \rangle \geq -|\tilde{w}_t| \|\boldsymbol{v}_*\| \|\boldsymbol{x}_i\| \geq -\eta \|\boldsymbol{x}_i\|, \tag{91}$$

where the last inequality uses $|\tilde{w}_t| \leq \eta$ from Lemma 2.5 and $\|\boldsymbol{v}_*\| = 1$. Therefore $a_s^i \geq 0$ when $s - t \geq 4n/\|\boldsymbol{x}_i\|$. Since $\|\boldsymbol{x}_i\| \geq \langle \boldsymbol{x}_i, \boldsymbol{w}_* \rangle \geq \gamma$, it suffices that $s - t \geq 4n/\gamma$.

Let t_0 be the smallest $s \geq t$ such that $a_s^i \geq 0$ for all i, so that by the above $t_0 \leq t + 4n/\gamma$. Note also that $a_s^i \geq 0$ for all $s \in \{t_0, \dots, t_{\text{osc}} - 1\}$, since Equation (87) holds for all such s, i and a_s^i is increasing for $i \in D_-$ by Lemma 2.4.

Now we consider $G(\boldsymbol{w}_s)$ for the second phase. For all $s \geq t_0$,

$$G(\boldsymbol{w}_{s+1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\exp(a_{s+1}^i)+1} \tag{92}$$

$$\overset{(i)}{\leq} \frac{\eta^3}{\eta^3-16}\frac{1}{n}\sum_{i\in D_-} \frac{1}{\exp(a_{s+1}^i)+1} \tag{93}$$

$$\leq \frac{\eta^3}{\eta^3-16}\frac{1}{n}\sum_{i\in D_-} \exp(-a_{s+1}^i) \tag{94}$$

$$\overset{(ii)}{\leq} \exp\left(-\frac{1}{2}\eta\gamma^2 G(\boldsymbol{w}_s)\right)\frac{\eta^3}{\eta^3-16}\frac{1}{n}\sum_{i\in D_-} \exp(-a_s^i) \tag{95}$$

$$\overset{(iii)}{\leq} \exp\left(-\frac{1}{2}\eta\gamma^2 G(\boldsymbol{w}_s)\right)\frac{2\eta^3}{\eta^3-16}\frac{1}{n}\sum_{i\in D_-} \frac{1}{\exp(a_s^i)+1} \tag{96}$$

$$\overset{(iv)}{\leq} \frac{129}{64}\exp\left(-\frac{1}{2}\eta\gamma^2 G(\boldsymbol{w}_s)\right)\frac{1}{n}\sum_{i\in D_-} \frac{1}{\exp(a_s^i)+1} \tag{97}$$

$$\leq \frac{129}{64}\exp\left(-\frac{1}{2}\eta\gamma^2 G(\boldsymbol{w}_s)\right) G(\boldsymbol{w}_s), \tag{98}$$

where (i) uses Lemma A.2 together with $B_{s+1} \subset D_s^- = D_-$, (ii) uses Lemma 2.4, (iii) uses $a_s^i \geq 0$, and (iv) uses $\eta \geq \eta_0 \geq 32\log(256)$. We can first bound $G(\boldsymbol{w}_{t_0+1})$ as follows. Denoting $b = \frac{1}{2}\eta\gamma^2$ and $T(z) = 129/64\exp(-bz)z$, we have $G(\boldsymbol{w}_{s+1}) \leq T(G(\boldsymbol{w}_s))$. T has a global maximum at $z = 1/b$, so that

$$G(\boldsymbol{w}_{t_0+1}) \leq T(1/b) = \frac{129}{64}\exp(-1)/b = \frac{129}{32e\eta\gamma^2} \leq \frac{3}{2\eta\gamma^2}. \tag{99}$$

Knowing this, we perform a similar, slightly stronger derivation as in Equations 92 through 98. For $s \geq t_0 + 1$,

$$G(\boldsymbol{w}_{s+1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\exp(a_{s+1}^i)+1} \tag{100}$$

$$\overset{(i)}{\leq} \left(1+\frac{16}{\eta^3-16}\right)\frac{1}{n}\sum_{i\in B_{s+1}} \frac{1}{\exp(a_{s+1}^i)+1} \tag{101}$$

$$\overset{(ii)}{\leq} \left(1+\frac{16}{\eta^3-16}\right)\frac{1}{n}\sum_{i\in B_{s+1}} \frac{1}{\exp(a_s^i)\exp(\eta\gamma^2 G(\boldsymbol{w}_s)/2)+1} \tag{102}$$

$$\overset{(iii)}{\leq} \left(1+\frac{16}{\eta^3-16}\right)\exp\left(-\frac{1}{8}\eta\gamma^2 G(\boldsymbol{w}_s)\right)\frac{1}{n}\sum_{i\in B_{s+1}} \frac{1}{\exp(a_s^i)+1} \tag{103}$$

$$\leq \left(1+\frac{16}{\eta^3-16}\right)\exp\left(-\frac{1}{8}\eta\gamma^2 G(\boldsymbol{w}_s)\right) G(\boldsymbol{w}_s), \tag{104}$$

where (i) uses Lemma A.2, (ii) uses Lemma 2.4 together with $B_{s+1} \subset D_s^- = D_-$, and (iii) can be justified as follows: denoting $C = \exp(\eta\gamma^2 G(\boldsymbol{w}_s)/2)$, we want to show that

$$C\exp(a_s^i) + 1 \geq C^{1/4}(\exp(a_s^i) + 1), \tag{105}$$

or equivalently,

$$C^{1/4} \leq \frac{C\exp(a_s^i) + 1}{\exp(a_s^i) + 1} \tag{106}$$

$$(\iff) \quad \frac{1}{4}\log C \leq \log\left(\frac{C\exp(a_s^i) + 1}{\exp(a_s^i) + 1}\right) \tag{107}$$

$$(\iff) \quad \frac{1}{4} \leq \log\left(1 + \frac{(C-1)\exp(a_s^i)}{\exp(a_s^i) + 1}\right) / \log C \tag{108}$$

$$(\impliedby) \quad \frac{1}{4} \overset{(iv)}{\leq} \log\left(1 + \frac{(C-1)\exp(a_s^i)}{\exp(a_s^i) + 1}\right) / (C-1) \tag{109}$$

$$(\impliedby) \quad \frac{1}{4} \overset{(v)}{\leq} \frac{(C-1)\exp(a_s^i)}{\exp(a_s^i) + 1} \frac{\exp(a_s^i) + 1}{C\exp(a_s^i) + 1} \frac{1}{C-1} \tag{110}$$

$$(\iff) \quad \frac{1}{4} \leq \frac{1}{C + \exp(-a_s^i)} \tag{111}$$

$$(\iff) \quad 4 \geq C + \exp(-a_s^i), \tag{112}$$

where (iv) uses $\log(1+z) \leq z$ and (v) uses $\log(1+z) \geq z/(1+z)$. The final condition is satisfied since $a_s^i \geq 0$ and $G(\boldsymbol{w}_s) \leq 3/(2\eta\gamma^2)$ from Equation (99) implies $C \leq \exp(3/4) \leq 3$. This justifies (iii).

From Equation (104),

$$G(\boldsymbol{w}_{s+1}) \leq \left(1 + \frac{16}{\eta^3 - 16}\right) \exp\left(-\frac{1}{8}\eta\gamma^2 G(\boldsymbol{w}_s)\right) G(\boldsymbol{w}_s) \tag{113}$$

$$\overset{(i)}{\leq} \left(1 + \frac{16}{\eta^3 - 16}\right) \left(1 - \frac{\eta\gamma^2 G(\boldsymbol{w}_s)/8}{1 + \eta\gamma^2 G(\boldsymbol{w}_s)/8}\right) G(\boldsymbol{w}_s) \tag{114}$$

$$\overset{(ii)}{\leq} \left(1 + \frac{16}{\eta^3 - 16}\right) \left(1 - \frac{1}{10}\eta\gamma^2 G(\boldsymbol{w}_s)\right) G(\boldsymbol{w}_s) \tag{115}$$

$$\leq \left(1 - \frac{1}{10}\eta\gamma^2 G(\boldsymbol{w}_s) + \frac{16}{\eta^3 - 16}\right) G(\boldsymbol{w}_s) \tag{116}$$

$$\overset{(iii)}{\leq} \left(1 - \frac{1}{10}\eta\gamma^2 G(\boldsymbol{w}_s) + \frac{32}{\eta^3}\right) G(\boldsymbol{w}_s) \tag{117}$$

$$= \left(1 - \left(\frac{1}{10} - \frac{32}{\eta^4\gamma^2 G(\boldsymbol{w}_s)}\right)\eta\gamma^2 G(\boldsymbol{w}_s)\right) G(\boldsymbol{w}_s) \tag{118}$$

$$\overset{(iv)}{\leq} \left(1 - \left(\frac{1}{10} - \frac{512}{\eta^3\gamma^2}\right)\eta\gamma^2 G(\boldsymbol{w}_s)\right) G(\boldsymbol{w}_s) \tag{119}$$

$$\overset{(v)}{\leq} \left(1 - \frac{1}{12}\eta\gamma^2 G(\boldsymbol{w}_s)\right) G(\boldsymbol{w}_s) \tag{120}$$

where (i) uses $\exp(z) \leq 1/(1-z) = 1 + z/(1-z)$, (ii) uses $G(\boldsymbol{w}_s) \leq 3/(2\eta\gamma^2)$ from Equation (99), (iii) uses $\eta \geq \eta_0 \geq 32$, (iv) uses $F(\boldsymbol{w}_s) > 1/(8\eta) \implies G(\boldsymbol{w}_s) \geq 1/(16\eta)$ by Lemma C.1, and (v) uses $\eta \geq \eta_0 \geq 32\log(128)/\gamma^2$. Finally, we can unroll this recurrence over s after some manipulation:

$$\frac{1}{G(\boldsymbol{w}_s)} \leq \frac{1}{G(\boldsymbol{w}_{s+1})} - \frac{1}{12}\eta\gamma^2\frac{G(\boldsymbol{w}_s)}{G(\boldsymbol{w}_{s+1})} \tag{121}$$

$$\frac{1}{G(\boldsymbol{w}_s)} \leq \frac{1}{G(\boldsymbol{w}_{s+1})} - \frac{1}{12}\eta\gamma^2 \tag{122}$$

$$\frac{1}{G(\boldsymbol{w}_{s+1})} \geq \frac{1}{G(\boldsymbol{w}_s)} + \frac{1}{12}\eta\gamma^2 \tag{123}$$

$$\frac{1}{G(\boldsymbol{w}_s)} \geq \frac{1}{G(\boldsymbol{w}_{t_0+1})} + \frac{1}{12}\eta\gamma^2(s-(t_0+1)) \tag{124}$$

$$\frac{1}{G(\boldsymbol{w}_s)} \geq \frac{1}{12}\eta\gamma^2(s-(t_0+1)) \tag{125}$$

$$G(\boldsymbol{w}_s) \leq \frac{12}{\eta\gamma^2(s-(t_0+1))}. \tag{126}$$

Now define $t_1 = 1 + t_0 + 192/\gamma^2$. If $\tau \geq t_1$ and no oscillation happens between t and t_1, then $G(\boldsymbol{w}_{t_1}) \leq 1/(16\eta)$, which implies $F(\boldsymbol{w}_{t_1}) \leq 1/(8\eta)$ by Lemma C.6, so that $\tau = t_1$. ■

Lemma 2.7 *For any t, if there exists an oscillation $t_k > t$ and $\eta \geq \eta_0$, then $\hat{w}_{t+1} \geq (1+\gamma^2/2)\hat{w}_t$.*

Proof Without loss of generality, assume that t_k is the earliest iteration where an oscillation occurs after t.

Similarly to the proof of Lemma 2.6, we can partition the dataset into D_+ and D_-, where $D_+ = D_s^+$ and $D_- = D_s^-$ for all $s \in \{t, \dots, t_k - 1\}$. We can then apply Lemma 2.4 for each iteration from t to t_k and conclude that $a_t^i \leq a_{t_k}^i$ for $i \in D_-$. Therefore

$$\hat{w}_{t+1} - \hat{w}_t = \eta\langle \boldsymbol{w}_*, -\nabla F(\boldsymbol{w}_t)\rangle \tag{127}$$

$$= \frac{\eta}{n}\sum_{i=1}^n \frac{\langle \boldsymbol{w}_*, \boldsymbol{x}_i\rangle}{\exp(a_t^i)+1} \tag{128}$$

$$\geq \frac{\eta}{n}\sum_{i\in D_-} \frac{\langle \boldsymbol{w}_*, \boldsymbol{x}_i\rangle}{\exp(a_t^i)+1} \tag{129}$$

$$\geq \frac{\eta}{n}\sum_{i\in D_-} \frac{\langle \boldsymbol{w}_*, \boldsymbol{x}_i\rangle}{\exp(a_{t_k}^i)+1} \tag{130}$$

$$= \eta\Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^n \frac{\langle \boldsymbol{w}_*, \boldsymbol{x}_i\rangle}{\exp(a_{t_k}^i)+1}}_{A_1} - \underbrace{\frac{1}{n}\sum_{i\in D_+} \frac{\langle \boldsymbol{w}_*, \boldsymbol{x}_i\rangle}{\exp(a_{t_k}^i)+1}}_{A_2}\Bigg). \tag{131}$$

Note that

$$A_1 \geq \frac{\gamma}{n}\sum_{i=1}^n \frac{1}{\exp(a_{t_k}^i)+1} = \gamma G(\boldsymbol{w}_{t_k}), \tag{132}$$

and

$$|A_2| \leq \frac{1}{n} \sum_{i \in D_+} \frac{1}{\exp(a_{t_k}^i) + 1} \overset{(i)}{\leq} \frac{1}{n} \sum_{i \notin B_{t_k}} \frac{1}{\exp(a_{t_k}^i) + 1} \overset{(ii)}{\leq} \frac{16}{\eta^3} G(\boldsymbol{w}_{t_k}), \tag{133}$$

where (i) uses $B_{t_k} \subset D_-$ and (ii) uses Lemma A.2. Using $\eta \geq \eta_0 \geq 32/\gamma^2$, this means

$$|A_2| \leq \frac{16}{\eta^3} G(\boldsymbol{w}_{t_k}) \leq \frac{1}{2\eta^2} \gamma^2 G(\boldsymbol{w}_{t_k}) \leq \frac{1}{2} \gamma G(\boldsymbol{w}_{t_k}) \leq \frac{1}{2} A_1. \tag{134}$$

Plugging back to Equation (131) yields

$$\begin{aligned}
\hat{w}_{t+1} - \hat{w}_t &\geq \frac{1}{2} \eta A_1 && (135)\\
&\geq \frac{\eta}{2n} \sum_{i=1}^{n} \frac{\langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle}{\exp(a_{t_k}^i) + 1} && (136)\\
&= \frac{\eta}{2} \langle \boldsymbol{w}_*, -\nabla F(\boldsymbol{w}_{t_k}) \rangle && (137)\\
&= \frac{1}{2} (\hat{w}_{t_k+1} - \hat{w}_{t_k}) && (138)\\
&\overset{(i)}{\geq} \frac{1}{2} \gamma^2 \hat{w}_{t_k} && (139)\\
&\overset{(ii)}{\geq} \frac{1}{2} \gamma^2 \hat{w}_t, && (140)
\end{aligned}$$

where (i) uses Lemma 2.2 on iteration t_k and (ii) uses that $\hat{w}_t$ is strictly increasing (Lemma 2.1). ■

Theorem 2.1 *If $d = 2$ and $\eta \geq \eta_0 := \max(n, 32/\gamma^2 \log(256/\gamma^2))$, then the transition time $\tau := \min\{t \geq 0 : F(\boldsymbol{w}_t) \leq 1/8\eta\}$ of GD for Equation (1) satisfies*

$$\tau \leq \mathcal{O}\left(\frac{n}{\gamma} + \frac{\log(1/\gamma)}{\gamma^2} \right). \tag{6}$$

Further, $F(\boldsymbol{w}_t) \leq \mathcal{O}\left(1/(\eta\gamma^2(t - \tau))\right)$ for all $t > \tau$.

Proof The upper bound of τ was shown in Lemma 2.8. For the upper bound of $F(\boldsymbol{w}_t)$, we first show that $F(\boldsymbol{w}_t) \leq 1/8\eta$ for all $t \geq \tau$. This is essentially a repetition of parts of the analysis from Wu et al. (2024) and Crawshaw et al. (2025a); we include it here for completeness.

We already know that $F(\boldsymbol{w}_\tau) \leq 1/8\eta$, so assume that $F(\boldsymbol{w}_t) \leq 1/8\eta$ for $t \geq \tau$. We can show that $F(\boldsymbol{w}_{t+1}) \leq F(\boldsymbol{w}_t)$ by applying the modified descent inequality of Lemma C.3, but to do so we need to verify the condition $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| \leq 1$:

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| = \eta \|\nabla F(\boldsymbol{w}_t)\| \overset{(i)}{\leq} \eta F(\boldsymbol{w}_t) \overset{(ii)}{\leq} 1/8, \tag{141}$$

where (i) uses $\|\nabla F(\boldsymbol{w}_t)\| \leq F(\boldsymbol{w}_t)$ from Lemma C.4 and (ii) uses the inductive hypothesis. So we can apply Lemma C.3:

$$F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}_t) \leq \langle \nabla F(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle + 4F(\boldsymbol{w}_t)\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \tag{142}$$

$$= -\eta\|\nabla F(\boldsymbol{w}_t)\|^2 + 4\eta^2 F(\boldsymbol{w}_t)\|\nabla F(\boldsymbol{w}_t)\|^2 \tag{143}$$

$$= -\eta\left(1 - 4\eta F(\boldsymbol{w}_t)\right)\|\nabla F(\boldsymbol{w}_t)\|^2 \tag{144}$$

$$\overset{(i)}{\leq} -\frac{1}{2}\eta\|\nabla F(\boldsymbol{w}_t)\|^2 \tag{145}$$

$$\overset{(ii)}{\leq} -\frac{1}{8}\eta\gamma^2 F(\boldsymbol{w}_t)^2, \tag{146}$$

where (i) uses the inductive hypothesis, and (ii) uses Lemma C.5. Note that the condition of Lemma C.5 is satisfied here, since $F(\boldsymbol{w}_t) \leq 1/8\eta \leq \log(2)/n$, which implies for every $i \in [n]$ that

$$\log(1 + \exp(-a_t^i)) = n \cdot \frac{1}{n}\log(1 + \exp(-a_t^i)) \leq n \cdot \frac{1}{n}\sum_{j=1}^{n}\log(1 + \exp(-a_t^j)) \tag{147}$$

$$= nF(\boldsymbol{w}_t) \overset{(i)}{\leq} n/(8\eta) \overset{(ii)}{\leq} \log 2, \tag{148}$$

where (i) uses the inductive hypothesis and (ii) uses $\eta \geq \eta_0 \geq n$, and therefore $a_t^i \geq 0$.

By Equation (146), we have $F(\boldsymbol{w}_{t+1}) < F(\boldsymbol{w}_t)$, which completes the induction. We can then use Equation (146) to get a convergence rate of $F(\boldsymbol{w}_t)$:

$$F(\boldsymbol{w}_{t+1}) - F(\boldsymbol{w}_t) \leq -\frac{1}{8}\eta\gamma^2 F(\boldsymbol{w}_t)^2 \tag{149}$$

$$\frac{1}{F(\boldsymbol{w}_t)} - \frac{1}{F(\boldsymbol{w}_{t+1})} \leq -\frac{1}{8}\eta\gamma^2\frac{F(\boldsymbol{w}_t)}{F(\boldsymbol{w}_{t+1})} \tag{150}$$

$$\frac{1}{F(\boldsymbol{w}_{t+1})} \geq \frac{1}{F(\boldsymbol{w}_t)} + \frac{1}{8}\eta\gamma^2\frac{F(\boldsymbol{w}_t)}{F(\boldsymbol{w}_{t+1})} \tag{151}$$

$$\frac{1}{F(\boldsymbol{w}_{t+1})} \geq \frac{1}{F(\boldsymbol{w}_t)} + \frac{1}{8}\eta\gamma^2, \tag{152}$$

and unrolling back to $t = \tau$,

$$\frac{1}{F(\boldsymbol{w}_t)} \geq \frac{1}{F(\boldsymbol{w}_\tau)} + \frac{1}{8}\eta\gamma^2(t - \tau) \tag{153}$$

$$F(\boldsymbol{w}_t) \leq \frac{8}{1/F(\boldsymbol{w}_\tau) + \eta\gamma^2(t - \tau)} \tag{154}$$

$$F(\boldsymbol{w}_t) \leq \frac{8}{\eta\gamma^2(t - \tau)} \tag{155}$$

■

Appendix B. Deferred Proofs from Section 3

Recall the definition $\eta_1 := \max\left\{n, 32/\gamma^2 \log(3/\gamma)\right\}$. Throughout the proof of Theorem 3.1, we require $\eta \geq \eta_1$.

Lemma 3.1 *[Time until Classification] Suppose that $\gamma \leq 1/6$, $n \geq 6$, and $\eta \geq \eta_1$. Then there exists some dataset satisfying Assumption 1.1, such that for every $t \leq n/(16\gamma)$, there exists $i \in [n]$ with $\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle < 0$.*

Proof Recall the dataset definition:

$$\boldsymbol{x}_1 = (\gamma, -\gamma) \tag{156}$$

$$\boldsymbol{x}_i = (\gamma, \sqrt{1-\gamma^2}), \quad \text{for } i \in \{2, \ldots, n\}, \tag{157}$$

The iterate $\boldsymbol{w}_1$ after the first step can be computed exactly:

$$\boldsymbol{w}_1 = \boldsymbol{w}_0 - \eta \nabla F(\boldsymbol{w}_0) \tag{158}$$

$$= \boldsymbol{w}_0 + \frac{\eta}{n} \sum_{i=1}^{n} \frac{\boldsymbol{x}_i}{\exp(\langle \boldsymbol{w}_0, \boldsymbol{x}_i \rangle) + 1} \tag{159}$$

$$= \frac{\eta}{2n} \sum_{i=1}^{n} \boldsymbol{x}_i \tag{160}$$

$$= \frac{1}{2}\eta\gamma \boldsymbol{e}_1 + \frac{1}{2}\eta \left(\left(1 - \frac{1}{n}\right) \sqrt{1-\gamma^2} - \frac{\gamma}{n} \right) \boldsymbol{e}_2. \tag{161}$$

The loss for each data point is determined by $\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle$, which for $t = 1$ can be computed exactly:

$$a_1^1 = \langle \boldsymbol{w}_1, \boldsymbol{x}_1 \rangle \tag{162}$$

$$= \frac{1}{2}\eta\gamma^2 - \frac{1}{2}\eta\gamma \left(\left(1 - \frac{1}{n}\right) \sqrt{1-\gamma^2} - \frac{\gamma}{n} \right), \tag{163}$$

and for $i \geq 2$,

$$a_1^i = \langle \boldsymbol{w}_1, \boldsymbol{x}_i \rangle \tag{164}$$

$$= \frac{1}{2}\eta\gamma^2 + \frac{1}{2}\eta\sqrt{1-\gamma^2} \left(\left(1 - \frac{1}{n}\right) \sqrt{1-\gamma^2} - \frac{\gamma}{n} \right). \tag{165}$$

Note that

$$\left(\left(1 - \frac{1}{n}\right) \sqrt{1-\gamma^2} - \frac{\gamma}{n} \right) \geq \frac{5}{6}\sqrt{1-(1/4)^2} - \frac{1}{6} \geq \frac{1}{2}. \tag{166}$$

since $\gamma \leq 1/4$ and $n \geq 6$. So

$$a_1^1 \leq \frac{1}{2}\eta\gamma^2 - \frac{1}{4}\eta\gamma \leq -\frac{1}{8}\eta\gamma. \tag{167}$$

Let t_c be the first timestep where $a_t^1 > 0$. Then for all t with $1 \le t < t_c$,

$$a_{t+1}^1 - a_t^1 = \langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \boldsymbol{x}_1 \rangle \tag{168}$$

$$= \frac{\eta}{n} \sum_{j=1}^{n} \frac{\langle \boldsymbol{x}_j, \boldsymbol{x}_1 \rangle}{\exp(a_t^j) + 1} \tag{169}$$

$$= \frac{\eta}{n} \frac{\|\boldsymbol{x}_1\|^2}{\exp(a_t^1) + 1} + \frac{\eta}{n} \sum_{j \ge 2} \frac{\langle \boldsymbol{x}_j, \boldsymbol{x}_1 \rangle}{\exp(a_t^j) + 1} \tag{170}$$

$$= \frac{2\eta\gamma^2}{n} \frac{1}{\exp(a_t^1) + 1} + \frac{\eta(\gamma^2 - \gamma\sqrt{1-\gamma^2})}{n} \sum_{j \ge 2} \frac{1}{\exp(a_t^j) + 1} \tag{171}$$

$$\le \frac{2\eta\gamma^2}{n} \frac{1}{\exp(a_t^1) + 1} - \frac{\eta\gamma}{2n} \sum_{j \ge 2} \frac{1}{\exp(a_t^j) + 1} \tag{172}$$

$$\le \frac{2\eta\gamma^2}{n} - \frac{\eta\gamma}{2n} \sum_{j \ge 2} \frac{1}{\exp(a_t^j) + 1} \tag{173}$$

$$\le \frac{2\eta\gamma^2}{n}. \tag{174}$$

Now we can recurse over t from 1 to t_c:

$$a_{t_c}^1 = a_1^1 + \sum_{t=1}^{t_c - 1} (a_{t+1}^1 - a_t^1) \tag{175}$$

$$0 \overset{(i)}{<} -\eta\gamma/8 + \sum_{t=1}^{t_c - 1} (a_{t+1}^1 - a_t^1) \tag{176}$$

$$\eta\gamma/8 \overset{(ii)}{\le} 2(t_c - 1)\eta\gamma^2/n \tag{177}$$

$$t_c \ge 1 + \frac{n}{16\gamma}, \tag{178}$$

where (i) uses Equation (167), and (ii) uses Equation (174). ■

Lemma 3.2 *[Time until Stability] Suppose that $\gamma \le 1/6, n \ge 2$, and $\eta \ge \eta_1$. Then there exists some dataset satisfying Assumption 1.1 such that $F(\boldsymbol{w}_t) > 2/\eta$ for all $t \le 1 + 1/(59\gamma^2)$.*

Proof Denote $F_i(\boldsymbol{w}) = \log(1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle))$, so that $F = \frac{1}{n} \sum_{i=1}^n F_i$.

Step 1: Constructing the dataset Let $\boldsymbol{v}_1 = (\gamma, -\delta)$ and $\boldsymbol{v}_2 = (\gamma, \sqrt{1-\gamma^2})$, where $\delta \in [0, \sqrt{1-\gamma^2}]$. We define a dataset as

$$\boldsymbol{x}_i = \begin{cases} \boldsymbol{v}_1 & i \le k := \lceil n/2 \rceil \\ \boldsymbol{v}_2 & i > k \end{cases} \tag{179}$$

It is easy to verify that this dataset satisfies Assumption 1.1 and has maximum margin γ with $\boldsymbol{w}_* = (1, 0)$. Denoting

$$\lambda = \log\left(\frac{1}{\exp(2n/k\eta) - 1}\right), \tag{180}$$

we want to choose δ so that $\langle \boldsymbol{w}_1, \boldsymbol{v}_1 \rangle = \lambda - 1$ and $\langle \boldsymbol{w}_1, \boldsymbol{v}_2 \rangle \geq \Omega(\eta)$. Notice that $\log(1+\exp(-\lambda)) = 2n/k\eta$, so if for some $\boldsymbol{w}$ we have $\langle \boldsymbol{w}, \boldsymbol{v}_1 \rangle < \lambda$, then

$$F(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)) \tag{181}$$

$$= \frac{k}{n} \log(1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{v}_1 \rangle)) + \frac{n-k}{n} \log(1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{v}_2 \rangle)) \tag{182}$$

$$> \frac{k}{n} \log(1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{v}_1 \rangle)) \tag{183}$$

$$> \frac{k}{n} \log(1 + \exp(-\lambda)) \tag{184}$$

$$> \frac{k}{n} \frac{2n}{k\eta} = 2/\eta. \tag{185}$$

We can find a δ that satisfies these conditions with a simple derivation. First, we write $\boldsymbol{w}_1$ in closed form:

$$\boldsymbol{w}_1 = -\eta \nabla F(\boldsymbol{0}) = \frac{\eta}{2n} \sum_{i=1}^{n} \boldsymbol{x}_i = \frac{\eta k}{2n} \boldsymbol{v}_1 + \frac{\eta(n-k)}{2n} \boldsymbol{v}_2 \tag{186}$$

We want δ to satisfy:

$$\langle \boldsymbol{w}_1, \boldsymbol{v}_1 \rangle = \lambda - 1 \tag{187}$$

$$\left\langle \frac{\eta k}{2n} \boldsymbol{v}_1 + \frac{\eta(n-k)}{2n} \boldsymbol{v}_2, \boldsymbol{v}_1 \right\rangle = \lambda - 1 \tag{188}$$

$$\frac{\eta k}{2n} \left(\|\boldsymbol{v}_1\|^2 + (n/k - 1)\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle \right) = \lambda - 1 \tag{189}$$

$$\frac{\eta k}{2n} \left(\gamma^2 + \delta^2 + (n/k - 1)(\gamma^2 - \delta\sqrt{1-\gamma^2}) \right) = \lambda - 1 \tag{190}$$

$$\delta^2 - (n/k - 1)\sqrt{1-\gamma^2}\delta + (n/k)\gamma^2 - (2n/\eta k)(\lambda - 1) = 0 \tag{191}$$

This equation has a solution, since the discriminant is

$$(n/k-1)^2(1-\gamma^2) - 4((n/k)\gamma^2 - (2n/\eta k)(\lambda - 1)) \overset{(i)}{\geq} (n/k-1)^2(1-\gamma^2) - 4(n/k)\gamma^2 \tag{192}$$

$$\overset{(ii)}{\geq} \frac{1}{4}(1-\gamma^2) - 8\gamma^2 \tag{193}$$

$$= \frac{1}{4} - \frac{33}{4}\gamma^2 \tag{194}$$

$$\overset{(iii)}{\geq} \frac{1}{4}\left(1 - \frac{33}{36}\right) > 0, \tag{195}$$

where (i) uses that $\lambda > 1$, which comes from

$$\lambda = \log\left(\frac{1}{\exp(2n/k\eta) - 1}\right) \geq \log\left(\frac{1}{\exp(4/\eta) - 1}\right) \geq \log\left(\frac{1}{\exp(4/1152) - 1}\right) \geq 1, \tag{196}$$

where the last inequality uses $\eta \geq \eta_1 \geq 32/\gamma^2 \geq 1152$, (ii) uses $k/n \in [1/2, 2/3]$, and (iii) uses $\gamma \leq 1/6$. We will let δ_* be the smaller of the two solutions to Equation (191), that is,

$$\delta_* = \frac{1}{2}\left((n/k-1)\sqrt{1-\gamma^2} - \sqrt{(n/k-1)^2(1-\gamma^2) - 4((n/k)\gamma^2 - (2n/\eta k)(\lambda-1)}\right) \tag{197}$$

$$= \frac{1}{2}\frac{(n/k-1)^2(1-\gamma^2) - \left((n/k-1)^2(1-\gamma^2) - 4((n/k)\gamma^2 - (2n/\eta k)(\lambda-1)\right)}{(n/k-1)\sqrt{1-\gamma^2} + \sqrt{(n/k-1)^2(1-\gamma^2) - 4((n/k)\gamma^2 - (2n/\eta k)(\lambda-1)}} \tag{198}$$

$$= 2\frac{(n/k)\gamma^2 - (2n/\eta k)(\lambda-1)}{(n/k-1)\sqrt{1-\gamma^2} + \sqrt{(n/k-1)^2(1-\gamma^2) - 4((n/k)\gamma^2 - (2n/\eta k)(\lambda-1)}} \tag{199}$$

Notice that

$$\delta_* \leq 2\frac{(n/k)\gamma^2}{(n/k-1)\sqrt{1-\gamma^2}} \overset{(i)}{\leq} \frac{6\gamma^2}{\sqrt{1-\gamma^2}} \overset{(ii)}{\leq} 7\gamma^2 \tag{200}$$

where (i) uses $n/k \geq 3/2$ and (ii) uses $\gamma \leq 1/6$. To lower bound δ_*, we'll need the following fact:

$$\log(k\eta/2n) \overset{(i)}{\leq} \log(\eta/3) \overset{(ii)}{\leq} \log(9/\gamma^2) + \frac{\gamma^2}{9}\left(\frac{\eta}{3} - \frac{9}{\gamma^2}\right) \tag{201}$$

$$\leq 2\log(3/\gamma) + \frac{1}{27}\eta\gamma^2 \overset{(iii)}{\leq} \frac{1}{8}\eta\gamma^2 + \frac{1}{27}\eta\gamma^2 \leq \frac{1}{6}\eta\gamma^2, \tag{202}$$

where (i) uses $k/n \leq 2/3$, (ii) uses concavity of $\log$, and (iii) uses $\eta \geq 16/\gamma^2 \log(3/\gamma)$. Now we can upper bound λ as

$$\lambda = \log\left(\frac{1}{\exp(2n/k\eta) - 1}\right) \overset{(i)}{\leq} \log(k\eta/2n) \leq \frac{1}{6}\eta\gamma^2, \tag{203}$$

where (i) uses $\exp(z) \geq 1 + z$. Finally, we can lower bound δ_*:

$$\delta_* \geq 2\frac{(n/k)\gamma^2 - (2n/\eta k)\lambda}{(n/k-1)\sqrt{1-\gamma^2} + \sqrt{(n/k-1)^2(1-\gamma^2) - 4((n/k)\gamma^2 - (2n/\eta k)\lambda}} \tag{204}$$

$$\overset{(i)}{\geq} 2\frac{(n/k)\gamma^2 - (n/3k)\gamma^2}{(n/k-1)\sqrt{1-\gamma^2} + \sqrt{(n/k-1)^2(1-\gamma^2) - 4((n/k)\gamma^2 - (n/3k)\gamma^2}} \tag{205}$$

$$= \frac{4}{3}\frac{(n/k)\gamma^2}{(n/k-1)\sqrt{1-\gamma^2} + \sqrt{(n/k-1)^2(1-\gamma^2) - (8/3)(n/k)\gamma^2}} \tag{206}$$

$$\geq \frac{2}{3}\frac{(n/k)\gamma^2}{(n/k-1)\sqrt{1-\gamma^2}} \overset{(ii)}{\geq} \frac{4}{3}\gamma^2, \tag{207}$$

where (i) uses Equation (203) and (ii) uses $n/k \geq 3/2$ and $\gamma > 0$. Together, Equation (200) and Equation (204) show that $\delta_* = \Theta(\gamma^2)$, which we will use in the following analysis. For the remainder of the proof, we will choose $\delta = \delta_*$, yielding the dataset $\boldsymbol{v}_1 = (\gamma, -\delta_*), \boldsymbol{v}_2 = (\gamma, \sqrt{1-\gamma^2})$.

Step 2: $\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle$ is increasing Let $t_s = \min\{t \geq 0 \mid \langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle \geq \lambda\}$. We want to show that $\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle$ is increasing from $t = 1$ to $t = t_s$. For $t < t_s$,

$$\langle \boldsymbol{w}_{t+1}, \boldsymbol{v}_1 \rangle - \langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle = \frac{\eta}{n} \sum_{i=1}^{n} \frac{\langle \boldsymbol{x}_i, \boldsymbol{v}_1 \rangle}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) + 1} \tag{208}$$

$$= \frac{k\eta}{n} \left(\frac{\|\boldsymbol{v}_1\|^2}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle) + 1} + \frac{(n/k - 1)\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle) + 1} \right) \tag{209}$$

$$\overset{(i)}{\geq} \frac{k\eta}{n} \left(\frac{\gamma^2}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle) + 1} - \frac{6(n/k - 1)\gamma^2}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle) + 1} \right) \tag{210}$$

$$\geq \frac{k\eta\gamma^2}{n} \left(\frac{1}{\exp(\lambda) + 1} - \frac{6}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle) + 1} \right), \tag{211}$$

where (i) uses

$$\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle = \gamma^2 - \delta_* \sqrt{1 - \gamma^2} \geq (1 - 7\sqrt{1 - \gamma^2})\gamma^2 \geq -6\gamma^2. \tag{212}$$

This means that, to show $\langle \boldsymbol{w}_{t+1}, \boldsymbol{v}_1 \rangle \geq \langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle$, it suffices that

$$\frac{1}{\exp(\lambda) + 1} \geq \frac{6}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle) + 1} \tag{213}$$

$$(\iff) \quad \exp(\lambda) + 1 \leq \frac{1}{6} \left(\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle) + 1 \right) \tag{214}$$

$$(\impliedby) \quad 2\exp(\lambda) \leq \frac{1}{6} \exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle) \tag{215}$$

$$(\iff) \quad \lambda + \log(12) \leq \langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle. \tag{216}$$

To lower bound $\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle$, we use $\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle \leq \lambda$ to bound $\langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle$:

$$\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle < \lambda \tag{217}$$

$$\gamma \langle \boldsymbol{w}_t, \boldsymbol{e}_1 \rangle - \delta_* \langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle < \lambda \tag{218}$$

$$\langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle > \frac{1}{\delta_*} \left(\gamma \langle \boldsymbol{w}_t, \boldsymbol{e}_1 \rangle - \lambda \right) \tag{219}$$

$$\langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle > \frac{1}{\delta_*} \left(\gamma \langle \boldsymbol{w}_t, \boldsymbol{w}_* \rangle - \lambda \right) \tag{220}$$

$$\langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle \overset{(i)}{>} \frac{1}{\delta_*} \left(\gamma \langle \boldsymbol{w}_1, \boldsymbol{w}_* \rangle - \lambda \right) \tag{221}$$

$$\langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle > \frac{1}{\delta_*} \left(\frac{1}{2} \eta\gamma^2 - \lambda \right) \tag{222}$$

$$\langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle \overset{(ii)}{>} \frac{1}{3\delta_*} \eta\gamma^2 \tag{223}$$

$$\langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle \overset{(iii)}{>} \frac{1}{21} \eta, \tag{224}$$

where (i) uses the monotonicity part of Lemma 2.1, (ii) uses Equation (203), and (iii) uses Equation (200). Now we can lower bound $\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle$:

$$\begin{aligned}
\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle &= \langle \boldsymbol{w}_t, \boldsymbol{e}_1 \rangle \gamma + \langle \boldsymbol{w}_t, \boldsymbol{e}_2 \rangle \sqrt{1-\gamma^2} && (225)\\
&\overset{(i)}{\geq} \frac{1}{21}\eta\sqrt{1-\gamma^2} && (226)\\
&\overset{(ii)}{\geq} \frac{1}{24}\eta && (227)\\
&= \frac{9}{240}\eta + \frac{1}{240}\eta && (228)\\
&\overset{(iii)}{\geq} \lambda + \log(12) && (229)
\end{aligned}$$

where (i) uses $\langle \boldsymbol{w}_t, \boldsymbol{w}_* \rangle \geq \langle \boldsymbol{w}_0, \boldsymbol{w}_* \rangle = 0$ from Lemma 2.1, (ii) uses $\gamma \leq 1/6$, and (iii) uses Equation (203), $\gamma \leq 1/6$, and $\eta \geq \eta_1 \geq 32/\gamma^2 \geq 1152$. This proves Equation (216), which implies that $\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle$ is increasing for $t < t_s$.

Step 3: Trajectory Analysis Recall from Equation (185) that $\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle < \lambda$ implies $F(\boldsymbol{w}_t) > 2/\eta$, so to prove the lemma, it suffices to show that $t_s \geq \Omega(1/\gamma^2)$. To lower bound t_s, we upper bound $\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle$: for all $1 \leq t < t_s$,

$$\begin{aligned}
\langle \boldsymbol{w}_{t+1}, \boldsymbol{v}_1 \rangle - \langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle &= \frac{\eta}{n}\sum_{i=1}^{n} \frac{\langle \boldsymbol{x}_i, \boldsymbol{v}_1 \rangle}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle) + 1} && (230)\\
&= \frac{k\eta}{n}\left(\frac{\|\boldsymbol{v}_1\|^2}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle) + 1} + \frac{(n/k-1)\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_2 \rangle) + 1} \right) && (231)\\
&\overset{(i)}{\leq} \frac{k\eta}{n} \frac{\|\boldsymbol{v}_1\|^2}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle) + 1} && (232)\\
&= \frac{k\eta}{n} \frac{\gamma^2 + \delta_*^2}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle) + 1} && (233)\\
&\overset{(ii)}{\leq} \frac{8k\eta\gamma^2}{n} \frac{1}{\exp(\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle) + 1} && (234)\\
&\overset{(iii)}{\leq} \frac{8k\eta\gamma^2}{n} \frac{1}{\exp(\langle \boldsymbol{w}_1, \boldsymbol{v}_1 \rangle) + 1} && (235)\\
&= \frac{8k\eta\gamma^2}{n} \frac{1}{\exp(\lambda - 1) + 1} \leq \frac{8ek\eta\gamma^2}{n}(\exp(2n/k\eta) - 1) && (236)\\
&\leq \frac{8ek\eta\gamma^2}{n}(\exp(4/\eta) - 1) \overset{(iv)}{\leq} \frac{8ek\eta\gamma^2}{n} \frac{4/\eta}{4/1152}(\exp(4/1152) - 1) && (237)\\
&\leq \frac{8ek}{n} 1152(\exp(4/1152) - 1)\gamma^2 \leq 59\gamma^2, && (238)
\end{aligned}$$

where (i) uses $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle \leq 0$ from Equation (212), (ii) uses $\delta_* \leq 7\gamma^2$ from Equation (200), (iii) uses that $\langle \boldsymbol{w}_t, \boldsymbol{v}_1 \rangle$ is increasing for $t < t_s$, and (iv) uses convexity of exp together with $\eta \geq \eta_1 \geq$

$32/\gamma^2 \geq 1152$. Finally, this means that

$$\langle \boldsymbol{w}_{t_s}, \boldsymbol{v}_1 \rangle \leq \langle \boldsymbol{w}_1, \boldsymbol{v}_1 \rangle + 59\gamma^2(t_s - 1) \tag{239}$$

$$\lambda \leq (\lambda - 1) + 59\gamma^2(t_s - 1) \tag{240}$$

$$t_s \geq 1 + \frac{1}{59\gamma^2}. \tag{241}$$

■

Theorem 3.1 *If $\gamma \leq 1/6, n \geq 2$, and $\eta \geq \eta_1 := \max\left\{n, 32/\gamma^2 \log(3/\gamma)\right\}$, then there exists a dataset satisfying Assumption 1.1 such that the transition time $\tau := \min\{t \geq 0 : F(\boldsymbol{w}_t) \leq 1/8\eta\}$ of GD for Equation (1) satisfies $\tau \geq \Omega\left(n/\gamma + 1/\gamma^2\right)$.*

Proof The result follows more or less immediately from Lemmas 3.1 and 3.2. Recalling

$$t_c = \min\left\{t \geq 0 : \langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle \geq 0 \text{ for all } i \in [n]\right\} \tag{242}$$

$$t_s = \min\left\{t \geq 0 : F(\boldsymbol{w}_t) \leq 2/\eta\right\}, \tag{243}$$

We know that $\tau \geq t_c$, since

$$\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle \leq 0 \implies F(\boldsymbol{w}_t) \geq \frac{1}{n}\log(1 + \exp(-\langle \boldsymbol{w}_t, \boldsymbol{x}_i \rangle)) \geq \frac{\log 2}{n} \geq \frac{1}{8\eta}, \tag{244}$$

where the last line uses $\eta \geq \eta_1 \geq n$. Also, $\tau \geq t_s$ is immediate from definitions.

The only detail needing consideration is the condition $n \geq 6$ for Lemma 3.1. If this condition is not met, then $n \leq 5 \leq 1/\gamma$, so Lemma 3.2 implies

$$\tau \geq t_s \geq \frac{1}{59\gamma^2} = \frac{1}{118\gamma^2} + \frac{1}{118\gamma^2} \geq \frac{1}{118\gamma^2} + \frac{n}{118\gamma}. \tag{245}$$

If $n \geq 6$, then Lemmas 3.1 and 3.2 imply

$$\tau \geq \max\left\{t_c, t_s\right\} \geq \max\left\{\frac{n}{16\gamma}, \frac{1}{59\gamma^2}\right\} \geq \frac{1}{2}\left(\frac{n}{16\gamma} + \frac{1}{59\gamma^2}\right) = \frac{n}{32\gamma} + \frac{1}{118\gamma^2}. \tag{246}$$

In all cases, we have $\tau \geq (n/\gamma + 1/\gamma^2)/118$. ■

Appendix C. Auxiliary Lemmas

Lemma C.1 *Denote $G(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\exp(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) + 1}$. If $F(\boldsymbol{w}) \geq c$, then $G(\boldsymbol{w}) \geq \frac{1 - \exp(-nc)}{n}$. If additionally $c \leq 1/n$, then $G(\boldsymbol{w}) \geq c/2$.*

Proof We want to lower bound the solution of the following:

$$\inf_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \frac{1}{\exp(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) + 1} \tag{247}$$

$$\text{s.t. } \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)) \geq c. \tag{248}$$

The solution of Equation (247) is lower bounded by the following:

$$\inf_{a_1,\ldots,a_n\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\exp(a_i)+1} \tag{249}$$

$$\text{s.t. } \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-a_i)) \geq c. \tag{250}$$

Changing variables to $\ell_i = \log(1+\exp(-a_i))$, we have

$$\frac{1}{\exp(a_i)+1} = 1-\exp(-\ell_i), \tag{251}$$

so Equation (249) can be rewritten as

$$\inf_{\ell_1,\ldots,\ell_n\geq 0} \frac{1}{n}\sum_{i=1}^{n}(1-\exp(-\ell_i)) \tag{252}$$

$$\text{s.t. } \frac{1}{n}\sum_{i=1}^{n}\ell_i \geq c, \tag{253}$$

or $G(\boldsymbol{w}) \geq 1-\phi$, where ϕ is the solution of

$$\sup_{\ell_1,\ldots,\ell_n\geq 0} \frac{1}{n}\sum_{i=1}^{n}\exp(-\ell_i) \tag{254}$$

$$\text{s.t. } \frac{1}{n}\sum_{i=1}^{n}\ell_i \geq c. \tag{255}$$

Note that Equation (254) is equivalent to

$$\sup_{\ell_1,\ldots,\ell_n\geq 0} \frac{1}{n}\sum_{i=1}^{n}\exp(-\ell_i) \tag{256}$$

$$\text{s.t. } \frac{1}{n}\sum_{i=1}^{n}\ell_i = c, \tag{257}$$

since the supremum of Equation (254) will not be achieved when $\frac{1}{n}\sum_{i=1}^{n}\ell_i > c$. Now, the supremum of Equation (256) is achieved by $\ell_1 = cn$ and $\ell_i = 0$ for $i \geq 2$, and this is shown by Karamata's inequality (Lemma C.2): for any other $\ell_1' \geq \ldots \geq \ell_n'$ with $\frac{1}{n}\sum_{i=1}^{n}\ell_i' = c$, all conditions of Lemma C.2 are satisfied, so that $\frac{1}{n}\sum_{i=1}^{n}\exp(-\ell_i) \geq \frac{1}{n}\sum_{i=1}^{n}\exp(-\ell_i')$. Therefore $\phi = 1-\frac{1-\exp(-cn)}{n}$, and

$$G(\boldsymbol{w}) \geq 1-\phi = \frac{1-\exp(-cn)}{n}, \tag{258}$$

which is exactly the desired conclusion.

If we additionally have $c \leq 1/n$, then

$$G(\boldsymbol{w}) \geq \frac{1-\exp(-cn)}{n} \overset{(i)}{\geq} cn\frac{1-\exp(-1)}{n} + (1-cn)\frac{1-\exp(0)}{n} = c(1-1/e) \geq c/2, \tag{259}$$

where (i) uses concavity of $-\exp(\cdot)$. ■

Lemma C.2 (Karamata's Inequality, Theorem 1 (Kadelburg et al., 2005)) *If $f : \mathbb{R} \to \mathbb{R}$ is convex, and $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$ are such that*

1. *$a_1 \geq \ldots \geq a_n$ and $b_1 \geq \ldots \geq b_n$,*
2. *$a_1 + \ldots + a_i \geq b_1 + \ldots + b_i$ for every $i \leq n$,*
3. *$a_1 + \ldots + a_n = b_1 + \ldots + b_n$,*

then $f(a_1) + \ldots + f(a_n) \geq f(b_1) + \ldots + f(b_n)$.

Lemma C.3 (Lemma 4.5 from Crawshaw et al. (2025a)) *For $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, if $\|\boldsymbol{w} - \boldsymbol{w}'\| \leq 1$ then*

$$F(\boldsymbol{w}') \leq F(\boldsymbol{w}) + \langle \nabla F(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w} \rangle + 4F(\boldsymbol{w})\|\boldsymbol{w} - \boldsymbol{w}'\|^2. \tag{260}$$

Lemma C.4 (Lemma 25 from Crawshaw et al. (2025b)) *$\|\nabla F(\boldsymbol{w})\| \leq F(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathbb{R}^d$.*

Lemma C.5 (Lemma 26 from Crawshaw et al. (2025b)) *For all $\boldsymbol{w} \in \mathbb{R}^d$, if $\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 0$ for all $i \in [n]$, then*

$$\|\nabla F(\boldsymbol{w})\| \geq \frac{\gamma}{2} F(\boldsymbol{w}). \tag{261}$$

Lemma C.6 *If $\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 0$ for all $i \in [n]$, then*

$$F(\boldsymbol{w}) \leq 2G(\boldsymbol{w}). \tag{262}$$

Note: This lemma appears as an intermediate step in the analysis of Wu et al. (2024). We include it here for completeness.

Proof

$$F(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)) \leq \frac{1}{n} \sum_{i=1}^{n} \exp(-\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \tag{263}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\exp(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)} \overset{(i)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \frac{2}{\exp(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) + 1} \tag{264}$$

$$\leq 2G(\boldsymbol{w}), \tag{265}$$

where (i) uses $\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \geq 0$. ■