**GRADIENT DESCENT MAXIMIZES THE MARGIN OF HOMOGENEOUS NEURAL NETWORKS**

Kaifeng Lyu & Jian Li  
Institute for Interdisciplinary Information Sciences  
Tsinghua University  
Beijing, China  
vfleaking@gmail.com, lijian83@mail.tsinghua.edu.cn

**ABSTRACT**

In this paper, we study the implicit regularization of the gradient descent algorithm in homogeneous neural networks, including fully-connected and convolutional neural networks with ReLU or LeakyReLU activations. In particular, we study the gradient descent or gradient flow (i.e., gradient descent with infinitesimal step size) optimizing the logistic loss or cross-entropy loss of any homogeneous model (possibly non-smooth), and show that if the training loss decreases below a certain threshold, then we can define a smoothed version of the normalized margin which increases over time. We also formulate a natural constrained optimization problem related to margin maximization, and prove that both the normalized margin and its smoothed version converge to the objective value at a KKT point of the optimization problem. Our results generalize the previous results for logistic regression with one-layer or multi-layer linear networks, and provide more quantitative convergence results with weaker assumptions than previous results for homogeneous smooth neural networks. We conduct several experiments to justify our theoretical finding on MNIST and CIFAR-10 datasets. Finally, as margin is closely related to robustness, we discuss potential benefits of training longer for improving the robustness of the model.

1 INTRODUCTION

A major open question in deep learning is why gradient descent or its variants, are biased towards solutions with good generalization performance on the test set. To achieve a better understanding, previous works have studied the implicit bias of gradient descent in different settings. One simple but insightful setting is linear logistic regression on linearly separable data. In this setting, the model is parameterized by a weight vector $w$, and the class prediction for any data point $x$ is determined by the sign of $w^\top x$. Therefore, only the direction $w/\|w\|_2$ is important for making prediction. Soudry et al. (2018a;b); Ji & Telgarsky (2018; 2019c); Nacson et al. (2019c) investigated this problem and proved that the direction of $w$ converges to the direction that maximizes the $L^2$-margin while the norm of $w$ diverges to $+\infty$, if we train $w$ with (stochastic) gradient descent on logistic loss. Interestingly, this convergent direction is the same as that of any regularization path: any sequence of weight vectors $\{w_t\}$ such that every $w_t$ is a global minimum of the $L^2$-regularized loss $\mathcal{L}(w) + \lambda_t \|w\|_2^2$ with $\lambda_t \to 0$ [Rosset et al., 2004]. Indeed, the trajectory of gradient descent is also pointwise close to a regularization path [Suggala et al., 2018].

The aforementioned linear logistic regression can be viewed as a single-layer neural network. A natural and important question is to what extent gradient descent has similar implicit bias for modern deep neural networks. For theoretical analysis, a natural candidate is to consider homogeneous neural networks. Here a neural network $\Phi$ is said to be (positively) homogeneous if there is a number $L > 0$ (called the order) such that the network output $\Phi(\theta; x)$, where $\theta$ stands for the parameter and $x$ stands for the input, satisfies the following:

$$\forall c > 0 : \Phi(c\theta; x) = c^L \Phi(\theta; x) \text{ for all } \theta \text{ and } x. \quad (1)$$

It is important to note that many neural networks are homogeneous [Neyshabur et al., 2015a; Du et al., 2018]. For example, deep fully-connected neural networks or deep CNNs with ReLU or...
LeakyReLU activations can be made homogeneous if we remove all the bias terms, and the order $L$ is exactly equal to the number of layers.

In (Wei et al., 2019), it is shown that the regularization path does converge to the max-margin direction for homogeneous neural networks with cross-entropy or logistic loss. This result suggests that gradient descent or gradient flow may also converge to the max-margin direction by assuming homogeneity, and this is indeed true for some sub-classes of homogeneous neural networks. For gradient flow, this convergent direction is proven for linear fully-connected networks (Ji & Telgarsky, 2019a). For gradient descent on linear fully-connected and convolutional networks, (Gunasekar et al., 2018b) formulate a constrained optimization problem related to margin maximization and prove that gradient descent converges to the direction of a KKT point or even the max-margin direction, under various assumptions including the convergence of loss and gradient directions. In an independent work, (Nacson et al., 2019a) generalize the result in (Gunasekar et al., 2018b) to smooth homogeneous models (we will discuss this work in more details in Section 3).

1.1 MAIN RESULTS

In this paper, we identify a minimal set of assumptions for proving our theoretical results for homogeneous neural networks on classification tasks. Besides homogeneity, we make two additional assumptions:

1. **Exponential-type Loss Function.** We require the loss function to have certain exponential tail (see Appendix A for the details). This assumption is not restrictive as it includes the most popular classification losses: exponential loss, logistic loss and cross-entropy loss.

2. **Separability.** The neural network can separate the training data during training (i.e., the neural network can achieve 100% training accuracy).

While the first assumption is natural, the second requires some explanation. In fact, we assume that at some time $t_0$, the training loss is smaller than a threshold, and the threshold here is chosen to be so small that the training accuracy is guaranteed to be 100% (e.g., for the logistic loss and cross-entropy loss, the threshold can be set to $\ln 2$). Empirically, state-of-the-art CNNs for image classification can even fit randomly labeled data easily (Zhang et al., 2017). Recent theoretical work on over-parameterized neural networks (Allen-Zhu et al., 2019; Zou et al., 2018) show that gradient descent can fit the training data if the width is large enough. Furthermore, in order to study the margin, ensuring the training data can be separated is inevitable; otherwise, there is no positive margin between the data and decision boundary.

**Our Contribution.** Similar to linear models, for homogeneous models, only the direction of parameter $\theta$ is important for making predictions, and one can see that the margin $\gamma(\theta)$ scales linearly with $\|\theta\|_2$, when fixing the direction of $\theta$. To compare margins among $\theta$ in different directions, it makes sense to study the normalized margin, $\bar{\gamma}(\theta) := \gamma(\theta)/\|\theta\|_2$.

In this paper, we focus on the training dynamics of the network after $t_0$ (recall that $t_0$ is a time that the training loss is less than the threshold). Our theoretical results can answer the following questions regarding the normalized margin.

First, *how does the normalized margin change during training?* The answer may seem complicated since one can easily come up with examples in which $\bar{\gamma}$ increases or decreases in a short time interval. However, we can show that the overall trend of the normalized margin is to increase in the following sense: there exists a *smoothed* version of the normalized margin, denoted as $\tilde{\gamma}$, such that (1) $|\tilde{\gamma} - \bar{\gamma}| \to 0$ as $t \to \infty$; and (2) $\tilde{\gamma}$ is non-decreasing for $t > t_0$.

Second, *how large is the normalized margin at convergence?* To answer this question, we formulate a natural constrained optimization problem which aims to directly maximize the margin. We show that every limit point of $\{\theta(t)/\|\theta(t)\|_2 : t > 0\}$ is along the direction of a KKT point of the max-margin problem. This indicates that gradient descent/gradient flow performs margin maximization implicitly in deep homogeneous networks. This result can be seen as a significant generalization of previous works (Soudry et al., 2018a; Ji & Telgarsky, 2019a; Gunasekar et al., 2018b) from linear classifiers to homogeneous classifiers.

---

1Note that this does NOT mean the training loss is 0.
As by-products of the above results, we derive tight asymptotic convergence/growth rates of the loss and weights. It is shown in [Soudry et al., 2018a;b; Ji & Telgarsky, 2018; 2019c] that the loss decreases at the rate of \( O(1/t) \), the weight norm grows as \( O((\log t)^{2/2-L}) \) for linear logistic regression. In this work, we generalize the result by showing that the loss decreases at the rate of \( O(1/(t(\log t)^{2/2-L})) \) and the weight norm grows as \( O((\log t)^{1/L}) \) for homogeneous neural networks with exponential loss, logistic loss, or cross-entropy loss.

**Experiments**

The main practical implication of our theoretical result is that training longer can enlarge the normalized margin. To justify this claim empirically, we train CNNs on MNIST and CIFAR-10 with SGD (see Section K.1). Results on MNIST are presented in Figure 1. For constant step size, we can see that the normalized margin keeps increasing, but the growth rate is rather slow (because the gradient gets smaller and smaller). Inspired by our convergence results for gradient descent, we use a learning rate scheduling method which enlarges the learning rate according to the current training loss, then the training loss decreases exponentially faster and the normalized margin increases significantly faster as well.

For feedforward neural networks with ReLU activation, the normalized margin on a training sample is closely related to the \( L^2 \)-robustness (the \( L^2 \)-distance from the training sample to the decision boundary). Indeed, the former divided by a Lipschitz constant is a lower bound for the latter. For example, the normalized margin is a lower bound for the \( L^2 \)-robustness on fully-connected networks with ReLU activation (see, e.g., Theorem 4 in [Sokolic et al., 2017]). This fact suggests that training longer may have potential benefits on improving the robustness of the model. In our experiments, we observe noticeable improvements of \( L^2 \)-robustness on both training and test sets (see Section K.2).

## 2 Related Work

**Implicit Bias in Training Linear Classifiers.** For linear logistic regression on linearly separable data, Soudry et al. [2018a,b] showed that full-batch gradient descent converges in the direction of the max \( L^2 \)-margin solution of the corresponding hard-margin Support Vector Machine (SVM). Subsequent works extended this result in several ways: Nacson et al. [2019c] extended the results to the case of stochastic gradient descent; Gunasekar et al. [2018a] considered other optimization methods; Nacson et al. [2019b] considered other loss functions including those with poly-exponential tails; Ji & Telgarsky [2018; 2019c] characterized the convergence of weight direction without assuming separability; Ji & Telgarsky [2019b] proved a tighter convergence rate for the weight direction.

Those results on linear logistic regression have been generalized to deep linear networks. Ji & Telgarsky [2019a] showed that the product of weights in a deep linear network with strictly decreasing loss converges in the direction of the max \( L^2 \)-margin solution. Gunasekar et al. [2018b] showed

---

**Figure 1:** (a) Training CNNs with and without bias on MNIST, using SGD with learning rate 0.01. The training loss (left) decreases over time, and the normalized margin (right) keeps increasing after the model is fitted, but the growth rate is slow (\( \approx 1.8 \times 10^{-4} \) after 10000 epochs). (b) Training CNNs with and without bias on MNIST, using SGD with the loss-based learning rate scheduler. The training loss (left) decreases exponentially over time (\( < 10^{-800} \) after 9000 epochs), and the normalized margin (right) increases rapidly after the model is fitted (\( \approx 1.2 \times 10^{-3} \) after 10000 epochs, 10× larger than that of SGD with learning rate 0.01). Experimental details are in Appendix K.

2Code available: [https://github.com/vfleaking/max-margin](https://github.com/vfleaking/max-margin)
more general results for gradient descent on linear fully-connected and convolutional networks with exponential loss, under various assumptions on the convergence of the loss and gradient direction.

Margin maximization phenomenon is also studied for boosting methods (Schapire et al. 1998; Schapire & Freund 2012; Shalev-Shwartz & Singer 2010; Telgarsky 2013) and Normalized Perceptron (Ramdas & Pena 2016).

Implicit Bias in Training Nonlinear Classifiers. Soudry et al. (2018a) analyzed the case where there is only one trainable layer of a ReLU network. Xu et al. (2018) characterized the implicit bias for the model consisting of one single ReLU unit. Our work is closely related to a recent independent work by (Nacson et al. 2019a) which we discuss in details below.

Comparison with (Nacson et al., 2019a). Very recently, (Nacson et al. 2019a) analyzed gradient descent for smooth homogeneous models and proved the convergence of parameter direction to a KKT point of the aforementioned max-margin problem. Compared with their work, our work adopt much weaker assumptions: (1) They assume the training loss converges to 0 along gradient flow (it moves around a circle when the Mexican hat function (Absil et al. 2005)), such that even the direction of the parameter does not converge to a fixed point; (2) They assume the convergence of parameter direction while we prove that KKT conditions hold for all limit points of \{\theta(t)/\|\theta(t)\|_2 : t > 0\}, without requiring any convergence assumption; (3) They assume the convergence of the direction of losses (the direction of the vector whose entries are loss values on every data point) and Linear Independence Constraint Qualification (LICQ) for the max-margin problem, while we do not need such assumptions. Besides the above differences in assumptions, we also prove the monotonicity of the normalized margin and provide tight convergence rate for training loss. We believe both results are interesting in their own right.

Another technical difference is that their work analyzes discrete gradient descent on smooth homogeneous models (which fails to capture ReLU networks). In our work, we analyze both gradient descent on smooth homogeneous models and also gradient flow on homogeneous models which could be non-smooth.

Other Works on Implicit Bias. Banburski et al. (2019) also studied the dynamics of gradient flow and among other things, provided mathematical insights to the implicit bias towards max margin solution for homogeneous networks. We note that their analysis of gradient flow decomposes the dynamics to the tangent component and radial component, which is similar to our proof of Theorem 4.1 in spirit. Wilson et al. (2017); Ali et al. (2019); Gunasekar et al. (2018a) showed that for the linear least-square problem gradient-based methods converge to the unique global minimum that is closest to the initialization in \( L^2 \) distance. Du et al. (2019); Jacot et al. (2018); Lee et al. (2019); Arora et al. (2019b) showed that over-parameterized neural networks of sufficient width (or infinite width) behave as linear models with Neural Tangent Kernel (NTK) with proper initialization and gradient descent converges linearly to a global minimum near the initial point. Other related works include (Ma et al. 2019; Gidel et al. 2019; Arora et al. 2019a; Suggala et al. 2018; Blanc et al. 2019; Neyshabur et al. 2018b).

3 Preliminaries

Basic Notations. For any \( N \in \mathbb{N} \), let \([N] = \{1, \ldots, N\}\). \( \|v\|_2 \) denotes the \( L^2 \)-norm of a vector \( v \). The default base of log is \( e \). For a function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \), \( \nabla f(x) \) stands for the gradient at \( x \) if it exists. A function \( f : X \rightarrow \mathbb{R}^d \) is \( C^k \)-smooth if \( f \) is \( k \) times continuously differentiable. A function \( f : X \rightarrow \mathbb{R} \) is locally Lipschitz if for every \( x \in X \) there exists a neighborhood \( U \) of \( x \) such that the restriction of \( f \) on \( U \) is Lipschitz continuous.

Non-smooth Analysis. For a locally Lipschitz function \( f : X \rightarrow \mathbb{R} \), the Clarke’s subdifferential (Clarke 1977; Clarke et al. 2008; Davis et al. 2020) at \( x \in X \) is the convex set

\[
\partial^c f(x) := \text{conv} \left\{ \lim_{k \to \infty} \nabla f(x_k) : x_k \to x, f \text{ is differentiable at } x_k \right\}.
\]

\(^3\) Assuming the convergence of the parameter direction may seem quite reasonable, however, the problem here can be quite subtle in theory. In Appendix B we present a smooth homogeneous function \( f \), based on the Mexican hat function (Absil et al. 2005), such that even the direction of the parameter does not converge along gradient flow (it moves around a circle when \( t \) increases).
For brevity, we say that a function $z : I \to \mathbb{R}^d$ on the interval $I$ is an arc if $z$ is absolutely continuous for any compact sub-interval of $I$. For an arc $z$, $z'(t)$ (or $\frac{dz}{dt}(t)$) stands for the derivative at $t$ if it exists. Following the terminology in (Davis et al., 2020), we say that a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ admits a chain rule if for any arc $z : [0, +\infty) \to \mathbb{R}^d$, $\forall h \in \partial^o f(z(t)) : (f \circ z)'(t) = (h, z'(t))$ holds for a.e. $t > 0$ (see also Appendix I).

Binary Classification. Let $\Phi$ be a neural network, assumed to be parameterized by $\theta$. The output of $\Phi$ on an input $x \in \mathbb{R}^d$ is a real number $\Phi(\theta; x)$, and the sign of $\Phi(\theta; x)$ stands for the classification result. A dataset is denoted by $D = \{ (x_n, y_n) : n \in [N] \}$, where $x_n \in \mathbb{R}^{d_n}$ stands for a data input and $y_n \in \{ \pm 1 \}$ stands for the corresponding label. For a loss function $\ell : \mathbb{R} \to \mathbb{R}$, we define the training loss of $\Phi$ on the dataset $D$ to be $L(\theta) := \sum_{n=1}^{N} \ell(y_n \Phi(\theta; x_n))$.

**Gradient Descent.** We consider the process of training this neural network $\Phi$ with either gradient descent or gradient flow. For gradient descent, we assume the training loss $L(\theta)$ is $C^2$-smooth and describe the gradient descent process as $\theta(t+1) = \theta(t) - \eta(t)\nabla L(\theta(t))$, where $\eta(t)$ is the learning rate at time $t$ and $\nabla L(\theta(t))$ is the gradient of $L$ at $\theta(t)$.

**Gradient Flow.** For gradient flow, we do not assume the differentiability but only some regularity assumptions including locally Lipschitz. Gradient flow can be seen as gradient descent with infinitesimal step size. In this model, $\theta$ changes continuously with time, and the trajectory of parameter $\theta$ during training is an arc $\theta : [0, +\infty) \to \mathbb{R}^d,$ $t \mapsto \theta(t)$ that satisfies the differential inclusion $\frac{d\theta(t)}{dt} \in -\partial^o L(\theta(t))$ for a.e. $t \geq 0$. The Clarke’s subdifferential $\partial^o L$ is a natural generalization of the usual differential to non-differentiable functions. If $L(\theta)$ is actually a $C^1$-smooth function, the above differential inclusion reduces to $\frac{d\theta(t)}{dt} = -\nabla L(\theta(t))$ for all $t \geq 0$, which corresponds to the gradient flow with differential in the usual sense.

## 4 GRADIENT DESCENT / GRADIENT FLOW ON HOMOGENEOUS MODEL

In this section, we first state our results for gradient flow and gradient descent on homogeneous models with exponential loss $\ell(q) := e^{-q}$ for simplicity of presentation. Due to space limit, we defer the more general results which hold for a large family of loss functions (including logistic loss and cross-entropy loss) to Appendix A/F/G.

### 4.1 Assumptions

**Gradient Flow.** For gradient flow, we assume the following:

(A1). (Regularity). For any fixed $x$, $\Phi(\cdot; x)$ is locally Lipschitz and admits a chain rule;

(A2). (Homogeneity). There exists $L > 0$ such that $\forall \alpha > 0 : \Phi(\alpha \theta; x) = \alpha^L \Phi(\theta; x)$;

(A3). (Exponential Loss). $\ell(q) = e^{-q}$;

(A4). (Separability). There exists a time $t_0$ such that $L(\theta(t_0)) < 1$.

(A1) is a technical assumption about the regularity of the network output. As shown in (Davis et al., 2020), the output of almost every neural network admits a chain rule (as long as the neural network is composed by definable pieces in an o-minimal structure, e.g., ReLU, sigmoid, LeakyReLU).

(A2) assumes the homogeneity, the main property we assume in this work. (A3), (A4) correspond to the two conditions introduced in Section I. The exponential loss in (A3) is main focus of this section. (A4) is a separability assumption: the condition $L(\theta(t_0)) < 1$ ensures that $\ell(y_n \Phi(\theta(t_0); x_n)) < 1$ for all $n \in [N]$, and thus $y_n \Phi(\theta(t_0); x_n) > 0$, meaning that $\Phi$ classifies every $x_n$ correctly.

**Gradient Descent.** For gradient descent, we assume (A2), (A3), (A4) similarly as for gradient flow, and the following two assumptions (S1) and (S5).

(S1). (Smoothness). For any fixed $x$, $\Phi(\cdot; x)$ is $C^2$-smooth on $\mathbb{R}^d \setminus \{0\}$.

(S5). (Learning rate condition. Informal). $\eta(t) = \eta_0$ for a sufficiently small constant $\eta_0$. In fact, $\eta(t)$ is even allowed to be as large as $O(L(t)^{-1} \text{polylog} \frac{1}{L(t)})$. See Appendix E.1 for the details.
Gradient Flow. Our first result is on the overall trend of the normalized margin which the following statements are true for gradient descent:

\[ \tilde{\gamma}(\theta) := \min_{\theta \in \Theta} \left( \frac{\theta}{\|\theta\|_2} \right) = \frac{\min_{\theta \in \Theta} (\theta)}{\|\theta\|_2}. \] (2)

We say \( f \) is an \( \epsilon \)-additive approximation for the normalized margin if \( \tilde{\gamma} - \epsilon \leq f \leq \tilde{\gamma} \), and \( \epsilon \)-multiplicative approximation if \( \gamma \leq f \leq \gamma \).

**Gradient Flow.** Our first result is on the overall trend of the normalized margin \( \tilde{\gamma}(\theta(t)) \). For both gradient flow and gradient descent, we identify a smoothed version of the normalized margin, and show that it is non-decreasing during training. More specifically, we have the following theorem for gradient flow.

**Theorem 4.1** (Corollary of Theorem A.7). Under assumptions (A1) - (A4), there exists an \( O(\|\theta\|_2^{-L}) \)-additive approximation function \( \bar{\gamma}(\theta) \) for the normalized margin such that the following statements are true for gradient flow:

1. For a.e. \( t > t_0 \), \( \frac{d}{dt} \bar{\gamma}(\theta(t)) \geq 0 \);
2. For a.e. \( t > t_0 \), either \( \frac{d}{dt} \bar{\gamma}(\theta(t)) > 0 \) or \( \frac{d}{dt} \frac{\theta(t)}{\|\theta(t)\|_2} = 0 \);
3. \( \mathcal{L}(\theta(t)) \rightarrow 0 \) and \( \|\theta(t)\|_2 \rightarrow \infty \) as \( t \rightarrow +\infty \); therefore, \( |\gamma(\theta(t)) - \bar{\gamma}(\theta(t))| \rightarrow 0 \).

More concretely, the function \( \bar{\gamma}(\theta) \) in Theorem 4.1 is defined as

\[ \bar{\gamma}(\theta) := \frac{\log \frac{1}{\mathcal{L}(\theta)}}{\|\theta\|_2} = -\log \left( \sum_{n=1}^{N} e^{-q_n(\theta)} \right). \] (3)

Note that the only difference between \( \tilde{\gamma}(\theta) \) and \( \bar{\gamma}(\theta) \) is that the margin \( q_{\min}(\theta) \) in \( \tilde{\gamma}(\theta) \) is replaced by \( \log \frac{1}{\mathcal{L}(\theta)} = \mathcal{LSE}(-q_1(\theta), \ldots, q_N(\theta)) \), where \( \mathcal{LSE}(a_1, \ldots, a_N) = \log(\exp(a_1) + \cdots + \exp(a_N)) \) is the LogSumExp function. This is indeed a very natural idea, and previous works on linear models (e.g., (Telgarsky 2013, Nacson et al. 2019b)) also approximate \( q_{\min} \) with \( \mathcal{LSE} \) in the analysis of margin. It is easy to see why \( \bar{\gamma}(\theta) \) is an \( O(\|\theta\|_2^{-L}) \)-additive approximation for \( \gamma(\theta) \): \( e^{a_{\max}} \leq \sum_{n=1}^{N} e^{a_n} \leq N e^{a_{\max}} \) holds for \( a_{\max} = \max\{a_1, \ldots, a_N\} \), so \( e^{a_{\max}} \leq \mathcal{LSE}(a_1, \ldots, a_N) \leq a_{\max} + \log N \); combining this with the definition of \( \bar{\gamma}(\theta) \) gives \( \gamma(\theta) - \|\theta\|_2^{-L} \log N \leq \bar{\gamma}(\theta) \leq \gamma(\theta) \).

**Gradient Descent.** For gradient descent, Theorem 4.1 holds similarly with a slightly different function \( \hat{\gamma}(\theta) \) that approximates \( \bar{\gamma}(\theta) \) multiplicatively rather than additively.

**Theorem 4.2** (Corollary of Theorem E.2). Under assumptions (S1), (A2) - (A4), (S5), there exists an \( (1 - O(1/\log \frac{1}{L})) \)-multiplicative approximation function \( \hat{\gamma}(\theta) \) for the normalized margin such that the following statements are true for gradient descent:

1. For all \( t > t_0 \), \( \hat{\gamma}(\theta(t + 1)) \geq \hat{\gamma}(\theta(t)) \);
2. For all \( t > t_0 \), either \( \hat{\gamma}(\theta(t + 1)) > \hat{\gamma}(\theta(t)) \) or \( \frac{\theta(t + 1)}{\|\theta(t + 1)\|_2} = \frac{\theta(t)}{\|\theta(t)\|_2} \);
3. \( \mathcal{L}(\theta(t)) \rightarrow 0 \) and \( \|\theta(t)\|_2 \rightarrow \infty \) as \( t \rightarrow +\infty \); therefore, \( |\hat{\gamma}(\theta(t)) - \hat{\gamma}(\theta(t))| \rightarrow 0 \).

Due to the discreteness of gradient descent, the explicit formula for \( \hat{\gamma}(\theta) \) is somewhat technical, and we refer the readers to Appendix E for full details.
Convergence Rates. It is shown in Theorem 4.1 that $\mathcal{L}(\theta(t)) \to 0$ and $\|\theta(t)\|_2 \to \infty$. In fact, with a more refined analysis, we can prove tight loss convergence and weight growth rates using the monotonicity of normalized margins.

**Theorem 4.3** (Corollary of Theorem A.10 and E.5). For gradient flow under assumptions (A1) - (A4) or gradient descent under assumptions (S1), (A2) - (A4), (S5), we have the following tight bounds for training loss and weight growth:

$$\mathcal{L}(\theta(t)) = \Theta \left( \frac{1}{T (\log T)^{2-2/L}} \right) \quad \text{and} \quad \|\theta(t)\|_2 = \Theta \left( (\log T)^{1/L} \right),$$

where $T = t$ for gradient flow and $T = \sum_{\tau=t_0}^{t-1} \eta(\tau)$ for gradient descent.

### 4.3 Main Theorem: Convergence to KKT Points

For gradient flow, $\bar{\gamma}$ is upper-bounded by $\bar{\gamma} \leq \bar{\gamma} \leq \sup\{ q_0(\theta) : \|\theta\|_2 = 1 \}$. Combining this with Theorem 4.1 and the monotone convergence theorem, it is not hard to see that $\lim_{t \to \infty} \bar{\gamma}(\theta(t))$ and $\lim_{t \to \infty} \gamma(\theta(t))$ exist and equal to the same value. Using a similar argument, we can draw the same conclusion for gradient descent.

To understand the implicit regularization effect, a natural question arises: what optimality property does the limit of normalized margin have? To this end, we identify a natural constrained optimization problem related to margin maximization, and prove that $\theta(t)$ directionally converges to its KKT points, as shown below. We note that we can extend this result to the finite time case, and show that gradient flow or gradient descent passes through an approximate KKT point after a certain amount of time. See Theorem A.9 in Appendix A and Theorem E.4 in Appendix E for the details. We will briefly review the definition of KKT points and approximate KKT points for a constraint optimization problem in Appendix C.1.

**Theorem 4.4** (Corollary of Theorem A.8 and E.3). For gradient flow under assumptions (A1) - (A4) or gradient descent under assumptions (S1), (A2) - (A4), (S5), any limit point $\bar{\theta}$ of $\{ \theta(t) : t \geq 0 \}$ is along the direction of a KKT point of the following constrained optimization problem (P):

$$\min \frac{1}{2} \|\theta\|^2_2 \quad \text{s.t.} \quad q_{\tilde{n}}(\theta) \geq 1 \quad \forall n \in [N]$$

That is, for any limit point $\bar{\theta}$, there exists a scaling factor $\alpha > 0$ such that $\alpha \bar{\theta}$ satisfies Karush-Kuhn-Tucker (KKT) conditions of (P).

Minimizing (P) over its feasible region is equivalent to maximizing the normalized margin over all possible directions. The proof is as follows. Note that we only need to consider all feasible points $\theta$ with $q_{\min}(\theta) > 0$. For a fixed $\theta$, $\alpha \theta$ is a feasible point of (P) iff $\alpha \geq q_{\min}(\theta)^{-1/L}$. Thus, the minimum objective value over all feasible points of (P) in the direction of $\theta$ is $\frac{1}{2} \|\theta/ q_{\min}(\theta)^{1/L}\|^2_2 = \frac{1}{2} \bar{\gamma}(\theta)^{-2/L}$. Taking minimum over all possible directions, we can conclude that if the maximum normalized margin is $\bar{\gamma}_s$, then the minimum objective of (P) is $\frac{1}{2} \bar{\gamma}_s^{-2/L}$.

It can be proved that (P) satisfies the Mangasarian-Fromovitz Constraint Qualification (MFCQ) (See Lemma C.7). Thus, KKT conditions are first-order necessary conditions for global optimality. For linear models, KKT conditions are also sufficient for ensuring global optimality; however, for deep homogeneous networks, $q_{\min}(\theta)$ can be highly non-convex. Indeed, as gradient descent is a first-order optimization method, if we do not make further assumptions on $q_{\min}(\theta)$, then it is easy to construct examples that gradient descent does not lead to a normalized margin that is globally optimal. Thus, proving the convergence to KKT points is perhaps the best we can hope for in our setting, and it is an interesting future work to prove stronger convergence results with further natural assumptions.

Moreover, we can prove the following corollary, which characterizes the optimality of the normalized margin using SVM with Neural Tangent Kernel (NTK, introduced in [Jacot et al., 2018]) defined at limit points. The proof is deferred to Appendix C.6.

**Corollary 4.5** (Corollary of Theorem A.4). Assume (S1). Then for gradient flow under assumptions (A2) - (A4) or gradient descent under assumptions (A2) - (A4), (S5), any limit point $\theta$ of $\{ \theta(t)/\|\theta(t)\|_2 : t \geq 0 \}$ is along the max-margin direction for the hard-margin SVM with kernel
\( K_\theta(x, x') = \langle \nabla \Phi_x(\bar{\theta}), \nabla \Phi_{x'}(\bar{\theta}) \rangle \), where \( \Phi_x(\theta) := \Phi(\theta; x) \). That is, for some \( \alpha > 0 \), \( \alpha \bar{\theta} \) is the optimal solution for the following constrained optimization problem:

\[
\min_{\bar{\theta}} \frac{1}{2} \| \theta \|_2^2 \quad \text{s.t.} \quad y_n \langle \theta, \nabla \Phi_{x_n}(\bar{\theta}) \rangle \geq 1 \quad \forall n \in [N]
\]

If we assume (A1) instead of (S1) for gradient flow, then there exists a mapping \( h(x) \in \partial^\circ \Phi_x(\bar{\theta}) \) such that the same conclusion holds for \( K_\theta(x, x') = \langle h(x), h(x') \rangle \).

4.4 Other Main Results

The above results can be extended to other settings as shown below.

Other Binary Classification Loss. The results on exponential loss can be generalized to a much broader class of binary classification loss. The class includes the logistic loss which is one of the most popular loss functions, \( \ell(q) = \log(1 + e^{-q}) \). The function class also includes other losses with exponential tail, e.g., \( \ell(q) = e^{-\gamma q}, \ell(q) = \log(1 + e^{-\gamma q}) \). For all those loss functions, we can use its inverse function \( \ell^{-1} \) to define the smoothed normalized margin as follows

\[
\tilde{\gamma}(\theta) = \frac{\ell^{-1}(L(\theta))}{\|\theta\|_2^2}.
\]

Then all our results for gradient flow continue to hold (Appendix A). Using a similar modification, we can also extend it to gradient descent (Appendix F).

Cross-entropy Loss. In multi-class classification, we can define \( q_n \) to be the difference between the classification score for the true label and the maximum score for the other labels, then the margin \( q_{\min} := \min_{n \in [N]} q_n \) and the normalized margin \( \tilde{\gamma}(\theta) := \frac{q_{\min}(\theta)}{\|\theta\|_2^2} \) can be similarly defined as before. In Appendix G, we define the smoothed normalized margin for cross-entropy loss to be the same as that for logistic loss (See Remark A.4). Then we show that Theorem 4.4 and Theorem 4.4 still hold (but with a slightly different definition of (P)) for gradient flow, and we also extend the results to gradient descent.

Multi-homogeneous Models. Some neural networks indeed possess a stronger property than homogeneity, which we call multi-homogeneity. For example, the output of a CNN (without bias terms) is 1-homogeneous with respect to the weights of each layer. In general, we say that a neural network \( \Phi(\theta; x) \) with \( \theta = (w_1, \ldots, w_m) \) is \((k_1, \ldots, k_m)\)-homogeneous if for any \( x \) and any \( c_1, \ldots, c_m > 0 \), we have \( \Phi(c_1 w_1, \ldots, c_m w_m; x) = \prod_{i=1}^m c_i^{k_i} \Phi(w_1, \ldots, w_m; x) \). In the previous example, an \( L \)-layer CNN with layer weights \( \theta = (w_1, \ldots, w_L) \) is \((1, \ldots, 1)\)-homogeneous.

One can easily see that that \((k_1, \ldots, k_m)\)-homogeneity implies \( L \)-homogeneity, where \( L = \sum_{i=1}^m k_i \), so our previous analysis for homogeneous models still applies to multi-homogeneous models. But it would be better to define the normalized margin for multi-homogeneous model as

\[
\tilde{\gamma}(w_1, \ldots, w_m) := q_{\min}\left( \frac{w_1}{\|w_1\|_2}, \ldots, \frac{w_m}{\|w_m\|_2} \right) = \frac{q_{\min}}{\prod_{i=1}^m \|w_i\|_2^{k_i}}.
\]

In this case, the smoothed approximation of \( \tilde{\gamma} \) for general binary classification loss (under some conditions) can be similarly defined for gradient flow:

\[
\tilde{\gamma}(w_1, \ldots, w_m) := \frac{\ell^{-1}(L)}{\prod_{i=1}^m \|w_i\|_2^{k_i}},
\]

It can be shown that \( \tilde{\gamma} \) is also non-decreasing during training when the loss is small enough (Appendix H). In the case of cross-entropy loss, we can still define \( \tilde{\gamma} \) by (5) while \( \ell(\cdot) \) is set to the logistic loss in the formula.

5 Proof Sketch: Gradient Flow on Homogeneous Model with Exponential Loss

In this section, we present a proof sketch in the case of gradient flow on homogeneous model with exponential loss to illustrate our proof ideas. Due to space limit, the proof for the main theorems on gradient flow and gradient descent in Section 4 are deferred to Appendix A and E respectively.
For convenience, we introduce a few more notations for a $L$-homogeneous neural network $\Phi(\theta, x)$. Let $S^{d-1} = \{ \theta \in \mathbb{R}^d : \|\theta\|_2 = 1 \}$ be the set of $L^2$-normalized parameters. Define $\rho := \|\theta\|_2$ and $\theta := \frac{\theta}{\|\theta\|_2} \in S^{d-1}$ to be the length and direction of $\theta$. For both gradient descent and gradient flow, $\theta$ is a function of time $t$. For convenience, we also view the functions of $\theta$, including $L(\theta), q_n(\theta), q_{\min}(\theta)$, as functions of $t$. So we can write $L(t) := L(\theta(t)), q_n(t) := q_n(\theta(t)), q_{\min}(t) := q_{\min}(\theta(t))$.

Lemma 5.1 below is the key lemma in our proof. It decomposes the growth of the smoothed normalized margin into the ratio of two quantities related to the radial and tangential velocity components of $\theta$ respectively. We will give a proof sketch for this later in this section. We believe that this lemma is of independent interest.

**Lemma 5.1** (Corollary of Lemma B.1). For a.e. $t > t_0$,

$$\frac{d}{dt} \log \rho > 0 \quad \text{and} \quad \frac{d}{dt} \log \tilde{\gamma} \geq L \left( \frac{d}{dt} \log \rho \right)^{-1} \| \frac{d\theta}{dt} \|_2^2.$$  

Using Lemma 5.1, the first two claims in Theorem 4.1 can be directly proved. For the third claim, we make use of the monotonicity of the margin to lower bound the gradient, and then show that $\rho \rightarrow +\infty$. Recall that $\tilde{\gamma}$ is an $O(\rho^{-L})$-additive approximation for $\gamma$. So this proves the third claim. We defer the detailed proof to Appendix B.

To show Theorem 4.4, we first change the time measure to log $\rho$, i.e., now we see $t$ as a function of log $\rho$. So the second inequality in Lemma 5.1 can be rewritten as $\frac{d}{d\log \rho} \log \tilde{\gamma} \geq L \| \frac{d\theta}{d\log \rho} \|_2^2$. Integrating on both sides and noting that $\tilde{\gamma}$ is upper-bounded, we know that there must be many instant $\gamma$ such that $\| \frac{d\theta}{d\log \rho} \|_2$ is small. By analyzing the landscape of training loss, we show that these points are “approximate” KKT points. Then we show that every convergent sub-sequence of $\{ \theta(t) : t \geq 0 \}$ can be modified to be a sequence of “approximate” KKT points which converges to the same limit. Then we conclude the proof by applying a theorem from [Dutta et al., 2013] to show that the limit of this convergent sequence of “approximate” KKT points is a KKT point. We defer the detailed proof to Appendix C.

Now we give a proof sketch for Lemma 5.1 in which we derive the formula of $\tilde{\gamma}$ step by step. In the proof, we obtain several clean close form formulas for several relevant quantities, by using the chain rule and Euler’s theorem for homogeneous functions extensively.

**Proof Sketch of Lemma 5.1** For ease of presentation, we ignore the regularity issues of taking derivatives in this proof sketch. We start from the equation $\frac{dL}{dt} = -\langle \partial^\rho L(\theta(t)), \frac{d\theta}{dt} \rangle = -\| \frac{d\theta}{dt} \|_2^2$ which follows from the chain rule (see also Lemma 1.3). Then we note that $\frac{d\theta}{dt}$ can be decomposed into two parts: the radial component $v := \theta^\top \frac{d\theta}{dt}$ and the tangent component $u := (I - \theta \theta^\top) \frac{d\theta}{dt}$.

The radial component is easier to analyze. By the chain rule, $\|v\|_2 = \theta^\top \frac{d\theta}{dt} = \frac{1}{\rho} \langle \theta, \frac{d\theta}{dt} \rangle = \frac{1}{\rho} \frac{1}{2} \frac{d\rho^2}{dt}$.

For $\frac{1}{2} \frac{d\rho^2}{dt}$, we have an exact formula:

$$\frac{1}{2} \frac{d\rho^2}{dt} = \langle \theta, \frac{d\theta}{dt} \rangle = \sum_{n=1}^{N} e^{-q_n} \partial^\rho q_n, \theta \rangle = L \sum_{n=1}^{N} e^{-q_n} q_n. \quad (6)$$

The last equality is due to $\langle \partial^\rho q_n, \theta \rangle = L q_n$ by homogeneity of $q_n$. This is sometimes called Euler’s theorem for homogeneous functions (see Theorem B.2). For differentiable $q_n$, it can be easily proved by taking the derivative over $c$ on both sides of $q_n(c \theta) = e^{L} q_n(\theta)$ and letting $c = 1$.

With (6), we can lower bound $\frac{1}{2} \frac{d\rho^2}{dt} \rho$ by

$$\frac{1}{2} \frac{d\rho^2}{dt} \geq L \sum_{n=1}^{N} e^{-q_n} q_n \geq L \sum_{n=1}^{N} e^{-q_n} q_{\min} \geq L \cdot L \cdot \log \frac{1}{L}, \quad (7)$$

where the last inequality uses the fact that $e^{-q_{\min}} \leq L$. (7) also implies that $\frac{1}{2} \frac{d\rho^2}{dt} > 0$ for $t > t_0$ since $L(t_0) < 1$ and $L$ is non-increasing. As $\frac{d}{dt} \log \rho = \frac{1}{2} \frac{d\rho^2}{dt}$, this also proves the first inequality of Lemma 5.1.
Now, we have $\|v\|^2 = \frac{1}{\rho^2} \left( \frac{1}{2} \frac{d\rho^2}{dt} \right)^2 = \frac{1}{2} \frac{d\rho^2}{dt} \cdot \frac{d}{dt} \log \rho$ on the one hand; on the other hand, by the chain rule we have

$$
-\frac{dL}{dt} = \frac{\|\hat{\theta}\|^2}{2} = \frac{1}{\rho^2} \left( \rho \frac{d\theta}{dt} - \frac{d\rho}{dt} \right) = \frac{1}{\rho^2} \left( \rho \frac{d\theta}{dt} - (\hat{\theta}^T \frac{d\theta}{dt}) \theta \right) = \frac{u}{\rho}.
$$

So we have

$$
-\frac{dL}{dt} = \frac{1}{2} \frac{d\rho^2}{dt} \cdot \frac{d}{dt} \log \rho + \rho^2 \frac{\|\hat{\theta}\|^2}{2}.
$$

Dividing $\frac{1}{2} \frac{d\rho^2}{dt}$ on the leftmost and rightmost sides, we have

$$
-\frac{dL}{dt} \cdot \left( \frac{1}{2} \frac{d\rho^2}{dt} \right)^{-1} = \frac{d}{dt} \log \rho + \left( \frac{d}{dt} \log \rho \right)^{-1} \frac{\|\hat{\theta}\|^2}{2}.
$$

By $-\frac{dL}{dt} \geq 0$ and (7), the LHS is no greater than $-\frac{dL}{dt} \cdot (L \cdot L \log \frac{1}{L})^{-1} = \frac{1}{L} \log \log \frac{1}{L}$. Thus we have

$$
\frac{d}{dt} \log \log \frac{1}{L} - L \frac{d}{dt} \log \rho \geq L \left( \frac{d}{dt} \log \rho \right)^{-1} \frac{\|\hat{\theta}\|^2}{2},
$$

where the LHS is exactly $\frac{d}{dt} \log \tilde{\gamma}$.

6 DISCUSSION AND FUTURE DIRECTIONS

In this paper, we analyze the dynamics of gradient flow/descent of homogeneous neural networks under a minimal set of assumptions. The main technical contribution of our work is to prove rigorously that for gradient flow/descent, the normalized margin is increasing and converges to a KKT point of a natural max-margin problem. Our results leads to some natural further questions:

- Can we generalize our results for gradient descent on smooth neural networks to non-smooth ones? In the smooth case, we can lower bound the decrement of training loss by the gradient norm squared, multiplied by a factor related to learning rate. However, in the non-smooth case, no such inequality is known in the optimization literature.

- Can we make more structural assumptions on the neural network to prove stronger results? In this work, we use a minimal set of assumptions to show that the convergent direction of parameters is a KKT point. A potential research direction is to identify more key properties of modern neural networks and show that the normalized margin at convergence is locally or globally optimal (in terms of optimizing (P)).

- Can we extend our results to neural networks with bias terms? In our experiments, the normalized margin of the CNN with bias also increases during training despite that its output is non-homogeneous. It is very interesting (and technically challenging) to provide a rigorous proof for this fact.

ACKNOWLEDGMENTS

The research is supported in part by the National Natural Science Foundation of China Grant 61822203, 61772297, 61632016, 61761146003 and the Zhongguancun Hahua Institute for Frontier Information Technology and Turing AI Institute of Nanjing. We thank Liwei Wang for helpful suggestions on the connection between margin and robustness. We thank Sanjeev Arora, Tianle Cai, Simon Du, Jason D. Lee, Zhiyuan Li, Tengyu Ma, Ruosong Wang for helpful discussions.

REFERENCES


A RESULTS FOR GENERAL LOSS

In this section, we state our results for a broad class of binary classification loss. A major consequence of this generalization is that the logistic loss, one of the most popular loss functions, \( \ell(q) = \log(1 + e^{-q}) \) is included. The function class also includes other losses with exponential tail, e.g., \( \ell(q) = e^{-q^q}, \ell(q) = \log(1 + e^{-q^q}) \).

A.1 ASSUMPTIONS

We first focus on gradient flow. We assume (A1), (A2) as we do for exponential loss. For (A3), (A4), we replace them with two weaker assumptions (B3), (B4). All the assumptions are listed below:

(A1). (Regularity). For any fixed \( x \), \( \Phi(\cdot; x) \) is locally Lipschitz and admits a chain rule;

(A2). (Homogeneity). There exists \( L > 0 \) such that \( \forall \alpha > 0 : \Phi(\alpha \theta; x) = \alpha \Phi(\theta; x) \);

(B3). The loss function \( \ell(q) \) can be expressed as \( \ell(q) = e^{-f(q)} \) such that

(B3.1). \( f : \mathbb{R} \rightarrow \mathbb{R} \) is \( C^1 \)-smooth.

(B3.2). \( f'(q) > 0 \) for all \( q \in \mathbb{R} \).

(B3.3). There exists \( b_f \geq 0 \) such that \( f'(q)q \) is non-decreasing for \( q \in (b_f, +\infty) \), and \( f'(q)q \rightarrow +\infty \) as \( q \rightarrow +\infty \).

(B3.4). Let \( g : [f(b_f), +\infty) \rightarrow [b_f, +\infty) \) be the inverse function of \( f \) on the domain \( [b_f, +\infty) \). There exists \( b_g \geq \max\{2f(b_f), f(2b_f)\} \), \( K \geq 1 \) such that \( g'(x) \leq Kg'(\theta x) \) and \( f'(y) \leq Kg'(\theta y) \) for all \( x \in (b_g, +\infty) \) and \( y \in (g(b_g), +\infty) \) and \( \theta \in [1/2, 1] \).

(B4). (Separability). There exists a time \( t_0 \) such that \( \mathcal{L}(t_0) < e^{-f(b_f)} = \ell(b_f) \).

(A1) and (A2) remain unchanged. (B3) is satisfied by exponential loss \( \ell(q) = e^{-q} \) (with \( f(q) = q \)) and logistic loss \( \ell(q) = \log(1 + e^{-q}) \) (with \( f(q) = -\log \log(1 + e^{-q}) \)). (B4) are essentially the same as (A4) but (B4) uses a threshold value that depends on the loss function. Assuming (B3), it is easy to see that (B4) ensures the separability of data since \( \ell(q_n) < e^{-f(b_f)} \) implies \( q_n > b_f \geq 0 \). For logistic loss, we can set \( b_f = 0 \) (see Remark A.2). So the corresponding threshold value in (B4) is \( \ell(0) = \log 2 \).

Now we discuss each of the assumptions in (B3). (B3.1) is a natural assumption on smoothness. (B3.2) requires \( \ell(\cdot) \) to be monotone decreasing, which is also natural since \( \ell(\cdot) \) is used for binary classification. The rest of two assumptions in (B3) characterize the properties of \( \ell'(q) \) when \( q \) is large enough. (B3.3) is an assumption that appears naturally from the proof. For exponential loss, \( f'(q)q = q \) is always non-decreasing, so we can set \( b_f = 0 \). In (B3.4), the inverse function \( g \) is defined. It is guaranteed by (B3.1) and (B3.2) that \( g \) always exists and \( g \) is also \( C^1 \)-smooth. Though (B3.4) looks very complicated, it essentially says that \( f'(\Theta(q)) = \Theta(f'(q)), g'(\Theta(q)) = \Theta(g'(q)) \) as \( q \rightarrow \infty \). (B3.4) is indeed a technical assumption that enables us to asymptotically compare the loss or the length of gradient at different data points. It is possible to base our results on weaker assumptions than (B3.4), but we use (B3.4) for simplicity since it has already been satisfied by many loss functions such as the aforementioned examples.

We summarize the corresponding \( f, g \) and \( b_f \) for exponential loss and logistic loss below:

**Remark A.1.** Exponential loss \( \ell(q) = e^{-q} \) satisfies (B3) with

\[
\begin{align*}
  f(q) &= q & f'(q) &= 1 & g(q) &= q & g'(q) &= 1 & b_f &= 0 & \ell(b_f) &= 1.
\end{align*}
\]

**Remark A.2.** Logistic loss \( \ell(q) = \log(1 + e^{-q}) \) satisfies (B3) with

\[
\begin{align*}
  f(q) &= -\log \log(1 + e^{-q}) = \Theta(q) & f'(q) &= \frac{e^{-q}}{(1 + e^{-q}) \log(1 + e^{-q})} = \Theta(1) \\
  g(q) &= -\log(e^{-q} - 1) = \Theta(q) & g'(q) &= \frac{e^{-q}}{e^{-q} - 1} = \Theta(1) \\
  b_f &= 0 & \ell(b_f) &= \log 2.
\end{align*}
\]

The proof for Remark A.1 is trivial. For Remark A.2 we give a proof below.
Proof for Remark A.2. By simple calculations, the formulas for \( f(q), f'(q), g(q), g'(q) \) are correct. (B3.1) is trivial. \( f'(q) = \frac{q}{(1+e^{\rho q}) \log(1+e^{-\rho q})} > 0 \), so (B3.2) is satisfied. For (B3.3), note that 
\[
\frac{d}{dq} \left( \frac{1}{1+e^{\rho q}} \log(1+e^{-\rho q}) \right) = e^{\rho q} \log(1+e^{-\rho q}) - e^{\rho q} \cdot e^{-\rho q} - 1 = 0.
\]
Thus, \( f'(q)q \) is a strictly increasing function on \( \mathbb{R} \). As \( b_f \) is required to be non-negative, we set \( b_f = 0 \). For proving that \( f'(q)q \rightarrow +\infty \) and (B4), we only need to notice that \( f'(q) \sim \frac{e^{-\rho q}}{1+e^{-\rho q}} = 1 \) and \( g'(x) = 1/f'(g(x)) \sim 1 \).

A.2 Smoothed Normalized Margin

For a loss function \( \ell(\cdot) \) satisfying (B3), it is easy to see from (B3.2) that its inverse function \( \ell^{-1}(\cdot) \) must exist. For this kind of loss functions, we define the smoothed normalized margin as follows:

Definition A.3. For a loss function \( \ell(\cdot) \) satisfying (B3), the smoothed normalized margin \( \tilde{\gamma}(\theta) \) of \( \theta \) is defined as
\[
\tilde{\gamma}(\theta) := \frac{\ell^{-1}(\mathcal{L})}{\rho^L} = \frac{g(\log \frac{1}{\rho^L})}{\rho^L} = \frac{g\left(-\log \left( \sum_{n=1}^{N} e^{-f(q_n(\theta))} \right) \right)}{\rho^L},
\]
where \( \ell^{-1}(\cdot) \) is the inverse function of \( \ell(\cdot) \) and \( \rho := ||\theta||_2 \).

Remark A.4. For logistic loss \( \ell(q) = \log(1+e^{-q}) \), \( \tilde{\gamma}(\theta) = \rho^{-L} \log \left( \frac{1}{\exp(L) - 1} \right) \); for exponential loss \( \ell(q) = e^{-q} \), \( \tilde{\gamma}(\theta) = \rho^{-L} \log \frac{1}{\theta} \), which is the same as (3).

Now we give some insights on how well \( \tilde{\gamma}(\theta) \) approximates \( \bar{\gamma}(\theta) \) using a similar argument as in Section 4.2. Using the LogSumExp function, the smoothed normalized margin \( \bar{\gamma}(\theta) \) can also be written as
\[
\bar{\gamma}(\theta) = g(-\text{LSE}(-f(q_1), \ldots, -f(q_N))) \frac{1}{\rho^L}.
\]
LSE is a (log \( N \))-additive approximation for max. So we can roughly approximate \( \bar{\gamma}(\theta) \) by
\[
\bar{\gamma}(\theta) \approx g\left(-\max\{ -f(q_1), \ldots, -f(q_N) \} \right) \frac{1}{\rho^L} = g\left( f(q_{\text{min}}) \right) \frac{1}{\rho^L} = \tilde{\gamma}(\theta).
\]
Note that (B3.3) is crucial to make the above approximation reasonable. Similar to exponential loss, we can show the following lemma asserting that \( \bar{\gamma} \) is a good approximation of \( \tilde{\gamma} \).

Lemma A.5. Assuming (B3)\footnote{Indeed, (B3.4) is not needed for showing Lemma A.5 and Theorem A.7.} we have the following properties about the margin:

(a) \( f(q_{\text{min}}) - \log N \leq \log \frac{1}{L} \leq f(q_{\text{min}}) \).

(b) If \( \log \frac{1}{L} > f(b_f) \), then there exists \( \xi \in \{ f(q_{\text{min}}) - \log N, f(q_{\text{min}}) \} \cap (b_f, +\infty) \) such that
\[
\bar{\gamma} \geq \frac{g'(\xi) \log N}{\rho^L} \leq \tilde{\gamma} \leq \bar{\gamma}.
\]

(c) For a sequence of parameters \( \{ \theta_m \in \mathbb{R}^d : m \in \mathbb{N} \} \), if \( \mathcal{L}(\theta_m) \rightarrow 0 \), then \( |\bar{\gamma}(\theta_m) - \tilde{\gamma}(\theta_m)| \rightarrow 0 \).

Proof. (a) can be easily deduced from \( e^{-f(q_{\text{min}})} \leq L \leq Ne^{-f(q_{\text{min}})} \). Combining (a) and the monotonicity of \( g(\cdot) \), we further have \( g(s) \leq g(\log \frac{1}{L}) \leq q_{\text{min}} \) for \( s := \max\{ f(b_f), f(q_{\text{min}}) - \log N \} \). By the mean value Theorem, there exists \( \xi \in (s, f(q_{\text{min}})) \) such that \( g(s) = g(f(q_{\text{min}})) - g'(\xi)(f(q_{\text{min}}) - s) \geq q_{\text{min}} - g'(\xi) \log N \). Dividing \( \rho^L \) on each side of \( q_{\text{min}} - g'(\xi) \log N \leq g(\log \frac{1}{L}) \leq q_{\text{min}} \) proves (b).

Now we prove (c). Without loss of generality, we assume \( \log \frac{1}{L} > f(b_f) \) for all \( \theta_m \). It follows from (b) that for every \( \theta_m \), there exists \( \xi_m \in (f(q_{\text{min}}(\theta_m)) - \log N, f(q_{\text{min}}(\theta_m))) \cap (b_f, +\infty) \) such that
\[
\bar{\gamma}(\theta_m) - (g'(\xi_m) \log N)/\rho(\theta_m)^L \leq \tilde{\gamma}(\theta_m) \leq \bar{\gamma}(\theta_m).
\]
Note that $\xi_m \geq f(q_{\text{min}}(\theta_m)) - \log N \geq \log \frac{1}{\ell(\theta_m)} - \log N \to +\infty$. So $\frac{g(\xi_m)}{g(\xi_m)} = f'(g(\xi_m))g(\xi_m) \to +\infty$ by (B.3.3). Also note that there exists a constant $B_0$ such that $\tilde{\gamma}(\theta_m) \leq B_0$ for all $m$ since $\tilde{\gamma}$ is continuous on the unit sphere $S^{d-1}$. So we have

$$\frac{g'(\xi_m)}{\rho(\theta_m)} \leq \frac{g'(\xi_m) q_{\text{min}}(\theta_m)}{g(\xi_m) \rho(\theta_m)} \leq \frac{g'(\xi_m)}{g(\xi_m)} \cdot \tilde{\gamma}(\theta_m) \leq \frac{g'(\xi_m)}{g(\xi_m)} \cdot B \to 0,$$

where the first inequality follows since $\xi_m \leq f(q_{\text{min}}(\theta_m))$. Together with [8], we have $|\tilde{\gamma}(\theta_m) - \tilde{\gamma}(\theta_m)| \to 0$. □

**Remark A.6.** For exponential loss, we have already shown in Section 4.2 that $\tilde{\gamma}(\theta)$ is an $O(\rho^{-L})$-additive approximation for $\tilde{\gamma}(\theta)$. For logistic loss, it follows easily from $g'(q) = \Theta(1)$ and (b) of Lemma A.5 that $\tilde{\gamma}(\theta)$ is an $O(\rho^{-L})$-additive approximation for $\tilde{\gamma}(\theta)$ if $L$ is sufficiently small.

### A.3 Theorems

Now we state our main theorems. For the monotonicity of the normalized margin, we have the following theorem. The proof is provided in Appendix [B].

**Theorem A.7.** Under assumptions (A1), (A2), (B3), (B4), the following statements are true for gradient flow:

1. For $t > t_0$, $\frac{d}{dt} \tilde{\gamma}(\theta(t)) \geq 0$;
2. For $t > t_0$, either $\frac{d}{dt} \tilde{\gamma}(\theta(t)) > 0$ or $\frac{d}{dt} \frac{\theta(t)}{\||\theta(t)||_2} = 0$;
3. $L(\theta(t)) \to 0$ and $\||\theta(t)||_2 \to \infty$ as $t \to +\infty$; therefore, $|\tilde{\gamma}(\theta(t)) - \tilde{\gamma}(\theta(t))| \to 0$.

For the normalized margin at convergence, we have two theorems, one for infinite-time limiting case, and the other being a finite-time quantitative result. Their proofs can be found in Appendix [C]. As in the exponential loss case, we define the constrained optimization problem (P) as follows:

$$\min \frac{1}{2} \||\theta||_2^2 \quad \text{s.t.} \quad q_{\text{min}}(\theta) \geq 1 \quad \forall n \in [N]$$

First, we show the directional convergence of $\theta(t)$ to a KKT point of (P).

**Theorem A.8.** Consider gradient flow under assumptions (A1), (A2), (B3), (B4). For every limit point $\bar{\theta}$ of $\{\theta(t) : t \geq 0\}$, $\bar{\theta} = \theta_{q_{\text{min}}(\theta)}^{1/L}$ is a KKT point of (P).

Second, we show that after finite time, gradient flow can pass through an approximate KKT point.

**Theorem A.9.** Consider gradient flow under assumptions (A1), (A2), (B3), (B4). For any $\epsilon, \delta > 0$, there exists $r := \Theta(\log \delta^{-1})$ and $\Delta := \Theta(\epsilon^{-2})$ such that $\bar{\theta} = \theta_{q_{\text{min}}(\theta)}^{1/L}$ is an $(\epsilon, \delta)$-KKT point at some time $t_*$ satisfying $\log \||\theta(t_*)||_2 \in (r, r + \Delta)$.

For the definitions for KKT points and approximate KKT points, we refer the readers to Appendix [D] for more details.

With a refined analysis, we can also provide tight rates for loss convergence and weight growth. The proof is given in Appendix [E].

**Theorem A.10.** Under assumptions (A1), (A2), (B3), (B4), we have the following tight rates for loss convergence and weight growth:

$$L(\theta(t)) = \Theta \left( \frac{g(\log t)^{2/L}}{t(\log t)^2} \right) \quad \text{and} \quad \||\theta(t)||_2 = \Theta(g(\log t)^{1/L}).$$

Applying Theorem A.10 to exponential loss and logistic loss, in which $g(x) = \Theta(x)$, we have the following corollary.

**Corollary A.11.** If $\ell(\cdot)$ is the exponential or logistic loss, then,

$$L(\theta(t)) = \Theta \left( \frac{1}{t(\log t)^{2-2/L}} \right) \quad \text{and} \quad \||\theta(t)||_2 = \Theta((\log t)^{1/L}).$$
\section{Margin Monotonicity for General Loss}

In this section, we consider gradient flow and prove Theorem A.7. We assume (A1), (A2), (B3), (B4) as mentioned in Appendix A.

We follow the notations in Section 5 to define \( \rho := \|\theta\|_2 \) and \( \hat{\theta} := \frac{\theta}{\|\theta\|_2} \in S^{d-1} \), and sometimes we view the functions of \( \theta \) as functions of \( t \).

\subsection{Proof for Proposition 1 and 2}

To prove the first two propositions, we generalize our key lemma (Lemma 5.1) to general loss.

\begin{lemma}
\textbf{Lemma B.1.} For \( \hat{\gamma} \) defined in Definition A.3, the following holds for all \( t > t_0 \),
\[ \frac{d}{dt} \log \rho > 0 \quad \text{and} \quad \frac{d}{dt} \log \hat{\gamma} \geq L \left( \frac{d}{dt} \log \rho \right)^{-1} \left\| \frac{d\theta}{dt} \right\|_2. \] \end{lemma}

Before proving Lemma B.1, we review two important properties of homogeneous functions. Note that these two properties are usually shown for smooth functions. By considering Clarke’s subdifferential, we can generalize it to locally Lipschitz functions that admit chain rules:

\begin{theorem}
\textbf{Theorem B.2.} Let \( F : \mathbb{R}^d \rightarrow \mathbb{R} \) be a locally Lipschitz function that admits a chain rule. If \( F \) is \( k \)-homogeneous, then
\begin{itemize}
\item[(a)] For all \( x \in \mathbb{R}^d \) and \( \alpha > 0 \),
\[ \partial^\alpha F(\alpha x) = \alpha^{k-1} \partial^\alpha F(x) \]
That is, \( \partial^\alpha F(\alpha x) = \{\alpha^{k-1} h : h \in \partial^\alpha F(x)\} \).
\item[(b)] (Euler’s Theorem for Homogeneous Functions). For all \( x \in \mathbb{R}^d \),
\[ \langle x, \partial^\alpha F(x) \rangle = k \cdot F(x) \]
That is, \( \langle x, h \rangle = k \cdot F(x) \) for all \( h \in \partial^\alpha F(x) \).
\end{itemize}
\end{theorem}

\begin{proof}
Let \( D \) be the set of points \( x \) such that \( F \) is differentiable at \( x \). According to the definition of Clarke’s subdifferential, for proving (a), it is sufficient to show that
\[ \{ k \rightarrow \infty \} \nabla F(\alpha x_k) : x_k \rightarrow x, \alpha x_k \in D \} = \{ \alpha^{k-1} \lim_{k \rightarrow \infty} \nabla F(x_k) : x_k \rightarrow x, x_k \in D \} \] (10)
Fix \( x_k \in D \). Let \( U \) be a neighborhood of \( x_k \). By definition of homogeneity, for any \( h \in \mathbb{R}^d \) and any \( y \in U \setminus \{x_k\} \),
\[ \frac{\nabla F(\alpha y) - \nabla F(\alpha x_k) - \langle \alpha y - \alpha x_k, \alpha^{k-1} h \rangle}{\|\alpha y - \alpha x_k\|_2} = \alpha^{k-1} \frac{\nabla F(\alpha y) - \nabla F(x_k) - \langle y - x_k, h \rangle}{\|y - x_k\|_2}. \]
Taking limits \( y \rightarrow x_k \) on both sides, we know that the LHS converges to 0 iff the RHS converges to 0. Then by definition of differentiability and gradient, \( F \) is differentiable at \( \alpha x_k \) iff it is differentiable at \( x_k \), and \( \nabla F(\alpha x_k) = \alpha^{k-1}h \) iff \( \nabla F(x_k) = h \). This proves (10).

To prove (b), we fix \( x \in \mathbb{R}^d \). Let \( z : [0, +\infty) \rightarrow \mathbb{R}^d, \alpha \rightarrow \alpha x \) be an arc. By definition of homogeneity, \( (F \circ z)(\alpha) = \alpha^k F(x) \) for \( \alpha > 0 \). Taking derivative with respect to \( \alpha \) on both sides (for differentiable points), we have
\[ \forall h \in \partial^\alpha F(\alpha x) : \langle x, h \rangle = k \alpha^{k-1} F(x) \] (11)
holds for a.e. \( \alpha > 0 \). Pick an arbitrary \( \alpha > 0 \) making (11) hold. Then by (a), (11) is equivalent to
\[ \forall h \in \partial^\alpha F(x) : \langle x, h \rangle = k \alpha^{k-1} F(x) \]
which proves (b).
\end{proof}

Applying Theorem B.2 to homogeneous neural networks, we have the following corollary:

\begin{corollary}
\textbf{Corollary B.3.} Under the assumptions (A1) and (A2), for any \( \theta \in \mathbb{R}^d \) and \( x \in \mathbb{R}^d \),
\[ \langle \theta, \partial^\alpha \Phi_x(\theta) \rangle = L \cdot \Phi_x(\theta), \]
where \( \Phi_x(\theta) = \Phi(\theta; x) \) is the network output for a fixed input \( x \).
Corollary [B.3] can be used to derive an exact formula for the weight growth during training.

**Theorem B.4.** For a.e. $t \geq 0$,

$$\frac{1}{2} \frac{d \rho^2}{dt} = L \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n)q_n.$$ 

**Proof.** The proof idea is to use Corollary [B.3] and chain rules (See Appendix I for chain rules in Clarke’s sense). Applying the chain rule on $t \mapsto \rho^2 = \|\theta\|^2_2$ yields $\frac{1}{2} \frac{d \rho^2}{dt} = -\langle \theta, h \rangle$ for all $h \in \partial_\rho L$ and a.e. $t > 0$. Then applying the chain rule on $\theta \mapsto L$, we have

$$-\partial_\rho L \subseteq \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n)\partial_\rho q_n = \left\{ \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n)h_n : h_n \in \partial_\rho q_n \right\}.$$ 

By Corollary [B.3] $\langle \theta, h_n \rangle = Lq_n$, and thus $\frac{1}{2} \frac{d \rho^2}{dt} = L \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n)q_n$. 

For convenience, we define $\nu(t) := \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n)q_n$, then Theorem B.4 can be rephrased as $\frac{1}{2} \frac{d \rho^2}{dt} = L \nu(t)$ for a.e. $t \geq 0$.

**Lemma B.5.** For all $t > t_0$,

$$\nu(t) \geq \frac{g(\log \frac{1}{L})}{g'(\log \frac{1}{L})} L.$$ 

**Proof.** By Lemma A.5, $q_n \geq g(\log \frac{1}{L})$ for all $n \in [N]$. Then by Assumption (B3),

$$f'(q_n)q_n \geq f'(g(\log \frac{1}{L})) \cdot g(\log \frac{1}{L}) = \frac{g(\log \frac{1}{L})}{L}.$$ 

Combining this with the definitions of $\nu(t)$ and $L$ gives

$$\nu(t) = \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n)q_n \geq \sum_{n=1}^{N} e^{-f(q_n)} \frac{g(\log \frac{1}{L})}{g'(\log \frac{1}{L})} = \frac{g(\log \frac{1}{L})}{g'(\log \frac{1}{L})} L.$$ 

**Proof for Lemma B.5**. Note that $\frac{d}{dt} \log \rho = \frac{1}{\rho^2} \frac{d \rho^2}{dt} = L \nu(t)$ by Theorem B.4. Then it simply follows from Lemma B.5 that $\frac{d}{dt} \log \rho > 0$ for a.e. $t > t_0$. For the second inequality, we first prove that $\log \tilde{\gamma} = \log \left( L^{-1}(t) \right)/\rho^2$ exists for all $t \geq t_0$. $L(t)$ is non-increasing with $t$. So $L(t) < e^{-f(b(t))}$ for all $t \geq t_0$. This implies that (1) $\log \frac{1}{L}$ is always in the domain of $g$; (2) $\rho > 0$ (otherwise $\rho \geq Ne^{-f(0)} > e^{-f(b(t))}$, contradicting (B4)). Therefore, $\tilde{\gamma} := g(\log \frac{1}{L})/\rho^2$ exists and is always positive for all $t \geq t_0$, which proves the existence of $\log \tilde{\gamma}$.

By the chain rule and Lemma B.5, we have

$$\frac{d}{dt} \log \tilde{\gamma} = \frac{d}{dt} \left( \log \left( g(\log \frac{1}{L}) \right) - L \log \rho \right) = \frac{g'(\log \frac{1}{L})}{g(\log \frac{1}{L})} \cdot \frac{1}{L} \cdot \frac{d L}{dt} - L^2 \cdot \frac{\nu(t)}{\rho^2} \geq \frac{1}{\nu(t)} \left( -\frac{d L}{dt} - L^2 \cdot \frac{\nu(t)}{\rho^2} \right).$$ 

On the one hand, $-\frac{d L}{dt} = \|d\theta\|^2_2$ for a.e. $t > 0$ by Lemma I.3 on the other hand, $L \nu(t) = \langle \theta, d\theta \rangle$ by Theorem B.4. Combining these together yields

$$\frac{d}{dt} \log \tilde{\gamma} \geq \frac{1}{\nu(t)} \left( \|d\theta\|^2_2 - \langle \theta, d\theta \rangle \right)^2 = \frac{1}{\nu(t)} \left( \|I - \theta\theta^\top \|d\theta\|^2_2 \right).$$
Lemma B.6. For all 

Proof for Lemma B.6. By Lemma I.3 and Theorem B.4, we follow the notations in Section 5 to define \((A1), (A2), (B3), (B4)\) as mentioned in Section A.

Therefore, \(\mathcal{L}(t) \rightarrow 0\) and \(\rho(t) \rightarrow +\infty\) as \(t \rightarrow \infty\).

Proof for Lemma B.6 By Lemma I.3 and Theorem B.4

Using Lemma B.5 to lower bound \(\nu\) and replacing \(\rho\) with \((g(\log 1/\bar{\gamma}))^{1/L}\) by the definition of \(\bar{\gamma}\), we have

Integrating on both sides from \(t_0\) to \(t\), we can conclude that

Note that \(1/\mathcal{L}\) is non-decreasing. If \(1/\mathcal{L}\) does not grow to \(+\infty\), then neither does \(G(1/\mathcal{L})\). But the RHS grows to \(+\infty\), which leads to a contradiction. So \(\mathcal{L} \rightarrow 0\).

To make \(\mathcal{L} \rightarrow 0\), \(q_{\min}\) must converge to \(+\infty\). So \(\rho \rightarrow +\infty\).

C CONVERGENCE TO THE MAX-MARGIN SOLUTION

In this section, we analyze the convergent direction of \(\theta\) and prove Theorem A.8 and A.9 assuming (A1), (A2), (B3), (B4) as mentioned in Section A.

We follow the notations in Section 5 to define \(\rho := \|\theta\|_2\) and \(\hat{\theta} := \frac{\theta}{\|\theta\|_2} \in S^{d-1}\), and sometimes we view the functions of \(\theta\) as functions of \(t\).

C.1 Preliminaries for KKT Conditions

We first review the definition of Karush-Kuhn-Tucker (KKT) conditions for non-smooth optimization problems following from [Dutta et al., 2013].

Consider the following optimization problem \((P)\) for \(x \in \mathbb{R}^d\):

\[
\begin{aligned}
\min & \quad f(x) \\
\text{s.t.} & \quad g_n(x) \leq 0 \quad \forall n \in [N]
\end{aligned}
\]

where \(f, g_1, \ldots, g_n : \mathbb{R}^d \rightarrow \mathbb{R}\) are locally Lipschitz functions. We say that \(x \in \mathbb{R}^d\) is a feasible point of \((P)\) if \(x\) satisfies \(g_n(x) \leq 0\) for all \(n \in [N]\).
Definition C.1 (KKT Point). A feasible point $x$ of (P) is a KKT point if $x$ satisfies KKT conditions: there exists $\lambda_1, \ldots, \lambda_N \geq 0$ such that

1. $0 \in \partial^o f(x) + \sum_{n \in [N]} \lambda_n \partial^o g_n(x)$;
2. $\forall n \in [N]: \lambda_n g_n(x) = 0$.

It is important to note that a global minimum of (P) may not be a KKT point, but under some regularity assumptions, the KKT conditions become a necessary condition for global optimality. The regularity condition we shall use in this paper is the non-smooth version of Mangasarian-Fromovitz Constraint Qualification (MFCQ) (see, e.g., the constraint qualification (C.Q.5) in [Giorgi et al., 2004]):

Definition C.2 (MFCQ). For a feasible point $x$ of (P), (P) is said to satisfy MFCQ at $x$ if there exists $v \in \mathbb{R}^d$ such that for all $n \in [N]$ with $g_n(x) = 0$,

$$\forall h \in \partial^o g_n(x) : \langle h, v \rangle > 0.$$ 

Following from [Dutta et al., 2013], we define an approximate version of KKT point, as shown below. Note that this definition is essentially the modified $\epsilon$-KKT point defined in their paper, but these two definitions differ in the following two ways: (1) First, in their paper, the subdifferential is allowed to be evaluated in a neighborhood of $x$, so our definition is slightly stronger; (2) Second, their paper fixes $\delta = \epsilon^2$, but in our definition we make them independent.

Definition C.3 (Approximate KKT Point). For $\epsilon, \delta > 0$, a feasible point $x$ of (P) is an $(\epsilon, \delta)$-KKT point if there exists $\lambda_n \geq 0, k \in \partial^o f(x), h_n \in \partial^o g_n(x)$ for all $n \in [N]$ such that

1. $\left\Vert k + \sum_{n \in [N]} \lambda_n h_n(x) \right\Vert_2 \leq \epsilon$;
2. $\forall n \in [N]: \lambda_n g_n(x) \geq -\delta$.

As shown in [Dutta et al., 2013], $(\epsilon, \delta)$-KKT point is an approximate version of KKT point in the sense that a series of $(\epsilon, \delta)$-KKT points can converge to a KKT point. We restate their theorem in our setting:

Theorem C.4 (Corollary of Theorem 3.6 in [Dutta et al., 2013]). Let $\{x_k \in \mathbb{R}^d : k \in \mathbb{N}\}$ be a sequence of feasible points of (P), $\{\epsilon_k > 0 : k \in \mathbb{N}\}$ and $\{\delta_k > 0 : k \in \mathbb{N}\}$ be two sequences. $x_k$ is an $(\epsilon_k, \delta_k)$-KKT point for every $k$, and $\epsilon_k \to 0, \delta_k \to 0$. If $x_k \to x$ as $k \to +\infty$ and MFCQ holds at $x$, then $x$ is a KKT point of (P).

C.2 KKT CONDITIONS FOR (P)

Recall that for a homogeneous neural network, the optimization problem (P) is defined as follows:

$$\min \frac{1}{2} \Vert \theta \Vert_2^2 \quad \text{s.t.} \quad q_n(\theta) \geq 1 \quad \forall n \in [N]$$

Using the terminologies and notations in Appendix [C.1] the objective and constraints are $f(x) = \frac{1}{2} \Vert x \Vert_2^2$ and $g_n(x) = 1 - q_n(x)$. The KKT points and approximate KKT points for (P) are defined as follows:

Definition C.5 (KKT Point of (P)). A feasible point $\theta$ of (P) is a KKT point if there exist $\lambda_1, \ldots, \lambda_N \geq 0$ such that

1. $\theta - \sum_{n=1}^N \lambda_n h_n = 0$ for some $h_1, \ldots, h_N$ satisfying $h_n \in \partial^o q_n(\theta)$;
2. $\forall n \in [N] : \lambda_n (q_n(\theta) - 1) = 0$.

Definition C.6 (Approximate KKT Point of (P)). A feasible point $\theta$ of (P) is an $(\epsilon, \delta)$-KKT point of (P) if there exists $\lambda_1, \ldots, \lambda_N \geq 0$ such that

1. $\left\Vert \theta - \sum_{n=1}^N \lambda_n h_n \right\Vert_2 \leq \epsilon$ for some $h_1, \ldots, h_N$ satisfying $h_n \in \partial^o q_n(\theta)$;
2. \( \forall n \in [N] : \lambda_n(q_n(\theta)) - 1 \leq \delta. \)

By the homogeneity of \( q_n \), it is easy to see that (P) satisfies MFCQ, and thus KKT conditions are first-order necessary condition for global optimality.

**Lemma C.7.** (P) satisfies MFCQ at every feasible point \( \theta \).

**Proof.** Take \( v := \theta \). For all \( n \in [N] \) satisfying \( q_n = 1 \), by homogeneity of \( q_n \),

\[
(v, h) = Lq_n(\theta) = L > 0
\]

holds for any \( h \in -\partial^c q_n(\theta) = \partial^c (1 - q_n(\theta)) \).

---

**C.3 KEY LEMMAS**

Define \( \beta(t) := \frac{1}{\|\theta\|_2} \left\langle \hat{\theta}, \frac{\partial \theta}{\partial t} \right\rangle \) to be the cosine of the angle between \( \theta \) and \( \frac{\partial \theta}{\partial t} \). Here \( \beta(t) \) is only defined for a.e. \( t > 0 \). Since \( q_n \) is locally Lipschitz, it can be shown that \( q_n \) is (globally) Lipschitz on the compact set \( S^{d-1} \), which is the unit sphere in \( \mathbb{R}^d \). Define

\[
B_0 := \sup \left\{ \frac{q_n}{\rho_n} : \theta \in \mathbb{R}^d \setminus \{0\} \right\} = \sup \{ q_n : \theta \in S^{d-1} \} < \infty.
\]

\[
B_1 := \sup \left\{ \frac{\|h\|_2}{\rho^{L-1}} : \theta \in \mathbb{R}^d \setminus \{0\}, h \in \partial^c q_n, n \in [N] \right\} = \sup \{ \|h\|_2 : \theta \in S^{d-1}, h \in \partial^c q_n, n \in [N] \} < \infty.
\]

For showing Theorem A.8 and Theorem A.9, we first prove Lemma C.8. In light of this lemma, if we aim to show that \( \theta \) is along the direction of an approximate KKT point, we only need to show \( \beta \to 1 \) (which makes \( \epsilon \to 0 \)) and \( \mathcal{L} \to 0 \) (which makes \( \delta \to 0 \)).

**Lemma C.8.** Let \( C_1, C_2 \) be two constants defined as

\[
C_1 := \frac{\sqrt{2}}{\tilde{\gamma}(t_0)^{1/L}}, \quad C_2 := \frac{2eNK^2}{L\tilde{\gamma}(t_0)^{2/L}} \left( \frac{B_1}{\tilde{\gamma}(t_0)} \right)^{\log_2 K},
\]

where \( K, b \) are constants specified in (B3.4). If \( \log \frac{1}{2} \geq b \) at time \( t > t_0 \), then \( \Theta := \theta / q_{\min}(\theta)^{1/L} \) is an \((\epsilon, \delta)\)-KKT point of (P), where \( \epsilon := C_1\sqrt{1 - \beta} \) and \( \delta := C_2/\log \frac{1}{2} \).

**Proof.** Let \( h(t) := \frac{\partial h}{\partial t} / \|h\|_2 \) for a.e. \( t > 0 \). By the chain rule, there exist \( h_1, \ldots, h_N \) such that

\[
h_n \in \partial^c q_n, \quad \text{and} \quad h = \sum_{n \in [N]} e^{-f(q_n)} f(q_n) h_n.
\]

Let \( \tilde{h}_n := h_n / q_{\min}(\theta)^{1/L} \in \partial^c q_n(\tilde{\theta}) \) (Recall that \( \tilde{\theta} := \theta / q_{\min}(\theta)^{1/L} \)). Construct \( \lambda_n := q_{\min}^{-1/L} \rho \cdot e^{-f(q_n)} / \|h\|_2 \). Now we only need to show

\[
\left\| \tilde{\theta} - \sum_{n=1}^N \lambda_n \tilde{h}_n \right\|_2^2 \leq 2 \frac{K}{\gamma} (1 - \beta)
\]

\[
\sum_{n=1}^N \lambda_n (q_n(\tilde{\theta}) - 1) \leq \left( \frac{2eNK^2}{L\tilde{\gamma}^{2/L}} \left( \frac{B_1}{\tilde{\gamma}} \right)^{\log_2 K} \right) \frac{1}{f(\tilde{\gamma}\rho^{t_0})}.
\]

Then \( \tilde{\theta} \) can be shown to be an \((\epsilon, \delta)\)-KKT point by the monotonicity \( \tilde{\gamma}(t) \geq \tilde{\gamma}(t_0) \) for \( t > t_0 \).

**Proof of (12).** From our construction, \( \sum_{n=1}^N \lambda_n \tilde{h}_n = q_{\min}^{-1/L} \rho \|h\|_2 \). So

\[
\left\| \tilde{\theta} - \sum_{n=1}^N \lambda_n \tilde{h}_n \right\|_2^2 = q_{\min}^{-2/L} \rho^2 \left\| \tilde{\theta} - \frac{h}{\|h\|_2} \right\|_2^2 = q_{\min}^{-2/L} \rho^2 (2 - 2\beta) \leq 2 \frac{K}{\gamma} (1 - \beta),
\]

where the last equality is by Lemma A.5.
Proof for (13). According to our construction,  
\[ \sum_{n=1}^{N} \lambda_n(q_n(\theta) - 1) = \frac{q_{\min}^2}{\|h\|_2^2} \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n)(q_n - q_{\min}). \]

Note that \( \|h\|_2 \geq \langle h, \hat{\theta} \rangle = L\nu/\rho \). By Lemma B.5 and Lemma D.1 we have  
\[ \nu \geq g(\log \frac{1}{L}) \mathcal{L} \geq \frac{1}{2K} \mathcal{L} \geq \frac{1}{2K} f(\gamma \rho^L)e^{-f(q_n)}, \]

where the last inequality uses \( f(\gamma \rho^L) = \log \frac{1}{L} \) and \( \mathcal{L} \geq e^{-f(q_n)} \). Combining these gives  
\[ \sum_{n=1}^{N} \lambda_n(q_n(\theta) - 1) \leq \frac{2Kq_{\min}^{-2/L} \rho^2}{L f(\gamma \rho^L)} \sum_{n=1}^{N} e^{-f(q_n) - f(q_n)} f'(q_n)(q_n - q_{\min}). \]

If \( q_n > q_{\min} \), then by the mean value theorem there exists \( \xi_n \in (q_{\min}, q_n) \) such that \( f(q_n) - f(q_{\min}) = f'() (\xi_n (q_n - q_{\min}) \). By Assumption (B.3.4), we know that \( f'(q_n) \leq K[\log_2(q_n/\xi_n)] f'(\xi_n) \). Note that \( [\log_2(q_n/\xi_n)] \leq \log_2(2B_1 \rho^L / q_{\min}) \leq \log_2(2B_1 / \gamma) \). Then we have  
\[ \sum_{n=1}^{N} \lambda_n(q_n(\theta) - 1) \leq \frac{2Kq_{\min}^{-2/L} \rho^2}{L f(\gamma \rho^L)} K^{\log_2(2B_1 / \gamma)} \sum_{n: q_n \neq q_{\min}} e^{-f(\xi_n) \xi_n (q_n - q_{\min}) f'(\xi_n)(q_n - q_{\min})} \]

where the second inequality uses \( q_{\min}^{-2/L} \rho^2 \leq \gamma^{-2/L} \) by Lemma A.5 and the fact that the function \( x \mapsto e^{-x} x \) on \((0, +\infty)\) attains the maximum value \( e \) at \( x = 1 \). \( \square \)

By Theorem A.7 we have already known that \( \mathcal{L} \to 0 \). So it remains to bound \( \beta(t) \). For this, we first prove the following lemma to bound the integral of \( \beta(t) \).

Lemma C.9. For all \( t_2 > t_1 \geq t_0 \),  
\[ \int_{t_1}^{t_2} \left( \beta(\tau)^{-2} - 1 \right) \cdot \frac{d}{d\tau} \log \rho(\tau) \cdot d\tau \leq \frac{1}{L} \log \frac{\gamma(t_2)}{\gamma(t_1)}. \]

Proof. By Lemma B.4 \( \frac{d}{dt} \log \rho = \frac{1}{\rho^2} \frac{d^2 \rho}{dx^2} = \frac{L \nu}{\rho} \) for a.e. \( t > 0 \). By Lemma B.1 for a.e. \( t \in (t_1, t_2) \),  
\[ \frac{d}{dt} \log \gamma \geq L \cdot \left\| \frac{\rho^2}{L \nu} \frac{d\theta}{dt} \right\|_2^2 \cdot \frac{d}{dt} \log \rho. \]  
(14)

By the chain rule, \( \frac{d^2 \theta}{dt^2} = \frac{1}{L}(I - \hat{\theta}^\top) \frac{d\theta}{dt} \) So we have  
\[ \left\| \frac{\rho^2}{L \nu} \frac{d\theta}{dt} \right\|_2^2 = \left\| \frac{\rho}{L \nu(t)} (I - \hat{\theta}^\top) \frac{d\theta}{dt} \right\|_2^2 = \left\| \frac{d\theta}{dt} \right\|^2 - \left\langle \frac{d\theta}{dt}, \hat{\theta} \right\rangle^2 = \beta^{-2} - 1. \]
(15)

where the last equality follows from the definition of \( \beta \). Combining (14) and (15) we have  
\[ \frac{d}{dt} \log \gamma \geq L \left( \beta^{-2} - 1 \right) \cdot \frac{d}{dt} \log \rho. \]

Integrating on both sides from \( t_1 \) to \( t_2 \) proves the lemma. \( \square \)

A direct corollary of Lemma C.9 is the upper bound for the minimum \( \beta^2 - 1 \) within a time interval:

Corollary C.10. For all \( t_2 > t_1 \geq t_0 \), then there exists \( t_* \in (t_1, t_2) \) such that  
\[ \beta(t_*)^{-2} - 1 \leq \frac{1}{L} \cdot \frac{\log \gamma(t_2) - \log \gamma(t_1)}{\log \rho(t_2) - \log \rho(t_1)}. \]
Proof. Denote the RHS as $C$. Assume to the contrary that $\beta(\tau)^{-2} - 1 > C$ for a.e. $\tau \in (t_1, t_2)$. By Lemma C.9, $\log \rho(\tau) > 0$ for a.e. $\tau \in (t_1, t_2)$. Then by Lemma C.8 we have
\[
\frac{1}{L} \log \frac{\hat{g}(t_2)}{\hat{g}(t_1)} > \int_{t_1}^{t_2} C \cdot \frac{d}{d\tau} \log \rho(\tau) \cdot d\tau = C \cdot (\log \rho(t_2) - \log \rho(t_1)) = \frac{1}{L} \log \frac{\hat{g}(t_2)}{\hat{g}(t_1)},
\]
which leads to a contradiction. \hfill \-box

In the rest of this section, we present both asymptotic and non-asymptotic analyses for the directional convergence by using Corollary C.10 to bound $\beta(t)$.  

C.4 ASYMPTOTIC ANALYSIS

We first prove an auxiliary lemma which gives an upper bound for the change of $\hat{\theta}$.

Lemma C.11. For a.e. $t > t_0$,
\[
\left\| \frac{d\hat{\theta}}{dt} \right\|_2 \leq \frac{B_1}{L \gamma} \cdot \frac{d}{dt} \log \rho.
\]

Proof. Observe that $\left\| \frac{d\hat{\theta}}{dt} \right\|_2 = \frac{1}{\rho} \left\| (I - \hat{\theta} \hat{\theta}^\top) \frac{d\hat{\theta}}{dt} \right\|_2 \leq \frac{1}{\rho} \left\| \frac{d\hat{\theta}}{dt} \right\|_2$. It is sufficient to bound $\left\| \frac{d\theta}{dt} \right\|_2$. By the chain rule, there exists $h_1, \ldots, h_N : [0, +\infty) \to \mathbb{R}^d$ satisfying that for a.e. $t > 0$, $\frac{d}{dt} h_n(t) \in \partial^c q_n$ and $\frac{d}{dt} = \sum_{n \in [N]} e^{-f(q_n)} f'(q_n) h_n(t)$. By definition of $B_1$, $\|h_n\|_2 \leq B_1 \rho^{L - 1}$ for a.e. $t > 0$. So we have
\[
\left\| \frac{d\theta}{dt} \right\|_2 \leq \sum_{n \in [N]} e^{-f(q_n)} f'(q_n) \|h_n\|_2 \leq \sum_{n \in [N]} e^{-f(q_n)} f'(q_n) q_n \cdot \frac{1}{q_n} \cdot B_1 \rho^{L - 1}.
\]

Note that every summand is positive. By Lemma A.5, $q_n$ is lower-bounded by $q_n \geq q_{\min} \geq g(\log \frac{1}{\rho})$, so we can replace $q_n$ with $g(\log \frac{1}{\rho})$ in the above inequality. Combining with the fact that $\sum_{n \in [N]} e^{-f(q_n)} f'(q_n) q_n$ is just $\nu$, we have
\[
\left\| \frac{d\theta}{dt} \right\|_2 \leq \frac{\nu}{g(\log \frac{1}{\rho})} \cdot B_1 \rho^{L - 1} = \frac{B_1 \nu}{\gamma \rho}.
\]

So we have $\left\| \frac{d\theta}{dt} \right\|_2 \leq \frac{1}{\rho} \left\| \frac{d\theta}{dt} \right\|_2 \leq \frac{B_1}{L \gamma} \cdot \frac{d}{dt} \log \rho$. \hfill \-box

To prove Theorem A.8, we consider each limit point $\hat{\theta}/q_{\min}(\hat{\theta})^{1/L}$, and construct a series of approximate KKT points converging to it. Then $\theta/q_{\min}(\theta)^{1/L}$ can be shown to be a KKT point by Theorem C.4. The following lemma ensures that such construction exists.

Lemma C.12. For every limit point $\hat{\theta}$ of $\left\{ \hat{\theta}(t) : t \geq 0 \right\}$, there exists a sequence of $\left\{ t_m : m \in \mathbb{N} \right\}$ such that $t_m \uparrow +\infty$, $\hat{\theta}(t_m) \to \hat{\theta}$, and $\beta(t_m) \to 1$.

Proof. Let $\{ \epsilon_m > 0 : m \in \mathbb{N} \}$ be an arbitrary sequence with $\epsilon_m \to 0$. Now we construct $\{ t_m \}$ by induction. Suppose $t_1 < t_2 < \cdots < t_{m-1}$ have already been constructed. Since $\hat{\theta}$ is a limit point and $\hat{g}(\tau) \uparrow \hat{g}(\infty)$ (recall that $\hat{g}(\infty) := \lim_{t \to +\infty} \hat{g}(t)$), there exists $s_m > t_{m-1}$ such that
\[
\left\| \theta(s_m) - \theta \right\|_2 \leq \epsilon_m \quad \text{and} \quad \frac{1}{L} \log \frac{\hat{g}(s_m)}{\hat{g}(s_m)} \leq \epsilon_m^3.
\]

Let $s'_m > s_m$ be a time such that $\log \rho(s'_m) = \log \rho(s_m) + \epsilon_m$. According to Theorem A.7, $\log \rho \to +\infty$, so $s'_m$ must exist. We construct $t_m \in (s_m, s'_m)$ to be a time that $\beta(t_m)^{-2} - 1 \leq \epsilon_m^2$, where the existence can be shown by Corollary C.10. 

25
Now we show that this construction meets our requirement. It follows from $\beta(t_m)^{-2} - 1 \leq \epsilon_m^2$ that $\beta(t_m) \geq 1/\sqrt{1 + \epsilon_m^2} \rightarrow 1$. By Lemma C.11 we also know that
\[
\|\theta(t_m) - \tilde{\theta}\|_2 \leq \|\tilde{\theta}(t_m) - \hat{\theta}(s_m)\|_2 + \|\hat{\theta}(s_m) - \tilde{\theta}\|_2 \leq \frac{B_1}{L\gamma(t_0)} \cdot \epsilon_m + \epsilon_m \rightarrow 0.
\]
This completes the proof. □

Proof of Theorem A.8 Let $\tilde{\theta} := \theta/q_{\min}(\theta)\|L\|^{1/2}$ for short. Let $\{t_m : m \in \mathbb{N}\}$ be the sequence constructed as in Lemma C.12. For each $t_m$, define $\epsilon(t_m)$ and $\delta(t_m)$ as in Lemma C.8. Then we know that $\theta(t_m)/q_{\min}(t_m)^{\|L\|/2} \rightarrow (\epsilon(t_m), \delta(t_m))$ is a KKT point of (P). By Theorem A.8, every limit point $\bar{\theta}$ satisfies KKT conditions. Thus, we only need to show $\bar{\theta}$ is a KKT point of (P).

C.5 Non-Asymptotic Analysis

Proof of Theorem A.9 Let $C_0 := \frac{1}{L} \log \frac{\epsilon_m}{\gamma(t_0)}$. Without loss of generality, assume $\epsilon < \frac{\sqrt{\epsilon}}{2} C_1$ and $\delta < C_2/f(bf)$. Let $t_1$ be the time such that $\log \rho(t_1) = \frac{1}{L} \log \frac{g(C_2\delta^{-1})}{\gamma(t_0)} = \Theta(\log \delta^{-1})$ and $t_2$ be the time such that $\log \rho(t_2) - \log \rho(t_1) = \frac{1}{L} C_0 C_2^{-1} \rho(t_1)^{-2} = \Theta(\epsilon^{-2})$. By Corollary C.10 there exists $t_s \in (t_1, t_2)$ such that $\beta(t_s)^{-2} - 1 \leq 2\epsilon^2 C_1^{-2}$.

Now we argue that $\tilde{\theta}(t_s)$ is an $(\epsilon, \delta)$-KKT point. By Lemma C.8 we only need to show $C_1 \sqrt{1 - \beta(t_s)} \leq \epsilon$ and $C_2/f(\gamma(t_s)) \rho(t_s)^L \leq \delta$. For the first inequality, by assumption $\epsilon < \frac{\sqrt{\epsilon}}{2} C_1$, we know that $\beta(t_s)^{-2} - 1 < 3$, which implies $|\beta(t_s)| < \frac{\sqrt{\epsilon}}{2}$. Then we have $1 - \beta(t_s) \leq 2\epsilon^2 C_1^{-2} \cdot \frac{\beta(t_s)^2}{1 + \beta(t_s)} \leq 2\epsilon^2 C_1^{-2}$. Therefore, $C_1 \sqrt{1 - \beta(t_s)} \leq \epsilon$ holds. For the second inequality, $\gamma(t_s) \rho(t_s)^L \geq \gamma(t_s) \cdot \frac{g(C_2\delta^{-1})}{\gamma(t_0)} \geq g(C_2\delta^{-1})$. Therefore, $C_2/f(\gamma(t_s)) \rho(t_s)^L \leq \delta$ holds. □

C.6 Proof for Corollary 4.5

By the homogeneity of $q_n$, we can characterize KKT points using kernel SVM.

Lemma C.13. If $\theta_*$ is KKT point of (P), then there exists $h_n \in \partial^c \Phi_{\mu_n}(\theta_*)$ for $n \in [N]$ such that $\frac{1}{L} \theta_*$ is an optimal solution for the following constrained optimization problem (Q):
\[
\min \frac{1}{2} \|\theta\|^2_2 \quad \text{s.t.} \quad y_n \langle \theta, h_n \rangle \geq 1 \quad \forall n \in [N]
\]

Proof. It is easy to see that (Q) is a convex optimization problem. For $\theta = \frac{1}{L} \theta_*$ from Theorem B.2 we can see that $y_n \langle \theta, h_n \rangle = 2q_n(\theta_*) \geq 2 > 1$, which implies Slater’s condition. Thus, we only need to show that $\frac{1}{L} \theta_*$ satisfies KKT conditions for (Q).

By the KKT conditions for (P), we can construct $h_n \in \partial^c q_n(\theta_*)$ for $n \in [N]$ such that $\theta_* - \sum_{n=1}^N \lambda_n y_n h_n = 0$ for some $\lambda_1, \ldots, \lambda_N \geq 0$ and $\lambda_n(q_n(\theta_*) - 1) = 0$. Thus, $\frac{1}{L} \theta_*$ satisfies

1. $\frac{1}{L} \theta_* - \sum_{n=1}^N \frac{1}{L} \lambda_n y_n h_n = 0$;
2. $\frac{1}{L} \lambda_n \left( y_n \left( \frac{1}{L} \theta_* + h_n \right) - 1 \right) = \frac{1}{L} \lambda_n(q_n(\theta_*) - 1) \geq 0$.

So $\frac{1}{L} \theta_*$ satisfies KKT conditions for (Q). □

Now we prove Corollary 4.5 in Section 4.3

Proof. By Theorem A.8, every limit point $\theta$ is along the direction of a KKT point of (P). Combining this with Lemma C.13 we know that every limit point $\theta$ is also along the max-margin direction of (Q).
For smooth models, $h_n$ in (Q) is exactly the gradient $\nabla \Phi_{x_n}(\tilde{\theta})$. So, (Q) is the optimization problem for SVM with kernel $K_\theta(x, x') = \langle \Phi_{x}(\tilde{\theta}), \Phi_{x'}(\tilde{\theta}) \rangle$. For non-smooth models, we can construct an arbitrary function $h(x) \in \partial \Phi_{x}(\tilde{\theta})$ that ensures $h(x_n) = h_n$. Then, (Q) is the optimization problem for SVM with kernel $K_\theta(x, x') = \langle h(x), h(x') \rangle$.

\section{Tight Bounds for Loss Convergence and Weight Growth}

In this section, we give proof for Theorem A.10, which gives tight bounds for loss convergence and weight growth under Assumption (A1), (A2), (B3), (B4).

\subsection{Consequences of (B3.4)}

Before proving Theorem A.10, we show some consequences of (B3.4).

**Lemma D.1.** For $f(\cdot)$ and $g(\cdot)$, we have

1. For all $x \in [b_g, +\infty)$, $\frac{g(x)}{g'(x)} \in \left[\frac{1}{2K}, 2Kx\right]$.

2. For all $y \in [g(b_y), +\infty)$, $\frac{f(y)}{f'(y)} \in \left[\frac{1}{2K}, 2Ky\right]$.

Thus, $g(x) = \Theta(xg'(x))$, $f(y) = \Theta(yf'(y))$.

**Proof.** To prove Item 1, it is sufficient to show that

$$g(x) = b_f + \int_{f(b_f)}^{x} g'(u)du \geq \int_{x/2}^{x} g'(u)du \geq (x/2) \cdot \frac{g'(x)}{K} = \frac{1}{2K} \cdot xg'(x).$$

$$x = f(b_f) + \int_{f(b_f)}^{x} g'(u)f'(g(u))du \geq \frac{f'(g(x))}{K} \int_{f(b_f)}^{x} g'(u)du \geq (g(x)/2) \cdot \frac{f'(g(x))}{K} = \frac{1}{2K} \cdot \frac{g(x)}{g'(x)}.$$

To prove Item 2, we only need to notice that Item 1 implies $yf'(y) = \frac{f'(g(y))}{g'(g(y))} \in \left[\frac{1}{2K}, 2Kf(y)\right]$ for all $y \in [g(b_y), +\infty)$.

Recall that (B3.4) directly implies that $f'(\Theta(x)) = \Theta(f'(x))$ and $g'(\Theta(x)) = \Theta(g'(x))$. Combining this with Lemma D.1, we have the following corollary:

**Corollary D.2.** $f(\Theta(x)) = \Theta(f(x))$ and $g(\Theta(x)) = \Theta(g(x))$.

Also, note that Lemma D.1 essentially shows that $(log f(x))' = \Theta(1/x)$ and $(log g(x))' = \Theta(1/x)$. So $log f(x) = \Theta(log x)$ and $log g(x) = \Theta(log x)$, which means that $f$ and $g$ grow at most polynomially.

**Corollary D.3.** $f(x) = x^{\Theta(1)}$ and $g(x) = x^{\Theta(1)}$.

\subsection{Proof for Theorem A.10}

We follow the notations in Section 5 to define $\rho := \|\theta\|_2$ and $\tilde{\theta} := \frac{\theta}{\|\theta\|_2} \in S^{d-1}$, and sometimes we view the functions of $\theta$ as functions of $t$. And we use the notations $B_0, B_1$ from Appendix C.3.

The key idea to prove Theorem A.10 is to utilize Lemma B.6, in which $\mathcal{L}(t)$ is bounded from above by $\frac{1}{\rho} \mathcal{L}(\frac{1}{\gamma(t)})$. So upper bounding $\mathcal{L}(t)$ reduces to lower bounding $\gamma(t)^{-1}$. In the following lemma, we obtain tight asymptotic bounds for $G(\cdot)$ and $G^{-1}(\cdot)$.
Lemma D.4. For function $G(\cdot)$ defined in Lemma B.6 and its inverse function $G^{-1}(\cdot)$, we have the following bounds:

$$G(x) = \Theta \left( \frac{g(\log x)^{2/L}}{(\log x)^2} x \right) \quad \text{and} \quad G^{-1}(y) = \Theta \left( \frac{(\log y)^2}{g(\log y)^{2/L} y} \right).$$

**Proof.** We first prove the bounds for $G(x)$, and then prove the bounds for $G^{-1}(y)$.

**Bounding for $G(x)$.** Let $C_G = \int_{\exp(b_y)}^{\exp(\beta_\rho)} \frac{g'(\log u)^2}{g(\log u)^{2-2/L}} du$. For $x \geq \exp(b_y)$,

$$G(x) = C_G + \int_{\exp(b_y)}^{x} \frac{g'(\log u)^2}{g(\log u)^{2-2/L}} du = C_G + \int_{\exp(b_y)}^{x} \left( \frac{g'(\log u) \log u}{g(\log u)} \right)^2 \frac{g(\log u)^{2/L}}{(\log u)^2} du,$$

By Lemma D.1,

$$G(x) \leq C_G + 4K^2 g(\log x)^{2/L} \int_{\exp(b_y)}^{x} \frac{1}{(\log u)^2} du = O \left( \frac{g(\log x)^{2/L}}{(\log x)^2} x \right).$$

On the other hand, for $x \geq \exp(2b_y)$, we have

$$G(x) \geq \frac{1}{4K^2} \int_{\exp(b_y)}^{x} \frac{g(\log u)^{2/L}}{(\log u)^2} du \geq \frac{g((\log x)/2)^{2/L}}{4K^2} \int_{\exp(b_y)}^{x} \frac{1}{(\log u)^2} du = \Omega \left( \frac{g(\log x)^{2/L}}{(\log x)^2} x \right).$$

**Bounding for $G^{-1}(y)$.** Let $x = G^{-1}(y)$ for $y \geq 0$. $G(x)$ always has a finite value whenever $x$ is finite. So $x \to +\infty$ when $y \to +\infty$. According to the first part of the proof, we know that $y = \Theta \left( \frac{g(\log x)^{2/L}}{(\log x)^2} x \right)$. Taking logarithm on both sides and using Corollary D.3, we have $\log y = \Theta(\log x)$. By Corollary D.2 $g(\log y) = \Theta(\log x) = \Theta(g(\log x))$. Therefore,

$$\frac{(\log y)^2}{g(\log y)^{2/L} y} = \Theta \left( \frac{(\log x)^2}{g(\log x)^{2/L} y} \right) = \Theta(x).$$

This implies that $x = \Theta \left( \frac{(\log y)^2}{g(\log y)^{2/L} y} \right).$ \hfill \Box

For other bounds, we derive them as follows. We first show that $g(\log \frac{1}{\gamma}) = \Theta(\rho_L)$. With this equivalence, we derive an upper bound for the gradient at each time $t$ in terms of $\mathcal{L}$, and take an integration to bound $\mathcal{L}(t)$ from below. Now we have both lower and upper bounds for $\mathcal{L}(t)$. Plugging these two bounds to $g(\log \frac{1}{\gamma}) = \Theta(\rho_L)$ gives the lower and upper bounds for $\rho(t)$.

**Proof for Theorem A.10.** We first prove the upper bound for $\mathcal{L}$. Then we derive lower and upper bounds for $\rho$ in terms of $\mathcal{L}$, and use these bounds to give a lower bound for $\mathcal{L}$. Finally, we plug in the tight bounds for $\mathcal{L}$ to obtain the lower and upper bounds for $\rho$ in terms of $t$.

**Upper Bounding $\mathcal{L}$.** By Lemma B.6 we have $\frac{1}{\mathcal{L}} \geq G^{-1}(\Omega(t))$. Using Lemma D.4, we have $\frac{1}{\mathcal{L}} = \Omega \left( \frac{(\log t)^2}{g(\log t)^{2-2/L} t} \right)$, which completes the proof.

**Bounding $\rho$ in Terms of $\mathcal{L}$.** Let $\hat{\gamma}(t) \geq \hat{\gamma}(t_0)$, then $\rho_L \leq \frac{1}{\hat{\gamma}(t_0)} g(\log \frac{1}{\mathcal{L}}) = O(g(\log \frac{1}{\mathcal{L}}))$. On the other hand, $g(\log \frac{1}{\gamma}) \leq q_{\min} \leq B_0 \rho_L$. So $\rho_L = \Omega(g(\log \frac{1}{\mathcal{L}}))$. Therefore, we have the following relationship between $\rho_L$ and $g(\log \frac{1}{\mathcal{L}})$:

$$\rho_L = \Theta(g(\log \frac{1}{\mathcal{L}})). \quad (17)$$

28
Lower Bounding $\mathcal{L}$. Let $h_1, \ldots, h_N$ be a set of vectors such that $h_n \in \partial g / \partial f$ and

$$\frac{d\theta}{dt} = \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n) h_n.$$

By \[17\] and Corollary \ref{corollary:smoothness}, $f'(q_n) = f'(O(\rho^t)) = f'(O(g(\log \frac{1}{L}))) = O(1/g'(\log \frac{1}{L}))$. Again by \[17\], we have $\|h_n\|_2 \leq B_1 \rho^{t-1} = O(g(\log \frac{1}{L})^{1-1/L})$. Combining these two bounds together, it follows from Corollary \ref{corollary:smoothness} that

$$f'(q_n) \|h_n\|_2 = O\left(\frac{g(\log \frac{1}{L})^{1-1/L}}{g'(\log \frac{1}{L})}\right) = O\left(\frac{\log \frac{1}{L}}{g(\log \frac{1}{L})^{1/L}}\right).$$

Thus,

$$\frac{d \mathcal{L}}{dt} = \left\| \frac{d\theta}{dt} \right\|^2_2 = \left\| \sum_{n=1}^{N} e^{-f(q_n)} f'(q_n) h_n \right\|^2_2 \leq \left( \sum_{n=1}^{N} e^{-f(q_n)} \cdot \max_{n \in [N]} \{ f'(q_n) \|h_n\|_2 \} \right)^2 \leq \mathcal{L}^2 \cdot O\left(\frac{(\log \frac{1}{L})^2}{g(\log \frac{1}{L})^{2/L}}\right).$$

By definition of $G(\cdot)$, this implies that there exists a constant $c$ such that $\frac{d}{dt} G(\frac{1}{L}) \leq c$ for any $\mathcal{L}$ that is small enough. We can complete our proof by applying Lemma \ref{lemma:discretization_error}.

Bounding $\rho$ in Terms of $t$. By \[17\] and the tight bound for $\mathcal{L}(t)$, $\rho^t = \Theta(g(\log \frac{1}{L})) = \Theta(g(\Theta(\log t)))$. Using Corollary \ref{corollary:smoothness}, we can conclude that $\rho^t = \Theta(g(\log t))$.

E \ GRADIENT DESCENT: SMOOTH HOMOGENEOUS MODELS WITH EXPONENTIAL LOSS

In this section, we discretize our proof to prove similar results for gradient descent on smooth homogeneous models. As usual, the update rule of gradient descent is defined as

$$\theta(t + 1) = \theta(t) - \eta(t) \nabla \mathcal{L}(t)$$

(18)

Here $\eta(t)$ is the learning rate, and $\nabla \mathcal{L}(t) := \nabla \mathcal{L}(\theta(t))$ is the gradient of $\mathcal{L}$ at $\theta(t)$.

We first focus on the exponential loss. At the end of this section (Appendix F), we discuss how to extend the proof to general loss functions with a similar assumption as (B3).

The main difficulty for discretizing our previous analysis comes from the fact that the original version of the smoothed normalized margin $\gamma'(\theta) := \rho^{-L} \log \frac{1}{L}$ becomes less smooth when $\rho \to +\infty$. Thus, if we take a Taylor expansion for $\gamma'(\theta(t + 1))$ from the point $\theta(t)$, although one can show that the first-order term is positive as in the gradient flow analysis, the second-order term is unlikely to be bounded during gradient descent with a constant step size. To get a smoothed version of the normalized margin that is monotone increasing, we need to define another one that is even smoother than $\gamma$.

Technically, recall that $\frac{d \mathcal{L}}{dt} = -\|\nabla \mathcal{L}\|_2^2$ does not hold exactly for gradient descent. However, if the smoothness can be bounded by $s(t)$, then it is well-known that

$$\mathcal{L}(t + 1) - \mathcal{L}(t) \leq -\eta(t)(1 - s(t)\eta(t))\|\nabla \mathcal{L}\|_2^2.$$

By analyzing the landscape of $\mathcal{L}$, one can easily find that the smoothness is bounded locally by $O(\mathcal{L} \cdot \text{polylog}(\frac{1}{L}))$. Thus, if we set $\eta(t)$ to a constant or set it appropriately according to the loss, then this discretization error becomes negligible. Using this insight, we define the new smoothed normalized margin $\tilde{\gamma}$ in a way that it increases slightly slower than $\gamma$ during training to cancel the effect of discretization error.
E.1 Assumptions

As stated in Section 4.1, we assume (A2), (A3), (A4) similarly as for gradient flow, and two additional assumptions (S1) and (S5).

(S1). Smoothness. For any fixed \(x\), \(\Phi(\cdot; x)\) is \(C^2\)-smooth on \(\mathbb{R}^d \setminus \{0\}\).

(A2). Homogeneity. There exists \(L > 0\) such that \(\forall \alpha > 0 : \Phi(\alpha \theta; x) = \alpha^L \Phi(\theta; x)\);

(A3). Exponential Loss. \(\ell(q) = e^{-q}\);

(A4). Separability. There exists a time \(t_0\) such that \(\mathcal{L}(\theta(t_0)) < 1\).

(S5). Learning rate condition. \(\sum_{t \geq 0} \eta(t) = +\infty\) and \(\eta(t) \leq H(\mathcal{L}(\theta(t)))\).

Here \(H(\mathcal{L})\) is a function of the current training loss. The explicit formula of \(H(\mathcal{L})\) is given below:

\[
H(x) := \frac{\mu(x)}{C \kappa(x)} = \Theta \left( \frac{1}{x (\log \frac{1}{x})^{3/2}/L} \right),
\]

where \(C\) is a constant, and \(\kappa(x), \mu(x)\) are two non-decreasing functions. For constant learning rate \(\eta(t) = \eta_0\), (S5) is satisfied when \(\eta_0\) is sufficiently small.

Roughly speaking, \(C \kappa(x)\) is an upper bound for the smoothness of \(\mathcal{L}\) in a neighborhood of \(\theta\) when \(x = \mathcal{L}(\theta)\). And we set the learning rate \(\eta(t)\) to be the inverse of the smoothness multiplied by a factor \(\mu(x) = o(1)\). In our analysis, \(\mu(x)\) can be any non-decreasing function that maps \((0, \mathcal{L}(t_0))\) to \((0, 1/2]\) and makes the integral \(\int_{0}^{1/2} \mu(x) dx\) exist. But for simplicity, we define \(\mu(x)\) as

\[
\mu(x) := \frac{1}{2 \log \frac{1}{x}} = \Theta \left( \frac{1}{\log \frac{1}{x}} \right).
\]

The value of \(C\) will be specified later. The definition of \(\kappa(x)\) depends on \(L\). For \(0 < L \leq 1\), we define \(\kappa(x)\) as

\[
\kappa(x) := x \left( \log \frac{1}{x} \right)^{2-2/L}
\]

For \(L > 1\), we define \(\kappa(x)\) as

\[
\kappa(x) := \begin{cases} 
  x \left( \log \frac{1}{x} \right)^{2-2/L} & x \in (0, e^{2/L-2}] \\
  \kappa_{\text{max}} & x \in (e^{2/L-2}, 1]
\end{cases}
\]

where \(\kappa_{\text{max}} := e^{(2-2/L)(\ln(2-2/L)-1)}\). The specific meaning of \(C, \kappa(x)\) and \(\mu(x)\) will become clear in our analysis.

E.2 Smoothed Normalized Margin

Now we define the smoothed normalized margins. As usual, we define \(\tilde{\gamma}(\theta) := \frac{\log \frac{\phi}{\rho w}}{\rho L}\). At the same time, we also define

\[
\tilde{\gamma}(\theta) := \frac{e^{\phi(\mathcal{L})}}{\rho L}.
\]

Here \(\phi : (0, \mathcal{L}(t_0)) \rightarrow (0, +\infty)\) is constructed as follows. Construct the first-order derivative of \(\phi(x)\) as

\[
\phi'(x) = -\sup \left\{ \frac{1 + 2(1 + \lambda(w)/L)\mu(w)}{w \log \frac{1}{w}} : w \in [x, \mathcal{L}(t_0)] \right\},
\]

where \(\lambda(x) := (\log \frac{1}{x})^{-1}\). And then we set \(\phi(x)\) to be

\[
\phi(x) = \log \log \frac{1}{x} + \int_{0}^{x} \left( \phi'(w) + \frac{1}{w \log \frac{1}{w}} \right) dw.
\]

It can be verified that \(\phi(x)\) is well-defined and \(\phi'(x)\) is indeed the first-order derivative of \(\phi(x)\). Moreover, we have the following relationship among \(\tilde{\gamma}, \tilde{\gamma}\) and \(\tilde{\gamma}\).
Lemma E.1.  \( \hat{\gamma}(\theta) \) is well-defined for \( \mathcal{L}(\theta) \leq \mathcal{L}(t_0) \) and has the following properties:

(a) If \( \mathcal{L}(\theta) \leq \mathcal{L}(t_0) \), then \( \hat{\gamma}(\theta) < \hat{\gamma}(\theta) \leq \hat{\gamma}(\theta) \).

(b) Let \( \{\theta_m \in \mathbb{R}^d : m \in \mathbb{N}\} \) be a sequence of parameters. If \( \mathcal{L}(\theta_m) \to 0 \), then

\[
\hat{\gamma}(\theta_m) = 1 - O \left( \left( \log \frac{1}{\mathcal{L}(\theta_m)} \right)^{-1} \right) \to 1.
\]

Proof. First we verify that \( \hat{\gamma} \) is well-defined. To see this, we only need to verify that

\[
I(x) := \int_0^x \left( \phi'(w) + \frac{1}{w \log \frac{1}{w}} \right) dw
\]

exists for all \( x \in (0, \mathcal{L}(t_0)) \), then it is trivial to see that \( \phi'(w) \) is indeed the derivative of \( \phi(w) \) by

\[
I'(x) = \phi'(x) + \frac{1}{x \log \frac{1}{x}}.
\]

Note that \( I(x) \) exists for all \( x \in (0, \mathcal{L}(t_0)) \) as long as \( I(x) \) exists for a small enough \( x > 0 \). By definition, it is easy to verify that \( r(w) := \frac{1 + 2(1 + \lambda(w)/L_m(w))}{w \log \frac{1}{w}} \) is decreasing when \( w \) is small enough.

Thus, for a small enough \( w > 0 \), we have

\[
-\phi'(w) = r(w) = \frac{1 + 2(1 + \frac{1}{w} \log \frac{1}{w})^{-1} \log \frac{1}{w} \log \frac{1}{w}}{w \log \frac{1}{w}^2} = \frac{1}{w \log \frac{1}{w}^2} + \frac{\log \frac{1}{w} \log \frac{1}{w}}{w \log \frac{1}{w}^2}.
\]

So we have the following for small enough \( x \):

\[
I(x) = \int_0^x -\log \frac{1}{\mathcal{L}(t_0)} \left( 1 + \frac{1}{w} \log \frac{1}{w}^{-1} \right) \log \frac{1}{w} \log \frac{1}{w} \frac{dw}{w \log \frac{1}{w}^2} = -\log \frac{1}{\mathcal{L}(t_0)} \left( \frac{1}{\log \frac{1}{w}} + \frac{2 \log \frac{1}{w} \log \frac{1}{w}}{2 \log \frac{1}{w} \log \frac{1}{w}} \right) \bigg|_0^x
\]

\[
= -\log \frac{1}{\mathcal{L}(t_0)} \left( \frac{1}{\log \frac{1}{x}} + \frac{2 \log \frac{1}{x} \log \frac{1}{x}}{2 \log \frac{1}{x} \log \frac{1}{x}} \right).
\]

This proves the existence of \( I(x) \).

Now we prove (a). By Lemma A.5, \( \hat{\gamma}(\theta) \leq \hat{\gamma}(\theta) \), so we only need to prove that \( \hat{\gamma}(\theta) < \hat{\gamma}(\theta) \). To see this, note that for all \( w \leq \mathcal{L}(t_0) \), \( r(w) > \frac{1}{w \log \frac{1}{w}} \). So we have \( -\phi'(w) > \frac{1}{w \log \frac{1}{w}} \), and this implies \( I(x) < 0 \) for all \( x \leq \mathcal{L}(t_0) \). Then it holds for all \( \theta \) with \( \mathcal{L}(\theta) \leq \mathcal{L}(t_0) \) that

\[
\frac{\hat{\gamma}(\theta)}{\hat{\gamma}(\theta)} = e^{\phi(\mathcal{L}(\theta))} = e^{\log \frac{1}{\mathcal{L}(\theta)}} = e^{(\log \frac{1}{\mathcal{L}(\theta)})} < 0.
\]

To prove (b), we combine (19) and (20), then for small enough \( \mathcal{L}(\theta_m) \), we have

\[
\hat{\gamma}(\theta_m) = \hat{\gamma}(\theta) \cdot \hat{\gamma}(\theta) = \exp \left( -\log \frac{1}{\mathcal{L}(t_0)} \left( \frac{1}{\log \frac{1}{\mathcal{L}(\theta_m)}} + \frac{1}{2 \log \frac{1}{\mathcal{L}(\theta_m)}} \right) \right) \cdot \frac{1}{\mathcal{L}(\theta_m)} \to 0.
\]

So \( \frac{\hat{\gamma}(\theta)}{\hat{\gamma}(\theta)} = 1 - O \left( \left( \log \frac{1}{\mathcal{L}(\theta_m)} \right)^{-1} \right) \to 1. \)

\[\square\]
Now we specify the value of $C_\eta$. By (S1) and (S2), we can define $B_0, B_1, B_2$ as follows:

\[
B_0 := \sup \{ q_n : \theta \in S_d, n \in [N] \}, \\
B_1 := \sup \{ \| \nabla q_n \|_2 : \theta \in S_d, n \in [N] \}, \\
B_2 := \sup \{ \| \nabla^2 q_n \|_2 : \theta \in S_d, n \in [N] \}.
\]

Then we set $C_\eta := \left( \frac{1}{2} (B_0^2 + \rho(t_0)^{-L} B_2) \right) \min \left\{ \tilde{\gamma}(t_0)^{-2+2/L}, B_0^{-2+2/L} \right\}$.

### E.3 Theorems

Now we state our main theorems for the monotonicity of the normalized margin and the convergence to KKT points. We will prove Theorem E.2 in Appendix E.4 and prove Theorem E.3 and E.4 in Appendix E.5.

**Theorem E.2.** Under assumptions (S1), (A2) - (A4), (S5), the following are true for gradient descent:

1. For all $t \geq t_0$, $\tilde{\gamma}(t + 1) \geq \tilde{\gamma}(t)$;
2. For all $t \geq t_0$, either $\tilde{\gamma}(t + 1) > \tilde{\gamma}(t)$ or $\tilde{\theta}(t + 1) = \tilde{\theta}(t)$;
3. $L(t) \to 0$ and $\rho(t) \to \infty$ as $t \to +\infty$; therefore, $|\tilde{\gamma}(t) - \tilde{\gamma}(t)| \to 0$.

**Theorem E.3.** Consider gradient flow under assumptions (S1), (A2) - (A4), (S5). For every limit point $\theta$ of \( \{ \tilde{\theta}(t) : t \geq 0 \} \), $\theta / q_{\min}(\theta)^{1/L}$ is a KKT point of (P).

**Theorem E.4.** Consider gradient descent under assumptions (S1), (A2) - (A4), (S5). For any $\epsilon, \delta > 0$, there exists $r := \Theta(\log \delta^{-1})$ and $\Delta := \Theta(\epsilon^{-2})$ such that $\theta / q_{\min}(\theta)^{1/L}$ is an $(\epsilon, \delta)$-KKT point at some time $t_*$ satisfying $\log \rho(t_*) \in (r, r + \Delta)$.

With a refined analysis, we can also derive tight rates for loss convergence and weight growth. We defer the proof to Appendix E.6.

**Theorem E.5.** Under assumptions (S1), (A2) - (A4), (S5), we have the following tight rates for training loss and weight norm:

\[
L(t) = \Theta \left( \frac{1}{T(\log T)^{2-2/L}} \right) \quad \text{and} \quad \rho(t) = \Theta((\log T)^{1/L}),
\]

where $T = \sum_{\tau=t_0}^{t-1} \eta(\tau)$.

### E.4 Proof for Theorem E.2

We define $\nu(t) := \sum_{n=1}^{N} e^{-q_n(t)} q_n(t)$ as we do for gradient flow. Then we can get a closed form for $(\theta(t), -\nabla L(t))$ easily from Corollary B.3. Also, we can get a lower bound for $\nu(t)$ using Lemma B.5 for exponential loss directly.

**Corollary E.6.** $(\theta(t), -\nabla L(t)) = L(\nu(t)$. If $L(t) < 1$, then $\nu(t) \geq \frac{\nu(t)}{L(t)}$.

As we are analyzing gradient descent, the norm of $\nabla L$ and $\nabla^2 L$ appear naturally in discretization error terms. To bound them, we have the following lemma when $\tilde{\gamma}(\theta)$ and $\rho(\theta)$ have lower bounds:

**Lemma E.7.** For any $\theta$, if $\tilde{\gamma}(\theta) \geq \tilde{\gamma}(t_0), \rho(\theta) \geq \rho(t_0)$, then

\[
\| \nabla L(\theta) \|_2^2 \leq 2C_\eta \kappa(L(\theta)) L(\theta) \quad \text{and} \quad \| \nabla^2 L(\theta) \|_2 \leq 2C_\eta \kappa(L(\theta)).
\]
For all gives a lower bound for the decrement of training loss. Finally, shows the monotonicity of .

Note that . Then if (P1) holds for , (P2) gives both lower and upper bounds for the weight growth using . (P3) gives a lower bound for the decrement of training loss. Finally, (P4) shows the monotonicity of , and it is trivial to deduce the first two propositions in Theorem E.2 from (P4).

Lemma E.8. For all , we only need to prove Lemma E.8. (P1) gives a lower bound for . (P2) gives both lower and upper bounds for the weight growth using . (P3) gives a lower bound for the decrement of training loss. Finally, (P4) shows the monotonicity of , and it is trivial to deduce the first two propositions in Theorem E.2 from (P4).

Lemma E.9. Suppose that holds for . Then all holds for where .

Proof for Lemma E.8 We prove this lemma by induction. For , (P1) also holds trivially.

Now we fix an integer and assume that (P1), (P2), (P3), (P4) hold for any (where is an integer and (P1) does not hold for ). If , then (P1) holds for since (P1) holds for ; if , we can also know that (P1) holds for by Lemma E.9. Suppose that . Then by the continuity of with respect to , we know that there exists such that for all , which contradicts to the definition of . Therefore, . Using Lemma E.9 again, we can conclude that (P1), (P2), (P3), (P4) hold for and .
Now we turn to prove Lemma E.9.

**Proof for Lemma E.9.** Applying (P3) on \((t, \alpha) \in \{t_0, \ldots, T - 1\} \times 1\), we have \(L(t) \leq L(t_0) < 1\). Then by Corollary E.6, we have \(\nu(t) > 0\). Applying (P2) on \((t, \alpha) \in \{t_0, \ldots, T - 1\} \times 1\), we can get \(\rho(t) \geq \rho(t_0)\).

Fix \(t = T\). By (P2) with \(\alpha \in [0, A)\) and the continuity of \(\gamma\), we have \(\gamma(t + A) \geq \gamma(t_0)\). Thus, \(\gamma(t + \alpha) \geq \gamma(t_0)\) for all \(\alpha \in [0, A)\).

**Proof for (P2).** By definition, \(\rho(t + \alpha)^2 = \|\theta(t) - \alpha \eta(t) \nabla L(t)\|_2^2 = \rho(t)^2 + 2\alpha \eta(t)^2 \|\nabla L(t)\|_2^2 - 2\alpha \eta(t) \langle \theta(t), \nabla L(t) \rangle\) by Corollary E.6 we have

\[
\rho(t + \alpha)^2 - \rho(t)^2 = \alpha^2 \eta(t)^2 \|\nabla L(t)\|_2^2 + 2L\alpha \eta(t) \nu(t).
\]

So \(\rho(t + \alpha)^2 - \rho(t)^2 \geq 2L\alpha \eta(t) \nu(t)\). For the other direction, we have the following using Corollary E.6 and Lemma E.7:

\[
\rho(t + \alpha)^2 - \rho(t)^2 = 2L\alpha \eta(t) \nu(t) \left(1 + \frac{\alpha \eta(t) \|\nabla L(t)\|_2}{2\nu(t)}\right)
\leq 2L\alpha \eta(t) \nu(t) \left(1 + \frac{C^{-1} \kappa(L(t))^{-1} \mu(L(t)) \cdot 2C \eta \kappa(L(t)) L(t)}{2L(t) \kappa(L(t))^{-1}}\right)
= 2L\alpha \eta(t) \nu(t) \left(1 + \lambda(L(t)) \mu(L(t))/L\right).
\]

**Proof for (P3).** (P3) holds trivially for \(\alpha = 0\) or \(\nabla L(t) = 0\). So now we assume that \(\alpha \neq 0\) and \(\nabla L(t) \neq 0\). By the update rule (18) and Taylor expansion, there exists \(\xi \in (0, \alpha)\) such that

\[
L(t + \alpha) = L(t) + (\theta(t + \alpha) - \theta(t))^\top \nabla L(t) + \frac{1}{2} (\theta(t + \alpha) - \theta(t))^\top \nabla^2 L(t + \xi) (\theta(t + \alpha) - \theta(t))
\leq L(t) - \alpha \eta(t) \|\nabla L(t)\|_2^2 + \frac{1}{2} \alpha^2 \eta(t)^2 \|\nabla^2 L(t + \xi)\|_2 \|\nabla L(t)\|_2^2
= L(t) - \alpha \eta(t) \left(1 - \frac{1}{2} \alpha \eta(t) \|\nabla^2 L(t + \xi)\|_2\right) \|\nabla L(t)\|_2^2.
\]

By Lemma E.7, \(\|\nabla^2 L(t + \xi)\|_2 \leq 2C \eta \kappa(L(t + \xi))\), so we have

\[
L(t + \alpha) \leq L(t) - \alpha \eta(t) \left(1 - \alpha C \eta \kappa(L(t + \xi))\right) \|\nabla L(t)\|_2^2.
\]

Now we only need to show that \(L(t + \alpha) < L(t)\) for all \(\alpha \in (0, A]\). Assuming this, we can have \(\kappa(L(t + \xi)) \leq \kappa(L(t))\) by the monotonicity of \(\kappa\), and thus

\[
L(t + \alpha) \leq L(t) - \alpha \eta(t) \left(1 - \alpha C \eta \kappa(L(t))\right) \|\nabla L(t)\|_2^2
\leq L(t) - \alpha \eta(t) \left(1 - \mu(L(t))\right) \|\nabla L(t)\|_2^2.
\]

Now we show that \(L(t + \alpha) < L(t)\) for all \(\alpha \in (0, A]\). Assume to the contrary that \(\alpha_0 := \inf \{\alpha' \in (0, A] : L(t + \alpha') \leq L(t)\}\) exists. If \(\alpha_0 = 0\), let \(p : [0, 1] \rightarrow \mathbb{R}, \alpha \mapsto L(t + \alpha)\), then \(p'(0) = \lim_{\alpha \downarrow 0} \frac{L(t + \alpha) - L(t)}{\alpha} \geq 0\), but it contradicts to \(p'(0) = -\eta(t) \|\nabla L(t)\|_2^2 < 0\). If \(\alpha_0 > 0\), then by the monotonicity of \(\kappa\) we have \(\kappa(L(t + \xi)) < \kappa(L(t))\) for all \(\xi \in (0, \alpha_0)\), and thus

\[
L(t + \alpha) \leq L(t) - \alpha \eta(t) \left(1 - \mu(L(t))\right) \|\nabla L(t)\|_2^2 < L(t),
\]

which leads to a contradiction.

**Proof for (P4).** We define \(v(t) := \hat{\theta}(t) \hat{\theta}(t)^\top (-\nabla L(t))\) and \(u(t) := (I - \hat{\theta}(t) \hat{\theta}(t)^\top) (-\nabla L(t))\) similarly as in the analysis for gradient flow. For \(v(t)\), we have

\[
\|v(t)\|_2 = \frac{1}{\rho(t)} \langle \theta(t), -L(t) \rangle = \frac{1}{\rho(t)} \cdot L\nu(t).
\]
Decompose $\|\nabla \mathcal{L}(t)\|_2^2 = \|v(t)\|_2^2 + \|u(t)\|_2^2$. Then by (P3), we have

$$\mathcal{L}(t + \alpha) - \mathcal{L}(t) \leq -\alpha \eta(t)(1 - \mu(\mathcal{L}(t))) \left( \frac{1}{\rho(t)^2} \cdot L^2 \nu(t)^2 + \|u\|_2^2 \right).$$

Multiplying $\frac{1 + \lambda(\mathcal{L}(t)) \mu(\mathcal{L}(t))/L}{(1 - \mu(\mathcal{L}(t))) \nu(t)}$ on both sides, we have

$$\frac{1 + \lambda(\mathcal{L}(t)) \mu(\mathcal{L}(t))/L}{(1 - \mu(\mathcal{L}(t))) \nu(t)}(\mathcal{L}(t + \alpha) - \mathcal{L}(t)) \leq -\alpha \eta(t) \left(1 + \frac{\lambda(\mathcal{L}(t)) \mu(\mathcal{L}(t))}{L}\right) \left( \frac{L^2 \nu(t)}{\rho(t)^2} + \|u\|_2^2 \right).$$

By Corollary E.6, we can bound $\nu(t)$ by $\nu(t) \geq \mathcal{L}(t)/\lambda(\mathcal{L}(t))$. By (P2), we have the inequality $\alpha \eta(t) \nu(t) \left(1 + \lambda(\mathcal{L}(t)) \mu(\mathcal{L}(t))/L\right) \geq \frac{\rho(t+\alpha)^2 - \rho(t)^2}{2L \nu(t)^2}$. So we further have

$$\frac{1 + \lambda(\mathcal{L}(t)) \mu(\mathcal{L}(t))/L}{(1 - \mu(\mathcal{L}(t))) \Lambda(t)/\lambda(\mathcal{L}(t))}(\mathcal{L}(t+\alpha) - \mathcal{L}(t)) \leq -\frac{1}{\rho(t)^2}(\rho(t+\alpha)^2 - \rho(t)^2) \left( \frac{L}{2} + \frac{\rho(t)^2}{2L \nu(t)^2} \|u\|_2^2 \right)$$

From the definition $\phi$, it is easy to see that $-\phi'(\mathcal{L}(t)) \geq \frac{1 + \lambda(\mathcal{L}(t)) \mu(\mathcal{L}(t))/L}{(1 - \mu(\mathcal{L}(t))) \Lambda(t)/\lambda(\mathcal{L}(t))}$.

Let $\psi(x) = -\log x$.

Combining these together gives

$$\phi'\mathcal{(L}(t))(\mathcal{L}(t) + \alpha) - \mathcal{L}(t) + \psi'(\rho(t)^2)(\rho(t+\alpha)^2 - \rho(t)^2) \left( \frac{L}{2} + \frac{\rho(t)^2}{2L \nu(t)^2} \|u\|_2^2 \right) \geq 0$$

Then by convexity of $\phi$ and $\psi$, we have

$$\left(\phi(\mathcal{L}(t+\alpha)) - \phi(\mathcal{L}(t))\right) + \left(\log \frac{1}{\rho(t+\alpha)} - \log \frac{1}{\rho(t)}\right) \left( \frac{L}{2} + \frac{\rho(t)^2}{2L \nu(t)^2} \|u\|_2^2 \right) \geq 0$$

And by definition of $\tilde{\gamma}$, this can be re-written as

$$\log \tilde{\gamma}(t+\alpha) - \log \tilde{\gamma}(t) = \left(\phi(\mathcal{L}(t+\alpha)) - \phi(\mathcal{L}(t))\right) + L \left(\log \frac{1}{\rho(t+\alpha)} - \log \frac{1}{\rho(t)}\right) \geq -\left(\log \frac{1}{\rho(t+\alpha)} - \log \frac{1}{\rho(t)}\right) \frac{\rho(t)^2}{2L \nu(t)^2} \|u\|_2^2 \geq \frac{\rho(t)^2}{L \nu(t)^2} \|u\|_2^2 \log \rho(t+\alpha).$$

**Proof for (P1).** By (P4), $\log \tilde{\gamma}(t+\alpha) \geq \log \tilde{\gamma}(t) \geq \log \tilde{\gamma}(t_0)$. Note that $\phi(x) \geq \log \frac{1}{x}$. So we have

$$\tilde{\gamma}(t+\alpha) > \tilde{\gamma}(t + \alpha) \geq \tilde{\gamma}(t_0),$$

which completes the proof.

For showing the third proposition in Theorem E.2, we use (P1) to give a lower bound for $\|\nabla \mathcal{L}(t)\|_2$. and use (P3) to show the speed of loss decreasing. Then it can be seen that $\mathcal{L}(t) \to 0$ and $\rho(t) \to +\infty$. By Lemma E.1, we then have $\|\nabla \mathcal{L}(t)\|_2 \to 0$.

**Lemma E.10.** Let $E_0 := \mathcal{L}(t_0)^2(\log \frac{1}{\varepsilon(t_0)^2})^{-2^L}$. Then for all $t > t_0$,

$$E(\mathcal{L}(t)) \geq \frac{1}{2} L^2 \tilde{\gamma}(t_0)^2/L \sum_{\tau = t_0}^{t-1} \eta(\tau) \quad \text{for } E(x) := \int_x^{\mathcal{L}(t_0)} \min\{u^2(\log \frac{1}{u})^{2^L} - E_0\}du.$$
By (P1), $\rho(t)^{-2} \leq \tilde{\gamma}(t_0)^{2/L} \left( \log \frac{1}{\xi(t)} \right)^{-2/L}$. By Corollary E.6, $\nu(t)^2 \geq \mathcal{L}(t)^2 \left( \log \frac{1}{\xi(t)} \right)^2$. Thus we have

$$
\mathcal{L}(t+1) - \mathcal{L}(t) \leq \frac{1}{2} \eta(t) \cdot L^2 \tilde{\gamma}(t_0)^{2/L} \mathcal{L}(t)^2 \left( \log \frac{1}{\xi(t)} \right)^{2-2/L}.
$$

It is easy to see that $1/(\log (1 + 2/L))$ is unimodal in $(0, 1)$, so $E'(x)$ is non-decreasing and $E(x)$ is convex. So we have

$$
E(\mathcal{L}(t+1)) - E(\mathcal{L}(t)) \geq E'(\mathcal{L}(t)) (\mathcal{L}(t+1) - \mathcal{L}(t)) \\
\geq \frac{1}{2} \eta(t) \cdot L^2 \tilde{\gamma}(t_0)^{2/L} \cdot \mathcal{L}(t)^2 \left( \log \frac{1}{\xi(t)} \right)^{2-2/L} \\
\geq \frac{1}{2} \eta(t) \cdot L^2 \tilde{\gamma}(t_0)^{2/L},
$$

which proves $E(\mathcal{L}(t)) \geq \frac{1}{2} L^2 \tilde{\gamma}(t_0)^{2/L} \sum_{t=t_0}^{t-1} \eta(\tau)$. Note that $\mathcal{L}$ is non-decreasing. If $\mathcal{L}$ does not decreases to 0, then neither does $E(\mathcal{L})$. But the RHS grows to $+\infty$, which leads to a contradiction. So $\mathcal{L} \to 0$. To make $\mathcal{L} \to 0$, $q_{\text{min}}$ must converge to $+\infty$. So $\rho \to +\infty$. \qed

### E.5 Proof for Theorem E.3 and E.4

The proofs for Theorem E.3 and E.4 are similar as those for Theorem A.8 and A.9 in Appendix C.

Define $\beta(t) := \frac{1}{\|\nabla L(t)^2\|_2} \left( \tilde{\theta}, -\nabla L(t) \right)$ as we do in Appendix C. It is easy to see that Lemma C.8 still holds if we replace $\gamma(t_0)$ with $\tilde{\gamma}(t_0)$. So we only need to show $\mathcal{L} \to 0$ and $\beta \to 1$ for proving convergence to KKT points. $\mathcal{L} \to 0$ can be followed from Theorem E.2. Similar as the proof for Lemma C.9, it follows from Lemma E.8 and (15) that for all $t_2 > t_1 \geq t_0$,

$$
\sum_{t=t_1}^{t_2-1} (\beta(t)^2 - 1) \cdot \log \frac{\rho(t+1)}{\rho(t)} \leq \frac{1}{L} \log \frac{\gamma(t_2)}{\gamma(t_1)}.
$$

Now we prove Theorem E.4.

**Proof for Theorem E.4.** We make the following changes in the proof for Theorem A.9. First, we replace $\gamma(t_0)$ with $\tilde{\gamma}(t_0)$, since $\tilde{\gamma}(t)$ ($t \geq t_0$) is lower bounded by $\tilde{\gamma}(t_0)$ rather than $\gamma(t_0)$. Second, when choosing $t_1$ and $t_2$, we make $\log \rho(t_1)$ and $\log \rho(t_2)$ equal to the chosen values approximately with an additive error $o(1)$, rather than make them equal exactly. This is possible because it can be shown from (P2) in Lemma E.8 that the following holds:

$$
\rho(t+1)^2 - \rho(t)^2 = O(\eta(t) \nu(t))
$$

$$
= o(\kappa(L(t))^{-1} \cdot \mathcal{L}(t) \rho(t)^L) = o \left( \frac{\rho(t)^L}{(\log \frac{1}{\xi(t)})^{2-2/L}} \right) = o(\rho(t)^{-L+2}).
$$

Dividing $\rho(t)^2$ on the leftmost and rightmost sides, we have

$$
\rho(t+1)^2 / \rho(t)^2 = 1 + o(\rho(t)^{-L}) = 1 + o(1),
$$

which implies that $\log \rho(t+1) - \log \rho(t) = o(1)$. Therefore, for any $R$, we can always find the minimum time $t$ such that $\log \rho(t) \geq R$, and it holds for sure that $\log \rho(t) - R \to 0$ as $R \to +\infty$. \qed

For proving Theorem E.3, we also need the following lemma as a variant of Lemma C.11.

**Lemma E.11.** For all $t \geq t_0$,

$$
\|\tilde{\theta}(t+1) - \tilde{\theta}(t)\|_2 \leq \left( \frac{B_1}{L \gamma(t)} \cdot \frac{\rho(t+1)}{\rho(t)} + 1 \right) \log \frac{\rho(t+1)}{\rho(t)}.
$$
Proof. According to the update rule, we have
\[
\left\| \hat{\theta}(t + 1) - \hat{\theta}(t) \right\|_2 = \frac{1}{\rho(t + 1)} \left\| \theta(t + 1) - \frac{\rho(t + 1)}{\rho(t)} \theta(t) \right\|_2 \\
\leq \frac{1}{\rho(t + 1)} \left( \left\| \theta(t + 1) - \theta(t) \right\|_2 + \left\| \frac{\rho(t + 1)}{\rho(t)} - 1 \right\| \theta(t) \right\|_2 - 1) \right) .
\]
By (16), \( \left\| \nabla L(t) \right\|_2 \leq \frac{B_1 \rho(t)}{\gamma(t) \rho(t)} \). So we can bound the first term as
\[
\frac{\eta(t)}{\rho(t + 1)} \left\| \nabla L(t) \right\|_2 \leq \frac{B_1}{\gamma(t)} \frac{\eta(t) \rho(t)}{\rho(t + 1) \rho(t)} \leq \frac{B_1}{\gamma(t)} \frac{\rho(t + 1)^2 - \rho(t)^2}{2L \rho(t + 1) \rho(t)} \\
= \frac{B_1}{2L \gamma(t)} \frac{\rho(t + 1)^2 - \rho(t)^2}{\rho(t + 1)^2} \frac{\rho(t + 1)}{\rho(t)} \\
\leq \frac{B_1}{L \gamma(t)} \frac{\rho(t + 1)}{\rho(t)} \log \frac{\rho(t + 1)}{\rho(t)}
\]
where the last inequality uses the inequality \( \frac{a - b}{a} \leq \log(a/b) \). Using this inequality again, we can bound the second term by
\[
1 - \frac{\rho(t)}{\rho(t + 1)} \leq \log \frac{\rho(t + 1)}{\rho(t)} .
\]
Combining these together gives
\[
\left\| \hat{\theta}(t + 1) - \hat{\theta}(t) \right\|_2 \leq \left( \frac{B_1}{L \gamma(t)} \frac{\rho(t + 1)}{\rho(t)} + 1 \right) \log \frac{\rho(t + 1)}{\rho(t)} . \quad \square
\]

Now we are ready to prove Theorem E.3.

Proof for Theorem E.3. As discussed above, we only need to show a variant of Lemma C.12 for gradient descent: for every limit point \( \hat{\theta} \) of \( \{ \hat{\theta}(t) : t \geq 0 \} \), there exists a sequence of \( \{ t_m : m \in \mathbb{N} \} \) such that \( t_m \uparrow +\infty, \hat{\theta}(t_m) \to \hat{\theta}, \) and \( \beta(t_m) \to 1 \).

We only need to change the choices of \( s_m, s_m', t_m \) in the proof for Lemma C.12. We choose \( s_m > t_m - 1 \) to be a time such that
\[
\left\| \hat{\theta}(s_m) - \hat{\theta} \right\|_2 \leq \epsilon_m \quad \text{and} \quad \frac{1}{L} \log \left( \lim_{t \to +\infty} \frac{\gamma(t)}{\gamma(s_m)} \right) \leq \epsilon_m^2 .
\]

Then we let \( s_m' > s_m \) be the minimum time such that \( \log \rho(s_m') \geq \log \rho(s_m) + \epsilon_m \). According to Theorem E.2, \( s_m \) and \( s_m' \) must exist. Finally, we construct \( t_m \in \{ s_m, \ldots, s_m' - 1 \} \) to be a time that \( \beta(t_m)^{-2} - 1 \leq \epsilon_m^2 \), where the existence can be shown by (22).

To see that this construction meets our requirement, note that \( \beta(t_m)^{-2} - 1 \leq \epsilon_m^2 \to 0 \) and
\[
\left\| \hat{\theta}(t_m) - \hat{\theta} \right\|_2 \leq \left\| \hat{\theta}(t_m) - \hat{\theta}(s_m) \right\|_2 + \left\| \hat{\theta}(s_m) - \hat{\theta} \right\|_2 \leq \left( \frac{B_1}{L \gamma(t_0)} \epsilon_m + 1 \right) \cdot \epsilon_m + \epsilon_m \to 0 ,
\]
where the last inequality is by Lemma E.11. \( \square \)

E.6 PROOF FOR THEOREM E.5

Proof. By a similar analysis as Lemma D.4, we have
\[
E(x) = \int_x^{L(t_0)} \Theta \left( \frac{1}{u^2 (\log \frac{1}{u})^{2-2/L}} \right) du = \int_{1/L(t_0)}^{1/x} \Theta \left( \frac{1}{(\log u)^{2-2/L}} \right) du \\
= \Theta \left( \frac{1}{x (\log \frac{1}{x})^{2-2/L}} \right) .
\]
We can also bound the inverse function $E^{-1}(y)$ by $\Theta\left(\frac{1}{\eta^2 \log \eta} \right)$. With these, we can use a similar analysis as Theorem A.10 to prove Theorem E.5.

First, using a similar proof as for (17), we have $\rho^L = \Theta\left(\log \frac{1}{\varepsilon} \right)$. So we only need to show $\mathcal{L}(t) = \Theta\left(\frac{1}{T^2} \log T \right)$. With a similar analysis as for (P3) in Lemma E.9, we have the following bound for $\mathcal{L}(\tau + 1) - \mathcal{L}(\tau)$:

$$
\mathcal{L}(\tau + 1) - \mathcal{L}(\tau) \geq -\eta(\tau)(1 + \mu(\mathcal{L}(\tau))) \parallel \nabla \mathcal{L}(\tau) \parallel _2^2 .
$$

Using the fact that $\eta(\tau) \leq \frac{1}{T^2}$, we have $\mathcal{L}(\tau + 1) - \mathcal{L}(\tau) \geq -\frac{1}{2} \eta(\tau) \parallel \nabla \mathcal{L}(\tau) \parallel _2^2$. By Lemma E.7, $\parallel \nabla \mathcal{L}(\tau) \parallel _2 \leq 2C, \kappa(\mathcal{L}(\tau)) \mathcal{L}(\tau)$. Using a similar proof as for Lemma E.10, we can show that $E(\mathcal{L}(t)) \leq O(T)$. Combining this with Lemma E.10, we have $E(\mathcal{L}(t)) = \Theta(T)$. Therefore, $\mathcal{L}(t) = \Theta\left(\frac{1}{T^2} \log T \right)$.

## F Gradient Descent: General Loss Functions

It is worth to note that the above analysis can be extended to other loss functions. For this, we need to replace (B3) with a strong assumption (S3), which takes into account the second-order derivatives of $f$.

(S3). The loss function $\ell(q)$ can be expressed as $\ell(q) = e^{-f(q)}$ such that

(S3.1) $f: \mathbb{R} \to \mathbb{R}$ is $C^2$-smooth.

(S3.2) $f'(q) \geq 0$ for all $q \in \mathbb{R}$.

(S3.3) There exists $b_f > 0$ such that $f'(q)q$ is non-decreasing for $q \in (b_f, +\infty)$, and $f'(q)q \to +\infty$ as $q \to +\infty$.

(S3.4) Let $g: [f(b_f), +\infty) \to [b_f, +\infty)$ be the inverse function of $f$ on the domain $[b_f, +\infty)$. There exists $p \geq 0$ such that for all $x > f(b_f)$, $y > b_f$,

$$
\frac{g'(x)}{g(x)} \leq \frac{p}{x} \quad \text{and} \quad \frac{f''(y)}{f'(y)} \leq \frac{p}{y} .
$$

It can be verified that (S3) is satisfied by exponential loss and logistic loss. Now we explain each of the assumptions in (S3). (S3.2) and (S3.3) are essentially the same as (B3.2) and (B3.3). (S3.1) and (B3.1) are the same except that (S3.1) assumes $f$ is $C^2$-smooth rather than $C^1$-smooth.

(S3.4) can also be written in the following form:

$$
\frac{d}{dx} \log g'(x) \leq p \cdot \frac{d}{dx} \log x \quad \text{and} \quad \frac{d}{dy} \log f'(y) \leq p \cdot \frac{d}{dy} \log y .
$$

That is, $\log g'(x)$ and $\log f'(y)$ grow no faster than $O(\log x)$ and $O(\log y)$, respectively. In fact, (B3.4) can be deduced from (S3.4). Recall that (B3.4) ensures that $\Theta(\log(\eta)) = g'(\Theta(\eta))$ and $\Theta(\log(\eta)) = f'(\Theta(\eta))$. Thus, (S3.4) also gives us the interchangeability between $f'$, $g'$ and $\Theta$.

**Lemma F.1.** (S3.4) implies (B3.4) with $b_g = \max\{2f(b_f), f(2b_f)\}$ and $K = 2^p$.

**Proof.** Fix $x \in (b_g, +\infty)$, $y \in (g(b_g), +\infty)$ and $\eta \in [1/2, 1)$. Integrating (23) on both sides of the inequalities from $\theta x$ to $x$ and $\theta y$ to $y$, we have

$$
| \log g'(x) - \log g'(\theta x) | \leq p \cdot \log \frac{1}{\theta} \quad \text{and} \quad | \log f'(y) - \log f'(\theta y) | \leq p \cdot \log \frac{1}{\theta} .
$$

Therefore, we have $g'(x) \leq \theta^{-p} g'(\theta x) \leq K \theta^{-p} g'(\theta x)$ and $f'(y) \leq \theta^{-p} f'(\theta y) \leq K f'(\theta x)$.

To extend our results to general loss functions, we need to redefine $\kappa(x)$, $\lambda(x)$, $\phi(x)$, $C, H(x)$ in order. For $\kappa(x)$ and $\lambda(x)$, we redefine them as follows:

$$
\kappa(x) := \sup \left\{\frac{w(\log \frac{1}{w})^2 - 2/L}{g'(\log \frac{1}{w})^2} : w \in (0, x)\right\} \quad \lambda(x) := \frac{g'(\log \frac{1}{x})}{g(\log \frac{1}{x})} .
$$
By Lemma D.1 \[ w(\log \frac{1}{\varepsilon} - 2/L) = O \left( w(\log \frac{1}{\varepsilon})^4 - 2/L / g(\log \frac{1}{\varepsilon})^2 \right) \rightarrow 0. \] So \( \kappa(x) \) is well-defined.

Using \( \lambda(x) \), we can define \( \phi(x) \) and \( \hat{\gamma}(\theta) \) as follows.

\[
\phi'(x) := -\sup \left\{ \frac{\lambda(w)}{w} \left( 1 + 2(1 + \lambda(w)/L)\mu(w) \right) : w \in [x, \mathcal{L}(t_0)] \right\}
\]

\[
\phi(x) := \log \left( g(\log \frac{1}{x}) \right) + \int_0^x \left( \phi'(w) + \frac{\lambda(w)}{w} \right) \, dw
\]

\[
\hat{\gamma}(\theta) := \frac{e^{\phi(\mathcal{L}(\theta))}}{\rho^L}.
\]

Using a similar argument as in Lemma E.1, we can show that \( \hat{\gamma}(\theta) \) is well-defined and \( \hat{\gamma}(\theta) < \gamma(\theta) := g(\log \frac{1}{L})/\rho^L \leq \hat{\gamma}(\theta) \). When \( \mathcal{L}(\theta) \rightarrow 0 \), we also have \( \gamma(\theta)/\hat{\gamma}(\theta) \rightarrow 1 \).

For \( C_n \), we define it to be the following.

\[
C_n := \frac{1}{2} B_1 \left( \frac{B_0}{\gamma(0)} \right)^p B_1 + \frac{2^{p+1}}{\log \frac{1}{\gamma(0)}} (pB_1 + B_2) \left( \frac{B_0}{\gamma(0)} \right)^p \min \left\{ \gamma(t_0)^{-2-2/L}, B_0^{-2-2/L} \right\}
\]

Finally, the definitions for \( \mu(x) := (\log \frac{1}{\mathcal{L}(x)})/(2\log \frac{1}{\varepsilon}) \) and \( H(x) := \mu(x)/(C_n\kappa(x)) \) in (S5) remain unchanged except that \( \kappa(x) \) and \( C_n \) now use the new definitions.

Similar as gradient flow, we define \( \nu(t) := \sum_{n=1}^N e^{-f(q_n(t))} f'(q_n(t)) q_n(t) \). The key idea behind the above definitions is that we can prove similar bounds for \( \nu(t), \|\nabla \mathcal{L}\|_2, \|\nabla^2 \mathcal{L}\|_2 \) as Corollary E.6 and Lemma E.7.

**Lemma F.2.** \( (\theta(t), -\nabla \mathcal{L}(t)) = \mathcal{L}(t) \). If \( \mathcal{L}(t) < e^{-f(b_f)} \), then \( \nu(t) \geq \frac{\mathcal{L}(t)}{\lambda(\mathcal{L}(t))} \) and \( \lambda(\mathcal{L}(t)) \) has the lower bound \( \lambda(\mathcal{L}(t)) \leq 2^{p+1} \left( \log \frac{1}{\mathcal{L}(t)} \right)^{-1} \).

**Proof.** It can be easily proved by combining Theorem B.4 Lemma B.5 and Lemma D.1 together. \( \square \)

**Lemma F.3.** For any \( \theta \), if \( \mathcal{L}(\theta) \leq \mathcal{L}(t_0), \hat{\gamma}(\theta) \geq \hat{\gamma}(t_0) \), then

\[
\|\nabla \mathcal{L}(\theta)\|_2^2 \leq 2C_n\kappa(\mathcal{L}(\theta))\mathcal{L}(\theta) \quad \text{and} \quad \|\nabla^2 \mathcal{L}(\theta)\|_2 \leq 2C_n\kappa(\mathcal{L}(\theta)).
\]

**Proof.** Note that a direct corollary of (23) is that \( f'(x_1) \leq f'(x_2)(x_1/x_2)^p \) for \( x_1 \geq x_2 > b_f \). So we have

\[
f'(q_n) \leq f' \left( g(\log \frac{1}{\mathcal{L}}) \left( \frac{q_n}{g(\log \frac{1}{\mathcal{L}})} \right)^p \right) = \frac{1}{\rho^L} \left( \frac{q_n}{\rho^L} \right)^p \leq \frac{R}{g'(\log \frac{1}{\mathcal{L}})},
\]

where \( R := (B_0/\gamma(t_0))^p \). Applying (S3.4), we can also deduce that

\[
|f''(q_n)| \leq \frac{pR}{g(\log \frac{1}{\mathcal{L}})g'(\log \frac{1}{\mathcal{L}})}.
\]

Now we bound \( \|\nabla \mathcal{L}\|_2 \) and \( \|\nabla^2 \mathcal{L}\|_2 \). By the chain rule, we have

\[
\|\nabla \mathcal{L}\|_2 = \left\| -\sum_{n=1}^N e^{-f(q_n)} f'(q_n) \nabla q_n \right\|_2 \leq B_1 R \cdot \frac{\mathcal{L}^L}{g'(\log \frac{1}{\mathcal{L}})}
\]

\[
\|\nabla^2 \mathcal{L}\|_2 = \left\| \sum_{n=1}^N e^{-f(q_n)} ((f'(q_n)^2 - f''(q_n)) \nabla q_n \nabla q_n^\top - f'(q_n) \nabla^2 q_n) \right\|_2 \leq \left\| \sum_{n=1}^N e^{-f(q_n)} \left( \frac{R^2}{g'(\log \frac{1}{\mathcal{L}})^2} + \frac{pR}{g(\log \frac{1}{\mathcal{L}})g'(\log \frac{1}{\mathcal{L}})} \right) B_1^2 \rho^{2L-2} + \frac{R}{g'(\log \frac{1}{\mathcal{L}})B_2 \rho^{L-2}} \right\|_2 \leq \frac{\mathcal{L}^L}{g'(\log \frac{1}{\mathcal{L}})^2} \left( (R + p\lambda(\mathcal{L})) B_1^2 + \frac{g'(\log \frac{1}{\mathcal{L}})}{\rho^L} B_2 \right) R.
\]

39
For $\rho$ we have $\hat{\gamma}(t_0)\rho^L \leq \log \frac{1}{L} \leq B_0\rho^L$, so we can bound $\rho^{2L-2}$ by $\rho^{2L-2} \leq M(\log \frac{1}{L})^{2L-2}$, where $M := \min \left\{ \hat{\gamma}(t_0)^{-2+2/L}, B_0^{-2+2/L} \right\}$. By Lemma F.2 we have $\lambda(L) \leq \frac{2^{p+1}}{\log L}$, and also $\frac{\lambda(L)}{\lambda(\rho)} = \lambda(L) \cdot \hat{\gamma}(t_0) \leq \frac{2^{p+1}B_1}{\log \frac{1}{L}}$. Combining all these formulas together gives

$$\|\nabla \mathcal{L}\|^2 \leq B_1^2 R^2 M \frac{\mathcal{L}^2}{g'(\log \frac{1}{L})^2} \left( \log \frac{1}{L} \right)^{2-2/L} \leq 2C_\eta \mathcal{L} \mathcal{L}$$

$$\|\nabla^2 \mathcal{L}\| \leq \left( \left( R + \frac{2^{p+1}}{\log \frac{1}{L}} \right) B_1^2 + \frac{2^{p+1}B_1B_2}{\log \frac{1}{L}} \right) R \mathcal{L} \frac{\mathcal{L}^2}{g'(\log \frac{1}{L})^2} \left( \log \frac{1}{L} \right)^{2-2/L} \leq 2C_\eta \mathcal{L},$$

which completes the proof.

With Corollary E.6 and Lemma E.7, we can prove Lemma E.8 with exactly the same argument. Then Theorem E.2, E.3, E.4 can also be proved similarly. For Theorem E.5 we can follow the argument for gradient flow to show that it holds with slightly different tight bounds:

**Theorem F.4.** Under assumptions (S1), (A2), (S3), (A4), (S5), we have the following tight rates for training loss and weight norm:

$$\mathcal{L}(t) = \Theta \left( \frac{g(\log T)^{2/L}}{T(\log T)^2} \right) \quad \text{and} \quad \rho(t) = \Theta(g(\log T)^{1/L}),$$

where $T = \sum_{\tau=t_0}^{t-1} \eta(\tau)$.

**G Extension: Multi-class Classification**

In this section, we generalize our results to multi-class classification with cross-entropy loss. This part of analysis is inspired by Theorem 1 in [Zhang et al. 2019](https://arxiv.org/abs/1909.04222), which gives a lower bound for the gradient in terms of the loss $\mathcal{L}$.

Since now a neural network has multiple outputs, we need to redefine our notations. Let $C$ be the number of classes. The output of a neural network $\Phi$ is a vector $\Phi(\theta; x) \in \mathbb{R}^C$. We use $\Phi_j(\theta; x) \in \mathbb{R}$ to denote the $j$-th output of $\Phi$ on the input $x \in \mathbb{R}^d$. A dataset is denoted by $\mathcal{D} = \{x_n, y_n\}_{n=1}^N = \{(x_n, y_n) : n \in [N]\}$, where $x_n \in \mathbb{R}^d$ is a data input and $y_n \in [C]$ is the corresponding label. The loss function of $\Phi$ on the dataset $\mathcal{D}$ is defined as

$$\mathcal{L}(\theta) := \sum_{n=1}^N \frac{1}{N} \log \sum_{c=1}^C e^{\Phi_c(\theta; x_n)}.$$

The margin for a single data point $(x_n, y_n)$ is defined to be $q_{\min}(\theta) := \Phi_{y_n}(\theta; x_n) - \max_{j \neq y_n} \Phi_j(\theta; x_n)$, and the margin for the entire dataset is defined to be $q_{\min}(\theta) = \min_{n \in [N]} q_{\min}(\theta)$. We define the normalized margin to be $\tilde{\gamma}(\theta) := \min_{\theta_{\min}(\theta)} = q_{\min}(\theta) / \rho T$, where $\rho := \sqrt{\sum_{j=1}^N \Phi^2_j(\theta; x_n)}$ and $\theta := \theta / \rho \in S^{d-1}$ as usual.

Let $\ell(q) := \log(1 + e^{-q})$ be the logistic loss. Recall that $\ell(q)$ satisfies (B3). Let $f(q) = -\log \ell(q) = -\log \log(1 + e^{-q})$. Let $g$ be the inverse function of $f$. So $g(q) = -\log(e^{q} - 1)$.

The cross-entropy loss can be rewritten in other ways. Let $s_{nj} := \Phi_{y_n}(\theta; x_n) - \Phi_j(\theta; x_n)$. Let $q_n := -\text{LSE}\left( \{s_{nj} : j \neq y_n\} \right) = -\log \left( \sum_{j \neq y_n} e^{-s_{nj}} \right)$. Then

$$\mathcal{L}(\theta) := \sum_{n=1}^N \log \left( 1 + \sum_{j \neq y_n} e^{-s_{nj}} \right) = \sum_{n=1}^N \log(1 + e^{-q_n}) = \sum_{n=1}^N \ell(q_n) = \sum_{n=1}^N e^{-f(q_n)}.$$

**Gradient Flow.** For gradient flow, we assume the following:

(M1). (Regularity). For any fixed $x$ and $j \in [C]$, $\Phi_j(\cdot; x)$ is locally Lipschitz and admits a chain rule;
(M2). (Homogeneity). There exists $L > 0$ such that $\forall j \in [C], \forall \alpha > 0 : \Phi_j(\alpha \theta; x) = \alpha^C \Phi_j(\theta; x)$;

(M3). (Cross-entropy Loss). $\mathcal{L}(\theta)$ is defined as the cross-entropy loss on the training set;

(M4). (Separability). There exists a time $t_0$ such that $\mathcal{L}(t_0) < \log 2$.

If $\mathcal{L} < \log 2$, then $\sum_{j \neq y_n} e^{-s_{nj}} < 1$ for all $n \in [N]$, and thus $s_{nj} > 0$ for all $n \in [N], j \in [C]$. So (M4) ensures the separability of training data.

**Definition G.1.** For cross-entropy loss, the smoothed normalized margin $\tilde{\gamma}(\theta)$ of $\theta$ is defined as

$$\tilde{\gamma}(\theta) := \frac{\ell^{-1}(\mathcal{L})}{\rho^L} = \frac{-\log(e^\mathcal{L} - 1)}{\rho^L},$$

where $\ell^{-1}(\cdot)$ is the inverse function of the logistic loss $\ell(\cdot)$.

Theorem 4.1 and 4.4 still hold. Here we redefine the optimization problem (P) to be

$$\min \quad \frac{1}{2} \|\theta\|_2^2$$

s.t. $s_{nj}(\theta) \geq 1 \quad \forall n \in [N], \forall j \in [C] \setminus \{y_n\}$

Most of our proofs are very similar as before. Here we only show the proof for the generalized version of Lemma 5.1

**Lemma G.2.** Lemma 5.1 is also true for the smoothed normalized margin $\tilde{\gamma}$ defined in Definition G.1.

**Proof.** Define $\nu(t)$ by the following formula:

$$\nu(t) := \sum_{n=1}^{N} \sum_{j \neq y_n} e^{-s_{nj}} s_{nj} \frac{1}{1 + \sum_{j \neq y_n} e^{-s_{nj}}}.$$

Using a similar argument as in Theorem B.4, it can be proved that $\frac{1}{2} \frac{dp^2}{dt} = L \nu(t)$ for a.e. $t > 0$.

It can be shown that Lemma B.5, which asserts that $\nu(t) \geq \frac{g(\log \frac{1}{\gamma})}{g'(\log \frac{1}{\gamma})} \mathcal{L}$, still holds for this new definition of $\nu(t)$. By definition, $s_{nj} \geq q_n$ for $j \neq y_n$. Also note that $e^{-\tilde{q}_n} \geq e^{-q_n}$. So $s_{nj} \geq q_n \geq \tilde{q}_n$. Then we have

$$\nu(t) \geq \sum_{n=1}^{N} \sum_{j \neq y_n} e^{-s_{nj}} \frac{e^{-\tilde{q}_n}}{1 + \sum_{j \neq y_n} e^{-s_{nj}}} \cdot \tilde{q}_n \geq \sum_{n=1}^{N} \frac{e^{-\tilde{q}_n}}{1 + e^{-\tilde{q}_n}} \cdot \tilde{q}_n = \sum_{n=1}^{N} e^{-f(\tilde{q}_n)} f'(\tilde{q}_n)\tilde{q}_n.$$

Note that $\mathcal{L} = \sum_{n=1}^{N} e^{-f(q_n)}$. Then using Lemma B.5 for logistic loss can conclude that $\nu(t) \geq \frac{g(\log \frac{1}{\gamma})}{g'(\log \frac{1}{\gamma})} \mathcal{L}$.

The rest of the proof for this lemma is exactly the same as that for Lemma 5.1.

**Gradient Descent.** For gradient descent, we only need to replace (M1) with (S1) and make the assumption (S5) on the learning rate.

(S1). (Smoothness). For any fixed $x$ and $j \in [C]$, $\Phi_j(\cdot; x)$ is $C^2$-smooth;

(S5). (Learning rate condition). $\sum_{t \geq 0} \eta(t) = +\infty$ and $\eta(t) \leq H(\mathcal{L}(\theta(t)))$.

Here $H(x) := \mu(x)/(C_\eta \kappa(x))$ is defined to be the same as in Appendix F (when $\ell(\cdot)$ is set to logistic loss) except that we use another $C_\eta$ which will be specified later.

We only need to show that Lemma F.2 and Lemma F.3 continue to hold. Using the same argument as we do for gradient flow, we can show that $\gamma(t)$ and $\lambda(x)$ do satisfy the propositions in Lemma F.2.
For Lemma F.3, we first note the original definition of \( C_\eta \) involves \( B_0, B_1, B_2 \), which are undefined in the multi-class setting. So now we redefine them as

\[
B_0 := \sup \left\{ s_{nj} : \theta \in S^{d-1}, n \in [N], j \in [C] \right\}, \\
B_1 := \sup \left\{ \| \nabla s_{nj} \|_2 : \theta \in S^{d-1}, n \in [N], j \in [C] \right\}, \\
B_2 := \sup \left\{ \| \nabla^2 s_{nj} \|_2 : \theta \in S^{d-1}, n \in [N], j \in [C] \right\}.
\]

By the property of LSE, we can use \( B_0, B_1, B_2 \) to bound \( \tilde{q}_n, \nabla \tilde{q}_n, \nabla^2 \tilde{q}_n \).

\[
\tilde{q}_n \leq q_n \leq B_0 \rho^L, \\
\| \nabla \tilde{q}_n \|_2 = \left\| \sum_{j=1}^C \frac{e^{-s_{nj}}}{\sum_{k=1}^C e^{-s_{nk}}} \nabla s_{nj} \right\|_2 \leq \sum_{j=1}^C \frac{e^{-s_{nj}}}{\sum_{k=1}^C e^{-s_{nk}}} \| \nabla s_{nj} \|_2 \leq B_1 \rho^{L-1}, \\
\| \nabla^2 \tilde{q}_n \|_2 \leq B_2 \rho^{L-2} + 2 B_1^{2} \rho^{2L-2}.
\]

Using these bounds, we can prove with the constant \( C_\eta \) defined as follows:

\[
C_\eta := \frac{1}{2} \left( \left( \frac{B_0}{\gamma(t_0)} \right)^p + \frac{p^{p+1}}{\log \frac{1}{L(t_0)}} \right) B_1^2 + (2 \log 2) \left( \frac{B_2}{\rho(t_0)^L} + 2 B_1^2 \right) \left( \frac{B_0}{\gamma(t_0)} \right)^p M,
\]

where \( M := \min \left\{ \gamma(t_0)^{-2+2/L}, B_0^{-2+2/L} \right\} \).

H EXTENSION: MULTI-HOMOGENEOUS MODELS

In this section, we extend our results to multi-homogeneous models. For this, the main difference from the proof for homogeneous models is that now we have to separate the norm of each homogeneous parts of the parameter, rather than consider them as a whole. So only a small part to proof needs to be changed. We focus on gradient flow, but it is worth to note that following the same argument, it is not hard to extend the results to gradient descent.

Let \( \Phi(w_1, \ldots, w_m; x) \) be \((k_1, \ldots, k_m)\)-homogeneous. Let \( \rho_i = \| w_i \|_2 \) and \( \hat{w}_i = \frac{w_i}{\| w_i \|_2} \). The smoothed normalized margin defined in \([F] \) can be rewritten as follows:

**Definition H.1.** For a multi-homogeneous model with loss function \( \ell(\cdot) \) satisfying (B3), the smoothed normalized margin \( \gamma(\theta) \) of \( \theta \) is defined as

\[
\gamma(\theta) := \frac{g(\log \frac{1}{L})}{\prod_{i=1}^m \rho_i} = \frac{\ell^{-1}(L)}{\prod_{i=1}^m \rho_i}.
\]

We only prove the generalized version of Lemma 5.1 here. The other proofs are almost the same.

**Lemma H.2.** For all \( t > t_0 \),

\[
\frac{d}{dt} \log \rho_i > 0 \quad \text{for all } i \in [m] \quad \text{and} \quad \frac{d}{dt} \log \gamma \geq \sum_{i=1}^m k_i \left( \frac{d}{dt} \log \rho_i \right)^{-1} \left\| \frac{d \hat{w}_i}{dt} \right\|_2^2. \quad (24)
\]

**Proof.** Note that \( \frac{d}{dt} \log \rho_i = \frac{1}{\rho_i} \frac{d\rho_i}{dt} = \frac{k_i \ell(t)}{\rho_i^2} \) by Theorem B.4. It simply follows from Lemma B.5 that \( \frac{d}{dt} \log \rho > 0 \) for a.e. \( t > t_0 \). And it is easy to see that \( \frac{d\gamma}{dt} = \log (g(\log \frac{1}{L})/\rho^L) \) exists for all
\( t \geq t_0 \). By the chain rule and Lemma [B.5] we have
\[
\frac{d}{dt} \log \tilde{\gamma} = \frac{d}{dt} \left( \log \left( g(\log \frac{1}{L}) - \sum_{i=1}^{m} k_i \log \rho_i \right) \right) = g'(\log \frac{1}{L}) \cdot \frac{1}{L} \cdot \left( \frac{dL}{dt} - \sum_{i=1}^{m} k_i^2 \nu(t) \rho_i^2 \right) \geq \frac{1}{\nu(t)} \cdot \left( \frac{dL}{dt} - \sum_{i=1}^{m} k_i^2 \nu(t) \right) \geq \frac{1}{\nu(t)} \cdot \left( \frac{dL}{dt} - \sum_{i=1}^{m} k_i^2 \nu(t)^2 \right).
\]

On the one hand, \( -\frac{dL}{dt} = \sum_{i=1}^{m} \| \frac{dw_i}{dt} \|^2 \) for a.e. \( t > 0 \) by Lemma I.3 in calculus. However, for non-differentiable functions, it is difficult to find a natural definition of the chain rule equation. To solve this issue, Clarke proposed Clarke’s subdifferential (Clarke, 1975; 1990; Clarke et al., 2008) for locally Lipschitz functions, for which the chain rule holds as an inclusion rather than an equation:
\[
\text{Theorem I.1} \quad \partial f \circ z \subseteq \text{conv} \left\{ \sum_{i=1}^{n} \alpha_i h_i : \alpha \in \partial f(z_1(x)), \ldots, z_n(x)), h_i \in \partial z_i(x) \right\}.
\]

The only place we need to change in the proof for Lemma [H.2] is that instead of using Lemma B.5, we need to prove \( \nu(t) \geq \frac{g'(\log \frac{1}{L})}{g'(\log \frac{1}{L})} \) in a similar way as in Lemma G.2. The other parts of the proof are exactly the same as before.

I Chain Rules for Non-differentiable Functions

In this section, we provide some background on the chain rule for non-differentiable functions. The ordinary chain rule for differentiable functions is a very useful formula for computing derivatives in calculus. However, for non-differentiable functions, it is difficult to find a natural definition of subdifferential so that the chain rule equation holds exactly. To solve this issue, Clarke proposed Clarke’s subdifferential (Clarke, 1975; 1990; Clarke et al., 2008) for locally Lipschitz functions, for which the chain rule holds as an inclusion rather than an equation:

Theorem 1.1 (Theorem 2.3.9 and 2.3.10 of Clarke, 1990). Let \( z_1, \ldots, z_n : \mathbb{R}^d \to \mathbb{R} \) and \( f : \mathbb{R}^n \to \mathbb{R} \) be locally Lipschitz functions. Let \( (f \circ z)(x) = f(z_1(x), \ldots, z_n(x)) \) be the composition of \( f \) and \( z \).

For analyzing gradient flow, the chain rule is crucial. For a differentiable loss function \( \mathcal{L}(\theta) \), we can see from the chain rule that the function value keeps decreasing along the gradient flow \( \frac{d\theta(t)}{dt} = -\nabla \mathcal{L}(\theta(t)) \):
\[
\frac{d\mathcal{L}(\theta(t))}{dt} = -\nabla \mathcal{L}(\theta(t)) \cdot \frac{d\theta(t)}{dt} = -\left\| \frac{d\theta(t)}{dt} \right\|^2.
\]
But for locally Lipschitz functions which could be non-differentiable, \[25\] may not hold in general since Theorem I.1 only holds for an inclusion.

Following (Davis et al. 2020; Drusvyatskiy et al. 2015), we consider the functions that admit a chain rule for any arc.

**Definition I.2 (Chain Rule).** A locally Lipschitz function \(f : \mathbb{R}^d \to \mathbb{R}\) admits a chain rule if for any arc \(z : [0, +\infty) \to \mathbb{R}^d\),
\[
\forall h_i \in \partial^\circ f(z(t)) : (f \circ z)'(t) = \langle h_i, z'(t) \rangle
\]
holds for a.e. \(t > 0\).

It is shown in (Davis et al. 2020; Drusvyatskiy et al. 2015) that a generalized version of \[25\] holds for such functions:

**Lemma I.3 (Lemma 5.2 [Davis et al. 2020]).** Let \(L : \mathbb{R}^d \to \mathbb{R}\) be a locally Lipschitz function that admits a chain rule. Let \(\theta : [0, +\infty) \to \mathbb{R}^d\) be the gradient flow on \(L\):
\[
\frac{d\theta(t)}{dt} \in -\partial^\circ L(\theta(t)) \quad \text{for a.e.} \ t \geq 0,
\]
Then
\[
\frac{dL(\theta(t))}{dt} = -\left\| \frac{d\theta(t)}{dt} \right\|^2_2 = -\min\{||h||^2_2 : h \in \partial^\circ L(\theta(t))\}
\]
holds for a.e. \(t > 0\).

We can see that \(C^1\)-smooth functions admit chain rules. As shown in (Davis et al. 2020), if a locally Lipschitz function is subdifferentiably regular or Whitney \(C^1\)-stratifiable, then it admits a chain rule. The latter one includes a large family of functions, e.g., semi-algebraic functions, semi-analytic functions, and definable functions in an o-minimal structure (Coste 2002; van den Dries & Miller 1996).

It is worth noting that the class of functions that admits chain rules is closed under composition. This is indeed a simple corollary of Theorem I.1

**Theorem I.4.** Let \(z_1, \ldots, z_n : \mathbb{R}^d \to \mathbb{R}\) and \(f : \mathbb{R}^n \to \mathbb{R}\) be locally Lipschitz functions and assume all of them admit chain rules. Let \((f \circ z)(x) = f(z_1(x), \ldots, z_n(x))\) be the composition of \(f\) and \(z\). Then \(f \circ z\) also admits a chain rule.

**Proof.** We can see that \(f \circ z\) is locally Lipschitz. Let \(x : [0, +\infty) \to \mathbb{R}^d, t \mapsto x(t)\) be an arc. First, we show that \(z \circ x : [0, +\infty) \to \mathbb{R}^d, t \mapsto z(x(t))\) is also an arc. For any closed sub-interval \(I, z(x(I))\) must be contained in a compact set \(U\). Then it can be shown that the locally Lipschitz continuous function \(z\) is (globally) Lipschitz continuous on \(U\). By the fact that the composition of a Lipschitz continuous and an absolutely continuous function is absolutely continuous, \(z \circ x\) is absolutely continuous on \(I\), and thus it is an arc.

Since \(f\) and \(z\) admit chain rules on arcs \(z \circ x\) and \(x\) respectively, the following holds for a.e. \(t > 0\),
\[
\forall \alpha_i \in \partial^\circ f(z(x(t))) : (f \circ (z \circ x))'(t) = \langle \alpha_i, (z \circ x)'(t) \rangle,
\]
\[
\forall h_i \in \partial^\circ z_i(x(t)) : (z_i \circ x)'(t) = \langle h_i, x'(t) \rangle.
\]
Combining these we obtain that for a.e. \(t > 0\),
\[
(f \circ z \circ x)'(t) = \sum_{i=1}^n \alpha_i \langle h_i, x'(t) \rangle,
\]
for all \(\alpha \in \partial^\circ f(z(x(t)))\) and for all \(h_i \in \partial^\circ z_i(x(t))\). The RHS can be rewritten as \(\sum_{i=1}^n \alpha_i (h_i, x'(t))\). By Theorem I.1 every \(k \in \partial^\circ (f \circ z)(x(t))\) can be written as a convex combination of a finite set of points in the form of \(\sum_{i=1}^n \alpha_i h_i\). So \((f \circ z \circ x)'(t) = \langle k, x'(t) \rangle\) holds for a.e. \(t > 0\). 

\(\square\)
J MEXICAN HAT

In this section, we give an example to illustrate that gradient flow does not necessarily converge in direction, even for $C^\infty$-smooth homogeneous models.

It is known that gradient flow (or gradient descent) may not converge to any point even when optimizing an $C^\infty$ function (Curry, 1944; Zoutendijk, 1976; Palis & De Melo, 2012; Absil et al., 2005).

One famous counterexample is the “Mexican Hat” function described in (Absil et al., 2005):

$$f(u, v) := f(r \cos \varphi, r \sin \varphi) := \begin{cases} e^{-\frac{r^2}{4(1-r^2)}} (1 - C(r) \sin \left( \varphi - \frac{1}{1-r^2} \right)) & r < 1 \\ 0 & r \geq 1 \end{cases}$$

where $C(r) = \frac{4r^4}{4r^4 + (1-r^2)^2} \in [0, 1]$. It can be shown that $f$ is $C^\infty$-smooth on $\mathbb{R}^2$ but not analytic. See Figure 2 for a plot for $f(u, v)$.

![Figure 2: A plot for the Mexican Hat function $f(u, v)$.](image)

However, the Mexican Hat function is not homogeneous, and [Absil et al., 2005] did not consider the directional convergence, either. To make it homogeneous, we introduce an extra variable $z$, and normalize the parameter before evaluating $f$. In particular, we fix $L > 0$ and define

$$h(\theta) = h(x, y, z) = \rho^z (1 - f(u, v)) \quad \text{where} \quad u = x/\rho, \quad v = y/\rho, \quad \rho = \sqrt{x^2 + y^2 + z^2}.$$  

We can show the following theorem.

**Theorem J.1.** Consider gradient flow on $\mathcal{C}(\theta) = \sum_{n=1}^{N} e^{-q_n(\theta)}$, where $q_n(\theta) = h(\theta)$ for all $n \in \mathbb{N}$. Suppose the polar representation of $(u, v)$ is $(r \cos \varphi, r \sin \varphi)$. If $0 < r < 1$ and $\varphi = \frac{1}{1-r^2}$ holds at time $t = 0$, then $\frac{\theta(t)}{\|\theta(t)\|_2}$ does not converge to any point, and the limit points of $\{\frac{\theta(t)}{\|\theta(t)\|_2} : t > 0\}$ form a circle $\{(x, y, z) \in S^2 : x^2 + y^2 = 1, z = 0\}$.

**Proof.** Define $\psi = \varphi - \frac{1}{1-r^2}$. Our proof consists of two parts, following from the idea in [Absil et al., 2005]. First, we show that $\frac{\partial \psi}{\partial r} = 0$ as long as $\psi = 0$. Then we can infer that $\psi = 0$ for all $t \geq 0$. Next, we show that $r \to 1$ as $t \to +\infty$. Using $\psi = 0$, we know that the polar angle $\varphi \to +\infty$ as $t \to +\infty$. Therefore, $(u, v)$ circles around $\{(u, v) : u^2 + v^2 = 1\}$, and thus it does not converge.

**Proof for $\frac{\partial \psi}{\partial r} = 0$.** For convenience, we use $w$ to denote $z/\rho$. By simple calculation, we have the following formulas for partial derivatives:

$$\begin{align*}
\frac{\partial h}{\partial x} &= \rho^{L-1} \left( Lu(1 - f) - (1 - u^2) \frac{\partial f}{\partial u} + uv \frac{\partial f}{\partial v} \right) \\
\frac{\partial h}{\partial y} &= \rho^{L-1} \left( Lv(1 - f) + uv \frac{\partial f}{\partial u} - (1 - v^2) \frac{\partial f}{\partial v} \right) \\
\frac{\partial h}{\partial z} &= \rho^{L-1} \left( Lw(1 - f) + uw \frac{\partial f}{\partial u} + vw \frac{\partial f}{\partial v} \right)
\end{align*}$$

$$\begin{align*}
\frac{\partial \psi}{\partial x} &= u \frac{\partial f}{\partial u} - v \frac{\partial f}{\partial v} \\
\frac{\partial \psi}{\partial y} &= \frac{\partial f}{\partial u} + u \frac{\partial f}{\partial v} \\
\frac{\partial \psi}{\partial z} &= \frac{\partial f}{\partial u} + v \frac{\partial f}{\partial v}
\end{align*}$$
For gradient flow, we have
\[
\frac{d\theta}{dt} = Ne^{-h} \nabla h
\]
\[
\frac{du}{dt} = \frac{1}{\rho} \left( (1-u^2) \frac{dx}{dt} - uw \frac{dy}{dt} - uw \frac{dz}{dt} \right) = Ne^{-h} \rho^{L-2} \left( -(1-u^2) \frac{\partial f}{\partial u} + uw \frac{\partial f}{\partial v} \right)
\]
\[
\frac{dv}{dt} = \frac{1}{\rho} \left( (1-v^2) \frac{dy}{dt} - uw \frac{dx}{dt} - vu \frac{dz}{dt} \right) = Ne^{-h} \rho^{L-2} \left( -(1-v^2) \frac{\partial f}{\partial v} + vu \frac{\partial f}{\partial u} \right)
\]

By writing down the movement of \((u, v)\) in the polar coordinate system, we have
\[
\frac{dr}{dt} = \frac{u}{r} \frac{du}{dt} + \frac{v}{r} \frac{dv}{dt} = -Ne^{-h} \rho^{L-2} \left( u(1-r^2) \frac{\partial f}{\partial u} + v(1-r^2) \frac{\partial f}{\partial v} \right)
\]
\[
\frac{d\varphi}{dt} = \frac{v}{r} \frac{du}{dt} - \frac{u}{r} \frac{dv}{dt} = Ne^{-h} \rho^{L-2} \left( v \frac{\partial f}{\partial u} - u \frac{\partial f}{\partial v} \right)
\]

For \(\psi = 0\), the partial derivatives of \(f\) with respect to \(r\) and \(\varphi\) can be evaluated as follows:
\[
\frac{\partial f}{\partial r} = \frac{d}{dr} \left( e^{-\frac{1}{1-r^2}} - e^{-\frac{1}{1-r^2}} C(r) \sin \psi - e^{-\frac{1}{1-r^2}} C(r) \cos \psi \cdot \frac{\partial \psi}{\partial r} \right)
\]
\[
= \frac{d}{dr} \left( e^{-\frac{1}{1-r^2}} \right) + e^{-\frac{1}{1-r^2}} C(r) \frac{d}{dr} \left( \frac{1}{1-r^2} \right)
\]
\[
= -\frac{2r}{(1-r^2)^2}(1-C(r))e^{-\frac{1}{1-r^2}}.
\]
\[
\frac{\partial f}{\partial \varphi} = -e^{-\frac{1}{1-r^2}} C(r) \cos \psi \cdot \frac{\partial \psi}{\partial \varphi}
\]
\[
= -e^{-\frac{1}{1-r^2}} C(r).
\]

So if \(\psi = 0\), then \(\frac{d\psi}{dt} = 0\) by the direct calculation below:
\[
\frac{d\psi}{dt} = \frac{d\varphi}{dt} - \frac{d}{dr} \left( \frac{1}{1-r^2} \right) \cdot \frac{dr}{dt}
\]
\[
= Ne^{-h} \rho^{L-2} \left( \frac{1}{r^2} \frac{\partial f}{\partial \varphi} + \frac{2r}{(1-r^2)^2} \frac{\partial f}{\partial r} \right)
\]
\[
= Ne^{-h} \rho^{L-2} e^{-\frac{1}{1-r^2}} \left( \frac{1}{r^2} + \frac{4r^2}{(1-r^2)^2} \right) C(r) - \frac{4r^2}{(1-r^2)^4} = 0.
\]

**Proof for** \(r \to 1\). Let \((\bar{u}, \bar{v})\) be a convergent point of \(\{(u, v) : t \geq 0\}\). Define \(\bar{r} = \sqrt{u^2 + v^2}\). It is easy to see that \(r \leq 1\) from the normalization of \(\theta\) in the definition. According to Theorem 4.4, we know that \((\bar{u}, \bar{v})\) is a stationary point of \(\hat{\gamma}(u(t), v(t)) = 1 - f(u(t), v(t))\). If \(\bar{r} = 0\), then \(f(\bar{u}, \bar{v}) > f(u(0), v(0))\), which contradicts to the monotonicity of \(\hat{\gamma}(t) = \hat{\gamma}(t) = 1 - f(u(t), v(t))\).

If \(\bar{r} < 1\), then \(\bar{\psi} = 0\) (defined as \(\psi\) for \((\bar{u}, \bar{v})\)). So \(\frac{df}{dr} = -\frac{2r}{(1-r^2)^2}(1-C(r))e^{-\frac{1}{1-r^2}} \neq 0\), which again leads to a contradiction. Therefore, \(\bar{r} = 1\), and thus \(r \to 1\).

**K Experiments**

To validate our theoretical results, we conduct several experiments. We mainly focus on MNIST dataset. We trained two models with Tensorflow. The first one (called the CNN with bias) is a standard 4-layer CNN with exactly the same architecture as that used in MNIST Adversarial Examples.
The layers of this model can be described as $conv-32$ with filter size $5 \times 5$, $max-pool$, $conv-64$ with filter size $3 \times 3$, $max-pool$, $fc-1024$, $fc-10$ in order. Notice that this model has bias terms in each layer, and thus does not satisfy homogeneity. To make its outputs homogeneous to its parameters, we also trained this model after removing all the bias terms except those in the first layer (the modified model is called the CNN without bias). Note that keeping the bias terms in the first layer prevents the model to be homogeneous in the input data while retains the homogeneity in parameters. We initialize all layer weights by He normal initializer (He et al., 2015) and all bias terms by zero. In training the models, we use SGD with batch size 100 without momentum. We normalize all the images to $[0, 1]^{32 \times 32}$ by dividing 255 for each pixel.

K.1 Evaluation for Normalized Margin

In the first part of our experiments, we evaluate the normalized margin every few epochs to see how it changes over time. From now on, we view the bias term in the first layer as a part of the weight in the first layer for convenience. Observe that the CNN without bias is multi-homogeneous in layer weights (see [4] in Section 4.3). So for the CNN without bias, we define the normalized margin $\bar{\gamma}$ as the margin divided by the product of the $L^2$-norm of all layer weights. Here we compute the $L^2$-norm of a layer weight parameter after flattening it into a one-dimensional vector. For the CNN with bias, we still compute the smoothed normalized margin in this way. When computing the $L^2$-norm of every layer weight, we simply ignore the bias terms if they are not in the first layer. For completeness, we include the plots for the normalized margin using the original definition (2) in Figure 3 and 4.

Figure 3: Training CNNs with and without bias on MNIST, using SGD with learning rate 0.01. The training accuracy (left) increases to 100% after about 100 epochs, and the normalized margin with the original definition (right) keeps increasing after the model is fitted.

Figure 4: Training CNNs with and without bias on MNIST, using SGD with the loss-based learning rate scheduler. The training accuracy (left) increases to 100% after about 20 epochs, and the normalized margin with the original definition (middle) increases rapidly after the model is fitted. The right figure shows the change of the relative learning rate $\alpha(t)$ (see (27) for its definition) during training.

https://github.com/MadryLab/mnist_challenge
SGD with Constant Learning Rate. We first train the CNNs using SGD with constant learning rate 0.01. After about 100 epochs, both CNNs have fitted the training set. After that, we can see that the normalized margins of both CNNs increase. However, the growth rate of the normalized margin is rather slow. The results are shown in Figure 1 in Section 1. We also tried other learning rates other than 0.01, and similar phenomena can be observed.

SGD with Loss-based Learning Rate. Indeed, we can speed up the training by using a proper scheduling of learning rates for SGD. We propose a heuristic learning rate scheduling method, called the loss-based learning rate scheduling. The basic idea is to find the maximum possible learning rate at each epoch based on the current training loss (in a similar way as the line search method). See Appendix L.1 for the details. As shown in Figure 1, SGD with loss-based learning rate scheduling decreases the training loss exponentially faster than SGD with constant learning rate. Also, a rapid growth of normalized margin is observed for both CNNs. Note that with this scheduling the training loss can be as small as $10^{-800}$, which may lead to numerical issues. To address such issues, we applied some re-parameterization tricks and numerical tricks in our implementation. See Appendix L.2 for the details.

Experiments on CIFAR-10. To verify whether the normalized margin is increasing in practice, we also conduct experiments on CIFAR-10. We use a modified version of VGGNet-16. The layers of this model can be described as conv-64 x2, max-pool, conv-128 x2, max-pool, conv-256 x3, max-pool, conv-512 x3, max-pool, conv-512 x3, max-pool, fc-10 in order, where each conv has filter size 3 x 3. We train two networks: one is exactly the same as the VGGNet we described, and the other one is the VGGNet without any bias terms except those in the first layer (similar as in the experiments on MNIST). The experiment results are shown in Figure 5 and 6. We can see that the normalize margin is increasing over time.

![Figure 5](image1.png) **Figure 5:** Training VGGNet with and without bias on CIFAR-10, using SGD with learning rate 0.1.

![Figure 6](image2.png) **Figure 6:** Training VGGNet with and without bias on CIFAR-10, using SGD with the loss-based learning rate scheduler.

Test Accuracy. Previous works on margin-based generalization bounds [Neyshabur et al., 2018; Bartlett et al., 2017; Golowich et al., 2018; Li et al., 2018a; Wei et al., 2019; Banburski et al.].
2019) usually suggest that a larger margin implies a better generalization bound. To see whether the generalization error also gets smaller in practice, we plot train and test accuracy for both MNIST and CIFAR-10. As shown in Figure 7, the test accuracy changes only slightly after training with loss-based learning rate scheduling for 10000 epochs, although the normalized margin does increase a lot. We leave it as a future work to study this interesting gap between margin-based generalization bound and generalization error. Concurrent to this work, [Wei & Ma, 2020] proposed a generalization bound based on a new notion of margin called all-layer margin, and showed via experiments that enlarging all-layer margin can indeed improve generalization. It would be an interesting research direction to study how different definitions of margin may lead to different generalization abilities.

![Figure 7: (Left). Training and test accuracy during training CNNs without bias on MNIST, using SGD with the loss-based learning rate scheduler. Every number is averaged over 10 runs. (Right). Training and test accuracy during training VGGNet without bias on CIFAR-10, using SGD with the loss-based learning rate scheduler. Every number is averaged over 3 runs.](image)

**K.2 Evaluation for Robustness**

Recently, robustness of deep learning has received considerable attention (Szegedy et al., 2013; Biggio et al., 2013; Athalye et al., 2018), since most state-of-the-arts deep neural networks are found to be very vulnerable against small but adversarial perturbations of the input points. In our experiments, we found that enlarging the normalized margin can improve the robustness. In particular, by simply training the neural network for a longer time with our loss-based learning rate, we observe noticeable improvements of $L^2$-robustness on both the training set and test set.

We first elaborate the relationship between the normalized margin and the robustness from a theoretical perspective. For a data point $z = (x, y)$, we can define the robustness (with respect to some norm $\| \cdot \|$) of a neural network $\Phi$ for $z$ to be

$$R_\theta(z) := \inf_{x' \in X} \{ \| x - x' \| : (x', y) \text{ is misclassified} \}.$$  

where $X$ is the data domain (which is $[0, 1]^{32 \times 32}$ for MNIST). It is well-known that the normalized margin is a lower bound of $L^2$-robustness for fully-connected networks (See, e.g., Theorem 4 in [Sokolic et al., 2017]). Indeed, a general relationship between those two quantities can be easily shown. Note that a data point $z$ is correctly classified iff the margin for $z$, denoted as $q_\theta(z)$, is larger than 0. For homogeneous models, the margin $q_\theta(z)$ and the normalized margin $\frac{q_\theta(z)}{\| x \|_2}$ for $x$ have the same sign. If $q_\theta(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ is $\beta$-Lipschitz (with respect to some norm $\| \cdot \|$), then it is easy to see that $R_\theta(z) \geq q_\theta(z)/\beta$. This suggests that improving the normalize margin on the training set can improve the robustness on the training set. Therefore, our theoretical analysis suggests that training longer can improve the robustness of the model on the training set.

This observation does match with our experiment results. In the experiments, we measure the $L^2$-robustness of the CNN without bias for the first time its loss decreases below $10^{-10}$, $10^{-15}$, $10^{-20}$, $10^{-120}$ (labelled as model-1 to model-4 respectively). We also measure the $L^2$-robustness for the final model after training for 10000 epochs (labelled as model-5), whose training loss is about $10^{-882}$. The normalized margin of each model is monotone increasing with respect to the number of epochs, as shown in Table I.

Published as a conference paper at ICLR 2020
Figure 8: \(L^2\)-robustness of the models of CNNs without bias trained for different number of epochs (see Table 1 for the statistics of each model). Figures on the first row show the robust accuracy on the training set, and figures on the second row show that on the test set. On every row, the left figure and the right figure plot the same curves but they are in different scales. From model-1 to model-4, noticeable robust accuracy improvements can be observed. The improvement of model-5 upon model-4 is marginal or nonexistent for some \(\epsilon\), but the improvement upon model-1 is always significant.

Table 1: Statistics of the CNN without bias after training for different number of epochs.

<table>
<thead>
<tr>
<th>model name</th>
<th>number of epochs</th>
<th>train loss</th>
<th>normalized margin</th>
<th>train acc</th>
<th>test acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>model-1</td>
<td>38</td>
<td>(10^{-10.04})</td>
<td>(5.65 \times 10^{-5})</td>
<td>100%</td>
<td>99.3%</td>
</tr>
<tr>
<td>model-2</td>
<td>75</td>
<td>(10^{-15.12})</td>
<td>(9.50 \times 10^{-5})</td>
<td>100%</td>
<td>99.3%</td>
</tr>
<tr>
<td>model-3</td>
<td>107</td>
<td>(10^{-20.07})</td>
<td>(1.30 \times 10^{-4})</td>
<td>100%</td>
<td>99.3%</td>
</tr>
<tr>
<td>model-4</td>
<td>935</td>
<td>(10^{-120.01})</td>
<td>(4.61 \times 10^{-4})</td>
<td>100%</td>
<td>99.2%</td>
</tr>
<tr>
<td>model-5</td>
<td>10000</td>
<td>(10^{-881.51})</td>
<td>(1.18 \times 10^{-3})</td>
<td>100%</td>
<td>99.1%</td>
</tr>
</tbody>
</table>

We use the standard method for evaluating \(L^2\)-robustness in \cite{Carl17} and the source code from the authors with default hyperparameters. We plot the robust accuracy (the percentage of data with robustness > \(\epsilon\)) for the training set in the figures on the first row of Figure 8. It can be seen from the figures that for small \(\epsilon\) (e.g., \(\epsilon < 0.3\)), the relative order of robust accuracy is just the order of model-1 to model-5. For relatively large \(\epsilon\) (e.g., \(\epsilon > 0.3\)), the improvement of model-5 upon model-2 to model-4 becomes marginal or nonexistent in certain intervals of \(\epsilon\), but model-1 to model-4 still have an increasing order of robust accuracy and the improvement of model-5 upon model-1 is always significant. This shows that training longer can help to improve the \(L^2\)-robust accuracy on the training set.

We also evaluate the robustness on the test set, in which a misclassified test sample is considered to have robustness 0, and plot the robust accuracy in the figures on the second row of Figure 8. It can be seen from the figures that for small \(\epsilon\) (e.g., \(\epsilon < 0.2\)), the curves of the robust accuracy of model-1 to model-5 are almost indistinguishable. However, for relatively large \(\epsilon\) (e.g., \(\epsilon > 0.2\)), again, model-1 to model-4 have an increasing order of robust accuracy and the improvement of

\[\text{https://github.com/carlini/nn_robust_attacks}\]
model-5 upon model-1 is always significant. This shows that training longer can also help to improve the $L^2$-robust accuracy on the test set.

We tried various different settings of hyperparameters for the evaluation method (including different learning rates, different binary search steps, etc.) and we observed that the shapes and relative positions of the curves in Figure 8 are stable across different hyperparameter settings.

It is worth to note that the normalized margin and robustness do not grow in the same speed in our experiments, although the theory suggests $R_\theta(z) \geq q_\theta(z)/\beta$. This may be because the Lipschitz constant $\beta$ (if defined locally) is also changing during training. Combining training longer with existing techniques for constraining Lipschitz number (Anil et al., 2019; Cisse et al., 2017) could potentially alleviate this issue, and we leave it as a future work.

L. ADDITIONAL EXPERIMENTAL DETAILS

In this section, we provide additional details of our experiments.

L.1 LOSS-BASED LEARNING RATE SCHEDULING

The intuition of the loss-based learning rate scheduling is as follows. If the training loss is $\alpha$-smooth, then optimization theory suggests that we should set the learning rate to roughly $1/\alpha$. For a homogeneous model with cross-entropy loss, if the training accuracy is 100% at $\theta$, then a simple calculation can show that the smoothness (the $L^2$-norm of the Hessian matrix) at $\theta$ is $O(\bar{L} \cdot \text{poly}(\rho))$, where $\bar{L}$ is the average training loss and $\text{poly}(\rho)$ is some polynomial. Motivated by this fact, we parameterize the learning rate $\eta(t)$ at epoch $t$ as

$$\eta(t) := \frac{\alpha(t)}{\bar{L}(t-1)}, \quad (27)$$

where $\bar{L}(t-1)$ is the average training loss at epoch $t-1$, and $\alpha(t)$ is a relative learning rate to be tuned (Similar parameterization has been considered in (Nacson et al., 2019b) for linear model). The loss-based learning rate scheduling is indeed a variant of line search. In particular, we initialize $\alpha(0)$ by some value, and do the following at each epoch $t$:

Step 1. Initially $\alpha(t) \leftarrow \alpha(t-1)$; Let $\bar{L}(t-1)$ be the training loss at the end of the last epoch;
Step 2. Run SGD through the whole training set with learning rate $\eta(t) := \alpha(t)/\bar{L}(t-1)$;
Step 3. Evaluate the training loss $\bar{L}(t)$ on the whole training set;
Step 4. If $\bar{L}(t) < \bar{L}(t-1)$, $\alpha(t) \leftarrow \alpha(t) \cdot r_u$ and end this epoch; otherwise, $\alpha(t) \leftarrow \alpha(t)/r_d$ and go to Step 2.

In all our experiments, we set $\alpha(0) := 0.1, r_u := 2^{1/5} \approx 1.149, r_d := 2^{1/10} \approx 1.072$. This specific choice of those hyperparameters is not important; other choices can only affect the computational efficiency, but not the overall tendency of normalized margin.

L.2 ADDRESSING NUMERICAL ISSUES

Since we are dealing with extremely small loss (as small as $10^{-800}$), the current Tensorflow implementation would run into numerical issues. To address the issues, we work as follows. Let $\bar{L}_B(\theta)$ be the (average) training loss within a batch $B \subseteq [N]$. We use the notations $C, s_{nj}, q_n, q_n$ from Appendix G. We only need to show how to perform forward and backward passes for $\bar{L}_B(\theta)$.

Forward Pass. Suppose we have a good estimate $\bar{F}$ for $\log \bar{L}_B(\theta)$ in the sense that

$$\mathcal{R}_B(\theta) := \bar{L}_B(\theta) e^{-\bar{F}} = \frac{1}{B} \sum_{n \in B} \log \left( 1 + \sum_{j \neq y_n} e^{-s_{nj}(\theta)} \right) e^{-\bar{F}} \quad (28)$$

is in the range of $\text{float64}$. $\mathcal{R}_B(\theta)$ can be thought of a relative training loss with respect to $\bar{F}$. Instead of evaluating the training loss $\bar{L}_B(\theta)$ directly, we turn to evaluate this relative training loss in a numerically stable way:
Step 1. Perform forward pass to compute the values of $s_{nj}$ with float32, and convert them into float64:

Step 2. Let $Q := 30$. If $q_n(\theta) > Q$ for all $n \in B$, then we compute

$$R_B(\theta) = \frac{1}{B} \sum_{n \in B} e^{-(\tilde{q}_n(\theta) + \tilde{\varphi})},$$

where $\tilde{q}_n(\theta) := -\text{LSE}\{ \{ -s_{nj} : j \neq y_n \} \} = -\log \left( \sum_{j \neq y_n} e^{-s_{nj}} \right)$ is evaluated in a numerically stable way; otherwise, we compute

$$R_B(\theta) = \frac{1}{B} \sum_{n \in B} \log 1 + \left( \sum_{j \neq y_n} e^{-s_{nj}(\theta)} \right) e^{-\tilde{\varphi}},$$

where $\log 1 + p(x)$ is a numerical stable implementation of $\log(1 + x)$.

This algorithm can be explained as follows. Step 1 is numerically stable because we observe from the experiments that the layer weights and layer outputs grow slowly. Now we consider Step 2. If $q_n(\theta) \leq Q$ for some $n \in [B]$, then $\tilde{L}_B(\theta) = \Omega(e^{-Q})$ is in the range of float64, so we can compute $R_B(\theta)$ by \ref{eq:loss} directly except that we need to use a numerical stable implementation of $\log(1 + x)$. For $q_n(\theta) > Q$, arithmetic underflow can occur. By Taylor expansion of $\log(1 + x)$, we know that when $x$ is small enough $\log(1 + x) \approx x$ in the sense that the relative error $\frac{|\log(1 + x) - x|}{\log(1 + x)} = O(x)$. Thus, we can do the following approximation

$$\log \left( 1 + \sum_{j \neq y_n} e^{-s_{nj}(\theta)} \right) e^{-\tilde{\varphi}} \approx \sum_{j \neq y_n} e^{-s_{nj}(\theta)} e^{-\tilde{\varphi}}$$

for $q_n(\theta) > Q$, and only introduce a relative error of $O(C e^{-Q})$ (recall that $C$ is the number of classes). Using a numerical stable implementation of LSE, we can compute $\tilde{q}_n$ easily. Then the RHS of \ref{eq:loss} can be rewritten as $e^{-(\tilde{q}_n(\theta) + \tilde{\varphi})}$. Note that computing $e^{-(\tilde{q}_n(\theta) + \tilde{\varphi})}$ does not have underflow or overflow problems if $\tilde{\varphi}$ is a good approximation for $\log \tilde{L}_B(\theta)$.

**Backward Pass.** To perform backward pass, we build a computation graph in Tensorflow for the above forward pass for the relative training loss and use the automatic differentiation. We parameterize the learning rate as $\eta = \tilde{\eta} \cdot e^{\tilde{\varphi}}$. Then it is easy to see that taking a step of gradient descent for $\tilde{L}_B(\theta)$ with learning rate $\eta$ is equivalent to taking a step for $R_B(\theta)$ with $\tilde{\eta}$. Thus, as long as $\tilde{\eta}$ can fit into float64, we can perform gradient descent on $R_B(\theta)$ to ensure numerical stability.

**The Choice of $\tilde{\varphi}$.** The only question remains is how to choose $\tilde{\varphi}$. In our experiments, we set $\tilde{\varphi}(t) := \log \tilde{L}(t - 1)$ to be the training loss at the end of the last epoch, since the training loss cannot change a lot within one single epoch. For this, we need to maintain $\log \tilde{L}(t)$ during training. This can be done as follows: after evaluating the relative training loss $\tilde{L}(t)$ on the whole training set, we can obtain $\log \tilde{L}(t)$ by adding $\tilde{\varphi}(t)$ and $\log \tilde{L}(t)$ together.

It is worth noting that with this choice of $\tilde{\varphi}$, $\tilde{\eta}(t) = \alpha(t)$ in the loss-based learning rate scheduling. As shown in the right figure of Figure \ref{fig:loss}, $\alpha(t)$ is always between $10^{-9}$ and $10^0$, which ensures the numerical stability of backward pass.