

# Fast Convergence in Learning Two-Layer Neural Networks with Separable Data

Hossein Taheri<sup>1</sup>, Christos Thrampoulidis<sup>2</sup>

<sup>1</sup>University of California, Santa Barbara

<sup>2</sup>University of British Columbia

hossein@ucsb.edu, cthrampo@ece.ubc.ca

## Abstract

Normalized gradient descent has shown substantial success in speeding up the convergence of exponentially-tailed loss functions (which includes exponential and logistic losses) on linear classifiers with separable data. In this paper, we go beyond linear models by studying normalized GD on two-layer neural nets. We prove for exponentially-tailed losses that using normalized GD leads to linear rate of convergence of the training loss to the global optimum. This is made possible by showing certain gradient self-boundedness conditions and a log-Lipschitzness property. We also study generalization of normalized GD for convex objectives via an algorithmic-stability analysis. In particular, we show that normalized GD does not overfit during training by establishing finite-time generalization bounds.

## Introduction

### Motivation

A wide variety of machine learning algorithms for classification tasks rely on learning a model using monotonically decreasing loss functions such as logistic loss or exponential loss. In modern practice these tasks are often accomplished using over-parameterized models such as large neural networks where the model can interpolate the training data, i.e., it can achieve perfect classification accuracy on the samples. In particular, it is often the case that the training of the model is continued until achieving approximately zero training loss (Zhang et al. 2021).

Over the last decade there has been remarkable progress in understanding or improving the convergence and generalization properties of over-parameterized models trained by various choices of loss functions including logistic loss and quadratic loss. For the quadratic loss it has been shown that over-parameterization can result in significant improvements in the training convergence rate of (stochastic)gradient descent on empirical risk minimization algorithms. Notably, quadratic loss on two-layer ReLU neural networks is shown to satisfy the Polyak-Łojasiewicz(PL) condition (Charles and Papailiopoulos 2018; Bassily, Belkin, and Ma 2018; Liu, Zhu, and Belkin 2022). In fact, the PL property is a consequence of the observation that the tangent kernel associated with

the model is a non-singular matrix. Moreover, in this case the PL parameter, which specifies the rate of convergence, is the smallest eigenvalue of the tangent kernel(Liu, Zhu, and Belkin 2022). The fact that over-parameterized neural networks trained by quadratic loss satisfy the PL condition, guarantees that the loss converges exponentially fast to a global optimum. The global optimum in this case is a model which “perfectly” interpolates the data, where we recall that perfect interpolation requires that the model output for every training input is precisely equal to the corresponding label.

On the other hand, gradient descent using un-regularized logistic regression with linear models and separable data is biased toward the max-margin solution. In particular, in this case the parameter converges in direction with the rate  $O(1/\log(t))$  to the solution of hard margin SVM problem, while the training loss converges to zero at the rate  $\tilde{O}(1/t)$  (Soudry et al. 2018; Ji and Telgarsky 2018). More recently, normalized gradient descent has been proposed as a promising approach for fast convergence of exponentially tailed losses (Nacson et al. 2019). In this method, at any iteration the step-size is chosen proportionally to the inverse of value of training loss function i.e.,  $\eta_t \propto 1/F(w_t)$ . This results in choosing unboundedly increasing step-sizes for the iterates of gradient descent as  $F(w_t)$  decays to zero. This choice of step-size leads to significantly faster rates for the parameter’s directional convergence. In particular, for linear models with separable data, it is shown that normalized GD with decaying step-size enjoys a rate of  $O(1/\sqrt{t})$  in directional parameter convergence to the max-margin separator (Nacson et al. 2019). This has been improved to  $O(1/t)$  with normalized GD using fixed step-size (Ji and Telgarsky 2021).

Despite remarkable progress in understanding the behavior of normalized GD with separable data, these results are only applicable to the implicit bias behavior of “linear models”. In this paper, we aim to discover for the first time, the dynamics of learning a two-layer neural network with normalized GD trained on separable data. For the training loss, we show that using normalized GD on an exponentially-tailed loss with a two layered neural network leads to exponentially fast convergence of the loss to the global optimum. This is significantly faster than the convergence rate of  $\tilde{O}(1/t)$  for GD with constant step-size under the same settings. Compared to the convergence analysis of standard GD which is usually carried out using smoothness of the loss function, here for

normalized GD we use the Taylor’s expansion of the loss and use the fact the operator norm of the Hessian is bounded by the loss. Next, we apply a lemma in our proof which shows that exponentially-tailed losses on a two-layered neural network satisfy a log-Lipschitzness condition throughout the iterates of normalized GD. Moreover, crucial to our analysis is showing that the  $\ell_2$  norm of the gradient at every point is upper-bounded and lower-bounded by constant factors of the loss under given assumptions on the activation function and the training data. Subsequently, the log-Lipschitzness property together with the bounds on the norm of Gradient and Hessian of the loss function ensures that normalized GD is a descent algorithm where the loss value decreases by a constant factor after each step of normalized GD, which leads to the promised geometric rate of decay for the loss. Our another goal is to characterize the iterate-wise test error performance of normalized GD for convex objectives. Notably, we show that despite the fast grow of weights norm, the generalization error remains bounded. Particularly, through an algorithmic-stability analysis, we derive optimal test-error rates for normalized GD with logistic loss. We summarize our contributions below.

## Contributions

We introduce conditions –namely log-Lipschitz and self-boundedness assumptions on the gradient and the Hessian– under which the training loss of the normalized GD algorithm converges exponentially fast to the global optimum. More importantly, in Section we prove that the aforementioned conditions are indeed satisfied by two-layer neural networks trained with an exponentially-tailed loss function. This yields the first theoretical guarantee on the convergence of normalized GD for non-linear models. We also study a stochastic variant of normalized GD and investigate its training loss convergence.

We study, for the first time, the finite-time test loss and test error performance of normalized GD for convex losses. In particular, we derive bounds of order  $O(1/n)$  on the expected generalization error of normalized GD, where  $n$  is the training-set size.

## Prior Works

The theoretical study of the optimization landscape of over-parameterized models trained by GD or SGD has been the subject of several recent works. The majority of these works study over-parameterized models with specific choices of loss functions, mainly quadratic or logistic loss functions. For quadratic loss, the exponential convergence rate of over-parameterized neural networks is proved in several recent works e.g., (Charles and Papailiopoulos 2018; Bassily, Belkin, and Ma 2018; Du et al. 2019; Allen-Zhu, Li, and Song 2019; Arora et al. 2019; Oymak and Soltanolkotabi 2019, 2020; Safran, Yehudai, and Shamir 2021; Liu, Zhu, and Belkin 2022). These results naturally relate to the Neural Tangent Kernel (NTK) regime of infinitely wide or sufficiently large initialized neural networks (Jacot, Gabriel, and Hongler 2018) in which the iterates of gradient descent stay close to the initialization. The NTK approach can not be applied to our

setting as the parameters’ norm in our setting is growing as  $\Theta(t)$  with the NGD updates.

The majority of the prior results apply to the quadratic loss. However, the state-of-the-art architectures for classification tasks use unregularized ERM with logistic/exponential loss functions. Notably, for these losses over-parameterization leads to infinite norm optimizers. As a result, the objective in this case does not satisfy strong convexity or the PL condition even for linear models. The analysis of loss and parameter convergence of logistic regression on separable data has attracted significant attention in the last five years. Notably, a line of influential works have shown that gradient descent provably converges in direction to the max-margin solution for linear models and two-layer homogenous neural networks. In particular, the study of training loss and implicit bias behavior of GD on logistic/exponential loss was first initiated in the settings of linear classifiers (Rosset, Zhu, and Hastie 2003; Telgarsky 2013; Soudry et al. 2018; Ji and Telgarsky 2018; Nacson et al. 2019). The implicit bias behavior of GD with logistic loss in two-layer neural networks was later studied by (Lyu and Li 2019; Chizat and Bach 2020; Ji and Telgarsky 2020). The loss landscape of logistic loss for over-parameterized neural networks and structured data is analyzed in (Zou et al. 2020; Chatterji, Long, and Bartlett 2021), where it is proved that GD converges to a global optima at the rate  $O(1/t)$ . The majority of these results hold for standard GD while we focus on normalized GD.

The generalization properties of GD/SGD with binary and multi-class logistic regression in the interpolation regime is studied in (Shamir 2021; Schliserman and Koren 2022) for linear models and in (Li and Liang 2018; Cao and Gu 2019, 2020) for neural networks. Recently, (Taheri and Thrampoulidis 2023) studied the generalization error of decentralized logistic regression through a stability analysis. For our generalization analysis we use an algorithmic stability analysis (Bousquet and Elisseeff 2002; Hardt, Recht, and Singer 2016; Lei and Ying 2020). However, unlike these prior works which focus on GD with constant step-size, we consider normalized GD and derive the first generalization analysis for this algorithm.

The benefits of normalized GD for speeding up the directional convergence of GD for linear models was suggested by (Nacson et al. 2019; Ji and Telgarsky 2021). Our paper contributes to this line of work. Compared to these prior works which are focused on implicit-bias behavior of linear models, we study non-linear models and derive training loss convergence rates. We also study, the generalization performance of normalized GD for convex objectives.

## Notation

We use  $\|\cdot\|$  to denote the operator norm of a matrix and also to denote the  $\ell_2$ -norm of a vector. The Frobenius norm of a matrix  $W$  is shown by  $\|W\|_F$ . The Gradient and the Hessian of a function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  are denoted by  $\nabla F$  and  $\nabla^2 F$ . Similarly, for a function  $F : \mathbb{R}^d \times \mathbb{R}^{d'} \rightarrow \mathbb{R}$  that takes two input variables, the Gradient and the Hessian with respect to the  $i$ th variable (where  $i = 1, 2$ ) are denoted by  $\nabla_i F$  and  $\nabla_i^2 F$ , respectively. For functions  $F, G : \mathbb{R} \rightarrow \mathbb{R}$ , we write  $F(t) = O(G(t))$  when  $|F(t)| \leq mG(t)$  after  $t \geq t_0$  for

positive constants  $m, t_0$ . We write  $F(t) = \tilde{O}(G(t))$  when  $F(t) = O(G(t)H(t))$  for a polylogarithmic function  $H$ . Finally, we denote  $F(t) = \Theta(G(t))$  if  $|F(t)| \leq m_1 G(t)$  and  $|F(t)| \geq m_2 G(t)$  for all  $t \geq t_0$  for some positive constants  $m_1, m_2, t_0$ .

### Problem Setup

We consider unconstrained and unregularized empirical risk minimization (ERM) on  $n$  samples,

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

The  $i$ th sample  $z_i := (x_i, y_i)$  consists of a data point  $x_i \in \mathbb{R}^d$  and its associated label  $y_i \in \{\pm 1\}$ . The function  $\Phi : \mathbb{R}^{\tilde{d}} \times \mathbb{R}^d \rightarrow \mathbb{R}$  represents the model taking the weights vector  $w$  and data point  $x$  to approximate the label. In this section, we take  $\Phi$  as a neural network with one hidden layer and  $m$  neurons,

$$\Phi(w, x) := \sum_{j=1}^m a_j \sigma(\langle w_j, x \rangle).$$

Here  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is the activation function and  $w_j \in \mathbb{R}^d$  denotes the input weight vector of the  $j$ th hidden neuron.  $w \in \mathbb{R}^{\tilde{d}}$  represents the concatenation of these weights i.e.,  $w = [w_1; w_2; \dots; w_m]$ . In our setting the total number of parameters and hence the dimension of  $w$  is  $\tilde{d} = md$ . We assume that only the first layer weights  $w_j$  are updated during training and the second layer weights  $a_j \in \mathbb{R}$  are initialized randomly and are maintained fixed during training. The function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is non-negative and monotonically decreases such that  $\lim_{t \rightarrow +\infty} f(t) = 0$ . In this section, we focus on the exponential loss  $f(t) = \exp(-t)$ , but we expect that our results apply to a broader class of loss functions that behave similarly to the exponential loss for large  $t$ , such as logistic loss  $f(t) = \log(1 + \exp(-t))$ .

We consider activation functions with bounded absolute value for the first and second derivatives.

**Assumption 1** (Activation function). *The activation function  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is smooth and for all  $t \in \mathbb{R}$*

$$|\sigma''(t)| \leq L.$$

Moreover, there are positive constants  $\alpha, \ell$  such that  $\sigma$  satisfies for all  $t \in \mathbb{R}$ ,

$$\alpha \leq \sigma'(t) \leq \ell.$$

An example satisfying the above condition is the activation function known as smoothed-leaky-ReLU which is a smoothed variant of the leaky-ReLU activation  $\sigma(t) = \ell t \mathbb{I}(t \geq 0) + \alpha t \mathbb{I}(t \leq 0)$ , where  $\mathbb{I}(\cdot)$  denotes the 0–1 indicator function.

Throughout the paper we let  $R$  and  $a$  denote the maximum norm of data points and second layer weights, respectively, i.e.,

$$R := \max_{i \in [n]} \|x_i\|, \quad a := \max_{j \in [m]} |a_j|.$$

Throughout the paper we assume  $R = \Theta(1)$  w.r.t. problem parameters and  $a = \frac{1}{m}$ .

We also denote the *training loss* of the model by  $F$ , defined in (1) and define the *train error* as misclassification error over the training data, or formally by  $F_{0-1}(w) := \frac{1}{n} \sum_{i=1}^n \mathbb{I}(\text{SIGN}(\Phi(w, x_i)) \neq y_i)$ .

**Normalized GD.** We consider the iterates of normalized GD as follows,

$$w_{t+1} = w_t - \eta_t \nabla F(w_t). \quad (2)$$

The step size is chosen inversely proportional to the loss value i.e.,  $\eta_t = \eta / F(w_t)$ , implying that the step-size is growing unboundedly as the algorithm approaches the optimum solution. Since the gradient norm decays proportionally to the loss, one can equivalently choose  $\eta_t = \eta / \|\nabla F(w_t)\|$ .

### Main Results

For convergence analysis in our case study, we introduce a few definitions.

**Definition 1** (log-Lipschitz Objective). *The training loss  $F : \mathbb{R}^{\tilde{d}} \rightarrow \mathbb{R}$  satisfies the log-Lipschitzness property if for all  $w, w' \in \mathbb{R}^{\tilde{d}}$ ,*

$$\max_{v \in [w, w']} F(v) \leq F(w) \cdot \tilde{c}_{w, w'},$$

where  $[w, w']$  denotes the line between  $w$  and  $w'$  and we define  $\tilde{c}_{w, w'} := \exp(c(\|w - w'\| + \|w - w'\|^2))$  where the positive constant  $c$  is independent of  $w, w'$ .

As we will see in the following sections, log-Lipschitzness is a property of neural networks trained with exponentially tailed losses with  $c = \Theta(\frac{1}{\sqrt{m}})$ . We also define the property “log-Lipschitzness in the gradient path” if for all  $w_t, w_{t-1}$  in Eq. (2) there exists a constant  $C$  such that,

$$\max_{v \in [w_t, w_{t+1}]} F(v) \leq C F(w_t).$$

**Definition 2** (Self lower-bounded gradient). *The loss function  $F : \mathbb{R}^{\tilde{d}} \rightarrow \mathbb{R}$  satisfies the self-lower bounded Gradient condition for a function, if there exists a constant  $\mu$  such that for all  $w$ ,*

$$\|\nabla F(w)\| \geq \mu F(w).$$

**Definition 3** (Self-boundedness of the gradient). *The loss function  $F : \mathbb{R}^{\tilde{d}} \rightarrow \mathbb{R}$  satisfies the self-boundedness of the gradient condition for a constant  $h$ , if for all  $w$*

$$\|\nabla F(w)\| \leq h F(w).$$

The above two conditions on the upper-bound and lower bound of the gradient norm based on loss can be thought as the equivalent properties of smoothness and the PL condition but for our studied case of exponential loss. To see this, note that smoothness and PL condition provide upper and lower bounds for the square norm of gradient. In particular, by  $L$ -smoothness one can deduce that  $\|\nabla F(w)\|^2 \leq 2L(F(w) - F^*)$  (e.g., (Nesterov 2003)) and by the definition of  $\mu$ -PL condition  $\|\nabla F(w)\|^2 \geq 2\mu(F(w) - F^*)$  (Polyak 1963; Lojasiewicz 1963).

The next necessary condition is an upper-bound on the operator norm of the Hessian of loss.

**Definition 4** (Self-boundedness of the Hessian). *The loss function  $F : \mathbb{R}^{\tilde{d}} \rightarrow \mathbb{R}$  satisfies the self-boundedness of the Hessian property for a constant  $H$ , if for all  $w$ ,*

$$\|\nabla^2 F(w)\| \leq H F(w),$$

where  $\|\cdot\|$  denotes the operator norm.

It is worthwhile to mention here that in the next sections of the paper, we prove all the self lower and upper bound in Definitions 3-4 are satisfied for a two-layer neural network under some regularity conditions.

### Convergence Analysis of Training Loss

The following theorem states that under the conditions above, the training loss converges to zero at an exponentially fast rate.

**Theorem 1** (Convergence of Training Loss). *Consider normalized gradient descent update rule with loss  $F$  and step-size  $\eta_t$ . Assume  $F$  and the normalized GD algorithm satisfy log-Lipschitzness in the gradient path with parameter  $C$ , as well as self-boundedness of the Gradient and the Hessian and the self-lower bounded Gradient properties with parameters  $h, H$  and  $\mu$ , respectively. Let  $\eta_t = \frac{\eta}{F(w_t)}$  for all  $t \in [T]$  and for any positive constant  $\eta$  satisfying  $\eta \leq \frac{\mu^2}{HC^2}$ . Then for the training loss at iteration  $T$  the following bound holds:*

$$F(w_T) \leq \left(1 - \frac{\eta\mu^2}{2}\right)^T F(w_0). \quad (3)$$

*Remark 1.* The proof of Theorem 1 is provided in Appendix, where we use a Taylor expansion of the loss and apply the conditions of the theorem. It is worth noting that the rate obtained for normalized GD in Theorem 1 is significantly faster than the rate of  $\tilde{O}(\frac{1}{T})$  for standard GD with logistic or exponential loss in neural networks (e.g., (Zou et al. 2020, Theorem 4.4)). Additionally, for a continuous-time perspective on the training convergence of normalized GD, we refer to Proposition 10 in the appendix, which presents a convergence analysis based on *normalized Gradient Flow*. The advantage of this approach is that it does not require the self-bounded Hessian property and can be used to show exponential convergence of normalized Gradient Flow for ReLU and leaky-ReLU activations.

### Two-Layer Neural Networks

In this section, we prove that the conditions that led to Theorem 1 are in fact satisfied by a two-layer neural network. Consequently, this implies that the training loss bound in Eq.(3) is valid for this class of functions. We choose  $f(t) = \exp(-t)$  for simpler proofs, however an akin result holds for the broader class of exponentially tailed loss functions.

First, we start with verifying the log-Lipschitzness condition (Definition 1). In particular, here we prove a variation of this property for the iterates of normalized GD i.e., where  $w, w'$  are chosen as  $w_t, w_{t+1}$ . The proofs are included in Appendix.

**Lemma 2** (log-Lipschitzness in the gradient path). *Let  $F$  be as in (1) for the exponential loss  $f$  and let  $\Phi$  be a two-layer*

*neural network with the activation function satisfying Assumption 1. Consider the iterates of normalized GD with the step-size  $\eta_t = \frac{\eta}{F(w_t)}$ . Then for any  $\lambda \in [0, 1]$  the following inequality holds:*

$$F(w_t + \lambda(w_{t+1} - w_t)) \leq \exp(\lambda c) F(w_t), \quad (4)$$

for a positive constant  $c$  independent of  $\lambda, w_t$  and  $w_{t+1}$ . As a direct consequence, it follows that,

$$\max_{v \in [w_t, w_{t+1}]} F(v) \leq C F(w_t), \quad (5)$$

for a numerical constant  $C$ .

The next two lemmas state sufficient conditions for  $F$  to satisfy the self-lower boundedness for its gradient (Def. 2).

**Lemma 3** (Self lower-boundedness of gradient). *Let  $F$  be as in (1) for the exponential loss  $f$  and let  $\Phi$  be a two-layer neural network with the activation function satisfying Assumption 1. Assume the training data is linearly separable with margin  $\gamma$ . Then  $F$  satisfies the self-lower boundedness of gradient with the constant  $\mu = \frac{\alpha\gamma}{\sqrt{m}}$  for all  $w$ , i.e.,  $\|\nabla F(w)\| \geq \mu F(w)$ .*

Next, we aim to show that the condition  $\|\nabla F(w)\| \geq \mu F(w)$ , holds for training data separable by a two-layer neural network during gradient descent updates. In particular, we assume the Leaky-ReLU activation function taking the following form,

$$\sigma(t) = \begin{cases} \ell t & t \geq 0, \\ \alpha t & t < 0. \end{cases} \quad (6)$$

for arbitrary non-negative constants  $\alpha, \ell$ . This includes the widely-used ReLU activation as a special case. Next lemma shows that when the weights are such that the neural network separates the training data, the self-lower boundedness condition holds.

**Lemma 4.** *Let  $F$  be in (1) for the exponential loss  $f$  and let  $\Phi$  be a two-layer neural network with activation function in Eq.(6). Assume the first layer weights  $w \in \mathbb{R}^{\tilde{d}}$  are such that the neural network separates the training data with margin  $\gamma$ . Then  $F$  satisfies the self-lower boundedness of gradient, i.e.,  $\|\nabla F(w)\| \geq \mu F(w)$ , where  $\mu = \gamma$ .*

A few remarks are in place. The result of Lemma 4 is relevant for  $w$  that can separate the training data. Especially, this implies the self lower-boundedness property after GD iterates succeed in finding an interpolator. However, we should also point out that the non-smoothness of leaky-ReLU activation functions precludes the self-bounded Hessian property and it remains an interesting future direction to prove the self lower-boundedness property with general smooth activations. On the other hand, the convergence of normalized "Gradient-flow" does not require the self-bounded Hessian property, as demonstrated in Proposition 10. This suggests that Lemma 4 can be applied to prove the convergence of normalized Gradient-flow with ReLU or leaky-ReLU activations.

Next lemma derives the self-boundedness of the gradient and Hessian (c.f. Definitions 3-4) for our studied case. The proof of Lemma 5 follows rather straight-forwardly from the closed-form expressions of gradient and Hessian and using properties of the activation function.

**Lemma 5** (Self-boundedness of the gradient and Hessian). *Let  $F$  be in (1) for the exponential loss  $f$  and let  $\Phi$  be a two-layer neural network with the activation function satisfying Assumption 1. Then  $F$  satisfies the self-boundedness of gradient and Hessian with constants  $h = \frac{\ell R}{\sqrt{m}}$ ,  $H := \frac{LR^2}{m^2} + \frac{\ell^2 R^2}{m}$  i.e.,*

$$\|\nabla F(w)\| \leq hF(w), \quad \|\nabla^2 F(w)\| \leq HF(w).$$

We conclude this section by offering a few remarks regarding our training convergence results. We emphasize that combining Theorem 1 and Lemmas 2-5 achieves the convergence of training loss of normalized Gradient Descent for two-layer networks. Moreover, we refer to Proposition 10 in Appendix which presents a continuous time convergence analysis of normalized GD based on Gradient Flow. This result is especially relevant in the context of ReLU/leaky-ReLU activation, where Proposition 10 together with Lemma 4 shows exponential convergence of normalized Gradient Flow.

### Generalization Error

In this section, we study the generalization performance of normalized GD algorithm. Formally, the *test loss* for the data distribution  $\mathcal{D}$  is defined as follows,

$$\tilde{F}(w) := \mathbb{E}_{(x,y) \sim \mathcal{D}} [f(y\Phi(w,x))].$$

Depending on the choice of loss  $f$ , the test loss might not always represent correctly the classification performance of a model. For this, a more reliable standard is the *test error* which is based on the 0 – 1 loss,

$$\tilde{F}_{0-1}(w) := \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathbb{I}(y \neq \text{SIGN}(\Phi(w,x)))].$$

We also define the *generalization loss* as the gap between training loss and test loss. Likewise, we define the *generalization error* based on the train and test errors.

With these definitions in place, we are ready to state our results. In particular, in this section we prove that under the normalized GD update rule, the generalization loss at step  $T$  is bounded by  $O(\frac{T}{n})$  where recall that  $n$  is the training sample size. While, the dependence of generalization loss on  $T$  seems unappealing, we show that this is entirely due to the fact that a convex-relaxation of the 0 – 1 loss, i.e. the loss function  $f$ , is used for evaluating the generalization loss. In particular, we can deduce that under appropriate conditions on loss function and data (c.f. Corollary 7.1), the test error is related to the test loss through,

$$\tilde{F}_{0-1}(w_T) = O\left(\frac{\tilde{F}(w_T)}{\|w_T\|}\right).$$

As we will see in the proof of Corollary 7.1, for normalized GD with exponentially tailed losses the weights norm  $\|w_T\|$  grows linearly with  $T$ . Thus, this relation implies that the test error satisfies  $\tilde{F}_{0-1}(w_T) = O(\frac{1}{n})$ . Essentially, this bound on the misclassification error signifies the fast convergence of normalized GD on test error and moreover, it shows that normalized GD never overfits during its iterations.

It is worthwhile to mention that our generalization analysis is valid for any model  $\Phi$  such that  $f(y\Phi(\cdot, x))$  is convex for any  $(x, y) \sim \mathcal{D}$ . This includes linear models i.e.,  $\Phi(w, x) = \langle w, x \rangle$  or the Random Features model (Rahimi and Recht 2007), i.e.,  $\Phi(w, x) = \langle w, \sigma(Ax) \rangle$  where  $\sigma(\cdot)$  is applied element-wise on its entries and the matrix  $A \in \mathbb{R}^{m \times d}$  is initialized randomly and kept fixed during train and test time. Our results also apply to kernel methods and in particular neural networks in the NTK regime due to the convex-like behavior of optimization landscape in the infinite-width limit.

We study the generalization performance of normalized GD, through a stability analysis (Bousquet and Elisseeff 2002). The existing analyses in the literature for algorithmic stability of  $\tilde{L}$ -smooth losses, rely on the step-size satisfying  $\eta_t = O(1/\tilde{L})$ . This implies that such analyses can not be employed for studying increasingly large step-sizes as in our case  $\eta_t$  is unboundedly growing. In particular, the common approach in the stability analysis (Hardt, Recht, and Singer 2016; Lei and Ying 2020) uses the “non-expansiveness” property of standard GD with smooth and convex losses, by showing that for  $\eta \leq 2/\tilde{L}$  and for any two points  $w, v \in \mathbb{R}^d$ , it holds that  $\|w - \eta \nabla F(w) - (v - \eta \nabla F(v))\| \leq \|w - v\|$ . Central to our stability analysis is showing that under the assumptions of self-boundedness of Gradient and Hessian, the normalized GD update rule satisfies the non-expansiveness condition with any step-size satisfying both  $\eta \lesssim \frac{1}{F(w)}$  and  $\eta \lesssim \frac{1}{F(v)}$ .

**Lemma 6** (Non-expansiveness of normalized GD). *Assume the loss  $F$  to satisfy convexity and self-boundedness for the gradient and the Hessian with parameter  $h \leq 1$  (Definitions 3-4). Let  $v, w \in \mathbb{R}^d$ . If  $\eta \leq \frac{1}{h \cdot \max(F(v), F(w))}$ , then*

$$\|w - \eta \nabla F(w) - (v - \eta \nabla F(v))\| \leq \|w - v\|.$$

The next theorem characterizes the test loss for both Lipschitz and smooth objectives. Before stating the theorem, we need to define  $\delta$ . For the leave-one-out parameter  $w_t^{-i}$  and loss  $F^{-i}(\cdot)$  defined as

$$w_{t+1}^{-i} = w_t^{-i} - \eta_t \nabla F^{-i}(w_t^{-i}),$$

and

$$F^{-i}(w) := \frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^n f(w, z_j),$$

we define  $\delta \geq 1$  to be any constant which satisfies for all  $t \in [T], i \in [n]$ , the following

$$F^{-i}(w_t^{-i}) \leq \delta F^{-i}(w_t).$$

While this condition seems rather restrictive, we prove in Lemma 9 in Appendix that the condition on  $\delta$  is satisfied by two-layer neural networks with sufficient over-parameterization. With these definitions in place, we are ready to state the main theorem of this section.

**Theorem 7** (Test loss). *Consider normalized GD update rule with  $\eta_t = \frac{\eta}{F(w_t)}$  where  $\eta \leq \frac{1}{h\delta}$ . Assume the loss  $F$  to be convex and to satisfy the self-bounded gradient and*

Hessian property with a parameter  $h$  (Definitions 3-4). Then the following statements hold for the test loss:

(i) if the loss  $F$  is  $G$ -Lipschitz, then the generalization loss at step  $T$  satisfies

$$\mathbb{E}[\tilde{F}(w_T) - F(w_T)] \leq \frac{2GT}{n}.$$

(ii) if the loss  $F$  is  $\tilde{L}$ -smooth, then the test loss at step  $T$  satisfies,

$$\mathbb{E}[\tilde{F}(w_T)] \leq 4\mathbb{E}[F(w_T)] + \frac{3\tilde{L}^2 T}{n},$$

where all expectations are over training sets.

As discussed earlier in this section, the test loss dependence on  $T$  is due to the rapid growth of the  $\ell_2$  norm of  $w_t$ . As a corollary, we show that the generalization error is bounded by  $O(\frac{1}{n})$ . For this, we assume the next condition.

**Assumption 2** (Distribution Absolute-Margin). *There exists a constant  $\tilde{\gamma}$  such that after sufficient iterations the model satisfies  $|\Phi(w_t, x)| \geq \tilde{\gamma}\|w_t\|$  almost surely over the data distribution  $(x, y) \sim \mathcal{D}$ .*

Assumption 2 implies that the absolute value of the margin is  $\tilde{\gamma}$  i.e.,  $\frac{|\Phi(w_t, x)|}{\|w_t\|} \geq \tilde{\gamma}$  for almost every  $x$  after sufficient iterations. This assumption is rather mild, as intuitively it requires that data distribution is not concentrating around the decision boundaries.

For the loss function, we consider the special case of logistic loss  $f(t) = \log(1 + \exp(-t))$  for simplicity of exposition and more importantly due to its Lipschitz property. The use of Lipschitz property is essential in view of Theorem 7.

**Corollary 7.1** (Test error). *Suppose the assumptions of Theorem 7 hold. Consider the neural network setup under Assumptions 1 and 2 and let the loss function  $f$  be the logistic loss. Then the test error at step  $T$  of normalized GD satisfies the following:*

$$\mathbb{E}[\tilde{F}_{0-1}(w_T)] = O\left(\frac{1}{T}\mathbb{E}[F(w_T)] + \frac{1}{n}\right)$$

In the proof, we use that  $\|w_t\|$  grows linearly with  $t$  as well as Assumption 2 to deduce  $\tilde{F}_{0-1}(w_T) = O(\frac{\tilde{F}(w_T)}{T})$ . Hence, the statement of the corollary follows from Theorem 7 (i). Finally, we remark the expected test error is decreasing with the rate  $1/n$ , which is known to be optimal in the realizable setting we consider in this paper.

## Stochastic Normalized GD

In this section we consider a stochastic variant of normalized GD algorithm and show that the benefits of normalized gradient methods extend to stochastic settings. Assume  $z_t$  to be the batch selected randomly from the dataset at iteration  $t$ . The stochastic normalized GD takes the form,

$$w_{t+1} = w_t - \eta_t \nabla F_{z_t}(w_t), \quad (7)$$

where  $\nabla F_{z_t}(w_t)$  is the gradient of loss at  $w_t$  by using the batch of training points  $z_t$  at iteration  $t$ . We assume  $\eta_t$  to be proportional to  $1/F(w_t)$ . Our result in this section states that under the following strong growth condition (Schmidt and Roux 2013; Vaswani, Bach, and Schmidt 2019), the training loss converges at an exponential rate to the global optimum.

**Assumption 3** (Strong Growth Condition). *The training loss  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  satisfies the strong growth condition with a parameter  $\rho$ ,*

$$\mathbb{E}_z[\|\nabla F_z(w)\|^2] \leq \rho \|\nabla F(w)\|^2.$$

Notably, we show in Appendix that the strong growth condition holds for our studied case under the self-bounded and self-lower bounded gradient property.

The next theorem characterizes the rate of decay for the training loss. The proof and numerical experiments are deferred to Appendix.

**Theorem 8** (Convergence of Training Loss). *Consider stochastic normalized GD update rule in Eq.(7). Assume  $F$  satisfies Assumption 3 as well as the log-Lipschitzness in the GD path, self-boundedness of the Gradient and the Hessian and the self-lower bounded gradient properties (Definitions 1-4). Let  $\eta_t = \eta/F(w_t)$  for all  $t \in [T]$  and for any positive constant  $\eta$  satisfying  $\eta \leq \frac{\mu^2}{HC\rho h^2}$ . Then for the training loss at iteration  $T$  the following bound holds:*

$$F(w_T) \leq \left(1 - \frac{\eta\mu^2}{2}\right)^T F(w_0).$$

## Numerical Experiments

In this section, we demonstrate the empirical performance of normalized GD. Figure 1 illustrates the training loss (Left), the test error % (middle), and the weight norm (Right) of GD with normalized GD. The experiments are conducted on a two-layer neural network with  $m = 50$  hidden neurons with leaky-ReLU activation function in (6) where  $\alpha = 0.2$  and  $\ell = 1$ . The second layer weights are chosen randomly from  $a_j \in \{\pm \frac{1}{m}\}$  and kept fixed during training and test time. The first layer weights are initialized from standard Gaussian distribution and then normalized to unit norm. We consider binary classification with exponential loss using digits “0” and “1” from the MNIST dataset ( $d = 784$ ) and we set the sample size to  $n = 1000$ . The step-size are fine-tuned to  $\eta = 30$  and 5 for GD and normalized GD, respectively so that each line represents the best of each algorithm. We highlight the significant speed-up in the convergence of normalized GD compared to standard GD. For the training loss, normalized GD decays exponentially fast to zero while GD converges at a remarkably slower rate. We also highlight that  $\|w_t\|$  for normalized GD grows at a rate  $\Theta(t)$  while it remains almost constant for GD. In fact this was predicted by Corollary 7.1 where in the proof we showed that the weight norm grows linearly with the iteration number. In Figure 2, we generate two synthetic dataset according to a realization of a zero-mean Gaussian-mixture model with  $n = 40$  and  $d = 2$  where the two classes have different covariance matrices (top) and a zero-mean Gaussian-mixture model with  $n = 40, d = 5$  (only the first two entires are depicted in the figure) where

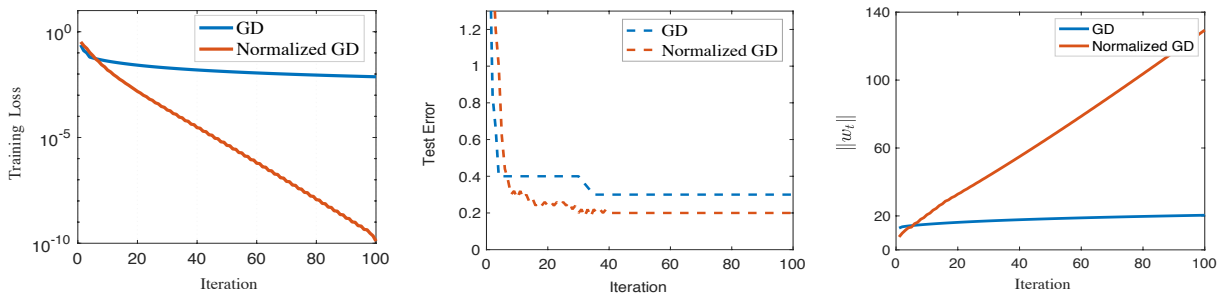


Figure 1: Comparison of the training loss, test error (in percentage), and weight norm (i.e.,  $\|w_t\|$ ) between gradient descent and normalized gradient descent algorithms. The experiments were conducted on two classes of the MNIST dataset using exponential loss and a two-layer neural network with  $m = 50$  hidden neurons. The results demonstrate the performance advantages of normalized gradient descent over traditional gradient descent in terms of both the training loss and test error.

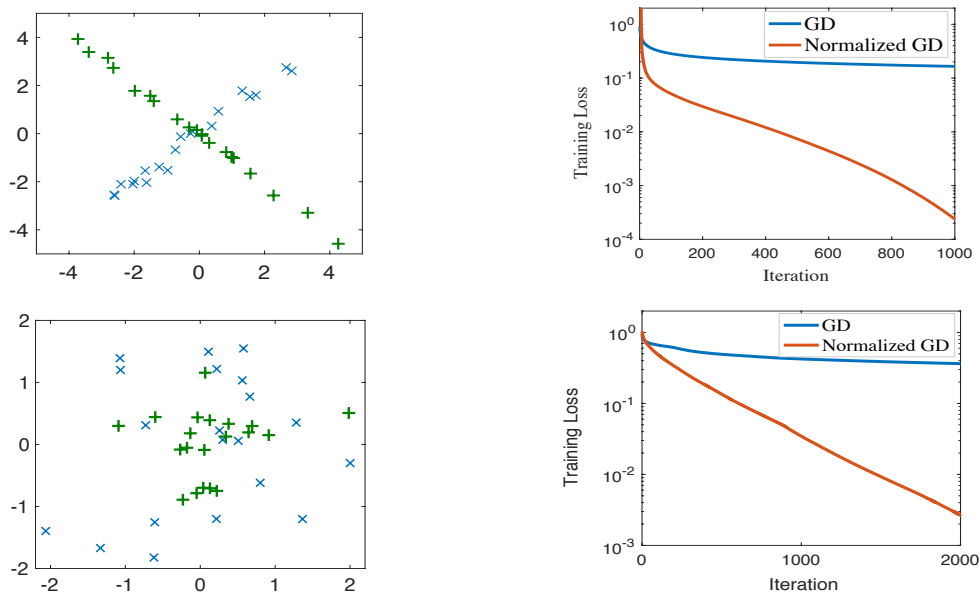


Figure 2: The left plot depicts two synthetic datasets, each consisting of  $n = 40$  data points. On the right, we present the training loss results of gradient descent and normalized gradient descent algorithms applied to a two-layer neural network with  $m = 50$  (top) and 100 (bottom) hidden neurons.

$\Sigma_1 = \mathbf{I}, \Sigma_2 = \frac{1}{4}\mathbf{I}$  (Bottom). Note that none of the datasets is linearly separable. We consider the same settings as in Figure 1 and compared the performance of GD and normalized GD in the right plots. The step-sizes are fine-tuned to  $\eta = 80, 350$  and  $30, 20$  for GD and normalized GD, respectively. Here again the normalized GD algorithm demonstrates a superior rate in convergence to the final solution.

## Conclusions

We presented the first theoretical evidence for the convergence of normalized GD in non-linear models. While previous results on standard GD for two-layer neural networks trained with logistic/exponential loss proved a rate of  $\tilde{O}(1/t)$  for the training loss, we showed that normalized GD enjoys an exponential rate. We also studied for the first time, the stability of normalized GD and derived bounds on its generalization performance for convex objectives including infinitely

wide two-layer neural networks. We also briefly discussed the stochastic normalized GD algorithm. As a future direction, we believe extensions of our results to deep neural networks is interesting. Notably, we expect several of our results to be still true for deep neural networks. Extending the self lower-boundedness property in Lemma 4 for smooth activation functions is another important direction.

## Acknowledgements

This work was partially supported by NSF under Grant CCF-2009030.

## References

Allen-Zhu, Z.; Li, Y.; and Song, Z. 2019. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 242–252. PMLR.

- Arora, S.; Du, S.; Hu, W.; Li, Z.; and Wang, R. 2019. Fine-grained analysis of optimization and generalization for over-parameterized two-layer neural networks. In *International Conference on Machine Learning*, 322–332. PMLR.
- Bassily, R.; Belkin, M.; and Ma, S. 2018. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*.
- Bousquet, O.; and Elisseeff, A. 2002. Stability and generalization. *The Journal of Machine Learning Research*, 2: 499–526.
- Cao, Y.; and Gu, Q. 2019. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32.
- Cao, Y.; and Gu, Q. 2020. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3349–3356.
- Charles, Z.; and Papailiopoulos, D. 2018. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, 745–754. PMLR.
- Chatterji, N. S.; Long, P. M.; and Bartlett, P. 2021. When does gradient descent with logistic loss interpolate using deep networks with smoothed ReLU activations? In *Conference on Learning Theory*, 927–1027. PMLR.
- Chizat, L.; and Bach, F. 2020. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, 1305–1338. PMLR.
- Du, S.; Lee, J.; Li, H.; Wang, L.; and Zhai, X. 2019. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, 1675–1685. PMLR.
- Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, 1225–1234. PMLR.
- Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.
- Ji, Z.; and Telgarsky, M. 2018. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*.
- Ji, Z.; and Telgarsky, M. 2020. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33: 17176–17186.
- Ji, Z.; and Telgarsky, M. 2021. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, 772–804. PMLR.
- Lei, Y.; and Ying, Y. 2020. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, 5809–5819. PMLR.
- Li, Y.; and Liang, Y. 2018. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31.
- Liu, C.; Zhu, L.; and Belkin, M. 2022. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116.
- Lojasiewicz, S. 1963. A topological property of real analytic subsets. *Coll. du CNRS, Les equations aux derive es partielles*.
- Lyu, K.; and Li, J. 2019. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*.
- Nacson, M. S.; Lee, J.; Gunasekar, S.; Savarese, P. H. P.; Srebro, N.; and Soudry, D. 2019. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3420–3428. PMLR.
- Nesterov, Y. 2003. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.
- Oymak, S.; and Soltanolkotabi, M. 2019. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, 4951–4960. PMLR.
- Oymak, S.; and Soltanolkotabi, M. 2020. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 84–105.
- Polyak, B. 1963. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3: 864–878.
- Rahimi, A.; and Recht, B. 2007. Random Features for Large-Scale Kernel Machines. In Platt, J.; Koller, D.; Singer, Y.; and Roweis, S., eds., *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.
- Rosset, S.; Zhu, J.; and Hastie, T. J. 2003. Margin Maximizing Loss Functions. In *NIPS*.
- Safran, I. M.; Yehudai, G.; and Shamir, O. 2021. The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. In *Conference on Learning Theory*, 3889–3934. PMLR.
- Schliserman, M.; and Koren, T. 2022. Stability vs Implicit Bias of Gradient Methods on Separable Data and Beyond. *arXiv preprint arXiv:2202.13441*.
- Schmidt, M.; and Roux, N. L. 2013. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*.
- Shamir, O. 2021. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85): 1–20.
- Soudry, D.; Hoffer, E.; Nacson, M. S.; Gunasekar, S.; and Srebro, N. 2018. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878.
- Taheri, H.; and Thrampoulidis, C. 2023. On Generalization of Decentralized Learning with Separable Data. *arXiv:2209.07116*.



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

Vaswani, S.; Bach, F.; and Schmidt, M. 2019. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, 1195–1204. PMLR.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zou, D.; Cao, Y.; Zhou, D.; and Gu, Q. 2020. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine learning*, 109(3): 467–492.