Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

When will gradient methods converge to max-margin classifier under ReLU models?

Tengyu Xu*[1] | Yi Zhou[2] | Kaiyi Ji[1] | Yingbin Liang[1]

[1]Department of Electrical and Computer Engineering, The Ohio State University, Ohio, United States

[2]Department of Electrical and Computer Engineering, University of Utah, Utah, United States

Correspondence

*Corresponding author.

Email: xu.3260@osu.edu

Present Address

2015 Neil Ave, Columbus, OH 43210, United States

Summary

We study the implicit bias of gradient descent methods in solving a binary classification problem over a linearly separable dataset. The classifier is described by a nonlinear ReLU model and the objective function adopts the exponential loss function. We first characterize the landscape of the loss function and show that there can exist spurious asymptotic local minimal besides asymptotic global minimal. We then show that gradient descent (GD) can converge to either a global or a local max-margin direction, or may diverge from the desired max-margin direction in a general context. For stochastic gradient descent (SGD), we show that it converges in expectation to either the global or the local max-margin direction if SGD converges. We further explore the implicit bias of these algorithms in learning a multi-neuron network under certain stationary conditions, and show that the learned classifier maximizes the margins of each sample pattern partition under the ReLU activation.

KEYWORDS:

algorithm, classification, linear model, machine learning, neural network

1 | INTRODUCTION

It has been observed in various machine learning problems recently that the gradient descent (GD) algorithm and the stochastic gradient descent (SGD) algorithm converge to solutions with certain properties even without explicit regularization in the objective function. Correspondingly, theoretical analysis has been developed to explain such implicit regularization property. For example, it has been shown in Gunasekar, Lee, Soudry, and Srebro (2018a); Gunasekar, Woodworth, Bhojanapalli, Neyshabur, and Srebro (2017) that GD converges to the solution with the minimum norm under certain initialization for regression problems, even without an explicit norm constraint.

Another type of implicit regularization, where GD converges to the max-margin classifier, has been recently studied in Gunasekar et al. (2018a); Ji and Telgarsky (2018); Nacson, Lee, et al. (2019); Soudry, Hoffer, and Srebro (2018) for classification problems as we describe below. Given a set of training samples $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$, where $\mathbf{x}_i$ denotes a feature vector and $y_i \in \{-1, +1\}$ denotes the corresponding label, the goal is to find a desirable linear model (i.e., a classifier) by solving the following empirical risk minimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \mathbf{w}^\intercal \mathbf{x}_i).$$

It has been shown in Nacson, Lee, et al. (2019); Soudry et al. (2018) that if the loss function $\ell(\cdot)$ is monotonically strictly decreasing and satisfies proper tail conditions (e.g., the exponential loss), and the data are linearly separable, then GD converges to the solution $\mathbf{w}$ with infinite norm and the maximum margin direction of the data, although there is no explicit regularization towards the max-margin direction in the objective function. Such a phenomenon is referred to as the implicit bias of GD, and can help to explain some experimental results. For example, even when the training error achieves zero (i.e., the resulting model enters into the linearly separable region that correctly classifies the data), the testing error continues to decrease, because the direction of the model parameter continues to have an improved margin. Such a study has been further generalized to hold

for various other types of gradient-based algorithms Gunasekar et al. (2018a). Moreover, Ji and Telgarsky (2018) analyzed the convergence of GD with no assumption on the data separability, and characterized the implicit regularization to be in a subspace-based form.

The focus of this paper is on the following two fundamental issues, which have not been well addressed by existing studies.

- Existing studies so far focused only on the linear classifier model. An important question one naturally asks is what happens for the more general nonlinear leaky ReLU and ReLU models. Will GD still converge, and if so will it converge to the max-margin direction? Our study here provides new insights for the ReLU model that have not been observed for the linear model in the previous studies.

- Existing studies mainly analyzed the convergence of GD with the only exceptions Ji and Telgarsky (2018); Nacson, Srebro, and Soudry (2019) on SGD. However, Ji and Telgarsky (2018) did not establish the convergence to the max-margin direction for SGD, and Nacson, Srebro, and Soudry (2019) established the convergence to the max-margin solution only epochwisely for cyclic SGD (not iterationwise for SGD under random sampling with replacement). Moreover, both studies considered only the linear model. Here, our interest is to explore the iterationwise convergence of SGD under random sampling with replacement to the max-margin direction, and our result can shed insights for online SGD. Furthermore, our study provides new understanding for the nonlinear ReLU and leaky ReLU models.

1.1 | Main contributions

We summarize our main contributions, where our focus is on the exponential loss function under ReLU model.

We first characterize the landscape of the empirical risk function under the ReLU model, which is nonconvex and nonsmooth. We show that such a risk function has asymptotic global minima and asymptotic spurious local minima. Such a landscape is in sharp contrast to that under the linear model previously studied in Soudry et al. (2018), where there exist only equivalent global minima.

Based on the landscape property, we show that the implicit bias property in the course of the convergence of GD can fall into four cases: converges to the asymptotic global minimum along the max-margin direction, converges to an asymptotic local minimum along a local max-margin direction, stops at a finite spurious local minimum, or oscillates between the linearly separable and misclassified regions without convergence. Such a diverse behavior is also in sharp difference from that under the linear model Soudry et al. (2018), where GD always converges to the max-margin direction.

We then take a further step to study the implicit bias of SGD. We show that the expected averaged weight vector normalized by its expected l_2 norm converges to the global max-margin direction or local max-margin direction, as long as SGD stays either in the linearly separable region or in a region of the local minima defined by a subset of data samples with positive label. The proof here requires considerable new technical developments, which are very different from the traditional analysis of SGD, e.g., F. Bach and Moulines (2013); F. R. Bach (2014); Bottou, Curtis, and Nocedal (2016); Duchi and Singer (2009); Nemirovskii, Yudin, and Dawson (1983); Shalev-Shwartz, Shamir, Srebro, and Sridharan (2009); Xiao (2010). This is because our focus here is on the exponential loss function without attainable global/local minima, whereas traditional analysis typically assumed that the minimum of the loss function is attainable. Furthermore, our goal is to analyze the implicit bias property of SGD, which is also beyond traditional analysis of SGD.

We further extend our analysis to the leaky ReLU model and multi-neuron networks.

2 | RELATED WORK

Implicit bias of gradient descent: Gunasekar et al. (2018a) studied the implicit bias of GD and SGD for minimizing the squared loss function under bounded global minimum, and showed that some of these algorithms converge to a global minimum that is closest to the initial point. Another collection of papers Gunasekar et al. (2018a); Ji and Telgarsky (2018); Nacson, Lee, et al. (2019); Soudry et al. (2018); Telgarsky (2013) characterized the implicit bias of algorithms for the loss functions without attainable global minimum. Telgarsky (2013) showed that AdaBoost converges to an approximate max-margin classifier. Soudry et al. (2018) studied the convergence of GD in logistic regression with linearly separable data and showed that GD converges in direction to the solution of support vector machine at a rate of $1/\ln(t)$. Nacson, Lee, et al. (2019) improved this rate to $\ln(t)/\sqrt{t}$ under the exponential loss via normalized gradient descent. Gunasekar et al. (2018a) further showed that steepest descent can lead to margin maximization under generic norms. Ji and Telgarsky (2018) analyzed the convergence of GD on an arbitrary dataset, and provided the convergence rates along the strongly convex subspace and the separable subspace. Later, the implicit bias of linear neural network was characterized in Gunasekar, Lee, Soudry, and Srebro (2018b); Moroshko et al. (2020). Our work studies the convergence of GD and SGD under the nonlinear ReLU model with the exponential loss, as opposed to the linear model studied by all the above previous work on the same type of loss functions.

Implicit bias of SGD: Ji and Telgarsky (2018) analyzed the average SGD (under random sampling) with fixed learning rate and proved the convergence of the population risk, but did not establish the parameter convergence of SGD in the max-margin direction. Nacson, Srebro, and Soudry

(2019) established the convergence of cyclic SGD epochwisely in direction to the max-margin classifier at a rate $\mathcal{O}(1/\ln t)$. Our work differs from these two studies first in that we study the ReLU model, whereas both of these studies analyzed the linear model. Furthermore, we showed that under SGD with random sampling, the expectation of the averaged weight vector converges in direction to the max-margin classifier at a rate $\mathcal{O}(1/\sqrt{\ln t})$.

Generalization of SGD: There have been extensive studies of the convergence and generalization performance of SGD under various models, of which we cannot provide a comprehensive list due to the space limitations. In general, these type of studies either characterize the convergence rate of SGD or provide the generalization error bounds at the convergence of SGD, e.g., Brutzkus, Globerson, Malach, and Shalev-Shwartz (2017); Li and Liang (2018); Wang, Giannakis, and Chen (2019), but did not characterize the implicit regularization property of SGD, such as the convergence to the max-margin direction as provided in our paper.

3 | RELU CLASSIFICATION MODEL

We consider the binary classification problem, in which we are given a set of training samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. Each training sample $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ contains an input data $\mathbf{x}_i$ and a corresponding binary label $y_i \in \{-1, +1\}$. We denote $I^+ := \{i : y_i = +1\}$ as the set of indices of samples with label $+1$ and denote $I^- := \{i : y_i = -1\}$ in a similar way. Their cardinalities are denoted as n^+ and n^-, respectively, and are assumed to be non-zero. We consider all datasets that are linearly separable, i.e., there exists a linear classifier $\mathbf{w}$ such that $y_i \mathbf{w}^\intercal \mathbf{x}_i > 0$ for all $i = 1, \ldots, n$.

We are interested in training a ReLU model for the classification task. In specific, for a given input data $\mathbf{x}$, the model outputs $\sigma(\mathbf{w}^\intercal \mathbf{x}_i)$, where $\sigma(v) = \max\{0, v\}$ is the ReLU activation function and $\mathbf{w}$ denotes the weight parameters. The predicted label is set to be $\operatorname{sgn}(\mathbf{w}^\intercal \mathbf{x})$. Our goal is to learn a classifier by solving the following empirical risk minimization problem, where we adopt the exponential loss.

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{z}_i), \text{ where } \ell(\mathbf{w}, \mathbf{z}_i) = \exp(-y_i \sigma(\mathbf{w}^\intercal \mathbf{x}_i)). \tag{P}$$

The ReLU activation causes the loss function in problem (P) to be nonconvex and nonsmooth. Therefore, it is important to first understand the landscape property of the loss function, which is critical for characterizing the implicit bias property of the GD and SGD algorithms.

4 | IMPLICIT BIAS OF GD IN LEARNING RELU MODEL

4.1 | Landscape of ReLU model

In order to understand the convergence of GD under the ReLU model, we first study the landscape of the loss function in problem (P), which turns out to be very different from that under the linear activation model. As been shown in Ji and Telgarsky (2018); Soudry et al. (2018), the loss function in problem (P) under linear activation is convex, and achieves asymptotic global minimum, i.e., $\nabla \mathcal{L}(\alpha \mathbf{w}^*) \xrightarrow{\alpha} \mathbf{0}$ and $\mathcal{L}(\alpha \mathbf{w}^*) \xrightarrow{\alpha} 0$ as the scaling constant $\alpha \to +\infty$, only if $\mathbf{w}^*$ is in the linearly separable region. In contrast, under the ReLU model, the asymptotic critical points can be either global minimum or (spurious) local minimum depending on the training datasets, and hence the convergence property of GD can be very different in nature from that under the linear model.

The following theorem characterizes the landscape properties of problem (P). Throughout, we denote the infimum of the objective function in problem (P) as $\mathcal{L}^* = \frac{n^-}{n}$. Furthermore, we call a direction $\mathbf{w}^*$ asymptotically critical if it satisfies $\nabla \mathcal{L}(\alpha \mathbf{w}^*) \to \mathbf{0}$ as $\alpha \to +\infty$.

Theorem 4.1 (Asymptotic landscape property)**.** For problem (P) under the ReLU model, any corresponding asymptotic critical direction $\mathbf{w}^*$ fall into one of the following cases:

- (Asymptotic global minimum): $y_i \mathbf{w}^{*\intercal} \mathbf{x}_i > 0$ for all $i \in I^+ \cup I^-$. Then,
$$\mathcal{L}(\alpha \mathbf{w}^*) \to \mathcal{L}^* \text{ as } \alpha \to +\infty.$$
- (Asymptotic local minimum): $\mathbf{w}^{*\intercal} \mathbf{x}_i > 0$ for all $i \in J^+$ and $\mathbf{w}^{*\intercal} \mathbf{x}_i \le 0$ for all $i \in (I^+ \setminus J^+) \cup I^-$, where $J^+ \subseteq I^+$. Then,
$$\mathcal{L}(\alpha \mathbf{w}^*) \to \mathcal{L}^* + \frac{n^+ - |J^+|}{n} \text{ as } \alpha \to +\infty.$$
- (Local minimum): $\mathbf{w}^{*\intercal} \mathbf{x}_i \le 0$ for all $i \in I^+ \cup I^-$. Then,
$$\mathcal{L}(\mathbf{w}^*) = \mathcal{L}^* + \frac{n^+}{n}.$$

To further elaborate Theorem 4.1, if $\mathbf{w}^*$ classifies all data correctly (i.e., item 1), then the objective function possibly achieves global minimum $\mathcal{L}^*$ along this direction. On the other hand, if $\mathbf{w}^*$ classifies some data with label $+1$ as -1 (item 2), then the objective function achieves a sub-optimal

value along this direction. In the worst case where all data samples are classified as -1 (item 3), the ReLU unit is never activated and hence the corresponding objective function has constant value 1. We note that the cases in items 2 and 3 may or may not take place depending on specific datasets, but if they do occur, the corresponding $\mathbf{w}^*$ are spurious (asymptotic) local minima. In summary, the landscape under the ReLU model can be partitioned into different regions, where gradient descent algorithms can have different implicit bias as we show next.

4.2 | Convergence of GD

In this subsection, we analyze the convergence of GD in learning the ReLU model. At each iteration t, GD performs the update

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t), \tag{GD}$$

where η denotes the stepsize. For the linear model whose loss function has infinitely many asymptotic global minima, it has been shown in Soudry et al. (2018) that GD always converges to the max-margin direction. Such a phenomenon is regarded as the implicit bias property of GD. Here, for the ReLU model, we are also interested in analyzing whether such an implicit-bias property still holds. Furthermore, since the loss function under the ReLU model possibly contains spurious asymptotic local minima, the convergence of GD under the ReLU model should be very different from that under the linear model.

Next, we introduce various notions of margin in order to characterize the implicit bias under the ReLU model. The global max-margin direction of samples in I^+ is defined as

$$\widehat{\mathbf{w}}^+ = \underset{\|\mathbf{w}\|=1}{\arg\max} \min_{i \in I^+} (\mathbf{w}^\intercal \mathbf{x}_i).$$

Such a notion of max-margin is natural because the ReLU activation function can suppress negative inputs. We note that here $\widehat{\mathbf{w}}^+$ may not locate in the linearly separable region, and hence it may not be parallel to any (asymptotic) global minimum. As we show next, only when $\widehat{\mathbf{w}}^+$ is in the linearly separable region, GD may converge in direction to such a max-margin direction under the ReLU model. Furthermore, for each given subset $J^+ \subseteq I^+$, we define the associated local max-margin direction $\widehat{\mathbf{w}}_J^+$ as

$$\widehat{\mathbf{w}}_J^+ = \underset{\|\mathbf{w}\|=1}{\arg\max} \min_{i \in J^+} (\mathbf{w}^\intercal \mathbf{x}_i).$$

We further denote the set of asymptotic local minima with respect to $J^+ \subseteq I^+$ (see Theorem 4.1 item 2) as

$$\mathcal{W}_J^+ := \{\mathbf{w}^\intercal \mathbf{x}_i > 0,\ \forall i \in J^+ \text{ and } \mathbf{w}^\intercal \mathbf{x}_i \le 0,\ \forall i \in (I^+ \setminus J^+) \cup I^-\}.$$

Of course, $\mathcal{W}_J^+$ may or may not be empty for a certain J^+, and $\widehat{\mathbf{w}}_J^+$ may or may not belong to $\mathcal{W}_J^+$ depending on the specific training dataset. As we show next, only when there exists a non-empty $\mathcal{W}_J^+$ and the corresponding $\widehat{\mathbf{w}}_J^+ \in \mathcal{W}_J^+$, GD may converge to such an asymptotic local minimum $\widehat{\mathbf{w}}_J^+$ direction under the ReLU model.

Next, we present the implicit bias of GD for learning the ReLU model in problem (P).

Theorem 4.2. Apply GD to solve problem (P) with arbitrary initialization and a small enough constant stepsize. Then, the sequence $\{\mathbf{w}_t\}_t$ generated by GD falls into one of the following cases.

- $\mathcal{L}(\mathbf{w}_t) \to \mathcal{L}^*$, and $\| \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} - \widehat{\mathbf{w}}^+ \| = \mathcal{O}(\frac{\ln \ln t}{\ln t})$, where $\widehat{\mathbf{w}}^+$ is in linearly separable region;
- the direction of $\mathbf{w}_t$ does not converge and oscillates between linearly separable and misclassified regions, where $\widehat{\mathbf{w}}^+$ is not in linearly separable region;
- $\mathcal{L}(\mathbf{w}_t) \to \mathcal{L}^* + \frac{n^+ - |J^+|}{n}$, and $\| \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} - \widehat{\mathbf{w}}_J^+ \| = \mathcal{O}(\frac{\ln \ln t}{\ln t})$, where $J^+ \neq \emptyset$, and $\widehat{\mathbf{w}}_J^+ \in \mathcal{W}_J^+$;
- $\mathcal{L}(\mathbf{w}_t) = \mathcal{L}^* + \frac{n^+}{n}$, and $\mathbf{w}_t = \hat{\mathbf{w}}_J^+$, where $J^+ = \emptyset$, i.e., GD terminates within finite steps.

Theorem 4.2 characterizes various instances of implicit bias of GD in learning the ReLU model, which the nature of the convergence is different from that in learning the linear model. In specific, GD can either converge in direction to the global max-margin direction $\widehat{\mathbf{w}}^+$ that leads to the global minimum, or converge to the local max-margin direction $\widehat{\mathbf{w}}_J^+$ that leads to a spurious local minimum. Furthermore, it may occur that GD oscillates between the linearly separable region and the misclassified region due to the suppression effect of ReLU function. In this case, GD does not have an implicit bias property and convergence guarantee. We provide two simple examples in the supplementary material to further elaborate these cases.

Next, we illustrate through some examples that GD can fail to learn a proper linear classifier on linearly separable data under the ReLU activation

Example 1 (Figure 1, left). The dataset consists of two samples with label $+1$ and one sample with label -1. These samples satisfy $\mathbf{x}_1^\intercal \mathbf{x}_3 < 0$ and $\mathbf{x}_1^\intercal \mathbf{x}_2 < 0$.

For this example, if we initialize GD at the green classifier, then GD converges to the max-margin direction of the sample $(\mathbf{x}_1, +1)$. Clearly, such a classifier misclassifies the data sample $(\mathbf{x}_2, +1)$.

Proof of Example 1. Consider the first iteration. Note that the sample $\mathbf{z}_3$ has label -1, and from the illustration of Figure 1 (left) we have $\mathbf{w}_0^\intercal \mathbf{x}_3 < 0$, $\mathbf{w}_0^\intercal \mathbf{x}_2 < 0$ and $\mathbf{w}_0^\intercal \mathbf{x}_1 > 0$. Therefore, only the sample $\mathbf{z}_1$ contributes to the gradient, which is given by

$$\nabla_{\mathbf{w}_0} \mathcal{L}(\mathbf{w}_0) = -\exp(-\mathbf{w}_0^\intercal \mathbf{x}_1)\mathbf{x}_1. \tag{1}$$

By the update rule of GD, we obtain that for all t

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \exp(-\mathbf{w}_t^\intercal \mathbf{x}_1)\mathbf{x}_1. \tag{2}$$

By telescoping eq. (2), it is clear that any $\mathbf{w}_t^\intercal \mathbf{x}_2 < 0$ for all t since $\mathbf{x}_1^\intercal \mathbf{x}_2 < 0$. This implies that the sample $\mathbf{z}_2$ is always misclassified. □

Example 2 (Figure 1, right). The dataset consists of one sample with label $+1$ and one sample with label -1. These two samples satisfy $0 < \mathbf{x}_1^\intercal \mathbf{x}_2 \leq 0.5\|\mathbf{x}_2\|^2$.

For this example, if we initialize at the green classifier, then GD oscillates around the direction $\mathbf{x}_2/\|\mathbf{x}_2\|$ and does not converge.

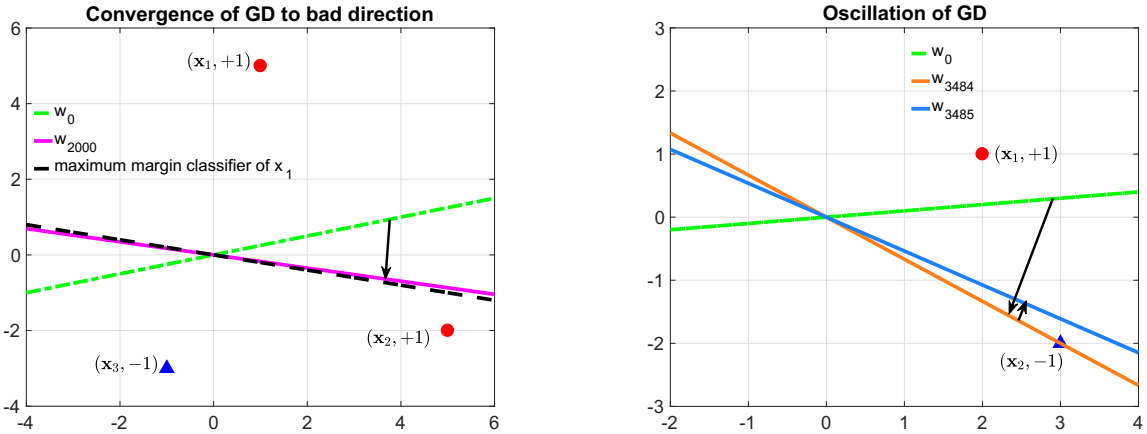

FIGURE 1 Failure of GD in learning ReLU models

Proof of Example 2. Since we initialize GD at $\mathbf{w}_0$ such that $\mathbf{w}_0^\intercal \mathbf{x}_1 > 0$ and $\mathbf{w}_0^\intercal \mathbf{x}_2 < 0$, the sample $\mathbf{z}_2$ does not contribute to the GD update due to the ReLU activation. Next, we argue that there must exists a t such that $\mathbf{w}_t^\intercal \mathbf{x}_2 > 0$. Suppose such t does not exist, we always have $\mathbf{w}_t^\intercal \mathbf{x}_1 = (\mathbf{w}_0 + \sum_{k=0}^{t-1} \exp(-\mathbf{w}_k^\intercal \mathbf{x}_1)\mathbf{x}_1)^\intercal \mathbf{x}_1 > 0$. Then, the linear classifier $\mathbf{w}_t$ generated by GD stays between $\mathbf{x}_1$ and $\mathbf{x}_2$, and the corresponding objective function reduces to a linear model that depends on the sample $\mathbf{z}_1$ (Note that $\mathbf{z}_2$ contributes a constant due to ReLU activation). Following from the results in Ji and Telgarsky (2018); Soudry et al. (2018) for linear model, we conclude that $\mathbf{w}_t$ converges to the max-margin direction $\frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$ as $t \to +\infty$. Since $\mathbf{x}_1^\intercal \mathbf{x}_2 > 0$, this implies that $\mathbf{w}_t^\intercal \mathbf{x}_2 > 0$ as $t \to +\infty$, contradicting with the assumption.

Next, we consider the t such that $\mathbf{w}_t^\intercal \mathbf{x}_1 > 0$ and $\mathbf{w}_t^\intercal \mathbf{x}_2 > 0$, the objective function is given by

$$\mathcal{L}(\mathbf{w}_t) = \exp(-\mathbf{w}_t^\intercal \mathbf{x}_1) + \exp(\mathbf{w}_t^\intercal \mathbf{x}_2),$$

and the corresponding gradient is given by

$$\nabla_{\mathbf{w}_t} \mathcal{L}(\mathbf{w}_t) = -\exp(-\mathbf{w}_t^\intercal \mathbf{x}_1)\mathbf{x}_1 + \exp(\mathbf{w}_t^\intercal \mathbf{x}_2)\mathbf{x}_2.$$

Next, we consider the case that $\mathbf{w}_t^\intercal \mathbf{x}_1 > 0$ for all t. Otherwise, both of $\mathbf{x}_1$ and $\mathbf{x}_2$ are on the negative side of the classifier and GD cannot make any progress as the corresponding gradient is zero. In the case that $\mathbf{w}_t^\intercal \mathbf{x}_1 > 0$ for all t, by the update rule of GD, we obtain that

$$\mathbf{w}_{t+1}^\intercal \mathbf{x}_2 - \mathbf{w}_t^\intercal \mathbf{x}_2 = \eta \exp(-\mathbf{w}_t^\intercal \mathbf{x}_1)\mathbf{x}_1^\intercal \mathbf{x}_2 - \eta \exp(\mathbf{w}_t^\intercal \mathbf{x}_2)\|\mathbf{x}_2\|^2 \leq -0.5\eta\|\mathbf{x}_2\|^2. \tag{3}$$

Clearly, the sequence $\{\mathbf{w}_t^\top \mathbf{x}_2\}_t$ is strictly decreasing with a constant gap, and hence within finite steps we must have $\mathbf{w}_t^\intercal \mathbf{x}_2 \leq 0$. □

4.3 | Implicit bias of SGD in learning ReLU models

In this subsection, we analyze the convergence property and the implicit bias of SGD for solving problem (P). At each iteration t, SGD samples an index $\xi_t \in \{1, \ldots, n\}$ uniformly at random with replacement, and performs the update

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, \mathbf{z}_{\xi_t}). \tag{SGD}$$

Similarly to the convergence of GD characterized in Theorem 4.2, SGD may oscillate between the linearly separable and misclassified regions. Therefore, our major interest here is the implicit bias of SGD when it does converge either to the asymptotic global minimum or local minimum. Thus, without loss of generality, we implicitly assume that $\widehat{\mathbf{w}}^+$ is in the linearly separable region, and the relevant $\widehat{\mathbf{w}}_J^+ \in \mathcal{W}_J^+$. Otherwise, SGD does not even converge.

The implicit bias of SGD with replacement sampling has not been studied in the existing literature, and the proof of the convergence and the characterization of the implicit bias requires substantial new technical developments. In particular, traditional analysis of SGD under convex functions requires the assumption that the variance of the gradient is bounded F. Bach and Moulines (2013); F. R. Bach (2014); Bottou et al. (2016). Instead of making such an assumption, we next prove that SGD enjoys a nearly-constant bound on the variance up to a logarithmic factor of t in learning the ReLU model.

Proposition 1 (Variance bound). Apply SGD to solve problem (P) with any initialization. If there exists $\mathcal{T}$ such that for all $t > \mathcal{T}$, $\mathbf{w}_t$ either stays in the linearly separable region, or in $\mathcal{W}_J^+$, then with stepsize $\eta_k = (k+1)^{-\alpha}$ where $0.5 < \alpha < 1$, the variances of the stochastic gradients sampled by SGD along the iteration path satisfy that for all t,

$$\sum_{k=0}^{t-1} \eta_k^2 \mathbb{E}\|\nabla \ell(\mathbf{w}_k, \mathbf{z}_{\xi_k})\|^2 \leq \mathcal{O}\left(\frac{\ln t}{\gamma^2}\right).$$

Proposition 1 shows that the summation of the norms of the stochastic gradients grows logarithmically fast. This implies that the variance of the stochastic gradients is well-controlled. In particular, if we choose $\eta_k = (k+1)^{-1/2}$, then the bound in Proposition 1 implies that the term $\mathbb{E}\|\nabla \ell(\mathbf{w}_k, \mathbf{z}_{\xi_k})\|^2$ stays at a constant level. Based on the variance bound in Proposition 1, we next establish the convergence rate of SGD for learning the ReLU model. Throughout, we denote $\overline{\mathbf{w}}_t := \frac{1}{t}\sum_{k=0}^{t-1} \mathbf{w}_k$ as the averaged iterates generated by SGD.

Theorem 4.3 (Convergence rate of loss). Apply SGD to solve problem (P) with any initialization. If there exist $\mathcal{T}$ such that for all $t > \mathcal{T}$, $\mathbf{w}_t$ either stays in the linearly separable region, then with the stepsize $\eta_k = (k+1)^{-\alpha}$, where $0.5 < \alpha < 1$, the averaged iterates generated by SGD satisfies

$$\mathbb{E}\mathcal{L}(\overline{\mathbf{w}}_t) - \mathcal{L}^* \leq \mathcal{O}\left(\frac{\ln^2 t}{t^{1-\alpha}}\right), \quad \|\mathbb{E}\overline{\mathbf{w}}_t\| \geq \mathcal{O}(\ln t).$$

If there exist $\mathcal{T}$ such that for all $t > \mathcal{T}$, $\mathbf{w}_t$ stays in $\mathcal{W}_J^+$, then with the same stepsize

$$\mathbb{E}\mathcal{L}(\overline{\mathbf{w}}_t) - \left(\mathcal{L}^* + \frac{n^+ - |J^+|}{n}\right) \leq \mathcal{O}\left(\frac{\ln^2 t}{t^{1-\alpha}}\right), \quad \|\mathbb{E}\overline{\mathbf{w}}_t\| \geq \mathcal{O}(\ln t).$$

Theorem 4.3 establishes the convergence rate of the expected risk of the averaged iterates generated by SGD. It can be seen that the convergence of SGD achieves different loss values corresponding to global and local minimum in different regions. The stepsize is set to be diminishing to compensate the variance introduced by SGD. In particular, if α is chosen to be sufficiently close to 0.5, then the convergence rate is nearly of the order $\mathcal{O}(\ln^2 t/\sqrt{t})$, which matches the standard result of SGD in convex optimization up to an logarithmic order. Theorem 4.3 also implies that the convergence of SGD is attained as $\|\mathbb{E}\overline{\mathbf{w}}_t\| \to +\infty$ at a rate of $\mathcal{O}(\ln t)$. We note that the analysis of Theorem 4.3 is different from that of SGD in traditional convex optimization, which requires the global minimum to be achieved at a bounded point and assumes the variance of the stochastic gradients is bounded by a constant Duchi and Singer (2009); Nemirovski, Juditsky, Lan, and Shapiro (2009); Shalev-Shwartz et al. (2009). These assumptions do not hold here.

Theorem 4.4 (Implicit bias of SGD). Apply SGD to solve problem (P) with any initialization. If there exist $\mathcal{T}$ such that for all $t > \mathcal{T}$, $\mathbf{w}_t$ stays in the linearly separable region, then with the stepsize $\eta_k = (k+1)^{-\alpha}$ where $0.5 < \alpha < 1$, the sequence of the averaged iterate $\{\overline{\mathbf{w}}_t\}_t$ generated by SGD satisfies

$$\left\| \frac{\mathbb{E}\overline{\mathbf{w}}_t}{\|\mathbb{E}\overline{\mathbf{w}}_t\|} - \widehat{\mathbf{w}}^+ \right\|^2 = \mathcal{O}\left(\frac{1}{\ln t}\right).$$

If there exist $\mathcal{T}$ such that for all $t > \mathcal{T}$, $\mathbf{w}_t$ stays in $\mathcal{W}_J^+$, then with the same stepsize

$$\left\| \frac{\mathbb{E}\overline{\mathbf{w}}_t}{\|\mathbb{E}\overline{\mathbf{w}}_t\|} - \widehat{\mathbf{w}}_J^+ \right\|^2 = \mathcal{O}\left(\frac{1}{\ln t}\right).$$

Theorem 4.4 shows that the direction of the expected averaged iterate $\mathbb{E}[\overline{\mathbf{w}}_t]$ generated by SGD converges to the max-margin direction $\widehat{\mathbf{w}}^+$, without any explicit regularizer in the objective function. The proof of Theorem 4.4 requires a detailed analysis of the SGD update under the ReLU model and is substantially different from that under the linear model Nacson, Lee, et al. (2019); Nacson, Srebro, and Soudry (2019); Soudry et al.

(2018). In particular, we need to handle the variance of the stochastic gradients introduced by SGD and exploit its classification properties under the ReLU model.

We next provide an example class of datasets (which has been studied in Combes, Pezeshki, Shabanian, Courville, and Bengio (2018)), for which we show that SGD stays stably in the linearly separable region.

Proposition 2. If the linear separable samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ satisfy the following conditions given in Combes et al. (2018):

- For all $(i, j) \in I^+ \times I^+ \cup I^- \times I^-$, it holds that $\mathbf{x}_i^\intercal \mathbf{x}_j > 0$;
- For all $(i, j) \in I^+ \times I^- \cup I^- \times I^+$, it holds that $\mathbf{x}_i^\intercal \mathbf{x}_j < 0$,

then there exists a $\bar{t} \in \mathbb{N}$ such that for all $t \geq \bar{t}$ the sequence generated by SGD stays in the linearly separable region, as long as SGD is not initialized at the local minima described in item 3 of Theorem 4.1.

We also want to point out that any linearly separable dataset can satisfy the condition in Proposition 2 after a proper transformation, e.g., data augmentation by padding 1s to the samples with label $+1$ and -1s to the samples with label -1. Such data transformation changes the landscape of the ReLU model into a more optimization-friendly version that facilitates to regularize the SGD path.

5 | FURTHER EXTENSIONS AND DISCUSSIONS

5.1 | Leaky ReLU models

The leaky ReLU activation takes the form $\sigma(v) = \max(\alpha v, v)$, where the parameter $(0 \leq \alpha \leq 1)$. Clearly, leaky ReLU takes the linear and ReLU models as two special cases, respectively corresponding to $\alpha = 0$ and $\alpha = 1$. Since the convergence of GD/SGD of the ReLU model is very different from that of the linear model, a natural question to ask is whether leaky ReLU with intermediate parameters $0 < \alpha < 1$ takes the same behavior as the linear or ReLU model.

It can be shown that the loss function in problem (P) under the leaky ReLU model has only asymptotic global minima achieved by $\mathbf{w}^*$ in the separable region with infinite norm (there does not exist asymptotic local minima). Hence, the convergence of GD is similar to that under the linear model, where the only difference is that the max-margin classifier needs to be defined based on leaky ReLU as follows.

For the given set of linearly separable data samples, we construct a new set of data $\mathbf{z}_i^* = (\mathbf{x}_i^*, y_i^*)$, in which $\mathbf{x}_i^* = \mathbf{x}_i$, $\forall i \in I^+$, $\mathbf{x}_i^* = \alpha \mathbf{x}_i$, $\forall i \in I^-$, and $y_i^* = y_i$, $\forall i \in I^+ \cup I^-$. Essentially, the data samples with label -1 are scaled by the parameter α of leaky ReLU. Without loss of generality, we assume that the max-margin classifier for data $\{\mathbf{x}_i^*\}$ passes through the origin after a proper translation. Then, we define the max-margin direction of data $\mathbf{X}^*$ as

$$\widehat{\mathbf{w}}^* = \arg\max_{\|\mathbf{w}\|=1} \min_{i \in I^+ \cup I^-} (y_i^* \mathbf{w}^\intercal \mathbf{x}_i^*).$$

Then, following the result under the linear model in Soudry et al. (2018), it can be shown that GD with arbitrary initialization and small constant stepsize for solving problem (P) under the leaky ReLU model satisfies that $\mathcal{L}(\mathbf{w})$ converges to zero, and $\mathbf{w}$ converges to the max-margin direction, i.e., $\lim_{t\to\infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} = \widehat{\mathbf{w}}^*$, with its norm going to infinity.

Furthermore, following our result of Theorem 4.4, it can be shown that for SGD applied to solve problem (P) with any initialization, if there exists $\mathcal{T}$ such that for all $t > \mathcal{T}$ $\mathbf{w}_t$ stays in the linearly separable region, then with the stepsize $\eta_k = (k+1)^{-\alpha}$, $0.5 < \alpha < 1$, the sequence of the averaged iterate $\{\overline{\mathbf{w}}_t\}_t$ generated by SGD satisfies

$$\left\| \frac{\mathbb{E}\overline{\mathbf{w}}_t}{\|\mathbb{E}\overline{\mathbf{w}}_t\|} - \widehat{\mathbf{w}}^* \right\|^2 = \mathcal{O}\left(\frac{1}{\ln t}\right).$$

Thus, for SGD under the leaky ReLU model, the normalized average of the parameter vector converges in direction to the max-margin classifier.

5.2 | Multi-neuron Networks

In this subsection, we extend our study of the ReLU model to the problem of training a one-hidden-layer ReLU neural network with K hidden neurons for binary classification. Here, we do not assume linear separability of the dataset. The output of the network is given by

$$f(\mathbf{x}) = \sum_{k=1}^{K} v_k \sigma(\mathbf{w}_k^\intercal \mathbf{x}) = \mathbf{v}^\intercal \sigma(\mathbf{W}^\top \mathbf{x}), \tag{4}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K]$ with each column $\mathbf{w}_k$ representing the weights of the kth neuron in the hidden layer, $\mathbf{v}^\intercal = [v_1, v_2, \cdots, v_K]$ denotes the weights of the output neuron, and $\sigma(\cdot)$ represents the entry-wise ReLU activation function. We assume that $\mathbf{v}$ is a fixed vector whose entries

are nonzero and have both positive and negative values. Such an assumption is natural as it allows the model to have enough capacity to achieve zero loss. The predicted label is set to be the sign of $f(\mathbf{x})$, and the objective function under the exponential loss is given by

$$\mathcal{L}(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n}\exp(-y_i f(\mathbf{x}_i)). \tag{5}$$

Our goal is to characterize the implicit bias of GD and SGD for learning the weight parameters $\mathbf{W}$ of the multi-neuron model. In general, such a problem is challenging, as we have shown that GD may not converge to a desirable classifier even under the single-neuron ReLU model. For this reason, we adopt the same setting as that in (Soudry et al. 2018, Corollary 8), which assumes that the activated neurons do not change their activation status and the training error converges to zero after a sufficient number of iterations, but our result presented below characterizes the implicit bias of GD and SGD in the original feature space, which is different from that in (Soudry et al. 2018, Corollary 8). We define a set of vectors $\{\mathbf{A}_i \in \mathbb{R}^{k\times 1}\}_{i=1}^{n}$, where $\mathbf{A}_i^j = 1$ if the sample $\mathbf{x}_i$ is activated on the jth neuron, i.e., $\mathbf{w}_j^\intercal \mathbf{x}_i > 0$, and set $\mathbf{A}_i^j = 0$ otherwise. Such an $\mathbf{A}_i$ vector is referred to as the activation pattern of $\mathbf{x}_i$. We then partition the set of all training samples into m subsets $\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_m$, so that the samples in the same subset have the same ReLU activation pattern, and the samples in different subsets have different ReLU activation patterns. We call $\mathcal{B}_h,\ h \in [m]$ as the h-th pattern partition. Let $\widetilde{\mathbf{w}}_h = \sum_{k\in\{j:\mathbf{A}_h^j=1\}} v_k \mathbf{w}_k$. Then, for any sample $\mathbf{x} \in \mathcal{B}_h$, the output of the network is given by

$$f(\mathbf{x}) = \sum_{k=1}^{K} v_k \sigma(\mathbf{w}_k^\intercal \mathbf{x}) = \sum_{k\in\{j:\mathbf{A}_h^j=1\}} v_k \mathbf{w}_k^\intercal \mathbf{x} = \widetilde{\mathbf{w}}_h^\intercal \mathbf{x}.$$

We next present our characterization of the implicit bias property of GD and SGD under the above ReLU network model. We define the corresponding max-margin direction of the samples in $\mathcal{B}_h$ as

$$\widehat{\mathbf{w}}_h = \underset{\|\mathbf{w}\|=1}{\arg\max} \min_{\mathbf{x}\in\mathcal{B}_h} (\mathbf{w}^\intercal \mathbf{x}).$$

Then the following theorem characterizes the implicit bias of GD under the multi-neuron network.

Theorem 5.1. Suppose that GD optimizes the loss $\mathcal{L}(\mathbf{W})$ in eq. (5) to zero and there exists $\mathcal{T}$ such that for all $t > \mathcal{T}$, the neurons in the hidden layer do not change their activation status. If $\mathbf{A}_{h_1} \wedge \mathbf{A}_{h_2} = \mathbf{0}$ (where "$\wedge$" denotes the entry-wise logic operator "AND" between digits zero or one) for any $h_1 \neq h_2$, then the samples in the same pattern partition of the ReLU activation have the same label, and

$$\left\| \frac{\widetilde{\mathbf{w}}_h^t}{\|\widetilde{\mathbf{w}}_h^t\|} - \widehat{\mathbf{w}}_h \right\| = \mathcal{O}\Big(\frac{\ln\ln t}{\ln t}\Big), \qquad \text{for all } h \in [m]\,.$$

Differently from (Soudry et al. 2018, Corollary 8) which studies the convergence of the vectorized weight matrix so that the implicit bias of GD is with respect to features being lifted to an extended dimensional space, Theorem 5.1 characterizes the convergence of the weight parameters and the implicit bias in the original feature space. In particular, Theorem 5.1 implies that although the ReLU neural network is a nonlinear classifier, $f(\mathbf{x})$ is equivalent to a ReLU classifier for the samples in the same pattern partition (that are from the same class), which converges in direction to the max-margin classifier $\widehat{\mathbf{w}}_h$ of those data samples. We next let $\breve{\mathbf{w}}_h^t := \frac{1}{t}\sum_{k=0}^{t-1} \widetilde{\mathbf{w}}_h(t)$. Then the following theorem establishes the implicit bias of SGD.

Theorem 5.2. Suppose that SGD optimizes the loss $\mathcal{L}(\mathbf{W})$ in eq. (5) so that there exists $\mathcal{T}$ such that for any $t > \mathcal{T}$, $\mathcal{L}(\mathbf{W}) < 1/n$, the neurons in the hidden layer do not change their activation status, and for any $h_1 \neq h_2$, $\mathbf{A}_{h_1} \wedge \mathbf{A}_{h_2} = \mathbf{0}$. Then, for the stepsize $\eta_k = (k+1)^{-\alpha}$, $0.5 < \alpha < 1$, the samples in the same pattern partition of the ReLU activation have the same label, and

$$\left\| \frac{\mathbb{E}\breve{\mathbf{w}}_h^t}{\|\mathbb{E}\breve{\mathbf{w}}_h^t\|} - \widehat{\mathbf{w}}_h \right\|^2 = \mathcal{O}\left(\frac{1}{\ln t}\right), \qquad \text{for all } h \in [m]\,.$$

Similarly to GD, the averaged SGD in expectation maximizes the margin for every sample partition. At the high level, Theorem 5.1 and Theorem 5.2 imply the following generalization performance of the ReLU network under study. After a sufficiently large number of iterations, the neural network partitions the data samples into different subsets, and for each subset, the distance from the samples to the decision boundary is maximized by GD and SGD. Thus, the learned classifier is robust to small perturbations of the data, resulting in good generalization performance.

6 | CONCLUSION

In this paper, we study the problem of learning a ReLU neural network via gradient descent methods, and establish the corresponding risk and parameter convergence under the exponential loss function. In particular, we show that due to the possible existence of spurious asymptotic local minima, GD and SGD can converge either to the global or local max-margin direction, which in the nature of convergence is very different from that under the linear model in the previous studies. We also discuss the extensions of our analysis to the more general leaky ReLU model and multi-neuron networks. In the future, it is worthy to explore the implicit bias of GD and SGD in learning multi-layer neural network models and under more general (not necessarily linearly separable) datasets.

ACKNOWLEDGMENTS

T. Xu, K. Ji, and Y. Liang were supported in part by the U. S. National Science Foundation under the grants CCF-1900145 and CCF-1801855.

SUPPORTING INFORMATION

Additional information for this article is available in Appendix A, Appendix B, Appendix C, Appendix D, Appendix E, Appendix F, Appendix G and Appendix H.

References

Agarwal, A., Wainwright, M. J., Bartlett, P. L., & Ravikumar, P. K. (2009). Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Proceedings of the Advances in Neural Information Processing System (NIPS)* (p. 1-9). Vancouver, Canada.

Bach, F., & Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Proceedings of the Advances in Neural Information Processing System (NIPS)* (p. 773-781). Harrahs and Harveys, Lake Tahoe.

Bach, F. R. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, *15*(2), 595–627.

Borwein, J., & Lewis, A. S. (2010). *Convex Analysis and Nonlinear Optimization: Theory and Examples*. New York: Springer Science & Business Media.

Bottou, L., Curtis, F. E., & Nocedal, J. (2016). Optimization methods for large-scale machine learning. *arXiv:1606.04838*.

Brutzkus, A., Globerson, A., Malach, E., & Shalev-Shwartz, S. (2017). SGD learns over-parameterized networks that provably generalize on linearly separable data. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Toulon, France.

Bubeck, S. (2015). Convex optimization: algorithms and complexity. *Foundations and Trends in Machine Learning*, *8*(3-4), 231–357. doi: 10.1561/2200000050

Combes, R. T., Pezeshki, M., Shabanian, S., Courville, A., & Bengio, Y. (2018). On the learning dynamics of deep neural networks. *arXiv:1809.06848*.

Duchi, J., & Singer, Y. (2009). Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, *10*(99), 2899–2934.

Ghadimi, S., & Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, *23*(4), 2341–2368. doi: 10.1137/120880811

Gunasekar, S., Lee, J., Soudry, D., & Srebro, N. (2018a). Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 1832–1841). Stockholm, SWEDEN.

Gunasekar, S., Lee, J., Soudry, D., & Srebro, N. (2018b). Implicit bias of gradient descent on linear convolutional networks. In *Proceedings of the Advances in Neural Information Processing System (NIPS)* (pp. 9461–9471). Montreal Canada.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., & Srebro, N. (2017). Implicit regularization in matrix factorization. In *Proceedings of the Advances in Neural Information Processing System (NIPS)* (p. 6151-6159). Long Beach, CA, USA.

Ji, Z., & Telgarsky, M. (2018). Risk and parameter convergence of logistic regression. *arXiv:1803.07300*.

Li, Y., & Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Proceedings of the Advances in Neural Information Processing System (NIPS)* (p. 8157-8166). Montreal Canada.

Moroshko, E., Gunasekar, S., Woodworth, B., Lee, J., Srebro, N., & Soudry, D. (2020). Implicit bias in deep linear classification: initialization scale vs training accuracy. *arXiv:2007.06738*.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., & Soudry, D. (2019). Convergence of gradient descent on separable data. In *Proceedings of the International Conference on Artificial Intelligence and statistics (AISTATS)* (p. 3420-3428). Okinawa, Japan.

Nacson, M. S., Srebro, N., & Soudry, D. (2019). Stochastic gradient descent on separable data: exact convergence with a fixed learning rate. In *Proceedings of the International Conference on Artificial Intelligence and statistics (AISTATS)* (p. 3051-3059). Okinawa, Japan.

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, *19*(4), 1574–1609. doi: 10.1137/070704277

Nemirovskii, A., Yudin, D. B., & Dawson, E. R. (1983). *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley.

Nesterov, Y. (2013). *Introductory Lectures on Convex Optimization: A Basic Course*. New York: Springer Science & Business Media.

Shalev-Shwartz, S., Shamir, O., Srebro, N., & Sridharan, K. (2009). Stochastic convex optimization. In *Proceedings of the Conference on Learning Theory (COLT)*. Montreal, Canada.

Soudry, D., Hoffer, E., & Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, *19*(70), 1–57.

Telgarsky, M. (2013). Margins, shrinkage, and boosting. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 307–315). Atlanta, GA, USA.

Wang, G., Giannakis, G. B., & Chen, J. (2019). Learning ReLU networks on linearly separable data: algorithm, optimality, and generalization. *IEEE Transactions on Signal Processing*, *67*(9), 2357–2370. doi: 10.1109/TSP.2019.2904921

Xiao, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, *11*(88), 2543–2596.

How to cite this article: T. Xu, Y. Zhou, K. Ji, and Y. Liang, When will gradient methods converge to max-margin classifier under ReLU models?, *xxxx*, *xxxx*;*xx*:*x*–*x*.