Stochastic Gradient Descent on Separable Data: Exact Convergence with a Fixed Learning Rate

Mor Shpigel Nacson[1] **Nathan Srebro**[2] **Daniel Soudry**[1]
[1]Technion, Israel, [2]TTI Chicago, USA

Abstract

Stochastic Gradient Descent (SGD) is a central tool in machine learning. We prove that SGD converges to zero loss, even with a fixed (non-vanishing) learning rate — in the special case of homogeneous linear classifiers with smooth monotone loss functions, optimized on linearly separable data. Previous works assumed either a vanishing learning rate, iterate averaging, or loss assumptions that do not hold for monotone loss functions used for classification, such as the logistic loss. We prove our result on a fixed dataset, both for sampling with or without replacement. Furthermore, for logistic loss (and similar exponentially-tailed losses), we prove that with SGD the weight vector converges in direction to the L_2 max margin vector as $O(1/\log(t))$ for almost all separable datasets, and the loss converges as $O(1/t)$ — similarly to gradient descent. Lastly, we examine the case of a fixed learning rate proportional to the minibatch size. We prove that in this case, the asymptotic convergence rate of SGD (with replacement) does not depend on the minibatch size in terms of epochs, if the support vectors span the data. These results may suggest an explanation to similar behaviors observed in deep networks, when trained with SGD.

1 INTRODUCTION

Deep neural networks (DNNs) are commonly trained using stochastic gradient descent (SGD), or one of its variants. During training, the learning rate is typically decreased according to some schedule (e.g., every T epochs we multiply the learning rate by some $\alpha < 1$). Determining the learning rate schedule, and its dependency on other factors, such as the minibatch size, has been the subject of a rapidly increasing number of recent empirical works (Hoffer et al. (2017); Goyal et al. (2017); Jastrzebski et al. (2017); Smith et al. (2018) are a few examples). Therefore, it is desirable to improve our understanding of such issues. However, somewhat surprisingly, we observe that we do not have even a satisfying answer to the basic question

Why do we need to decrease the learning rate during training?

At first, it may seem that this question has already been answered. Many previous works have analyzed SGD theoretically (e.g., see Robbins and Monro (1951); Bertsekas (1999); Geary and Bertsekas (2001); Bach and Moulines (2011); Ben-David and Shalev-Shwartz (2014); Ghadimi et al. (2013); Bubeck (2015); Bottou et al. (2016); Ma et al. (2017) and references therein), under various assumptions. In all previous works, to the best of our knowledge, one must assume a vanishing learning rate schedule, averaging of the SGD iterates, partial strong convexity (i.e., strong convexity in some subspace), or the Polyak-Lojasiewicz (PL) condition (Bassily et al., 2018) — so that the SGD increments or the loss (in the convex case) will converge to zero for generic datasets. However, even near its global minima, a neural network loss is not partially strongly convex, and the PL condition does not hold. Therefore, without a vanishing learning rate or iterate averaging, the gradients are only guaranteed to decrease below some constant value, proportional to the learning rate. Thus, in this case, we may fluctuate near a critical point, but never converge to it.

Consequently it may seem that in neural networks we should always decrease the learning rate in SGD or average the weights, to enable the convergence of the weights to a critical point, and to decrease the loss. However, this reasoning does not hold empirically. In many datasets, even with a fixed learning rate and without averaging, we observe that the training loss can converge to zero. For example, we examine the learning dynamics of a ResNet-18 trained on CIFAR10 in Figure 1. Even though the learning rate is fixed, the training loss converges to zero (and so does the classification error).

Notably, we do not observe any convergence issues, as

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

we may have suspected from previous theoretical results. In fact, if we decrease the learning rate at any point, this only decreases the convergence rate of the training loss to zero. The main benefit of decreasing the learning rate is that it typically improves generalization performance. Such contradiction between existing theoretical and empirical results may indicate a significant gap in our understanding. We are therefore interested in closing this gap.

To do so, we first examine the network dynamics in Figure 1. Since the training error has reached zero after a certain number of iterations, by then the last hidden layer must have become linearly separable. Since the network is trained using the monotone cross-entropy loss (with softmax outputs), by increasing the norm of the weights we decrease the loss. Therefore, if the loss is minimized then the weights would tend to diverge to infinity — as indeed happens. This weight divergence does not affect the scale-insensitive validation (classification) error, which continues to decrease during training. In contrast, the validation loss starts to increase.

To explain this behavior, Soudry et al. (2018b,a) focused on the dynamics of the last layer, for a fixed separable input and no bias. For Gradient Descent (GD) dynamics, Soudry et al. (2018b,a) proved that the training loss converges to zero as $1/t$, the direction of the weight vector converges to the max margin as $1/\log(t)$, and the validation loss increase as $\log(t)$. This had similar dynamics to those observed in Figure 1. However, the dynamics of GD are simpler than those of SGD. Notably, it is well known that on smooth functions, for the iterates of GD, the gradient converges to zero even with a fixed learning rate — just as long as this learning rate is below some fixed threshold (which depends on the smoothness of the function).

Our contributions. In this paper we examine SGD optimization of homogeneous linear classifiers with smooth monotone loss functions, where the data is sampled either with replacement (the sampling regime typically examined in theory), or without replacement (the sampling regime typically used in practice). For simplicity, we focus on binary classification (e.g., logistic regression). First, we prove three basic results:

- The norm of the weights diverges to infinity for any learning rate.
- For a sufficiently small *fixed* learning rate, the loss and gradients converge to zero.
- This upper bound we derived for the maximal learning rate is proportional to the minibatch size, when the data in SGD is sampled with replacement.

Similar behavior to the last property is also observed in deep networks (Goyal et al., 2017; Smith et al., 2018). Next, given an additional assumption that the loss function has an exponential tail (e.g., logistic regression), we prove that for almost all linearly separable datasets (i.e., except for measure zero cases):

- The direction of the weight vector converges to that of the L_2 max margin solution.
- The margin converges as $O(1/\log(t))$, while the training loss converges as $O(1/t)$.

These conclusions for SGD are the same as for GD (Soudry et al., 2018b) — the only difference is the value of the maximal learning rate, which depends on the minibatch size. Therefore, we believe our SGD results might be similarly extended, as GD, to multi-class (Soudry et al., 2018a), other loss functions (Nacson et al., 2018), other optimization methods (Gunasekar et al., 2018b), linear convolutional neural networks (Gunasekar et al., 2018a), and hopefully to nonlinear deep networks.

Finally, under the assumption that the SVM support vectors span the dataset, we further characterize SGD iterate asymptotic behavior. Specifically, we show that, if we keep the learning rate proportional to the minibatch size, then:

- The minibatch size does not affect the asymptotic convergence rate of SGD, in terms of epochs.
- In terms of SGD iterations, the fastest asymptotic convergence rate, is obtained at full batch size, i.e. GD.

These results suggest the large potential of parallelism in separable problems, as observed in deep networks (Goyal et al., 2017; Smith et al., 2018).

2 PRELIMINARIES

Consider a dataset $\{\mathbf{x}_n, y_n\}_{n=1}^N$, with binary labels $y_n \in \{-1, 1\}$. We analyze learning by minimizing an empirical loss of homogeneous linear predictors (i.e., without bias), of the form

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} \ell\left(y_n \mathbf{w}^\top \mathbf{x}_n\right) , \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector. To simplify notation, we assume that $\forall n : \ y_n = 1$ — this is true without loss of generality, since we can always re-define $y_n \mathbf{x}_n$ as $\mathbf{x}_n$.

We are particularly interested in problems that are linearly separable and with a smooth strictly decreasing and non-negative loss function. Therefore, we assume:

Assumption 1. *The dataset is strictly linearly separable:* $\exists \mathbf{w}_*$ *such that* $\forall n : \ \mathbf{w}_*^\top \mathbf{x}_n > 0$.

Given that the data is linearly separable, the maximal L_2 margin is strictly positive

$$\gamma = \max_{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| = 1} \min_n \mathbf{w}^\top \mathbf{x}_n > 0 \tag{2}$$

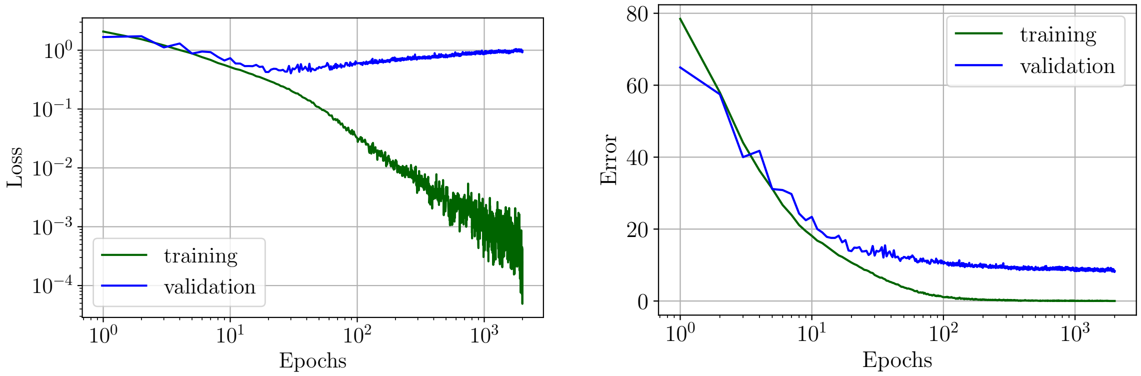

Figure 1: Training of a convolutional neural network on CIFAR10 using stochastic gradient descent with constant learning rate, softmax output and a cross entropy loss. We observe that, approximately: (1) The training loss and (classification) error both decays to zero; (2) after a while, the validation loss starts to increase; and (3) in contrast, the validation (classification) error slowly improves. In Soudry et al. (2018b), the authors observed similar results with momentum.

Assumption 2. *$\ell(u)$ is a positive, differentiable, β-smooth function (*i.e.*, its derivative is β-Lipshitz), monotonically decreasing to zero, (so*[1] *$\forall u : \ell(u) > 0, \ell'(u) < 0$ and $\lim_{u\to\infty} \ell(u) = \lim_{u\to\infty} \ell'(u) = 0$), and $\limsup_{u\to-\infty} \ell'(u) \neq 0$.*

Many common loss functions, including the logistic and probit losses, follow Assumption 2. Assumption 2 also straightforwardly implies that $\mathcal{L}(\mathbf{w})$ is a $\beta\sigma^2_{\max}$-smooth function, where the columns of $\mathbf{X}$ are all samples, and $\sigma_{\max}$ is the maximal singular value of $\mathbf{X}$.

Under these conditions, the infimum of the optimization problem is zero, but it is not attained at any finite $\mathbf{w}$. Furthermore, no finite critical point $\mathbf{w}$ exists. We consider minimizing eq. 1 using Stochastic Gradient Descent (SGD) with a fixed learning rate η, *i.e.,* with steps of the form:

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \frac{\eta}{B} \sum_{n\in\mathcal{B}(t)} \ell'\left(\mathbf{w}(t)^\top \mathbf{x}_n\right)\mathbf{x}_n, \quad (3)$$

where $\mathcal{B}(t) \subset \{1,\dots,N\}$ is a minibatch of B distinct indices, chosen so $K = N/B$ is an integer, and that it satisfies one of the following assumptions. The first option is the assumption of random sampling with replacement:

Assumption 3a. [Random sampling with replacement] *At each iteration t we randomly and uniformly sample a minibatch $\mathcal{B}(t)$ of B distinct indices, i.e. so each sample has an identical probability to be selected.*

For example, this assumption holds if at each iteration we uniformly sample the indices without replacement from $\{1,\dots,N\}$, or uniformly sample $k \in \{1,\dots,K\}$ and select $\mathcal{B}(t) = \mathcal{B}_k$, where $\{\mathcal{B}_k\}_{k=0}^{K-1}$ is some fixed partition of the data indices, i.e.,

$$\cup_{k=0}^{K-1} \mathcal{B}_k = \{1,\dots,N\}\,.$$

This assumption is rather common in theoretical analysis, but less common in practice. The next alternative sampling method is more common in practice:

Assumption 3b (Sampling without replacement)**.** *At each epoch, the minibatches partition the data:*

$$\forall u \in \{0,1,2,\dots\} : \cup_{k=0}^{K-1} \mathcal{B}(Ku+k) = \{1,\dots,N\}\,.$$

This way, each sample is chosen exactly once at each epoch, and SGD completes balanced passes over the data. An important special case of this assumption is random sampling without replacement, which is the practically common method. Other special cases are periodic sampling (round-robin), and even adversarial selection of the order of the samples.

3 MAIN RESULT 1: THE LOSS CONVERGES TO A GLOBAL INFIMUM

The weight norm always diverges to infinity, for any learning rate, as we prove next.

Lemma 1. *Given assumptions 1 and 2, and any starting point $\mathbf{w}(0)$, the iterates of SGD on $\mathcal{L}(\mathbf{w})$ (eq. 3), with either sampling regimes (Assumption 3a or 3b), diverge to infinity, i.e. $\|\mathbf{w}(t)\| \to \infty$.*

Proof. Since the data is linearly separable, $\exists \mathbf{w}_*$ such that $\forall n : \mathbf{w}_* \mathbf{x}_n > 0$. We examine the dot product of $\mathbf{w}^*$ with

[1] The requirement of nonnegativity and that the loss asymptotes to zero is purely for convenience. It is enough to require the loss is monotone decreasing and bounded from below. Any such loss asymptotes to some constant, and is thus equivalent to one that satisfies this assumption, up to a shift by that constant.

the iterates of SGD

$$\mathbf{w}_*^\top \mathbf{w}\left(t\right) = \mathbf{w}_*^\top \mathbf{w}\left(0\right) - \frac{\eta}{B}\sum_{u=0}^{t-1}\sum_{n\in\mathcal{B}(u)} \ell'\left(\mathbf{x}_n^\top \mathbf{w}\left(u\right)\right)\mathbf{w}_*^\top \mathbf{x}_n \,.$$

Since $\forall n:\ \mathbf{w}_*\mathbf{x}_n > 0$ and $-\ell'(u) > 0$ for any finite u, we get that either $\mathbf{w}_*^\top \mathbf{w}\left(t\right) \to \infty$ or $\ell'\left(\mathbf{x}_n^\top \mathbf{w}\left(u\right)\right) \to 0$. In the first case, from Cauchy-Shwartz

$$\left\|\mathbf{w}\left(t\right)\right\| \geq \left\|\mathbf{w}_*^\top \mathbf{w}\left(t\right)\right\| / \left\|\mathbf{w}_*\right\| \to \infty \,.$$

In the second case, since $-\ell'\left(u\right)$ is strictly positive for any finite value, and achieves zero only at $u \to \infty$, we must have $\mathbf{x}_n^\top \mathbf{w}\left(t\right) \to \infty$, which again implies

$$\left\|\mathbf{w}\left(t\right)\right\| \geq \left\|\mathbf{x}_n^\top \mathbf{w}\left(t\right)\right\| / \left\|\mathbf{x}_n\right\| \to \infty \,.$$

Combing both cases, we prove the theorem. □

As the weights go to infinity, we wish to understand the asymptotic behavior of the loss. As the next theorem shows, if the fixed learning rate η is sufficiently small, then we get that the loss converges to zero.

Theorem 1. *Let $\mathbf{w}\left(t\right)$ be the iterates of SGD (eq. 3) from any starting point $\mathbf{w}(0)$, where samples are either* (case 1) *selected randomly with replacement (Assumption 3a)) and with learning rate*

$$\frac{\eta}{B} < \frac{2\gamma^2}{\beta\sigma_{\max}^2}\,, \tag{4}$$

or (case 2) *sampled without replacement (Assumption 3b)) and with learning rate*

$$\frac{\eta}{B} < \min\left[\frac{1}{2K\beta\sigma_{\max}^2}, \frac{\gamma}{2\beta\sigma_{\max}^3\left(K+\gamma^{-1}\sigma_{\max}\right)}\right]\,. \tag{5}$$

For linearly separable data (Assumption 1), and smooth-monotone loss function (Assumption 2), we have the following, almost surely (with probability 1*) in the first case, and surely in the second case:*

1. *The loss converges to zero:*

$$\lim_{t\to\infty}\mathcal{L}\left(\mathbf{w}\left(t\right)\right) = 0,$$

2. *All samples are correctly classified, given sufficiently long time:*

$$\forall n: \lim_{t\to\infty}\mathbf{w}\left(t\right)^\top \mathbf{x}_n = \infty,$$

3. *The iterates of SGD are square summable:*

$$\sum_{t=0}^{\infty}\left\|\mathbf{w}\left(t+1\right) - \mathbf{w}\left(t\right)\right\|^2 < \infty.$$

The complete proof of this theorem is given in section A in the appendix. The proof relies on the following key lemma

Lemma 2. *The L_2 max margin lower bounds the minimal "non-negative right eigenvalue" of* $\mathbf{X}$

$$\gamma = \max_{\mathbf{w}\in\mathbb{R}^d:\|\mathbf{w}\|=1}\ \min_n \mathbf{w}^\top \mathbf{x}_n \leq \min_{\mathbf{v}\in\mathbb{R}^d_{\geq 0}:\|\mathbf{v}\|=1}\left\|\mathbf{X}\mathbf{v}\right\| \tag{6}$$

Proof. In this proof we define $\mathbf{v}^*$ as the minimizer of the right hand side of eq. 6, and $\mathbf{w}_*$ as the maximizer of the optimization problem on the left hand side of the same equation. On the one hand

$$\mathbf{w}_*^\top \mathbf{X}\mathbf{v}^* \overset{(1)}{\leq} \left\|\mathbf{w}_*\right\|\left\|\mathbf{X}\mathbf{v}^*\right\| \overset{(2)}{=} \min_{\mathbf{v}\in\mathbb{R}^d_{\geq 0}:\|\mathbf{v}\|=1}\left\|\mathbf{X}\mathbf{v}\right\|\,, \tag{7}$$

where in (1) we used Cauchy-Shwartz inequality, and in (2) we used the definition of $\mathbf{v}^*$, and that $\left\|\mathbf{w}_*\right\| = 1$. On the other hand,

$$\mathbf{w}_*^\top \mathbf{X}\mathbf{v}^* \overset{(1)}{\geq} \gamma\sum_{n=1}^{N} v_n^* \overset{(2)}{\geq} \gamma\sqrt{\sum_{n=1}^{N}(v_n^*)^2} \overset{(3)}{=} \gamma\,, \tag{8}$$

where in (1) we used the definition of the L_2 max margin from the left hand side of eq. 6 and $\mathbf{v}^* \in \mathbb{R}^d_+$, in (2) we used that $v_n \geq 0$ and the triangle inequality, and in (3) we used that $\left\|\mathbf{v}\right\| = 1$. Together, eqs. 7 and 8 imply the Lemma. □

This Lemma is useful since the SGD weight increments in eq. 3 have the form $\mathbf{X}\mathbf{v}$, where $\mathbf{v}$ is some vector with non-negative components. This enables us to bound the norm of the SGD updates using the norm of the full gradient, which allows us to use similar analysis as for GD. Additionally, we note the regime we analyze in Theorem 1 is somewhat unusual, as the weight vector goes to infinity. In many previous works it is assumed that there exists a finite critical point, or that the weights are bounded within a compact domain.

Theorem 1 Implications. In both sampling regimes, we obtained that a fixed (non-vanishing) learning rate results in convergence to zero error. In the case of random sampling with replacement (Assumption 3a) we got a better upper bound on the learning rate (eq. 4), which does not depend on K. Interestingly, this bound matches the empirical findings of Goyal et al. (2017); Smith et al. (2018), which observed that in a large range $\eta \propto B$. Interestingly, in our case the relation $\eta \propto B$ holds exactly for all B in the maximum learning rate (eq. 4). In contrast, for linear regression, the relation becomes sub-linear for large B (Ma et al., 2017).

We also considered here the case when the datapoints are sampled without replacement (Assumption 3b). This is in contrast to most theoretical SGD results, which typically assume sampling with replacement (which is less common in practice). There are a few notable exceptions (Geary and Bertsekas (2001); Bertsekas (2011); Shamir (2016), and references therein). Perhaps the most similar previous result is the classical result of (Proposition 2.1 in Geary and Bertsekas (2001)), which has a similar sampling schedule, and in which the weights can go to infinity. However, in this result the learning rate must go to zero for the SGD iterates to converge. In our case, we are able to relax this assumption since we focus on linear classification with a monotone loss and separable data.

When assuming sampling without replacement (Assumption 3b the learning rate bound (eq. 5) becomes significantly lower — roughly proportional to $1/K$. This is because such a sampling assumption is very pessimistic (e.g., the samples can be selected by an adversary). Therefore, a small (yet non vanishing) learning rate is required to guarantee convergence. Such a dependence on K is expected, since in this case we need to use a incremental gradient method type of proof, where such low learning rates are common. For example, in Bertsekas (2011) Proposition 3.2b, to get a low final error we must have a learning rate $\eta \ll 1/K^2$.

4 MAIN RESULT 2: THE WEIGHT VECTOR DIRECTION CONVERGES TO THE MAX MARGIN

Next, we focus on a special case of monotone loss functions:

Definition 1. *A function $f(u)$ has a "tight exponential tail", if there exist positive constants μ_+, μ_-, and $\bar{u}$ such that $\forall u > \bar{u}$:*

$$(1 - \exp(-\mu_- u))e^{-u} \leq f(u) \leq (1 + \exp(-\mu_+ u))e^{-u}$$

Assumption 4. *The negative loss derivative $-\ell'(u)$ has a tight exponential tail.*

Specifically, this applies to the logistic loss function. Given this additional assumption, we prove that SGD converges to the L_2 max margin solution.

Theorem 2. *For almost all datasets for which the assumptions of Theorem 1 hold, if $-\ell'(u)$ has a tight exponential tail (Assumption 4), then the iterates of SGD, for any $\mathbf{w}(0)$, will behave as:*

$$\mathbf{w}(t) = \hat{\mathbf{w}} \log\left(\frac{\eta}{B} \cdot \frac{t}{K}\right) + \boldsymbol{\rho}(t), \tag{9}$$

where $\hat{\mathbf{w}}$ is the following L_2 max margin separator:

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \ \text{ s.t. } \mathbf{w}^\top \mathbf{x}_n \geq 1, \tag{10}$$

and the residual $\|\boldsymbol{\rho}(t)\|$ is bounded almost surely in the first case of Theorem 1 (random sampling with replacement), or surely in the second case (sampling without replacement).

Thus, from Theorem 2, for almost any linearly separable data set (e.g., with probability 1 if the data is sampled from an absolutely continuous distribution) , the normalized weight vector converges to the normalized max margin vector, i.e.,

$$\lim_{t \to \infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}$$

with rate $1/\log(t)$, identically to GD (Soudry et al., 2018b). Interestingly, the number of minibatches per epoch K affects only the constants. Intuitively, this is reasonable, since if we rescale the time units, then the log term in eq. 9 will only add a constant to the residual $\rho(t)$.

Proof idea. The theorem is proved in appendix section B.1. The proof builds on the results of Soudry et al. (2018b) for GD: as the weights diverge, the loss converges to zero, and only the gradients of the support vector remain significant. This implies that the gradient direction, as a positive linear combination of support vectors converges to the direction of the max margin. The main difficulty in extending the proof to the case of SGD is that at each iteration, $\mathbf{w}(t)$ is updated using only a subset of the data points. This could potentially lead to large difference from the GD solution. However, conceptually, we show that this difference of $\mathbf{w}(t)$ from the GD dynamics solution is $O(1)$ in t. The main novel idea here is that in order to calculate this $O(1)$ difference at time t, we use information on sampling selections made in the future, i.e. at times larger than t.

Convergence Rates. Theorem 2 directly implies the same convergence rates as in GD (Soudry et al., 2018b). Specifically, in the L_2 distance

$$\left\| \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} - \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|} \right\| = O\left(\frac{1}{\log t}\right), \tag{11}$$

in the angle

$$1 - \frac{\mathbf{w}(t)^\top \hat{\mathbf{w}}}{\|\mathbf{w}(t)\| \, \|\hat{\mathbf{w}}\|} = O\left(\frac{1}{\log^2 t}\right), \tag{12}$$

and in the margin gap

$$\frac{1}{\|\hat{\mathbf{w}}\|} - \frac{\min_n \mathbf{x}_n^\top \mathbf{w}(t)}{\|\mathbf{w}(t)\|} = O\left(\frac{1}{\log t}\right). \tag{13}$$

On the other hand, the loss itself decreases as

$$\mathcal{L}(\mathbf{w}(t)) = O\left(\frac{1}{t}\right). \tag{14}$$

In Figure 2 we visualize these results. Additionally, in Figure 3 we observe that the convergence rates remain nearly the same for different minibatch sizes — as long as we linearly scale the learning rate with the minibatch size, i.e. $\eta \propto B$. This behavior fits with the behavior of the maximal learning rate for which SGD converge in the case of sampling with replacement (eq. 4). However, it is not clear from

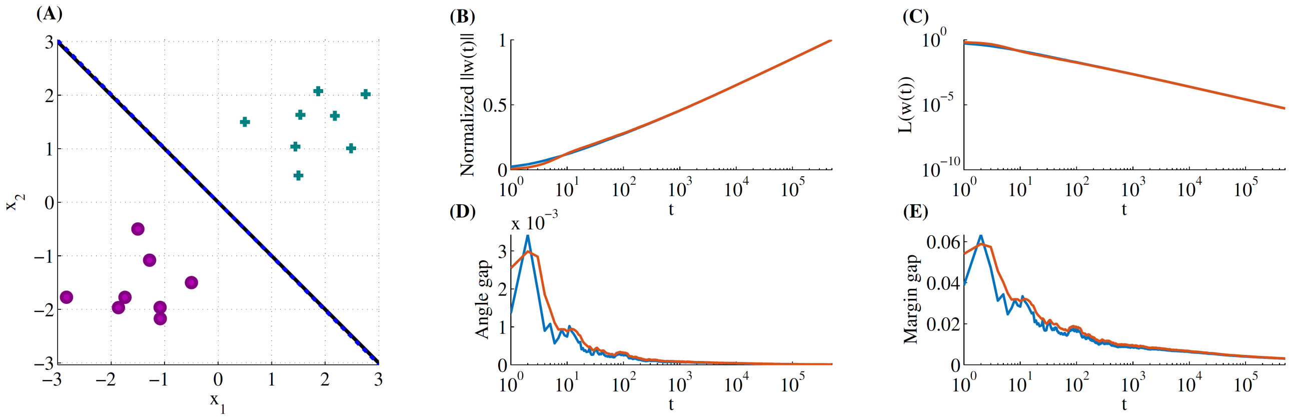

Figure 2: **Visualization of Theorem 2 on a synthetic dataset in which the L_2 max margin vector $\hat{\mathbf{w}}$ is precisely known**. **(A)** The dataset (positive and negatives samples ($y = \pm 1$) are respectively denoted by $'+'$ and $'\circ'$), max margin separating hyperplane (black line), and the asymptotic solution of SGD (dashed blue). For both SGD (blue line) and SGD with momentum (orange line), we show: **(B)** The norm of $\mathbf{w}\,(t)$, normalized so it would equal to 1 at the last iteration, to facilitate comparison. As expected (from eq. 9), the norm increases logarithmically; **(C)** the training loss. As expected, it decreases as t^{-1} (eq. 14); and **(D&E)** the angle and margin gap of $\mathbf{w}\,(t)$ from $\hat{\mathbf{w}}$ (eqs. 12 and 13). As expected, these are logarithmically decreasing to zero. Figure reproduced from Soudry et al. (2018b). We also observe similar behavior with different input dimension d. This is demonstrated in Figure 4

Theorem 2 why the convergence rate stays almost exactly the same with such a linear scaling, since we do not know how does $\rho(t)$ depends on η and B. In the special case where the SVM support vectors span the dataset, we can further characterize $\rho(t)$ asymptotic dependence on η and B. We define $\mathbf{P} \in \mathbb{R}^{d\times d}$ as the orthogonal projection matrix to the subspace spanned by the support vectors, and $\bar{\mathbf{P}} = \mathbf{I} - \mathbf{P}$ as the complementary projection. In addition, we denote α_n as the SVM dual variables so $\hat{\mathbf{w}} = \sum_{n\in\mathcal{S}} \alpha_n \mathbf{x}_n$.

Theorem 3. *Under the conditions and notation of Theorem 2, for almost all datasets, if in addition the support vectors span the data (*i.e. $\operatorname{rank}(\mathbf{X}_{\mathcal{S}}) = \operatorname{rank}(\mathbf{X})$*), where $\mathbf{X}_{\mathcal{S}}$ is a matrix whose columns are only those data points $\mathbf{x}_n$ s.t. $\hat{\mathbf{w}}^\top \mathbf{x}_n = 1$), then $\lim_{t\to\infty} \boldsymbol{\rho}(t) = \tilde{\mathbf{w}}$, where $\tilde{\mathbf{w}}$ is a solution to*

$$\forall n \in \mathcal{S}: \ \exp\left(-\mathbf{x}_n^\top \tilde{\mathbf{w}}\right) = \alpha_n\,,\ \bar{\mathbf{P}}\left(\tilde{\mathbf{w}} - \mathbf{w}(0)\right) = 0\,, \tag{15}$$

The theorem is proved in appendix section B.2. Note that $\tilde{\mathbf{w}}$ is only dependent on the dataset and the initialization. This fact enables us to state the following result for the asymptotic behavior of SGD.

Corollary 1. *Under the conditions and notation of Theorem 3, GD iterate will behave as:*

$$\mathbf{w}(t) = \hat{\mathbf{w}} \log\left(\frac{\eta}{B}\cdot\frac{t}{K}\right) + \tilde{\mathbf{w}} + o(1)\,,$$

where $\hat{\mathbf{w}}$ is the maximum-margin separator, $\tilde{\mathbf{w}}$ is the solution of eq. 15 (which does not depend on K, η and B), and $o(1)$ is a vanishing term. Therefore, if the step size is kept proportional to the minibatch size, i.e., $\eta \propto B$, changing the number of minibatches K is equivalent to linearly re-scaling the time units of t.

From the corollary, we expect the same asymptotic convergence rates for all batch sizes B as long as we scale the learning rate linearly with the batch size, *i.e.*, keep $\eta \propto B$. This is exactly the behavior we observe in Figure 3. Since changing the number of minibatches is equivalent to linearly re-scaling the time units, smaller K implies faster asymptotic convergence assuming full parallelization capabilities (i.e. the minibatch size does not affect the iterate time). Additionally, note that the corollary only guarantees the same asymptotic behavior. Particularly, different initializations and datasets can exhibit different behavior initially. It remains an interesting direction for future work to understand $\rho(t)$ dependence on η and B, in the case when the support vectors do not span the dataset.

Lastly, for logistic regression loss, the validation loss (calculated on an independent validation set $\mathcal{V}$) increases as

$$\mathcal{L}_{\text{val}}\left(\mathbf{w}\,(t)\right) = \sum_{\mathbf{x}\in\mathcal{V}} \ell\left(\mathbf{w}\,(t)^\top \mathbf{x}\right) = \Omega(\log(t)).$$

Notably, as was observed in Soudry et al. (2018b), these asymptotic rates also match what we observe numerically for the convnet in Figure 1: the training loss decreases as $1/t$, the validation loss increases as $\log(t)$, and the validation (classification) improves very slowly, similarly to the logarithmic decay of the angle gap (so the convnet might have a similarly slow decay to its respective implicit bias).

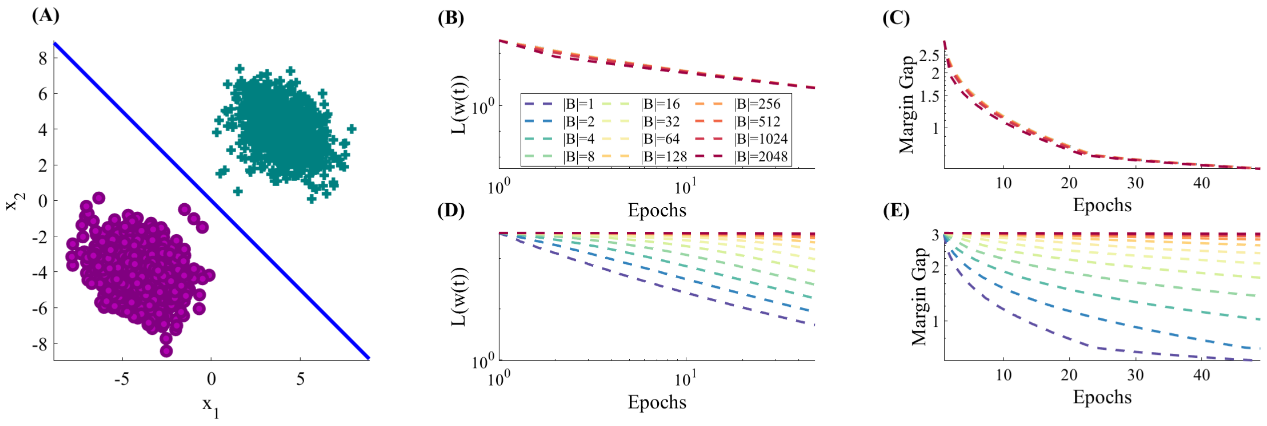

Figure 3: **We observe the convergence rate of SGD remains almost exactly the same for all minibatch sizes when the learning rate is proportional to the minibatch size** ($\eta = \frac{2\gamma^2}{\beta\sigma^2_{\max}}B$ in panels **B** and **C**, vs. $\eta = \frac{2\gamma^2}{\beta\sigma^2_{\max}}$ in panels **D** and **E**). We initialized $\mathbf{w}(0)$ to be a standard normal vector. We used a dataset **(A)** with $N = 2048$ samples divided into two classes, and with the same support vectors as in Figure 2. The convergence of the loss **(B)** and margin **(C)** is practically identical for all minibatch sizes. When we used a fixed learning rate, the convergence rate was different **(D-E)**.

5 DISCUSSION AND RELATED WORKS

In Theorem 1 we proved that for monotone smooth loss functions on linearly separable data, the iterates of SGD with a sufficiently small (but non-vanishing) learning rate converge to zero loss. In contrast to typical convergence to finite critical points, in this case, the "noise" inherent in SGD vanishes asymptotically. Therefore, we do not need to decrease the learning rate, or average the SGD iterates, to ensure exact convergence. Decaying the learning rate during training will only decrease the convergence speed of the loss.

To the best of our knowledge, such exact convergence result previously required that either (1) the loss function is partially strongly convex, i.e. strongly convex except on some subspace (where the dynamics are frozen), as shown in (Ma et al., 2017) for the case over-parameterized linear regression (with more parameters then samples); or (2) that the Polyak-Lojasiewicz (PL) condition applies (Bassily et al., 2018). However, in this paper we do not require such conditions, which does not hold for deep networks, even in the vicinity of the (finite or infinite) critical points. Moreover, the dependence of the learning rate on the minibatch size is different, as we discuss next.

We proved Theorem 1 both for random sampling with replacement (Assumption 3a) and for sampling without replacement (Assumption 3b). In the first case, eq. 4 implies that, to guarantee convergence, we need to increase the learning rate proportionally to the minibatch size. In the second case (sampling without replacement) the learning rate bound (eq. 5) is more pessimistic, since our assumption is more general (e.g., it includes adversarial sampling).

In Theorem 2, we proved, given the additional assumption of an exponential tail (e.g., as in logistic regression), that for almost all datasets the weight vector converges to the L_2 max margin in direction as $1/\log(t)$, and that the training loss converges to zero as $1/t$. We believe these results could be extended for every dataset, using the techniques of Soudry et al. (2018a). Again, decaying the learning rate will only degrade the convergence speed to the max margin direction. In fact, the results of Nacson et al. (2018) indicate that we may need to *increase* the learning rate to improve convergence: For GD, Nacson et al. (2018) proved that this can drastically improve the convergence rate from $1/\log(t)$ to $\log(t)/\sqrt{t}$. It is yet to be seen if such results might also be applied to deep networks.

In Theorem 3 we further characterized the weights asymptotic behaviour under the additional assumption that the SVM support vectors span the dataset. Combining the results from Theorem 2 and Theorem 3 we obtain Corollary 1. This corollary states that under, linear scaling of the learning rate with the batch size, the asymptotic convergence rate of SGD, in terms of epochs, is not affected by the mini-batch size.

Thus, we have shown that exact linear scaling of the learning rate with the minibatch size ($\eta \propto B$) is beneficial in two ways: (a) in Theorem 1 for the upper bound of the learning rate in the case of of random sampling with replacement (b) in Corollary 1 for the asymptotic behaviour of the weights assuming tight exponential loss function and that the SVM support vectors span the data. This exact linear scaling, stands in contrast to previous theoretical results with exact convergence (Ma et al., 2017), in which there exists a "saturation limit". Above this limit we should not increase the learning rate linearly with the minibatch size, or the convergence rate will be degraded, and eventually we will loose the convergence guarantee. As predicted by Corollary 1, in Figure 3 we observe that with a linear scaling $\eta \propto B$, the convergence plots exactly match: as we can see, there is almost no asymptotic difference between different minibatch sizes. Therefore, in contrast to Ma et al. (2017), there is no "optimal" minibatch size. In this case, to minimize the number of SGD iterations we should use the largest minibatch possible. This will speed up convergence in wall clock time (as was done in Goyal et al. (2017); Smith et al. (2018)) if it is possible parallelize the calculation of a minibatch — so one SGD update with a minibatch of size MB takes less time then M updates of SGD with minibatch of size B.

An early version of this manuscript previously appeared on arxiv. However, it had only the results in the case of sampling without replacement, and no Theorem 3. Two other related SGD results appeared on arXiv in parallel (with less than a week difference).

First, Ji and Telgarsky (2018) analyzed logistic regression optimized by SGD on separable data (in addition to other results on GD when the data is non-separable). Ji and Telgarsky (2018) also assume a fixed learning rate, but use averaging of the iterates (which is known to enable exact convergence). They focus on the case in which the datapoints are independently sampled from a separable distribution, while we focused on the case of sampling from a fixed dataset. They show, that with high probability, the population risk converges to zero as $\tilde{O}(1/t)$. As explained in Ji and Telgarsky (2018), such a fast rate was proven before only for strongly convex loss functions (the logistic loss is not strongly convex). We showed a similar rate, but for the empirical risk (eq. 14). We additionally showed that the weight vector converges in direction to the direction of the L_2 max margin.

Second, among other results, Xu et al. (2018) also examined optimizing logistic regression with SGD on a fixed dataset using random sampling with replacement, iterate averaging and a vanishing learning rate. There, in Theorems 3.2 and 3.3, it is shown that the expectation of the loss converges as $\tilde{O}(1/t)$ and the expectation of the averaged iterates converges in the norm as $O(1/\sqrt{\log(t)}$, which is slower than our result. Thus, in contrast to both works Ji and Telgarsky (2018); Xu et al. (2018), we did not assume iterate averaging or decreasing learning rate. Additionally, our new results on sampling with replacement give a linear relationship between the learning rate and the minibatch size, and Corollary 1 shows the affect of the minibatch size on the asymptotic convergence rate.

6 CONCLUSIONS

We found that for logistic regression with no bias on separable data, SGD behaves similarly to GD in terms of the implicit bias and convergence rate. The only difference is the maximum possible learning rate should change proportionally to the minibatch size. It remains to be seen if this also holds for deep networks.

Acknowledgements

The authors are grateful to C. Zeno, and I. Golan for helpful comments on the manuscript. This research was supported by the Israel Science foundation (grant No. 31/1031), and by the Taub foundation. A Titan Xp used for this research was donated by the NVIDIA Corporation. NS was partially supported by NSF awards IIS-1302662 and IIS-1764032.

References

Francis Bach and Eric Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *NIPS*, pages –, 2011.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of SGD in non-convex overparametrized learning. pages 1–7, 2018.

Shai Ben-David and Shai Shalev-Shwartz. *Understanding Machine Learning: From Theory to Algorithms*. 2014.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

Dimitri P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, jul 2011.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. 2016.

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

A. Geary and D.P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No.99CH36304)*, 1(1):907–912, 2001.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Minibatch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization. *Math. Prog.*, 155 (1-2):267–305, 2013.

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

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit Bias of Gradient Descent on Linear Convolutional Networks. In *NIPS*, jun 2018a.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018b.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1729–1739, 2017.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three Factors Influencing Minima in SGD. *arXiv*, pages 1–21, 2017.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300v2*, 2018.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning. 2017.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Nathan Srebro, and Daniel Soudry. Convergence of Gradient Descent on Separable Data. *arXiv*, pages 1–45, 2018.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Ohad Shamir. Without-Replacement Sampling for Stochastic Gradient Methods: Convergence Results and Application to Distributed Optimization. pages 1–36, 2016.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't Decay the Learning Rate, Increase the Batch Size. In *ICLR*, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint: 1710.10345v3*, 2018a.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. *ICLR*, 2018b.

Tengyu Xu, Yi Zhou, Kaiyi Ji, and Yingbin Liang. When Will Gradient Methods Converge to Max-margin Classifier under ReLU Models? *arXiv*, 2018.