ON THE NONCONVEX CONVERGENCE OF SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic gradient descent (SGD) and its variants are the main workhorses for solving large-scale optimization problems with nonconvex objective functions. Although the convergence of SGDs in the (strongly) convex case is well-understood, their convergence for nonconvex functions stands on weak mathematical foundations. Most existing studies on the nonconvex convergence of SGD show the complexity results based on either the minimum of the expected gradient norm or the functional sub-optimality gap (for functions with extra structural property) by searching over the entire range of iterates. Hence the last iterations of SGDs do not necessarily maintain the same complexity guarantee. This paper shows that the ϵ-stationary point exists in the final iterates of SGDs, not just anywhere in the entire range of iterates—A much stronger result than the existing one. Additionally, our analyses allow us to measure the *density of the ϵ-stationary points* in the final iterates of SGD, and we recover the classical $O(\frac{1}{\sqrt{T}})$ asymptotic rate under various existing assumptions on the regularity of the objective function and the bounds on the stochastic gradient.

1 INTRODUCTION

We consider the *empirical risk minimization* (ERM) problem:

$$\min_{x\in\mathbb{R}^d}\left[F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right], \tag{1}$$

where $f_i(x) := \mathbb{E}_{z_i\sim\mathcal{D}_i} l(x; z_i)$ denotes the loss function evaluated on input, z_i sampled from its distribution, $\mathcal{D}_i$. Additionally, let F be nonconvex, lower bounded, with Lipschitz continuous gradient; see Section 3. ERM problems appear frequently in statistical estimation and machine learning, where the parameter, x, is estimated by the SGD updates (Bottou et al., 2018; Xu et al., 2021). For a sequence of iterates, $\{x_t\}_{t\geq 0}$ and a stepsize parameter, $\gamma_t > 0$, SGD updates are of the form:

$$x_{t+1} = x_t - \gamma_t g_t, \tag{2}$$

where g_t is an unbiased estimator of ∇F_t, the gradient of F at x_t; that is, $\mathbb{E}(g_t|x_t) = \nabla F_t$. This approach, as given in (2), selects an index independently and uniformly *with replacement* from the set $[n]$, and processes its corresponding stochastic gradient; this way, the same index can be selected again. However, the existing programming interfaces in ML toolkits such as PyTorch (Pytorch.org, 2019) and TensorFlow (tensorflow.org) use a different approach—*random reshuffling* or *randomness without replacement* (Mishchenko et al., 2020; Gürbüzbalaban et al., 2021). In this case, at each cycle, t, a random permutation σ_t of the set $[n]$ is selected, and one complete run of all indices from σ_t, which guarantees that each function in (1) contributes exactly once. Formally, RR-SGD updates are of the form:

$$x_{(t-1)n+i} = x_{(t-1)n+i-1} - \gamma_t g_{\sigma_t(i)}(x_{(t-1)n+i-1}),\ i = 1, 2, ..., n;\ t = 1, 2, 3, ..., \tag{3}$$

where $g_{\sigma_t(i)}(x_j)$ is the stochastic gradient calculated at x_j. RR-SGD posses faster convergence than regular SGD (Mishchenko et al., 2020; Gürbüzbalaban et al., 2021), leaves less stress on the memory (cf. Section 19.2.1 in Bengio (2012)), and hence more practical.

The convergence of SGD and RR-SGD for the strongly convex functions is mostly well understood (Shalev-Shwartz et al., 2009; Gower et al., 2019; Shamir & Zhang, 2013; Mishchenko et al., 2020), but their convergence of the *last* iterates for nonconvex functions remains an open problem. For

a nonconvex function, F, the existing convergence analyses of SGD show, as $T \to \infty$, either (*i*) the minimum of the norm of the gradient function, $\min_{t\in[T]} \mathbb{E}\|\nabla F_t\| \to 0$ (Ghadimi & Lan, 2013; Khaled & Richtárik, 2020; Stich & Karimireddy, 2020); [1] or (*ii*) the minimum sub-optimality gap, $\min_{t\in[T]} (\mathbb{E}(F(x_t)) - F_\star) \to 0$ (Gower et al., 2021; Lei et al., 2020). Notably, the first-class uses the classical L-smoothness, and the size of the gradient function, $\mathbb{E}\|\nabla F_t\|$ measures the convergence. Whereas the second class considers F to have extra structural property, such as Polyak-Łojasiewicz (PL) condition (Gower et al., 2021; Lei et al., 2020) and the minimum sub-optimality gap is the measure of convergence. Nevertheless, in both cases the notion of ϵ-stationary point [2] is weak as they only consider the minimum of the quantity $\mathbb{E}\|\nabla F_t\|$ or $(\mathbb{E}(F(x_t)) - F_\star)$ approaching to 0 (as $T \to \infty$) by searching over the entire range of iterates, $[T]$. Alongside, by adding one more sampling step at the end, $x_\tau \sim \{x_t\}_{t\in[T]}$, some works show $\mathbb{E}\|\nabla F_\tau\| \to 0$ instead; see (Ghadimi & Lan, 2013; Stich & Karimireddy, 2020; Wang & Srebro, 2019).

In practice, it is common to run SGD for T iterations (that in the order of millions for DNN training) and return the last iterate (Shalev-Shwartz et al., 2011). Therefore, we may ask: How practical is the notion of an ϵ-stationary point? For example, training ResNet-50 (He et al., 2016) on ImageNet dataset (Deng et al., 2009) requires roughly $600,000$ iterations. The present nonconvex convergence analysis of SGD tells us that at one of these $600,000$ iterations, $\mathbb{E}\|\nabla F_t\| \approx 0$. Indeed, this is impractical. That is, the existing results treat all the iterations similarly, do not motivate why we need to keep producing more iterations, but only reveal that as long we are generating more iterations, one of them will be ϵ-stationary point. However, numerical experiments suggest that the final iterates of SGD (and RR-SGD) will for sure contain more ϵ-stationary points. This motivated us to ask: *Can we guarantee the existence of ϵ-stationary point for SGD and RR-SGD for the nonconvex case on the final iterates?* But guaranteeing the final iterates of SGD contain one of the ϵ-stationary points alone does not conclude the task. Thus, we also would like to quantify the denseness of these ϵ-stationary points among the last iterates. In all cases, SGD achieves an $O(\frac{1}{\sqrt{T}})$ asymptotic convergence rate for nonconvex, L-smooth functions which is optimal (Carmon et al., 2020). Therefore, a more refined analysis is required to capture this asymptotic rate.

To answer these questions, we make the following contributions:

(*i*) **Convergence analysis of SGD for nonconvex functions.** By controlling the stepsize parameter (using either constant or decreasing stepsize), we show the existence of ϵ-stationary points in the final iterates of SGD and RR-SGD for nonconvex functions; see Section 4. This is the first result to guarantee that the ϵ-stationary points *exist* in the final iterates of nonconvex SGD, compared to the existing classical convergence results in (Ghadimi & Lan, 2013; Khaled & Richtárik, 2020; Stich & Karimireddy, 2020) or *high probability* convergence results in (Harvey et al., 2019a; Lei et al., 2020). In contrast to the existing works of Shamir & Zhang (2013); Jain et al. (2019), on showing the final iterates of SGD converge for convex and strongly convex cases, we perform our analysis without suffix averaging. Additionally, our techniques can be extended to the convergence of SGD for nonconvex and nonsmooth objective; see A.3.

(*ii*) **Classic asymptotic rate under various assumptions on bounds of the stochastic gradient.** Assumptions on the bounds of the stochastic gradient are an important factor, but we do not judge which assumption is better over the other as the literature has established potentially many interplays between them; see Section 3. The focus of this study is to give a proper mathematical convergence guarantee of SGD and RR-SGD for nonconvex functions. Therefore, our analyses is based on one of the most general assumptions on the bounds of stochastic gradient—the expected smoothness by Khaled & Richtárik (2020). This bound encompasses most commonly used bounds such as, the (M, σ^2)-bounded gradient noise (Stich & Karimireddy, 2020), strong and weak growth condition Vaswani et al. (2019). We recovepr the classic $O(\frac{1}{\sqrt{T}})$ convergence rate of nonconvex SGD under no additional assumptions, where T is the number of iterations. For RR-SGD, the convergence rate is $O(\frac{1}{\sqrt{nT}})$, similar to Mishchenko et al. (2020), where T is the number of epochs.

[1] Some works show, the average of the expected gradient norm, $\frac{1}{T}\sum_t \mathbb{E}\|\nabla F_t\| \to 0$ as $T \to \infty$ for fixed stepsize, or the weighted average of the expected gradient norm, $\frac{1}{\sum_t \gamma_t}\sum_t \gamma_t \mathbb{E}\|\nabla F_t\| \to 0$ as $T \to \infty$ for variable stepsize; see (Bottou et al., 2018).

[2] A stationary point, in general, is either a local minimum, a local maximum, or a saddle point. In nonconvex convergence of SGD, x is an ϵ-stationary point if $\mathbb{E}\|\nabla F(x)\| \leq \epsilon$ or $\mathbb{E}(F(x)) - F_\star) \leq \epsilon$.

(*iii*) **Density of ϵ-stationary points.** An interesting consequence of our convergence analyses is that they allow us to measure the *density of the ϵ-stationary points* in the final iterates of SGD—A first standalone result. That is, we show that the density of the ϵ-stationary points over the tail portion for the SGD iterates, x_t is almost 1 for large T, where $t \in [(1-\eta)T, T]$ and $\eta \in (0, 1]$; see Section 5.

Finally, we support our theoretical results by performing numerical experiments on nonconvex functions, both smooth (logistic regression with nonconvex penalty) and nonsmooth (feed forward neural network with ReLU activation); see Section 6. Our code and results are publicly available; see B.

2 BRIEF LITERATURE REVIEW

The convergence of SGD for convex and strongly convex functions is well understood; see Appendix A for a few related work. Additionally, see Appendix A for the stability and generalization bound of SGD. We start with nonconvex convergence of SGD.

Nonconvex convergence of SGD was first proposed by Ghadimi & Lan (2013) for nonlinear (possibly nonconvex) stochastic programming. Inspired by (Nesterov, 2003; Gratton et al., 2008), Ghadimi & Lan (2013) showed that SGD achieves $\min_{t\in[T]} \mathbb{E}\|\nabla F_t\|^2 \leq \epsilon$ after running for at most $O(\epsilon^{-2})$ steps—same complexity as the gradient descent for solving (1). Ghadimi et al. (2016) extended their results to a class of constrained stochastic composite optimization problems with loss function as the sum of a differentiable (possibly nonconvex) function and a non-differentiable, convex function. Recently, Vaswani et al. (2019) proposed a *strong growth condition (SGC)* of the stochastic gradient, and showed that under SGC with a constant ρ, SGD with a constant step-size can attain the optimal rate, $O(\epsilon^{-1})$ for nonconvex functions; see Theorem 3 which is an improvement over Ghadimi & Lan (2013). Stich & Karimireddy (2020) proposed the *(M, σ^2) noise bound* for stochastic gradients and proposed a convergence analysis for (compressed and/or) error-compensated SGD; see (Stich et al., 2018; Sahu et al., 2021). For nonconvex functions, Stich & Karimireddy (2020) showed $\mathbb{E}\|\nabla F(x_\tau)\| \to 0$, where $x_\tau \sim \{x_t\}_{t\in[T]}$. At about the same time, Khaled & Richtárik (2020) proposed a new assumption, *expected smoothness (ES)*, see Assumption 5, for modelling the second moment of the stochastic gradient and achieved the optimal $O(\epsilon^{-2})$ rate for SGD in finding stationary points for nonconvex L-smooth functions. Among others, Lei et al. (2020) used Holder's continuity on gradients and showed the nonconvex convergence of SGD. Additionally, they showed the loss, F converges almost surely to a random variable. By using mini-batches to control the loss of iterates to non-attracted regions, Fehrman et al. (2020) proved the convergence of SGD to a minimum for *not necessarily* locally convex nor contracting objective functions. The highlight of the above works is that they show nonconvex convergence of SGD with different conditions on the second moment of the stochastic gradient, and use the minimum of the expected gradient norm, $\min_{t\in[T]} \mathbb{E}\|\nabla F_t\| \to 0$, or the average of the expected gradient norm, $\frac{1}{T}\sum_t \mathbb{E}\|\nabla F_t\| \to 0$ as $T \to \infty$, to show this. Additionally, for convergence of proximal stochastic gradient algorithms (with or without variance reduction) for nonconvex, nonsmooth finite-sum problems, see (J Reddi et al., 2016; Li & Li, 2018); for non-convex problems with a non-smooth and non-convex regularizer, see (Xu et al., 2019).

Adaptive gradient methods such as ADAM (Kingma & Ba, 2015), AMSGrad (Reddi et al., 2018), AdaGrad (Duchi et al., 2011) are extensively used for DNN training. Although the nonconvex convergence of these algorithms are more involved than SGD, they focus on the same quantities as SGD to show convergence. See nonconvex convergence of ADAM and AdaGrad (with or without momentum) in (Défossez et al., 2020), nonconvex convergence for AdaGrad in (Ward et al., 2019), Theorem 2.1; also, see Theorem 3 in (Zhou et al., 2020), and Theorem 2 in (Yang et al., 2016) for an unified analysis of stochastic momentum methods for nonconvex functions. Recently, Jin et al. (2022) proved almost sure asymptotic convergence of momentum SGD and ADAGRAD.

Compressed and distributed SGD is widely studied to remedy the network bottleneck in bandwidth limited training of large DNN models, such as federated learning (Konečný et al., 2016; Kairouz et al., 2019). The convergence analyses of compressed and distributed SGD for nonconvex loss (Dutta et al., 2020; Xu et al., 2021; Alistarh et al., 2017; Sahu et al., 2021; Stich & Karimireddy, 2020) follow the same structure of the existing nonconvex convergence of SGD.

Structured nonconvex convergence analysis of SGD and similar methods. Gower et al. (2021) used extra structural assumptions on the nonconvex functions and showed SGD converges to a global

minimum. Gorbunov et al. (2021) showed when F is cocoercive (monotonic and L-Lipschitz gradient), the last-iterate for extra gradient (Korpelevich, 1976) and optimistic gradient method (Popov, 1980) converge at $O(\frac{1}{T})$ rate.

3 ASSUMPTIONS

Assumption 1. ***(Smoothness)*** *For every $i \in [n]$, the function, $f_i : \mathbb{R}^d \to \mathbb{R}$, is L-smooth, i.e. $f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$ for all $x, y \in \mathbb{R}^d$.*

Remark 1. The above implies that F is L-smooth.

Assumption 2. ***(Global minimum)*** *There exists $x_\star$ such that, $F(x_\star) = F_\star \leq F(x)$, for all $x \in \mathbb{R}^d$.*

Bound on stochastic gradient. There have been different assumptions to bound the stochastic gradient. One may follow the model of Stich & Karimireddy (2020). Let the stochastic gradient, g_t, at iteration t is of the form, $g_t = \nabla F_t + \xi_t$, with $\mathbb{E}[\xi_t | x_t] = 0$. The above model leads to the following assumption as given by Stich & Karimireddy (2020).

Assumption 3. ***((M, σ^2) bounded noise)*** *There exist constants, $M, \sigma^2 \geq 0$, such that for all $x_t \in \mathbb{R}^d$, the stochastic noise, ξ_t, follows*

$$\mathbb{E}[\|\xi_t\|^2 \mid x_t] \leq M\|\nabla F_t\|^2 + \sigma^2.$$

Remark 2. Assumption 3 implies that $\mathbb{E}[\|g_t\|^2 \mid x_t] \leq (M + 1)\|\nabla F_t\|^2 + \sigma^2$.

Assumption 4. ***((M, σ^2) bounded stochastic gradient)*** *Let F follow Assumption 1. Then there exist constants $M, \sigma^2 \geq 0$, such that for all $x_t \in \mathbb{R}^d$, the stochastic gradient, g_t follows*

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \leq 2L(1 + M)(F_t - F_\star) + \sigma^2.$$

Among different assumptions considered on bounding the stochastic gradient in the literature (see Bottou et al. (2018); Ghadimi & Lan (2013); Stich & Karimireddy (2020); Lei et al. (2020); Gower et al. (2021); Vaswani et al. (2019); Dutta et al. (2020)), recently, Khaled & Richtárik (2020) noted that the expected smoothness is the weakest among them and is as follow. Assumption 5 contains Assumption 3 and 4 as special case.

Assumption 5. ***(expected smoothness)*** *There exist constants, $A, B, C \geq 0$, such that, for all $x_t \in \mathbb{R}^d$ we have*

$$\mathbb{E}[\|g_t\|^2 \mid x_t] \leq 2A(F_t - F_\star) + B\|\nabla F_t\|^2 + C.$$

Finally, we state the bounded variance assumption of gradients from Mishchenko et al. (2020) that the authors used in proving the nonconvex descent lemma of RR-SGD, Lemma 1. For details of how Assumptions 5 and 6 are connected see Mishchenko et al. (2020).

Assumption 6. ***(bounded variance of gradients)*** *There exist constants, $\mathcal{A}, \mathcal{B} \geq 0$, such that, for all $x \in \mathbb{R}^d$, the variance of gradients follow*

$$\frac{1}{n}\sum_{i\in[n]}\|\nabla f_i(x) - \nabla F(x)\|^2 \leq 2\mathcal{A}(F(x) - F_\star) + \mathcal{B}.$$

4 MAIN CONVERGENCE RESULT

We present our main convergence results in this section. Section 4.1 presents the nonconvex convergence results by using the expected smoothness. Also, see Appendix A.3 on how our techniques can be extended to the convergence of SGD for nonconvex and nonsmooth objective. Finally, in Section 4.2, we show the nonconvex convergence of RR-SGD.

4.1 CONVERGENCE OF SGD USING EXPECTED SMOOTHNESS

Khaled & Richtárik (2020) demonstrated that the expected smoothness in Assumption 5 contains Assumption 3 and 4 as special cases. We will use this assumption to derive our main result. To help the reader, we sketch the key steps that lead to our main results. In the nonconvex case of convergence of SGD, by using the L-smoothness of F and *expected smoothness* of the stochastic gradients, we arrive at the following key inequality; see Lemma 2 by Khaled & Richtárik (2020):

$$\gamma_t(1 - \frac{LB\gamma_t}{2})\mathbb{E}\|\nabla F_t\|^2 \leq (1 + L\gamma_t^2 A)(\mathbb{E}(F_t) - F_\star) - (\mathbb{E}(F_{t+1}) - F_\star) + \frac{L\gamma_t^2 C}{2}. \tag{4}$$

Convergence using a fiexed step size $\gamma_t = \gamma$. Denote $r_t = \mathbb{E}\|\nabla F_t\|^2$, $\delta_t = \mathbb{E}(F_t) - F_\star$, $D := (1 + L\gamma^2 A)$, $E := \gamma(1 - \frac{LB\gamma}{2})$, $F := \frac{L\gamma^2 C}{2}$, and rewrite (4) as

$$\delta_{t+1} \quad \leq \quad D\delta_t - Er_t + F, \tag{5}$$

which after unrolling the recurrence becomes

$$\delta_{T+1} \leq D^{T+1}\delta_0 - E\sum_{t=0}^{T} D^{T-t} r_t + F\sum_{t=0}^{T} D^t. \tag{6}$$

Denote $W = \sum_{t=0}^{T} D^t$. Rearranging the terms again and dividing both sides by E we have

$$\sum_{t=0}^{T} D^{T-t} r_t + \frac{\delta_{T+1}}{E} \leq \frac{D^{T+1}}{E}\delta_0 + \frac{FW}{E}. \tag{7}$$

Note that

$$W = \sum_{t=0}^{T} D^t = \frac{(1 + L\gamma^2 A)^{T+1} - 1}{L\gamma^2 A}, \quad \frac{FW}{E} = \frac{C[(1 + L\gamma^2 A)^{T+1} - 1]}{\gamma A(2 - LB\gamma)},$$

and

$$\frac{D^{T+1}}{E} = \frac{2(1 + L\gamma^2 A)^{T+1}}{\gamma(2 - LB\gamma)}.$$

Therefore, (7) can be written as

$$\sum_{t=0}^{T} (1 + L\gamma^2 A)^{T-t} r_t + \frac{2\delta_{T+1}}{\gamma(2 - L\gamma B)} \leq \frac{2(1 + L\gamma^2 A)^{T+1}}{\gamma(2 - LB\gamma)}\delta_0 + \frac{C[(1 + L\gamma^2 A)^{T+1} - 1]}{\gamma A(2 - LB\gamma)}. \tag{8}$$

Let $\eta \in (0, 1]$. Then the left-hand side in the inequality (8) is bounded from below by

$$\min_{(1-\eta)T \leq t \leq T} r_t \sum_{(1-\eta)T \leq t \leq T} (1 + L\gamma^2 A)^{T-t} \geq (\eta T - 1) \min_{(1-\eta)T \leq t \leq T} r_t;$$

if $LB\gamma \leq 1$ and $(1 + L\gamma^2 A)^{T+1} \leq 3$ then the right-hand side of (8) could be bounded from above by

$$\frac{6\delta_0}{\gamma} + \frac{2C}{\gamma A}. \tag{9}$$

Hence, we obtain

$$\min_{(1-\eta)T \leq t \leq T} r_t \leq 2\left(3\delta_0 + \frac{C}{A}\right) \frac{1}{(\eta T - 1)\gamma}. \tag{10}$$

Now, letting $\gamma := \sqrt{\frac{\ln 3}{(T+1)LA}}$, we are able to show the following result; see Appendix A.1 for the proof.

Theorem 1. *Let F follow Assumptions 1, 2, and 5. Let $\epsilon > 0$ and $\eta \in (0, 1]$. If the number of iterations $T \geq 1$ satisfies*

$$T \geq \max\left\{\left(\frac{4\sqrt{2LA}(3\delta_0 + C/A)}{\varepsilon\eta\sqrt{\ln 3}}\right)^2, \frac{LB^2 \ln 3}{A} - 1, \frac{2}{\eta}\right\},$$

then, there exists an index $t \geq (1 - \eta)T$ such that $\mathbb{E}\|\nabla F_t\|^2 \leq \epsilon$.

Remark 3. Let $\epsilon > 0$. For $\eta \to 0$, and $T = \Omega\left(\max\{\frac{1}{\eta}, \frac{1}{\eta^2\epsilon^2}\}\right)$, there exists a $t \in [(1 - \eta)T, T]$, such that, $\mathbb{E}\|\nabla F_t\|^2 \leq \epsilon$. For example, take $\eta = 0.05$ in the Theorem above. Then we know that the last 5% steps in the T iterations will produce at least one ε-stationary point. For $\eta = 1$, we recover the classical asymptotic convergence rate of SGD, that is, $\min_{t \in [T]} \mathbb{E}\|\nabla F_t\|^2 = O\left(\frac{1}{\sqrt{T}}\right)$.

Remark 4. Using the inequality that $1 + x \geq e^{x/2}$ for $x \in [0, 1]$, we can see that

$$(1 + L\gamma^2 A)^{T+1} \geq \sqrt{3},$$

as $\ln 3/(T+1) < 1$ when $T > 1$. So, our choice of γ makes sure that the expression $(1+L\gamma^2 A)^{T+1}$ is contained in an interval $[\sqrt{3}, 3]$ to the right side of 1 on the real line. Indeed, any choice of stepsize such that the expression is contained in an interval on the right side of 1 will work—the only difference will be in the constants in the estimations.

Convergence using a decreasing step size γ_t. We consider stepsize $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ with $\gamma_0 > 0$, and adopt a slightly different technique. Inspired by Stich & Karimireddy (2020), we define a non-negative, decreasing weighting sequence, $\{w_t\}_{t=0}^T$, such that $w_{-1} = 1$ and $w_t := \frac{w_{t-1}}{(1+L\gamma_t^2 A)}$. With these weights, and by using the notations before, we can rewrite (4) as:

$$w_t\gamma_t(1 - \frac{LB\gamma_t}{2})r_t \leq w_t(1 + L\gamma_t^2 A)\delta_t - w_t\delta_{t+1} + w_t\frac{L\gamma_t^2 C}{2}. \tag{11}$$

Taking summation on (11) from $t = 0$ to $t = T$, we have

$$\sum_{t=0}^{T} w_t\gamma_t(1 - \frac{LB\gamma_t}{2})r_t \leq \delta_0 + \frac{LC}{2}\sum_{t=0}^{T} w_t\gamma_t^2. \tag{12}$$

The right hand side of (12) is bounded above by

$$\delta_0 + \frac{LC}{2}\gamma_0^2(\ln(T + 1) + 1). \tag{13}$$

Following the same technique as in the constant stepsize case, the left hand side of (12) is bounded from below by

$$(1 - LA\gamma_0^2 \ln(T + 1)) \min_{(1-\eta)T \leq t \leq T} r_t(\gamma_0(1 - \sqrt{1-\eta})\sqrt{T+1} - \frac{LB\gamma_0^2}{2}\ln(T + 1) + \frac{LB\gamma_0^2}{2}\ln([(1 - \eta)T] + 1)). \tag{14}$$

Combining (13) and (14), we can state the following Theorem; see Appendix A.1 for the proof.

Theorem 2. *Let F follow Assumptions 1, 2, and 5. Let $\eta \in (0, 1]$. By choosing the stepsize $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ with $\gamma_0^2 < \frac{1}{LA\ln(T+1)}$, there exists a step $t \geq (1 - \eta)T$ such that*

$$\mathbb{E}\|\nabla F_t\|^2 \leq \frac{F_0 - F_\star + \frac{LC\gamma_0^2}{2}(\ln(T+1)+1)}{(1-LA\gamma_0^2\ln(T+1))\left(\gamma_0\eta\sqrt{T+1} - \frac{LB\gamma_0^2}{2}\ln(T+1) + \frac{LB\gamma_0^2}{2}\ln([(1-\eta)T]+1)\right)}.$$

Remark 5. For $\eta = 1$, we recover the classical asymptotic convergence rate of SGD, that is, $\min_{t\in[T]} \mathbb{E}\|\nabla F_t\|^2 = O\left(\frac{\ln(T+1)}{\sqrt{T+1}}\right)$. For $\eta \to 0$, there exists a $t \in [(1 - \eta)T, T]$, such that $\mathbb{E}\|\nabla F_t\|^2 = O\left(\frac{\ln(T+1)}{\eta\sqrt{T+1}}\right)$.

4.2 Convergence of RR-SGD

RR-SGD, one of the closest variants of SGD, outperforms SGD in many aspects, as it is significantly faster than SGD in practice (Mishchenko et al., 2020; Gürbüzbalaban et al., 2021). Recently, Mishchenko et al. (2020) showed a better nonconvex convergence of RR-SGD compared to prior work of Nguyen et al. (2021) without the bounded gradient assumption. Mishchenko et al. (2020) followed Assumption 6—bounded variance of gradients. We start by quoting the key descent Lemma used for the convergence of RR-SGD from (Mishchenko et al., 2020). We focus on constant stepsize case, results for decreasing stepsize follow the similar arguments.

Lemma 1. *Let F follow Assumptions 1, 2, and 6, and the update rule in* (3) *is run for T epochs. Then for $\gamma \leq \frac{1}{2Ln}$ and $t \in \{0, 1, \cdots T - 1\}$, the iterates of* (3) *satisfy*

$$(\mathbb{E}(F_{t+1}) - F_\star) \leq (1 + \mathcal{A}L^2n^2\gamma^3)(\mathbb{E}(F_t) - F_\star) - \frac{\gamma n}{2}(1 - \gamma^2L^2n^2)\mathbb{E}\|\nabla F_t\|^2 + \frac{L^2\gamma^3n^2\mathcal{B}}{2}, \tag{15}$$

where T denotes the total number of epochs.

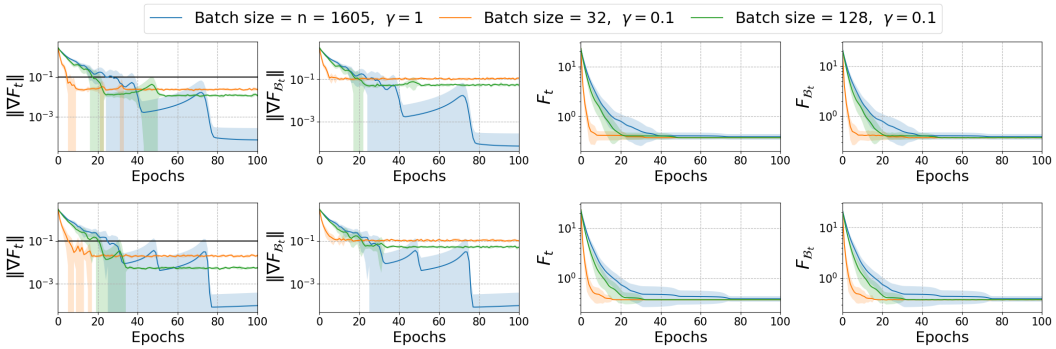

Figure 1: Average of 10 runs of SGD (top row) and RR−SGD (bottom row) on logistic regression with nonconvex regularization. Batch size $n = 1605$ represents full batch.

Proceeding similarly as before, and letting $\gamma_t = \gamma := \left(\frac{\ln 3}{(T+1)AL^2n^2}\right)^{\frac{1}{3}}$, we can show the following result. See the detailed derivation in Appendix A.2.

Theorem 3. *Let F follow Assumptions 1, 2, and 6, and the update rule in* (3) *is run for T epochs. Let $\epsilon > 0$ and $\eta \in (0, 1]$. If the number of epochs $T > 1$ satisfies*

$$T \geq \max\left\{27\left(3\delta_0 + \frac{\mathcal{B}}{\mathcal{A}}\right)^3 \frac{^2}{n\eta^2\varepsilon^2}, \frac{8Ln\ln 3}{\mathcal{A}} - 1, \frac{2}{\eta}\right\},$$

then, there exists an index, $t \geq (1-\eta)T$, such that $\mathbb{E}\|\nabla F_t\|^2 \leq \epsilon$.

5 Concentration of the ϵ-stationary points

In nonconvex convergence of SGD, x is an ϵ-stationary point if the expected gradient norm, $\mathbb{E}\|\nabla F(x)\| \leq \epsilon$ or the functional sub-optimality gap (for functions with extra structural property), $\mathbb{E}(F(x)) - F_\star) \leq \epsilon$. The existing studies show that either of these quantities approaches 0 by searching over the entire range of iterates; hence claim an ϵ-stationary point exists. Nevertheless, this does not confirm that the last iterations of SGDs would maintain the same complexity guarantee. In modern machine learning applications, generally, the last iterates of the SGD are the most significant; see (Shalev-Shwartz et al., 2011). In this section, we argue that by using our analyses we can close the existing gap between theory and practice. An interesting consequence of our convergence analyses is that they allow us to measure the *density of the ϵ-stationary points* in the final iterates of SGD without any additional assumptions. This result strengthens our claim in Theorem 1 and corroborates with the practical aspect of SGD.

Denote the set of indices of ϵ-stationary points, $S_\epsilon := \{t : r_t \leq \epsilon\}$. For $\eta \in (0, 1]$, let $S_{\epsilon,\eta} = S_\epsilon \cap [(1-\eta)t, T]$.

Constant step-size. We know $S_{\epsilon,\eta} \neq \emptyset$ by Theorem 1. On one hand, we have

$$\sum_{t=(1-\eta)T}^{T}(1+L\gamma^2A)^{T-t}r_t > \sum_{\substack{t\in S_\epsilon^c \\ t\geq(1-\eta)T}}(1+L\gamma^2A)^{T-t}r_t > \sum_{t=(1-\eta)T+|S_{\epsilon,\eta}|}^{T}(1+L\gamma^2A)^{T-t}\epsilon \geq \frac{(1+L\gamma^2A)^{\eta T-|S_{\epsilon,\eta}|}-1}{L\gamma^2A}\epsilon, \quad (16)$$

where $|S_{\epsilon,\eta}|$ denotes the cardinality of the set $S_{\epsilon,\eta}$. Note that, $\sum_{\substack{t\in S_\epsilon^c \\ t\geq(1-\eta)T}}(1+L\gamma^2A)^{T-t}r_t$ has $(\eta T - |S_{\epsilon,\eta}| + 1)$ terms; so, we lower bound them with smallest of those many terms. On the other hand, using (8) and (9) to bound the left hand side of (16), and rearranging the terms we obtain

$$(1+L\gamma^2A)^{\eta T-|S_{\epsilon,\eta}|} \leq \frac{6\delta_0 L\gamma A + 2CL\gamma}{\epsilon} + 1,$$

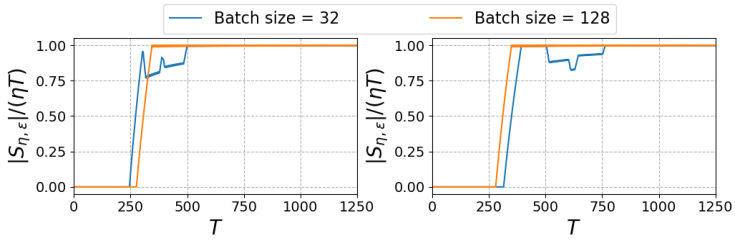

Figure 2: Concentration of the ϵ-stationary points, $\frac{|S_{\epsilon,\eta}|}{\eta T}$ vs. Iterations for nonconvex logistic regression problem running SGD (left) and RR-SGD (right). We set $\epsilon = 10^{-2}$ and $\eta = 0.2$.

Taking logarithm to the previous inequality and rearranging the terms we get

$$\frac{|S_{\epsilon,\eta}|}{\eta T} \geq 1 - \frac{1}{\eta T} \frac{\ln\left(\frac{6\delta_0 L\gamma A + 2CL\gamma}{\epsilon} + 1\right)}{\ln\left(1 + L\gamma^2 A\right)}. \tag{17}$$

We see that the density of the ϵ-stationary points in the top η portion of the tails approaches to 1 as T increases, which roughly speaking, tells us that almost all the iterations x_t for $t \in [(1-\eta)T, T]$ are ϵ-stationary points.

Recall, from Theorem 1, we know that if the total number of iterations, T be such that $T \geq \max\left\{\left(\frac{4\sqrt{2LA}(3\delta_0 + C/A)}{\varepsilon\eta\sqrt{\ln 3}}\right)^2, \frac{LB^2 \ln 3}{A} - 1, \frac{2}{\eta}\right\}$, then there will be iterate, $t \in [(1-\eta)T, T]$ to produce $\mathbb{E}\|\nabla F_t\|^2 \leq \epsilon$, where $\eta \in (0, 1]$. That is, Theorem 1 guarantees the existence of (at least one) stationary point(s) in the final iterates, which is indeed an improvement over the existing results; see Remark 5. However, we may also notice that the stationary point(s) exists in the final iterates, which requires that the SGD is run for a sufficiently large number of iterations. Whereas the claim $\frac{|S_{\epsilon,\eta}|}{\eta T} \to 1$ as $T \to \infty$, says that running SGD for a large number of iterations, T is not necessarily problematic, as one can now surely know that the density of the stationary points in the tail portion will approach to 1, guaranteeing the entire tail comprising mostly of ϵ-stationary points.

Decreasing step-size. In this case, similarly, by Theorem 2, we have $S_{\epsilon,\eta} \neq \emptyset$. For T large enough, we can lower bound the left side of the inequality (12) as

$$\begin{aligned}
&\sum_{t=0}^{T} w_t \gamma_t (1 - \frac{LB\gamma_t}{2}) r_t \geq \epsilon w_T \sum_{\substack{t \in S_\epsilon^c \\ t \geq (1-\eta)T}} \left(\gamma_t - \frac{LB\gamma_t^2}{2}\right) \\
\geq\ & \epsilon(1 + L\gamma_0^2 A(T+1)) \left(\gamma_0\sqrt{T+1} - \gamma_0\sqrt{(1-\eta)T + |S_{\epsilon,\eta}|} - \tfrac{LB\gamma_0^2}{2}\ln(T+1)\right).
\end{aligned}$$

The above, combined with the upper bound in (13) can be written as

$$\gamma_0(\sqrt{T+1} - \sqrt{(1-\eta)T + |S_{\epsilon,\eta}|}) \leq \underbrace{\frac{\delta_0 + \frac{LC}{2}\gamma_0^2(\ln(T+1)+1)}{(1 + L\gamma_0^2 A(T+1))} + \frac{LB\gamma_0^2}{2}\ln(T+1)}_{:=\mathcal{D}}, \tag{18}$$

which can be further reduced to

$$\frac{|S_{\epsilon,\eta}|}{\eta T} \geq 1 - 2\frac{\mathcal{D}}{\gamma_0\eta\sqrt{T}} + \frac{\mathcal{D}^2}{\gamma_0^2\eta T}. \tag{19}$$

Similar to the argument for constant stepsize case, we conclude that the density of the ϵ-stationary points in the top η portion of the tail approaches to 1 as T increases.

6 Numerical Evidence

We conduct experiments on nonconvex functions with L-smooth and non-smooth (for DNNs) loss to substantiate our theoretical results that are based on stochastic gradient, g_t. In practice, g_t can be calculated by sampling and processing minibatches of data. Therefore, besides $\|\nabla F_t\|$ and F_t, we

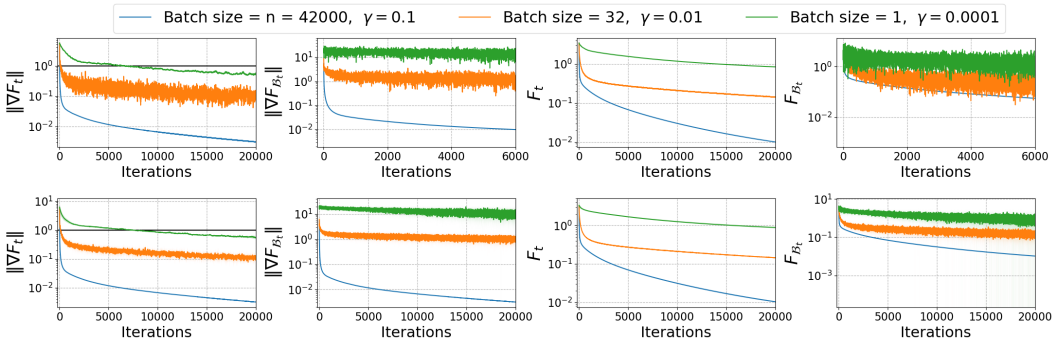

Figure 3: Performance of SGD on MNIST digit classification. The top row shows the result of 1 single run of SGD while the bottom row shows the result of the average of 10 runs.

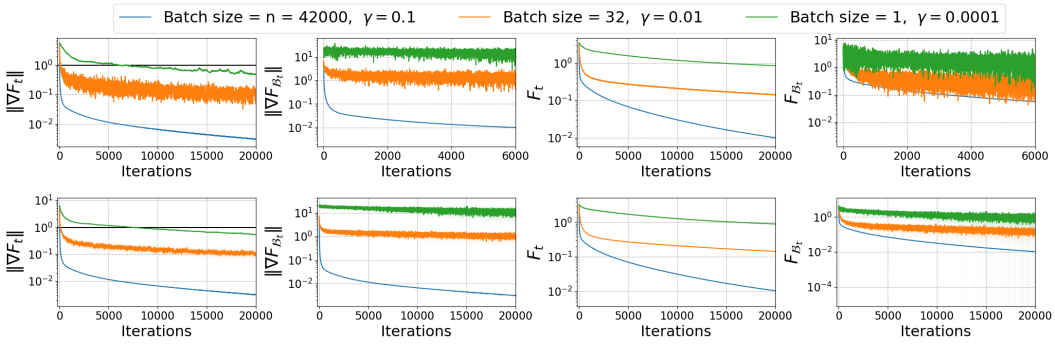

Figure 4: Performance of RR−SGD on MNIST digit classification. The top row shows the result of 1 single run of RR−SGD while the bottom row shows the result of the average of 10 runs.

also track, the norm of the minibatch stochastic gradient, $\|\nabla F_{\mathcal{B}_t}\|$, and minibatch stochastic loss, $F_{\mathcal{B}_t}$. Note that, $\mathcal{B}_t$ is the selected minibatch of data at iteration t and $F_{\mathcal{B}_t} := \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} f_i(x_t)$.

Nonconvex and L-smooth loss. We consider logistic regression with nonconvex regularization:

$$\min_{x\in\mathbb{R}^d} F(x) := \frac{1}{n}\sum_{i=1}^{n} \ln(1+\exp(-a_i^\top x)) + \lambda \sum_{j=1}^{d} \frac{x_j^2}{1+x_j^2},$$

where $a_1, a_2, ..., a_n \in \mathbb{R}^d$ are the given data, and $\lambda > 0$ is the regularization parameter. We run the experiments on the a1a dataset from LIBSVM (Chang & Lin, 2011), where $n = 1605, d = 123$, and set $\lambda = 0.5$. Figures 1 shows the average of 10 runs of SGD and RR−SGD, respectively, with different minibatch sizes. The shaded area is given by $\pm\sigma$ where σ is the standard deviation. For the plots in the first column, the horizontal lines correspond to the precision, $\epsilon = 10^{-1}$, and conform our theoretical results—If the total number of iterations is large enough then the density of the stationary points in the final iterates is 1, guaranteeing the entire tail comprising of the ϵ-stationary points.

Concentration of ϵ-stationary points. For $\epsilon = 10^{-2}$ and $\eta = 0.2$, in Figure 2, we plot the density of the ϵ-stationary points, $\frac{|S_{\epsilon,\eta}|}{\eta T}$ as a function of iteration, T for nonconvex logistic regression problems. As T increases, $\frac{|S_{\epsilon,\eta}|}{\eta T} \to 1$ from below, and conform our theoretical result in Section 5.

Nonconvex and nonsmooth loss. We use a feed forward neural network (FNN) for MNIST digit (LeCun et al., 1998) classification. The FNN has one hidden layer with 256 neurons activated by ReLU activation, and an 10 dimensional output layer activated by the softmax function. The loss function is the categorical cross entropy. We calculate the loss and the stochastic gradient during the training by using different minibatches. The entire loss and the full gradient are computed using all $n = 42 \times 10^3$ samples. For the average of 10 runs, the shaded area is given by $\pm\sigma$, where $\sigma \geq 0$ is the standard deviation, and γ is the learning rate. In Figures 3 and 4 the plots in the first column, the horizontal lines correspond to the precision, $\epsilon = 1$—For SGD and RR-SGD, if the total number of iterations is large enough then the entire tail comprising of the ϵ-stationary points.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Proc. of NeurIPS*, pp. 1709–1720, 2017.

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

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake E. Woodworth. Lower bounds for non-convex stochastic optimization. *CoRR*, abs/1912.02365, 2019. URL `http://arxiv.org/abs/1912.02365`.

Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pp. 437–478. 2012.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

C-C Chang and C-J Lin. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), 2011.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34, 2021.

Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International conference on machine learning*, pp. 2332–2341. PMLR, 2015.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of ADAM and AdaGrad. *arXiv preprint arXiv:2003.02395*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of CVPR*, pp. 248–255, 2009.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Aritra Dutta, El Houcine Bergou, Ahmed M. Abdelmoniem, Chen-Yu Ho, Atal Narayan Sahu, Marco Canini, and Panos Kalnis. On the Discrepancy between the Theoretical Analysis and Practical Implementations of Compressed Communication for Distributed Deep Learning. In *Proc. of AAAI*, volume 34, pp. 3817–3824, 2020.

Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21:136, 2020.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2): 267–305, 2016.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: $o(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. *25th International Conference on Artificial Intelligence and Statistics*, 151, 2021.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation . In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 1315–1323, 2021.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5200–5209, 2019.

Serge Gratton, Annick Sartenaer, and Philippe L Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1):414–444, 2008.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo A Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pp. 1579–1613, 2019a.

Nicholas J. A. Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. *Advances in neural information processing systems*, 29, 2016.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pp. 1752–1755, 2019.

Ruinan Jin, Yu Xing, and Xingkang He. On the convergence of msgd and adagrad for stochastic optimization. In *International Conference on Learning Representations*, 2022.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Bennis, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *Proc. of NeurIPS Workshop on Private Multi-Party Machine Learning*, 2016.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819. PMLR, 2020.

Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4394–4400, 2020.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

Konstantin Mishchenko, Ahmed Khaled Ragab Bayoumi, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33, 2020.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Lam M Nguyen, Quoc Tran-Dinh, Dzung T Phan, Phuong Ha Nguyen, and Marten van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Pytorch.org. PyTorch, 2019. URL `https://pytorch.org/`.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *ICLR*, 2018.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pp. 506–514, 1978.

Atal Sahu, Aritra Dutta, Ahmed M. Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. Rethinking gradient sparsification as total error minimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8133–8146, 2021.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, pp. 5, 2009.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 71–79, 2013.

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21: 1–36, 2020.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Proc. of NeurIPS*, pp. 4447–4458, 2018.

tensorflow.org. TensorFlow. `https://tensorflow.org/`.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204. PMLR, 2019.

Weiran Wang and Nathan Srebro. Stochastic nonconvex optimization with large minibatches. In *Algorithmic Learning Theory*, pp. 857–882. PMLR, 2019.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686. PMLR, 2019.

Hang Xu, Chen-Yu Ho, Ahmed M Abdelmoniem, Aritra Dutta, Konstantinos Karatsenidis El Houcine Bergou, Marco Canini, and Panos Kalnis. GRACE: A Compressed Communication Framework for Distributed Machine Learning. In *Proc. of ICDCS*, 2021.

Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *OPT2020: 12th Annual Workshop on Optimization for Machine Learning*, 2020.