SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation

Robert M. Gower*
LTCI, Télécom Paris,
Institut Polytechnique de Paris

Othmane Sebbouh
ENS Paris,
CREST-ENSAE, CNRS

Nicolas Loizou*
Mila and DIRO,
Université de Montréal

Abstract

Stochastic Gradient Descent (SGD) is being used routinely for optimizing non-convex functions. Yet, the standard convergence theory for SGD in the smooth non-convex setting gives a slow sublinear convergence to a stationary point. In this work, we provide several convergence theorems for SGD showing convergence to a global minimum for non-convex problems satisfying some extra structural assumptions. In particular, we focus on two large classes of structured non-convex functions: (i) Quasar (Strongly) Convex functions (a generalization of convex functions) and (ii) functions satisfying the Polyak-Lojasiewicz condition (a generalization of strongly-convex functions). Our analysis relies on an *Expected Residual* condition which we show is a strictly weaker assumption than previously used growth conditions, expected smoothness or bounded variance assumptions. We provide theoretical guarantees for the convergence of SGD for different step-size selections including constant, decreasing and the recently proposed stochastic Polyak step-size. In addition, all of our analysis holds for the arbitrary sampling paradigm, and as such, we give insights into the complexity of minibatching and determine an optimal mini-batch size. Finally, we show that for models that interpolate the training data, we can dispense of our Expected Residual condition and give state-of-the-art results in this setting.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. *Equal contribution.

1 INTRODUCTION

We consider the unconstrained finite-sum optimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right]. \qquad (1)$$

We use $\mathcal{X}^*\subset\mathbb{R}^d$ to denote the set of minimizers x^* of (1) and assume that $\mathcal{X}^*$ is not empty and that $f(x)$ is lower bounded. This problem is prevalent in machine learning tasks where x corresponds to the model parameters, $f_i(x)$ represents the loss on the training point i and the aim is to minimize the average loss $f(x)$ across training points.

When n is large, stochastic gradient descent (SGD) and its variants are the preferred methods for solving (1) mainly because of their cheap per iteration cost. The standard convergence theory for SGD (Robbins and Monro, 1951; Nemirovski and Yudin, 1978, 1983; Shalev-Shwartz et al., 2007; Nemirovski et al., 2009; Arjevani et al., 2019; Hardt et al., 2016) in the smooth nonconvex setting shows slow sub-linear convergence to a stationary point. Yet in contrast, when applying SGD to many practical nonconvex problems of the form (1) such as matrix completion (Sa et al., 2015), deep learning (Ma et al., 2018), and phase retrieval (Tan and Vershynin, 2019) the iterates converge globally, and sometimes, even linearly. This is because these problems often have additional structure and properties, such as all local minimas are global minimas (Sa et al., 2015; Kawaguchi, 2016), the model interpolates the data (Ma et al., 2018) or the function under study is unimodal on all lines through a minimizer (Hinder et al., 2019). By exploiting these structures and properties one can prove significantly tighter convergence bounds.

Here we present a general analysis of SGD for two large classes of structured nonconvex functions: (i) the Quasar (Strongly) Convex functions and (ii) functions satisfying the Polyak-Lojasiewicz (PL) condition. In all of our results we provide convergence guarantees for SGD to the *global minimum.* We also develop several corollaries for functions that interpolate the data.

1.1 Background and Main Contributions

Classes of structured nonconvex functions. The last few years has seen an increased interest in exploiting additional structure prevalent in large classes of nonconvex functions. Such conditions include error bound properties (Fabian et al., 2010), essential strong convexity (Liu et al., 2014), quasi strong convexity (Necoara et al., 2018; Gower et al., 2019), the restricted secant inequality (Zhang and Yin, 2013), and the quadratic growth (QG) condition (Anitescu, 2000; Loizou, 2019). We focus on two of the weakest conditions: the quasar (strongly) convex functions (Hinder et al., 2019; Hardt et al., 2018; Guminov and Gasnikov, 2017) and functions satisfying the PL condition (Polyak, 1987; Lojasiewicz, 1963; Karimi et al., 2016). The class of quasar-convex functions include all convex functions as a special case, but it also includes several nonconvex functions. Recently there is also some evidences suggesting that the loss function of neural networks have a quasar-convexity structure (Zhou et al., 2019; Kleinberg et al., 2018).

Contributions. We show that SGD converges at a $\mathcal{O}(1/\sqrt{k})$ rate on the *quasar-convex functions* and prove linear convergence to a neighborhood for PL functions without any bounded variance assumption or growth assumptions on the stochastic gradients. Instead, we rely on the recently introduced *expected residual* (ER) condition (Gower et al., 2018).

Assumptions on the gradient. The standard convergence analysis of SGD in the nonconvex setting relies on the bounded gradients assumption $\mathbb{E}_i\|\nabla f_i(x^k)\|^2 < c$ (Recht et al., 2011; Hazan and Kale, 2014; Rakhlin et al., 2012) or a growth condition $\mathbb{E}_i\|\nabla f_i(x^k)\|^2 \leq c_1 + c_2\mathbb{E}\|\nabla f(x^k)\|^2$ (Bertsekas and Tsitsiklis, 1996; Bottou et al., 2018; Schmidt et al., 2017). There is now a line of recent works (Nguyen et al., 2018; Vaswani et al., 2019a; Gower et al., 2019; Khaled and Richtarik, 2020; Lei et al., 2019; Koloskova et al., 2020; Loizou et al., 2020) which aims at relaxing these assumptions.

Contributions. We use the recently introduced Expected Residual (ER) condition (Gower et al., 2018). We give the first convergence proofs for SGD under the ER condition and we show that ER is a strictly weaker assumption than the Strong Growth Condition (SGC) (Schmidt and Roux, 2013), Weak Growth (WGC) (Vaswani et al., 2019a) or the Expected Smoothness (ES) (Gower et al., 2019) assumptions. Furthermore, we show that the ER condition holds for a large class of nonconvex functions including 1) smooth and interpolated functions 2) smooth and x^*- *convex* functions[1]. Not only does the ER assumption hold for a larger class of functions, our resulting convergence rates under ER either match or exceed the state-of-the-art for quasar-convex and PL functions.

PL condition. The PL condition (Polyak, 1987; Lojasiewicz, 1963) was introduced as a sufficient condition for the linear convergence of Gradient Descent for nonconvex functions. Assuming bounded gradients, it was shown in Karimi et al. (2016) that SGD with a decreasing step size converges sublinearly at a rate of $\mathcal{O}(1/\sqrt{k})$ for PL functions. In contrast, by using a step size which depends on the total number of iterations, the same convergence rate can be achieved without the need for the bounded gradient assumption (Khaled and Richtarik, 2020). Assuming in addition the interpolation condition and SGC Vaswani et al. (2019a) showed that SGD converges linearly for PL functions, but the specialization of this last result to gradient descent results in a suboptimal dependence on the condition number[2] of the function.

Contributions. We provide a complete minibatch analysis of SGD for PL functions which recovers the best known dependence on the condition number for Gradient Descent (Karimi et al., 2016) while also matching the current state-of-the-art rate derived in Vaswani et al. (2019a); Lei et al. (2019) for SGD for interpolated functions. All of which relies on the weaker ER condition. Moreover, we propose a switching step size scheme similar to Gower et al. (2019) which does not require knowledge of the last iterate of the algorithm. Using this step size, we prove that SGD converges sublinearly at a rate of $\mathcal{O}(1/k)$ for PL functions without any additional bounded gradient of bounded variance assumption or growth assumption.

Step-size selection for SGD. The most important parameter that one should select to guarantee the convergence of SGD is the step-size or learning rate. There are several choices that one can use including constant step-size (Moulines and Bach, 2011; Needell et al., 2016; Gower et al., 2019; Needell and Ward, 2017; Nguyen et al., 2018), decreasing step-size (Robbins and Monro, 1951; Ghadimi and Lan, 2013; Gower et al., 2019; Nemirovski et al., 2009; Karimi et al., 2016) and adaptive step-size Duchi et al. (2011); Liu et al. (2019); Kingma and Ba (2015); Bengio (2015); Vaswani et al. (2019b); Ward et al. (2019).

Contributions. We provide convergence theorems for SGD under several step-size rules for minimizing quasar-convex functions and functions satisfying the PL condi-

[1]The x^*- convexity includes all convex functions and several nonconvex functions.

[2]Theorem 4 in Vaswani et al. (2019a) specialized to GD gives a rate of μ^2/L^2 where L is the smoothness constant and μ the PL constant.

tion, including constant and decreasing step-sizes and a recently introduced adaptive learning rate called the stochastic Polyak step-size (Loizou et al., 2020).

Over-parameterized models and Interpolation. Recently it was shown that SGD converges considerably faster when the underlying model is sufficiently over-parameterized as to interpolate the data. This includes problems such as deep matrix factorization (Rolinek and Martius, 2018; Rahimi and Recht, 2017), binary classification using kernels (Loizou et al., 2020), consistent linear systems (Gower and Richtárik, 2015; Richtárik and Takác, 2020; Loizou and Richtárik, 2017, 2019) and multi-class classification using deep networks (Vaswani et al., 2019a; Loizou et al., 2020).

Contributions. As a corollary of our main theorems we show that for models that interpolate the training data, we can further relax our assumptions, dispense of the ER condition altogether and instead, simply assume that each f_i is smooth. Our results here match the state-of-the-art convergence results (Vaswani et al., 2019a) but again under strictly weaker assumptions.

1.2 SGD and Arbitrary Sampling

We assume we are given access to unbiased estimates $g(x) \in \mathbb{R}^d$ of the gradient such that $\mathbb{E}\left[g(x)\right] = \nabla f(x)$. For example, we can use a minibatch to form an estimate of the gradient such as $g(x) = \frac{1}{b}\sum_{i\in B} \nabla f_i(x)$, where $B \subset \{1,\ldots,n\}$ will be chosen uniformly at random and $|B| = b$. To allow for any form of minibatching we use the *arbitrary sampling* notation

$$g(x) = \nabla f_v(x) := \frac{1}{n}\sum_{i=1}^{n} v_i \nabla f_i(x), \tag{2}$$

where $v \in \mathbb{R}^n_+$ is a random *sampling vector* such that $\mathbb{E}\left[v_i\right] = 1$, for $i = 1,\ldots,n$ and $f_v(x) := \frac{1}{n}\sum_{i=1}^{n} v_i f_i(x)$. Note that it follows immediately from this definition of sampling vector that $\mathbb{E}\left[g(x)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[v_i\right]\nabla f_i(x) = \nabla f(x)$. In this work we mostly focus on the b–minibatch sampling, however we highlight that our analysis holds for every form of minibatching.

Definition 1.1 (Minibatch sampling)**.** Let $b \in [n]$. We say that $v \in \mathbb{R}^n$ is a b–minibatch sampling if for every subset $S \in [n]$ with $|S| = b$ we have that

$$\mathbb{P}\left[v = \frac{n}{b}\sum_{i\in S} e_i\right] = 1/\binom{n}{b} := \frac{b!(n-b)!}{n!}$$

By using a double counting argument you can show that if v is a b–minibatch sampling, it is also a valid sampling vector ($\mathbb{E}\left[v_i\right] = 1$) (Gower et al., 2019). See Gower et al. (2019) for other choices of sampling vectors v.

With an unbiased estimate of the gradient $g(x)$, we can now use Stochastic gradient descent (SGD) to solve (1) by sampling $g(x^k)$ i.i.d and iterating

$$\boxed{x^{k+1} = x^k - \gamma^k g(x^k)} \tag{3}$$

We also make the following mild assumption on the gradient noise.

Assumption 1.2. The *gradient noise* σ^2 is finite

$$\sigma^2 := \sup_{x^*\in\mathcal{X}^*} \mathbb{E}\left[\left\|g(x^*)\right\|^2\right] < \infty.$$

2 CLASSES OF STRUCTURED NONCONVEX FUNCTIONS

We work with two classes of nonconvex problems: the quasar-convex functions and the functions that satisfy the Polyak-Lojasiewicz (PL) condition.

Definition 2.1 (Quasar convex)**.** Let $\zeta \in (0, 1]$ and $x^* \in \mathcal{X}^*$. We say that f is ζ- quasar-convex with respect to x^* if for all $x \in \mathbb{R}^n$,

$$f(x^*) \geq f(x) + \frac{1}{\zeta}\langle \nabla f(x), x^* - x\rangle. \tag{4}$$

For shorthand we write $f \in QC(\zeta)$ to mean (4). The class of quasar-convex functions are parameterized by a positive constant $\zeta \in (0, 1]$. In the case that $\zeta = 1$ then (4) is known as star convexity (Nesterov and Polyak, 2006) (generalization of convexity). One can think of ζ as the value that controls the non-convexity of the function. As ζ becomes smaller the function becomes "more nonconvex" (Hinder et al., 2019).

One of weakest possible assumptions that guarantee a global convergence of gradient descent to the global minimum is the PL condition (Karimi et al., 2016). Indeed, all local minimas of a function satisfying the PL condition are also global minimas.

Definition 2.2 (Polyak-Lojasiewicz (PL) Condition)**.** There exists $\mu > 0$ such that

$$\|\nabla f(x)\|^2 \geq 2\mu\left[f(x) - f^*\right] \tag{5}$$

We write $f \in PL(\mu)$ if function f satisfies (5).

In addition we will also consider in several corollaries the following interpolation condition.

Assumption 2.3. We say that the interpolation condition holds if there exists $x^* \in \mathcal{X}^*$ such that

$$\min_{x\in\mathbb{R}^n} f_i(x) = f_i(x^*) \quad \text{for} \quad i = 1,\ldots,n. \tag{6}$$

This interpolation condition has drawn much attention recently because many overparametrized deep neural networks achieve a zero loss over all training data points (Ma et al., 2018) and thus satisfy (6).

3 EXPECTED RESIDUAL (ER)

In all of our analysis of SGD we rely on the *Expected Residual (ER)* assumption. In this section we formally define ER, provide new sufficient conditions for it to hold and relate it to the existing gradient assumptions.

ER measures how far the gradient estimate $g(x)$ is from the true gradient in the following sense.

Assumption 3.1 (Expected residual)**.** We say that the ER condition holds or $g \in \mathrm{ER}(\rho)$ if

$$\mathbb{E}\left[\|g(x) - g(x^*) - (\nabla f(x) - \nabla f(x^*))\|^2\right] \leq 2\rho\left(f(x) - f(x^*)\right), \quad \forall x \in \mathbb{R}^d. \tag{ER}$$

Note that ER depends on both how $g(x)$ is sampled and the properties of the $f(x)$ function.

As a direct consequence of Assumption 3.1 we have the following bound on the variance of $g(x)$.

Lemma 3.2. If $g \in \mathrm{ER}(\rho)$ then

$$\mathbb{E}\left[\|g(x)\|^2\right] \leq 4\rho(f(x) - f^*) + \|\nabla f(x)\|^2 + 2\sigma^2. \tag{7}$$

It is this bound on the variance (7) that we use in our proofs and allows us to avoid the stronger bounded gradient or bounded variance assumptions.

Connections to other Assumptions. Let us provide some more familiar sufficient conditions which guarantee that the ER condition holds. In doing so, we will also provide simple and informative bounds on the expected residual constant ρ when using minibatching.

We say that f_i is L_i–smooth if $\forall x, z \in \mathbb{R}^d$ holds that:

$$f_i(z) - f_i(x) \;\leq\; \langle \nabla f_i(x), z - x\rangle + \frac{L_i}{2}\|z - x\|^2. \tag{8}$$

Let $L_{\max} := \max_{i=1,\ldots,n} L_i$. For $x^* \in \mathcal{X}^*$, we say that f_i is x^*–convex if

$$f_i(x^*) - f_i(x) \;\leq\; \langle \nabla f_i(x^*), x^* - x\rangle, \quad \forall x \in \mathbb{R}^d. \tag{9}$$

These two assumptions are sufficient for the $ER(\rho)$ condition to hold and give a useful bound on ρ, as we show in the following proposition.

Proposition 3.3. Let v be a sampling vector. If f_i is L_i–smooth and there exists $x^* \in \mathcal{X}^*$ such that f_i is x^*–convex then $g \in \mathrm{ER}(\rho)$. If in addition v is the b–minibatch sampling then

$$\rho(b) = L_{\max}\frac{n-b}{(n-1)b}, \quad \sigma^2(b) = \frac{1}{b}\frac{n-b}{n-1}\sigma_1^2, \tag{10}$$

where $\sigma_1^2 := \sup_{x^*\in\mathcal{X}^*}\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2$.

The bounds in Proposition 3.3 have been proven before but under the stronger assumption that each f_i is convex[3]. In this work by dropping the requirement that each f_i is convex we are able to consider interesting classes of nonconvex functions.

Indeed, the following theorem establishes that only smoothness and the interpolation condition are sufficient for the ER to hold. Furthermore, we place the ER within a hierarchy of the following assumptions used in analysing SGD for smooth nonconvex functions:

SGC: Strong Growth Condition ($\rho_{SGC} > 0$)

$$\mathbb{E}\left[\|g(x)\|^2\right] \leq \rho_{SGC}\|\nabla f(x)\|^2. \tag{11}$$

WGC: Weak Growth Condition($\rho_{WGC} > 0$)

$$\mathbb{E}\left[\|g(x)\|^2\right] \leq 2\rho_{WGC}(f(x) - f(x^*)). \tag{12}$$

ES: Expected Smoothness ($\mathcal{L} > 0$)

$$\mathbb{E}\left[\|g(x) - g(x^*)\|^2\right] \leq 2\mathcal{L}(f(x) - f(x^*)). \tag{13}$$

Next in Theorem 3.4 we show that the ER condition is (strictly) the weakest condition from the above list.

Theorem 3.4. Let ES, WGC and SGC denote Assumption 2.1 in Gower et al. (2019), Eq (7) and Eq (2) in Vaswani et al. (2019a), respectively. Let L_i and x^*–convex abbreviate (8) and (9), respectively. Then the following hierarchy holds,

$$\begin{tikzcd}
\boxed{SGC + L\text{–smooth}} \arrow[d] & \\
\boxed{WGC} \arrow[d] & \boxed{L_i + \text{Interpolated}} \arrow[d] \\
\boxed{ES} \arrow[d] & \boxed{L_i + x^*\text{–convex}} \arrow[l] \\
\boxed{\mathrm{ER}} \arrow[r] & \boxed{ABC}
\end{tikzcd}$$

where L–smooth is shorthand for function f being L–smooth, and ABC refers to the condition (14) proposed in the concurrent work, Khaled and Richtarik (2020). Finally, there are problems such that ER holds and ES *does not* hold.

[3]See Proposition 3.10 item (iii) in Gower et al. (2019) and Lemma F.3 in Sebbouh et al. (2019).

The important assumptions for analyzing SGD in the nonconvex setting are the ones that are downstream from L_i + Interpolated. This is because there exists a rich class of nonconvex functions that are smooth and satisfy the interpolation condition. In contrast, the WGC is only known to hold for smooth and convex functions satisfying the interpolation assumptions (Proposition 2 in Vaswani et al. (2019a)).

An important distinction between the ES (13) and the ER condition, is that (ER) always holds trivially for full batch sampling ($g(x) = \nabla f(x)$). In contrast ES may not hold. We found that this simple fact prevented us from obtaining the correct rates of convergence of SGD with ES in the full batch setting (see Appendix D.2).

In concurrent work, Khaled and Richtarik (2020) propose an analysis of SGD for general smooth non-convex functions (and functions satisfying the PL condition[4]) under the following ABC condition:

ABC. Let $A, B, C \geq 0$. We say that ABC condition holds if

$$\mathbb{E}\left[\|g(x)\|^2\right] \leq 2A(f(x) - f(x^*) + B\|\nabla f(x)\|^2 + C. \quad (14)$$

We note that by properly choosing the constants A, B and C in the ABC condition we can recover the assumptions SGC, WGC, and ES appearing in Theorem 3.4. In Appendix B.3 we show how condition (7) which is a consequence of ER is also a special case of the ABC assumption.

4 CONVERGENCE ANALYSIS

In this section, we present the main convergence results. Proofs of all key results can be found in the Appendix C. In Appendix D, we present additional convergence results on quasar-strongly convex functions (Section D.1) and on convergence under expected smoothness (Section D.2).

4.1 Quasar Convex functions

4.1.1 Constant and Decreasing Step-sizes

Now we present our results for quasar-convex functions for SGD with a constant, finite horizon and decreasing step sizes.

Theorem 4.1. Assume $f(x)$ is L–smooth, ζ–quasar-convex with respect to x^* and $g \in ER(\rho)$. Let $0 < \gamma_k < \frac{\zeta}{2\rho+L}$ for all $k \in \mathbb{N}$ and let $r_0 := \|x^0 - x^*\|^2$. Then iterates of SGD given by (3) satisfy:

$$\min_{t=0,\ldots,k-1} \mathbb{E}\left[f(x^t) - f(x^*)\right] \leq \frac{1}{\sum_{i=0}^{k-1}\gamma_i(\zeta - \gamma_i(2\rho+L))}\left[\frac{r^0}{2} + \sigma^2\sum_{t=0}^{k-1}\gamma_t^2\right]. \quad (15)$$

Moreover, for $\gamma < \frac{\zeta}{2\rho+L}$ we have that

1. If $\forall k \in \mathbb{N}$, $\gamma_k = \gamma \equiv \frac{1}{2}\frac{\zeta}{(2\rho+L)}$ then $\forall k \in \mathbb{N}$,

$$\min_{t=0,\ldots,k-1} \mathbb{E}\left[f(x^t) - f(x^*)\right] \leq 2r_0\frac{2\rho+L}{\zeta^2 k} + \frac{\sigma^2}{2\rho+L}. \quad (16)$$

2. Suppose SGD (3) is run for T iterations. If $\forall k = 0, \ldots, T-1$, $\gamma_k = \frac{\gamma}{\sqrt{T}}$ then

$$\min_{t=0,\ldots,T-1} \mathbb{E}\left[f(x^t) - f(x^*)\right] \leq \frac{r_0 + 2\gamma^2\sigma^2}{\gamma\sqrt{T}}. \quad (17)$$

3. If $\forall k \in \mathbb{N}$, $\gamma_k = \frac{\gamma}{\sqrt{k+1}}$ then $\forall k \in \mathbb{N}$,

$$\min_{t=0,\ldots,k-1} \mathbb{E}\left[f(x^t) - f(x^*)\right] \leq \frac{1}{4\gamma}\frac{r_0 + 2\gamma^2\sigma^2(\log(k)+1)}{\zeta(\sqrt{k}-1) - \gamma(\rho + L/2)(\log(k)+1)}, \quad (18)$$

which is a convergence rate of $\mathcal{O}\left(\frac{\log(k)}{\sqrt{k}}\right)$.

To the best of our knowledge, the only prior result for the convergence of SGD for smooth quasar-convex functions was a finite horizon result similar to (17) but under the strong assumption of bounded gradient variance (Hardt et al., 2018). Of particular importance is (18) which is the first $\mathcal{O}\left(\log(k)/\sqrt{k}\right)$ any time convergence rate for quasar-convex functions. Indeed, this rate has only been achieved before under the *strictly stronger assumption* that the f_i's are smooth, convex and $g(x)$ has bounded variance (Nemirovski et al., 2009). Indeed, strictly stronger since due to Theorem 3.4 the ER condition holds when the f_i's are smooth and convex without any bounded gradient assumption.

When considering interpolated functions, we can completely drop the ER condition due to Theorem 3.4. In this next corollary we highlight this and show how the complexity of SGD is affected by increasing the minibatch size.

Corollary 4.2. Let f be ζ-quasar-convex with respect to x^*. Let the interpolation Assumption 2.3 hold and let each f_i be L_i–smooth. If v is a b-minibatch

[4]Under different step-size selection than the one we propose in our theorems for PL functions.

sampling and $\gamma_k \equiv \frac{1}{2}\frac{\zeta(n-1)b}{2L_{\max}(n-b)+L(n-1)b}$ then

$$\min_{t=0,\ldots,k-1} \mathbb{E}\left[f(x^t)-f(x^*)\right] \leq \frac{2L_{\max}(n-b)+L(n-1)b}{\zeta^2(n-1)b}\frac{2r_0}{k}. \tag{19}$$

This shows that $TC(b)$, the *total complexity* as a function of the minibatch size, to bring $\min_{i=1,\ldots,k-1} \mathbb{E}\left[f(x^i)-f^*\right] \leq \epsilon$ is given by

$$TC(b) \;\leq\; \frac{2(n-b)L_{\max}+(n-1)bL}{\zeta^2(n-1)}\frac{2r_0}{\epsilon}. \tag{20}$$

Thus the optimal minibatch size b^* that minimizes this total complexity is given by

$$b^* = \begin{cases} 1 & \text{if } (n-1) \geq 2\frac{L_{\max}}{L} \\ n & \text{if } (n-1) < 2\frac{L_{\max}}{L}. \end{cases} \tag{21}$$

Specializing (19) to the full batch setting ($n=b$), we have that gradient descent (GD) with step size $\gamma = \frac{\zeta}{4L}$ converges as follows[5]: $f(x^t)-f(x^*) \leq \frac{2L\|x^0-x^*\|^2}{\zeta^2 k}$. This is exactly the rate given recently for GD for quasar-convex functions in Guminov and Gasnikov (2017), with the exception that we have a squared dependency on ζ the quasar-convex parameter.

4.1.2 Stochastic Polyak Step-size (SPS) - Guarantee Convergence without tuning

The stochastic Polyak step size (SPS) is a recently proposed adaptive step size selection for SGD (Loizou et al., 2020). SPS is a natural extension of the classical Polyak step-size (Polyak, 1987) (commonly used in the deterministic subgradient method) to the stochastic setting.

In this work, we generalize the SPS to the arbitrary sampling regime and provide a novel convergence analysis of SGD with SPS for the class of smooth, quasar (strongly) convex functions.

Let v be a sampling vector and let $f_v = \sum_{i=1}^n f_i(x)v_i$. Let $f_v^* = \min_{x\in\mathbb{R}^n} f_v(x)$ which we assume exists. Just like the gradient, we have that f_v is an unbiased estimate of f. Now given a sampling vector v, we define the *Stochastic Polyak Step-size* (SPS) as

$$\text{SPS:}\quad \gamma_k = \frac{f_v(x^k)-f_v^*}{c\,\|\nabla f_v(x^k)\|^2}, \tag{22}$$

where $0 < c \in \mathbb{R}$. As explained in Loizou et al. (2020), the SPS rule is particularly effective when training over-parameterized models capable of interpolating the training data (when the interpolation Assumption 2.3 holds). In this case, SGD with SPS converges to the exact minimum (not to a neighborhood of the solution) (Loizou et al., 2020). In addition, if $f_i^* := \min_{x\in\mathbb{R}^n} f_i(x)$ then for machine learning problems using standard unregularized surrogate loss functions (e.g. squared loss for regression, hinge loss for classification) it holds that $f_i^* = 0$ (Loizou et al., 2020). If on top of this, we assume that interpolation Assumption 2.3 holds (that is, $f_i^* = f_i(x^*)$, $\forall i \in [n]$), then we have that $f_i^* = f_v^* = f_v(x^*) = 0$ for every $i \in [n]$ and for every v.

By assuming that every f_i is L_i–smooth, we have that f_v is L_v–smooth with $L_v := \frac{1}{n}\sum_{i=1}^n v_i L_i$. This smoothness combined with Lemma A.2 and Jensen's inequality gives a lower bound on SPS (22):

$$\frac{1}{2c\mathbb{E}\left[L_v\right]} \overset{Jensen}{\leq} \mathbb{E}\left[\frac{1}{2cL_v}\right] \leq \mathbb{E}\left[\gamma_k = \frac{f_v(x^k)-f_v^*}{c\|\nabla f_v(x^k)\|^2}\right]. \tag{23}$$

This lower bound combining with the following new bound allows us to establish the forthcoming theorem for quasar-convex functions.

Lemma 4.3. Assume interpolation 2.3 holds. Let f_i be L_i–smooth and let v be a sampling vector. It follows that there exists $\mathcal{L}_{\max} > 0$ such that

$$\frac{1}{2\mathcal{L}_{\max}}(f(x)-f^*) \;\leq\; \mathbb{E}\left[\frac{(f_v(x)-f_v^*)^2}{\|\nabla f_v(x)\|^2}\right]. \tag{24}$$

Furthermore, for $B \subset \{1,\ldots,n\}$ let L_B be the smoothness constant of $f_B := \frac{1}{b}\sum_{i\in B} f_i$. If v is the b–minibatch sampling then

$$\mathcal{L}_{\max} \;=\; \mathcal{L}_{\max}(b) \;=\; \max_{i=1,\ldots,n} \frac{\binom{n-1}{b-1}}{\sum_{B:i\in B} L_B^{-1}}.$$

With the above lemma we can now establish our main theorem.

Theorem 4.4. Let v be a sampling vector. Assume interpolation 2.3 holds. Assume that each f_i is ζ-quasar-convex with respect to x^* and L_i-smooth. Then SGD with SPS (22) and $c > \frac{1}{2\zeta}$ converges as follows:

$$\min_{i=0,\ldots,K-1} \mathbb{E}\left[f(x^i)-f^*\right] \;\leq\; \frac{2c^2}{2c\zeta-1}\frac{\mathcal{L}_{\max}}{K}\|x^0-x^*\|^2,$$

where $\mathcal{L}_{\max}$ is defined in Lemma (4.3).

We now use $\mathcal{L}_{\max}(b)$ given in Lemma 4.3 to derive the importance sampling complexity. To the best of our knowledge, this is the first importance sampling result for SGD with SPS in any setting.

[5]Here we use that the smoothness of f guarantees that $f(x^1),\ldots,f(x^t)$ for GD is a decreasing sequence.

Corollary 4.5. Consider the setting of Theorem 4.4 with $c = 1/4\zeta$. Given $\epsilon > 0$ we have that

$$k \geq \frac{\mathcal{L}_{\max}}{4\zeta^2} \frac{\|x^0 - x^*\|^2}{\epsilon} = \mathcal{O}\left(\frac{\mathcal{L}_{\max}}{\zeta^2 \epsilon}\right) \Rightarrow \min_{i=0,\dots,K-1} \mathbb{E}\left[f(x^i) - f^*\right] < \epsilon. \quad (25)$$

1. (Full batch) If we use full batch sampling we have that $\mathcal{L}_{\max} = L$ and (25) becomes $\mathcal{O}(L/\epsilon\zeta^2)$
2. (Importance sampling). If we use single element sampling with $p_i = L_i / \sum_j L_j$ we have that $\mathcal{L}_{\max} = \frac{1}{n}\sum_{j=1} L_j := \overline{L}$ and (25) becomes $\mathcal{O}(\overline{L}/\epsilon\zeta^2)$.

We highlight that the result on importance sampling of Corollary 25 requires the knowledge of the smoothness parameters L_i. This comes in contradiction with the parameter-free nature of the stochastic Polyak step-size. However, such result was missing from the literature and we believe that it could work as a first step towards the understanding of efficient (parameter-free) non-uniform sampling variants of SGD with SPS. We leave such extensions for future work.

4.2 PL Condition

Here we present our convergence results for functions satisfying the PL condition (5).

4.2.1 Constant Step-size

Let us start by presenting convergence guarantees for SGD with constant step-size.

Theorem 4.6. Let f be L-smooth. Assume $f \in PL(\mu)$ and $g \in \mathrm{ER}(\rho)$. Let $\gamma_k = \gamma \leq \frac{1}{1+2\rho/\mu}\frac{1}{L}$, for all k, then SGD given by (3) converges as follows:

$$\mathbb{E}[f(x^k) - f^*] \leq (1-\gamma\mu)^k \left[f(x^0) - f^*\right] + \frac{L\gamma\sigma^2}{\mu}. \quad (26)$$

Hence, given $\epsilon > 0$ and using the step size $\gamma = \frac{1}{L}\min\left\{\frac{\mu\epsilon}{2\sigma^2}, \frac{1}{1+2\rho/\mu}\right\}$ we have that

$$k \geq \frac{L}{\mu}\max\left\{\frac{2\sigma^2}{\mu\epsilon}, 1 + \frac{2\rho}{\mu}\right\} \log\left(\frac{2(f(x^0) - f^*)}{\epsilon}\right) \implies \mathbb{E}\left[f(x^k) - f^*\right] \leq \epsilon. \quad (27)$$

When the function is able to interpolate the data (interpolation condition 2.3 is satisfied), SGD with constant step size convergences with a linear rate to the exact solution (no neighborhood of convergence), as we show next.

Corollary 4.7. Consider the setting of Theorem 4.6 and assume interpolation 2.3 holds. Then SGD with $\gamma_k = \gamma \leq \frac{1}{1+2\rho/\mu}\frac{1}{L}$ converges linearly at a rate of $(1-\gamma\mu)$. Consequently for every $\epsilon > 0$, the iteration complexity of SGD to achieve $\mathbb{E}\left[f(x^k) - f^*\right] \leq \epsilon$ is

$$k \;\geq\; \frac{L}{\mu}\left(1 + 2\frac{\rho}{\mu}\right)\log\left(\frac{f(x^0) - f^*}{\epsilon}\right). \quad (28)$$

If v is a b–minibatch sampling then $TC(b)$, the *total complexity* with respect to the minibatch size, is

$$TC(b) \;\leq\; \frac{L}{\mu}\left(b + 2\frac{L_{\max}}{\mu}\frac{n-b}{n-1}\right)\log\left(\frac{f(x^0) - f^*}{\epsilon}\right). \quad (29)$$

Finally, let $\kappa_{\max} := L_{\max}/\mu$. The minibatch size b^* that optimizes the total complexity is given by

$$b^* = \begin{cases} 1 & \text{if } n-1 \geq 2\kappa_{\max} \\ n & \text{if } n-1 < 2\kappa_{\max}. \end{cases} \quad (30)$$

Note that Corollary 4.7 recovers the linear convergence rate of the gradient descent algorithm under the PL condition (Karimi et al., 2016) as a special case. Indeed for gradient descent we have that $\sigma = 0 = \rho$. Thus by choosing $\gamma = \frac{1}{L}$ the resulting iteration complexity is $\frac{L}{\mu}\log(\epsilon^{-1})$ which is currently the tightest known convergence result for gradient descent under the PL condition Karimi et al. (2016). On the other extreme, we see that for $b = 1$, that is SGD without minibatching, we obtain the convergence rate $1 - \mu^2/3LL_{\max}$ which matches the current state-of-the-art rate (Vaswani et al., 2019a, Thm. 4), (Khaled and Richtarik, 2020, Thm. 3) and (Lei et al., 2019, Thm. 4) known under the exact same assumptions. Thus we recover the best known rate on either end ($b = n$ and $b = 1$), and give the first rates for everything in between $1 < b < n$. To the best of our knowledge our result is the first analysis of SGD for PL functions that recovers the deterministic gradient descent convergence as special case.

The closest work to our result, on the convergence of SGD for PL functions is Khaled and Richtarik (2020). There the authors provide similar convergence result to Theorem 4.6 but using different step-size selection and under the slightly more general ABC condition (14). In Appendix C.5.1 we present a detailed comparison of our Theorem 4.6 and Theorem 3 in Khaled and Richtarik (2020).

4.2.2 Decreasing Step-size

As an extension of Theorem 4.6, we also show how to obtain a $\mathcal{O}(1/k)$ convergence for SGD using an insightful *stepsize-switching rule*. This stepsize-switching rule describes when one should switch from a constant to a decreasing step-size regime.

Theorem 4.8 (Decreasing step sizes/switching strategy)**.** Let f be an L-smooth. Assume $f \in PL(\mu)$ and $g \in \text{ER}(\rho)$. Let $k^* := 2\frac{L}{\mu}\left(1+2\frac{\rho}{\mu}\right)$ and

$$\gamma^k = \begin{cases} \dfrac{\mu}{L(\mu+2\rho)}, & \text{for} \quad k \leq \lceil k^* \rceil \\ \dfrac{2k+1}{(k+1)^2\mu} & \text{for} \quad k > \lceil k^* \rceil \end{cases} \tag{31}$$

If $k \geq \lceil k^* \rceil$, then SGD given by (3) satisfies:

$$\mathbb{E}[f(x^k) - f^*] \leq \frac{4L\sigma^2}{\mu^2}\frac{1}{k} + \frac{(k^*)^2}{k^2 e^2}[f(x^0) - f^*]. \tag{32}$$

Stochastic Polyak-Step-size (SPS). For the convergence of SGD with SPS for solving functions satisfying the PL condition we refer the interested reader to Theorem 3.5 in Loizou et al. (2020). There the authors focus on analyzing SGD with single-element uniform sampling. By assuming interpolation, their convergence results can be trivially extended to the arbitrary sampling paradigm using the lower bound (23) and Lemma 4.3.

5 EXAMPLES

In this section we provide some examples of classes of nonconvex functions that satisfy the assumptions of our main theorems.

System Identification. In optimal control sometimes we need to learn the underlining dynamics of the system we are trying to control. For instance, consider the system governed by the *linear dynamics*

$$h_{t+1} = Ah_t + Bw_t \tag{33}$$
$$y_t = Ch_t + Dw_t + \xi_t, \tag{34}$$

where $w_t \in \mathbb{R}$ and $y_t \in \mathbb{R}$ are the input and output at time t, $h_t \in \mathbb{R}^d$ is the hidden state, and $\xi_t \in \mathbb{R}$ is a random variable sampled i.i.d at each iteration. The parameters we want would to learn are the matrices $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times 1}$, $C \in \mathbb{R}^{1\times d}$ and $D \in \mathbb{R}$ that govern the dynamics. Furthermore, we can only observe the input-output pairs (w_t, y_t) by simulating the dynamics.

Our goal is to use the collected samples of the simulation (w_t, y_t) to then *fit* a linear model

$$\begin{aligned} h_{t+1} &= \hat{A}h_t + \hat{B}w_t \\ \hat{y}_t &= \hat{C}h_t + \hat{D}w_t, \end{aligned} \tag{35}$$

governed by the matrices $x := (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ such that the output of our model $\hat{y}_t$, and that of the simulation y_t are close. That is we want to solve

$$\min_{x=(\hat{A},\hat{B},\hat{C},\hat{D})} f(x) := \mathbb{E}_{w_t,\xi_t}\left[\frac{1}{T}\sum_{i=1}^{T}\|y_t - \hat{y}_t\|^2\right]. \tag{36}$$

As done in Hardt et al. (2018), we assume that the states w_t are sampled from some fixed distribution.

This objective function (36) is highly non-convex due to repeated multiplications of the parameters, as we can see by substituting out the hidden states and unrolling the recurrence (35) since

$$\hat{y}_t \;=\; \hat{D}w_t + \sum_{k=t}^{t-1}\hat{C}\hat{A}^{t-k-1}\hat{B}w_k + \hat{C}\hat{A}^{t-1}h_0. \tag{37}$$

Despite this non-convexity, the objective function (36) is quasar-convex (4) and L–weakly smooth[6], that is

$$\|\nabla f(x)\|^2 \;\leq\; 2L(f(x) - f(x^*)). \tag{WS}$$

By also bounding the domain of the parameters, Hardt et al. (2018) show that the stochastic gradients $g(x)$ have bounded variance

$$\mathbb{E}\left[\|\nabla f(x) - g(x)\|^2\right] \;\leq\; \sigma^2. \tag{BV}$$

Hardt et al. (2018) then use quasar convexity, (WS) and (BV) to show that the linear dynamics (34) can be learned with SGD in polynomial time.

As a consequence of Hardt et al. (2018) results, first we show that the objective function (36) satisfies the assumptions of our Theorem 4.1.

Theorem 5.1. The following hierarchy holds

$$\boxed{BV + WS} \longrightarrow \boxed{ES} \longrightarrow \boxed{\text{ER}}$$

Furthermore, there are functions for which (ER) holds and (BV) does not.

Consequently, since (36) satisfies (BV), (WS) and (4) we have that it satisfies (ER) and (4), and thus by Theorem 4.1 SGD applied to (36) converges at a rate of $O(1/\sqrt{t})$.

We conjecture that the linear dynamics (34) could be learned without the bounded gradient assumption by only relying on the (ER) condition. This would be significant because, it would mean that the costly projection step onto the constrained set of parameters, required so that (BV) holds, may not be necessary. We leave this conjecture to be verified in future work.

[6]To be precise the objective function is well approximated and upper bounded by a quasar-convex and weakly-smoooth function, which also requires some domain restrictions. SGD is then applied to this upper bound. See (Hardt et al., 2018) for details.

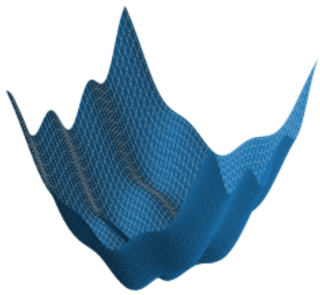

Figure 1: Surface plot of $x^2 + 3\sin^2(x) + 1.5y^2 + 4\sin^2(y)$

Contrived Illustrative Example. To given an example of a visually non-convex functions that satisfies both the PL and ER condition we consider the separable functions $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x_i)$. If each $f_i(x_i)$ satisfies the PL condition with constant μ_i then $f(x)$ satisfies the PL condition with $\mu = \min_{i=1,\ldots,n} \frac{\mu_i}{n}$. If in addition each f_i is a smooth function then according to Theorem 3.4 we have that the ER condition holds, and thus Theorem 4.7 holds. As an example, consider the nonconvex function

$$f(x) = \frac{1}{n}\sum_{i=1}^n a_i(x_i^2 + 4b_i \sin^2(x_i)) := f_i(x), \qquad (38)$$

where $a_i > 0$ and $1 > b_i > 0$ for $i = 1, \ldots, n$, so that each f_i satisfies the PL condition (see Karimi et al. (2016)[7]). The function (38) is interpolated since $x^* = 0$ is a global minima for each f_i. Furthermore f_i is smooth since $|f_i''(x)| \leq 2a_i + 6b_i$. By the above arguments, so does f satisfy the PL condition. Thus by Theorem 4.7 we know that SGD converges linearly when applied to (38). To illustrate that such functions (38) are nonconvex, we have a surface plot for $n = 2$ in Figure 1.

Nonlinear least squares. Let $F : \mathbb{R}^d \to \mathbb{R}^n$ be a differentiable function where $DF(x) \in \mathbb{R}^{n\times d}$ is its Jacobian. Now consider the nonlinear least squares problem $\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{2n}\|F(x) - y\|^2 = \frac{1}{2n}\sum_{i=1}^n (F_i(x) - y_i)^2$, where $y \in \mathbb{R}^n$.

Lemma 5.2. Assume there exists $x^* \in \mathbb{R}^d$ such that $F(x^*) = y$. If the $F_i(x)$ functions are Lipschitz and the $DF(x)$ has full row rank then F satisfies the PL and the ER condition.

Star/quasar-convex. Several nonconvex empirical risk problems are quasar-convex functions (Lee and Valiant, 2016). Let $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ be a smooth star-convex (quasar-convex with $\zeta = 1$) centered at 0. Let $\mathbf{A} \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ such that there exists $\mathbf{A}x^* = b$. Since compositions of affine maps with star convex functions are star convex (Lee and Valiant, 2016, Section A.4) we have that $f_i(\mathbf{A}x - b)$ is star convex centered at x^*. Furthermore the average of star convex functions that share the same center are star convex. Thus, $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{A}x - b)$, is a star-convex function which also satisfies the interpolation condition.

[7] In Karimi et al. (2016) the authors claim that $x^2 + 3\sin^2(x)$ is PL. We then used computer aided analysis to show that $x^2 + 3b\sin^2(x)$ satisfies the PL condition for $0 < b < 4$.

6 CONCLUSION

We establish a hierarchy between the expected residual (ER) condition and a host of other assumptions previously used in the analysis of SGD in the smooth setting, showing that ER is a strictly weaker condition. Using the ER, we present the first convergence results for SGD under different step-size selections (constant, decreasing, and stochastic Polyak step-size) on quasar-convex functions (4) without the bounded gradient or bounded variance assumption. For functions satisfying the PL condition (5) we provide tight theoretical convergence guarantees for minibatch SGD that recover the best-known convergence results for deterministic gradient descent and single-element sampling SGD as special cases, and all minibatch sizes in between.

Acknowledgements

Nicolas Loizou acknowledges support by the IVADO post-doctoral funding program.

The work of Othmane Sebbouh was supported in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). Othmane Sebbouh also acknowledges the support of a "Chaire d'excellence de l'IDEX Paris Saclay".

References

Anitescu, M. (2000). Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2019). Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*.

Bengio, Y. (2015). Rmsprop and equilibrated adaptive learning rates for nonconvex optimization. *corr abs/1502.04390*.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, 1st edition.

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

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Fabian, M. J., Henrion, R., Kruger, A. Y., and Outrata, J. V. (2010). Error bounds: necessary and sufficient conditions. *Set-Valued and Variational Analysis*, 18(2):121–149.

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

Gower, R. and Richtárik, P. (2015). Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: General analysis and improved rates. In *ICML*.

Gower, R. M., Richtárik, P., and Bach, F. (2018). Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *arxiv:1805.02632*.

Guminov, S. and Gasnikov, A. (2017). Accelerated methods for α-weakly-quasi-convex problems. *arXiv preprint arXiv:1710.00797*.

Hardt, M., Ma, T., and Recht, B. (2018). Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: stability of stochastic gradient descent. In *ICML*.

Hazan, E. and Kale, S. (2014). Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512.

Hinder, O., Sidford, A., and Sohoni, N. S. (2019). Near-optimal methods for minimizing star-convex functions and beyond. *arXiv:1906.11985*.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*.

Khaled, A. and Richtarik, P. (2020). Better theory for SGD in the nonconvex world. *arXiv:2002.03329*.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Kleinberg, B., Li, Y., and Yuan, Y. (2018). An alternative view: When does SGD escape local minima? In *ICML*.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. (2020). A unified theory of decentralized SGD with changing topology and local updates. *arXiv preprint arXiv:2003.10422*.

Lee, J. C. H. and Valiant, P. (2016). Optimizing star-convex functions. In Dinur, I., editor, *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS*, pages 603–614.

Lei, Y., Hu, T., Li, G., and Tang, K. (2019). Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, J., Wright, S., Ré, C., Bittorf, V., and Sridhar, S. (2014). An asynchronous parallel stochastic coordinate descent algorithm. In *ICML*.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. (2019). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Loizou, N. (2019). Randomized iterative methods for linear systems: momentum, inexactness and gossip. *arXiv preprint arXiv:1909.12176*.

Loizou, N. and Richtárik, P. (2017). Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*.

Loizou, N. and Richtárik, P. (2019). Convergence analysis of inexact randomized iterative methods. *arXiv preprint arXiv:1903.07971*.

Loizou, N., Vaswani, S., Laradji, I., and Lacoste-Julien, S. (2020). Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. *arXiv preprint arXiv:2002.10542*.

Lojasiewicz, S. (1963). A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89.

Ma, S., Bassily, R., and Belkin, M. (2018). The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *ICML*.

Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*.

Necoara, I., Nesterov, Y., and Glineur, F. (2018). Linear convergence of first order methods for nonstrongly convex optimization. *Mathematical Programming*, pages 1–39.

Needell, D., Srebro, N., and Ward, R. (2016). Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming, Series A*, 155(1):549–573.

Needell, D. and Ward, R. (2017). Batched stochastic gradient descent with weighted sampling. In *Approximation Theory XV, Springer*, volume 204 of *Springer Proceedings in Mathematics & Statistics,*, pages 279 – 306.

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

Nemirovski, A. and Yudin, D. B. (1978). On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Mathetmatics Doklady*, 19.

Nemirovski, A. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley Interscience.

Nesterov, Y. E. and Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. (2018). SGD and hogwild! Convergence without the bounded gradients assumption. In *ICML*.

Polyak, B. (1987). Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*.

Rahimi, A. and Recht, B. (2017). Reflections on random kitchen sinks.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*.

Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*.

Richtárik, P. and Takác, M. (2020). Stochastic reformulations of linear systems: algorithms and convergence theory. *SIAM Journal on Matrix Analysis and Applications*, 41(2):487–524.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

Rolinek, M. and Martius, G. (2018). L4: Practical loss-based stepsize adaptation for deep learning. In *Advances in Neural Information Processing Systems*.

Sa, C. D., Re, C., and Olukotun, K. (2015). Global convergence of stochastic gradient descent for some non-convex matrix problems. In *ICML*.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

Schmidt, M. and Roux, N. (2013). Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*.

Sebbouh, O., Gazagnadou, N., Jelassi, S., Bach, F., and Gower, R. (2019). Towards closing the gap between the theory and practice of SVRG. In *Advances in Neural Information Processing Systems*.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. (2007). Pegasos: primal estimated subgradient solver for SVM. In *ICML*.

Tan, Y. S. and Vershynin, R. (2019). Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *arXiv preprint arXiv:1910.12837*.

Vaswani, S., Bach, F., and Schmidt, M. (2019a). Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *AISTATS*.

Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., and Lacoste-Julien, S. (2019b). Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*.

Ward, R., Wu, X., and Bottou, L. (2019). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *ICML*.

Wright, S. and Nocedal, J. (1999). Numerical optimization. *Springer Science*, 35(67-68):7.

Zhang, H. and Yin, W. (2013). Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*.

Zhou, Y., Yang, J., Zhang, H., Liang, Y., and Tarokh, V. (2019). SGD converges to global minimum in deep learning via star-convex path. In *ICLR*.