Stochastic Gradient Descent for Nonconvex Learning Without Bounded Gradient Assumptions

Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang, *Senior Member, IEEE*

***Abstract*— Stochastic gradient descent (SGD) is a popular and efficient method with wide applications in training deep neural nets and other nonconvex models. While the behavior of SGD is well understood in the convex learning setting, the existing theoretical results for SGD applied to nonconvex objective functions are far from mature. For example, existing results require to impose a nontrivial assumption on the uniform boundedness of gradients for all iterates encountered in the learning process, which is hard to verify in practical implementations. In this article, we establish a rigorous theoretical foundation for SGD in nonconvex learning by showing that this boundedness assumption can be removed without affecting convergence rates, and relaxing the standard smoothness assumption to Hölder continuity of gradients. In particular, we establish sufficient conditions for almost sure convergence as well as optimal convergence rates for SGD applied to both general nonconvex and gradient-dominated objective functions. A linear convergence is further derived in the case with zero variances.**

***Index Terms*— Learning theory, nonconvex optimization, Polyak–Łojasiewicz condition, stochastic gradient descent (SGD).**

I. Introduction

Stochastic gradient descent (SGD) is an efficient iterative method suitable to tackle large-scale data sets due to its low computational complexity per iteration and its promising practical behavior, which has found wide applications to solve optimization problems in a variety of areas including machine learning and signal processing [1], [2]. At each iteration, SGD first calculates a gradient based on a randomly selected example and updates the model parameter along the minus gradient direction of the current iterate. This strategy of processing a single training example makes SGD very popular in the big data era, which enjoys a great computational advantage over its batch counterpart [2], [3].

Theoretical properties of SGD are well understood for optimizing both convex and strongly convex objectives, the latter of which can be relaxed to other assumptions on objective functions, e.g., error bound conditions and Polyak–Łojasiewicz conditions [1], [4]. As a comparison, SGD applied to nonconvex objective functions is much less studied. Indeed, there is a huge gap between the theoretical understanding of SGD and its very promising practical behavior in the nonconvex learning setting, as exemplified in the setting of training highly nonconvex deep neural networks. For example, while theoretical analysis can only guarantee that SGD may get stuck in local minima, in practice, it often converges to special ones with good generalization ability even in the absence of early stopping or explicit regularization [3].

Motivated by the popularity of SGD in training deep neural networks and nonconvex models as well as the huge gap between the theoretical understanding and its practical success, theoretical analysis of SGD has received increasing attention recently. The first nonasymptotical convergence rate of nonconvex SGD was established in [5], which was extended to stochastic variance reduction [6] and stochastic proximal gradient descent [7]. However, these results require to impose a nontrivial boundedness assumption on the gradients at all iterates encountered in the learning process [4]–[6], [8], which, however, depends on the realization of the optimization process and is hard to check in practice. It still remains unclear whether this assumption holds or not when learning takes place in an unbounded domain, in which scenario the existing analysis is not rigorous. Furthermore, an essential assumption on the smoothness of loss functions is required in the literature to derive meaningful convergence rates [4]–[6].

In this article, we aim to contribute a better understanding of SGD. We highlight our contributions as follows.

1) We build a sound theoretical foundation for SGD by showing that bounded gradient assumptions can be removed without affecting the convergence rates for nonconvex learning. Our analysis covers both general nonconvex and gradient-dominated objectives.
2) We relax the standard smoothness assumption on loss functions to a milder Hölder continuity on gradients.
3) We introduce a zero-variance condition that allows us to derive linear convergence of SGD. Sufficient conditions are also established for the almost sure convergence measured by both function values and gradient norms.

This article is organized as follows. We formulate problems and present theoretical results in Section II. Related work and discussions are presented in Section III. We present simulation results and conclude this article in Sections IV and V.

II. Problem Formulation and Main Results

Let ρ be a probability measure defined on a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with an input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space $\mathcal{Y} \subset \mathbb{R}$. We are interested in building a prediction rule $h : \mathcal{X} \mapsto \mathcal{Y}$ based on training examples $\{z_t\}_{t\in\mathbb{N}}$ independently drawn from ρ. The quality of h at $z = (x, y)$ is measured by $\ell(h(x), y)$, where $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_+$ is a differentiable loss function, with which we define the objective function as

$$\mathcal{E}(h) = \mathbb{E}_z[\ell(h(x), y)] = \int \ell(h(x), y) d\rho. \tag{1}$$

Manuscript received March 10, 2019; revised August 21, 2019; accepted November 2, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1003102, in part by the National Natural Science Foundation of China under Grant 11571078, Grant 11671307, Grant 61672478, and Grant 61806091, in part by the Program for University Key Laboratory of Guangdong Province under Grant 2017KSYS008, in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X386, and in part by the Alexander von Humboldt Foundation. *(Corresponding author: Guiying Li.)*

Y. Lei, G. Li, and K. Tang are with the University Key Laboratory of Evolving Intelligent Systems of Guangdong Province, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: leiyw@sustech.edu.cn; ligy@sustech.edu.cn; tangk3@sustech.edu.cn).

T. Hu is with the School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China (e-mail: tinghu@whu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2952219

We consider models $h_{\mathbf{w}}$ that can be indexed by $\mathbf{w} \in \mathbb{R}^d$ and denote $\mathcal{E}(\mathbf{w}) = \mathcal{E}(h_{\mathbf{w}})$. We implement learning by SGD to minimize the objective function $\mathcal{E}(\mathbf{w})$. Let $\mathbf{w}_1 = 0$ and $z_t = (x_t, y_t)$ be the example sampled according to ρ at the tth iteration. We update the model sequence $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ in $\mathbb{R}^d$ by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell(h_{\mathbf{w}_t}(x_t), y_t) \nabla_{\mathbf{w}} h_{\mathbf{w}_t}(x_t) = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t, z_t) \tag{2}$$

where $\nabla \ell$ denotes the gradient of ℓ with respect to the first argument, $\{\eta_t\}_{t\in\mathbb{N}}$ is a positive step size sequence, and we introduce $f(\mathbf{w}, z) = \ell(h_{\mathbf{w}}(x), y)$ for brevity. We denote $\|\cdot\|_2$ the ℓ_2 norm, i.e., $\|\mathbf{w}\|_2 = (\sum_{j=1}^d w_j^2)^{(1/2)}$ for $\mathbf{w} = (w_1, \ldots, w_d)$.

This article is based on a fundamental assumption on the regularity of loss functions. Assumption 1 with $\alpha = 1$ is the standard smooth assumption, which we extend to a general Hölder continuity assumption on the gradient of loss functions.

Assumption 1: Let $\alpha \in (0, 1]$ and $L > 0$. We assume that the gradient of $f(\cdot, z)$ is α-Hölder continuous in the sense that

$$\|\nabla f(\mathbf{w}, z) - \nabla f(\tilde{\mathbf{w}}, z)\|_2 \leq L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^\alpha \quad \forall \mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d \ z \in \mathcal{Z}.$$

We have the following useful lemma under Assumption 1.

Lemma 2: Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ be a differentiable function. Let $\alpha \in (0, 1]$ and $L > 0$. If for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d, z \in \mathcal{Z}$

$$\|\nabla\phi(\mathbf{w}) - \nabla\phi(\tilde{\mathbf{w}})\|_2 \leq L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^\alpha \tag{3}$$

then, we have

$$\phi(\tilde{\mathbf{w}}) - \big(\phi(\mathbf{w}) + \langle \tilde{\mathbf{w}} - \mathbf{w}, \nabla\phi(\mathbf{w})\rangle\big) \leq \frac{L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^{1+\alpha}}{1+\alpha}. \tag{4}$$

Furthermore, if $\phi(\mathbf{w}) \geq 0$ for all $\mathbf{w} \in \mathbb{R}^d$, then

$$\|\nabla\phi(\mathbf{w})\|_2^{\frac{1+\alpha}{\alpha}} \leq \frac{(1+\alpha)L^{\frac{1}{\alpha}}}{\alpha}\phi(\mathbf{w}) \quad \forall \mathbf{w} \in \mathbb{R}^d. \tag{5}$$

Lemma 2 to be proved in the Appendix is an extension of [9, Proposition 1] from univariate functions to multivariate functions. It should be noted that (5) improves [9, Proposition 1(d)] by removing a factor of $(1 + \alpha)^{(1/\alpha)}$. The left-hand side of (4) quantifies the error incurred by approximating $\phi(\tilde{\mathbf{w}})$ with its first-order approximation at $\mathbf{w}$. Therefore, Hölder continuity on gradients amounts to saying that this approximation error can be bounded by a polynomial function of the distance $\|\mathbf{w} - \tilde{\mathbf{w}}\|_2$ with exponent in $(1, 2]$. A similar restricted smoothness condition was considered in [10] by assuming this approximation error to be bounded by $L_1\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 + L_2 \log(d) \ \|\mathbf{w} - \tilde{\mathbf{w}}\|_1^2$, where L_1 and L_2 are two constants (L_2 may be very small and depend on the sample size). Equation (5) provides a self-bounding property, meaning that the norm of gradients can be controlled by function values.

A. General Nonconvex Objective Functions

We now present theoretical results for SGD with general nonconvex functions to be proved in the Appendix. In this case, we measure the progress of SGD by gradients. Part 1) gives a nonasymptotic convergence rate by step sizes, while Parts 2) and 3) provide sufficient conditions on the asymptotic convergence measured by function values and gradient norms.

Theorem 3: Suppose that Assumption 1 holds. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be produced by (2) with the step sizes satisfying $C_1 := \sum_{t=1}^\infty \eta_t^{1+\alpha} < \infty$. Then, the following three statements hold.

1) There is a constant C independent of t such that

$$\min_{t=1,\ldots,T} \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2] \leq C\left(\sum_{t=1}^T \eta_t\right)^{-1}. \tag{6}$$

2) $\{\mathcal{E}(\mathbf{w}_t)\}_t$ converges to an almost surely (a.s.) bounded random variable.
3) If Assumption 1 holds with $\alpha = 1$ and $\sum_{t=1}^\infty \eta_t = \infty$, then $\lim_{t\to\infty} \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] = 0$.

Remark 4: Part 1) was derived in [5] under the assumption

$$\mathbb{E}_z\big[\|\nabla f(\mathbf{w}_t, z) - \nabla\mathcal{E}(\mathbf{w}_t)\|_2^2\big] \leq \sigma^2 \quad \forall t \in \mathbb{N} \tag{7}$$

and a constant $\sigma > 0$. This boundedness assumption depends on the realization of the optimization process and it is therefore difficult to check in practice. It was removed in our analysis. Although Parts 2), 3) do not give rates, an appealing property is that they consider individual iterates. As a comparison, the convergence rates in (6) only hold for the minimum of the first T iterates. The analysis for individual iterates is much more challenging than that for the minimum over all the iterates. Indeed, Part 3) is based on a careful contradiction strategy.

Explicit rates follow by instantiating the step sizes. If $\alpha = 1$, the convergence rate in Part 2) becomes $O(T^{-(1/2)} \log^{(\beta/2)} T)$ which matches the best known rate for SGD under the bounded gradient assumption up to a logarithmic factor.

Corollary 5: Suppose that Assumption 1 holds. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be the sequence produced by (2). Then, the following statements hold.

1) If $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in (1/(1+\alpha), 1)$, then $\min_{t=1,\ldots,T} \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2] = O(T^{\theta-1})$.
2) If $\eta_t = \eta_1(t \log^\beta(t+1))^{-(1/1+\alpha)}$ with $\beta > 1$, then $\min_{t=1,\ldots,T} \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2] = O(T^{-(\alpha/\alpha+1)} \log^{(\beta/1+\alpha)} T)$.

Remark 6: As shown in Corollary 5, the convergence rate improves from $O(1)$ to $O(T^{-(1/2)})$, if α increases from 0 to 1 (up to a logarithmic factor). The decrease of the convergence rate is the cost we pay for considering a more general condition.

B. Objective Functions With Polyak–Łojasiewicz Inequality

We now proceed with our convergence analysis by imposing an assumption referred to as PL inequality named after Polyak and Łojasiewicz [1]. Intuitively, this inequality means that the suboptimality of iterates measured by function values can be bounded by gradient norms. PL condition is also referred to as the gradient-dominated condition in the literature [6], and widely adopted in the analysis in both the convex and nonconvex optimization setting [4], [11], [12]. Examples of functions satisfying PL condition include neural networks with one-hidden layers, ResNets with linear activation, and objective functions in matrix factorization [12]. It should be noted that functions satisfying the PL condition are not necessarily convex.

Assumption 7: Let $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathcal{E}(\mathbf{w})$. We assume that $\mathcal{E}$ satisfies the PL inequality with the parameter $\mu > 0$

$$\mathcal{E}(\mathbf{w}) - \mathcal{E}(\mathbf{w}^*) \leq (2\mu)^{-1}\|\nabla\mathcal{E}(\mathbf{w})\|_2^2 \quad \forall \mathbf{w} \in \mathbb{R}^d.$$

Under Assumption 7, we can state convergence results measured by the suboptimality of function values to be proved in the Appendix. Part 1) provides a sufficient condition for almost sure convergence measured by function values and gradient norms, while Part 2) establishes explicit convergence rates for step sizes reciprocal to the iteration number. If $\alpha = 1$, we derive convergence rates $O(t^{-1})$ after t iterations, which is minimax optimal even when the objective function is strongly convex. Part 3) shows that a linear convergence can be achieved if $\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|_2^2] = 0$, which extends the linear convergence of gradient descent [4] to the stochastic setting. The assumption $\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|_2^2] = 0$ means that variances of the stochastic gradients vanish at $\mathbf{w} = \mathbf{w}^*$ since $\mathrm{Var}(f(\mathbf{w}^*, z)) = \mathbb{E}[\|f(\mathbf{w}^*, z) - \nabla\mathcal{E}(\mathbf{w}^*)\|_2^2] = 0$.

Theorem 8: Let Assumptions 1 and 7 hold. Let $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ be produced by (2). Then, the following statements hold.

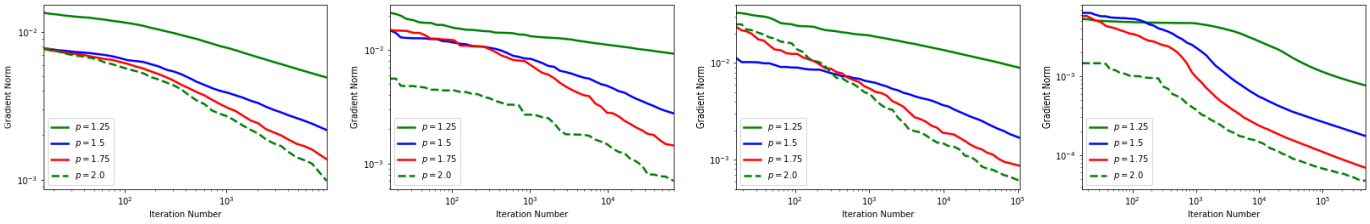

Fig. 1. Gradient norm as a function of iteration number for different data sets. (a) Splice. (b) Usps. (c) Phishing. (d) Shuttle.

1) If $\sum_{t=1}^{\infty}\eta_t^{1+\alpha} < \infty$ and $\sum_{t=1}^{\infty}\eta_t = \infty$, then a.s. $\lim_{t\to\infty}\mathcal{E}(\mathbf{w}_t) = \mathcal{E}(\mathbf{w}^*)$ and $\lim_{t\to\infty}\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2 = 0$.
2) If $\eta_t = 2/((t+1)\mu)$, then for any $t \geq t_0 := 2L^{(2/\alpha)}\mu^{-(1+\alpha/\alpha)}$ we have $\mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] - \mathcal{E}(\mathbf{w}^*) \leq \widetilde{C}t^{-\alpha}$, where $\widetilde{C}$ is a constant independent of t (explicitly given in the proof).
3) If $\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|_2^2] = 0$, Assumption 1 holds with $\alpha = 1$ and $\eta_t = \eta \leq \mu/L^2$, then

$$\mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] - \mathcal{E}(\mathbf{w}^*) \leq (1-\mu\eta)^t(\mathcal{E}(\mathbf{w}_1) - \mathcal{E}(\mathbf{w}^*)).$$

Remark 9: Conditions as $\sum_{t=1}^{\infty}\eta_t^2 < \infty$ and $\sum_{t=1}^{\infty}\eta_t = \infty$ were established for almost sure convergence with strongly convex objectives [3], which are extended here to nonconvex learning under PL conditions. Convergence rates $O(t^{-1})$ were established for nonconvex optimization under PL conditions, bounded gradient assumption as $\mathbb{E}[\|\nabla f(\mathbf{w}_t, z)\|_2^2] \leq \sigma^2$ and smoothness assumptions [4]. We derive the same convergence rates without the bounded gradient assumption, and relax the smoothness assumption to a Hölder continuity of $\nabla f(\mathbf{w}, z)$.

III. Related Work and Discussion

A. SGD for Convex Learning

SGD has been comprehensively studied in the convex setting. For generally convex objective functions, regret bounds $O(\sqrt{T})$ were established for SGD with T iterates [2], which imply convergence rates $O(1/\sqrt{T})$ [3]. For strongly convex functions, regret bounds can be improved to $O(\log T)$ [13] which imply convergence rates $O(\log T/T)$. These results were extended to learning in reproducing kernel Hilbert spaces [14]–[17] and learning with a mirror map to capture the geometry of problems [18].

B. SGD for Nonconvex Learning

As compared to the maturity of understanding in convex optimization, convergence analysis for SGD in the nonconvex setting is far from satisfactory. Asymptotic convergence of SGD was established under the assumption $\mathbb{E}_z[\|\nabla f(\mathbf{w}_t, z) - \nabla\mathcal{E}(\mathbf{w}_t)\|_2^2] \leq A\big(1 + \|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2\big)$ for $A > 0$ and all $t \in \mathbb{N}$ [19]. Nonasymptotic convergence rates similar to (6) were established in [5] under the boundedness assumption (7). For objective functions satisfying PL conditions, convergence rates $O(1/T)$ were established for SGD under the boundedness assumption (7) [4]. This boundedness assumption in the literature depends on the realization of the optimization process, which is hard to check in practical application domains including computer vision [20]. In this article, we show that the same convergence rates can be established without any boundedness assumptions. This establishes a rigorous foundation to safeguard SGD. Existing discussions also require to impose a smoothness assumption on $f(\mathbf{w}, z)$, which is relaxed to a Hölder continuity of $\nabla f(\mathbf{w}, z)$. Both the PL condition and Hölder continuity condition do not depend on the iterates and can be checked by objective function themselves, which are satisfied by many nonconvex models [4], [6], [12].

TABLE I
Description of the Data Sets Used in the Experiments

Dataset	n	d	Dataset	n	d
Splice	1000	60	Usps	7291	256
Phishing	11,055	68	Shuttle	43,500	7

It should be noted that convergence analysis was also performed when $f(\mathbf{w}, z)$ is either convex [21] or nonconvex [22] without bounded gradient assumptions, both of which, however, require $\mathcal{E}(\mathbf{w})$ to be strongly convex and $f(\mathbf{w}, z)$ to be smooth. Furthermore, we establish a linear convergence of SGD in the case with zero variances, while this linear convergence was only derived for batch gradient descent applied to gradient-dominated objective functions [4]. Necessary and sufficient conditions as $\sum_{t=1}^{\infty}\eta_t = \infty, \sum_{t=1}^{\infty}\eta_t^2 < \infty$ were established for convergence of online mirror descent in a strongly convex setting [21], which are extended here to convergence of SGD for gradient-dominated objective functions.

C. Variants of SGD

Recently, the trick of variance reduction has been introduced to accelerate the performance of SGD in both the convex [23], [24] and nonconvex setting [6], [25]. A novel nonnegative momentum was incorporated into the variance-reduction-based algorithm to derive a stochastic gradient method with an optimal convergence rate in the convex setting [26]. For a specific class of strongly nonconvex problems, an epoch-wise stochastic method was also developed to exploit the strongly nonconvex parameter [25]. These variance-reduction-based algorithms [25], [26] assume a finite-sum structure on the objective function, i.e., ρ is an empirical measure based on training examples, and require the loss function to be smooth. As a comparison, our analysis does not impose a finite-sum structure on the objective function and considers a more general class of loss functions with Hölder continuous gradients. Our analysis also does not impose a strongly nonconvex assumption on the loss function [25].

IV. Simulation Validation

In this section, we present some experimental results to confirm our theoretical findings. We consider generalized linear models of the form $h_{\mathbf{w}}(x) = \sigma(\langle\mathbf{w}, x\rangle)$ to learn the relationship underlying the training examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $y_i \in \{0, 1\}$, where $\sigma(s) = (1 + e^{-s})^{-1}$ is the logistic link function defined on $\mathbb{R}$ [12]. We consider loss functions $\ell_p(a, y) = |a - y|^p$ with $p \in [1, 2]$, for which the empirical objective function becomes $\mathcal{E}_p(\mathbf{w}) = (1/n)\sum_{i=1}^{n} f_p(\mathbf{w}, z_i)$ where $f_p(\mathbf{w}, z) = |\sigma(\langle\mathbf{w}, x\rangle) - y|^p$. For $p = 2$ and $p = 1$, the loss function becomes the least squares loss and the absolute loss, respectively. The robustness of the trained model would improve as p decreases. It can be checked that $f_p(\mathbf{w}, z)$ is nonconvex with respect to $\mathbf{w}$ and its gradient $\nabla f(\mathbf{w}; z) = p \cdot \text{sgn}(\sigma(\langle\mathbf{w}, x\rangle) - y)|\sigma(\langle\mathbf{w}, x\rangle) - y|^{p-1}\nabla_{\mathbf{w}}\sigma(\langle\mathbf{w}, x\rangle)$ is α_p-Hölder continuous with $\alpha_p = p - 1$. According to Corollary 5, if we apply SGD

with step sizes $\eta_t = \eta_1 t^{-(1/p)}$, then $\min_{t=1,\ldots,T} \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2] = O(T^{(1-p)/p})$ (up to logarithmic factors). We aim to verify this result by considering different p in $\{5/4, 3/2, 7/4, 2\}$. For each temporarily fixed p, we consider $\eta_t = \eta_1 t^{-(1/p)}$ and use ten-fold cross-validation to tune η_1 in the set $\{1, 2, 4, \ldots, 256\}$. We then compute the quantity $E_{t,p} := \min_{\tilde{t}=1,\ldots,t} \|\nabla\mathcal{E}_p(\mathbf{w}_{\tilde{t},p})\|_2$, where $\{\mathbf{w}_{\tilde{t},p}\}_t$ is a sequence of iterates produced by SGD to minimize $\mathcal{E}_p(\mathbf{w})$. We consider four classification data sets from the LIBSVM data set [27] with information given in Table I. We transform multi-class data sets into binary classification data sets by grouping the first half of class labels into positive labels, and the remaining class labels into negative labels. We repeat the experiment 20 times and report the average of results.

In Fig. 1, we plot $E_{t,p}$ versus t for $p \in \{5/4, 3/2, 7/4, 2\}$ for four data sets. For all data sets, we see that $E_{t,p}$ decays with an increasingly fast rate as we increase p from 5/4 to 2. This is consistent with Corollary 5 showing that convergence rates improve if the smoothness (reflected by the order α of Hölder continuity) of the objective function increases.

V. Conclusion

We present a solid analysis of SGD for nonconvex learning by showing that the bounded gradient assumption imposed in the literature can be removed without affecting learning rates. Our analysis relaxes the smoothness assumption to the Hölder continuity of gradients, and covers both general nonconvex and gradient-dominated functions. Interesting future work includes the extension to distributed learning and sparse learning.

Appendix
Proofs

A. Proof of Theorem 3

We present here the proofs of Theorem 3 and Corollary 5. To this aim, we first prove Lemma 2 and introduce Doob's forward convergence theorem on almost sure convergence [28].

Proof of Lemma 2: Equation (4) can be proved in the same way as the proof of Part 1) of [9, Proposition 1]. We now prove (5) for nonnegative ϕ. We only need to consider the case $\nabla\phi(\mathbf{w}) \neq 0$. In this case, set

$$\tilde{\mathbf{w}} = \mathbf{w} - L^{-\frac{1}{\alpha}} \|\nabla\phi(\mathbf{w})\|_2^{\frac{1}{\alpha}} \|\nabla\phi(\mathbf{w})\|_2^{-1} \nabla\phi(\mathbf{w})$$

in (4). We derive

$$\begin{aligned} 0 \leq \phi(\tilde{\mathbf{w}}) &\leq \phi(\mathbf{w}) - \left\langle L^{-\frac{1}{\alpha}} \|\nabla\phi(\mathbf{w})\|_2^{\frac{1}{\alpha}} \frac{\nabla\phi(\mathbf{w})}{\|\nabla\phi(\mathbf{w})\|_2}, \nabla\phi(\mathbf{w}) \right\rangle \\ &\quad + \frac{L}{1+\alpha} L^{-\frac{1+\alpha}{\alpha}} \|\nabla\phi(\mathbf{w})\|_2^{\frac{1+\alpha}{\alpha}} \\ &= \phi(\mathbf{w}) - L^{-\frac{1}{\alpha}} \|\nabla\phi(\mathbf{w})\|_2^{\frac{1+\alpha}{\alpha}} + L^{-\frac{1}{\alpha}} (1+\alpha)^{-1} \|\nabla\phi(\mathbf{w})\|_2^{\frac{1+\alpha}{\alpha}} \\ &= \phi(\mathbf{w}) - \frac{\alpha L^{-\frac{1}{\alpha}}}{1+\alpha} \|\nabla\phi(\mathbf{w})\|_2^{\frac{1+\alpha}{\alpha}} \end{aligned}$$

from which the stated bound (5) follows. □

Lemma 10: Let $\{\widetilde{X}_t\}_{t\in\mathbb{N}}$ be a sequence of nonnegative random variables with $\mathbb{E}[\widetilde{X}_1] < \infty$ and let $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a nested sequence of sets of random variables with $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all $t \in \mathbb{N}$. If $\mathbb{E}[\widetilde{X}_{t+1}|\mathcal{F}_t] \leq \widetilde{X}_t$ for all $t \in \mathbb{N}$, then $\widetilde{X}_t$ converges to a nonnegative random variable $\widetilde{X}$ a.s. and $\widetilde{X} < \infty$ a.s.

Proof of Theorem 3: By Assumption 1, we know

$$\begin{aligned} \|\nabla\mathcal{E}(\mathbf{w}) - \nabla\mathcal{E}(\tilde{\mathbf{w}})\|_2 &= \|\mathbb{E}[\nabla f(\mathbf{w}, z)] - \mathbb{E}[\nabla f(\tilde{\mathbf{w}}, z)]\|_2 \\ &\leq \mathbb{E}[\|\nabla f(\mathbf{w}, z) - \nabla f(\tilde{\mathbf{w}}, z)\|_2] \\ &\leq L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^{\alpha}. \end{aligned}$$

$\nabla\mathcal{E}(\mathbf{w})$ is α-Hölder continuous. By (4) with $\phi = \mathcal{E}$ and (2)

$$\begin{aligned} \mathcal{E}(\mathbf{w}_{t+1}) &\leq \mathcal{E}(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \nabla\mathcal{E}(\mathbf{w}_t)\rangle + \frac{L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^{1+\alpha}}{1+\alpha} \\ &= \mathcal{E}(\mathbf{w}_t) - \eta_t \langle \nabla f(\mathbf{w}_t, z_t), \nabla\mathcal{E}(\mathbf{w}_t)\rangle \\ &\quad + \frac{L\eta_t^{1+\alpha}}{1+\alpha} \|\nabla f(\mathbf{w}_t, z_t)\|_2^{1+\alpha} \\ &\leq \mathcal{E}(\mathbf{w}_t) - \eta_t \langle \nabla f(\mathbf{w}_t, z_t), \nabla\mathcal{E}(\mathbf{w}_t)\rangle \\ &\quad + \frac{L^2\eta_t^{1+\alpha}}{1+\alpha} \Big(\frac{1+\alpha}{\alpha}\Big)^{\alpha} f^{\alpha}(\mathbf{w}_t, z_t) \end{aligned} \tag{8}$$

where the last inequality is due to (5). With the Young's inequality for all $\mu, \upsilon \in \mathbb{R}$, $p^{-1} + q^{-1} = 1$, $p \geq 0$

$$\mu\upsilon \leq p^{-1}|\mu|^p + q^{-1}|\upsilon|^q \tag{9}$$

we get $(((1+\alpha) f(\mathbf{w}_t, z_t))/\alpha)^{\alpha} \leq \alpha(((1+\alpha) f(\mathbf{w}_t, z_t))/\alpha)^{\alpha(1/\alpha)} + 1 - \alpha$. Plugging the aforementioned inequality into (8) shows

$$\begin{aligned} \mathcal{E}(\mathbf{w}_{t+1}) \leq \mathcal{E}(\mathbf{w}_t) &- \eta_t \langle \nabla f(\mathbf{w}_t, z_t), \nabla\mathcal{E}(\mathbf{w}_t)\rangle \\ &+ \frac{L^2\eta_t^{1+\alpha}}{1+\alpha} ((1+\alpha) f(\mathbf{w}_t, z_t) + 1 - \alpha). \end{aligned}$$

Taking conditional expectation with respect to z_t, we derive

$$\begin{aligned} &\mathbb{E}_{z_t}[\mathcal{E}(\mathbf{w}_{t+1})] \\ &\quad \leq \mathcal{E}(\mathbf{w}_t) - \eta_t \|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2 + L^2\eta_t^{1+\alpha}(\mathcal{E}(\mathbf{w}_t) + 1 - \alpha) \qquad (10) \\ &\quad \leq (1 + L^2\eta_t^{1+\alpha})\mathcal{E}(\mathbf{w}_t) - \eta_t \|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2 + L^2(1-\alpha)\eta_t^{1+\alpha}. \qquad (11) \end{aligned}$$

It then follows that

$$\mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] \leq (1 + L^2\eta_t^{1+\alpha})\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] + L^2(1-\alpha)\eta_t^{1+\alpha}$$

from which we derive

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] + L^2(1-\alpha) \sum_{k=t+1}^{\infty} \eta_k^{1+\alpha} &\leq (1 + L^2\eta_t^{1+\alpha}) \\ &\quad \times \left(\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] + L^2(1-\alpha) \sum_{k=t}^{\infty} \eta_k^{1+\alpha} \right). \end{aligned}$$

Introduce $A_t = \mathbb{E}[\mathcal{E}(\mathbf{w}_t)] + L^2(1-\alpha)\sum_{k=t}^{\infty} \eta_k^{1+\alpha}, \forall t \in \mathbb{N}$. Then, it follows from the inequality $1 + a \leq \exp(a)$ that $A_{t+1} \leq (1 + L^2\eta_t^{1+\alpha})A_t \leq \exp(L^2\eta_t^{1+\alpha})A_t$. An application of the aforementioned inequality recursively then gives

$$A_{t+1} \leq \exp\left(L^2 \sum_{k=1}^{t} \eta_k^{1+\alpha}\right) A_1 \leq \exp\left(L^2 \sum_{k=1}^{\infty} \eta_k^{1+\alpha}\right) A_1 := C_2$$

from which we know $\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] \leq C_2, \forall t \in \mathbb{N}$. Plugging the aforementioned inequality back into (11) gives

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] \leq \mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - \eta_t \mathbb{E}\big[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2\big] + L^2\eta_t^{1+\alpha} \\ \times (C_2 + 1 - \alpha). \end{aligned} \tag{12}$$

A summation of the aforementioned inequality then implies

$$\begin{aligned} &\sum_{t=1}^{T} \eta_t \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2] \\ &\quad \leq \sum_{t=1}^{T} \big(\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - \mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})]\big) + L^2(C_2 + 1 - \alpha) \sum_{t=1}^{T} \eta_t^{1+\alpha} \\ &\quad \leq \mathcal{E}(\mathbf{w}_1) + L^2(C_2 + 1 - \alpha) C_1 \end{aligned}$$

from which we directly get (6) with $C := \mathcal{E}(\mathbf{w}_1) + L^2 C_1 (C_2 + 1 - \alpha)$. This proves Part 1).

We now prove Part 2). Multiplying both sides of (11) by $\prod_{k=t+1}^{\infty}(1+L^2\eta_k^{1+\alpha})$, the term $\prod_{k=t+1}^{\infty}(1+L^2\eta_k^{1+\alpha})\mathbb{E}_{z_t}[\mathcal{E}(\mathbf{w}_{t+1})]$ can be upper bounded by

$$\prod_{k=t}^{\infty}(1+L^2\eta_k^{1+\alpha})\mathcal{E}(\mathbf{w}_t)+L^2(1-\alpha)\prod_{k=t+1}^{\infty}(1+L^2\eta_k^{1+\alpha})\eta_t^{1+\alpha} \leq \prod_{k=t}^{\infty}(1+L^2\eta_k^{1+\alpha})\mathcal{E}(\mathbf{w}_t)+C_3\eta_t^{1+\alpha} \quad (13)$$

where we introduce $C_3 = L^2(1-\alpha)\prod_{k=1}^{\infty}(1+L^2\eta_k^{1+\alpha}) < \infty$. Introduce the stochastic process

$$\widetilde{X}_t = \prod_{k=t}^{\infty}\big(1+L^2\eta_k^{1+\alpha}\big)\mathcal{E}(\mathbf{w}_t)+C_3\sum_{k=t}^{\infty}\eta_k^{1+\alpha}.$$

Equation (13) amounts to saying $\mathbb{E}_{z_t}[\tilde{X}_{t+1}] \leq \tilde{X}_t$ for all $t \in \mathbb{N}$, which shows that $\{\widetilde{X}_t\}_{t\in\mathbb{N}}$ is a nonnegative supermartingale. Furthermore, the assumption $\sum_{t=1}^{\infty}\eta_t^{1+\alpha} < \infty$ implies that $\widetilde{X}_1 < \infty$. We can apply Lemma 10 to show that $\lim_{t\to\infty}\widetilde{X}_t = \widetilde{X}$ for a nonnegative random variable $\widetilde{X}$ a.s. This together with the assumption $\sum_{t=1}^{\infty}\eta_t^{1+\alpha} < \infty$ implies $\lim_{t\to\infty}\widetilde{Y}_t \to \widetilde{Y}$ for a nonnegative random variable $\widetilde{Y}$, where $\widetilde{Y}_t = \prod_{k=t}^{\infty}(1+L^2\eta_k^{1+\alpha})\mathcal{E}(\mathbf{w}_t)$ for all $t \in \mathbb{N}$ and $\widetilde{Y} < \infty$ a.s. Furthermore, it is clear a.s. that

$$\begin{aligned}|\mathcal{E}(\mathbf{w}_t)-\widetilde{Y}| &= \left|\left(1-\prod_{k=t}^{\infty}\big(1+L^2\eta_k^{1+\alpha}\big)\right)\mathcal{E}(\mathbf{w}_t) + \prod_{k=t}^{\infty}\big(1+L^2\eta_k^{1+\alpha}\big)\mathcal{E}(\mathbf{w}_t)-\widetilde{Y}\right| \\ &\leq \left|\left(1-\prod_{k=t}^{\infty}\big(1+L^2\eta_k^{1+\alpha}\big)\right)\right|\mathcal{E}(\mathbf{w}_t) + \left|\prod_{k=t}^{\infty}\big(1+L^2\eta_k^{1+\alpha}\big)\mathcal{E}(\mathbf{w}_t)-\widetilde{Y}\right| \xrightarrow[t\to\infty]{} 0\end{aligned}$$

where we have used the fact $\lim_{t\to\infty}\prod_{k=t}^{\infty}(1+L^2\eta_k^{1+\alpha}) = 1$ due to $\sum_{t=1}^{\infty}\eta_t^{1+\alpha} < \infty$. That is, $\mathcal{E}(\mathbf{w}_t)$ converges to $\widetilde{Y}$ a.s.

We now prove Part 3). By Assumption 1, Lemma 2, we get

$$\begin{aligned}\|\nabla f(\mathbf{w}_k,z_k)\|_2 &\leq \left(\frac{(1+\alpha)L^{\frac{1}{\alpha}}f(\mathbf{w}_k,z_k)}{\alpha}\right)^{\frac{\alpha}{1+\alpha}} \\ &\leq L^{\frac{1}{\alpha}}f(\mathbf{w}_k,z_k)+(1+\alpha)^{-1}\end{aligned}$$

where we have used the Young's inequality (9). Taking expectations over both sides and using $\mathbb{E}[\mathcal{E}(\mathbf{w}_k)] \leq C_2$, we derive

$$\begin{aligned}\mathbb{E}[\|\nabla f(\mathbf{w}_k,z_k)\|_2] &\leq L^{\frac{1}{\alpha}}\mathbb{E}[\mathcal{E}(\mathbf{w}_k)]+(1+\alpha)^{-1} \\ &\leq L^{\frac{1}{\alpha}}C_2+(1+\alpha)^{-1} := C_4. \quad (14)\end{aligned}$$

Suppose to contrary that $\limsup_{t\to\infty}\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] > 0$. By Part 1) and the assumption $\sum_{t=1}^{\infty}\eta_t = \infty$, we know

$$\liminf_{t\to\infty}\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] \leq \liminf_{t\to\infty}\sqrt{\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2]} = 0.$$

Then there exists an $\epsilon > 0$ such that $\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] < \epsilon$ for infinitely many t and $\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] > 2\epsilon$ for infinitely many t. Let $\mathcal{T}$ be a subset of integers such that for every $t \in \mathcal{T}$ we can find an integer $k(t) > t$ such that

$$\begin{aligned}&\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] < \epsilon, \quad \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_{k(t)})\|_2] > 2\epsilon \quad \text{and} \\ &\qquad \epsilon \leq \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_k)\|_2] \leq 2\epsilon \quad \text{for all } t < k < k(t). \quad (15)\end{aligned}$$

Furthermore, we can assert that $\eta_t \leq \epsilon/(2LC_4)$ for every t larger than the smallest integer in $\mathcal{T}$ since $\lim_{t\to\infty}\eta_t = 0$.

By (14), (15), and Assumption 1 with $\alpha = 1$, we know

$$\begin{aligned}\epsilon &\leq \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_{k(t)})\|_2]-\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] \\ &\leq \sum_{k=t}^{k(t)-1}\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_{k+1})-\nabla\mathcal{E}(\mathbf{w}_k)\|_2] \\ &\leq L\sum_{k=t}^{k(t)-1}\mathbb{E}[\|\mathbf{w}_{k+1}-\mathbf{w}_k\|_2] = L\sum_{k=t}^{k(t)-1}\eta_k\mathbb{E}[\|\nabla f(\mathbf{w}_k,z_k)\|_2] \\ &\leq LC_4\sum_{k=t}^{k(t)-1}\eta_k. \quad (16)\end{aligned}$$

Analogously, one can show

$$\begin{aligned}&\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_{t+1})\|_2]-\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] \\ &\quad\leq \mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_{t+1})-\nabla\mathcal{E}(\mathbf{w}_t)\|_2] \\ &\quad\leq L\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2] \leq L\eta_t\mathbb{E}[\|\nabla f(\mathbf{w}_t,z_t)\|_2] \leq LC_4\eta_t\end{aligned}$$

from which, (15) and $\eta_t \leq \epsilon/(2LC_4)$ for any t larger than the smallest integer in $\mathcal{T}$ we get

$$\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_k)\|_2] \geq \epsilon/2 \quad \text{for every } k = t,\ t+1,\ldots,k(t)-1$$

and all $t \in \mathcal{T}$. It then follows that

$$\mathbb{E}\big[\|\nabla\mathcal{E}(\mathbf{w}_k)\|_2^2\big] \geq (\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_k)\|_2])^2 \geq \epsilon^2/4 \quad (17)$$

for every $k = t, t+1,\ldots,k(t)-1$ and all $t \in \mathcal{T}$. Putting (17) back into (12), $\mathbb{E}[\mathcal{E}(\mathbf{w}_{k(t)})]$ can be upper bounded by

$$\begin{aligned}\mathbb{E}[\mathcal{E}(\mathbf{w}_t)]-\sum_{k=t}^{k(t)-1}\eta_k\mathbb{E}\big[\|\nabla\mathcal{E}(\mathbf{w}_k)\|_2^2\big]+L^2C_2\sum_{k=t}^{k(t)-1}\eta_k^2 \\ \leq \mathbb{E}[\mathcal{E}(\mathbf{w}_t)]-\frac{\epsilon^2}{4}\sum_{k=t}^{k(t)-1}\eta_k+L^2C_2\sum_{k=t}^{k(t)-1}\eta_k^2.\end{aligned}$$

This together with (16) implies that

$$\begin{aligned}\epsilon^3/(4LC_4) \leq \frac{\epsilon^2}{4}\sum_{k=t}^{k(t)-1}\eta_k \leq \mathbb{E}[\mathcal{E}(\mathbf{w}_t)]-\mathbb{E}[\mathcal{E}(\mathbf{w}_{k(t)})] \\ +L^2C_2\sum_{k=t}^{k(t)-1}\eta_k^2 \quad \forall t \in \mathcal{T}. \quad (18)\end{aligned}$$

Part 2) implies that $\{\mathbb{E}[\mathcal{E}(\mathbf{w}_t)]\}_t$ converges to a nonnegative value, which together with the assumption $\sum_{t=1}^{\infty}\eta_t^2 < \infty$, shows that the right-hand side of (18) vanishes to zero as $t \to \infty$, while the left-hand side is a positive number. This leads to a contradiction and $\limsup_{t\to\infty}\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2] = 0$. □

Proof of Corollary 5: Since $\theta > 1/(1+\alpha)$, we know $\sum_{t=1}^{\infty}\eta_t^{1+\alpha} = \eta_1^{1+\alpha}\sum_{t=1}^{\infty}t^{-\theta(1+\alpha)} < \infty$.

Equation (6) and

$$\frac{1}{1-\gamma}[(T+1)^{1-\gamma}-1] \leq \sum_{t=1}^{T}t^{-\gamma} \leq \frac{1}{1-\gamma}T^{1-\gamma}, \quad \gamma \in (0,1)$$

immediately imply $\min_{t=1,\ldots,T}\mathbb{E}[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2] = O(T^{\theta-1})$. Part 2) can be proved analogously and we omit the proof. □

B. Proof of Theorem 8

Lemma 11 [8]: Let $\{\eta_t\}_{t\in\mathbb{N}}$ be a sequence of nonnegative numbers such that $\lim_{t\to\infty}\eta_t = 0$ and $\sum_{t=1}^{\infty}\eta_t = \infty$. Let $\alpha, a > 0$

and $t_1 \in \mathbb{N}$ such that $\eta_t < a^{-1}$ for any $t \geq t_1$. Then we have $\lim_{T\to\infty} \sum_{t=t_1}^{T} \eta_t^{1+\alpha} \prod_{k=t+1}^{T}(1-a\eta_k) = 0$.

Proof of Theorem 8: We first prove Part 1). We introduce $B_t := \mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - \mathcal{E}(\mathbf{w}^*), \forall t \in \mathbb{N}$. By (11) and Assumption 7

$$\mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] \leq (1 + L^2\eta_t^{1+\alpha})\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - 2\mu\eta_t\big(\mathbb{E}[\mathcal{E}(\mathbf{w}_t) - \mathcal{E}(\mathbf{w}^*)]\big) + L^2(1-\alpha)\eta_t^{1+\alpha}.$$

Subtracting $\mathcal{E}(\mathbf{w}^*)$ from both sides gives

$$\begin{aligned}\mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] - \mathcal{E}(\mathbf{w}^*) &\leq (1 + L^2\eta_t^{1+\alpha})(\mathcal{E}(\mathbf{w}_t) - \mathcal{E}(\mathbf{w}^*)) + L^2\eta_t^{1+\alpha}\mathcal{E}(\mathbf{w}^*) \\ &\quad - 2\mu\eta_t(\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - \mathcal{E}(\mathbf{w}^*)) + L^2(1-\alpha)\eta_t^{1+\alpha} \\ &= (1 + L^2\eta_t^{1+\alpha} - 2\mu\eta_t)(\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - \mathcal{E}(\mathbf{w}^*)) + C_5\eta_t^{1+\alpha}\end{aligned}$$

where we introduce $C_5 := L^2\big(\mathcal{E}(\mathbf{w}^*) + 1 - \alpha\big)$. The assumption $\sum_{t=1}^{\infty} \eta_t^{1+\alpha} < \infty$ implies $\lim_{t\to\infty} \eta_t = 0$, which further implies the existence of t_1 such that $\eta_t^{\alpha} \leq \mu/L^2$ and $\eta_t \leq 1/\mu$ for all $t \geq t_1$. Therefore, it follows that

$$B_{t+1} \leq (1 - \mu\eta_t)B_t + C_5\eta_t^{1+\alpha} \quad \forall t \geq t_1. \tag{19}$$

A recursive application of this inequality then shows

$$B_{T+1} \leq \prod_{t=t_1}^{T}(1-\mu\eta_t)B_{t_1} + C_5\sum_{t=t_1}^{T}\eta_t^{1+\alpha}\prod_{k=t+1}^{T}(1-\mu\eta_k) \tag{20}$$

where we denote $\prod_{k=T+1}^{T}(1-\mu\eta_k) = 1$. The first term of (20) can be estimated by the standard inequality $1 - a \leq \exp(-a)$ for $a > 0$ together with the assumption $\sum_{t=1}^{\infty}\eta_t = \infty$ as

$$\prod_{t=t_1}^{T}(1-\mu\eta_t)B_{t_1} \leq \exp\left(-\mu\sum_{t=t_1}^{T}\eta_t\right)B_{t_1} \xrightarrow[T\to\infty]{} 0. \tag{21}$$

An application of Lemma 11 with $a = \mu$ then shows that

$$\lim_{T\to\infty}\sum_{t=t_1}^{T}\eta_t^{1+\alpha}\prod_{k=t+1}^{T}(1-\mu\eta_k) = 0. \tag{22}$$

Combining (20), (21), and (22) together shows

$$\lim_{T\to\infty}\mathbb{E}[\mathcal{E}(\mathbf{w}_T)] = \mathcal{E}(\mathbf{w}^*). \tag{23}$$

According to Part 2) of Theorem 3, we know that $\{\mathcal{E}(\mathbf{w}_t)\}_t - \mathcal{E}(\mathbf{w}^*)$ converges to a random variable $\widetilde{X}$ a.s., which is nonnegative by the definition of $\mathbf{w}^*$. This together with Fatou's lemma and (23) implies that

$$\mathbb{E}[\widetilde{X}] = \mathbb{E}[\lim_{t\to\infty}\mathcal{E}(\mathbf{w}_t)] - \mathcal{E}(\mathbf{w}^*) \leq \liminf_{t\to\infty}\mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - \mathcal{E}(\mathbf{w}^*) = 0.$$

Since $\widetilde{X}$ is nonnegative, we have $\lim_{t\to\infty}\mathcal{E}(\mathbf{w}_t) = \mathcal{E}(\mathbf{w}^*)$ a.s. Let $\phi(\mathbf{w}) = \mathcal{E}(\mathbf{w}) - \mathcal{E}(\mathbf{w}^*)$. It is clear that $\phi(\mathbf{w})$ satisfies (3) and is nonnegative. Now, we can apply Lemma 2 to show $\alpha\|\nabla\phi(\mathbf{w})\|_2^{(1+\alpha/\alpha)} \leq (1+\alpha)L^{\frac{1}{\alpha}}\phi(\mathbf{w})$, from which we know

$$\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^{\frac{1+\alpha}{\alpha}} \leq \frac{(1+\alpha)L^{\frac{1}{\alpha}}}{\alpha}[\mathcal{E}(\mathbf{w}_t) - \mathcal{E}(\mathbf{w}^*)] \quad \forall t \in \mathbb{N}.$$

This together with $\lim_{t\to\infty}\mathcal{E}(\mathbf{w}_t) = \mathcal{E}(\mathbf{w}^*)$ a.s. immediately implies $\lim_{t\to\infty}\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2 = 0$ a.s. This proves Part 1).

We now prove Part 2). It is clear from the definition of t_0 that $L^2\eta_t^{1+\alpha} \leq \mu\eta_t$ for all $t \geq t_0$. Therefore, (19) holds with $t_1 = t_0$. Taking $\eta_t = 2/(\mu(t+1))$ in (19), we derive

$$B_{t+1} \leq \frac{t-1}{t+1}B_t + C_5\left(\frac{2}{(t+1)\mu}\right)^{1+\alpha} \quad \forall t \geq t_0. \tag{24}$$

Multiplying both sides of (24) with $t(t+1)$ gives

$$t(t+1)B_{t+1} \leq t(t-1)B_t + C_5(2\mu^{-1})^{1+\alpha}t(t+1)^{-\alpha} \quad \forall t \geq t_0.$$

Taking a summation from $t = t_0$ to $t = T$ gives

$$T(T+1)B_{T+1} \leq t_0(t_0-1)B_{t_0} + C_5(2\mu^{-1})^{1+\alpha}\sum_{t=t_0}^{T}t(t+1)^{-\alpha}.$$

It is clear that

$$\sum_{t=t_0}^{T}t(t+1)^{-\alpha} \leq \sum_{t=t_0}^{T}t^{1-\alpha} \leq \sum_{t=t_0}^{T}\int_t^{t+1}x^{1-\alpha}dx \leq \frac{(T+1)^{2-\alpha}}{2-\alpha}$$

from which and $(T+1)/T \leq (1+t_0^{-1})$ for all $T \geq t_0$, we derive the following inequality for all $T \geq t_0$:

$$\begin{aligned}B_{T+1} &\leq \frac{t_0(t_0-1)B_{t_0}}{T(T+1)} + \frac{C_5(2\mu^{-1})^{1+\alpha}(T+1)^{1-\alpha}}{(2-\alpha)T} \\ &\leq \frac{t_0(t_0-1)B_{t_0}}{T(T+1)} + \frac{(1+t_0^{-1})^{1-\alpha}C_5(2\mu^{-1})^{1+\alpha}}{(2-\alpha)T^{\alpha}}.\end{aligned}$$

This gives the stated result with

$$\widetilde{C} = (t_0-1)(\mathbb{E}[\mathcal{E}(\mathbf{w}_{t_0})] - \mathcal{E}(\mathbf{w}^*)) + \frac{(1+t_0^{-1})^{1-\alpha}C_5(2\mu^{-1})^{\frac{1}{\alpha+1}}}{2-\alpha}.$$

We now consider Part 3). Analogous to (8), we derive

$$\mathcal{E}(\mathbf{w}_{t+1}) \leq \mathcal{E}(\mathbf{w}_t) - \eta\langle\nabla f(\mathbf{w}_t, z_t), \nabla\mathcal{E}(\mathbf{w}_t)\rangle + 2^{-1}L\eta^2\|\nabla f(\mathbf{w}_t, z_t)\|_2^2. \tag{25}$$

Since $\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|_2^2] = 0$, we know $\nabla f(\mathbf{w}^*, z) = 0$ a.s. Therefore, $\mathbf{w}^*$ is a minimizer of the function $\mathbf{w} \mapsto f(\mathbf{w}, z)$ for almost every z, and the function $\phi_z(\mathbf{w}) = f(\mathbf{w}, z) - f(\mathbf{w}^*, z)$ is nonnegative a.s. We can apply Lemma 2 to show $\|\nabla\phi_z(\mathbf{w})\|_2^2 \leq 2L\phi_z(\mathbf{w})$ a.s., which is equivalent to $\|\nabla f(\mathbf{w}, z)\|_2^2 \leq 2L(f(\mathbf{w}, z) - f(\mathbf{w}^*, z))$ a.s. Plugging this inequality back to (25) gives the following inequality a.s.:

$$\mathcal{E}(\mathbf{w}_{t+1}) \leq \mathcal{E}(\mathbf{w}_t) - \eta\langle\nabla f(\mathbf{w}_t, z_t), \nabla\mathcal{E}(\mathbf{w}_t)\rangle + L^2\eta^2(f(\mathbf{w}_t, z_t) - f(\mathbf{w}^*, z_t)).$$

Taking expectation over both sides then gives

$$\mathbb{E}[\mathcal{E}(\mathbf{w}_{t+1})] - \mathcal{E}(\mathbf{w}^*) \leq \mathbb{E}[\mathcal{E}(\mathbf{w}_t)] - \mathcal{E}(\mathbf{w}^*) - \mathbb{E}\big[\|\nabla\mathcal{E}(\mathbf{w}_t)\|_2^2\big] + L^2\eta^2\mathbb{E}[\mathcal{E}(\mathbf{w}_t) - \mathcal{E}(\mathbf{w}^*)].$$

It follows from Assumption 7 and $\eta \leq \mu/L^2$ that

$$B_{t+1} \leq B_t - 2\mu\eta B_t + L^2\eta^2 B_t \leq (1-\mu\eta)B_t.$$

Applying this result recursively gives the stated result. □

ACKNOWLEDGMENT

The authors would like to thank the editor and referees for their helpful comments.

REFERENCES

[1] B. T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal Vychislitel Matematiki Matematicheskoi Fiziki*, vol. 3, no. 4, pp. 643–653, 1963.

[2] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2004, pp. 919–926.

[3] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018. [Online]. Available: http://leon.bottou.org/papers/bottou-curtis-nocedal-2018

[4] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition," in *Proc. Eur. Conf. Mach. Learn.*, Sep. 2016, pp. 795–811.

[5] S. Ghadimi and G. H. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013.

[6] S. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 314–323.

[7] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Math. Program.*, vol. 155, nos. 1–2, pp. 267–305, Jan. 2016.

[8] Y. Ying and D. X. Zhou, "Online regularized classification algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, Nov. 2006.

[9] Y. Ying and D. X. Zhou, "Unregularized Online learning algorithms with general loss functions," *Appl. Comput. Harmon. Anal.*, vol. 42, no. 2, pp. 224–244, Mar. 2017.

[10] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima," *J. Mach. Learn. Res.*, vol. 16, pp. 559–616, Mar. 2015.

[11] D. Chang, M. Lin, and C. Zhang, "On the generalization ability of Online gradient descent algorithm under the quadratic growth condition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 5008–5019, Oct. 2018.

[12] D. J. Foster, A. Sekhari, and K. Sridharan, "Uniform convergence of gradients for non-convex learning and optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8759–8770.

[13] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for Online convex optimization," *Mach. Learn.*, vol. 69, nos. 2–3, pp. 169–192, 2007.

[14] J. Lin and D.-X. Zhou, "Online learning algorithms can converge comparably fast as batch learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2367–2378, Jun. 2018.

[15] T. Hu and D.-X. Zhou, "Online learning with samples drawn from non-identical distributions," *J. Mach. Learn. Res.*, vol. 10, pp. 2873–2898, Dec. 2009.

[16] S.-B. Lin and D.-X. Zhou, "Distributed kernel-based gradient descent algorithms," *Constructive Approximation*, vol. 47, no. 2, pp. 249–276, 2018.

[17] H. Sun and Q. Wu, "Sparse Representation in Kernel Machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2576–2582, Oct. 2015.

[18] Y. Lei and K. Tang, "Stochastic composite mirror descent: Optimal bounds with high probabilities," in *Proc. Advance Neural Inf. Process. Syst.*, 2018, pp. 1524–1534.

[19] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM J. Optim.*, vol. 10, no. 3, pp. 627–642, 2000.

[20] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, and Y. Zhang, "A fast uyghur text detector for complex background images," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3389–3398, Dec. 2018.

[21] Y. Lei, L. Shi, and Z.-C. Guo, "Convergence of unregularized online learning algorithms," *J. Mach. Learn. Res.*, vol. 18, no. 171, pp. 6269–6301, Jan. 2018.

[22] L. Nguyen, P. Nguyen, M. van Dijk, P. Richtárik, K. Scheinberg, and M. Takáč, "SGD and hogwild! Convergence without the bounded gradients assumption," in *Proc. Int. Conf. Mach. Learn.*, Feb. 2018, pp. 3747–3755.

[23] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 315–323.

[24] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Math. Program.*, vol. 162, nos. 1–2, pp. 83–112, 2017.

[25] Z. Allen-Zhu, "Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 89–97.

[26] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *J. Mach. Learn. Res.*, vol. 18, no. 221, pp. 8194–8244, Jan. 2018.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, Art. no. 27.

[28] J. L. Doob, *Measure Theory*. New York, NY, USA: Springer, 1994.