Journal of Machine Learning Research 24 (2023) 1-49 Submitted 12/19; Revised 8/22; Published 4/23

On the Convergence of Stochastic Gradient Descent with Bandwidth-based Step Size

Xiaoyu Wang WXY@LSEC.CC.AC.CN
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190, China
University of Chinese Academy of Sciences
No.19A Yuquan Road, Beijing 100049, China

Ya-xiang Yuan YYX@LSEC.CC.AC.CN
State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190, China

Editor: Simon Lacoste-Julien

Abstract

We first propose a general step-size framework for the stochastic gradient descent(SGD) method: bandwidth-based step sizes that are allowed to vary within a banded region. The framework provides efficient and flexible step size selection in optimization, including cyclical and non-monotonic step sizes (*e.g.*, triangular policy and cosine with restart), for which theoretical guarantees are rare. We provide state-of-the-art convergence guarantees for SGD under mild conditions and allow a large constant step size at the beginning of training. Moreover, we investigate the error bounds of SGD under the bandwidth step size where the boundary functions are in the same order and different orders, respectively. Finally, we propose a $1/t$ *up-down policy* and design novel non-monotonic step sizes. Numerical experiments demonstrate these bandwidth-based step sizes' efficiency and significant potential in training regularized logistic regression and several large-scale neural network tasks.

Keywords: stochastic gradient descent, bandwidth-based step size, non-asymptotic convergence, non-monotonic step size, machine learning

1. Introduction

We consider the following stochastic optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) = \mathbb{E}_{\xi\sim\Xi}[f(x;\xi)], \tag{1}$$

where ξ is a random variable drawn from an unknown source distribution Ξ and $f(x;\xi)$ is the instantaneous loss function over the variable $x \in \mathbb{R}^d$. This problem is often encountered in machine learning and statistics and attracts much attention along with big data and artificial intelligence. The corresponding empirical risk problem is to minimize $f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x;\xi_i)$, where each ξ_i $(i \in \{1, 2, \ldots, n\})$ denotes a realization of ξ.

The stochastic gradient descent (SGD) algorithm (Robbins and Monro, 1951) is widely used to solve the machine learning problem (1). The iterates of SGD are given by

$$x_{t+1} = x_t - \eta(t) g_t, \tag{2}$$

where $\eta(t) > 0$ is step size and the stochastic gradient g_t is an unbiased estimator of its true gradient $\nabla f(x_t)$(i.e., $\mathbb{E}[g_t \mid \mathcal{F}_t{}^1] = \nabla f(x_t)$). However, its performance is highly dependent on the choice of step size due to the natural noise from the stochastic gradient. In this paper, we investigate the performance of the SGD algorithm defined in (2) under a general class of step size (possibly non-monotonic).

1.1 Theoretical Analysis of SGD Under Various Step Sizes

The asymptotic results of SGD are given in (Chung, 1954; Leen and Orr, 1994; Leen et al., 1998). Leen and Orr (1994) analyzed the asymptotic properties around the locally optimal solution x^* with $\eta(t) = \eta_0/t$ and show that if $\eta_0 > 1/(2\lambda_{\min})$ ($\lambda_{\min}$ is the smallest eigenvalue of $\nabla^2 f(x^*)$), the error $\mathbb{E}[\|x_t - x^*\|^2]$ has order $\mathcal{O}(1/t)$, which is an optimal (minimax) rate (Polyak and Juditsky, 1992; Agarwal et al., 2009; Ghadimi and Lan, 2012).

Recently, the focus has been shifted to studying the non-asymptotic convergence results. Moulines and Bach (2011) established the convergence rate of SGD for a class of step sizes $\eta(t) = \eta_0/t^p$ for $p \in (0, 1]$. For strongly convex and L-smooth functions, SGD exhibits an optimal error bound $\mathcal{O}(1/T)$ (T is the total number of iterations) with $\eta(t) = \eta_0/t$ (Moulines and Bach, 2011; Rakhlin et al., 2012; Nguyen et al., 2019b). However, the results become complicated if the function is not L-smooth. The best known result on the last iterate is $\mathbb{E}[f(x_T) - f(x^*)] \le O(\log T/T)$ with $\eta(t) = 1/(\mu t)$ (Shamir and Zhang, 2013), which is proved to be tight by (Harvey et al., 2019). Many averaging techniques such as suffix averaging (Rakhlin et al., 2012) and polynomial-decay averaging (Shamir and Zhang, 2013; Lacoste-Julien et al., 2012) are incorporated into SGD and obtain an optimal $\mathcal{O}(1/T)$ rate. Hazan and Kale (2014) achieved an $\mathcal{O}(1/T)$ convergence rate by exponentially decreasing the step size after a consecutive period which grows exponentially, and adopting a simple modification where the inner iterations are averaged as an output. Jain et al. (2019) designed the piece-wise decay step size with the form of $\mathcal{O}(1/t)$ per period and obtained an optimal error bound $\mathbb{E}[f(x_T) - f(x^*)] \le O(1/T)$ on the last iterate. But for non-smooth problems, these papers rely on the uniform boundedness of stochastic gradient (*i.e.*,$\mathbb{E}[\|g_t\|^2] \le G^2$). This restricts the trajectory of the iterates to be bounded (see Section 2 for details).

The step decay schedule (constant and then cut) has attracted much interest due to its excellent performance in training deep neural networks (Ge et al., 2019; Li et al., 2021). Ge et al. (2019) analyzed a step decay step size which decays exponentially after $T/\log T$ iterations and achieved a near-optimal $\mathcal{O}(\log T/T)$ convergence rate for least squares problems. Li et al. (2021) proposed a continuous step decay schedule and proved a near-optimal convergence rate under the Polyak-Łójasiewicz condition and smoothness.

To the best of our knowledge, there are many other efficient (possibly non-monotonic) step sizes preferred in deep learning, *e.g.*, adaptive methods (Duchi et al., 2011; Tieleman and Hinton, 2012; Zeiler, 2012; Kingma and Ba, 2015; Loizou et al., 2021), Barzilai-Borwein based (Tan et al., 2016; Yang et al., 2018), line-search based (Keskar and Saon, 2015;

[1] We use $\mathcal{F}_t$ to denote σ-algebra of the random information at iteration t.

Vaswani et al., 2019b), cyclical learning rate (step size) (Smith, 2017; Loshchilov and Hutter, 2017; An et al., 2017). Some recent works (Oymak, 2021; Goujaud et al., 2022) show that the gradient-based algorithms under cyclical step sizes have a faster convergence for a class of functions whose Hessian spectrum has special structures.

1.2 Motivation

In this paper, we focus on the non-asymptotic convergence of the SGD method in which the step size $\{\eta(t)\}$ varies in a bounded region rather than any fixed schedules. The lower and upper bounds of the region are defined by two monotonic but non-increasing functions $\delta_1(t)$ and $\delta_2(t)$ w.r.t. the iteration number t. More specifically, we assume there exist two positive constants $m \leq M$ such that

$$m\delta_1(t) \leq \eta(t) \leq M\delta_2(t), \forall\, t \geq 1, \tag{BD}$$

and $d\delta_1(t)/dt \leq 0$ and $d\delta_2(t)/dt \leq 0$. Especially, when $\delta_1(t) = \delta_2(t) = 1/t$, we call it *$1/t$-band*. Such an idea is originally motivated by the piece-wise decay and step-decay step sizes (Hazan and Kale, 2014; Jain et al., 2019; Ge et al., 2019), which is a step function whose graph consists of some line segments lying within two curves (*i.e.*, their lower and upper bounds). The diminishing step size $\eta(t) = \eta_0/t$, piece-wise decay step size proposed by Jain et al. (2019), and step-decay step size in Hazan and Kale (2014) can be regarded as the special cases of $1/t$-band.

Dauphin et al. (2014) pointed out that a great obstacle to minimizing deep neural networks with high possibility arose from saddle points instead of poor local minima. The proposed non-monotonic scheduling (BD), admitting some intermediate increase in step size, might help rapidly traverse the saddle points and find flat minima. Smith (2017) described a type of cyclical learning rate (step size) that varied within a band of minimum and maximum values and showed the potential benefits of training deep neural networks. Similarly, An et al. (2017) proposed a sine-wave learning rate framework. Their boundaries decay exponentially after a few fixed epochs. The policy lets the step size locally vary within a reasonable band. Although their mechanisms might have a short-term negative effect, it is beneficial overall.

Another motivation comes from the constant step size, which achieves linear convergence to the neighborhood of the optimal solution with constant noise (Gower et al., 2019). As long as the iterates are not diverging, a relatively larger constant step size can achieve a faster linear convergence at the beginning but finally lead to a higher noise error (also see Corollary 1 with $M = m$). One intuition is that we reach the upper bound of the bandwidth step size in the early stages of training to speed up the convergence and then drop the step size to touch the lower bound to reduce the noise error. We give a simple example in Corollary 1 to address how this bandwidth framework can be useful.

We are interested in the class of bandwidth-based step size described in (BD), which gives us a lot of freedom and a novel insight to design more efficient step sizes in practice. Although many specific and effective schedules are mentioned in Section 1.1, it is still a very interesting and challenging topic to analyze the convergence properties of the SGD method based on such a generic class of step size. Moreover, some popular step sizes, *e.g.*, cyclical learning rate (Smith, 2017; Loshchilov and Hutter, 2017; An et al., 2017), perform well in practice but lack non-asymptotic convergence guarantees. To overcome these limitations, we explore their connections in theory and practice using the novel bandwidth-based step size framework (BD).

1.3 Main Contributions

Inspired by the above potential benefits of this bandwidth-based framework, we are the first to provide uniform convergence guarantees of SGD for strongly convex problems under different classes of bandwidth step sizes and make the following contributions:

First, we explore a class of step sizes lying in a bandwidth-based region to achieve state-of-the-art results $\mathbb{E}[\|x_{T+1} - x^*\|^2]$ or $\mathbb{E}[f(\hat{x}_T{}^2) - f(x^*)] \leq \mathcal{O}(1/T)$ on strongly convex problems. The main results are briefly summarized in Table 1 and will come in Section 3 afterward. Specifically,

- We extend the typical $1/t$ *stepsize* (*i.e.*, $\eta(t) = \eta_0/t$) to $1/t$-band which allows step size to vary locally in any way and covers some interesting modes *e.g.*, constant and then cut (Hazan and Kale, 2014), triangular (Smith, 2017), and sine-wave (An et al., 2017). The convergence results for $1/t$-band are comparable to those of Moulines and Bach (2011), Shamir and Zhang (2013) and Lacoste-Julien et al. (2012) for $\mathcal{O}(1/t)$ step size. However, throughout the paper, we use a weak growth condition with noise $\mathbb{E}[\|g_t\|^2] \leq 4L_f(f(x_t) - f^*) + 2\sigma^2$ where $L_f > 0$ is a constant (Nguyen et al., 2018; Vaswani et al., 2019a), which is milder than the traditional L-smoothness (Moulines and Bach, 2011), and the individual function is not needed to be convex.

- We relax the lower bound of $1/t$-band, which is on average greater than (or equal to) $\mathcal{O}(1/t)$. We also prove that the lower bound is essential to achieve an $\mathcal{O}(1/T)$ rate if the upper bound is M/t. This covers the piece-wise decay step size proposed by Jain et al. (2019). A relatively large step size often performs well at the beginning of training from both theory and practice (Gower et al., 2019). We turn to extend the upper bound of $1/t$-band and provide theoretical guarantees for the policy which allows the constant step size in initial iterations (see the last case of Table 1). Our results demonstrate that there are wide classes of step sizes that can achieve a state-of-the-art $\mathcal{O}(1/T)$ rate beyond the classic $1/t$-stepsize.

- In particular, the cyclical step sizes developed in Smith (2017); An et al. (2017); Loshchilov and Hutter (2017), which lack convergence guarantees, can achieve the optimal $\mathcal{O}(1/T)$ and near-optimal convergence rates by properly choosing their boundaries. We elaborate on the applications of the bandwidth framework for the cyclical step sizes in Section 3.1.

Second, we provide unified worst-case convergence guarantees for a class of bandwidth step size that $\delta_1(t) = \delta_2(t) = \delta(t)$. The results are briefly shown in Table 2 and will be given in Section 4. Especially, in the degenerated case that $\eta(t) = 1/t^p$ ($p \in (0, 1]$), our result is comparable to those in the prior literature (Moulines and Bach, 2011), while we

[2] Here $\hat{x}_T$ is a type of averaging of the previous iterations x_t from $t = 1, 2, \cdots, T$.

[3] We use $[C_1 T^p]$ to denote a positive integer set from 1 to $C_1 T^p$ where $C_1 > 0$ is a constant and $p \in (0, 1)$. This notation is also suited for $[T]$

$\delta_1(t)$	$\delta_2(t)$	Theorem
$1/t$	$1/t$	1
$\sum_{t=t^*}^{T} \delta_1(t) \geq \mathcal{O}\left(\ln\left((T+1)/t^*\right)\right)$	$1/t$	3
1, for $t \in [C_1 T^p]^3$ $1/t$, for $t \in [T] \backslash [C_1 T^p]$	1, for $t \in [C_1 T^p]$ $1/t$, for $t \in [T] \backslash [C_1 T^p]$	4

Table 1: The bandwidth step sizes in Section 3 to achieve $\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \mathcal{O}(1/T)$

use the weak growth condition on the stochastic gradient instead of L-smoothness for each individual function. When $\lim_{t\to\infty} \delta(t)t = 0$, our result is novel. This includes the case that $\eta(t) = \mathcal{O}(1/(t \ln t))$ that has not been discussed before. In this analysis, we add a new condition $-d\delta(t)/dt \leq c_1 \delta(t)^2$ which clarifies "in the most general case" mentioned in Nguyen et al. (2019a) and we give a more rigorous proof. Moreover, our analysis can provide better upper bounds in some cases, such as $\eta(t) = 1/\sqrt{t}$ and $1/(t\log(t))$ than those of theorem 10 in Nguyen et al. (2019a).

Conditions		$\mathbb{E}[\|x_{T+1} - x^*\|^2]$	Theorem
$\delta_1(t) = \delta_2(t)$ $= \delta(t)$	$\lim_{t\to\infty} t\delta(t) = 1$	$\mathcal{O}(1/T^{\mu m}) + \mathcal{O}(1/T)$	1
	$\lim_{t\to\infty} t\delta(t) = 0$	$\mathcal{O}(\exp(-\mu m \sum_{t=1}^{T} \delta(t)))$	5
	$\lim_{t\to\infty} t\delta(t) = \infty$	$\mathcal{O}(\delta(t)) + \mathcal{O}(\exp(-\mu m \sum_{t=1}^{T} \delta(t)))$	
$\delta_1(t) \neq \delta_2(t)$	$\delta_1(t) = 1/t$ $\delta_2(t) = \log(t)/t$	$\mathcal{O}(\log^2(T)/T)$	6
	$\delta_1(t) = 1/t$ $\delta_2(t) = 1/t^\alpha$	$\mathcal{O}(1/T^{2\alpha-1})$	7
	$\delta_1(t) = 1/(t\log(t))$ $\delta_2(t) = 1/t^\alpha$	$\mathcal{O}(1/\log(T)^{\mu m})$	8

Table 2: A brief summary of convergence results in Sections 4 and 5 where μ is the strongly convexity parameter and $\alpha \in (1/2, 1]$.

Third, we also discuss the cases of the lower and upper bounds being in different orders (i.e., $\delta_1(t) \neq \delta_2(t)$), listed in Table 2, and the main results are given in Section 5. The theoretical results explore the connections between the band and its boundaries and broaden the boundaries of the step size for analyzing the convergence behaviors of SGD.

Finally, we propose a $1/t$ up-down policy and design four non-monotonic step sizes including $1/t$ Fix-period, $1/t$ Grow-period, $1/t$ Grow-Exp, and $1/t$ Fix-Exp. The proposed bandwidth step size, *e.g.*, $1/t$ Fix-period and $1/t$ Grow-period, have potential benefits due to the larger enclosed area of their graph compared to their baseline (see Remark 6).

- We test regularized logistic regression and some nonconvex problems (*e.g.*, deep neural networks, VGG-16 (Simonyan and Zisserman, 2015) and ResNet-18 (He et al., 2016)) on the real data sets (MNIST, CIFAR-10, and CIFAR-100). Numerical experiments demonstrate the efficiency of these bandwidth step sizes compared to their baselines: $\eta(t) = \eta_0/t$ and exponential decaying step size (Hazan and Kale, 2014), respectively.
- We implement the bandwidth-based step sizes with other default algorithms in deep learning, *e.g.*, averaged SGD (Polyak and Juditsky, 1992), SGD with momentum (Polyak, 1964; Sutskever et al., 2013) and Adam (Kingma and Ba, 2015). The results show that the proposed $1/t$ up-down policy and these step sizes also work for averaged SGD and momentum acceleration. Moreover, we compare the proposed step size strategies to other popular step sizes, such as triangular policy (Smith, 2017) and cosine annealing (Loshchilov and Hutter, 2017). A great potential is shown when the step size satisfies the bandwidth, especially for nonconvex optimization.

Organization: in Section 2, we present some necessary definitions and lemmas used in the downstream analysis. In Section 3, we investigate the conditions for the bandwidth-based step size of SGD to achieve the $\mathcal{O}(1/T)$ convergence rate. Section 4 discusses the scenario where the ending points of the bandwidth step size are in the same order, which covers most cases we met. Section 5 considers the situation where the bands have different lower and upper boundaries. In Section 6, we perform numerical experiments based on bandwidth for the proposed step sizes. Then we make a conclusion in Section 7.
Notation. Let x^* be the unique minimizer of f, that is $x^* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. We use $\mathcal{F}_t$ to denote σ-algebra of the random information at t-th iteration. In default, the expectation is taken with respect to the source distribution Ξ, that is $\mathbb{E}[\cdot] = \mathbb{E}_\Xi[\cdot] := \mathbb{E}_{\xi\sim\Xi}[\cdot]$. Other notations include: $\|\cdot\| := \|\cdot\|_2$; $[n] = \{1, 2, \ldots, n\}$; $[n]\backslash[n_1] = \{n_1 + 1, n_1 + 2, \ldots, n\}$ for any $n_1 < n \in \mathbb{N}$.

2. Preliminaries

This part will give some definitions and basic lemmas used in the later sections.

Assumption 1 (μ-strongly convex) *The objective function $f(\cdot) : \mathbb{R}^d \longmapsto \mathbb{R}$ is μ-strongly convex if there exists a constant $\mu > 0$ such that*

$$f(x) - f(\hat{x}) \geq \langle \nabla f(\hat{x}), x - \hat{x} \rangle + \frac{\mu}{2} \|x - \hat{x}\|^2, \tag{3}$$

for all $x, \hat{x} \in \mathbb{R}^d$.

Note that $f(x;\xi)$ for each ξ is not guaranteed convex even when we assume that $f(x)$ is μ-strongly convex.

Assumption 2 *(Unbiased gradient estimator) For any input vector x, the stochastic gradient oracle returns a vector $g \in \mathbb{R}^d$ such that $\mathbb{E}[g] = \nabla f(x)$.*

Next, we assume the stochastic gradient g_t of the SGD formula in (2) satisfies the following assumption, which is a direct consequence of expected smoothness (Gower et al.,

2020) if the noise $\mathbb{E}[\|\nabla f(x^*;\xi)\|^2)]$ for each $\xi \in \Xi$ is finite. When $\sigma = 0$, the inequality (4) is known as the weak growth condition (Vaswani et al., 2019a). We may also call it a weak growth condition with noise.

Assumption 3 *There exists a constant $L_f > 0$ such that*

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

The definition of expected smoothness in Gower et al. (2020) is about the individual functions *w.r.t* a distribution. Here we make the assumption on the stochastic gradient g_t, which is usually computed by some mini-batch strategies on the individual functions.

Uniformly bounded gradient. The assumption of uniformly bounded gradient (*i.e.*, $\mathbb{E}[\|g_t\|^2] \leq G^2$ for some fixed $G > 0$) is used in some recent papers (Shamir and Zhang, 2013; Rakhlin et al., 2012; Hazan and Kale, 2014; Jain et al., 2019). However, this is clearly false if f is strongly convex, which has been pointed out by Nguyen et al. (2018); Leblond et al. (2018). If f is μ-strongly convex and $\mathbb{E}[\|g_t\|^2] \leq G^2$, by *Jensen inequality* in expectation that $\|\mathbb{E}[X]\|^2 \leq \mathbb{E}[\|X\|^2]$, we have

$$\mu^2 \|x_t - x^*\|^2 \leq 2\mu(f(x_t) - f(x^*)) \leq \|\nabla f(x_t)\|^2 = \|\mathbb{E}[g_t]\|^2 \leq \mathbb{E}[\|g_t\|^2] \leq G^2.$$

In this case, $f(x_t) - f(x^*)$ and $\|x_t - x^*\|^2$ should be bounded on the whole space $\mathbb{R}^d$. However, this leads to a contradiction when $\|x_t - x^*\|$ is sufficiently large. Thus we assume the stochastic gradient of SGD satisfies Assumption 3 (Gower et al., 2020; Nguyen et al., 2018) rather than uniformly bounded.

L-smooth property vs expected smoothness. Suppose that f is μ-strongly convex. By (5), the L-smooth property used in Moulines and Bach (2011)

$$\|\nabla f(x;\xi) - \nabla f(x^*;\xi)\|^2 \leq L^2 \|x - x^*\|^2 \leq \frac{2L^2}{\mu}[f(x) - f^*], \tag{5}$$

implies expected smoothness with $L_f = L^2/\mu$ (assume $\mathbb{E}[\|\nabla f(x^*;\xi)\|^2]$ is a finite constant), but the opposite does not hold (see Nguyen et al. (2019a)). Moreover, if f is convex and L-smooth, the expected smoothness assumption can be satisfied with $L_f = 2L$, but the opposite is not true. Indeed, example 2.2 of Gower et al. (2019) shows that Assumption 3 holds even when $f(x;\xi)$ or f is not convex.

Lemma 1 *Suppose that f is μ-strongly convex then*

$$\langle \nabla f(x), x - x^* \rangle \geq f(x) - f(x^*) + \frac{\mu}{2}\|x - x^*\|^2, \textit{ for } x \in \mathbb{R}^d. \tag{6}$$

All proofs of the lemmas in this section are provided in Appendix A.

Lemma 2 *Suppose that the objective function f satisfies Assumption 1. Considering the SGD method defined by (2) where the stochastic gradient g_t satisfies Assumptions 2 and 3, we have $\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t]$ is at most*

$$(1 - \mu\eta(t)) \|x_t - x^*\|^2 + 2\eta(t)^2\sigma^2 + (4L_f\eta(t)^2 - 2\eta(t))[f(x_t) - f(x^*)]. \tag{7}$$

Besides, let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$ *and* $f_{n_0} := \max_{1 \le t \le n_0} \{f(x_t) - f(x^*)\}$*. If* n_0 *is a finite constant and is independent of* T *(the budget of the iteration* t*), then for* $t > n_0$*, we have* $\mathbb{E}[\|x_{t+1} - x^*\|^2]$ *is at most*

$$\exp\left(-\mu\sum_{l=1}^{t}\eta(l)\right)\Delta_{n_0}^0 + 2\sigma^2\sum_{l=1}^{t}\eta(l)^2\exp\left(-\mu\sum_{u>l}^{t}\eta(u)\right), \tag{8}$$

where $\Delta_{n_0}^0 = \|x_1 - x^*\|^2 + \frac{n_0\chi_{n_0}f_{n_0}}{\exp(-\mu\sum_{l=1}^{n_0}\eta(l))}$ *and* $\chi_{n_0} = \max_{1 \le t \le n_0}\left\{4L_f\eta(t)^2 - 2\eta(t)\right\}$*.*

In Lemma 2, we propose a unified analysis framework for the SGD algorithm under strong convexity and weak growth condition with noise which is milder than L-smooth used in Moulines and Bach (2011). Different from Gower et al. (2019); Nguyen et al. (2019a), in Lemma 2, we do not require the step size to be equal to or smaller than $1/(2L_f)$ for all iterations but allow it to be larger than $1/(2L_f)$ at the initial and finite n_0 iterations. Therefore, such a framework is more flexible in dealing with general situations. For instance, $\eta(t) = 1/(\mu t)$ is larger than $1/(2L_f)$ in the first few iterations. We will address the motivation of introducing n_0 in Remark 1.

Remark 1 **(Justification of** n_0**)** *In the strongly convex case, the optimal rate* $\mathcal{O}(1/T)$ *of SGD can be achieved when the step size* $\eta(t) = 1/(\mu t)$ *(Moulines and Bach, 2011; Shamir and Zhang, 2013). However, this step size may not satisfy* $\eta(t) \le 1/(2L_f)$ *at the first few iterations* $n_0 = 2L_f/\mu > 1$*. To avoid this potential conflict, Lemma 2 allows the step size to be larger than* $1/(2L_f)$ *at the first* n_0 *iterations and assumes that* n_0 *is a finite constant and independent on* T*. The restriction on* n_0 *can be easily guaranteed by the commonly used step sizes. For example, the polynomial diminishing step size* $\eta(t) = \eta_0/t^p$ *(*$p \in (0,1]$*), which finally decreases to zero, obviously satisfies the restriction of* n_0 *when* $n_0 = \lceil(2\eta_0 L_f)^{1/p}\rceil$ *with sufficient large* $T \ge n_0$*.*

Remark 2 *For simplicity, let*

$$\Gamma_T^1 := \exp\left(-\mu\sum_{l=1}^{T}\eta(l)\right)\Delta_{n_0}^0, \tag{9a}$$

$$\Gamma_T^2 := 2\sigma^2\sum_{l=1}^{T}\eta(l)^2\exp\left(-\mu\sum_{u>l}^{T}\eta(u)\right). \tag{9b}$$

From Lemma 2, let $t = T$*, we have*

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \le \Gamma_T^1 + \Gamma_T^2. \tag{10}$$

Based on (10), the upper bound of $\mathbb{E}[\|x_{T+1} - x^*\|^2]$ *is divided into two parts* Γ_T^1 *and* Γ_T^2*. Once the summation* $\sum_{l=1}^{T}\eta(l)$ *is evaluated,* Γ_T^1 *can be easily estimated by (9a). Therefore, the challenge of the following analysis for different bandwidth step sizes falls on the evaluation of* Γ_T^2*.*

As shown by Lemma 2, the error bound of $\mathbb{E}[\|x_{T+1} - x^*\|^2]$ is determined by two error terms Γ_T^1 and Γ_T^2. Clearly, the optimization error Γ_T^1 will decay faster for some larger step sizes whose graph encloses larger area (*i.e.*, $\sum_{t=1}^T \eta(t)$). In the bandwidth-based step size scenario, this implies that the upper bound $M\delta_2(t)$ is faster than the lower bound $m\delta_1(t)$ in reducing the optimization error Γ_T^1. However, this does not indicate that the upper bound $\eta(t) = M\delta_2(t)$ always performs better than the lower bound $\eta(t) = m\delta_1(t)$, especially when the noise error term Γ_T^2 leads the bound of (10). We give a specific example of the bandwidth constant step size ($\delta(t) = 1$) to show its advantages over the constant step size $\eta(t) = \eta > 0$.

Corollary 1 *(A motivating example of the bandwidth-based step size) Under the conditions of Lemma 2, we consider the step size $\eta(t) = M$ for $t \in [1, T/2)$ and $\eta(t) = m$ for $t \in [T/2, T]$ where $0 < m \leq M \leq 1/(2L_f)$, then for $T \geq 1$, we have $\mathbb{E}[\|x_{T+1} - x^*\|^2]$ is at most*

$$\exp\left(-\frac{\mu(m+M)T}{2}\right)\|x_1 - x^*\|^2 + \frac{2\sigma^2 m}{\mu} + 2\sigma^2\frac{M-m}{\mu}\exp\left(-\frac{\mu m T}{2}\right).$$

Let $M = m = \eta$, and then the above corollary recovers the result for constant step size, which linearly converges to the neighborhood of the solution with a constant noise (see theorem 3.1 of (Gower et al., 2019)). As we can see, a relatively large constant step size ($\eta = M$) is faster in reducing the optimization error but gives rise to a large noise error $2\sigma^2 M/\mu$. The example provided in Corollary 1 gives us the first intuition that we can choose the upper bound $\eta(t) = M$ at the beginning to achieve the faster convergence but finally reduce the noise error Γ_T^2 by hitting the lower bound $\eta(t) = m$.

3. Non-Asymptotic Analysis of SGD for An Optimal Convergence Rate

In this section, we will first analyze the non-asymptotic convergence rate of the classical SGD algorithm where the step size $\eta(t)$ satisfies the following conditions

(A) there exists a constant $m > 0$ such that $\eta(t) \geq \frac{m}{t}$,

(B) there exists a constant $M \geq m$ such that $\eta(t) \leq \frac{M}{t}$.

This is a special case of (BD) with $\delta_1(t) = \delta_2(t) = 1/t$. The step size under these conditions is more general and possibly non-monotonic compared with the common choice $\eta(t) = \eta_0/(a+t)$ (Rakhlin et al., 2012; Moulines and Bach, 2011; Shamir and Zhang, 2013; Lacoste-Julien et al., 2012; Bottou et al., 2018; Gower et al., 2019).

The natural questions arising are the convergence of SGD and, if the convergence holds, the corresponding convergence rate (*e.g.*, $\mathcal{O}(1/T)$ rate). It is easy to see that SGD converges under (A) and (B) since they satisfy the well-known conditions ($1'$) $\sum_{t=1}^\infty \eta(t) = \infty$ and ($2'$) $\sum_{t=1}^\infty \eta(t)^2 < \infty$ given by Robbins and Monro (1951). The remaining question is which cases can ensure that SGD obtains the optimal $\mathcal{O}(1/T)$ convergence rate under condition (BD). Here the optimal rate under (BD) means the state-of-the-art $\mathcal{O}(1/T)$ convergence rate, not the best results achieved *w.r.t.* bandwidth (BD). All proofs in this section are given in Appendix B.

Theorem 1 *Let Assumptions 1, 2, and 3 hold. We consider the step size $\eta(t)$ satisfy the conditions (A) and (B) for all $1 \leq t \leq T$ and let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. After at most $T > n_0$ iterates, we have*

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \begin{cases} \frac{\Delta^0_{n_0}}{(T+1)^{\mu m}} + \mathcal{O}\left(\frac{M^2\sigma^2}{(T+1)^{\mu m}}\right) & \text{if } m < \frac{1}{\mu}; \\ \frac{\Delta^0_{n_0}}{T+1} + \mathcal{O}\left(M^2\sigma^2 \cdot \frac{\ln T}{T+1}\right) & \text{else if } m = \frac{1}{\mu}; \\ \frac{\Delta^0_{n_0}}{(T+1)^{(\mu m)}} + \mathcal{O}\left(\frac{M^2\sigma^2}{T+1}\right) & \text{else} \, m > \frac{1}{\mu}. \end{cases}$$

where $\Delta^0_{n_0}$ has the same definition as Lemma 2.

First, we would like to clarify what this finite constant n_0 in Theorem 1 is. Under conditions (A) and (B), to make sure that $\eta(t) \leq 1/(2L_f)$ after n_0 iterations, we let $n_0 \geq 2ML_f + 1$. We might as well set $n_0 = 2ML_f + 1$ which is a finite constant and then choose a sufficiently large budget $T > n_0$. Especially, if $m > 1/\mu$, then $n_0 = 2ML_f + 1 > 2L_f/\mu + 1$.

Theorem 1 provides the unified worst-case convergence guarantees for all step sizes belonging to $1/t$-band. It reveals the variation of the convergence rates with the coefficient m of the lower bound $\delta_1(t)$. When $m > 1/\mu$, an optimal $O(1/T)$ convergence rate of SGD under strong convexity is obtained, which is comparable to that of Moulines and Bach (2011). Still, the weak growth condition on the gradient is milder than L-smooth used in Moulines and Bach (2011), and each individual function is not necessarily to be convex. Note that $m = 1/\mu$ is a special case that achieves a near-optimal $\mathcal{O}(\ln(T)/T)$ convergence rate. Besides, if $m < 1/\mu$, it greatly slows down the convergence of SGD with the rate $\mathcal{O}(1/T^{\mu m})$. Thus the value of m is critical. The similar behaviors have been also observed in Leen and Orr (1994); Nemirovski et al. (2009); Moulines and Bach (2011) for $\eta(t) = \eta_0/t$.

We then give a specific example of $1/t$-band to show its theoretical potential benefits compared to the typical step size $\eta(t) = \eta_0/t$.

Corollary 2 *(A special example of $1/t$-band) Under the same conditions as Theorem 1, we consider the step size $\eta(t) = M/t$ for $t \in [1, T/2)$ and $\eta(t) = m/t$ for $t \in [T/2, T]$ where $1/\mu < m \leq M$, then*

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \frac{\Delta^0_{n_0}}{2^{\mu(m-M)}T^{\mu M}} + \frac{2\sigma^2 m^2}{\mu m - 1} \cdot \frac{1}{T} + \left(\frac{M^2}{\mu M - 1} - \frac{m^2}{\mu m - 1}\right)\frac{2\sigma^2}{T2^{\mu m - 1}}$$

Let $m = M = \eta_0$, we then recover the convergence $\mathcal{O}\left(\Delta^0_{n_0}/T^{(\mu\eta_0)} + \sigma^2\eta_0^2/T\right)$ for $\eta(t) = \eta_0/t$ (Moulines and Bach, 2011). In the noise-less setting ($\sigma^2 = 0$), the upper bound $\eta(t) = M/t$ is fastest among the $1/t$-band in converging to the solution with an $\mathcal{O}(1/T^{\mu M})$ rate. However, it is incorrect in the noise setting where the noise error term (w.r.t. σ^2) finally dominates the convergence. In turn, a larger M results in a larger constant factor M^2 of the noise error. For $m > 1/\mu$, the constant factor with respect to σ^2 in Corollary 2 is smaller than the constant $2M^2/(\mu M - 1)$ achieved by the upper bound M/t, *i.e.*,

$$\frac{2\sigma^2 m^2}{\mu m - 1} + \left(\frac{M^2}{\mu M - 1} - \frac{m^2}{\mu m - 1}\right)\frac{2\sigma^2}{2^{\mu m - 1}} < \frac{2M^2\sigma^2}{\mu M - 1}.$$

This example in Corollary 2 makes a compromise that the step-size first reaches the upper bound M/t which can accelerate the convergence at the beginning, and then moves to the lower bound m/t to reduce the noise error.

Theorem 2 *Let Assumptions 1, 2, and 3 hold. We consider the step size $\eta(t)$ to satisfy the conditions (A) and (B) for all $1 \leq t \leq T$. Let $n_1 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(4L_f)\}$ and $f_{n_1} = \max_{1 \leq t \leq n_1}\{f(x_t) - f(x^*)\}$. If $m \geq 1/\mu$, for $T > n_1$, we have that $\mathbb{E}[f(\hat{x}_T) - f(x^*)]$ is bounded by*

$$\mathcal{O}\left(\frac{\Delta^0_{n_1}}{T(T+t_0)} + \frac{f_{n_1}}{T(T+t_0)} + \frac{M^2\sigma^2}{mT} + \frac{M^2\sigma^2}{m}\frac{\ln T}{T(T+t_0)}\right) \tag{11}$$

where $\hat{x}_T = \frac{\sum_{t=1}^T (t+t_0)x_t}{S_1}, t_0 \in \mathbb{N}, S_1 = \frac{T(T+t_0)(t_0+1)}{2}$, $\Delta^0_{n_1} = \frac{\|x_1 - x^*\|^2}{(n_1+1)^{\mu m}} + 4\sigma^2 M^2 + n_1 \chi_{n_1} f_{n_1}$.

Moreover, we derive the error bound for $1/t$-band on the functions values of order $\mathcal{O}(1/T + \ln(T)/T^2)$, which is comparable to those of Rakhlin et al. (2012); Lacoste-Julien et al. (2012); Shamir and Zhang (2013) using similar averaging techniques (see Remark 3) for $1/t$-stepsize. However, we use the much-relaxed growth condition (Assumption 3) instead of the uniform boundedness of stochastic gradient which is troublesome when the iterates are not restricted to be bounded (Rakhlin et al., 2012; Lacoste-Julien et al., 2012; Shamir and Zhang, 2013; Hazan and Kale, 2014). From (11), we know that the noise error (related to σ^2) depends on M^2/m. Compared to Theorem 1 (when $m > 1/\mu$), we find that if $M \approx m$, the averaging technique reduces the dependence of M from quadratic to linear.

Remark 3 (Other averaging techniques) *In (11), for any $T > 0$, let $\hat{x}_T = \sum_{t=1}^T \alpha(t)x_t$, where $\alpha(t) = (t+t_0)/S_1$, we have*

$$\frac{\alpha(t)}{\alpha(t+1)} = \frac{t+t_0}{t+t_0+1}.$$

If $t_0 = 1$, the weight scheme in (11) is exactly the same as Lacoste-Julien et al. (2012). For different $t_0 > 1$, $\hat{x}_T$ produces a generalized weighted average iterate, different from those in Lacoste-Julien et al. (2012) and Shamir and Zhang (2013). We can see that for fixed $0 < t < T$, the ratio between the weights $\alpha(t)/\alpha(t+1) = t/(t+\eta)$ (Shamir and Zhang, 2013) is smaller than $(t+t_0)/(t+t_0+1)$ if $\eta \geq 1$ and $t_0 \geq 1$. This means that the weight of (11) from t to $t-1$ decays slower than that in Shamir and Zhang (2013). Moreover, if $\alpha(t) = (t+t_0)^k / \sum_t (t+t_0)^k$ for some $k \in \mathbb{N}^+$, we have

$$\frac{\alpha(t)}{\alpha(t+1)} = \frac{(t+t_0)^k}{(t+t_0+1)^k}.$$

This form is actually equivalent to that of Shamir and Zhang (2013), and the integer k corresponds to η. These averaging techniques are also related to the tool of factorial powers proposed in Defazio and Gower (2021).

We further relax the lower or upper bound of $\eta(t)$ and figure out in which cases the state-of-the-art $\mathcal{O}(1/T)$ convergence rate can also be obtained. To better understand how the lower or upper bound affects the convergence rate, we only change one of them at one time. In general, if we fix the upper bound $\delta_2(t)$, the lower bound of $\eta(t)$ can be extended to (A_1) (see Theorem 3). Moreover, in Remark 4, we reveal that the condition (A_1) is essential to reach the optimal $\mathcal{O}(1/T)$ convergence rate.

Theorem 3 *Suppose that Assumptions 1, 2, and 3 hold. We consider the step size $\eta(t)$ satisfy the following conditions*

(A_1) *there exists a constant $C > 0$ such that for all $t^* \in \{1, 2, \cdots, T\}$, we have*

$$\sum_{t=t^*}^{T} \eta(t) \geq C \ln\left(\frac{T+1}{t^*}\right); \tag{12}$$

(B) *there exists a constant $M > 0$ such that $\eta(t) \leq \frac{M}{t}$ for all $1 \leq t \leq T$.*

Let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. If $C > 1/\mu$, for $t > n_0$, we have $\mathbb{E}[\|x_{T+1} - x^*\|^2]$ is at most

$$\frac{\Delta_{n_0}^0}{(T+1)^{(\mu C)}} + \mathcal{O}\left(\frac{\sigma^2 M^2}{\mu C - 1} \cdot \frac{1}{T+1}\right).$$

The theorem shows that if the upper bound $\delta_2(t)$ is of order $1/t$, the lower bound of $\eta(t)$ can be extended to be of order $1/t$ on average to obtain an $\mathcal{O}(1/T)$ rate. Note that condition (A_1) does not require $\eta(t)$ to be larger than C/t for all $1 \leq t \leq T$. For example, if $\eta(t)$ is larger than m/t for $t \in [1, \alpha T]$ where $\alpha \in (0, 1)$ and satisfies condition (B), we still can derive an $\mathcal{O}(1/T)$ bound for SGD under this step size.

Compared to the step size from Jain et al. (2019). The following piece-wise decay step size which is modified by Jain et al. (2019) for strongly convex problems (see (4) of Jain et al. (2019))

$$\eta(t) = 2^{-i} \cdot \frac{1}{\mu t}, \text{ for } T_i < t \leq T_{i+1},\ T_i = T - \lceil T \cdot 2^{-i} \rceil,$$

satisfies (A_1) and (B). From Theorem 3, we are able to achieve an $\mathcal{O}(1/T)$ rate measured by $\mathbb{E}[\|x_{T+1} - x^*\|^2]$, which is slightly weaker than that of Jain et al. (2019) measured by functions values on the final iterate ($\mathbb{E}[f(x_T) - f^*] \leq \mathcal{O}(1/T)$). Jain et al. (2019) assumes that the objective function is Lipschitz continuous ($\|\nabla f(x)\|$ is bounded) and the stochastic gradient is bounded (a.s.). However, our assumption of the gradient is much weaker. As we know, the $1/t$-stepsize only achieves $\mathbb{E}[f(x_T) - f^*] \leq \mathcal{O}(\log T/T)$ for non-smooth problems (Shamir and Zhang, 2013). Thus, the piece-wise example, in turn, indicates that the bandwidth-based framework can be useful and has the potential to design a step size that is better than the typical $\eta(t) = \eta_0/t$ step size. It is interesting to know whether we can achieve $\mathbb{E}[f(x_T) - f^*] \leq \mathcal{O}(1/T)$ as Jain et al. (2019) under the general conditions (A) and (B_1). But to keep our focus, we will not give the analysis here and leave it to the future.

Remark 4 *To analyze the convergence rate of SGD, the key step is to estimate Γ_T^2 defined by (9b). If $\eta(t)$ has an upper bound M/t for all $1 \leq t \leq T$, we have*

$$\Gamma_T^2 = 2\sigma^2 \sum_{t=1}^{T} \eta(t)^2 \exp\left(-\mu \sum_{u>t}^{T} \eta(u)\right) \leq 2\sigma^2 M^2 \sum_{t=1}^{T} \frac{1}{t^2} \cdot \exp\left(-\mu \sum_{u>t}^{T} \eta(u)\right).$$

Considering the partial summation of $\frac{1}{t^2}\exp\left(-\mu\sum_{u>t}^{T}\eta(u)\right)$ *from* t^* *to* T*, for all* $1 \le t^* \le T$*, we have*

$$\sum_{t=t^*}^{T}\frac{1}{t^2}\cdot\exp\left(-\mu\sum_{u>t}^{T}\eta(u)\right) \ge \sum_{t=t^*}^{T}\frac{1}{t^2}\cdot\exp\left(-\mu\sum_{u=t^*}^{T}\eta(u)\right).$$

In order to achieve the convergence rate such that $\mathbb{E}[\|x_{T+1}-x^*\|^2] \le \mathcal{O}(1/T)$*, we have to require that*

$$2\sigma^2 M^2\sum_{t=t^*}^{T}\frac{1}{t^2}\cdot\exp\left(-\mu\sum_{u=t^*}^{T}\eta(u)\right) \le \mathcal{O}\left(\frac{1}{T}\right).$$

Then

$$2\sigma^2 M^2\exp\left(-\mu\sum_{u=t^*}^{T}\eta(u)\right)\left(\frac{1}{t^*}-\frac{1}{T}\right) \le \mathcal{O}\left(\frac{1}{T}\right) \Longrightarrow \sum_{u=t^*}^{T}\eta(u) \ge \frac{1}{\mu}\ln\left(\frac{T}{t^*}-1\right)+\mathcal{O}(1).$$

Thus we see that condition (A_1) *in Theorem 3 is essential to achieve the optimal* $\mathcal{O}(1/T)$ *convergence rate under condition* (B)*.*

A relatively large step size, as long as the iterate is stable, is often preferred in practice, especially at the initial training (Huang et al., 2017; Loshchilov and Hutter, 2017). The upper bound of p is always smaller than 1 for $m > 1/\mu$ and $r \in (1/2, 1)$, so the value of p is reasonable. A few attempts have been made by Gower et al. (2019); Allen-Zhu (2018) to allow the constant step sizes at the earlier training. In the following theorem, the step size $\eta(t)$ is allowed to vary within a constant band whose lower and upper bounds consist of two positive constants in the early C_1T^p ($p \in (0,1)$) iterations. After C_1T^p iterations, the step size turns to the second stage within a $1/t$-band. For simplicity, we assume that C_1T^p is an integer.

Theorem 4 *We assume that Assumptions 1, 2, and 3 hold. If the step size* $\eta(t)$ *satisfies the following conditions: there are some constants* $p \in (0,1)$*,* $C_1 > 0$*,* $0 < m_1 \le M_1$*,* $0 < m_2 \le M_2$ *such that*

(A_2) $\eta(t) \ge m_1$ *for* $t \in [C_1T^p]$ *and* $\eta(t) \ge \frac{m_2}{t}$ *for* $t \in [T]\backslash[C_1T^p]$*;*

(B_2) $\eta(t) \le M_1$ *for* $t \in [C_1T^p]$ *and* $\eta(t) \le \frac{M_2}{t}$ *for* $t \in [T]\backslash[C_1T^p]$*.*

Let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$*. If* $\kappa = (\mu m_2)(1-p) \ge 1$ *and* n_0 *is a finite constant and is independent of* T*, then for* $T > n_0$*,* $\mathbb{E}[\|x_{T+1}-x^*\|^2]$ *is at most*

$$\mathcal{O}\left(\frac{\Delta_{n_0}^0}{T^{(\kappa+p)}}\right)+\mathcal{O}\left(\frac{M_1^2\sigma^2}{\mu m_1 T^{\kappa}}\right)+\mathcal{O}\left(\frac{M_2^2\sigma^2}{T+1}\right).$$

Let $\kappa = (\mu m_2)(1-p) \ge 1$, *i.e.*, $m_2 \ge 1/(\mu(1-p))$ for $p \in (0,1)$, we can achieve a unified $\mathcal{O}(1/T)$ convergence rate for a class of step-size that satisfies conditions (A_2) and (B_2). Suppose that the constant n_0 comes earlier than the turning point (*i.e.*, $n_0 < C_1T^p$), to ensure that $\eta(t) \le 1/(2L_f)$ after n_0 iterates, we require that the total number of iterations T is sufficient large such that $M_2/(C_1T^p) \le 1/(2L_f)$, *i.e.*, $C_1T^p \ge 2M_2L_f \ge 2L_f/(\mu(1-p))$.

We might as well let $n_0 \leq 2M_2L_f$. By properly choosing the step size such that $\eta(t) \leq 1/(2L_f)$ for $t \in [n_0, C_1T^p]$, then the constant n_0 is well-defined.

When the iteration budget T is very large, for example, $T \gg (4K/C_1)^{1/p}$ where $K = L_f/\mu$, we can see that $C_1T^p \gg 4K$, our result allows more iterations where the step size can be a constant at the early stage of training, which extends the existing result of Gower et al. (2019) only equipped with a constant step size at the initial $4K$ steps. Note that Allen-Zhu (2018) proposes an algorithm SGD^{sc}(a.k.a. SGD after SGD), in which the step size $\eta(t) = 1/(2L)$ for the initial $\lfloor T/2 \rfloor$ iterates, where L is the parameter of smoothness. However, the output of each inner loop is an average of all inner iterates, which is different from the SGD algorithm discussed in this paper. Thus we will not give a further comparison.

3.1 Guarantees for Cyclical Step Sizes

The last part of this section will address the applications of the bandwidth framework to provide guarantees for the cyclical step sizes. We focus on the cyclical step size, which repeats the same pattern (*e.g.*, constant, triangular, cosine, sine-wave) at each cycle $i \geq 1$, given the budget of iteration $T \geq 1$ and the length of each cycle T_i:

$$\eta^i_{\min} \leq \eta(t) \leq \eta^i_{\max} \tag{13}$$

where $t \in [1 + \sum_{l=1}^{i-1} T_l, \sum_{l=1}^{i} T_l]$, $\sum_i T_i = T$ and each $T_i \geq 1$.

Hazan and Kale (2014) proposed the piece-wise decay step size within the i-th run

$$\eta(t) = \eta_i = \frac{\eta_{i-1}}{2},\ t \in [T_i, T_{i+1}),\ T_{i+1} = 2T_i, \tag{14}$$

where $\sum_i T_i = T$, $T_0 \geq 1$, and $\eta(1) = \eta_0$. The step size drops half per cycle, but the period T_i of each cycle doubles. Clearly, it satisfies the conditions (A) and (B) with $m = T_0\eta_0/2$ and $M = T_0\eta_0$. So we can obtain an $\mathcal{O}(1/T)$ convergence rate for (14) (see Theorems 1 and 2) which is comparable to Hazan and Kale (2014) but under milder condition (we use expected smooth rather than uniformly bounded gradient in (Hazan and Kale, 2014)). We address one result derived from Theorem 1 below.

Corollary 3 *Let $k_0 \geq 1$ be the stage number of (14) where $n_0 = T_0(2^{k_0} - 1)$. To ensure that $\eta(t) \leq 1/(2L_f)$ after n_0 iterations and $m = \eta_0T_0/2 > 1/\mu$, we set $k_0 = 1 + \log_2(4L_f/(\mu T_0))$. Under the same setting as Theorem 1, for the SGD algorithm with step size (14), after $T > n_0$ iterations, we have $\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \mathcal{O}(1/T)$.*

In addition, for any step size (*e.g.*, locally varied like triangle, cosine, and sine-wave) which belong to this class of bandwidth schedule (*i.e.*, lower and upper bounds are based on (14)), the optimal rate can be guaranteed immediately by properly choosing the parameters.

The exponentially decaying step size is popular and defaulted in some deep learning libraries (*e.g.*, PyTorch and TensorFlow), that is

$$\eta(t) = \eta_0\alpha^{\lfloor t/T_0 \rfloor}, \tag{15}$$

where $\alpha \in (0, 1)$ is a constant which is independent of T and T_0 accounts for how many iterations have been performed since the last run. For simplicity, we let $\alpha = 1/2$. Here we consider the period T_0 be fixed and the same per cycle, and make the following discussions:

- $T_0 = 1$, or a constant (independent of T). Its non-asymptotic convergence can not be guaranteed because $\sum_{t=1}^{\infty} \eta(t) = +\infty$ is not satisfied.

- $T_0 = \lfloor T^r \rfloor$, $r \in (0, 1)$. When $k_0 = \lfloor t/T_0 \rfloor = \lfloor r \log_2(T) \rfloor$, the partial summation $\sum_{t=k_0T_0}^{(k_0+1)T_0} \eta(t)^2 \exp(-\mu \sum_{u=(k_0+1)T_0}^{T} \eta(u)) \geq \exp(-2\mu\eta_0) \sum_{t=k_0T_0}^{(k_0+1)T_0} \eta(t)^2 = \mathcal{O}(1/T^r)$. From Lemma 2, it hardly obtains the non-asymptotic $\mathcal{O}(1/T)$ convergence rate.

- $T_0 = \lfloor T/ \log_2 T \rfloor$. Let $k_0 = \lfloor t/T_0 \rfloor = \lfloor \log_2 T - \log_2 \log_2(T) \rfloor$. In this case we have $\sum_{t=k_0T_0}^{(k_0+1)T_0} \eta(t)^2 \exp(-\mu \sum_{u=k_0T_0}^{T} \eta(u)) \geq \exp(-2\mu\eta_0) \sum_{t=k_0T_0}^{(k_0+1)T_0} \eta(t)^2 = \mathcal{O}(\log_2 T/T)$. We can see that the best result will not exceed $\mathcal{O}(\log_2 T/T)$ from Lemma 2. This rate has been demonstrated by Ge et al. (2019) for the least squares problems.

- $T_0 = \lfloor T/k \rfloor$, where $k \in \mathbb{N}^+$ is a constant(independent of T). In this case, the final step size is $2^{-k} \gg 1/T$. It is impossible to achieve the non-asymptotic $\mathcal{O}(1/T)$ rate.

Therefore, we can conclude that SGD hardly achieves the ideal $\mathcal{O}(1/T)$ convergence rate under (15) for strongly convex problems if the period T_0 of each cycle is fixed and the same.

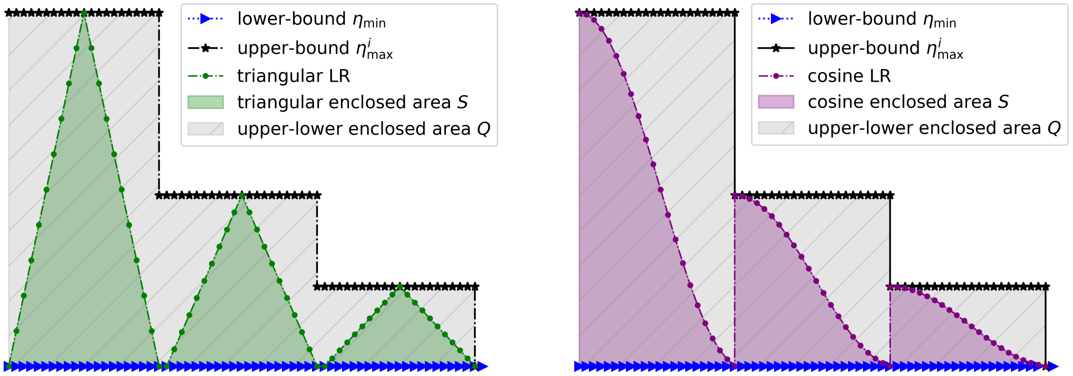
A sine-wave learning rate was proposed (An et al., 2017) where the step size decays exponentially (the continuous form of (15)) and local oscillations within a range of values. This cyclical step size can be treated within the bandwidth framework (BD) based on (15). Unfortunately, from the discussion on exponential decaying step size (15), the non-asymptotically state-of-the-art $\mathcal{O}(1/T)$ convergence rate can not be guaranteed in theory. Nevertheless, if the boundary functions $\delta_1(t)$ and $\delta_2(t)$ are taken as (14), that is, the boundary functions drop by half and the length of the cycle increases after each cycle. It results in the sine-wave learning rate achieving the optimal $\mathcal{O}(1/T)$ convergence rate. Moreover, if the sine-wave policy or their boundaries is chosen as the following corollary, our previous analysis can guarantee the optimal $\mathcal{O}(1/T)$ convergence rate.

Corollary 4 *For any cyclical step size whose lower and upper bounds satisfy, for example, (A) and (B) of Theorem 1 and its variants,* e.g., *(A_1) and (B), and (A_2) and (B_2), we have* $\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \mathcal{O}(1/T)$ *under some proper conditions.*

The triangular policy was proposed by Smith (2017) where the original idea is to linearly increase and then decrease the step size within a band. In the simulations, the author fixes the lower bound $\eta_{\min}^i$ of the band as a constant and adopts the decaying schedule (15) as the upper bound $\eta_{\max}^i$. The previous theorems can not be used anymore due to the fixed lower bound. According to the similarity of cyclical step sizes per cycle, we can apply Lemma 2 to show the convergence of this class of step size with extra carefulness. The formal description of the result is addressed below.

Corollary 5 *Consider the cyclical step size defined by (13) (see Figure 1) where the length of each cycle $T_i = T_0 \geq 1$ is fixed, number of cycle $N = \lceil T/T_0 \rceil$, the lower bound $\eta_{\min} = m > 0$ is fixed as a small constant and the upper bound $\eta_{\max}^i$ decays with cycle i. At each cycle i, let S_i and Q_i denote the enclosed area of the cyclical step size with its lower bound $\eta_{\min}$, and the enclosed area between the upper bound $\eta_{\max}^i$ and lower bound $\eta_{\min}$, respectively. We assume that $S_i/Q_i \geq \psi$ ($i \in [N]$) where $\psi \in (0, 1]$ is a constant. Under the same conditions of Lemma 2, let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$, then for $T \geq n_0$, we have*

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \exp\left(-\mu\psi \sum_{i=1}^{N} Q_i - \mu m T\right) \Delta_{n_0}^0 + 2\sigma^2 T_0 \sum_{i=1}^{N} (\eta_{\max}^i)^2 \exp\left(-\mu \sum_{l>i}^{N} (\psi Q_l + m T_0)\right).$$



(a) triangular

(b) cosine

Figure 1: The visualization of two popular cyclical step-sizes

In particular, we consider the two decaying patterns for upper bound $\eta^i_{\max}$*:*

(1) $\eta^i_{\max} = M/2^{i-1}$ *where* $t \in [1+(i-1)T_0, iT_0]$ *(*$1 \le i \le N$*). Especially, if* $N = \lceil \log_2 T \rceil$*, we have* $\mathbb{E}[\|x_{T+1} - x^*\|^2] \le \mathcal{O}(\frac{\log_2 T}{T})$*.*

(2) $\eta^i_{\max} = M/(iT_0)$ *for all* $i \in [N]$*. Then* $\mathbb{E}[\|x_{T+1} - x^*\|^2] \le \mathcal{O}\left(\left(\frac{T_0}{T}\right)^{\mu M \psi} + \frac{1}{T}\right)$*.*

This corollary establishes a unified analysis framework for the cyclical step size whose lower bound is fixed, and the upper bound is decreasing. The proof is given in Appendix B. To be more intuitive about the step-sizes of Corollary 5, we depict two popular cyclical schemes: triangular and cosine step-sizes in Figure 1. In each cycle, an example of the two enclosed areas S_i, Q_i can be found in the shaded area. The assumption on $S_i/Q_i \ge \psi$ ($i \in [N]$) is easily satisfied by triangular step size (Smith, 2017) with $\psi = 1/2$ and cosine step size (Loshchilov and Hutter, 2017) with $\psi = 1/2$. For the triangular step size (Smith, 2017), we can achieve a near-optimal (up to $\log_2 T$) rate based on the exponential decaying upper bound (see Case (1) of Corollary 5). This rate matches the result of the step-decay step size achieved by Ge et al. (2015) for strongly convex least squares problems. Furthermore, if we select the piece-wise $1/i$ as the upper bound (see Case (2) from Corollary 5), an $\mathcal{O}(1/T)$ rate can be achieved with triangular step size (Smith, 2017) under proper conditions: for instance when $T_0 \ll T$ is a constant and $M \ge 1/(\mu\psi)$ or $T_0 = \mathcal{O}(1/T^r)$ for $r \in (0,1)$ and $M \ge 1/((1-r)\mu\psi))$. We notice that the cosine with restart policy proposed by (Loshchilov and Hutter, 2017) only considers the fixed upper and lower bounds. But the authors mentioned that it is interesting to study cosine step size with decaying upper or lower bounds. The analysis above can provide convergence guarantees for the cosine step size with restart if the upper and lower bounds are under proper conditions.

4. Convergence Analysis Under the Same Boundary Order

In this section, we will investigate the convergence rate of the SGD algorithm where the bandwidth-based step size (BD) has the same boundary order, *i.e.*, $\delta_1(t) = \delta_2(t)$.

The well-known convergence conditions on step size for standard SGD were proposed by Robbins and Monro (1951)

$$(1')\ \sum_{t=1}^{\infty}\eta(t)=+\infty; \qquad (2')\ \sum_{t=1}^{\infty}\eta(t)^2<+\infty. \tag{H1}$$

Obviously, the polynomial decaying step size $\eta(t)=\mathcal{O}(1/t^p)$ for $p\in(\frac{1}{2},1]$ satisfies (H1). However, (H1) does not hold for $\eta(t)=\mathcal{O}(1/t^p)$ with $0<p\leq 1/2$ which has been proven to converge (Leen et al., 1998; Ljung, 1977; Moulines and Bach, 2011). Moreover, one interesting thing is that the step size under (H1) is possibly non-monotonic. For example, the step size may oscillate between two boundaries $\eta(t)=1/t$ and $\eta(t)=1/\sqrt{t}$.

Ljung (1977) proposed the following convergence conditions (H2) for the recursive stochastic algorithms

$$\begin{array}{ll}(1')\ \sum_{t=1}^{\infty}\eta(t)=+\infty; & (2')\ \sum_{t=1}^{\infty}\eta(t)^p<+\infty,\text{ for some } p>0;\\ (3')\ \eta(\cdot)\text{ is a decreasing sequence;} & (4')\ \lim_{t\to\infty}\sup[1/\eta(t)-1/\eta(t-1)]<\infty.\end{array} \tag{H2}$$

Compared to (H1), (H2) seems cover more generic cases, *e.g.*, $\eta(t)=\eta_0/t^p$ for all $p\in(0,1]$. However, there are some cases which satisfy (H1) but are not admitted by (H2), for example $\eta(t)=1/(t\log(t+1))$. Moreover, the step size $\eta(t)$ of (H2) is assumed to be decreasing, which is not essential for (H1).

Recently, Nguyen et al. (2019a) extended (H1) and (H2) to the following cases (H3)

$$(1')\ \sum_{t=1}^{\infty}\eta(t)=+\infty; \qquad (2')\ \lim_{t\to+\infty}\eta(t)=0; \qquad (3')\ \frac{d\eta(t)}{dt}\leq 0. \tag{H3}$$

As we can see, the common choices $\eta(t)=1/t^p$ for $p\in(0,1]$ and $1/(t\ln(t))$ all satisfy (H3). In addition, $\eta(t)=1/\ln(t)$, which decays slower than any polynomial decaying step sizes, satisfies the above conditions. The authors proved the convergence of SGD and derived a uniform formula to describe the convergence rates for the step sizes satisfying (H3) (see theorem 9 and 10 in Nguyen et al. (2019a)).

In the rest of this section, we focus on the sequence of step size $\{\eta(t)\}$ that satisfies

$$m\delta(t)\leq\eta(t)\leq M\delta(t), \tag{BD-S}$$

where $m\leq M$ are two positive constants and the boundary function $\delta(t)$ satisfies (H3). The main theorem is presented as follows, covering most of the abovementioned cases. The proofs in this section are provided in Appendix C.

Theorem 5 *Suppose Assumptions 1, 2, and 3 hold. The step size sequence $\{\eta(t)\}$ satisfies condition (BD-S) and the boundary function $\delta(t)$ is differentiable and satisfies (H3). Let $n_0:=\sup\{t\in\mathbb{N}^+:\eta(t)>1/(2L_f)\}$ and we assume that n_0 is a constant which is independent of T. For $t>n_0$,*

1. *if $\lim_{t\to\infty}t\delta(t)=0$, we have that $\mathbb{E}[\|x_{t+1}-x^*\|^2]$ is at most*

$$\left(\Delta^0_{n_0}+\varepsilon_2\frac{\delta(1)^2(t_\epsilon-1)+2\epsilon^2}{\exp\left(-\mu m\int_{u=1}^{t_\epsilon}\delta(u)du\right)}\right)\exp\left(-\mu m\int_{u=1}^{t+1}\delta(u)du\right),$$

where ϵ and t_ϵ are constants appeared in the proof, $\varepsilon_2=2\sigma^2M^2\exp(\mu m\delta(1))$.

2. *If* $\lim_{t\to\infty} t\delta(t) = 1$*, the results of Theorem 1 can be applied.*

3. *If* $\lim_{t\to\infty} t\delta(t) = +\infty$ *and there exist constants* $c_1 \le \frac{\mu m}{2}$ *and* $T_M \in \mathbb{N}$ *such that* $-\frac{d\delta(t)}{dt} \le c_1\delta(t)^2$ *for all* $t \ge T_M$*, then* $\mathbb{E}[\|x_{t+1} - x^*\|^2]$ *is at most*

$$\frac{\varepsilon_2}{\mu m - c_1}\delta(t+1) + \left(\Delta^0_{n_0} + \frac{\varepsilon_2\delta(1)^2 T_M}{\exp\left(-\mu m \int_{u=1}^{t_M}\delta(u)du\right)}\right)\exp\left(-\mu m\int_{u=1}^{t+1}\delta(u)du\right),$$

where ε_2 *is the same as the first case.*

Corresponding to the limit of $\delta(t)t$, we discuss three interesting cases in Theorem 5. As we know, the result is new when $\lim_{t\to\infty} t\delta(t) = 0$. It covers the cases in which the step size drops faster than $1/t$, *e.g.*, $\delta(t) = 1/(t\ln(t))$. In the third case, to make the proof precise, we add a condition that $-d\delta(t)/dt \le c_1\delta(t)^2$ ($\forall t \ge T_M$) for some c_1 and T_M (details are provided in Remark 5). The common choices, *e.g.*, $\delta(t) = 1/t^p$ for all $p \in (0, 1]$ and $\delta(t) = 1/\ln(t)$, all satisfy the condition. Especially, for $\eta(t) = \eta_0/t^p$ with $p \in (0, 1)$, we can achieve an $\mathcal{O}(1/t^p)$ convergence rate, which is comparable to that of Moulines and Bach (2011); however, we use the weak growth condition with noise which is milder than L-smooth, and we do not assume the convexity (a.s.) of each individual function (Moulines and Bach, 2011). More cases such as $\delta(t) = \ln(t+1)/t^p$ for all $p \in (0, 1]$ can also be included in the discussions. It is worthwhile to mention that when t is continuous, (4′) of (H2) can be reformulated as

$$\lim_{t\to\infty}\sup[1/\eta(t) - 1/\eta(t-1)] = \lim_{t\to\infty}\sup\left[\frac{\eta(t-1)-\eta(t)}{\eta(t)\eta(t-1)}\right] = \lim_{t\to\infty}\sup\frac{\frac{-d\eta(t)}{dt}}{\eta(t)^2} < +\infty.$$

This exactly implies that there exists a constant $c_1 > 0$ such that $-d\eta(t)/dt \le c_1\eta(t)^2$ for sufficiently large t. In the third case of Theorem 5, the scalar c_1 is supposed to be smaller than $\mu m/2$. The following lemma reveals that as long as such $c_1 > 0$ exists, there must be a constant $c_1 > 0$ such that $c_1 \le \mu m/2$.

Lemma 3 *We suppose that* $\lim_{t\to\infty} t\delta(t) = +\infty$*. If there exist constants* $c_1 > 0$ *and* $T_M \in \mathbb{N}^+$ *such that* $-\frac{d\delta(t)}{dt} \le c_1\delta(t)^2$ *for all* $t \ge T_M$*, there must be such a constant* c_1 *that satisfies* $c_1 \le \frac{\mu m}{2}$*.*

Remark 5 *Theorem 5 shows the convergence rate of SGD where the bandwidth-based step size satisfies (BD-S). We emphasize that*

1. *In the proof of the third case, an important step is to use integral* $\int_{l=1}^{t} P(l)dl$ *to evaluate the summation* $\sum_{l=1}^{t} P(l)$ *where* $P(l)$ *is the product of* $\delta(l)^2$ *and* $\exp(-\mu m\int_{u=l}^{t+1}\delta(u)du)$*. Even though* $\delta(l)$ *is decreasing and* $\exp(-\mu m\int_{u=l}^{t+1}\delta(u)du)$ *is increasing, there can be many possibilities for their product. Nguyen et al. (2019a) considered three cases for the product that,* e.g.*, decreases and then increases, keeps on increasing or decreasing (see the proof of theorem 9 in Nguyen et al. (2019a)). However, as we know the product of* $\delta(l)^2$ *and* $\exp(-\mu m\int_{u=l}^{t+1}\delta(u)du)$ *increases and then decreases in Ge et al. (2019). In Theorem 5, we add a condition* $-d\delta(t)/dt \le c_1\delta(t)^2$ *to describe "most general cases" mentioned in Nguyen et al. (2019a) and make the proof more rigorous.*

2. *Theorem 5 reveals the convergence rate of SGD, which is totally determined by* $\delta(t+1)$ *or* $\exp(-\mu m \int_{u=1}^{t+1} \delta(u)du)$*. Our result provides better upper bounds in many cases compared to that of Nguyen et al. (2019a). For example, when* $\eta(t) = 1/(t\ln(t))$*, theorem 10 of Nguyen et al. (2019a) no longer gives an upper bound but Theorem 5 shows that it is bounded by* $\exp(-\mu m \int_{u=1}^{t+1} \delta(u)du)$*. In the case that* $\eta(t) = 1/\sqrt{t}$*, the first term of the upper bound in theorem 10 (Nguyen et al., 2019a) is actually larger than* $\eta(t+1)$*, which is worse than the result of Theorem 5.*

3. *The step size* $\eta(t)$ *in Theorem 5 can be non-monotonic, rather than monotonic (Ljung, 1977; Nguyen et al., 2019a) or given in monotonic forms (*e.g., η_0/T *or* η_0/t^p *for* $p \in (0,1]$*) in most of the literature analyzing the convergence rate of SGD (Rakhlin et al., 2012; Moulines and Bach, 2011; Shamir and Zhang, 2013; Lacoste-Julien et al., 2012; Bottou et al., 2018; Gower et al., 2019; Jain et al., 2019).*

5. Convergence Analysis Based on the Different Boundary Orders

This section will present the convergence rate of SGD where the lower bound function $\delta_1(t)$ and the upper bound function $\delta_2(t)$ are in different orders. From Section 4, if the lower and upper bounds of the step size $\eta(t)$ are in the same order, their convergence rate is consistent with their boundaries. In the following part, we want to find out the convergence behaviors of SGD when the boundaries of the step size are in different orders.

First, we are interested in the case $\delta_2(t) = \ln(t+1)/(t+1)$ which decays slower than the lower bound $\delta_1(t) = 1/(t+1)$.

Theorem 6 *Suppose that Assumptions 1, 2, and 3 hold. Let the step size sequence* $\{\eta(t)\}$ *satisfy that*

$$\frac{m}{t+1} \le \eta(t) \le \frac{M\ln(t+1)}{t+1}, \ t \ge 1,$$

for $0 < m \le M$*. Let* $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$*. For* $T > n_0$*, we have*

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \le \begin{cases} \mathcal{O}\left(\frac{\Delta_{n_0}^0}{T+2} + \frac{M^2\sigma^2\ln^3 T}{T+2}\right) & \text{if } m = \frac{1}{\mu}; \\ \mathcal{O}\left(\frac{\Delta_{n_0}^0}{(T+2)^{(\mu m)}} + \frac{M^2\sigma^2}{(T+2)^{(\mu m)}}\right) & \text{else if } m < \frac{1}{\mu}; \\ \mathcal{O}\left(\frac{\Delta_{n_0}^0}{(T+2)^{(\mu m)}} + \frac{M^2\ln^2 T}{T+2}\right) & \text{else } m > \frac{1}{\mu}. \end{cases}$$

The theorem reveals that when $m > 1/\mu$, SGD can achieve an $\mathcal{O}(\ln^2(T)/T)$ bound, which is nearly optimal. The proofs in this section are given in Appendix D.

As we know, (H1) is sufficient for the convergence of SGD, but the convergence rate under (H1) is unknown yet. If we keep the lower bound $\delta_1(t) = 1/t$ and continue to extend the upper bound $\delta_2(t)$, what kinds of results will we get? The following result answers this interesting question.

Theorem 7 *We assume that Assumptions 1, 2, and 3 hold. If step size* $\eta(t)$ *satisfies that*

$$\frac{m}{t} \le \eta(t) \le \frac{M}{t^\alpha} \tag{16}$$

for $\alpha \in (1/2, 1]$*. Let* $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$*. For* $t > n_0$*, we have*

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \begin{cases} \frac{\Delta^0_{n_0}}{(T+1)^{(2\alpha-1)}} + \mathcal{O}\left(\frac{M^2\sigma^2 \ln T}{(T+1)^{(2\alpha-1)}}\right) & \textit{if } \mu m = 2\alpha - 1; \\ \mathcal{O}\left(\frac{\Delta^0_{n_0}}{(T+1)^{(\mu m)}}\right) + \mathcal{O}\left(\frac{M^2}{(T+1)^{(2\alpha-1)}}\right) & \textit{else } \mu m \neq 2\alpha - 1. \end{cases}$$

In Theorem 7, the upper bound $\delta_2(t)$ in (16) is extended to $1/t^{\alpha}$ for $\alpha \in (1/2, 1]$. It is straightforward to see that (H1) holds for the step size $\eta(t)$ that satisfies (16). The corresponding convergence rate is $\mathcal{O}(1/(T+1)^{2\alpha-1})$ which is relied on α when $\mu m > 1$. Obviously, this result is worse than those achieved at its lower and upper bounds. Unfortunately, we cannot improve Theorem 7. On the other direction, we reduce the lower bound $\delta_1(t)$ to $1/((t+1)\ln(t+1))$, which decreases faster than the case $\delta_1(t) = 1/t$ in Theorem 7.

Theorem 8 *Suppose that Assumptions 1, 2, and 3 hold. Let the step size* $\eta(t)$ *satisfy*

$$\frac{m}{(t+1)\ln(t+1)} \leq \eta(t) \leq \frac{M}{(t+1)^{\alpha}},\ t \geq 1, \tag{17}$$

for $\alpha \in (1/2, 1]$*. Then for sufficiently large* T*, we have*

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \mathcal{O}\left(\frac{1}{(\ln(T+2))^{\mu m}}\right).$$

Theorem 8 shows that the convergence rate of SGD where the step size satisfies (17) is consistent with the result achieved at the lower bound $\eta(t) = m/((t+1)\ln(t+1))$.

6. Numerical Experiments

In this section, we propose several non-monotonic step sizes within $1/t$-band to show the effectiveness compared to their baselines, *e.g.*, $\eta(t) = \eta_0/t$ (called **$1/t$-stepsize**) and exponentially decaying step size. The typical $1/t$ stepsize decays very fast at the beginning, so we update all step sizes after one epoch[4] shown as Algorithm 1 (called Epoch-SGD).

Algorithm 1 Epoch-SGD

```
Initialization: initial point $x_0 = x^1_1$, # inner loop $m^{'}$, # outer loop $N$
for  $t = 1 : N$  do
    Update the step size $\eta(t)$
    for  $i = 1 : m^{'}$  do
        Choose a subset $\Omega_i \subseteq [n]$ randomly, where $|\Omega_i| = b$
        Compute $g^t_i = \frac{1}{b}\sum_{l \in \Omega_i} \nabla f(x^t_i; \xi_l)$
        $x^t_{i+1} = x^t_i - \eta(t) g^t_i$
    end for
    $x^{t+1}_1 = x^t_{m^{'}+1}$
end for
Return $x^N_{m^{'}+1}$
```

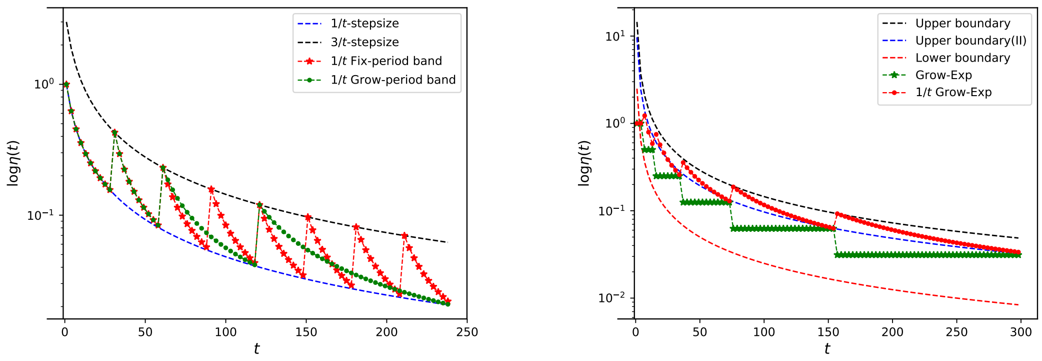
[4] One epoch means to traverse all sample data once.

6.1 $1/t$-band Step Sizes

We formulate some non-monotonic step sizes $\eta(t)$ which belongs to a banded region $[\eta_0/t, s\eta_0/t]$ (named **$1/t$-band**), where $s > 1$. The boundary function $\eta(t) = \eta_0/t$ is called $1/t$-stepsize. Let $t_i\,(i = 1, 2, \cdots, 1 \le t_1 < t_2 < t_3 < \cdots)$ be the nodes where the step size might be non-monotonic or non-differentiable. For $t \in [t_i, t_{i+1})$, let

$$\eta(t) = \frac{\hat{A}_i}{\hat{B}_i t + 1}, \tag{18}$$

where $\hat{A}_i, \hat{B}_i$ are constants such that $\eta(t_i) = s\eta_0/t_i$ and $\eta(t_{i+1}) = \eta_0/t_{i+1}$. In reality, other forms of $\eta(t)$ exist, *e.g.*, linear decay and concave decay. In the paper, we are interested in the case that $\eta(t)$ has the form of (18). We consider the two cases: **(1)** $t_{i+1} - t_i$ is fixed and the same. We call this **$1/t$ Fix-period band**; **(2)** $t_{i+1} - t_i$ grows exponentially. We call this **$1/t$ Grow-period band**. For an intuitive explanation, we plot the two cases and their boundaries $1/t$-stepsize ($s = 3$) in Figure 2(a).



(a)

(b)

Figure 2: Different kinds of $1/t$-band step sizes

More general, the step size varies between the minimum $\eta_{\min} = \{\eta^i_{\min}\}_{i\in\mathbb{N}}$ and maximum $\eta_{\max} = \{\eta^i_{\max}\}_{i\in\mathbb{N}}$, and locally has the form that

$$\eta(t) = \frac{\hat{A}_i}{\hat{B}_i t + 1} \in [\eta^i_{\min}, \eta^i_{\max}], t \in [t_i, t_{i+1}]. \tag{19}$$

Especially, we consider $\eta^i_{\max} > \eta^{i-1}_{\min}$, which is called **$1/t$ up-down policy**. For $1/t$ Fix-period band and $1/t$ Grow-period band, the baseline of the step size is $\eta_{\min} = \eta_0/t$. Based on the known exponentially decaying step size with a growing period (called **Grow-Exp**)

$$\eta(t) = \eta_i = \eta_0/2^i, t \in [t_i, t_{i+1}],\ T_i = t_{i+1} - t_i = T_0 2^i, \tag{20}$$

which has been studied by Hazan and Kale (2014). Let $\eta^i_{\min} = \eta_i$ in (20) and we define $\eta^i_{\max} = \theta\eta^{i-1}_{\min}$ where the up-down ratio $\theta > 1$ (called **$1/t$ Grow-Exp**). If θ is too large, a sudden increase in step size might lead to a negative effect. Therefore, we restrict the ratio $\theta \in (1, 1.5]$. The Grow-Exp step size, $1/t$ Grow-Exp step size, and their boundaries are plotted in Figure 2(b) where $T_0 = 5$ and $\theta = 1.5$. Regardless of Grow-Exp or $1/t$ Grow-Exp, we can easily find that they all belong to $1/t$-band.

We then include the comments on the proposed bandwidth step sizes to show how these step sizes also guide from our theory.

Remark 6 *(***Theoretical benefits from bandwidth***) From Figure 2(a), we see that the area enclosed by $1/t$ Fixed-period band and x-axis is larger than that of its lower boundary. According to Lemma 2, based on $1/t$ Fixed-period band, we can achieve a lower error bound for Γ_T^1 than that of the boundary $\eta(t) = \eta_0/t$. Thus $1/t$ Fixed-period band could be faster than $1/t$-stepsize ($\eta(t) = \eta_0/t$) at the initial iterations when Γ_T^1 dominates the error bound of $\mathbb{E}[\|x_{T+1} - x^*\|^2]$. At the end of each cycle, the step sizes hit the lower bound, which finally reduces the noise error. We have the similar conclusions for $1/t$ Grow-period band and $1/t$ Grow-Exp.*

Next, some numerical experiments are performed to demonstrate the efficiency of the proposed non-monotonic step sizes. All experiments are implemented in python 3.7.0 on a single node of LSSC-IV[5], which is a high-performance computing cluster maintained at the State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences. The operating system of LSSC-IV is Red Hat Enterprise Linux Server 7.3.

6.2 Parameters Tuning

This subsection discusses how to choose the parameters when designing the step sizes.

The initial step size η_0 is chosen from $\{0.1, 0.5, 1, 5, 10, 15\}$ for the Epoch-SGD algorithm on all step size schedules. Generally speaking, for the $1/t$-band, we do not know exactly the coefficients m and M for the lower and upper boundaries. In the experiments, the coefficient m is tuned properly using a similar approach as the initial step size η_0. Instead of finding the coefficient M of the upper bound, we tune the bandwidth $s = M/m \in \{2, 3, 4, 5\}$ for $1/t$ Fix-period band and $1/t$ Grow-period band. The distance of the adjacent nodes $t_i (i \in \mathbb{N}^+)$ depends on a budget of the outer loop N. In our experiments we set $t_{i+1} - t_i = 30$, $t_1 = 30$ for $1/t$ Fix-period band and $t_{i+1} = 2t_i$, $t_1 = 30$ for $1/t$ Grow-period. From Figure 2(a), we can see that $1/t$ Fix-period , $1/t$ Grow-period and $1/t$-stepsize coincide in the first cycle and $1/t$ Fix-period also coincides with $1/t$ Grow-period in the second cycle.

The Grow-Exp step size drops by half, and the period of each cycle is doubled. The initial period T_0 is chosen from $\{1, 2, 3, 5, 10, 20\}$. For $1/t$ Grow-Exp, we tune the up-down ratio $\theta \in \{1.1, 1.2, 1.3, 1.4, 1.5\}$ and the length of T_0 is the same as Grow-Exp.

In Section 3, to achieve the optimal rate, η_0 must be larger than $1/\mu$ where μ is the strongly-convex parameter. One may doubt the initial step size η_0 selected above is too small compared to the scalar $1/\mu = 1/\Lambda$ (*e.g.*, in the following logistic regression problems $\Lambda = 10^{-4}$). However, this is not the case. In the experiments, we modify the step size in every epoch instead of every iteration. Let $\tilde{t}_0$ be the number of iterations of each epoch, *i.e.*, $\tilde{t}_0 = n/b$ where n is the data size, and b is the mini-batch size. In the first epoch ($t \in [1, n/b]$), the step size is a constant, which can be covered by Theorem 4. After the first

[5] `http://lsec.cc.ac.cn/chinese/lsec/LSSC-IVintroduction.pdf.`

epoch ($t \geq \tilde{t}_0$), we then can compute $\eta_0 = \eta(\tilde{t}_0) = m_2/\tilde{t}_0$, *i.e.*, the scalar of lower bound m_2/t in Theorem 4 will be $m_2 = \tilde{t}_0\eta_0 \in \tilde{t}_0\{0.1, 0.5, 1, 5, 10, 15\}$. After one epoch or a few epochs, the scalar m_2 of the lower bound is competitive to $1/\Lambda$.

6.3 Regularized Logistic Regression

First, we empirically test the above step sizes on the regularized logistic regression problems, which is strongly convex for regularization parameter $\Lambda > 0$

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-b_i \langle a_i, x\rangle)) + \frac{\Lambda}{2}\|x\|^2,$$

where $\{a_i, b_i\}_{i=1}^n$ is a training sample set with $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, +1\}$. We use the two binary classification data sets w8a ($n = 49749, d = 300$) and rcv1.binary ($n = 20242, d = 47236$) from LIBSVM[6], where the 0.75 partition of the data is used for training, and the remaining is for testing. The regularizer parameter $\Lambda = 10^{-4}$, batch size $b = 128$, the outer loop $N = 120$ and the inner loop $m^{'} = n/128$.

We plot the average results of 5 runs on w8a in Figure 3. For the x-axis, we always use the number of epochs calculated. The y-axis is the value of the loss function on the training data set (left) and the accuracy (the percent of correctly classified data sets) on the testing data set (right). For $1/t$-stepsize, the best initial step size $\eta_0 = 5$, and we apply the same initial step size for the other step sizes. Other important parameters are set as $s = 3$, $T_0 = 2$, and $\theta = 1.2$. From Figure 3, we can see that the exponentially decaying step size (Grow-Exp) performs better than $1/t$-stepsize on training loss and accuracy. Our proposed $1/t$ Fix-period and $1/t$ Grow-period both achieve good performance than $1/t$-stepsize. In addition, $1/t$ Grow-Exp gets higher accuracy than Grow-Exp.

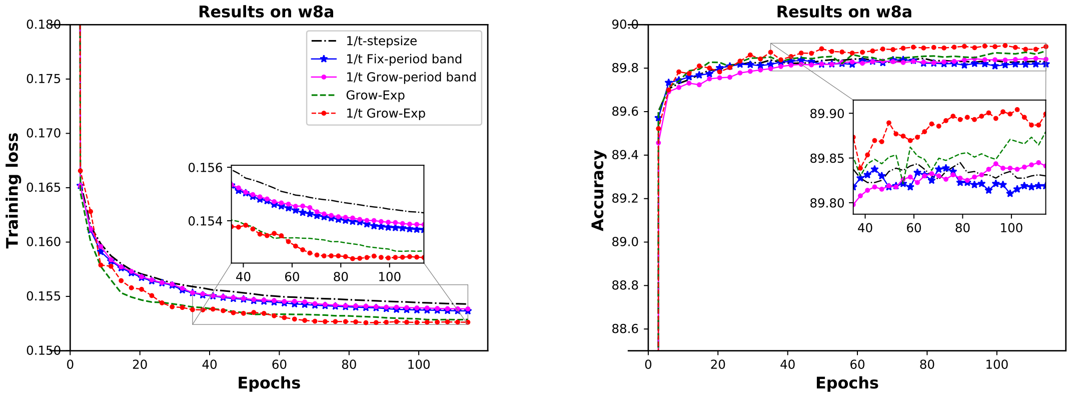


Figure 3: Results for regularized logistic regression

In Figure 4, we report the average results of 5 runs on rcv1.binary. The best-tuned initial step size η_0 is 10 for $1/t$-stepsize, and we use the same initial step size for other step size schedules. The value of θ is 1.3 for $1/t$ Grow-Exp, and other parameters are the same as w8a. We achieve a similar performance as Figure 3. From Figures 3 and 4, the

[6] `https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/`

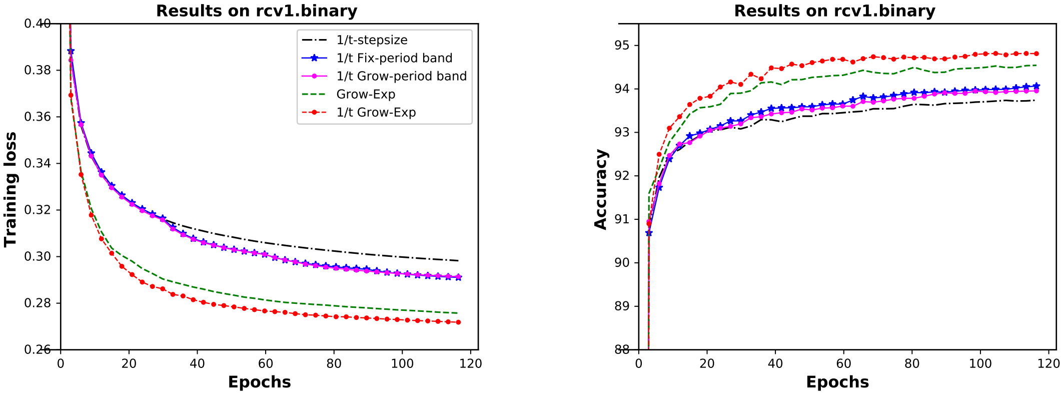


Figure 4: Results for regularized logistic regression

Grow-Exp step sizes significantly improve the performance of Epoch-SGD over $1/t$-stepsize. This implies that the relatively large step size at the initial iterations possibly makes the algorithm drop rapidly. We also observe that the proposed $1/t$ Grow-Exp step size, based on the $1/t$ up-down policy, yields better performance compared to the Grow-Exp step size.

6.4 Deep Neural Network and Residual Neural Network

In this subsection, we conduct experiments on some standard data sets, *e.g.*, MNIST and CIFAR-100.

First of all, we test on a fully-connected 3-layer (784-500-300-10) neural network to train **MNIST**[7], consisting of a training set of 60000 images with 28x28 pixels and a testing set of 10000 images in 10 classes. The batch size $b = 128$, the outer loop $N = 120$ and the inner loop $m^{'} = n/128$. For the $1/t$-stepsize, the best η_0 is achieved at $\eta_0 = 0.5$ based on its accuracy. For the $1/t$ Fix-period band and $1/t$ Grow-period band, η_0 is the same as that of the $1/t$-stepsize. We choose $s = 3$, that is $\eta(t) \in [\eta_0/t, 3\eta_0/t]$. For Grow-Exp, the parameters are set as $\eta_0 = 0.5, T_0 = 10$. For $1/t$ Grow-Exp, we set $\theta = 1.3$, and other parameters are the same as Grow-Exp. The average results of 5 runs are given in Figure 5. It is easy to see that the Grow-Exp type step size achieves better performance compared to $1/t$-stepsize, $1/t$ Fix-period band, and $1/t$ Grow-period band. Besides, our proposed $1/t$ Grow-Exp achieves lower training loss than Grow-Exp.

Next, we implement the above five step sizes on **ResNet-18** (He et al., 2016) with **CIFAR-100**[8]. The CIFAR-100 data set consists of 60000 32x32 color images in 100 classes, 50000 images for training, and the remaining 10000 images for testing. For $1/t$-stepsize, we set $\eta(t) = \eta_0/(1 + t/10)$, where $\eta_0 \in \{0.1, 0.5, 1, 5, 10, 15\}$. The best performance of $1/t$-stepsize is achieved at $\eta_0 = 1$. In this case, the bandwidth $s = 3$. Other important parameters are the same as the experiment in DNNs. For Grow-Exp, $\eta_0 = 0.5$ and $T_0 = 10$. For $1/t$ Grow-Exp, $\eta_0 = 0.5$, $T_0 = 10$ and $\theta = 1.3$.

[7] `http://deeplearning.net/data/mnist/`

[8] `https://www.cs.toronto.edu/~kriz/cifar.html`

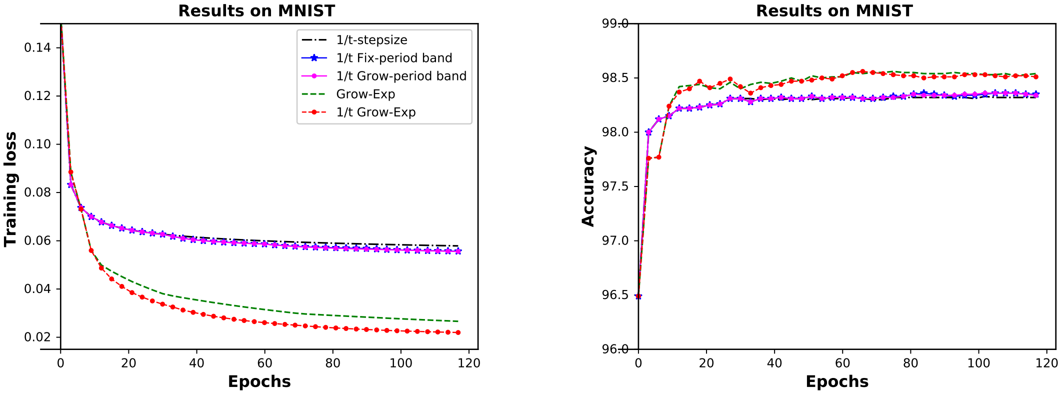


Figure 5: Results on deep neural networks (DNNs)

We repeat the training process 5 times, and the average results (the left is the testing loss function, and the right is the accuracy of the testing data set) are presented in Figure 6. In this case, we see that the sudden increase of the $1/t$ Fix-period band and $1/t$ Grow-period band may lead to a short-term negative effect but overall helps these step-sizes outperform the $1/t$-stepsize at the long-term training. Especially the $1/t$ Grow-period band performs better than the $1/t$ Fix-period band. The frequently going up and down makes the $1/t$ Fix-period band less stable than the $1/t$ Grow-period band. This may be the main reason for this phenomenon.

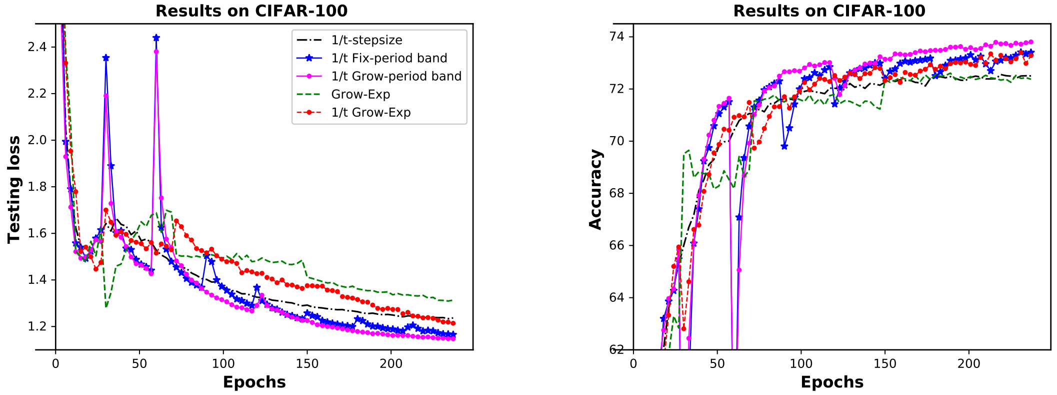


Figure 6: Results on ResNet-18

Another observation from Figure 6 is that the Grow-Exp step size does not work well as Section 6.3. This may be because a growing number of epochs in Grow-Exp might reduce its generalization at the final stage of each cycle. Nevertheless, we find that $1/t$ Grow-Exp yields better performance than Grow-Exp. Indeed, the $1/t$-stepsize scheme may not be the best baseline for solving nonconvex problems. We take it as an example and empirically demonstrate that the step size based on bandwidth is potential and often helps in practice.

6.5 Additional Experiments on Other Algorithms and Step Sizes

For further investigation, more experiments are carried out to compare different step sizes on Epoch-SGD and other default algorithms in deep learning, including SGD with momentum (called Momentum for short), averaged SGD (called ASGD) (Polyak and Juditsky, 1992) and Adam (Kingma and Ba, 2015). We use two popular data sets: CIFAR-10[9] and CIFAR-100 for image classifications. The CNN architectures VGG-16 (Simonyan and Zisserman, 2015) and ResNet-18 (He et al., 2016) are adopted for training CIFAR-10 and CIFAR-100, respectively.

In addition to the step sizes tested in the above subsections, we implement the popular exponentially decaying step size with a fixed period T_0 (called **Fix-Exp**), which has been discussed in Section 3.1:

$$\eta(t) = \eta_i = \eta_0/10^i, t \in [T_i, T_{i+1}),\ T_{i+1} - T_i = T_0, i \in \mathbb{N}. \tag{21}$$

Let $\eta_{\min}^i = \eta_i$ for $i \in \mathbb{N}^+$ and we define $\eta_{\max}^i = \theta \eta_{\min}^{i-1}$ where $\theta \in (1, 1.5]$. Based on (21), we propose the following step size (called **$1/t$ Fix-Exp**):

$$\eta(t) = \frac{\hat{A}_i}{\hat{B}_i t + 1} \in [\eta_{\min}^i, \eta_{\max}^i],\ t \in [T_i, T_{i+1})\,, T_{i+1} - T_i = T_0. \tag{22}$$

This is similar to $1/t$ Grow-Exp, but the number of epochs per cycle is fixed and is the same. Besides, we also implement the two cyclical step sizes: triangular policy (Smith, 2017) and cosine annealing (Loshchilov and Hutter, 2017).

Firstly, we test on VGG-16 for training CIFAR-10. The baseline initial step size is set as $\eta_0 = 1$ for SGD and ASGD, $\eta_0 = 0.1$ for Momentum, and $\eta_0 = 0.001$ for Adam. For Momentum, $\beta = 0.9$. In Adam, $(\beta_1, \beta_2) = (0.9, 0.99)$ is used. The best-tuned value of weight decay is 10^{-4} for SGD and ASGD, 5×10^{-4} for Momentum and 10^{-5} for Adam. The common parameters $N = 120$ and $b = 128$ for all algorithms. We perform the above algorithms with Fix-Exp ($T_0 = 30$) and $1/t$ Fix-Exp ($T_0 = 30, \theta = 1.3$). The average results of five runs are presented in Figure 7. We find that $1/t$ Fix-Exp overall shows better performance than Fix-Exp on SGD, Momentum, and ASGD. However, the results of Adam based on Fix-Exp and $1/t$ Fix-Exp almost coincide, which implies that the up-down policy may not work well for Adam.

Besides, we test Momentum with the following step sizes: (**1**) $1/t$-stepsize ($\eta(t) = \eta_0/(1+t/5)$); (**2**) $1/t$ Fix-period band ($t_{i+1} - t_i = 30, s = 3$); (**3**) Fix-Exp ($T_0 = 30$); (**4**) $1/t$ Fix-Exp ($T_0 = 30, \theta = 1.3$); (**5**) triangular policy based on (21), called "Triangular" (rise and fall ratio is 1.5); (**6**) cosine annealing, called "Cosine" (we use the last iterations as the initial point of restart cycle). All step sizes are best tuned with $\eta_0 = 0.1$, and the period of each cycle is 30 for triangular policy and cosine annealing. The average results of 5 runs are shown in Figure 8. We observe that $1/t$ Fix-Exp shows its advantages over $1/t$-stepsize, $1/t$ Fix-period band, Fix-Exp, and triangular policy after 80 epochs, and the final results are comparable to cosine annealing.

Next, we implement the above algorithms with Fix-Exp and $1/t$ Fix-Exp on ResNet-18 for training CIFAR-100. We report the average results of five runs in Figure 9. The budget

[9] `http://www.cs.toronto.edu/~kriz/cifar.html`

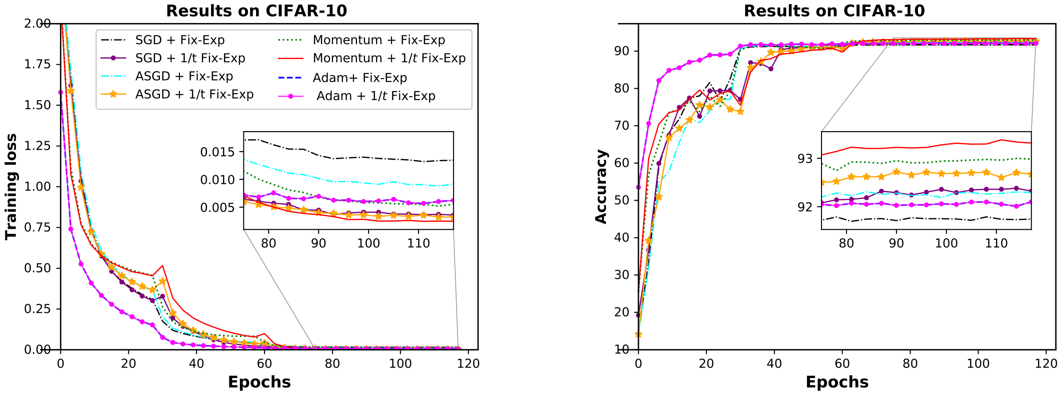


Figure 7: Results on VGG-16 for CIFAR-10

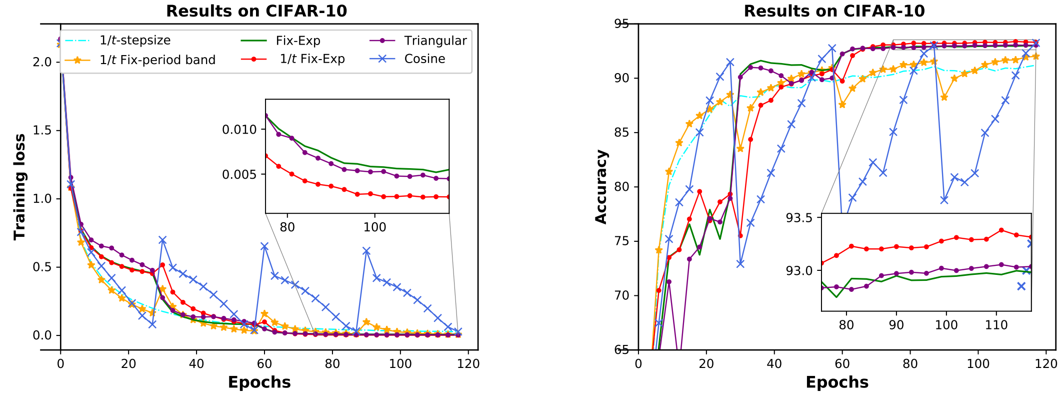


Figure 8: Results of different step sizes for CIFAR-10

of the outer iteration $N = 240$ and the period of each cycle $T_0 = 60$. The other parameters are chosen the same as the experiments on CIFAR-10. Similarly, we can conclude that the up-down policy in $1/t$ Fix-Exp leads to improvements after the second cycle over Fix-Exp on SGD, ASGD, and Momentum, respectively. We also observe that the up-down policy does not work for Adam but at least does not make Adam worse.

In Figure 10, we report the average results of five runs on the above step sizes for Momentum. The period for $1/t$ Fix-period band is $t_{i+1} - t_i = 60$. For Fix-Exp,$1/t$ Fix-Exp, triangular policy (the ratio of rising and fall is 2), cosine annealing, the period per cycle $T_0 = 60$ and other parameters are the same as those of CIFAR-10. As the figures show, $1/t$ Fix-Exp can reach lower testing loss and higher accuracy than the other step sizes after about 150 epochs.

7. Conclusion

We have proposed a bandwidth-based framework for SGD that allows the step size to vary in a banded region and be non-monotonic. Our purpose is not to focus on one specific step

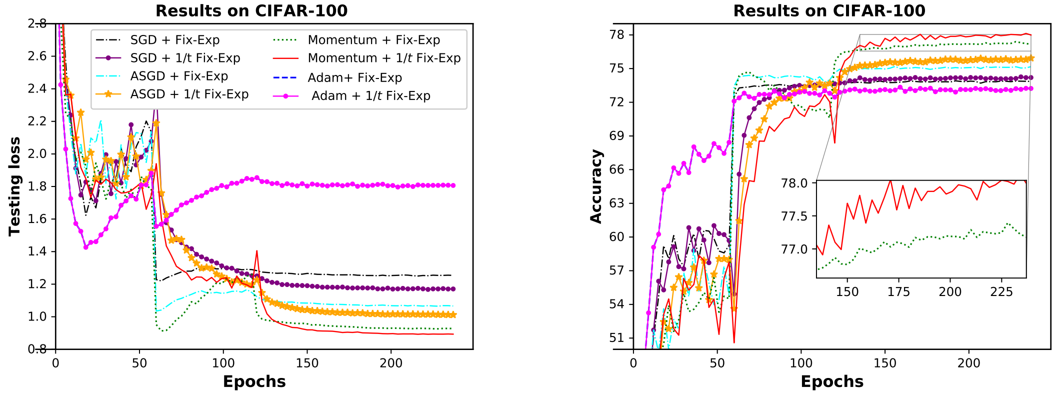


Figure 9: Results on ResNet-18 for CIFAR-100

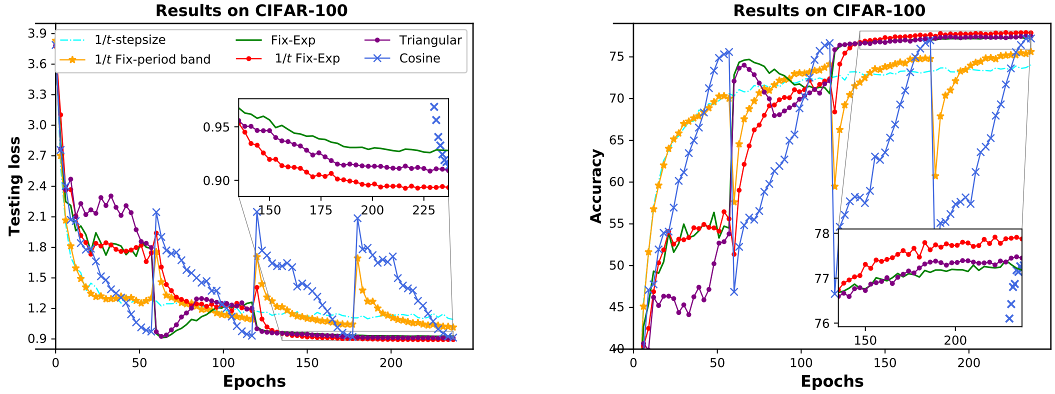


Figure 10: Results of different step sizes on CIFAR-100

size but to provide a uniform convergence framework for many non-monotonic step sizes within one class. We have investigated the conditions where the SGD method achieves an $\mathcal{O}(1/T)$ convergence rate and have extended its boundaries at the initial iterations, which could be useful in practical applications. Moreover, we have discussed three situations covering most general cases and provided explicit error bounds. In some cases, such as $\eta(t) = \eta_0/(t\ln(t))$ and $\eta_0/\sqrt{t}$, we have achieved better upper bounds than theorem 10 of Nguyen et al. (2019a). The bandwidth-based step size with different lower and upper bounds orders often gets worse convergence rates than its boundaries. The convergence rate for some existing step sizes such as exponentially decaying step size (Hazan and Kale, 2014), cyclical policy (Smith, 2017), sine-wave annealing (An et al., 2017) and cosine with restart (Loshchilov and Hutter, 2017) can be revealed by our analysis if their boundaries satisfy the conditions discussed in this paper.

The bandwidth-based framework gives us a lot of freedom when designing the step size with additional advantages. We have proposed four non-monotonic step sizes based on $1/t$-stepsize and exponentially decaying step size. The numerical results empirically demonstrate their efficiency and potential for solving convex and nonconvex problems, especially for nonconvex problems (*e.g.*, deep neural networks, and convolutional neural networks). Besides, we found that the bandwidth-based step size works for averaged SGD and momentum. It is worthwhile to explore SGD and its variants (*e.g.*, momentum) with bandwidth-based step size on nonconvex optimization in the future. We believe that the bandwidth scheme can inspire possibilities for designing more effective step sizes for nonconvex optimization. Besides, in the current experiments, the faction $\tau = M/m$ is tuned in a heuristic way. It will be interesting to portray the relationship or find the exact values for the lower and upper bounds in the future.

The proposed schedule leads to a new prospect based on step size, which might help avoid the saddle points. As we can see, a great effort has been made to avoid saddle points by incorporating the noise into the search direction per iteration (Ge et al., 2015; Jin et al., 2017; Du et al., 2017). Whether incorporating the noise or intermediate increasing to step size would help avoid the saddle points and bad local minimizers will be an exciting subject for future research.

Acknowledgments

The authors would like to thank three anonymous referees and the editor for handling this paper. We are grateful for their comments and suggestions, which led to some important improvements. The authors thank Dr. Shuxiong Wang for polishing the original version of the manuscript. This research is partially supported by the National Natural Science Foundation of China 11331012, 11688101, and 12288201.

Appendix A. Proofs of the Results in Section 2

Proof (of Lemma 1) Due to the μ-strongly convex property of the objective function $f(x)$ for $x \in \mathbb{R}^d$ and $\nabla f(x^*) = 0$, let $x = x$ and $\hat{x} = x^*$ in (3), we have

$$\begin{aligned} f(x) &\geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\mu}{2} \|x - x^*\|^2 \\ &\geq f(x^*) + \frac{\mu}{2} \|x - x^*\|^2 . \end{aligned} \tag{23}$$

Besides, letting $x = x^*$ and $\hat{x} = x$ in (3) gives

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2} \|x - x^*\|^2 .$$

Re-arranging the above inequality, we have

$$\langle \nabla f(x), x - x^* \rangle \geq f(x) - f(x^*) + \frac{\mu}{2} \|x - x^*\|^2 . \tag{24}$$

as required. ■

Proof (**of Lemma 2**) Considering the SGD algorithm defined by (2), we have

$$\begin{aligned}\mathbb{E}[\|x_{t+1}-x^*\|^2\,|\mathcal{F}_t] &= \mathbb{E}[\|x_t-\eta(t)g_t-x^*\|^2\,|\mathcal{F}_t] \\ &= \mathbb{E}[\|x_t-x^*\|^2\,|\mathcal{F}_t]-\mathbb{E}[2\eta(t)\langle g_t,x_t-x^*\rangle\,|\mathcal{F}_t]+\eta(t)^2\mathbb{E}[\|g_t\|^2\,|\mathcal{F}_t] \quad (25) \\ &= \|x_t-x^*\|^2-2\eta(t)\langle\nabla f(x_t),x_t-x^*\rangle+\eta(t)^2\mathbb{E}[\|g_t\|^2\,|\mathcal{F}_t],\end{aligned}$$

where the last equality uses the fact that the stochastic gradient g_t is an unbiased estimation of $\nabla f(x_t)$ at x_t. Assumption 3 holds that there exists a constant $L_f > 0$ such that

$$\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \le 4L_f(f(x_t)-f^*)+2\sigma^2. \tag{26}$$

Since f is μ-strongly convex, by Lemma 1, the inequality (24) holds. Let $x=x_t$ in (24), together with (26), then (25) can be evaluated by

$$\begin{aligned}\mathbb{E}[\|x_{t+1}-x^*\|^2 \mid \mathcal{F}_t] \le\, &(1-\mu\eta(t))\|x_t-x^*\|^2 \\ &+2\eta(t)^2\sigma^2+(4L_f\eta(t)^2-2\eta(t))[f(x_t)-f(x^*)].\end{aligned} \tag{27}$$

Let $n_0 := \sup\{t\in\mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. For $t > n_0$, we have $4L_f\eta(t)^2-2\eta(t)\le 0$. Then the inequality (27) can be

$$\mathbb{E}[\|x_{t+1}-x^*\|^2 \mid \mathcal{F}_t] \le (1-\mu\eta(t))\|x_t-x^*\|^2+2\eta(t)^2\sigma^2. \tag{28}$$

Let $\chi_{n_0}=\max\limits_{1\le t\le n_0}\left\{4L_f\eta(t)^2-2\eta(t)\right\}$ and $f_{n_0}=\max\limits_{1\le t\le n_0}\{f(x_t)-f(x^*)\}$. Because n_0 is supposed to be a constant which is independent of T, the sequence $\{f(x_t)-f(x^*)\}_{t=1}^{n_0}$ is bounded by a constant f_{n_0}. For $1\le t\le n_0$, we have

$$\mathbb{E}[\|x_{t+1}-x^*\|^2 \mid \mathcal{F}_t] \le (1-\mu\eta(t))\|x_t-x^*\|^2+2\eta(t)^2\sigma^2+\chi_{n_0}f_{n_0}. \tag{29}$$

For $t > n_0$, taking expectations again and applying the recursion of (28) and (29) from 1 to t, we have

$$\begin{aligned}&\mathbb{E}[\|x_{t+1}-x^*\|^2] \\ &\le \prod_{l=1}^{t}(1-\mu\eta(l))\|x_1-x^*\|^2+2\sigma^2\sum_{l=1}^{t}\eta(l)^2\prod_{u>l}^{t}(1-\mu\eta(u))+\chi_{n_0}f_{n_0}\sum_{l=1}^{n_0}\prod_{u>l}^{t}(1-\mu\eta(u)) \\ &\le \exp\left(-\mu\sum_{l=1}^{t}\eta(l)\right)\Delta_{n_0}^0+2\sigma^2\sum_{l=1}^{t}\eta(l)^2\exp\left(-\mu\sum_{u>l}^{t}\eta(u)\right),\end{aligned} \tag{30}$$

where $\Delta_{n_0}^0=\|x_1-x^*\|^2+\frac{n_0\chi_{n_0}f_{n_0}}{\exp\left(-\mu\sum_{l=1}^{n_0}\eta(l)\right)}$. The last inequality of (30) uses the fact that $1+x\le\exp(x)$ for all $x\in\mathbb{R}$. Note that the coefficient $1-\mu\eta(l)$ of $\mathbb{E}[\|x_l-x^*\|^2]$ may be negative for the previous finite terms $1\le l\le t$, so the recursive process starting from $t=1$ is not appropriate. However, because $\exp(-\mu\eta(l))$ is always positive, we might as well relax the upper bound of $\mathbb{E}[\|x_{t+1}-x^*\|^2]$ as (30).

■

Appendix B. Proofs of the Results in Section 3

Proof (**of Theorem 1**) In this case, the sequence of step size $\eta(t)$ satisfies that

$$0 < \frac{m}{t} \leq \eta(t) \leq \frac{M}{t}, \text{ for } 1 \leq t \leq T.$$

It is known that

$$\ln(t+1) \leq \sum_{l=1}^{t} \frac{1}{l} \leq \ln(t) + 1 \tag{31a}$$

and

$$\int_{u=l}^{t+1} \frac{du}{u} \leq \sum_{u=l}^{t} \frac{1}{u} \leq \int_{u=l-1}^{t} \frac{du}{u}, \text{ for any } l > 1. \tag{31b}$$

Then we have

$$\sum_{l=1}^{t} \eta(l) \geq \sum_{l=1}^{t} \frac{m}{l} \geq m \ln(t+1) \tag{32a}$$

and

$$\sum_{u>l}^{t} \eta(u) \geq \sum_{u>l}^{t} \frac{m}{u} = \sum_{u=1}^{t} \frac{m}{u} - \sum_{u=1}^{l} \frac{m}{u} \geq m(\ln(t+1) - \ln(l) - 1). \tag{32b}$$

Let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. In this case, when $t \geq 2ML_f$, we have

$$\eta(t) \leq \frac{M}{t} \leq \frac{1}{2L_f}. \tag{33}$$

Thus, $n_0 \leq 2ML_f$ which is independent of T.

From Lemma 2, we know that for $T > n_0$, $\mathbb{E}[\|x_{T+1} - x^*\|^2]$ can be estimated as

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \leq \Gamma_T^2 + \Gamma_T^2, \tag{34}$$

where

$$\Gamma_T^1 := \exp\left(-\mu \sum_{l=1}^{T} \eta(l)\right) \Delta_{n_0}^0, \quad \Gamma_T^2 := 2\sigma^2 \sum_{l=1}^{T} \eta(l)^2 \exp\left(-\mu \sum_{u>l}^{T} \eta(u)\right).$$

Applying (32a) into Γ_T^1, we can achieve that

$$\Gamma_T^1 \leq \exp\left(-\mu m \ln(T+1)\right) \Delta_{n_0}^0 = \frac{\Delta_{n_0}^0}{(T+1)^{\mu m}}. \tag{35}$$

Now, we proceed to obtain the upper bound for Γ_T^2. Using the upper bound of $\eta(t)$ and (32b) gives

$$\begin{aligned}
\Gamma_T^2 &= 2\sigma^2 \sum_{l=1}^{T} \eta(l)^2 \exp\left(-\mu \sum_{u>l}^{T} \eta(u)\right) \\
&\leq 2\sigma^2 \sum_{l=1}^{T} \eta(l)^2 \exp(-\mu m(\ln(T+1) - \ln(l) - 1)) \\
&\leq \frac{2\sigma^2 M^2 \exp(\mu m)}{(T+1)^{\mu m}} \sum_{l=1}^{T} \frac{1}{l^2} \cdot \exp(\mu m \ln(l)) \leq \frac{2\sigma^2 M^2 \exp(\mu m)}{(T+1)^{\mu m}} \sum_{l=1}^{T} \frac{l^{\mu m}}{l^2}.
\end{aligned}$$

If $m = \frac{1}{\mu}$ then

$$\Gamma_T^2 \leq 2\sigma^2 M^2 \exp(1) \cdot \frac{\ln(T) + 1}{T + 1}.$$

However, when $m \neq \frac{1}{\mu}$, we have that

$$\sum_{l=1}^{T} \frac{l^{\mu m}}{l^2} = \sum_{l=1}^{T} l^{(\mu m - 2)} \leq \int_{l=1}^{T+1} l^{(\mu m-2)} dl + 1, \tag{36}$$

then

$$\Gamma_T^2 \leq \frac{2\sigma^2 M^2 \exp(\mu m)}{(\mu m - 1)} \cdot \frac{(T+1)^{\mu m - 1} + \mu m - 2}{(T+1)^{\mu m}}.$$

Substituting the upper bounds of Γ_1^T and Γ_2^T into (34), we get the desired result.

■

Proof (of Theorem 2) Let $n_1 := \sup \left\{ t \in \mathbb{N}^+ : \eta(t) > \frac{1}{4L_f} \right\}$. In this case, $\frac{m}{t} \leq \eta(t) \leq \frac{M}{t}$ which implies that $\delta_1(t) = \delta_2(t) = 1/t$. When $t \geq 4ML_f$, we have $\eta(t) \leq 1/(4L_f)$. Thus we know $n_1 \leq 4ML_f$, which is independent of T. Let $\chi_{n_1} = \max\limits_{1 \leq t \leq n_1} \left\{ 4L_f \eta(t)^2 - 2\eta(t) \right\}$ and $f_{n_1} = \max\limits_{1 \leq t \leq n_1} \{ f(x_t) - f(x^*) \}$. Because n_1 is a constant, the sequence $\{ f(x_t) - f(x^*) \}_{t=1}^{n_1}$ can be bounded by f_{n_1} which is a constant. For $t > n_1$, $4L\eta(t)^2 - 2\eta(t) \leq -\eta(t)$, then the inequality (27) in Lemma 2 will be

$$\begin{aligned}
&\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \\
\leq &(1 - \mu\eta(t)) \|x_t - x^*\|^2 + 2\eta(t)^2\sigma^2 + (4L\eta(t)^2 - 2\eta(t))[f(x_t) - f(x^*)] \\
\leq &(1 - \mu\eta(t)) \|x_t - x^*\|^2 + 2\eta(t)^2\sigma^2 - \eta(t)[f(x_t) - f(x^*)].
\end{aligned} \tag{37}$$

Shifting $[f(x_t) - f(x^*)]$ to the left side and $\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t]$ to the right side, we obtain

$$\eta(t)[f(x_t) - f(x^*)] \leq (1 - \mu\eta(t)) \|x_t - x^*\|^2 - \mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] + 2\eta(t)^2\sigma^2.$$

Applying the lower bound of $\eta(t)$ into the left side and then dividing the above inequality by $m\delta_1(t)\delta_1(t + t_0)$ $(t_0 \in \mathbb{N})$ gives

$$\begin{aligned}
\frac{f(x_t) - f(x^*)}{\delta_1(t + t_0)} \leq &\frac{1}{m} \left\{ \left(\frac{1}{\delta_1(t)\delta_1(t + t_0)} - \frac{\mu m}{\delta_1(t + t_0)} \right) \|x_t - x^*\|^2 - \frac{\mathbb{E}[\|x_{t+1} - x^*\|^2]}{\delta_1(t)\delta_1(t + t_0)} \right\} \\
&+ \frac{2\eta(t)^2\sigma^2}{m\delta_1(t)\delta_1(t + t_0)}.
\end{aligned}$$

Summing the above inequality for t from n_1 to T, we get that

$$
\begin{aligned}
&\mathbb{E}\left[f\left(\frac{\sum_{t=1}^{T}\frac{1}{\delta_1(t+t_0)}x_t}{\sum_{t=1}^{T}\frac{1}{\delta_1(t+t_0)}}\right)-f(x^*)\right]\\
\leq\ & \frac{1}{\sum_{t=1}^{T}\frac{1}{\delta_1(t+t_0)}}\left(\sum_{t=1}^{n_1}\mathbb{E}\left[\frac{f(x_t)-f(x^*)}{\delta_1(t+t_0)}\right]+\sum_{t=n_1+1}^{T}\mathbb{E}\left[\frac{f(x_t)-f(x^*)}{\delta_1(t+t_0)}\right]\right)\\
\leq\ & \frac{1}{\sum_{t=1}^{T}\frac{m}{\delta_1(t+t_0)}}\sum_{t=n_1+1}^{T}\left\{\left(\frac{1}{\delta_1(t)\delta_1(t+t_0)}-\frac{\mu m}{\delta_1(t+t_0)}\right)\mathbb{E}[\|x_t-x^*\|^2]-\frac{\mathbb{E}[\|x_{t+1}-x^*\|^2]}{\delta_1(t)\delta_1(t+t_0)}\right\}\\
&+\frac{1}{\sum_{t=1}^{T}\frac{1}{\delta_1(t+t_0)}}\sum_{t=1}^{n_1}\frac{f_{n_1}}{\delta_1(t+t_0)}+\frac{1}{\sum_{t=1}^{T}\frac{m}{\delta_1(t+t_0)}}\sum_{t=n_1+1}^{T}\frac{2\eta(t)^2\sigma^2}{\delta_1(t)\delta_1(t+t_0)},
\end{aligned}
\tag{38}
$$

where the first inequality follows from the well-known *Jensen inequality* if f is convex. If μm satisfies the following condition:

$$
\mu m \geq \frac{1}{\delta_1(t+1)}-\frac{\delta_1(t+t_0+1)}{\delta_1(t)\delta_1(t+t_0)}\,(\forall\, t>n_1),
\tag{39}
$$

by simple calculations, we can show that the coefficient of $\mathbb{E}[\|x_t-x^*\|^2]$ $(t>n_1)$ is non-positive. Taking the form $\delta_1(t)=1/t$, if $\mu m\geq 1$, the condition (39) will hold. Then let $\hat{x}_T=\frac{\sum_{t=1}^{T}(t+t_0)x_t}{S_1}$ and $S_1=\sum_{t=1}^{T}(t+t_0)$, applying the inequality (38), we get

$$
\begin{aligned}
\mathbb{E}\left[f\left(\hat{x}_T\right)-f(x^*)\right]\leq\ & \frac{(n_1+t_0+1)}{mS_1}\left(n_1+1-\mu m\right)\mathbb{E}[\|x_{n_1+1}-x^*\|^2]+\frac{(1+t_0)(n_1+t_0)f_{n_1}}{2S_1}\\
&+\frac{2\sigma^2M^2}{mS_1}\sum_{t=n_1+1}^{T}\frac{t(t+t_0)}{t^2}.
\end{aligned}
\tag{40}
$$

By Lemma 2, for $1\leq t\leq n_1$, we have that

$$
\mathbb{E}[\|x_{t+1}-x^*\|^2\mid\mathcal{F}_t]\leq\left(1-\mu\eta(t)\right)\|x_t-x^*\|^2+2\eta(t)^2\sigma^2+\chi_{n_1}f_{n_1}.
\tag{41}
$$

Applying the recursion of (41) for t from 1 to n_1 and taking expectation again gives

$$
\begin{aligned}
&\mathbb{E}[\|x_{n_1+1}-x^*\|^2]\\
\leq\ & \exp\left(-\mu\sum_{t=1}^{n_1}\eta(t)\right)\|x_1-x^*\|^2+2\sigma^2\sum_{l=1}^{n_1}\eta(l)^2\exp\left(-\mu\sum_{u>l}^{n_1}\eta(u)\right)\\
&+\chi_{n_1}f_{n_1}\sum_{l=1}^{n_1}\exp\left(-\mu\sum_{u>l}^{n_1}\eta(u)\right)\\
\leq\ & \exp\left(-\mu m\ln(n_1+1)\right)\|x_1-x^*\|^2+2\sigma^2M^2\sum_{l=1}^{n_1}\frac{1}{l^2}+n_1\chi_{n_1}f_{n_1}\\
\leq\ & \frac{\|x_1-x^*\|^2}{(n_1+1)^{\mu m}}+4\sigma^2M^2+n_1\chi_{n_1}f_{n_1}.
\end{aligned}
$$

Incorporating the above bound of $\mathbb{E}[\|x_{n_1+1} - x^*\|^2]$ into (40), we can obtain that

$$\begin{aligned}\mathbb{E}\left[f\left(\hat{x}_T\right) - f(x^*)\right] \le & \frac{\left(n_1 + t_0 + 1\right)\left(n_1 + 1 - \mu m\right)}{mS_1}\left[\frac{\|x_1 - x^*\|^2}{(n_1+1)^{\mu m}} + 4\sigma^2 M^2 + n_1 \chi_{n_1} f_{n_1}\right] \\ & + \frac{(1+t_0)(n_1+t_0)f_{n_1}}{2S_1} + \frac{2\sigma^2 M^2}{mS_1}(T - n_1 + t_0 \ln(T/n_1)) \\ & = \frac{1}{mS_1}\left[\upsilon_1 \Delta^0_{n_1} + \frac{\upsilon_2}{2} m f_{n_1} + 2\sigma^2 M^2 (T - n_1 + t_0 \ln(T/n_1))\right],\end{aligned}$$

where $\hat{x}_T = \frac{\sum_{t=1}^{T}(t+t_0)x_t}{S_1}$, $S_1 = \frac{T(T+t_0)(t_0+1)}{2}$, $\Delta^0_{n_1} = \frac{\|x_1 - x^*\|^2}{(n_1+1)^{\mu m}} + 4\sigma^2 M^2 + n_1 \chi_{n_1} f_{n_1}$, $\upsilon_1 = \left(n_1 + t_0 + 1\right)\left(n_1 + 1 - \mu m\right)$ and $\upsilon_2 = (1+t_0)(n_1+t_0)$.

■

Proof (of Theorem 3) In this case, we assume that $\eta(t)$ satisfies conditions (A_1) and (B). Similar to Theorem 1, let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. We know $n_0 \le 2ML_f$, which is independent of T. Then for $T > n_0$, the conclusion of Lemma 2 is true.

Let $t^* = 1$ in (A_1), we have

$$\sum_{t=1}^{T} \eta(t) \ge C \ln(T+1),$$

then Γ^1_T defined by (9a) can be evaluated as follows

$$\Gamma^1_T = \exp\left(-\mu \sum_{l=1}^{T} \eta(l)\right) \Delta^0_{n_0} \le \frac{1}{(T+1)^{(\mu C)}} \Delta^0_{n_0}. \tag{42}$$

Recalling the definition of Γ^2_T in (9b), we have

$$\begin{aligned}\Gamma^2_T = 2\sigma^2 \sum_{t=1}^{T} \eta(t)^2 \exp\left(-\mu \sum_{u>t}^{T} \eta(u)\right) & \le 2\sigma^2 M^2 \sum_{t=1}^{T} \frac{1}{t^2} \cdot \exp\left(-\mu \sum_{u>t}^{T} \eta(u)\right) \\ \le 2\sigma^2 M^2 \sum_{t=1}^{T} \frac{1}{t^2} \cdot \exp\left(-\mu C \ln\left(\frac{T+1}{t+1}\right)\right) & = 2\sigma^2 M^2 \sum_{t=1}^{T} \frac{(t+1)^2}{t^2} \cdot \frac{(t+1)^{(\mu C - 2)}}{(T+1)^{(\mu C)}} \\ \le 8\sigma^2 M^2 \frac{\sum_{t=1}^{T}(t+1)^{(\mu C-2)}}{(T+1)^{(\mu C)}}, &\end{aligned}$$

where the first inequality uses condition (B), the second inequality follows from condition (A_1) for $t + 1 = t^*$, and the third inequality is derived from $(t+1)^2/t^2 \le 4$ for all $t \ge 1$.

No matter whether $\mu C > 2$ or not, we have $\sum_{t=1}^{T} t^{(\mu C-2)} \le \int_{t=1}^{T+1} t^{(\mu C-2)} dt + 1$. When $C > \frac{1}{\mu}$, then Γ_2^T can be estimated by

$$\Gamma^2_T \le \frac{8\sigma^2 M^2}{(\mu C - 1)} \cdot \frac{(T+2)^{(\mu C-1)} + \mu C - 2}{(T+1)^{(\mu C)}} \le \frac{8\sigma^2 M^2 \exp(1)}{(\mu C - 1)} \cdot \frac{1}{T+1} + \frac{8\sigma^2 M^2}{(T+1)^{(\mu C)}}. \tag{43}$$

Combining (42) and (43) together, we have

$$\begin{aligned}\mathbb{E}[\|x_{T+1}-x^*\|^2] &= \Gamma_T^1+\Gamma_T^2\\ &\le \frac{\Delta_{n_0}^0}{(T+1)^{(\mu C)}}+\frac{8\sigma^2M^2\exp(1)}{(\mu C-1)}\cdot\frac{1}{T+1}+\frac{8\sigma^2M^2}{(T+1)^{(\mu C)}}\\ &\le \frac{\Delta_{n_0}^0+8\sigma^2M^2}{(T+1)^{(\mu C)}}+\frac{8\sigma^2M^2\exp(1)}{(\mu C-1)}\cdot\frac{1}{T+1}.\end{aligned}$$

■

Proof (**of Theorem 4**) In this case, we assume that

$$\begin{aligned} m_1\le \eta(t)\le M_1, &\text{for } t\in[C_1T^p] \text{ and}\\ \frac{m_2}{t}\le\eta(t)\le\frac{M_2}{t}, &\text{for } t\in[T]\backslash[C_1T^p],\end{aligned}$$

where $p\in(0,1)$. Then we have

$$m_1C_1T^p\le\sum_{t=1}^{C_1T^p}\eta(t)\le M_1C_1T^p, \tag{44a}$$

$$m_2[\ln(T+1)-\ln(C_1T^p)-1]\le\sum_{C_1T^p+1}^{T}\eta(t)\le M_2[\ln(T)-\ln(C_1T^p)], \tag{44b}$$

where (44b) follows from inequalities (31b) and (32b). Let $n_0:=\sup\{t\in\mathbb{N}^+:\eta(t)>1/(2L_f)\}$. In this case, we assume that n_0 is a constant that is independent of T. Thus the results of Lemma 2 hold.

Recalling the definition of Γ_1^T in (9a) and applying (44a) and (44b), we have

$$\begin{aligned}\Gamma_1^T &= \exp\left(-\mu\sum_{t=1}^T\eta(t)\right)\Delta_{n_0}^0\\ &\le \exp\left(-\mu\left(m_1C_1T^p+m_2(\ln(T+1)-\ln(C_1T^p)-1)\right)\right)\Delta_{n_0}^0\\ &\le \frac{\exp(\mu m_2)\Delta_{n_0}^0}{T^{(\mu m_2(1-p))}\exp(\mu m_1C_1T^p)}\le\frac{\exp(\mu m_2)\Delta_{n_0}^0}{T^{(\mu m_2(1-p))}\,(\mu m_1C_1T^p+1)}\\ &\le \frac{\exp(\mu m_2)}{\mu m_1C_1}\cdot\frac{\Delta_{n_0}^0}{T^{(\mu m_2(1-p)+p)}}, \end{aligned}\tag{45}$$

where the last inequality dues to the fact that $\exp(x)\ge1+x$ for $x\in\mathbb{R}$. After that, we estimate Γ_2^T, divided into two parts as follows.

$$\begin{aligned}\Gamma_2^T &= 2\sigma^2\sum_{t=1}^T\eta(l)^2\exp(-\mu\sum_{u>t}^T\eta(u))\\ &\le 2\sigma^2\left[\sum_{t=1}^{C_1T^p}\eta(l)^2\exp(-\mu\sum_{u>t}^T\eta(u))+\sum_{t=C_1T^p+1}^T\eta(l)^2\exp(-\mu\sum_{u>t}^T\eta(u))\right].\end{aligned}$$

Let

$$\Theta_1 = \sum_{t=1}^{C_1T^p} \eta(l)^2 \exp(-\mu \sum_{u>t}^{T} \eta(u)),\ \Theta_2 = \sum_{t=C_1T^p+1}^{T} \eta(l)^2 \exp(-\mu \sum_{u>t}^{T} \eta(u)). \tag{46}$$

Then we have

$$\Gamma_2^T \leq 2\sigma^2(\Theta_1 + \Theta_2).$$

To get the upper bound of Γ_2^T, we will separately estimate Θ_1 and Θ_2. Let us evaluate Θ_1 first.

$$\begin{aligned}
\Theta_1 &= \sum_{t=1}^{C_1T^p} \eta(l)^2 \exp\left(-\mu \sum_{u>t}^{T} \eta(u)\right) \leq M_1^2 \sum_{t=1}^{C_1T^p} \exp\left(-\mu \sum_{u>t}^{T} \eta(u)\right) \\
&\leq M_1^2 \sum_{t=1}^{C_1T^p} \frac{\exp(\mu m_1 t)}{\exp\left(\mu m_1 C_1 T^p\right)} \exp\left(-\mu \sum_{u>C_1T^p}^{T} \eta(u)\right) \\
&\leq M_1^2 \exp\left(-\mu \sum_{u>C_1T^p}^{T} \eta(u)\right) \sum_{t=1}^{C_1T^p} \frac{\exp(\mu m_1 t)}{\exp\left(\mu m_1 C_1 T^p\right)} \\
&\leq \frac{M_1^2 \exp(\mu m_2)(C_1T^p)^{(\mu m_2)}}{(T+1)^{(\mu m_2)}} \cdot \frac{\int_{t=1}^{C_1T^p+1} \exp(\mu m_1 t) dt}{\exp(\mu m_1 C_1 T^p)} \\
&\leq \frac{M_1^2 \exp(\mu m_2)(C_1T^p)^{(\mu m_2)}}{(T+1)^{(\mu m_2)}} \cdot \frac{\exp(\mu m_1(C_1T^p+1)) - \exp(\mu m_1)}{\mu m_1 \exp(\mu m_1 C_1 T^p)} \leq \frac{M_1^2 \exp(\mu m_2) C_1^{(\mu m_2)}}{\mu m_1 T^{(\mu m_2)(1-p)}},
\end{aligned}$$

where the fourth inequality follows from (44b). Next, we bound Θ_2 as follows.

$$\begin{aligned}
\Theta_2 &= \sum_{t=C_1T^p+1}^{T} \eta(l)^2 \exp\left(-\mu \sum_{u>t}^{T} \eta(u)\right) \leq M_2^2 \sum_{t=C_1T^p+1}^{T} \frac{1}{t^2} \cdot \exp\left(-\mu m_2 \sum_{u>t}^{T} \frac{1}{u}\right) \\
&\leq M_2^2 \sum_{t=C_1T^p+1}^{T} \left(\frac{1}{t}\right)^2 \exp\left(-\mu m_2(\ln(T+1) - \ln(t+1) - 1)\right) \\
&\leq \frac{M_2^2 \exp(\mu m_2)}{(T+1)^{(\mu m_2)}} \cdot \sum_{t=C_1T^p+1}^{T} \frac{t^{(\mu m_2)}}{t^2} \leq \frac{M_2^2 \exp(\mu m_2)}{(T+1)^{(\mu m_2)}} \cdot \int_{t=C_1T^p}^{T+1} t^{(\mu m_2 - 2)} dt \\
&\leq \frac{M_2^2 \exp(\mu m_2)}{(T+1)^{(\mu m_2)}} \cdot \frac{(T+1)^{(\mu m_2 - 1)} - (C_1T^p)^{(\mu m_2 - 1)}}{\mu m_2 - 1} \\
&\leq \frac{M_2^2 \exp(\mu m_2)}{(\mu m_2 - 1)} \cdot \frac{1}{T+1},
\end{aligned}$$

where the fourth inequality follows from the fact that no matter whether $\mu m_2 > 2$ or not, we always have $\sum_{t=C_1T^p+1}^{T} \frac{t^{(\mu m_2)}}{t^2} \leq \int_{t=C_1T^p}^{T+1} t^{(\mu m_2-2)} dt$. The last inequality holds since $\kappa := (\mu m_2)(1-p) \geq 1$ and $p \in (0,1)$, we have $\mu m_2 \geq \frac{1}{(1-p)} > 1$. Thus

$$\Gamma_2^T = 2\sigma^2(\Theta_1 + \Theta_2) \leq \frac{2\sigma^2 M_1^2 \exp(\mu m_2) C_1^{(\mu m_2)}}{\mu m_1 T^{\kappa}} + \frac{2\sigma^2 M_2^2 \exp(\mu m_2)}{(\mu m_2 - 1)} \cdot \frac{1}{T+1}. \tag{47}$$

Hence, combining (45) and (47), we obtain the desired result. ■

Proof (**of Corollary 5**) In this case, we consider a class of cyclical step size defined by (13) where the period T_0 is fixed and the number of cycle $N = T/T_0$. The lower bound $\eta^i_{\min} = \eta_{\min} = m$ is a fixed constant, and the upper bound $\eta^i_{\max}$ is a decaying function with cycle i. Let S_i denote the enclosed area of the cyclical step size with its lower bound $\eta_{\min}$ and Q_i be the area between the upper bound and lower bound each cycle. We assume that $S_i/Q_i \geq \psi$ ($i \in [N]$) where $\psi \in (0, 1]$ is a constant. This assumption is easily satisfied by, for example, the triangular step size (Smith, 2017) with $\psi = 1/2$ and the cosine decaying step size (Loshchilov and Hutter, 2017) with $\psi = 1/2$. Under the same setting as Lemma 2, we have

$$\begin{aligned}
\mathbb{E}[\|x_{T+1} - x^*\|^2] &\leq \exp\left(-\mu \sum_{i=1}^{N} S_i - \mu m T\right) \Delta^0_{n_0} + 2\sigma^2 T_0 \sum_{i=1}^{N} (\eta^i_{\max})^2 \exp\left(-\mu \sum_{l>i} (S_i + mT_0)\right) \\
&\leq \exp\left(-\mu\psi \sum_{i=1}^{N} Q_i - \mu m T\right) \Delta^0_{n_0} + 2\sigma^2 T_0 \sum_{i=1}^{N} (\eta^i_{\max})^2 \exp\left(-\mu \sum_{l>i}^{N} (\psi Q_i + mT_0)\right).
\end{aligned}$$

If the upper bound $\eta^i_{\max} = M/2^{i-1}$, then we have $Q_i = T_0(M/2^{i-1} - m)$ and $\sum_{i=1}^{N} Q_i = 2MT_0(1 - 1/2^N) - mT$. Let $N = \lceil \log_2 T \rceil$, then $T_0 = T/N$. As we know, to ensure that $\eta_{\min} = m \leq \eta^i_{\max}$ for all t, the lower bound m is ought to sufficiently small that $m \leq 2M/T$.

$$\begin{aligned}
\mathbb{E}[\|x_{T+1} - x^*\|^2] &\leq \exp\left(-\mu\psi \sum_{i=1}^{N} Q_i - \mu m T\right) \Delta^0_{n_0} + 2\sigma^2 T_0 \sum_{i=1}^{N} (\eta^i_{\max})^2 \exp\left(-\mu \sum_{l>i}^{N} (\psi Q_i + mT_0)\right) \\
&\leq \frac{\Delta^0_{n_0}}{\exp\left(2\mu\psi M(T-1)/\log_2 T\right)} + \sigma^2 M^2 T_0 \sum_{i=1}^{N} 2^{-2i} \exp\left(-\frac{\mu\psi MT}{\log_2 T} \frac{2^{-i+1} - 2^{-N})}{1 - 2^{-1}}\right) \\
&\leq \frac{\Delta^0_{n_0}}{\exp\left(2\mu\psi M(T-1)/\log_2 T\right)} + 2\sigma^2 M^2 T_0 \sum_{i=1}^{N} 2^{-2i} \exp\left(-\frac{2\mu\psi MT}{\log_2 T}(2^{-i+1} - 2^{-N})\right) \\
&\overset{(a)}{\leq} \frac{\Delta^0_{n_0}}{\exp\left(2\mu\psi M(T-1)/\log_2 T\right)} + \mathcal{O}\left(\frac{\log_2 T}{T}\right). \qquad (48)
\end{aligned}$$

where (a) follows the fact that the individual term of the sum is at most $\mathcal{O}((\log^2 T/T^2)$ when $i = \max\left\{0, \lfloor \log_2\left(4\mu\psi M \cdot \frac{T}{\log_2 T}\right) \rfloor\right\}$.

Next we consider the decaying pattern of $\eta^i_{\max}$ is based on $1/t$, that is $\eta^i_{\max} = \frac{M}{iT_0}$ for $i \in [N]$. To make sure that $\eta_{\min} = m \leq \eta^i_{\max}$ for all i, we need $m \leq M/T$. Similar to the above case for the exponential decaying upper bound, we have

$$\begin{aligned}
\mathbb{E}[\|x_{T+1} - x^*\|^2] &\leq \exp\left(-\mu\psi M \ln N\right) \Delta^0_{n_0} + 2\sigma^2 M^2 \sum_{i=1}^{N} \frac{1}{i^2 T_0} \exp\left(-\mu\psi M(\ln N - \ln i)\right) \\
&\leq \frac{\Delta^0_{n_0}}{N^{\mu M\psi}} + \frac{2\sigma^2 M^2}{T_0 N^{\mu M\psi}} \sum_{i=1}^{N} \frac{i^{\mu M\psi}}{i^2} \leq \frac{\Delta^0_{n_0}}{N^{\mu M\psi}} + \frac{2\sigma^2 M^2}{T_0 N^{\mu M\psi}} \left(\int_{i=1}^{N} \frac{i^{\mu M\psi}}{i^2} + 1\right) \\
&\leq \frac{\Delta^0_{n_0}}{N^{\mu M\psi}} + \frac{2\sigma^2 M^2}{T_0 N^{\mu M\psi}} \left(\frac{N^{\mu M\psi - 1} - 1}{\mu M\psi - 1} + 1\right)
\end{aligned}$$

$$\leq \Delta_{n_0}^0 \left(\frac{T_0}{T}\right)^{\mu M \psi} + \frac{2\sigma^2 M^2}{\mu M \psi - 1} \frac{1}{T}. \tag{49}$$

The proof is complete. ■

Appendix C. Proofs of the Results in Section 4

Proof (of Theorem 5) In this case, we assume that $\eta(t)$ satisfies the following condition:

$$m\delta(t) \leq \eta(t) \leq M\delta(t),$$

where $\delta(t)$ satisfies (H3). Since $\frac{d\delta(t)}{dt} \leq 0$, it follows that

$$\sum_{u=1}^{t} \delta(u) \geq \int_{u=1}^{t+1} \delta(u) du, \tag{50a}$$

$$\sum_{u=l}^{t} \delta(u) \geq \int_{u=l}^{t+1} \delta(u) du. \tag{50b}$$

Let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. We assume that n_0 is a constant. Thus the conclusion of Lemma 2 holds. Now we invoke (8) and incorporate the lower and upper bounds of $\eta(t)$ into (8), then apply (50a) and (50b), consequently, for $t > n_0$, we have

$$\begin{aligned}
&\mathbb{E}[\|x_{t+1} - x^*\|^2] \\
\leq& \exp\left(-\mu m \sum_{l=1}^{t} \delta(t)\right) \Delta_{n_0}^0 + 2\sigma^2 M^2 \sum_{l=1}^{t} \delta(l)^2 \exp\left(-\mu m \sum_{u>l}^{t} \delta(u)\right) \\
\leq& \exp\left(-\mu m \sum_{l=1}^{t} \delta(t)\right) \Delta_{n_0}^0 + 2\sigma^2 M^2 \sum_{l=1}^{t} \delta(l)^2 \exp\left(-\mu m \left(\sum_{u=l}^{t} \delta(u) - \delta(l)\right)\right) \\
\leq& \exp\left(-\mu m \int_{u=1}^{t+1} \delta(u) du\right) \Delta_{n_0}^0 + 2\sigma^2 M^2 \sum_{l=1}^{t} \frac{\delta(l)^2 \exp(\mu m \delta(l))}{\exp\left(\mu m \int_{u=l}^{t+1} \delta(u) du\right)} \\
\leq& \frac{\Delta_{n_0}^0}{\exp\left(\mu m \int_{u=1}^{t+1} \delta(u) du\right)} + 2\sigma^2 M^2 \exp(\mu m \delta(1)) \sum_{l=1}^{t} \frac{\delta(l)^2}{\exp\left(\mu m \int_{u=l}^{t+1} \delta(u) du\right)}.
\end{aligned} \tag{51}$$

We consider the following three cases.

1. $\lim_{t\to\infty} \delta(t)t = 0$, that is for all $\epsilon > 0$, there exists an integer constant $t_\epsilon > 0$ such that $\delta(t)t < \epsilon$ for all $t \geq t_\epsilon$. To attain such a convergence rate, firstly, we want to prove that for all $t \geq t_\epsilon$, there exists $\alpha \in (0, \frac{1}{2}]$ such that the following inequality holds

$$\exp\left(\mu m \int_{t_\epsilon}^{t} \delta(l) dl\right) < t^\alpha. \tag{52}$$

Otherwise, there exists $t_1 \geq t_\epsilon$ such that for all $\alpha_1 \in (0, \frac{1}{2}]$ such that

$$\exp\left(\mu m \int_{t_\epsilon}^{t_1} \delta(l) dl\right) \geq t_1^{\alpha_1}.$$

Thus, we have

$$\mu m \int_{t_\epsilon}^{t_1} \delta(l) dl \geq \alpha_1 \ln(t_1). \tag{53}$$

We know that the integral of $\delta(t)$ from t_ϵ to t_1 can be rewritten as

$$\int_{t_\epsilon}^{t} \delta(l) dl = \int_{t_\epsilon}^{t} \delta(l) \cdot l \cdot \frac{1}{l} dl.$$

Since $\delta(t)t < \epsilon$ for $t \geq t_\epsilon$, then $\int_{t_\epsilon}^{t} \delta(l) \cdot l \cdot \frac{1}{l} dl < \epsilon \ln(\frac{t_1}{t_\epsilon})$. This is contradictory with (53) for small $\epsilon < \frac{\alpha_1}{\mu m}$. Thus for all $t \geq t_\epsilon$, the inequality (52) holds for a constant $\alpha \in (0, \frac{1}{2}]$. Then

$$\begin{aligned}
&\sum_{l=1}^{t} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u) du\right) \\
=\ & \sum_{l=1}^{t_\epsilon - 1} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u) du\right) dl + \sum_{t_\epsilon}^{t} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u) du\right) \\
\leq\ & \delta(1)^2 \exp\left(-\mu m \int_{u=t_\epsilon - 1}^{t+1} \delta(u) du\right)(t_\epsilon - 1) + \frac{\sum_{t_\epsilon}^{t} \left(\frac{\epsilon}{l}\right)^2 \exp\left(\mu m \int_{u=t_\epsilon}^{l} \delta(u) du\right)}{\exp\left(\mu m \int_{l=t_\epsilon}^{t+1} \delta(l) dl\right)} \\
\leq\ & \delta(1)^2 (t_\epsilon - 1) \exp\left(-\mu m \int_{u=t_\epsilon - 1}^{t+1} \delta(u) du\right) + \frac{\sum_{t_\epsilon}^{t} (\frac{\epsilon}{l})^2 (l+1)^\alpha}{\exp\left(\mu m \int_{l=t_\epsilon}^{t+1} \delta(l) dl\right)} \\
\leq\ & \left[\delta(1)^2 (t_\epsilon - 1) + 2\epsilon^2\right] \exp\left(-\mu m \int_{l=t_\epsilon}^{t+1} \delta(l) dl\right) \\
\leq\ & \frac{\delta(1)^2 (t_\epsilon - 1) + 2\epsilon^2}{\exp\left(-\mu m \int_{l=1}^{t_\epsilon} \delta(l) dl\right)} \exp\left(-\mu m \int_{l=1}^{t+1} \delta(l) dl\right),
\end{aligned}$$

where the third inequality follows from the fact that $\sum_{t_\epsilon}^{t} (\frac{\epsilon}{l})^2 (l+1)^\alpha \leq 2\epsilon^2$. Thus, in this case, for $t > n_0$, $\mathbb{E}[\|x_{t+1} - x^*\|^2]$ is at most

$$\left(\Delta_{n_0}^0 + 2\sigma^2 M^2 \exp(\mu m \delta(1)) \frac{\delta(1)^2 (t_\epsilon - 1) + 2\epsilon^2}{\exp(-\mu \int_{l=1}^{t_\epsilon} \delta(l) dl)}\right) \exp\left(-\mu m \int_{u=1}^{t+1} \delta(u) du\right).$$

2. $\lim_{t \to \infty} \delta(t)t = 1$.

In this case, it is easy to show there exist m and M such that $\frac{m}{t} \leq \eta(t) \leq \frac{M}{t}$. Hence the theorem follows from Theorem 1.

3. $\lim_{t\to\infty} \delta(t)t = +\infty$, that is for any $M_1 > 0$, there exists a constant $T_M \in \mathbb{N}^+$ such that for all $t \geq T_M$, $\delta(t)t > M_1$.

We suppose that there exists a constant $c_1 \leq \frac{\mu m}{2}$ such that for all $t \geq T_M$

$$-\frac{d\delta(t)}{dt} \leq c_1 \delta(t)^2. \tag{54}$$

Let $P(l) := \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right)$ for $1 \leq l \leq t$, then

$$\begin{aligned} \frac{dP(l)}{dl} &= 2\delta(l)\frac{d\delta(l)}{dl}\exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) + \mu m \delta(l)^3 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) \\ &= \delta(l)\exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right)\left[2\frac{d\delta(l)}{dl} + \mu m \delta(l)\delta(l)\right]. \end{aligned} \tag{55}$$

Let $Q(l) := 2\frac{d\delta(l)}{dl} + \mu m\delta(l)\delta(l)$. By (55), we know that the sign of $\frac{dP(l)}{dl}$ is determined by the sign of $Q(l)$. If $c_1 \leq \frac{\mu m}{2}$, from (54), we have $Q(l) \geq 0$, then the sequence of $P(l)$ is increasing when $l \geq T_M$.

If $P(u)$ is increasing for $u \in [l, t]$, then

$$\sum_{u=l}^{t} P(u) \leq \int_{u=l}^{t+1} P(u)du. \tag{56}$$

Otherwise, if $P(u)$ is decreasing for $u \in [l, t]$, then

$$\sum_{u=l}^{t} P(u) \leq P(l) + \int_{u=l}^{t} P(u)du. \tag{57}$$

By (56), we have

$$\sum_{l=1}^{t} P(l) = \sum_{l=1}^{T_M} P(l) + \sum_{l=T_M+1}^{t} P(l) \leq \sum_{l=1}^{T_M} P(l) + \int_{l=T_M}^{t+1} P(l)dl. \tag{58}$$

By integration by parts, $\int_{l=T_M}^{t+1} P(l)dl$ can be written as

$$\begin{aligned} &\mu m \int_{l=T_M}^{t+1} P(l)dl = \mu m \int_{l=T_M}^{t+1} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) dl \\ =&\delta(t+1) - \delta(T_M)\exp\left(-\mu m \int_{u=T_M}^{t+1} \delta(u)du\right) - \int_{l=T_M}^{t+1} \frac{d\delta(l)}{dl} \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) dl \\ \leq&\delta(t+1) - \delta(T_M)\exp\left(-\mu m \int_{u=T_M}^{t+1} \delta(u)du\right) + c_1 \int_{l=T_M}^{t+1} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) dl, \end{aligned}$$

where the above inequality holds because (54) satisfies. When $c_1 < \mu m$, rearranging the above inequality, we have

$$\int_{l=T_M}^{t+1} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) dl \leq \frac{\delta(t+1) - \delta(T_M)\exp\left(-\mu m \int_{u=T_M}^{t+1} \delta(u)du\right)}{(\mu m - c_1)}.$$

Hence,

$$
\begin{aligned}
&\sum_{l=1}^{t+1} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) dl \leq \sum_{l=1}^{T_M} P(l) + \int_{l=T_M}^{t+1} P(l)dl \\
=&\sum_{l=1}^{T_M} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) + \int_{l=T_M}^{t+1} \delta(l)^2 \exp\left(-\mu m \int_{u=l}^{t+1} \delta(u)du\right) dl \\
\leq&\frac{\delta(1)^2 T_M}{\exp\left(\mu m \int_{u=T_M}^{t+1} \delta(u)du\right)} + \frac{\delta(t+1) - \delta(T_M)\exp(-\mu m \int_{u=T_M}^{t+1} \delta(u)du)}{(\mu m - c_1)} \\
=&\frac{\delta(t+1)}{(\mu m - c_1)} + \frac{\delta(1)^2 T_M - \frac{\delta(T_M)}{(\mu m - c_1)}}{\exp\left(\mu m \int_{u=T_M}^{t+1} \delta(u)du\right)} \\
\leq&\frac{\delta(t+1)}{(\mu m - c_1)} + \frac{\delta(1)^2 T_M}{\exp\left(-\mu m \int_{u=1}^{T_M} \delta(u)du\right)} \exp\left(-\mu m \int_{u=1}^{t+1} \delta(u)du\right).
\end{aligned}
$$

Finally, incorporating the above inequality into (51), we can show that $\mathbb{E}[\|x_{t+1} - x^*\|^2]$ is bounded by

$$
\frac{\varepsilon_2}{(\mu m - c_1)}\delta(t+1) + \left[\Delta_{n_0}^0 + \frac{\varepsilon_2 \delta(1)^2 T_M}{\exp(-\mu m \int_{u=1}^{T_M} \delta(u)du)}\right] \exp\left(-\mu m \int_{u=1}^{t+1} \delta(u)du\right),
$$

where $\varepsilon_2 = 2\sigma^2 M^2 \exp(\mu m \delta(1))$.

■

Proof (**of Lemma 3**) Suppose that there exists a constant $c_1 > 0$ such that

$$-\frac{d\delta(t)}{dt} \leq c_1 \delta(t)^2.$$

Let $\hat{\delta}(t) = a\delta(t)$ for $a > 0$. Of course, for the new function $\hat{\delta}(t)$, there must be a constant $\hat{c}_1 > 0$ such that

$$-\frac{d\hat{\delta}(t)}{dt} \leq \hat{c}_1 \hat{\delta}(t)^2.$$

Then we have

$$-\frac{d\hat{\delta}(t)}{dt} = -a\frac{d\delta(t)}{dt} \leq \hat{c}_1 \hat{\delta}(t)^2 = a^2 \hat{c}_1 \delta(t)^2.$$

Thus,

$$-\frac{d\delta(t)}{dt} \leq a\hat{c}_1 \delta(t)^2.$$

Let $0 < a \leq \frac{\mu m}{2\hat{c}_1}$, we have $a\hat{c}_1 \leq \frac{\mu m}{2}$, which shows that there must be a constant $c_1 = a\hat{c}_1 \leq \frac{\mu m}{2}$.

■

Appendix D. Proofs of the Results in Section 5

Proof (**of Theorem 6**) We assume that $\eta(t)$ satisfies the following condition

$$\frac{m}{t+1} \leq \eta(t) \leq \frac{M \ln(t+1)}{t+1}, \forall\, 1 \leq t \leq T.$$

Let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. For $t \geq (2L_f M)^2$, we have

$$\eta(t) \leq \frac{M \ln(t+1)}{t+1} \leq \frac{M\sqrt{t+1}}{t+1} \leq \frac{1}{2L_f}. \tag{59}$$

Then n_0 must exist and is a constant that is independent of T. Thus the inequality (8) of Lemma 2 holds, then we get

$$\begin{aligned}
&\mathbb{E}[\|x_{t+1} - x^*\|^2] \\
\leq &\exp\left(-\mu \sum_{l=1}^{t} \eta(l)\right) \Delta_{n_0}^0 + 2\sigma^2 \sum_{l=1}^{t} \eta(l)^2 \exp\left(-\mu \sum_{u>l}^{t} \eta(u)\right) \\
\leq &\exp\left(-\mu m \sum_{l=1}^{t} \frac{1}{l+1}\right) \Delta_{n_0}^0 + 2\sigma^2 M^2 \sum_{l=1}^{t} \frac{\ln^2(l+1)}{(l+1)^2} \exp\left(-\mu m \sum_{u>l}^{t} \frac{1}{u+1}\right) \\
\leq &\frac{\Delta_{n_0}^0}{\exp(\mu m(\ln(t+2) - \ln 2))} + 2\sigma^2 M^2 \exp(\mu m) \sum_{l=1}^{t} \frac{\ln^2(l+1)}{(l+1)^2} \cdot \frac{\exp(\mu m \ln(l+1))}{\exp(\mu m \ln(t+2))} \\
\leq &\frac{2^{(\mu m)} \Delta_{n_0}^0}{(t+2)^{(\mu m)}} + \frac{2\sigma^2 M^2 \exp(\mu m)}{(t+2)^{(\mu m)}} \sum_{l=1}^{t} \frac{\ln^2(l+1)}{(l+1)^2} (l+1)^{(\mu m)} \\
\leq &\frac{2^{(\mu m)} \Delta_{n_0}^0}{(t+2)^{(\mu m)}} + \frac{2\sigma^2 M^2 \exp(\mu m)}{(t+2)^{(\mu m)}} \left[\frac{\ln(2)}{2} + \int_{l=2}^{t+2} \frac{\ln^2(l)}{l^2} \cdot (l)^{(\mu m)} dl\right],
\end{aligned} \tag{60}$$

where the third inequality follows from (32a) and (32b), and the last inequality is obtained from (58). If $\mu m = 1$, we have

$$\int_{l=2}^{t+2} \frac{\ln^2(l)}{l^2} \cdot (l)^{(\mu m)} dl = \frac{\ln^3(t+2)}{3} - \ln^3 2 < \frac{\ln^3(t+2)}{3}.$$

Otherwise, if $\mu m \neq 1$, integrating by parts we get

$$\begin{aligned}
&\int_{l=2}^{t+2} \frac{\ln^2(l)}{l^2} \cdot l^{(\mu m)} dl \\
\leq &\frac{(t+2)^{(\mu m-1)} \ln^2(t+2) - 2^{(\mu m-1)} \ln^2 2}{(\mu m - 1)} + \frac{2^{(\mu m)} \ln 2}{(\mu m - 1)^2} + \frac{2[(t+2)^{(\mu m-1)} - (2)^{(\mu m-1)}]}{(\mu m - 1)^3}.
\end{aligned}$$

From the above inequality, we can see that if $\mu m < 1$, such an integral can be bounded by a scalar

$$\int_{l=2}^{t+2} \frac{\ln^2(l)}{l^2} \cdot l^{(\mu m)} dl \leq \frac{2^{(\mu m)}}{(1-\mu m)^3} + \frac{2^{(\mu m)} \ln 2}{(1-\mu m)^2} + \frac{2^{(\mu m-1)} \ln^2 2}{(1-\mu m)} \leq \frac{2 + 2\ln 2 + \ln^2 2}{(1-\mu m)^3}.$$

While $\mu m > 1$, then

$$\int_{l=2}^{t+2} \frac{\ln^2(l)}{l^2} \cdot l^{(\mu m)} dl \leq \left[\frac{\ln^2(t+2)}{(\mu m - 1)} + \frac{2}{(\mu m - 1)^3} \right] (t+2)^{(\mu m - 1)} + \frac{2^{(\mu m)} \ln 2}{(\mu m - 1)^2}.$$

Thus, collecting the results obtained above, let $t = T$, we can get the result as desired. ■

Proof (of Theorem 7) In this case, $\eta(t)$ satisfies that

$$\frac{m}{t} \leq \eta(t) \leq \frac{M}{t^{\alpha}},$$

for $\alpha \in (1/2, 1]$. Let $n_0 := \sup\{t \in \mathbb{N}^+ : \eta(t) > 1/(2L_f)\}$. For $t \geq (2L_f M)^{(1/\alpha)}$, we have

$$\eta(t) \leq \frac{M}{t^{\alpha}} \leq \frac{1}{2L_f}. \tag{61}$$

Then n_0 must exist and is a constant that is independent of T. Thus, in this case, the inequality (8) of Lemma 2 holds. From (8) in Lemma 2, we have

$$\begin{aligned}
\mathbb{E}[\|x_{t+1} - x^*\|^2] &\leq \exp\left(-\mu \sum_{l=1}^{t} \eta(l)\right) \Delta_{n_0}^0 + 2\sigma^2 \sum_{l=1}^{t} \eta(l)^2 \exp\left(-\mu \sum_{u>l}^{t} \eta(u)\right) \\
&\leq \exp\left(-\mu m \sum_{l=1}^{t} \frac{1}{l}\right) \Delta_{n_0}^0 + 2\sigma^2 M^2 \sum_{l=1}^{t} \frac{1}{l^{2\alpha}} \exp\left(-\mu m \sum_{u>l}^{t} \frac{1}{u}\right) \\
&\leq \frac{\Delta_{n_0}^0}{(t+1)^{(\mu m)}} + \frac{2\sigma^2 M^2 \exp(\mu m)}{(t+1)^{(\mu m)}} \sum_{l=1}^{t} l^{(\mu m - 2\alpha)} \\
&\leq \frac{\Delta_{n_0}^0}{(t+1)^{(\mu m)}} + \frac{2\sigma^2 M^2 \exp(\mu m)}{(t+1)^{(\mu m)}} \left(\int_{l=1}^{t+1} l^{(\mu m - 2\alpha)} dl + 1 \right) \\
&\leq \frac{\Delta_{n_0}^0 + 2\sigma^2 M^2 \exp(\mu m)}{(t+1)^{(\mu m)}} + \frac{2\sigma^2 M^2 \exp(\mu m)}{(t+1)^{(\mu m)}} \int_{l=1}^{t+1} l^{(\mu m - 2\alpha)} dl.
\end{aligned}$$

If $\mu m = 2\alpha - 1 > 0$, it follows that

$$\int_{l=1}^{t+1} l^{(\mu m - 2\alpha)} dl = \int_{l=1}^{t+1} \frac{dl}{l} = \ln(t+1).$$

Consequently,

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \frac{\Delta_{n_0}^0 + 2\sigma^2 M^2 \exp(2\alpha - 1)}{(t+1)^{(2\alpha - 1)}} + \frac{2\sigma^2 M^2 \exp(2\alpha - 1) \ln(t+1)}{(t+1)^{(2\alpha - 1)}}.$$

If $\mu m \neq 2\alpha - 1$, we have

$$\int_{l=1}^{t+1} l^{(\mu m - 2\alpha)} dl = \frac{(t+1)^{(\mu m - 2\alpha + 1)} - 1}{(\mu m - 2\alpha + 1)},$$

then $\mathbb{E}[\|x_{t+1}-x^*\|^2]$ is at most

$$\frac{\Delta^0_{n_0}+2\sigma^2M^2\exp(2\alpha-1)}{(t+1)^{(\mu m)}}+\frac{2\sigma^2M^2\exp(\mu m)}{(\mu m-2\alpha+1)}\left[\frac{1}{(t+1)^{(2\alpha-1)}}-\frac{1}{(t+1)^{(\mu m)}}\right].$$

Combing the above results and letting $t=T$, we obtain the desired result. ■

Proof (of Theorem 8) In this case, we assume that $\eta(t)$ satisfies that

$$\frac{m}{(t+1)\ln(t+1)}\le\eta(t)\le\frac{M}{(t+1)^\alpha}$$

for $\alpha\in(1/2,1]$. Let $n_0:=\sup\{t\in\mathbb{N}^+:\eta(t)>1/(2L_f)\}$. For $t\ge(2L_fM)^{(1/\alpha)}-1$, we have

$$\eta(t)\le\frac{M}{(t+1)^\alpha}\le\frac{1}{2L_f}. \tag{62}$$

Therefore n_0 must exist and is a constant. In this case, the inequality (8) of Lemma 2 holds. By (8), we have

$$\begin{aligned}
&\mathbb{E}[\|x_{t+1}-x^*\|^2]\\
\le&\exp\left(-\mu\sum_{l=1}^t\eta(l)\right)\Delta^0_{n_0}+2\sigma^2\sum_{l=1}^t\eta(l)^2\exp\left(-\mu\sum_{u>l}^t\eta(u)\right)\\
\le&\exp\left(-\mu m\sum_{l=1}^t\frac{1}{(l+1)\ln(l+1)}\right)\Delta^0_{n_0}+2\sigma^2M^2\sum_{l=1}^t\frac{\exp\left(-\mu m\sum_{u>l}^t\frac{1}{(l+1)\ln(l+1)}\right)}{(l+1)^{2\alpha}}\\
\le&\frac{(\ln 2)^{(\mu m)}\Delta^0_{n_0}}{(\ln(t+2))^{(\mu m)}}+\frac{2\sigma^2M^2(\ln 2)^{(\mu m)}}{(\ln(t+2))^{(\mu m)}}\sum_{l=1}^t\frac{(\ln(l+1))^{(\mu m)}}{(\ln(t+2))^{2\alpha}}\\
\le&\frac{(\ln 2)^{(\mu m)}\Delta^0_{n_0}}{(\ln(t+2))^{(\mu m)}}+\frac{2\sigma^2M^2(\ln 2)^{(\mu m)}}{(\ln(t+2))^{(\mu m)}}\left[\frac{(\ln 2)^{\mu m}}{2^{2\alpha}}+\int_{l=1}^{t+1}\frac{(\ln(l+1))^{(\mu m)}}{(l+1)^{2\alpha}}dl\right],
\end{aligned} \tag{63}$$

where the third inequality dues to the fact that $\sum_{l=1}^t\frac{1}{l\ln(l)}\ge\int_{l=1}^{t+1}\frac{1}{(l+1)\ln(l+1)}dl=\ln\ln(t+2)-\ln\ln 2$ and the last inequality follows from (58).

We know that for any $\beta\in(0,1)$, there must be a constant t_β such that $\ln(t+1)\le(t+1)^\beta$ for all $t\ge t_\beta$. Here we choose that $0<\beta<\frac{2\alpha-1}{\mu m}$. There exists a constant t_β such that $\ln(t+1)\le(t+1)^\beta$ for all $t\ge t_\beta$. For sufficiently large $t\ge t_\beta$, we have

$$\begin{aligned}
\int_{l=1}^{t+1}\frac{(\ln(l+1))^{(\mu m)}}{(l+1)^{2\alpha}}dl&\le\int_{l=1}^{t_\beta}\frac{(\ln(l+1))^{(\mu m)}}{(l+1)^{2\alpha}}dl+\int_{t_\beta}^{t+1}\frac{(\ln(l+1))^{(\mu m)}}{(l+1)^{2\alpha}}dl\\
&\le(\ln(t_\beta+1))^{(\mu m)}\int_{l=1}^{t_\beta}\frac{dl}{(l+1)^{2\alpha}}+\int_{t_\beta}^{t+1}(l+1)^{(\beta\mu m-2\alpha)}dl\\
&\le\frac{2^{(1-2\alpha)}}{2\alpha-1}+\frac{(t+1)^{(\beta\mu m-2\alpha+1)}-(t_\beta+1)^{(\beta\mu m-2\alpha+1)}}{(\beta\mu m+1-2\alpha)}.
\end{aligned} \tag{64}$$

Thus, applying (64) into (63) and let $t = T$, we can bound $\mathbb{E}[\|x_{T+1} - x^*\|^2]$ by

$$\frac{(\ln 2)^{(\mu m)}\Delta_{n_0}^0}{(\ln(t+2))^{(\mu m)}} + \frac{2\sigma^2 M^2 (\ln 2)^{(\mu m)}}{(\ln(t+2))^{(\mu m)}}\left[\frac{(\ln 2)^{(\mu m)}}{2^{2\alpha}} + \frac{2^{(1-2\alpha)}}{2\alpha-1} + \frac{(t_\beta+1)^{(\beta\mu m - 2\alpha+1)}}{(2\alpha - 1 - \beta\mu m)}\right].$$

Therefore, there exists a constant $C_2 > 0$ such that

$$\mathbb{E}[\|x_{T+1} - x^*\|^2] \le \frac{C_2}{(\ln(t+2))^{(\mu m)}}.$$

■

References

Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems*, pages 1157–1167, 2018.

Wangpeng An, Haoqian Wang, Yulun Zhang, and Qionghai Dai. Exponential decay sine wave learning rate for fast deep neural network training. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

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

Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

Aaron Defazio and Robert M. Gower. The power of factorial powers: New parameter settings for (stochastic) optimization. In *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 49–64, 2021.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

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

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Advances in Neural Information Processing Systems*, pages 14977–14988, 2019.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Baptiste Goujaud, Damien Scieur, Aymeric Dieuleveut, Adrien B. Taylor, and Fabian Pedregosa. Super-acceleration with cyclical step-sizes. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3028–3065, 2022.

Robert M Gower, Peter Richtárik, and Francis Bach. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Mathematical Programming*, 2020. doi: https://doi.org/10.1007/s10107-020-01506-0.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209, 2019.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613, 2019.

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

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

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *International Conference on Learning Representations*, 2017.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *Conference on Learning Theory*, pages 1752–1755, 2019.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of International Conference on Machine Learning*, pages 1724–1732, 2017.

Nitish Shirish Keskar and George Saon. A nonmonotone learning rate strategy for SGD training of deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4974–4978. IEEE, 2015.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *International Conference on Learning Representations*, 2015.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $\mathcal{O}(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Remi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Improved asynchronous parallel optimization analysis for stochastic incremental methods. *Journal of Machine Learning Research*, 19(81):1–68, 2018.

Todd K Leen and Genevieve B Orr. Optimal stochastic search and adaptive momentum. In *Advances in Neural Information Processing Systems*, pages 477–484, 1994.

Todd K Leen, Bernhard Schottky, and David Saad. Two approaches to optimal annealing. In *Advances in Neural Information Processing Systems*, pages 301–307, 1998.

Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pages 6553–6564. PMLR, 2021.

Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575, 1977.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2017.

Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

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

Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. SGD and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR, 2018.

Lam M Nguyen, Phuong Ha Nguyen, Peter Richtárik, Katya Scheinberg, and Marten van Dijk. New convergence aspects of stochastic gradient algorithms. *Journal of Machine Learning Research*, 20(176):1–49, 2019a.

Phuong_Ha Nguyen, Lam Nguyen, and Marten van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in SGD. In *Advances in Neural Information Processing Systems*, pages 3665–3674, 2019b.

Samet Oymak. Provable super-convergence with a large cyclical learning rate. *IEEE Signal Processing Letters*, 28:1645–1649, 2021.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(2):1–17, 1964.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, pages 1139–1147, 2013.

Conghui Tan, Shiqian Ma, Yu-Hong Dai, and Yuqiu Qian. Barzilai-Borwein step size for stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 685–693, 2016.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp, COURSERA: Neural networks for machine learning. *Technical Report, University of Toronto*, 2012.

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

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, pages 3732–3745, 2019b.

Zhuang Yang, Cheng Wang, Zhemin Zhang, and Jonathan Li. Random Barzilai-Borwein step size for mini-batch algorithms. *Engineering Applications of Artificial Intelligence*, 72:124–135, 2018.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.