On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes

Xiaoyu Li
Department of Applied Mathematics & Statistics
Stony Brook University
Stony Brook, NY 11790
xiaoyu.li@stonybrook.edu

Francesco Orabona
Department of Computer Science
Stony Brook University
Stony Brook, NY 11790
francesco@orabona.com

Abstract

Stochastic gradient descent is the method of choice for large scale optimization of machine learning objective functions. Yet, its performance is greatly variable and heavily depends on the choice of the stepsizes. This has motivated a large body of research on adaptive stepsizes. However, there is currently a gap in our theoretical understanding of these methods, especially in the non-convex setting. In this paper, we start closing this gap: we theoretically analyze the use of adaptive stepsizes, like the ones in AdaGrad, in the non-convex setting. We show sufficient conditions for almost sure convergence to a stationary point when the adaptive stepsizes are used, proving the first guarantee for AdaGrad in the non-convex setting. Moreover, we show explicit rates of convergence that automatically interpolates between $O(1/T)$ and $O(1/\sqrt{T})$ depending on the noise of the stochastic gradients, in both the convex and non-convex setting.

1 Introduction

In the recent years, Stochastic Gradient Descent (SGD) has become the tool of choice to train machine learning models. In particular, in the Deep Learning community it is widely used to minimize the training error of deep networks. In this setting, the stochasticity arises from the use of so-called *mini-batches*, that allows to keep the complexity per iteration constant with respect to the size of the training set.

Classic convergence analysis of the SGD algorithm relies on conditions on the positive stepsizes η_t [Robbins and Monro, 1951]. In particular, sufficient conditions are that

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty \ . \tag{1}$$

The first condition is necessary and very intuitive too: we need the algorithm to be able to travel arbitrary distances, in order to reach the stationary point from the initial point. On the other hand, the second one is not necessary and many popular choices of the stepsize, e.g. the one in AdaGrad [Duchi et al., 2011], do not satisfy it while still guaranteeing convergence in the convex setting.

However, for a large number of SGD variations employed by practitioners the conditions above are not satisfied and not much is known about their convergence in the non-convex setting. In fact, these algorithms are often designed and analyzed for the convex domain, [e.g. Duchi et al., 2011], or they do not provide convergence guarantees at all, [e.g. Zeiler, 2012], or even worse they are known to fail to converge on simple one-dimensional convex stochastic optimization problems [Reddi et al., 2018]. This lack of understanding is particularly unsettling, especially when we consider the fact

Preprint. Work in progress.

that we do not even know under which conditions these popular variations of SGD converge to a stationary point with an *infinite* number of iterations.

In this paper we start closing this gap, finding necessary conditions under which popular variations of SGD converge to a stationary point. In particular, we focus on the *adaptive stepsizes* popularized by AdaGrad [Duchi et al., 2011]. This kind of updates has become the basis of all other adaptive optimization algorithms used in machine learning, e.g. [Zeiler, 2012, Tieleman and Hinton, 2012, Kingma and Ba, 2015, Reddi et al., 2018]. We analyze the coordinate-wise AdaGrad stepsize and a global version too. Also, we show novel theoretical properties of the adaptive global stepsizes in both the convex and non-convex setting.

More in details, the contributions of this paper are the following:

- In Section 5, in the convex setting we show that global adaptive stepsizes give rise to convergence rates that are adaptive to the noise level, interpolating between the convergence rates of Gradient Descent (GD) and SGD. In doing so, we also remove the strong assumption of having a bounded domain present in many previous analyses.
- In Section 6, in the non-convex setting we prove almost sure convergence of SGD with adaptive stepsizes, for both coordinate-wise and global adaptive stepsizes. As far as we know, this is the *first* theoretical justification for the use of AdaGrad in the non-convex setting.
- In Section 7, in the non-convex setting we show a finite-time convergence rate to stationary points, adaptive to the level of noise for the global stepsizes.

Next Section discusses more in details the related work, while Section 3 introduces formally the setting, and Section 4 the adaptive stepsizes considered in this work.

2 Related Work

In the convex setting, adaptive stepsizes have a long history. They were first proposed in the online learning literature [Auer et al., 2002] and adopted into the stochastic one later [Duchi et al., 2011]. Yet, most of these works assumed the optimization to be constrained in a convex bounded set. While this is a reasonable assumption in some settings, it is completely unreasonable in many applications of optimization for machine learning. Yousefian et al. [2012] analyze different adaptive stepsizes, but only for strongly convex optimization. Recently, Wu et al. [2018] have analyzed a choice of adaptive stepsizes similar to the global stepsizes we consider, but their result in the convex setting requires the very strong assumption of having the norm of the gradients strictly greater than zero.

The convergence of a random iterate of SGD for non-convex smooth functions has been proved by Ghadimi and Lan [2013], and it was already implied by the results in Bottou [1991]. With additional regularity assumptions, these results imply almost sure convergence of the gradient to zero [Bottou, 1991, Bottou et al., 2016]. In alternative to the regularity assumptions, Bottou [1998] proposed to assume that beyond a certain horizon the update always moves the iterate closer to the origin on average, that implies the confinement in a bounded domain and, in turn, the almost sure convergence. On the other hand, the weakest assumptions for the almost sure convergence of SGD for non-convex smooth functions have been established in Bertsekas and Tsitsiklis [2000]: the variance of the noise on the gradient in $\boldsymbol{x}_t$ can grow as $1 + \|\nabla f(\boldsymbol{x}_t)\|^2$, f is lower bounded, and the stepsizes satisfy (1). However, these approaches do not cover adaptive stepsizes.

The only work we know on adaptive stepsizes for non-convex stochastic optimization is Kresoja et al. [2017]. They study the convergence of a choice of adaptive stepsizes that require access to the function values, under strict conditions on the direction of the gradients. Wu et al. [2018] also consider adaptive stepsizes, but they only consider deterministic gradients in the non-convex setting.

A different route is to assume some properties of the non-convex function that allow to prove convergence rates. A number of such conditions has been proposed, such as the Polyak-Łojasiewicz condition [Polyak, 1963] (see Karimi et al. [2016] for a recent review on these conditions). However, all these conditions are a substitute of strong convexity and they are used to prove linear convergence rate in the non-convex deterministic setting through a contractive mapping. In this view, these conditions are actually very strong and it is still unclear how useful they are to model the problems we encounter in optimization problems in machine learning.

A very weak condition for almost sure convergence to the global optimum of non-convex functions was proposed in Bottou [1998] and recently independently reproposed in Zhou et al. [2017]. However, this condition implies the very strong assumption that the gradients never point in the opposite direction of the global optimum.

In this paper, in our most restrictive case, we will only assume the function to be smooth and Lipschitz.

3 Problem Set-Up

Notation. We denote vectors and matrices by bold letters, e.g. $\boldsymbol{x} \in \mathbb{R}^d$. The coordinate j of a vector $\boldsymbol{x}$ is denoted by x_j and as $(\nabla f(\boldsymbol{x}))_j$ for the gradient $\nabla f(\boldsymbol{x})$. We denote by $\mathbb{E}[\cdot]$ the expectation with respect to the underlying probability space and by $\mathbb{E}_t[\cdot]$ the conditional expectation with respect to the past, that is, with respect to $\xi_1, \cdots, \xi_{t-1}$. All the norms are L2 norms.

Setting and Assumptions. We consider the following optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}),$$

where $f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ is a function bounded from below. We will make different assumptions on the objective function f, depending on the setting. In particular, we will always assume that

- (**H1**) f is *M-smooth*, that is, f is differentiable and its gradient is M-Lipschitz, i.e. $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq M\|\boldsymbol{x} - \boldsymbol{y}\|, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

Note that (**H1**), for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, implies [Nesterov, 2003, Lemma 1.2.3]

$$|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle| \leq \frac{M}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 . \qquad (2)$$

Sometimes, we will also assume that

- (**H2**) f is *L-Lipschitz*, i.e. $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

We assume that we have access to a stochastic first-order black-box oracle, that returns a noisy estimate of the gradient of f at any point $\boldsymbol{x} \in \mathbb{R}^d$. That is, we will use the following assumption

- (**H3**) We receive a vector $\boldsymbol{g}(\boldsymbol{x}, \xi)$ such that $\mathbb{E}[\boldsymbol{g}(\boldsymbol{x}, \xi)] = \nabla f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$.

We will also make alternatively one of the following assumptions on the variance of the noise.

- (**H4**) The noise in the stochastic gradient has bounded support, that is $\|\boldsymbol{g}(\boldsymbol{x}, \xi) - \nabla f(\boldsymbol{x})\| \leq S, \ \forall \boldsymbol{x}$.
- (**H4'**) The stochastic gradient satisfies $\mathbb{E}\left[\exp\left(\frac{\|\nabla f(\boldsymbol{x}) - g(\boldsymbol{x}, \xi)\|^2}{\sigma^2}\right)\right] \leq \exp(1), \ \forall \boldsymbol{x}$.

Assumption (**H4'**) has been already used by Nemirovski et al. [2009] to prove high probability convergence guarantees. Note that this condition implies a bounded variance, in fact

$$\exp\left(\mathbb{E}\left[\frac{\|\nabla f(\boldsymbol{x}) - g(\boldsymbol{x}, \xi)\|^2}{\sigma^2}\right]\right) \leq \mathbb{E}\left[\exp\left(\frac{\|\nabla f(\boldsymbol{x}) - g(\boldsymbol{x}, \xi)\|^2}{\sigma^2}\right)\right] \leq \exp(1) .$$

This condition will be needed to control the expectation of the maximum of the terms $\|\nabla f(\boldsymbol{x}_t) - g(\boldsymbol{x}_t, \xi_t)\|^2$. Note that (**H4**) implies (**H4'**).

Stochastic Gradient Descent. The optimization algorithm we consider is SGD, that iteratively updates the solution as $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)$, starting from an arbitrary point $\boldsymbol{x}_1$. Differently from previous work, we allow the stepsizes η_t to depend on the past, effectively making them stochastic variables. Also, we will consider the more general setting in which the stepsizes are diagonal matrices whose elements on the diagonal are $\eta_{t,j}$, and the update becomes $\boldsymbol{x}_{t+1,j} = \boldsymbol{x}_{t,j} - \eta_{t,j} g(\boldsymbol{x}_t, \xi_t)_j$ for $j = 1, \cdots, d$.

4 Adaptive Stepsizes

The adaptive stepsizes we analyze are a generalization of ones widely used in the online and stochastic optimization literature. As such, their good performance have been already validated in numerous empirical results. In particular, we consider the following stepsizes

$$\eta_t = \frac{\alpha}{\left(\beta + \sum_{i=1}^{t-1} \|\boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2\right)^{1/2+\epsilon}} \quad (3) \qquad \text{and} \qquad \eta_{t,j} = \frac{\alpha}{\left(\beta + \sum_{i=1}^{t-1} (g(\boldsymbol{x}_i, \xi_i)_j)^2\right)^{1/2+\epsilon}} \quad (4)$$

where $\alpha, \beta > 0$ and $\epsilon \geq 0$. Depending on the particular setting, we might have more constraints on α, β, ϵ. Note that, with $\epsilon = 0$, (4) are the coordinate-wise stepsizes used in AdaGrad [Duchi et al., 2011], while (3) have been used in online convex optimization to achieve adaptive regret guarantees, [e.g. Rakhlin and Sridharan, 2013, Orabona and Pál, 2018].

A key difference of (3) and (4) with the standard adaptive stepsizes is the fact that $\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)$ is not used in η_t. This is a key property for the theoretical analysis, because it allows to calculate the conditional expectation of quantities involving η_t and $\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)$. Also, we claim this should be the right way to implement adaptive stepsizes. Indeed, as we show in the Example below, if the stepsize does depend on the current gradient, things can go wrong. The details can be found in the Appendix.

Example 1. *There exist a convex differentiable function satisfying (**H1**), an additive noise on the gradients satisfying (**H4**), and a sequence of gradients such that for a given t we have $\mathbb{E}_{\xi_t}[\langle \eta_{t+1} \boldsymbol{g}(\boldsymbol{x}_t, \xi_t), \nabla f(\boldsymbol{x}_t)\rangle] < 0$.*

In words, the example says that including the current noisy gradient in η_t (that is, using η_{t+1}) can make the algorithm deviate in expectation more than 90 degrees from the correct direction. While in the convex bounded case the algorithm can recover, it is intuitive that this could have catastrophic consequences in the unconstrained non-convex setting.

On the other hand, this difference makes the analysis more involved, because the quantity $\sum_{t=1}^T \eta_t^2 \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2$ cannot be bounded anymore in a straightforward way, see Lemma 2 in the next Section. Previous analyses, [e.g. Duchi et al., 2011], solved this issue by assuming the knowledge of the Lipschitz constant of the function f, while we will assume the function to be Lipschitz only to prove the asymptotic guarantee and no knowledge of it.

In the following, we will show that this stepsize allows to prove adaptive guarantees in the convex and non-convex setting.

5 Adaptive Convergence Rates for Convex Functions

In this section, we show that the global stepsizes (3) give adaptive rates of convergence that interpolate between the rate of GD and SGD, without knowledge of the variance of the noise. Differently from the other proofs on SGD with adaptive rates [e.g. Duchi et al., 2011], we do not assume to use projections. This makes the proof more technically challenging, but at the same time it mirrors the setting of many applications of SGD in machine learning optimization problems.

We first state some technical lemmas, whose proofs are in the Appendix.

Lemma 1. *Assume (**H1**). Then $\|\nabla f(\boldsymbol{x})\|^2 \leq 2M(f(\boldsymbol{x}) - \min_{\boldsymbol{y}} f(\boldsymbol{y})),\ \forall \boldsymbol{x}$.*

Lemma 2. *Assume (**H1, H3, H4'**). The stepsizes are chosen as* (3)*, where $\alpha, \beta, \epsilon > 0$. Then,*

$$\mathbb{E}\left[\sum_{t=1}^T \eta_t^2 \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right] \leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \frac{4\alpha^2}{\beta^{1+2\epsilon}}(1 + \log(T))\sigma^2 + \frac{4\alpha}{\beta^{1/2+\epsilon}} \mathbb{E}\left[\sum_{t=1}^T \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2\right] .$$

We can now state the adaptive convergence guarantee.

Theorem 1. *Assume (**H1**, **H3**, **H4'**) and f convex. Let $\delta \in (0,1)$ and the stepsizes set as in* (3)*, where $\alpha, \beta, \epsilon > 0$, and $4\alpha M < \beta^{1/2+\epsilon}$. Then, with probability at least $1 - \delta$, the iterates of SGD satisfies the following bound*

$$f(\bar{\boldsymbol{x}}_T) - f(\boldsymbol{x}^\star) \leq \frac{1}{T} \max\left(\left(\frac{C_T\, (8M)^{1/2+\epsilon}}{\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\alpha\delta} \right)^{\frac{1}{1/2-\epsilon}}, \frac{C_T}{\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\alpha\delta} \left(2\beta + \frac{8\sigma^2 T}{\delta}\right)^{1/2+\epsilon} \right),$$

where $\bar{\boldsymbol{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$*,* $\boldsymbol{x}^\star =$ *argmin*$_{\boldsymbol{x}}$ $f(\boldsymbol{x})$*, and* $C_T = \|\boldsymbol{x}^\star - \boldsymbol{x}_1\|^2 + \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \frac{4\alpha^2}{\beta^{1+2\epsilon}}(1 + \log(T))\sigma^2$.

Proof. From the update of SGD we have that

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 = -2\eta_t\langle \boldsymbol{g}(\boldsymbol{x}_t, \xi_t), \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle + \eta_t^2\|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \ .$$

Taking the conditional expectation with respect to $\xi_1, \cdots, \xi_{t-1}$, we have that

$$E_t[\langle \boldsymbol{g}(\boldsymbol{x}_t, \xi_t), \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle] = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle \geq f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star),$$

where in the inequality we used the fact that f is convex. Hence, summing over $t = 1$ to T, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \eta_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))\right] \leq \frac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}_1\|^2 + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \eta_t^2\|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right] \ .$$

From Lemma 1 and Lemma 2, we have that

$$\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\mathbb{E}\left[\sum_{t=1}^{T} \eta_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))\right] \leq \frac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}_1\|^2 + \frac{\alpha^2}{4\epsilon\beta^{2\epsilon}} + \frac{2\alpha^2}{\beta^{1+2\epsilon}}(1 + \log(T))\sigma^2 \ . \quad (5)$$

We can also lower bound the l.h.s. of (5) with

$$\mathbb{E}\left[\sum_{t=1}^{T} \eta_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))\right] \geq \mathbb{E}\left[\eta_T \sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))\right] \ .$$

Putting all together and using Markov's inequality, we have, with probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) \leq \frac{2}{\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\delta\eta_T}\left(\frac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}_1\|^2 + \frac{\alpha^2}{4\epsilon\beta^{2\epsilon}} + \frac{2\alpha^2}{\beta^{1+2\epsilon}}(1 + \log(T))\sigma^2\right) \ .$$

Using the expression of η_T, we have

$$\begin{aligned}\frac{1}{\eta_T} &= \frac{1}{\alpha}\left(\beta + \sum_{t=1}^{T-1}\|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right)^{1/2+\epsilon} \leq \frac{1}{\alpha}\left(\beta + 2\sum_{t=1}^{T-1}\left(\|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 + \|\nabla f(\boldsymbol{x}_t)\|^2\right)\right)^{1/2+\epsilon} \\ &\leq \frac{1}{\alpha}\left(\beta + 2\sum_{t=1}^{T-1}\left(\|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 + 2M(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))\right)\right)^{1/2+\epsilon},\end{aligned}$$

where in the first inequality we used the elementary inequality $\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2$ and Lemma 1 in the second one. We use again Markov's inequality, to have, with probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T-1}\|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \leq \frac{2\sigma^2(T-1)}{\delta} \ .$$

Hence, putting all together, using the notation of the theorem, and overapproximating, we have

$$\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) \leq \frac{C_T}{\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\alpha\delta}\left(\beta + \frac{4\sigma^2 T}{\delta} + 4M\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star))\right)^{1/2+\epsilon} \ .$$

Through a case analysis, we have that

$$\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) \leq \max\left(\left(\frac{C_T\,(8M)^{1/2+\epsilon}}{\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\alpha\delta}\right)^{\frac{1}{1/2-\epsilon}}, \frac{C_T}{\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\alpha\delta}\left(2\beta + \frac{8\sigma^2 T}{\delta}\right)^{1/2+\epsilon}\right) \ .$$

Using Jensen's inequality on the l.h.s. of last inequality together with the union bound concludes the proof. □

Up to polylog terms, if $\sigma = 0$ we recover the GD rate, $O(\frac{1}{T})$, and otherwise we get the rate of SGD, $O(\frac{1}{\sqrt{T}})$. The same behavior was proved in Dekel et al. [2012]. However, here we do not need to know the noise level nor assuming a bounded domain. In the case the constants of the slow term are small compared with the ones of the first term, we can expect a first quick convergent phase, followed by a slow one, as it is often observed in empirical experiments.

Observe that this bound is not in expectation but in probability. While all the bounds on the expected sub-optimality can be expressed in the same way, here the use of the Markov's inequality is actually necessary to be able to solve the implicit inequality in the proof.

6 Almost Sure Convergence for Non-Convex Functions

In this section, we show that SGD with the adaptive stepsizes in (3) and (4) converges to a stationary point almost surely, that is, with probability 1. Note that the stepsizes in (3) and (4) *do not satisfy* (1), not even in expectation, because the $\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)$ could decrease fast enough to have $\sum_{t=1}^{\infty} \eta_t^2 = \infty$. Hence, the results here cannot be obtain from the classic results in stochastic approximation [e.g. Bertsekas and Tsitsiklis, 2000].

Here, we will have to assume our strongest assumptions. In particular, we will need the function to be Lipschitz and the noise to have bounded support. This is mainly needed in order to be sure the sum of the stepsizes diverges.

We first state some technical lemmas we will use in the following, all the proofs are in the Appendix.

Lemma 3. *[Mairal, 2013, Lemma A.5] Let $(a_t)_{t\geq 1}, (b_t)_{t\geq 1}$ be two non-negative real sequences such that $(b_t)_{t\geq 1}$ is bounded, $\sum_{t=1}^{\infty} a_t b_t$ converges and $\sum_{t=1}^{\infty} a_t$ diverges, and there exists $K \geq 0$ such that $|b_{t+1} - b_t| \leq K a_t$. Then b_t converges to 0.*

Lemma 4. *Let $a_0 > 0$, $a_i \geq 0$, $i = 1, \cdots, T$ and $\beta > 1$. Then $\sum_{t=1}^{T} \frac{a_t}{(a_0 + \sum_{i=1}^{t} a_i)^{\beta}} \leq \frac{1}{(\beta-1) a_0^{\beta-1}}$.*

Lemma 5. *Assume (**H1, H3**). Then, the iterates of SGD satisfy the following inequality*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle\right] \leq f(\boldsymbol{x}_1) - f^* + \frac{M}{2}\mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right] .$$

We now state the almost sure convergence of SGD with adaptive stepsizes.

Theorem 2. *Assume (**H1, H2, H3, H4**). The stepsizes are chosen as in* (3)*, where $\alpha, \beta > 0$ and $\epsilon \in (0, \frac{1}{2}]$. Then, SGD converges to a stationary point almost surely, i.e. with probability 1. Moreover,* $\liminf_{t\to\infty} \|\nabla f(\boldsymbol{x}_t)\|^2 t^{1/2-\epsilon} = 0$ *with probability 1.*

Proof. From the result in Lemma 5, taking the limit for $T \to \infty$ and exchanging the expectation and the limits because the terms are non-negative, we have

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq f(\boldsymbol{x}_1) - f^\star + \frac{M}{2}\mathbb{E}\left[\sum_{t=1}^{\infty} \|\eta_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|_2^2\right] .$$

Observe that

$$\begin{aligned}
\sum_{t=1}^{\infty} \|\eta_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 &= \sum_{t=1}^{\infty} \eta_{t+1}^2 \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 + \sum_{t=1}^{\infty} (\eta_t^2 - \eta_{t+1}^2) \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \\
&\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \max_{t\geq 1} \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \sum_{t=1}^{\infty} (\eta_t^2 - \eta_{t+1}^2) \leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \max_{t\geq 1} \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \eta_1^2 \\
&\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\eta_1^2 \max_{t\geq 1} \|\nabla f(\boldsymbol{x}_t)\|^2 + \|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \\
&\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\frac{\alpha^2}{\beta^{1+2\epsilon}}(L^2 + S^2) < \infty,
\end{aligned} \tag{6}$$

where in the first inequality we have used Lemma 4, and in the third one the elementary inequality $\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2$.

Hence, we have $\mathbb{E}\left[\sum_{t=1}^{\infty} \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2\right] < \infty$. Now, note that $\mathbb{E}[X] < \infty$, where X is a non-negative random variable, implies that $X < \infty$ with probability 1. In fact, otherwise $\mathbb{P}[X = \infty] > 0$ implies $\mathbb{E}[X] \geq \int_{X=\infty} x d\mathbb{P}(X) = \infty$, contradicting our assumption. Hence, with probability 1, we have $\sum_{t=1}^{\infty} \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2 < \infty$.

Now, observe that the Lipschitzness of f and the bounded support of the noise on the gradients gives

$$\sum_{t=1}^{\infty} \eta_t = \sum_{t=1}^{\infty} \frac{\alpha}{(\beta + \sum_{i=1}^{t-1} \|g(\boldsymbol{x}_i, \xi_i)\|^2)^{1/2+\epsilon}} \geq \sum_{t=1}^{\infty} \frac{\alpha}{(\beta + 2(t-1)(L^2 + S^2))^{1/2+\epsilon}} = \infty \ .$$

Using the fact the f is L-Lipschitz and M-smooth, we have

$$\begin{aligned}\big|\|\nabla f(\boldsymbol{x}_{t+1})\|^2 - \|\nabla f(\boldsymbol{x}_t)\|^2\big| &= (\|\nabla f(\boldsymbol{x}_{t+1})\| + \|\nabla f(\boldsymbol{x}_t)\|) \cdot |\|\nabla f(\boldsymbol{x}_{t+1})\| - \|\nabla f(\boldsymbol{x}_t)\|| \\ &\leq 2LM\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = 2LM\|\eta_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\| \leq 2LM(L+S)\eta_t \ .\end{aligned}$$

Also, $\|\nabla f(\boldsymbol{x}_t)\|^2 \leq L^2, \ \forall t$. Hence, we can use Lemma 3 to obtain $\lim_{t\to\infty} \|\nabla f(\boldsymbol{x}_t)\|^2 = 0$.

For the second statement, observe that, with probability 1,

$$\sum_{t=1}^{\infty} \|\nabla f(\boldsymbol{x}_t)\|^2 t^{1/2-\epsilon} \frac{\alpha}{t(2L^2 + 2S^2 + \beta)^{1/2+\epsilon}} \leq \sum_{t=1}^{\infty} \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2 < \infty,$$

where in the first inequality we used the Lipschitzness of f and the bounded support of the noise on the gradients. Hence, noting that $\sum_{t=1}^{\infty} \frac{1}{t} = \infty$, we have that $\liminf_{t\to\infty} \|\nabla f(\boldsymbol{x}_t)\|^2 t^{1/2-\epsilon} = 0$. □

We now state a similar result for a version of SGD similar to AdaGrad [Duchi et al., 2011]. We use coordinate-wise adaptive stepsizes (4) as in AdaGrad, but with the power of the denominator $\frac{1}{2} + \epsilon$ with $\epsilon > 0$, rather than $\frac{1}{2}$. Also, differently from what is stated in the original AdaGrad paper, here we do not project onto a bounded closed convex set. This mirrors the actual implementation of AdaGrad in machine learning libraries, e.g. Tensorflow [Abadi et al., 2015]. Given that the proof is virtually identical to the one of Theorem 2, we defer its proof to the Appendix.

Theorem 3. *Assume (**H1, H2, H3, H4**). The stepsizes are given by a diagonal matrix $\boldsymbol{\eta}_t$ whose diagonal values are defined in* (4)*, where $\alpha, \beta > 0$ and $\epsilon \in (0, \frac{1}{2}]$. Then, SGD converges to a stationary point almost surely, i.e. with probability 1. Moreover,* $\liminf_{t\to\infty} \|\nabla f(\boldsymbol{x}_t)\|^2 t^{1/2-\epsilon} = 0$ *with probability 1.*

As far as we know, the above theorem is the first result on the convergence of AdaGrad to a stationary point, assuming $\epsilon > 0$. Also, the almost sure asymptotic convergence is the first theoretical support to the common heuristic of selecting the last iterate, rather than the minimum over the iterations.

Yet, in the above convergence guarantees the rate with which the gradient converges to zero is only asymptotic. In the next Section, we show a finite-time convergence rate for the minimum gradient over the iterates that precisely quantifies the effect of the noise on the rate.

7 Non-Asymptotic Adaptive Convergence Rates for Non-Convex Functions

We now prove non-asymptotic adaptive convergence rates to stationary points using the global stepsizes (3). This result is complementary to the one in the previous section. Given that SGD is not a descent method, we are not aware of any result of convergence with an explicit rate for the last iterate for non-convex functions. Hence, here we will prove a convergence guarantee for the *best iterate* over T iterations rather than for the *last one*. Note that choosing a random stopping time as in Ghadimi and Lan [2013] would be equivalent in expectation to choose the best iterate. For simplicity, we choose to state the theorem for the best iterate.

Theorem 4. *Assume (**H1, H3, H4'**). Let $\delta \in (0, 1)$ and the stepsizes set as* (3)*, where $\alpha, \beta > 0$, $\epsilon \in (0, \frac{1}{2})$, and $2\alpha M < \beta^{\frac{1}{2}+\epsilon}$. Then, with probability at least $1 - \delta$, the iterates of SGD satisfies the following bound*

$$\min_{1\leq t\leq T} \|\nabla f(\boldsymbol{x}_t)\|^2 \leq \frac{1}{T} \max\left(2^{\frac{1/2+\epsilon}{1/2-\epsilon}} \left(\frac{C}{\delta}\right)^{\frac{1}{1/2-\epsilon}}, \frac{4^{1/2+\epsilon}C}{\delta^{3/2+\epsilon}} \left(\frac{\beta}{2} + T\sigma^2\right)^{1/2+\epsilon}\right),$$

where $C_T = \frac{2^{1/2+\epsilon}}{\alpha\left(1-\frac{2\alpha M}{\beta^{1/2+\epsilon}}\right)} \left(f(\boldsymbol{x}_1) - f^* + \frac{\alpha^2 M}{4\epsilon\beta^{2\epsilon}} + \frac{2\alpha^2\sigma^2 M}{\beta^{1+2\epsilon}}(1+\log(T))\right)$ *and* $f^* = \min_{\boldsymbol{x}} f(\boldsymbol{x})$.

Proof. From Lemma 5, we have

$$\sum_{t=1}^{T} \mathbb{E}[\eta_t \|\nabla f(\boldsymbol{x}_t)\|^2] \leq f(\boldsymbol{x}_1) - f^\star + \frac{M}{2}\mathbb{E}\left[\sum_{t=1}^{T} \eta_t^2 \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right] .$$

Using Lemma 2, we can upper bound the expected sum in the r.h.s. of last inequality, to have

$$\mathbb{E}\left[\sum_{t=1}^{T} \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{1}{1-\frac{2\alpha M}{\beta^{1/2+\epsilon}}} \left(f(\boldsymbol{x}_1) - f^\star + \frac{\alpha^2 M}{4\epsilon\beta^{2\epsilon}} + \frac{2\alpha^2\sigma^2 M}{\beta^{1+2\epsilon}}(1+\log(T))\right) .$$

Denoting by $A = \sum_{t=1}^{T} \|\nabla f(\boldsymbol{x}_t)\|^2, B = \frac{\beta}{2} + \sum_{t=1}^{T} \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t) - \nabla f(\boldsymbol{x}_t)\|^2$, we have

$$\begin{aligned}\sum_{t=1}^{T} \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2 \geq \eta_T \sum_{t=1}^{T} \|\nabla f(\boldsymbol{x}_t)\|^2 = \eta_T A &\geq \frac{\alpha A}{\left(\beta + \sum_{t=1}^{T} \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right)^{1/2+\epsilon}} \\ \geq \frac{\alpha A}{\left(\beta + 2\sum_{t=1}^{T} (\|\nabla f(\boldsymbol{x}_t)\|^2 + \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t) - \nabla f(\boldsymbol{x}_t)\|^2)\right)^{1/2+\epsilon}} &= \frac{\alpha A}{(2A+2B)^{1/2+\epsilon}},\end{aligned}$$

where in the second inequality we used the elementary inequality $\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2$. Using the definition of C_T in the Theorem and defining $\gamma = \frac{1}{2} + \epsilon$, we have that

$$\mathbb{E}\left[\frac{A}{(A+B)^\gamma}\right] \leq C_T,$$

We now consider two cases: $B \leq A$ and $B > A$. In the first case, we have that

$$\mathbb{E}\left[\frac{A}{(2A)^\gamma}\right] \leq \mathbb{E}\left[\frac{A}{(A+B)^\gamma}\right] \leq C_T .$$

Using Markov's inequality, we have that with probability at least $1 - \delta$, $A^{1-\gamma} \leq 2^\gamma \frac{C_T}{\delta}$, that is $A \leq 2^{\frac{\gamma}{1-\gamma}} \left(\frac{C_T}{\delta}\right)^{\frac{1}{1-\gamma}}$. In the second case, we have

$$\mathbb{E}\left[\frac{A}{(2B)^\gamma}\right] \leq \mathbb{E}\left[\frac{A}{(A+B)^\gamma}\right] \leq C_T .$$

Using Markov's inequality, we have that with probability at least $1 - \delta/2$, $\frac{A}{(2B)^\gamma} \leq \frac{C_T}{\delta}$ that implies $A \leq 2^\gamma B^\gamma \frac{C_T}{\delta}$. Using again Markov inequality, we have with probability at least $1-\delta/2$, $B \leq \frac{2\mathbb{E}[B]}{\delta}$ that gives us $A \leq \frac{4^\gamma C_T}{\delta^{1+\gamma}} (\mathbb{E}[B])^\gamma$.

Putting all together and using the union bound, we have the stated bound. □

This theorem mirrors Theorem 1, proving again a convergence rate that is adaptive to the noise level. Hence, the same observations on adaptation to the noise level and convergence hold here as well. The main difference w.r.t. Theorem 1 is that here we only prove convergence to a stationary point because we do not assume convexity.

Note that such bounds were already known with an oracle tuning of the stepsizes, in particular with the knowledge of the variance of the noise, see, e.g., Ghadimi and Lan [2013]. In fact, the required stepsize in the deterministic case must be constant, while it has to be of the order of $O(\frac{1}{\sigma\sqrt{t}})$ in the stochastic case. However, here we obtain the same behaviour automatically, without having to estimate the variance of the noise, thanks to the adaptive stepsizes.

8 Discussion and Future Work

We have presented an analysis of adaptive stepsizes for stochastic gradient descent, with convex and non-convex functions. In the convex setting, our result overcomes the limitations of previous results, removing the assumption of a bounded domain, yet showing an adaptive convergence rate. In the non-convex setting, we show almost sure convergence and adaptive convergence rates to stationary points. Moreover, we show for the first time a convergence guarantee for non-convex functions for a minor variation of AdaGrad.

In the future, we would like to understand if the conditions we impose can be weakened. For example, the almost sure convergence require a bounded support noise, that, while it might be verified in many practical scenarios, still seems unsatisfying from a theoretical point of view. Also, we would like to address the issue of proving high probability finite-time convergence guarantees.

Acknowledgments

The authors thank Dávid Pál and Léon Bottou for the comments and discussions. This material is based upon work supported by the National Science Foundation under grant no. 1740762 "Collaborative Research: TRIPODS Institute for Optimization and Learning" and by a Google Research Award.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.

P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *J. Comput. Syst. Sci.*, 64(1):48–75, 2002.

D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

L. Bottou. *Une Approche théorique de l'Apprentissage Connexioniste; Applications à la reconnaissance de la Parole*. PhD thesis, Universite de Paris Sud, Centre d'Orsay, 1991.

L. Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17 (9):142, 1998.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

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

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

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

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

M. Kresoja, Z. Lužanin, and I. Stojkovska. Adaptive stochastic approximation algorithm. *Numerical Algorithms*, 76(4):917–937, Dec 2017.

J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291, 2013.

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

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2003.

F. Orabona and D. Pál. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. Special Issue on ALT 2015.

B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

X. Wu, R. Ward, and L. Bottou. WNGrad: Learn the learning rate in gradient descent. *arXiv preprint arXiv:1803.02865*, 2018.

F. Yousefian, A. Nedić, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

M. D. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Z. Zhou, P. Mertikopoulos, N. Bambos, S. Boyd, and P. W. Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, pages 7043–7052, 2017.

A Appendix

Here, we report the proofs missing from the main text.

A.1 Details of Example 1

Consider the function $f(x) = \frac{1}{2}x^2$. The gradient in t-th iteration is $\nabla f(x_t) = x_t$. Let the stochastic gradient be defined as $\boldsymbol{g}_t = \nabla f(x_t) + \xi_t$, where $P(\xi_t = \sigma_t) = P(\xi_t = -\sigma_t) = \frac{1}{2}$.

Let $A \triangleq \sum_{i=1}^{t-1} g_i^2 + \beta$. Then

$$\langle \mathbb{E}_t \eta_{t+1} \boldsymbol{g}_t, \nabla f(x_t) \rangle = \frac{\alpha}{2} \left[\frac{(x_t + \sigma_t)x_t}{[A + (x_t + \sigma_t)^2]^{\frac{1}{2}+\epsilon}} + \frac{(x_t - \sigma_t)x_t}{[A + (x_t - \sigma_t)^2]^{\frac{1}{2}+\epsilon}} \right] .$$

This expression can be negative, for example, setting $x_t = -3$, $\sigma_t = 5$, $\epsilon = 0.1$ and $A = 1$.

A.2 Proof of Lemma 4

Lemma 6. *Let $a_i \geq 0, \cdots, T$ and $f : [0, +\infty) \to [0, +\infty)$ nonincreasing function. Then*

$$\sum_{t=1}^{T} a_t f\left(a_0 + \sum_{i=1}^{t} a_i\right) \leq \int_{a_0}^{\sum_{t=0}^{T} a_t} f(x) dx .$$

Proof. Denote by $s_t = \sum_{i=0}^{t} a_i$.

$$a_i f(s_i) = \int_{s_{i-1}}^{s_i} f(s_i) dx \leq \int_{s_{i-1}}^{s_i} f(x) dx .$$

Summing over $i = 1, \cdots, T$, we have the stated bound. □

Proof of Lemma 4. The proof is immediate from Lemma 6. □

A.3 Proofs of Section 5

Proof of Lemma 1. From (2), for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} \rangle + \frac{M}{2} \|\boldsymbol{y}\|^2 .$$

Take $\boldsymbol{y} = -\frac{1}{M} \nabla f(\boldsymbol{x})$, to have

$$f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + \left(\frac{1}{2M} - \frac{1}{M}\right) \|\nabla f(\boldsymbol{x})\|^2 .$$

Hence,

$$\|\nabla f(\boldsymbol{x})\|^2 \leq 2M(f(\boldsymbol{x}) - f(\boldsymbol{x} + \boldsymbol{y})) \leq 2M(f(\boldsymbol{x}) - \min_{\boldsymbol{u}} f(\boldsymbol{u})) .$$ □

Proof of Lemma 2. Using the assumption on the noise, we have

$$\begin{aligned}
\exp\left(\frac{\mathbb{E}\left[\max_{1 \leq i \leq T} \|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2\right]}{\sigma^2}\right) &\leq \mathbb{E}\left[\exp\left(\frac{\max_{1 \leq i \leq T} \|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2}{\sigma^2}\right)\right] \\
&= \mathbb{E}\left[\max_{1 \leq i \leq T} \exp\left(\frac{\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2}{\sigma^2}\right)\right] \leq \sum_{i=1}^{T} \mathbb{E}\left[\exp\left(\frac{\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2}{\sigma^2}\right)\right] \\
&= \sum_{i=1}^{T} \mathbb{E}\left[\mathbb{E}_i\left[\exp\left(\frac{\|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2}{\sigma^2}\right)\right]\right] \leq Te,
\end{aligned}$$

that implies

$$\mathbb{E}\left[\max_{1 \leq i \leq T} \|\nabla f(\boldsymbol{x}_i) - \boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2\right] \leq \sigma^2(1 + \log T) . \tag{7}$$

Hence, we have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \eta_t^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right] &= \mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 + \sum_{t=1}^T \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2(\eta_t^2-\eta_{t+1}^2)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 + \sum_{t=1}^T \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2(\eta_t+\eta_{t+1})(\eta_t-\eta_{t+1})\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T \eta_{t+1}^2 \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 + \sum_{t=1}^T 2\eta_t \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2(\eta_t-\eta_{t+1})\right] \\
&\le \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\eta_1 \mathbb{E}\left[\max_{1\le t\le T} \eta_t \|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right] \\
&\le \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 4\eta_1 \mathbb{E}\left[\max_{1\le t\le T} \eta_t \left(\|\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)-\nabla f(\boldsymbol{x}_t)\|^2 + \|\nabla f(\boldsymbol{x}_t)\|^2\right)\right] \\
&\le \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 4\eta_1^2(1+\log(T))\sigma^2 + 4\eta_1 \mathbb{E}\left[\sum_{t=1}^T \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2\right] \\
&= \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \frac{4\alpha^2}{\beta^{1+2\epsilon}}(1+\log(T))\sigma^2 + \frac{4\alpha}{\beta^{\frac{1}{2}+\epsilon}} \mathbb{E}\left[\sum_{t=1}^T \eta_t \|\nabla f(\boldsymbol{x}_t)\|^2\right],
\end{aligned}
$$

where in second inequality we used Lemma 4 and in fourth one we used (7). □

A.4 Proofs of Section 7

Proof of Lemma 5. From (2), we have

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}) &\le f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1}-\boldsymbol{x}_t\rangle + \frac{M}{2}\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\|^2 \\
&= f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t(\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t,\xi_t))\rangle - \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle + \frac{M}{2}\|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2 .
\end{aligned}
$$

Taking the conditional expectation with respect to $\xi_1, \cdots, \xi_{t-1}$, we have that

$$E_t[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t(\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t,\xi_t))\rangle] = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t) - \boldsymbol{\eta}_t \mathbb{E}_t[\boldsymbol{g}(\boldsymbol{x}_t,\xi_t)]\rangle = 0 .$$

Hence, from the law of total expectation, we have

$$\mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle\right] \le \mathbb{E}\left[f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) + \frac{M}{2}\|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right] .$$

Summing over $t = 1$ to T and lower bounding $f(\boldsymbol{x}_{T+1})$ with $f^\star$, we have the stated bound. □

Proof of Lemma 3. Since the series $\sum_{t=1}^\infty a_t$ diverges, given that $\sum_{t=1}^\infty a_t b_t$ converges, we necessarily have $\liminf_{t\to\infty} b_t = 0$. Hence, we have to prove that $\limsup_{t\to\infty} b_t = 0$.

Let us proceed by contradiction and assume that $\limsup_{t\to\infty} b_t = \lambda > 0$. Note that $\lambda < \infty$ because the sequence is bounded. Given the values of the $\liminf$ and $\limsup$, we can then build two sequences of indices $(m_j)_{j\ge1}$ and $(n_j)_{j\ge1}$ such that

- $m_j < n_j < m_{j+1}$,
- $\frac{\lambda}{3} < b_k$, for $m_j \le k < n_j$,
- $b_k \le \frac{\lambda}{3}$, for $n_j \le k < m_{j+1}$.

Let $\epsilon = \frac{\lambda^2}{9K}$ and $\tilde{j}$ be large enough such that

$$\sum_{t=m_{\tilde{j}}}^{\infty} a_t b_t < \epsilon .$$

Then, we have for all $j \geq \tilde{j}$ and all m with $m_j \leq m \leq n_j - 1$,

$$|b_{n_j} - b_m| \leq \sum_{k=m}^{n_j-1} |b_{k+1} - b_k| \leq \frac{3K}{\lambda} \sum_{k=m}^{n_j-1} a_k \frac{\lambda}{3} \leq \frac{3K}{\lambda} \sum_{k=m}^{n_j-1} a_k b_k \leq \frac{3K}{\lambda} \sum_{k=m}^{\infty} a_k b_k \leq \frac{3K}{\lambda} \epsilon \leq \frac{\lambda}{3} \,.$$

Therefore, using the triangle inequality,

$$b_m \leq b_{n_j} + \frac{\lambda}{3} \leq \frac{2\lambda}{3} \,.$$

And finally for all $m \geq \tilde{j}$,

$$b_m \leq \frac{2\lambda}{3},$$

which contradicts $\limsup_{t\to\infty} b_t = \lambda > 0$. Therefore, b_t goes to zero. □

Proof of Theorem 3. We proceed similarly to the proof of Theorem 2, to get

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle\right] \leq f(\boldsymbol{x}_1) - f(\boldsymbol{x}^\star) + \frac{M}{2}\mathbb{E}\left[\sum_{t=1}^{\infty} \|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|_2^2\right] \,.$$

Observe that

$$\sum_{t=1}^{\infty} \|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 = \sum_{t=1}^{\infty}\sum_{i=1}^{d} \eta_{t,i}^2 \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)_i^2 = \sum_{i=1}^{d}\sum_{t=1}^{\infty} \eta_{t,i}^2 \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)_i^2 < \infty,$$

where the last inequality comes from the same reasoning in (6). Hence, we have

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle\right] < \infty \,.$$

Hence, with probability 1, we have

$$\sum_{t=1}^{\infty} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{\eta}_t \nabla f(\boldsymbol{x}_t)\rangle = \sum_{t=1}^{\infty}\sum_{j=1}^{d} \eta_{t,j} \nabla f(\boldsymbol{x}_t)_j^2 = \sum_{j=1}^{d}\sum_{t=1}^{\infty} \eta_{t,j} \nabla f(\boldsymbol{x}_t)_j^2 < \infty \,.$$

and, for any $j = 1, \cdots, d$,

$$\sum_{t=1}^{\infty} \eta_{t,j} (\nabla f(\boldsymbol{x}_t))_j^2 < \infty \,.$$

Now, observe that the Lipschitzness of f and the bounded support of the noise on the gradients gives

$$\sum_{t=1}^{\infty} \eta_{t,j} = \sum_{t=1}^{\infty} \frac{\alpha}{(\beta + \sum_{i=1}^{t-1} (g(\boldsymbol{x}_i, \xi_i)_j)^2)^{1/2+\epsilon}} \geq \sum_{t=1}^{\infty} \frac{\alpha}{(\beta + 2(t-1)(L^2 + S^2))^{1/2+\epsilon}} = \infty \,.$$

Using the fact the f is L-Lipschitz and M-smooth, we also have

$$\begin{aligned}\left|((\nabla f(\boldsymbol{x}_{t+1}))_j)^2 - ((\nabla f(\boldsymbol{x}_t))_j)^2\right| &= ((\nabla f(\boldsymbol{x}_{t+1}))_j + (\nabla f(\boldsymbol{x}_t))_j) \cdot |(\nabla f(\boldsymbol{x}_{t+1}))_j - (\nabla f(\boldsymbol{x}_t))_j| \\ &\leq 2LM\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = 2LM\|\boldsymbol{\eta}_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\| \leq 2LM(L+S)\eta_t \,.\end{aligned}$$

Hence, we case use Lemma 3 to obtain

$$\lim_{t\to\infty} ((\nabla f(\boldsymbol{x}_t))_j)^2 = 0 \,.$$

For the second statement, observe that, with probability 1,

$$\sum_{t=1}^{\infty} ((\nabla f(\boldsymbol{x}_t))_j)^2 t^{1/2-\epsilon} \frac{\alpha}{t(2L^2 + 2S^2 + \beta)^{1/2+\epsilon}} \leq \sum_{t=1}^{\infty} \eta_{t,j} ((\nabla f(\boldsymbol{x}_t))_j)^2 < \infty \,.$$

Hence, noting that $\sum_{t=1}^{\infty} \frac{1}{t} = \infty$, we have that $\liminf_{t\to\infty} ((\nabla f(\boldsymbol{x}_t))_j)^2 t^{1/2-\epsilon} = 0$. □