On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes

Xiaoyu Li
Boston University

Francesco Orabona
Boston University

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 in the convex and non-convex settings a generalized version of the AdaGrad stepsizes. We show sufficient conditions for these stepsizes to achieve almost sure asymptotic convergence of the gradients to zero, proving the first guarantee for generalized AdaGrad stepsizes in the non-convex setting. Moreover, we show that these stepsizes allow to automatically adapt to the level of noise of the stochastic gradients in both the convex and non-convex settings, interpolating between $O(1/T)$ and $O(1/\sqrt{T})$, up to logarithmic terms.

1 INTRODUCTION

In 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.

More in details, SGD 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 $\boldsymbol{x}_1$, where $\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)$ is a stochastic gradient in $\boldsymbol{x}_t$ that depends on the stochastic variable ξ_t. Classic convergence analysis of the SGD algorithm for non-convex smooth functions relies on conditions on the positive stepsizes η_t (Robbins and Monro, 1951). In particular, sufficient conditions are that $(\eta_t)_{t=1}^{\infty}$ is a deterministic sequence of non-negative numbers that satisfies

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

However, state-of-the-art SGD variants use *adaptive stepsizes*, that is η_t is a function of past stochastic gradients. These stepsizes are believed to require less tweaking to achieve good performance in machine learning applications and we have some partial explanations in the convex setting, i.e. sparsity of the gradients (Duchi et al., 2011). However, in the non-convex setting, we do not have any theory explaining the better performance.

Indeed, for a large number of SGD variants employed by practitioners the conditions above are not satisfied. In fact, these algorithms are often designed and analyzed for the convex domain under restrictive conditions, e.g. bounded domains, 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). Even considering an *infinite* number of iterations, the behavior of these algorithms is often unknown.

We focus on a generalized version of the adaptive stepsizes popularized by AdaGrad (Duchi et al., 2011). This kind of stepsizes 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 two types of step size: a global step size

$$\eta_t = \frac{\alpha}{\left(\beta + \sum_{i=1}^{t-1} \|\boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2\right)^{1/2+\epsilon}} \tag{2}$$

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

and a coordinate-wise one

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

where $\alpha > 0$ and $\beta, \epsilon \geq 0$. Note that, with $\epsilon = 0$, (3) are the coordinate-wise stepsizes used in AdaGrad (Duchi et al., 2011), while (2) have been used in online convex optimization to achieve adaptive regret guarantees, (e.g. Rakhlin and Sridharan, 2013; Orabona and Pál, 2018). The additional parameter ϵ allows us to increase the decrease rate of the stepsize and it will be critical to obtain our almost sure convergence results.

In this paper, we want to answer two basic questions: 1) Are there conditions under which the generalized AdaGrad stepsize converge almost surely with an infinite number of iterations in the non-convex setting? 2) Are there conditions under which the rate is better than the one of the plain SGD with decreasing stepsizes?

We answer positively to both questions. More in details, the contributions of this paper are the following:

- In Section 5, we prove for the *first* time in the non-convex setting almost sure asymptotic convergence to zero of the gradients of SGD with both coordinate-wise and global adaptive stepsizes.
- In Section 6.1, we prove that in the convex setting the generalized global AdaGrad stepsizes adapts to the noise level, through a finite-time convergence rate. In particular, we show that, depending on the noise level, SGD with the generalized AdaGrad updates automatically interpolates between the convergence rates of Gradient Descent (GD) and SGD, up to polylogarithmic terms. We do so *removing the strong assumptions* present in previous analyses.
- In Section 6.2, similarly to the results of Section 6.1, we show that in the non-convex setting the generalized global AdaGrad stepsizes adapts to the noise level, through a novel finite-time convergence rate. A low noise will result in an automatic faster convergence. As far as we know, these are the *first* theoretical results for the advantage of AdaGrad-like stepsizes over the plain SGD in the non-convex setting.

The next Section discusses more in details the related work, while Section 3 introduces formally the setting, and Section 4 discusses the details of 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 optimization one later (Duchi et al., 2011). In particular, in (Duchi et al., 2011) they prove that AdaGrad can converge faster if the gradients are sparse and the function is convex. Yet, most of these studies assumed the optimization to be constrained in a convex bounded set. This assumption is often false 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 norm of the gradients strictly greater than zero. Levy et al. (2018) propose an acceleration method with adaptive stepsizes which are also similar to our global ones, proving the $\tilde{O}(1/T^2)$ convergence in the deterministic smooth case and $\tilde{O}(1/\sqrt{T})$ in both general deterministic case and stochastic smooth case, but requiring a bounded-domain assumption.

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, both approaches do not cover adaptive stepsizes.

The first 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. Later, Ward et al. (2018), independently and the same time with us, improved their guarantees proving results similar to our Theorems 3 and 4. They use the original AdaGrad stepsizes, but with the assumption of bounded expected squared norm of the

stochastic gradients. Some other related works were proposed after our submission. Zhou et al. (2018) analyze an adaptive gradient method in the non-convex setting, but their bounds give advantages only in very sparse case.

A 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 in Section 5, 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 $\xi_1, \cdots, \xi_{t-1}$. We use 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)

$$|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle| \leq \frac{M}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \ . \quad (4)$$

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

(**H4'**) has been already used by Nemirovski et al. (2009) to prove high probability convergence guarantees. This condition allows to control the expectation of the maximum of the terms $\|\nabla f(\boldsymbol{x}_t) - g(\boldsymbol{x}_t, \xi_t)\|^2$. Note that, using Jensen's inequality, this condition implies a bounded variance. Also, (**H4**) implies (**H4'**).

4 KEEPING THE UPDATE DIRECTION UNBIASED

A key difference between the generalized AdaGrad stepsizes in (2) and (3) with the AdaGrad stepsizes in Duchi et al. (2011) is the fact that $\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)$ is not used in η_t. It is easy to see that doing otherwise introduces a spurious bias in the update direction. 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, especially when the function is not Lipschitz. So, in the following, we will analyze this minor variant of the AdaGrad stepsizes.

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 8 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. We believe that removing the assumption of knowing the Lipschitz constant more closely follow the use in real-world applications.

In the following, we will show that these stepsize allow to prove adaptive guarantees in the convex and non-convex settings.

5 ALMOST SURE CONVERGENCE FOR NON-CONVEX FUNCTIONS

In this section, we show that SGD with the generalized AdaGrad stepsizes in (2) and (3) allows to decrease the gradients to zero almost surely, that is, with probability 1. This is considered a required basic property for any optimization algorithm.

The stepsizes in (2) and (3) *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 obtained 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 that the sum of the stepsizes diverges.

We now state our almost sure convergence results.

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

We also state a similar result for the coordinate-wise stepsizes in (3). We remind the reader that these stepsizes closely mirror the ones used 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).

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

As far as we know, the above theorems are the first results on the almost sure convergence of the gradients using generalized AdaGrad stepsizes and assuming $\epsilon > 0$. In particular, Theorem 2 is the first theoretical support to the common heuristic of selecting the last iterate, rather than the minimum over the iterations.

For the proofs, we will need some technical lemmas, whose proofs are in the Appendix.

Lemma 1. *(Alber et al., 1998, Proposition 2)(Mairal, 2013, Lemma A.5) Let* $(a_t)_{t\geq 1}, (b_t)_{t\geq 1}$ *be two non-negative real sequences. Assume that* $\sum_{t=1}^{\infty} a_t b_t$ *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 2. *Let* $a_0 > 0$*,* $a_i \geq 0$*,* $i = 1, \cdots, T$ *and* $\beta > 1$*. Then* $\sum_{t=1}^{T} \frac{a_t}{(a_0 + \sum_{i=1}^{t} a_i)^{\beta}} \leq \frac{1}{(\beta-1) a_0^{\beta-1}}$*.*

We now state a Lemma that allows us to study the progress made in T steps. The proof is in the Appendix.

Lemma 3. *Assume (**H1**, **H3**). Then, the iterates of SGD with stepsizes* $\boldsymbol{\eta}_t \in \mathbb{R}^{d\times d}$ *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 can prove Theorem 1. Given that the proof of Theorem 2 is virtually identical to the one of Theorem 1, we defer its proof to the Appendix.

Proof of Theorem 1. From the result in Lemma 3, 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

$$\begin{aligned}
&\sum_{t=1}^{\infty} \|\eta_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \\
&= \sum_{t=1}^{\infty} \eta_{t+1}^2 \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 + \sum_{t=1}^{\infty} (\eta_t^2 - \eta_{t+1}^2) \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \\
&\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \max_{t\geq 1} \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \sum_{t=1}^{\infty} (\eta_t^2 - \eta_{t+1}^2) \\
&\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \max_{t\geq 1} \|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \eta_1^2 \\
&\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\eta_1^2 \max_{t\geq 1} \|\nabla f(\boldsymbol{x}_t)\|^2 + \|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \\
&\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\frac{\alpha^2}{\beta^{1+2\epsilon}}(L^2 + S^2) < \infty, \qquad (5)
\end{aligned}$$

where in the first inequality we have used Lemma 2, 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

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

Hence, we can use Lemma 1 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 $\liminf_{t\to\infty} \|\nabla f(\boldsymbol{x}_t)\|^2 t^{1/2-\epsilon} = 0$. □

Even if the above results hold with probability 1, the above convergence guarantees rates are only asymptotic. So, in the next Section, we show finite-time convergence rates in expectation. Moreover, we will show that the generalized AdaGrad stepsizes adapt to the level of noise for both the convex and non-convex case.

6 ADAPTIVE CONVERGENCE RATES

We will now show that the global generalized AdaGrad stepsizes give rises to adaptive convergence rates. In particular, we will show that for a large range of the parameters α, β, ϵ and independently from the noise variance σ, the algorithms will have a faster convergence when σ is small and worst-case optimal convergence when σ is large. Note that to achieve the same behavior with SGD we should use a different stepsize for each level of noise.

In the following, we will consider the convex and non-convex case.

6.1 Adaptive Convergence for Convex Functions

As a warm-up, in this section, we show that the global stepsizes (2) give adaptive rates of convergence that interpolate between the rate of GD and SGD, for a wide range of the parameters α, β, and ϵ and without knowledge of the variance of the noise. Note that, differently from the other proofs on SGD with adaptive rates (e.g. Duchi et al., 2011), we do not assume to use projections onto bounded domains. This makes our novel proof more technically challenging, but at the same time, it mirrors the setting of many applications of SGD in machine learning optimization problems.

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

$$\mathbb{E}\left[(f(\bar{\boldsymbol{x}}_T) - f(\boldsymbol{x}^\star))^{1/2-\epsilon}\right] \leq \frac{1}{T^{1/2-\epsilon}} \max\left(2^{\frac{1}{1/2-\epsilon}} M^{1/2+\epsilon} \gamma, \left(\beta + T\sigma^2\right)^{1/4-\epsilon^2} \gamma^{1/2-\epsilon}\right),$$

where $\bar{\boldsymbol{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$ and

$$\gamma = \begin{cases} O\left(\frac{1+\alpha^2 \ln T}{\alpha(1-\frac{4\alpha M}{\sqrt{\beta}})}\right), & \text{for } \epsilon = 0 \\ O\left(\frac{1+\alpha^2(\frac{1}{\epsilon}+\sigma^2 \ln T)}{\alpha(1-\frac{4\alpha M}{\beta^{1/2+\epsilon}})}\right), & \text{for } \epsilon > 0. \end{cases}$$

Remark. Using Markov's inequality, from the above bound it is immediate to get that, with probability at least $1-\delta$, we have

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

Up to polylog terms, if $\sigma = 0$ this recovers the GD rate, $O(\frac{1}{T})$, and otherwise we get the worst-case optimal rate of SGD, $O(\frac{1}{\sqrt{T}})$. The same behavior was proved in Dekel et al. (2012) with the knowledge of σ and stepsize depending on it. Instead, 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.

For the proof, we first state some technical lemmas, whose proofs are in the Appendix.

Lemma 4. *Assume (**H1**). Then* $\|\nabla f(\boldsymbol{x})\|^2 \leq 2M(f(\boldsymbol{x}) - \min_{\boldsymbol{y}} f(\boldsymbol{y}))$, $\forall \boldsymbol{x}$.

Lemma 5. *If* $x \geq 0$ *and* $x \leq C(A + Bx)^{\frac{1}{2}+\epsilon}$, *then* $x < \max([C(2B)^{\frac{1}{2}+\epsilon}]^{\frac{1}{1/2-\epsilon}}, C(2A)^{\frac{1}{2}+\epsilon})$.

Lemma 6. *If* $x \geq 0$, $A, C, D \geq 0$, $B > 0$, *and* $x^2 \leq (A + Bx)(C + D\ln(A + Bx))$, *then* $x < 32B^3D^2 + 2BC + 8B^2D\sqrt{C} + A/B$.

Lemma 7. *If* $x, y \geq 0$ *and* $0 \leq p \leq 1$, *then* $(x+y)^p \leq x^p + y^p$.

Lemma 8. *Assume (**H1**, **H3**, **H4'**). The stepsizes are chosen as* (2), *where* $\alpha, \beta, \epsilon \geq 0$. *Then,*

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

where in the case of $\epsilon > 0$, $K = \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}}$; *when* $\epsilon = 0$, $K = 2\alpha^2 \ln\left(\sqrt{\beta + 2T\sigma^2} + \sqrt{2}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla f(\boldsymbol{x}_t)\|^2}\right]\right)$.

We can now prove the theorem.

Proof of Theorem 3. For simplicity, denote by $\delta_t := f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)$ and by $\Delta := \sum_{t=1}^{T} \delta_t$.

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

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

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

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

From Lemma 4 and Lemma 8, when $\epsilon > 0$ we have that

$$\left(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\right)\mathbb{E}\left[\sum_{t=1}^{T} \eta_t \delta_t\right] \leq \frac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}_1\|^2 + \frac{\alpha^2}{4\epsilon\beta^{2\epsilon}} + \frac{2\alpha^2}{\beta^{1+2\epsilon}}(1 + \ln T)\sigma^2. \tag{6}$$

On the other hand, when $\epsilon = 0$ we have

$$\left(1 - \frac{4\alpha M}{\beta^{1/2}}\right)\mathbb{E}\left[\sum_{t=1}^{T} \eta_t \delta_t\right] \leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \frac{2\alpha^2}{\beta}(1 + \ln T)\sigma^2 + \alpha^2 \ln\left(\sqrt{\beta + 2T\sigma^2} + 2\sqrt{M}\mathbb{E}\left[\sqrt{\Delta}\right]\right). \tag{7}$$

We can also lower bound the l.h.s. of (6) and (7) with

$$\mathbb{E}\left[\sum_{t=1}^{T} \eta_t \delta_t\right] \geq \mathbb{E}\left[\eta_T \Delta\right] \geq \frac{\left(\mathbb{E}\left[\Delta^{1/2-\epsilon}\right]\right)^{\frac{1}{1/2-\epsilon}}}{\left(\mathbb{E}\left[\left(\frac{1}{\eta_T}\right)^{\frac{1/2-\epsilon}{1/2+\epsilon}}\right]\right)^{\frac{1/2+\epsilon}{1/2-\epsilon}}},$$

where the second inequality is due to Hölder's inequality, i.e. $\mathbb{E}[B^p] \geq \frac{\mathbb{E}[AB]^p}{\mathbb{E}[A^q]^{p/q}}$, with $\frac{1}{p} = \frac{1}{2} - \epsilon$, $\frac{1}{q} = \frac{1}{2} + \epsilon$, $A = (\frac{1}{\eta_T})^{\frac{1}{p}}$, and $B = [\eta_T \Delta]^{\frac{1}{p}}$. We also have

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

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 4 in the second one.

Define

$$\gamma = \frac{1}{\alpha(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}})}\left(\|\boldsymbol{x}^\star - \boldsymbol{x}_1\|^2 + \frac{4\alpha^2}{\beta^{1+2\epsilon}}(1 + \ln T)\sigma^2\right) + K,$$

where K will be defined in the following for the case $\epsilon = 0$ and $\epsilon > 0$.

When $\epsilon > 0$, we have

$$\begin{aligned}
&\frac{1}{\gamma^{\frac{1/2-\epsilon}{1/2+\epsilon}}}\left(\mathbb{E}\left[\Delta^{1/2-\epsilon}\right]\right)^{\frac{1}{1/2+\epsilon}} \leq \alpha^{\frac{1/2-\epsilon}{1/2+\epsilon}}\mathbb{E}\left[\left(\frac{1}{\eta_T}\right)^{\frac{1/2-\epsilon}{1/2+\epsilon}}\right] \\
&\leq \mathbb{E}\left[\left(\beta + 2\sum_{t=1}^{T-1}(\|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 + 2M\delta_t)\right)^{1/2-\epsilon}\right] \\
&\leq \mathbb{E}\left[\left(\beta + 2\sum_{t=1}^{T-1} \|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right)^{1/2-\epsilon}\right] \\
&\quad + \mathbb{E}\left[\left(4M\sum_{t=1}^{T-1} \delta_t\right)^{1/2-\epsilon}\right] \\
&\leq \left(\beta + 2(T-1)\sigma^2\right)^{1/2-\epsilon} + (4M)^{1/2-\epsilon}\mathbb{E}\left[\Delta^{1/2-\epsilon}\right],
\end{aligned} \tag{8}$$

where in the third inequality we used Lemma 7 and we define $K = \frac{\frac{\alpha^2}{2\epsilon\beta^{2\epsilon}}}{\alpha(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}})}$. Proceeding in the same way, for the case $\epsilon = 0$ we get

$$\left(\mathbb{E}\left[\sqrt{\Delta}\right]\right)^2 \leq \left(A + B\mathbb{E}\left[\sqrt{\Delta}\right]\right) \times \left(C + D\ln\left(A + B\mathbb{E}\left[\sqrt{\Delta}\right]\right)\right),$$

where $A = \sqrt{\beta + 2T\sigma^2}$, $B = 2\sqrt{M}$, $D = \frac{\alpha}{1-\frac{4\alpha M}{\sqrt{\beta}}}$ and $C = \frac{\beta\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2+4\alpha^2(1+\ln T)\sigma^2}{2\alpha\beta(1-\frac{4\alpha M}{\sqrt{\beta}})}$. Using Lemma 6, we have that

$$\mathbb{E}\left[\sqrt{\Delta}\right] \leq 32B^3D^2 + 2BC + 8B^2D\sqrt{C} + \frac{A}{B}.$$

We use this upper bound in the logarithmic term, so that for $\epsilon \geq 0$, we have (8) again, this time with $K = D\ln(2A+32B^4D^2+2B^2C+8B^3D\sqrt{C}) = O(\frac{\ln T}{1-\frac{4\alpha M}{\sqrt{\beta}}})$.

Hence, we proceed using Lemma 5 to have for $\epsilon \geq 0$

$$\mathbb{E}\left[\Delta^{1/2-\epsilon}\right] \leq \max\left(2^{\frac{1/2+\epsilon}{1/2-\epsilon}}(4M)^{1/2+\epsilon}\gamma, 2^{1/2+\epsilon}\gamma^{1/2-\epsilon}\left(\beta+2T\sigma^2\right)^{1/4-\epsilon^2}\right).$$

Using Jensen's inequality on the l.h.s. of last inequality concludes the proof. □

6.2 Adaptive Convergence for Non-Convex Functions

We now prove that the generalized AdaGrad stepsizes in (2) allow a faster convergence of the gradients to zero when the noise over the gradients is small.

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 the stepsizes set as* (2)*, where $\alpha, \beta > 0$, $\epsilon \in (0, \frac{1}{2})$, and $2\alpha M < \beta^{\frac{1}{2}+\epsilon}$. Then, the iterates of SGD satisfies the following bound*

$$\mathbb{E}\left[\min_{1\leq t\leq T}\|\nabla f(\boldsymbol{x}_t)\|^{1-2\epsilon}\right] \leq \frac{1}{T^{1/2-\epsilon}}\max\left(2^{\frac{1/2+\epsilon}{1/2-\epsilon}}\gamma, 2^{1/2+\epsilon}\left(\beta+2T\sigma^2\right)^{1/4-\epsilon^2}\gamma^{1/2-\epsilon}\right),$$

$$\text{where } \gamma = \begin{cases} O\left(\frac{1+\alpha^2\ln T}{\alpha(1-\frac{2\alpha}{\sqrt{\beta}})}\right) & \text{for } \epsilon = 0 \\ O\left(\frac{1+\alpha^2(\frac{1}{\epsilon}+\sigma^2\ln T)}{\alpha(1-\frac{2\alpha}{\beta^{1/2+\epsilon}})}\right) & \text{for } \epsilon > 0. \end{cases}$$

Remark. As in the previous Section, using Markov's inequality it's easy to get that, with probability at least $1-\delta$,

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

This theorem mirrors Theorem 3, 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 3 is that here we only prove that the gradients are converging to zero rather than the suboptimality gap, 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 behavior automatically, without having to estimate the variance of the noise, thanks to the adaptive stepsizes. This shows for the first time a clear advantage of the global generalized AdaGrad stepsizes over plain SGD.

Proof of Theorem 4. For simplicity, denote by $\Delta := \sum_{t=1}^T \|\nabla f(\boldsymbol{x}_t)\|^2$.

From Lemma 3, 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 8, we can upper bound the expected sum in the r.h.s. of last inequality. When $\epsilon > 0$, we have

$$\left(\frac{1}{1-\frac{2\alpha M}{\beta^{1/2+\epsilon}}}\right)\mathbb{E}\left[\sum_{t=1}^T \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq f(\boldsymbol{x}_1)-f^\star + \frac{\alpha^2 M}{4\epsilon\beta^{2\epsilon}} + \frac{2\alpha^2\sigma^2 M}{\beta^{1+2\epsilon}}(1+\ln T). \tag{9}$$

When $\epsilon = 0$, we have

$$\left(\frac{1}{1-\frac{2\alpha M}{\sqrt{\beta}}}\right)\mathbb{E}\left[\sum_{t=1}^T \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq f(\boldsymbol{x}_1)-f^\star + M\alpha^2\ln\left(\sqrt{\beta+2T\sigma^2}+\sqrt{2}\mathbb{E}\left[\sqrt{\sum_{t=1}^T \eta_t\|\nabla f(\boldsymbol{x}_t)\|^2}\right]\right) + \frac{2\alpha M}{\beta}(1+\ln T)\sigma^2. \tag{10}$$

With similar methods in the proof of Theorem 3, we lower bound the l.h.s. of both (9) and (10) with

$$\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\|\nabla f(\boldsymbol{x}_t)\|^2\right] \geq \mathbb{E}\left[\eta_T\Delta\right] = \mathbb{E}\left[\eta_T\Delta\right] \geq \frac{\left(\mathbb{E}\left[\Delta^{1/2-\epsilon}\right]\right)^{\frac{1}{1/2-\epsilon}}}{\left(\mathbb{E}\left[\left(\frac{1}{\eta_T}\right)^{\frac{1/2-\epsilon}{1/2+\epsilon}}\right]\right)^{\frac{1/2+\epsilon}{1/2-\epsilon}}}.$$

We also have

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

Define

$$\gamma = \frac{1}{\alpha(1-\frac{2\alpha M}{\beta^{1/2+\epsilon}})}\left(f(\boldsymbol{x}_1) - f^\star + \frac{2\alpha^2 M}{\beta^{1+2\epsilon}}\sigma^2\right) + K,$$

where K will be defined separately for the case $\epsilon = 0$ and $\epsilon > 0$.

When $\epsilon > 0$, we have

$$\begin{aligned}
&\left(\mathbb{E}\left[\Delta^{1/2-\epsilon}\right]\right)^{\frac{1}{1/2-\epsilon}} \\
&\leq \alpha\gamma\left(\mathbb{E}\left[\left(\frac{1}{\eta_T}\right)^{\frac{1/2-\epsilon}{1/2+\epsilon}}\right]\right)^{\frac{1/2+\epsilon}{1/2-\epsilon}} \\
&\leq \gamma\left(\mathbb{E}\left[\left(\beta + 2\sum_{t=1}^{T-1}\|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_t,\xi_t)\|^2\right)^{1/2-\epsilon}\right]\right. \\
&\quad \left. +2\mathbb{E}\left[\left(\sum_{t=1}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|^2\right)^{1/2-\epsilon}\right]\right)^{\frac{1/2+\epsilon}{1/2-\epsilon}} \\
&\leq \gamma\left(\left(\beta + 2T\sigma^2\right)^{1/2-\epsilon} + 2\mathbb{E}\left[\Delta^{1/2-\epsilon}\right]\right)^{\frac{1/2+\epsilon}{1/2-\epsilon}}. \qquad (11)
\end{aligned}$$

where in this case we define $K = \frac{\frac{\alpha M}{4\epsilon\beta^{2\epsilon}}}{\alpha(1-\frac{2\alpha M}{\beta^{1/2+\epsilon}})}$. Proceeding in the same way, when $\epsilon = 0$, we have

$$\left(\mathbb{E}\left[\sqrt{\Delta}\right]\right)^2 \leq \left(A + B\mathbb{E}\left[\sqrt{\Delta}\right]\right) \times \left(C + D\ln\left(A + B\mathbb{E}\left[\sqrt{\Delta}\right]\right)\right),$$

where $A = \sqrt{\beta + 2T\sigma^2}$, $B = \sqrt{2}$, $D = \frac{\alpha M}{1-\frac{2\alpha M}{\sqrt{\beta}}}$, $C = \frac{\beta(f(\boldsymbol{x}_1)-f^\star)+2\alpha(1+\ln T)\sigma^2}{\alpha\beta(1-\frac{2\alpha M}{\sqrt{\beta}})}$.

Using Lemma 6, we have that

$$\mathbb{E}\left[\sqrt{\Delta}\right] \leq 32B^3D^2 + 2BC + 8B^2D\sqrt{C} + \frac{A}{B}.$$

Similar with Theorem 3, we use this upper bound in the logarithmic term so that for $\epsilon \geq 0$, we have (11) again, this time with $K = D\ln(2A+32B^4D^2+2B^2C+8B^3D\sqrt{C}) = O(\frac{\ln T}{1-\frac{2\alpha M}{\beta}})$.

Hence, we proceed using Lemma 5 to have for $\epsilon \geq 0$

$$\mathbb{E}\left[\Delta^{1/2-\epsilon}\right] \leq \max\left(2^{\frac{1/2+\epsilon}{1/2-\epsilon}}\gamma, 2^{1/2+\epsilon}\left(\beta + 2T\sigma^2\right)^{1/4-\epsilon^2}\gamma^{1/2-\epsilon}\right).$$

Lower bounding $\mathbb{E}\left[\Delta^{1/2-\epsilon}\right]$ by $T^{1/2-\epsilon}\mathbb{E}\left[\min_{1\leq t\leq T}\|\nabla f(\boldsymbol{x}_t)\|^{1-2\epsilon}\right]$, we have the stated bound. □

7 DISCUSSION AND FUTURE WORK

We have presented an analysis of adaptive stepsizes based on the generalized AdaGrad stepsizes for stochastic gradient descent, with convex and non-convex functions. In the convex setting, our result shows an adaptive convergence rate, also overcoming the limitations of previous results. In the non-convex setting, we show almost sure convergence and adaptive convergence rates. Moreover, we show for the first time sufficient condition for a convergence guarantee for non-convex functions for a minor variation of AdaGrad.

We believe these results have twofold importance. First, we go in the direction of closing the gap between theory and practice for widely used optimization algorithms. Second, our adaptive rates provide a possible explanation for the empirical success of these kinds of algorithms in practical machine learning applications.

One of the limitations of the current analysis is the fact that our analysis implies high probability bounds that depends polynomially on $\frac{1}{\delta}$, due to the application of Markov's inequality. It would be better to prove bounds that depend on $\ln(\frac{1}{\delta})$, as they are possible for SGD under conditions (1). However, the generalized AdaGrad updates do not satisfy the conditions (1) and the analysis is not straightforward. Our future work will focus on shedding light on this issue.

In the future, we would also like to understand if the conditions we impose can be weakened. For example, the almost sure convergence requires a bounded support noise, that, while it might be verified in many practical scenarios, still seems unsatisfying from a theoretical point of view. Moreover, we would like to adapt the recent approaches for parameter-free online optimization (Orabona and Pal, 2016; Cutkosky and Orabona, 2018) to the non-convex setting.

Acknowledgements

The authors thank Dávid Pál for the comments and discussions and Léon Bottou for the comments on prior work. 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/`.

Ya. I. Alber, A. N. Iusem, and M. V Solodov. On the projected subgradient method for nonsmooth convex optimization in a hilbert space. *Mathematical Programming*, 81(1):23–35, 1998.

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.

A. Cutkosky and F. Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Proc. of COLT*, 2018.

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.

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.

K. Y. Levy, A. Yurtsever, and V. Cevher. Online adaptive methods, universality and acceleration, 2018. arXiv:1809.02864.

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. Pal. Coin betting and parameter-free online learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 577–585. Curran Associates, Inc., 2016.

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

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.

R. Ward, X. Wu, and L. Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv preprint arXiv:1806.01811*, 2018.

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.

D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. On the convergence of adaptive gradient methods for nonconvex optimization, 2018. arXiv:1808.05671.

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.