Spencer Shortt
APPM 5650: Randomized Algorithms
Dr. Becker
Final Paper
October 22, 2024

On Convergence of Stochastic Gradient Descent with Adaptive Step Sizes Li and Orabona '19

1 Introduction

The objective of this project is to redo some of the proofs presented in Li and Orabona '19 and provide extra details and analysis when necessary. I was motivated to learn more about this problem after watching a video on stochastic gradient descent (SGD). The professor in the video gave visuals for an SGD simulation and was able to vary the step size of the simulation with a slider, illustrating the behavior of the optimization method for a range of step sizes. I was intrigued by this topic and quickly found this paper. I chose a more theoretical paper because I wanted more practice reading through mathematical papers, since I hope to write some of my own while here at CU Boulder. This paper is very self-contained, which made it easier to work through for someone still getting accustomed to reading publications.

This project is directly related to a topic (SGD) studied in class. In class, we discussed the original Robbins-Monro proof, as well as some variations. This project investigates the subject further but looking at a different type of variation, which is a variation of step size choice. The proofs in this paper also make use of many of the inequalities discussed in the beginning of the semester. We will now start by discussing some of the motivation for the problems addressed in the paper.

Stochastic gradient descent is a popular optimization method, especially within the machine learning community. It is most recognizable by some form of the following equation:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}(\mathbf{x}_t, \xi_t),$$

where η_t is the step size and $\mathbf{g}(\mathbf{x}_t, \xi_t)$ is what we're calling the stochastic gradient (we'll see why later). From a theoretical perspective especially, how to find the best choice of step size η_t is still an open problem (best often referring to the balance between accuracy and computational efficiency). This is why I took interest in this topic, but it is also the motivation given in the abstract. This paper works towards bridging the gap between theory and application of choosing step sizes.

In the setting Robbins and Monro gave in 1951, $(\eta_t)_{t=1}^{\infty}$ was a positive, deterministic sequence such that

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

Presently, 70 years later, there have been advancements in SGD theory, but often not to the point of explaining the state-of-the-art variations being used in practice. One class of these variations is those that use *adaptive* stepsizes.

Adaptive stepsize refer to when η_t is a function of the previous stochastic gradients. Since my historical knowledge of adaptive stepsizes is limited, I quote Li and Orabona: "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)." The paper then proceeds to give two examples of adaptive stepsizes, which are generalized versions from Duchi et al., 2011:

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

which is referred to as global generalized AdaGrad, and the following coordinate-wise generalized AdaGrad:

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

where $\alpha > 0$ and $\beta, \epsilon \geq 0$. The proofs throughout the paper are work for both adaptive stepsize choices, so here we will only focus on the global generalized AdaGrad.

The paper works towards answering the following two questions:

- Under what conditions (if any) do we get almost sure convergence using this adaptive step size after infinite iterations in the non-convex setting?
- Are there conditions for which the convergence rate is better than our standard SGD method?

2 Necessary Assumptions

We will be using the $L2$ norm unless stated otherwise. The following assumptions are used frequently throughout the paper:

- H1: f is M-*smooth*, so its gradient is M-Lipschitz, meaning $||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \leq M||\mathbf{x} - \mathbf{y}||$, for all $x, y \in \mathbb{R}^d$.
- H2: f is L-Lipschitz, so $|f(\mathbf{x}) - f(\mathbf{y})| \leq L||\mathbf{x} - \mathbf{y}||$
- H3: $\mathbb{E}_{\xi_t}[\mathbf{g}(\mathbf{x}_t, \xi_t)] = \nabla f(\mathbf{x}_t)$ (hence why it is called stochastic gradient)
- H4: The noise of the stochastic gradient has bounded support, meaning for all $\mathbf{x}$,

$$||\mathbf{g}(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_t)|| \leq S$$

- H4': For all x , the stochastic gradient satisfies

$$\mathbb{E}_{\xi_t}[\exp(||\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t, \xi_t)||^2/\sigma^2)] \leq \exp(1)],$$

where $\mathbb{E}_{\xi_t}$ is the conditional expectation relative to the sigma algebra generated by the first $t-1$ random variables ξ.

3 Almost sure convergence for non-convex functions

This isn't intended to be full summary of the paper, so I will try to focus on mostly math from here on out. As a disclaimer, I can't imagine I typed this whole paper without some typos. First, we need to state a few lemmas before we can tackle the main proof of this section.

Lemma 3.1. *Let $(a_t)_{t\geq 1}$, $(b_t)_{t\geq 1}$ be two non-negative real sequences. Assume $\sum_{t=1}^{\infty} a_t b_t$ 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 3.2. *Let $a_0 > 0$, $a_i \geq 0$, $i = 1, \cdots, T$, and $\beta > 1$. Then $\sum_{t=1}^{T} \frac{a_t}{(a_0 + \sum_{i=1}^{t} a_i)^{\beta}} \leq \frac{1}{(\beta-1)a_0^{\beta-1}}$.*

Lemma 3.3. *Assume (H1 , H3). Then, the iterates of SGD with stepsizes $\eta_t \in \mathbb{R}^{d\times d}$ satisfy the following inequality:*

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

This leads to the first theorem:

Theorem 3.4. *Assume (H1, H2, H3, H4). Let η_t be our global generalized AdaGrad stepsize from before, where $\alpha, \beta > 0$ and $\epsilon \in (0, 1/2]$. Then the gradients of SGD converge almost surely to zero. Moreover, $\liminf_{t\to\infty} ||\nabla f(\boldsymbol{x}_t)||^2 t^{1/2-\epsilon} = 0$ almost surely.*

Proof. We begin with the result of Lemma 3.3, noting that $\langle \nabla f(\mathbf{x}_t), \eta_t \nabla f(\mathbf{x}_t)\rangle = \eta_t ||\nabla f(\mathbf{x}_t)||^2$. Since these are nonnegative terms, we use monotone convergence to take the limit as $T \to \infty$ and commute this limit with the expectation:

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

We use Lemma 2 and telescoping series to show:

$$\sum_{t=1}^{\infty} ||\eta_t \mathbf{g}(\mathbf{x}_t, \xi_t)||^2 = \sum_{t=1}^{\infty} (\eta_t^2 - \eta_{t+1}^2 + \eta_{t+1}^2)||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2 = \sum_{t=1}^{\infty} \eta_{t+1}^2 ||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2 + \sum_{t=1}^{\infty} (\eta_t^2 - \eta_{t+1}^2)||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2$$

$$\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \max_{t\geq 1} ||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2 \sum_{t=1}^{\infty} (\eta_t^2 - \eta_{t+1}^2)$$

$$\leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \max_{t\geq 1} ||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2 \eta_1^2,$$

where we can ignore the last term of the telescoping series $\lim_{i\to\infty} \eta_{i+1}$ since it is being subtracted. From here, we use the inequality $||\mathbf{x} + \mathbf{y}||^2 \leq 2||\mathbf{x}||^2 + 2||\mathbf{y}||^2$ to show

$$\frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + \max_{t\geq 1} ||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2 \eta_1^2 \leq \frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\eta_1^2 \max_{t\geq 1} \big(||\nabla f(\mathbf{x}_t)||^2 + ||\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t, \xi_t)||^2\big).$$

Finally, we use H2, H4, and the fact that $\eta_t \leq \frac{\alpha}{\beta^{1/2+\epsilon}}$ from our definition to bound the above equation by

$$\frac{\alpha^2}{2\epsilon\beta^{2\epsilon}} + 2\frac{\alpha^2}{\beta^{1+2\epsilon}}(L^2 + S^2) < \infty.$$

Note that we're using the fact that $|f(\mathbf{x}) - f(\mathbf{y})|/||\mathbf{x} - \mathbf{y}|| \leq L$. To recap, we have just essentially shown $\mathbb{E}\Big[\sum_{t=1}^{\infty} \eta_t ||\nabla f(\mathbf{x}_t)||^2\Big] < \infty$ since $\mathbb{E}\Big[\sum_{t=1}^{\infty} \eta_t ||\nabla f(\mathbf{x}_t)||^2\Big] \leq f(\mathbf{x}_1) - f^* + \frac{M}{2}\mathbb{E}\Big[\sum_{t=1}^{\infty} ||\eta_t \mathbf{g}(\mathbf{x}_t, \xi_t)||^2\Big]$. In general, if X is a non-negative random variable and $\mathbb{E}[X] < \infty$, then $\mathbb{P}[X < \infty] = 1$, since otherwise would lead to a contradiction. Thus, we can say that with probability 1, $\sum_{t=1}^{\infty} \eta_t ||\nabla f(\mathbf{x}_t)||^2 < \infty$.

Recall our definition of η_t. We see:

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

$$\geq \sum_{t=1}^{\infty} \frac{\alpha}{(\beta + 2(t-1)(L^2 + S^2))^{1/2+\epsilon}} = \infty,$$

where again, we use $||\mathbf{g}(\mathbf{x}_i, \xi_i)||^2 = ||\mathbf{g}(\mathbf{x}_i, \xi_i) - \nabla f(\mathbf{x}_i) + \nabla f(\mathbf{x}_i)||^2 \leq 2||\mathbf{g}(\mathbf{x}_i, \xi_i) - \nabla f(\mathbf{x}_i)||^2 + ||\nabla f(\mathbf{x}_i)||^2 \leq 2(S^2 + L^2)$ by H2 and H4. (On a side note, I wish I could show my Calc 2 students how important series tests are!)

By using the fact that f is M-smooth and L-Lipschitz (and the triangle inequality), we get

$$||\nabla f(\mathbf{x}_{t+1})||^2 - ||\nabla f(\mathbf{x}_t)||^2 = (||\nabla f(\mathbf{x}_{t+1})|| + ||\nabla f(\mathbf{x}_t)||)(\big|||\nabla f(\mathbf{x}_{t+1})|| - ||\nabla f(\mathbf{x}_t)|||\big|)$$

$$\leq (||\nabla f(\mathbf{x}_{t+1})|| + ||\nabla f(\mathbf{x}_t)||)(||\nabla f(\mathbf{x}_{t+1})|| + ||\nabla f(\mathbf{x}_t)||)$$

$$\leq 2LM||x_{t+1} - x_t|| = 2LM||\eta_t \mathbf{g}(\mathbf{x}_t, \xi_t)|| \leq 2LM(L+S)\eta_t.$$

We now have all the conditions we need to apply Lemma 1 and obtain that $\lim_{t\to\infty} ||\nabla f(\mathbf{x}_t)||^2 = 0$. Note that all of this is happening with probability 1, giving us almost sure convergence.

As for the second part of our statement, we have a simple argument from here. Note that

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

by using the same argument from H2 and H4. Since $\sum_{t=1}^{\infty} \frac{1}{t} = \infty$, it must be that $\liminf_{t\to\infty} ||\nabla f(\mathbf{x}_t)||^2 t^{1/2-\epsilon} = 0$. Again, we had $\sum_{t=1}^{\infty} \eta_t ||\nabla f(\mathbf{x}_t)||^2 < \infty$ with probability 1.

□

This is one of the first results on the almost sure convergence of the gradients using generalized AdaGrad stepsizes with $\epsilon > 0$. Note we are looking at things asymptoticly, so the next section is about finite-time convergence rates in expectation.

4 Starting with the convex case:

In the section we show that adaptive stepsizes can lead to adaptice convergence rates (dependent on the noise of the stochastic gradient, σ).

Theorem 4.1. *Assume (H1, H3, H4') and f convex. Let η_t be our global generalized AdaGrad stepsize from before, where $\alpha, \beta > 0$ and $\epsilon \in [0, 1/2)$, and $4\alpha M < \beta^{1/2+\epsilon}$. Then the iterates of SGD satisfy the following bound:*

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

where

$$\overline{\boldsymbol{x}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$$

and

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

I noticed that in the proof of this theorem, the authors are not consistent with some factors of 2. Although not mentioned, this doesn't seem to be a big deal since M can absorb $2^{\frac{1/2+\epsilon}{1/2-\epsilon}}$ since $\frac{1/2+\epsilon}{1/2-\epsilon} \geq 1$ (but perhaps I am overlooking something here). Before proving the above theorem, the authors mention that Markov's inequality can be immediately applied. We write this explicitly below for the case when $\max\left(\gamma M^{1/2+\epsilon}, (\beta + T\sigma^2)^{1/4-\epsilon^2}\gamma^{1/2-\epsilon}\right) = \gamma M^{1/2+\epsilon}$:

$$\mathbb{P}\left(f(\overline{\mathbf{x}}) - f(\mathbf{x}^*) > \frac{1}{\delta^{\frac{1}{1/2-\epsilon}} T} M^{\frac{1/2+\epsilon}{1/2-\epsilon}} \gamma^{\frac{1}{1/2-\epsilon}}\right)$$

$$= \mathbb{P}\left((f(\overline{\mathbf{x}}) - f(\mathbf{x}^*))^{1/2-\epsilon} > (\frac{1}{\delta^{\frac{1}{1/2-\epsilon}} T} M^{\frac{1/2+\epsilon}{1/2-\epsilon}} \gamma^{\frac{1}{1/2-\epsilon}})^{1/2-\epsilon}\right)$$

$$\leq \frac{\mathbb{E}[(f(\overline{\mathbf{x}}) - f(\mathbf{x}^*))^{1/2-\epsilon}]}{\frac{1}{\delta T^{1/2-\epsilon}} M^{1/2+\epsilon}\gamma} \leq \frac{\frac{1}{T^{1/2-\epsilon}} M^{1/2+\epsilon}\gamma}{\frac{1}{\delta T^{1/2-\epsilon}} M^{1/2+\epsilon}\gamma}$$

$$= \delta$$

Thus, with probability $1 - \delta$,

$$f(\overline{\mathbf{x}}) - f(\mathbf{x}^*) \leq \frac{1}{\delta^{\frac{1}{1/2-\epsilon}} T} \max\left(M^{\frac{1/2+\epsilon}{1/2-\epsilon}} \gamma^{\frac{1}{1/2-\epsilon}}, (\beta + T\sigma^2)^{1/2+\epsilon}\gamma\right)$$

This means that up to polylog terms, if $\sigma = 0$, our convergence rate is $O(\frac{1}{T})$, which matches that of Gradient Descent, and otherwise we get worst-case optimal rate of SGD, $O(\frac{1}{\sqrt{T}})$.

Before proving the theorem, we need to state several lemmas:

Lemma 4.2. *Assume H1. Then* $||\nabla f(\boldsymbol{x}||^2 \leq 2M(f(\boldsymbol{x}) - \min_{\boldsymbol{y}} f(\boldsymbol{y}))$*, for all* $\boldsymbol{x}$*.*

Lemma 4.3. *If* $x \geq 0$ *and* $x \leq C(A + Bx)^{1/2+\epsilon}$*, then* $x < \max([C(2B)^{1/2+\epsilon}]^{\frac{1}{1/2-\epsilon}}, C(2A)^{1/2+\epsilon})$*.*

Lemma 4.4. *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 4.5. *If* $x, y \geq 0$ *and* $0 \leq p \leq 1$*, then* $(x + y)^p \leq x^p + y^p$*.*

Lemma 4.6. *Assume H1, H3, H4'. The stepsizes are chosen to be the global generalized AdaGrad, where* $\alpha, \beta, \epsilon \geq 0$*. Then*

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

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

Now that we have our lemmas, we can prove Theorem 4.1:

Proof. For simplicity, the let $\delta_t = f(\mathbf{x}_t) - f(\mathbf{x}^*)$ and $\Delta = \sum_{t=1}^{T} \delta_t$. We start with the following equation:

$$||\mathbf{x}_t - \mathbf{x}^*||^2 - 2\eta_t\langle \mathbf{g}(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}^*\rangle + \eta_t^2||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2 = ||(\mathbf{x}_t - \mathbf{x}^*) - \eta_t \mathbf{g}(\mathbf{x}_t, \xi_t)||^2$$

$$= ||\mathbf{x}_{t+1} - \mathbf{x}^*||^2,$$

and thus $||\mathbf{x}_{t+1} - \mathbf{x}^*||^2 - ||\mathbf{x}_t - \mathbf{x}^*||^2 = -2\eta_t\langle \mathbf{g}(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}^*\rangle + \eta_t^2||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2$. By taking the conditional probability with respect to the sigma algebra generated by $\xi_1, \cdots, \xi_{t-1}$, we see

$$\mathbb{E}[\langle \mathbf{g}(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}^*\rangle] = \langle f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle \geq \delta_t$$

where the inequality comes from convexity of f. Now by summing $t = 1$ to T over our expectation, we see

$$2\mathbb{E}\big[\sum_{t=1}^{T} \eta_t \delta_t\big] \leq \sum_{t=1}^{T}\left(||\mathbf{x}_t - \mathbf{x}^*||^2 - ||\mathbf{x}_{t+1} - \mathbf{x}^*||^2 + \mathbb{E}\big[\eta_t^2||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2\big]\right)$$

$$\implies \mathbb{E}\big[\sum_{t=1}^{T} \eta_t \delta_t\big] \leq \frac{1}{2}(||\mathbf{x}_1 - \mathbf{x}^*||^2 - ||x_{T+1} - x^*||^2) + \frac{1}{2}\mathbb{E}\big[\sum_{t=1}^{T} \eta_t^2||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2\big]$$

$$\leq \frac{1}{2}||\mathbf{x}_1 - \mathbf{x}^*||^2 + \frac{1}{2}\mathbb{E}\big[\sum_{t=1}^{T} \eta_t^2||\mathbf{g}(\mathbf{x}_t, \xi_t)||^2\big]\Big)$$

For $\epsilon > 0$, Lemma 4.2 and Lemma 4.6 give us

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

$$\leq K + \frac{4\alpha^2}{\beta^{1+2\epsilon}}(1+\ln T)\sigma^2 + \frac{4\alpha}{\beta^{1/2+\epsilon}}\mathbb{E}\Big[\sum_{t=1}^{T} \eta_t 2M(\min_{\mathbf{y}} f(\mathbf{y}))\Big].$$

Thus, after some algebra we get

$$\Big(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\Big)\mathbb{E}\big[\sum_{t=1}^{T} \eta_t \delta_t\big] \leq \frac{1}{2}||\mathbf{x}_1 - \mathbf{x}^*||^2 + \frac{\alpha^2}{4\epsilon\beta^{2\epsilon}} + \frac{2\alpha^2}{\beta^{1+2\epsilon}}(1+\ln T)\sigma^2.$$

Using similar logic in the $\epsilon = 0$ case, we get

$$\Big(1 - \frac{4\alpha M}{\beta^{1/2+\epsilon}}\Big)\mathbb{E}\big[\sum_{t=1}^{T} \eta_t \delta_t\big] \leq \frac{1}{2}||\mathbf{x}_1 - \mathbf{x}^*||^2 + \frac{2\alpha^2}{\beta}(1+\ln T)\sigma^2 + \alpha^2 \ln\Big(\sqrt{\beta + 2T\sigma^2} + 2\sqrt{M}\mathbb{E}[\sqrt{\Delta}]\Big)$$

The next step is to apply Holder's inequality to find a lower bound for the cases above. We will use the following version

$$\mathbb{E}[B^p] \geq \frac{\mathbb{E}[AB]^p}{\mathbb{E}[A^q]^{p/q}},$$

where $1/p + 1/q = 1$. We will let $1/p = 1/2 - \epsilon$, and $1/q = 1/2 + \epsilon$, $A = (\frac{1}{\eta_T})^{1/p}$, and $B = (\eta_T \Delta)^{1/p}$, a

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

since $\eta_t \geq \eta_t$ for $t \leq T$ because $\sum_{i=1}^{t-1} ||\mathbf{g}(\mathbf{x}_i, \xi_i)||^2 \leq \sum_{i=1}^{T-1} ||\mathbf{g}(\mathbf{x}_i, \xi_i)||^2$.
Staying on this note, recall from the definition of our stepsize,

$$\frac{1}{\eta_T} = \frac{1}{\alpha}\Big(\beta + \sum_{t=1}^{T-1} ||\mathbf{g}(\mathbf{x}_i, \xi_i)||^2\Big)^{1/2+\epsilon}$$

We use the inequality $||\mathbf{x} + \mathbf{y}||^2 \leq 2||\mathbf{x}||^2 + 2||\mathbf{y}||^2$ and Lemma 4.2 again to show

$$\frac{1}{\eta_T} = \frac{1}{\alpha}\Big(\beta + \sum_{t=1}^{T-1} ||\mathbf{g}(\mathbf{x}_i, \xi_i)||^2\Big)^{1/2+\epsilon}$$

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

$$\leq \frac{1}{\alpha}\Big(\beta + 2\sum_{t=1}^{T-1}(||\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t, \xi_t)||^2 + 2M\delta_t)\Big)^{1/2+\epsilon}.$$

We now define

$$\gamma = \frac{1}{\alpha(1 - \frac{4}{\beta^{1/2+\epsilon}})}\big(||\mathbf{x}^* - \mathbf{x}_1||^2 + \frac{4\alpha^2}{\beta^{1+2\epsilon}}(1+\ln T)\sigma^2\big) + K,$$

where K will be described below depending on the value of ϵ. We look at our two cases again. When $\epsilon > 0$,

$$\frac{1}{\gamma^{\frac{1/2-\epsilon}{1/2+\epsilon}}}\left(\mathbb{E}\big[\Delta^{1/2-\epsilon}\big]\right)^{\frac{1}{1/2+\epsilon}} \leq \alpha^{\frac{1/2-\epsilon}{1/2+\epsilon}}\mathbb{E}\Big[(\frac{1}{\eta_T})^{\frac{1/2-\epsilon}{1/2+\epsilon}}\Big]$$

$$\leq \mathbb{E}\Big[\big(\beta + 2\sum_{t=1}^{T-1}(||\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t,\xi_t)||^2 + 2M\delta_t)\big)^{1/2-\epsilon}\Big]$$

$$\leq \mathbb{E}\Big[\big(\beta + 2\sum_{t=1}^{T-1}(||\nabla f(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t,\xi_t)||^2\big)^{1/2-\epsilon}\Big] + \mathbb{E}\Big[\Big(4M\sum_{t=1}^{T-1}\delta\Big)^{1/2-\epsilon}\Big]$$

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

where the third inequality comes from Lemma 4.5, and the last is partly due to H4'. Here we are defining $K = \frac{\frac{\alpha^2}{2\epsilon\beta^{2\epsilon}}}{\alpha(1-\frac{4}{\beta^{1/2+\epsilon}})}$. Similarly, for the case when $\epsilon = 0$ we get

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

where $A = \sqrt{\beta + 2T\sigma^2}$, $B = 2\sqrt{M}$, and $D = \frac{\alpha}{1-\frac{4\alpha M}{\sqrt{\beta}}}$, and $C = \frac{\beta||\mathbf{x}_1 - \mathbf{x}^*||^2 + 4\alpha^2(1+\ln T)\sigma^2}{2\alpha\beta(1-\frac{4\alpha M}{\sqrt{\beta}})}$. Now we can use Lemma 4.4 to see

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

We use this as an upper bound for the logarithmic term, generalizing our result earlier for $\epsilon = 0$, this time with $K = D\ln(2A + 32B^4D^2 + 2B^2C + 8B^3D\sqrt{C}) = O(\frac{\ln T}{1-\frac{4\alpha M}{\sqrt{\beta}}})$

Finally, we use Lemma 4.3 so that for $\epsilon \geq 0$,

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

For our last step we apply Jenson's inequality to get our result:

$$T^{1/2-\epsilon}\mathbb{E}[(\Delta/T)^{1/2-\epsilon}] \leq \mathbb{E}[(T\Delta/T)^{1/2-\epsilon}] = \mathbb{E}[\Delta^{1/2-\epsilon}]$$

□

5 The non-convex case

Since this paper is getting long, I won't go through the proof of the last theorem, but will state it since it is important to one of the goals of the paper, which is to improve upon theory in the non-convex case.

Theorem 5.1. *Assume (H1, H3, H4'). Let η_t be our global generalized AdaGrad stepsize from before, where $\alpha, \beta > 0$ and $\epsilon \in (0, 1/2)$, and $4\alpha M < \beta^{1/2+\epsilon}$. Then the iterates of SGD satisfy the following bound:*

$$\mathbb{E}[\min_{1\leq t\leq T}||\nabla f(\boldsymbol{x}_t)||^{1-2\epsilon}] \leq$$

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

As you can probably imagine, you can apply the same Markov inequality steps to the conclusion of the above theorem. Note that here they prove convergence for the best iterate over T iterations. It is also important to note that this theorem proves that the generalized AdaGrad step sizes allow faster convergence to zero when the noise over the stochastic gradients is small.

6 Conclusion

This paper illustrates an advantage of the global generalized AdaGrad stepsizes over SGD. In addition, the paper makes progress in bridging the gap between theoretical understanding and practical success of adaptive stepsizes. It also claims to show for the first time sufficient conditions for convergence for non-convex functions when using adaptive stepsizes. They mention briefly that the limitations of this current analysis, one of them being that the high probability bounds depend polynomially on $1/\delta$ because of the use of Markov's Inequality.

Overall, I enjoyed working through this paper and felt like it was good experience for me. I also had a great time in my first APPM class, so thanks for the great semester!