Optimization for Deep Learning Final Project Report

B09902055 Weiping Li, B09902073 Chun-Neng Chu

1 Introduction

1.1 Background

Compared to full gradient methods, Stochastic Gradient Methods (SGD) efficiently handles large-scale datasets by randomly selecting subsets of training samples, and thus its computational and storage complexity is not heavily affected by the dataset size, while it maintains decent convergence properties. The derivation of the stepsize and the update procedure are crucial aspects to consider in the context of Stochastic Gradient Descent (SGD), as they have prominent effects on both the convergence guarantee and convergence rate.

1.2 Motivation

In this study, we surveyed two articles: ”Xiaoyu Li, Francesco Orabona. On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes” and ”Rachel Ward, Xiaoxia Wu, Léon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes”. They are variants of AdaGrad-Norm with similar convergence rates, yet have slight differences in their respective algorithms regarding the update sequence and the constraints required for their proofs to hold. Given these properties, by studying the two aforementioned papers, we can gain insight on how the update sequence and constraints contribute to the convergence guarantee and convergence rate.

1.3 Contribution

In this study, we strive to complete the following:

1. Unify the notation of the two articles and extract a five-step procedure for proving the convergence rate from the shared techniques.
2. Find the reason for the different update sequence between the two algorithms.
3. Summarize the reasons that Xiaoyu Li et al. require prior knowledge of specific smoothness constant in the proof.
4. Summarize the reasons that Rachel Ward et al. require Lipschitz constraint in proof.
5. Analyze the different usages of the constant 'a' from the five-step proof procedure.
6. Examine if we can acquire the same convergence rate with Rachel Ward et al.'s algorithm, but using the constraints of Xiaoyu Li et al.. Vice versa.

2 Notation & Preliminaries

2.1 Notation

1. $f : \mathbb{R}^d \to \mathbb{R}$ is the objective function to minimize.
2. $G : \mathbb{R}^d \times \xi \to \mathbb{R}^d$ is the unbiased stochastic gradient of f.
3. $x_t \in \mathbb{R}^d$ is the input of f (model weights) after t updates.
4. η_t is the stepsize of the t^{th} update.
5. $f_t = f(x_t), G_t = G(x_t, \xi_t), \Delta = \sum_{t=1}^{T} \|\nabla f_{t-1}\|^2$

2.2 Preliminaries

- L-Lipschitz:

  f is L-Lipschitz iff

$$\forall x, y \in dom\ f,\ ||f(y) - f(x)|| \leq L||y - x||$$

- M-smooth:

  f is M-smooth iff f is differentiable and satisfied

$$\forall x, y \in dom\ f, f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}||x - y||_2^2$$

- Convex:

  f is convex iff $dom\ f$ is convex. and Jensen's inequality is satisfied, i.e.

$$\forall x, y \in dom\ f,\ \alpha, \beta \geq 0,\ s.t\ \alpha + \beta = 1,$$
$$\alpha f(x) + \beta f(y) \geq f(\alpha x + \beta y)$$

- Concave:

  f is concave iff $-f$ is convex

- Stochastic gradient noise:

$$E_{\xi_t}[||G(x_t, \xi_t) - \nabla f(x_t)||^2]$$

- Best-iterate convergence rate with probability $1 - \delta$:

$$P\left(min_{0 \leq t \leq T}||\nabla f(x_t)||^2 \leq H(\delta, T, D^*) = O\left(\frac{1}{T^\alpha}\right)\right) \geq 1 - \delta$$

  where H is a function that takes δ, T and data related hyperparameters D^* as input.

3 Paper Comparisons and Observations

3.1 Difference between Algorithms

When comparing the algorithms used in the works of Xiaoyu Li et al. and the works of Rachel Ward et al., we notice that they are identical, other than the sequence in which the weights and learning rate are updated. Xiaoyu Li updates the weights before the learning rate, while Rachel Ward does the opposite.

Algorithm 1 ADAGRAD-Norm (Xiaoyu Li et al.)

Input: Initialize $x_0 \in R^d, b_0 > 0, \eta > 0$
for $t = 1, 2, \ldots$ **do**
    Generate $\xi_{t-1}, G_{t-1} = G(x_{t-1}, \xi_{t-1})$
    $\boldsymbol{x_t \leftarrow x_{t-1} - \frac{\eta}{b_{t-1}} G_{t-1}}$
    $\boldsymbol{b_t^2 \leftarrow b_{t-1}^2 + ||G_{t-1}||^2}$
end for

Algorithm 2 ADAGRAD-Norm (Rachel Ward et al.)

Input: Initialize $x_0 \in R^d, b_0 > 0, \eta > 0$
for $t = 1, 2, \ldots$ **do**
    Generate $\xi_{t-1}, G_{t-1} = G(x_{t-1}, \xi_{t-1})$
    $\boldsymbol{b_t^2 \leftarrow b_{t-1}^2 + ||G_{t-1}||^2}$
    $\boldsymbol{x_t \leftarrow x_{t-1} - \frac{\eta}{b_t} G_{t-1}}$
end for

3.2 Different Constraints Required

The two algorithms, with their individual constraints, can be proven to have the same complexity for convergence with respect to iteration T: $O(\frac{1}{\sqrt{T}})$ However, their constraints differ.

Constraints		
Constraints↓ Algorithm →	Xiaoyu Li et al.	Rachel Ward et al.
M-smooth	✓	✓
$E\left[\left\|\nabla f\left(x_t\right) - G\left(x_t, \xi_t\right)\right\|^2\right] \leq \sigma^2$	✓	✓
$E_\xi[G(x, \xi)] = \nabla f(x)$	✓	✓
$f > -\infty$	✓	✓
know smoothness constant M	✓	
L-Lipschitz		✓

Xiaoyu Li requires concrete knowledge on the value of smoothness constant M, while Rachel Ward requires the L-Lipschitz constraint, but does not need to know the value of L.

3.3 Standard SGD Proof Format

Intuitively speaking, when optimizing, the goal is to find a complexity bound for $||\nabla f(x_t)||^2$ in the form of $O(\frac{1}{T^\alpha})$, where a and α are chosen constants. By observing the proofs in the work of both Xiaoyu Li and Rachel Ward, we extract the overlapping techniques and thought processes into a five-step procedure that we call the Standard SGD Proof Format, which is as follows:

① $P(min_{0\leq t\leq T}||\nabla f(x_t)||^2 = O(\frac{1}{T^\alpha})) \geq 1-\delta$

② $E\left[min_{0\leq t\leq T}||\nabla f(x_t)||^{\frac{2a-2}{a}}\right]^{\frac{a}{a-1}} = O(\frac{1}{T^\alpha})$

③ $E\left[(\sum_{t=1}^{T}||\nabla f(x_{t-1})||^2)^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}} = O(\frac{1}{T^{\alpha-1}})$

④ $\sum_{t=1}^{T} E\left[\eta_t^* \left\|\nabla f\left(x_{t-1}\right)\right\|^2\right] \leq f\left(x_0\right) - f^* + \sum_{t=1}^{T} E\left[\frac{\eta_t^2||G(x_{t-1},\xi_{t-1})||^2 M}{2}\right]$

⑤ $|f(x_t) - f(x_{t-1}) - \langle\nabla f(x_{t-1}), x_t - x_{t-1}\rangle| \leq \frac{M}{2}\|x_t - x_{t-1}\|^2$

① $\xleftarrow{\text{Markov's}}$ ② $\leftarrow$ ③ $\xleftarrow{\text{Trick}}$ ④ $\xleftarrow{\text{Trick}}$ ⑤

Note that η_t^* may be η_t or the estimation of η_t, and f^* is the optimal target function value.

The increment of numbers (①) represents the thought process of proof, while the arrows are logical or arithmetic implications. In other words, ① is the target inequality that guarantees convergence with a probability greater than $1-\delta$, while ⑤ is from the definition of M-smooth. The technicalities of the logical implications from ⑤ to ① will be fleshed out in the sections below.

4 Algorithm - Constraint Relation: Xiaoyu Li et al.

4.1 Necessity of Changing Update Sequence - Lemma 3

In this section, we discuss why it is necessary for Xiaoyu Li et al.'s proof that the weights are updated before the learning rate in each iteration.

Lemma 3 in Xiaoyu Li et al.'s work states the following: Assume f is M-smooth and $E[G(x_{t-1}, \xi_{t-1})] = \nabla f(x_{t-1})$. Then, the iterates of SGD with stepsizes $\eta_t \in R$ satisfy the below inequality.

$$E\left[\sum_{t=1}^{T}\left\langle\nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right)\right\rangle\right] \leq f\left(x_{t-1}\right) - f^* + \frac{M}{2} E\left[\sum_{t=1}^{T}\left\|\eta_t G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right]$$

In the proof of Lemma 3, there is an intermediary step that requires the following:

$$E_{\xi_{t-1}}\left[\left\langle \nabla f\left(x_{t-1}\right), \eta_t\left(\nabla f\left(x_{t-1}\right) - G\left(x_{t-1}, \xi_{t-1}\right)\right)\right\rangle\right]$$
$$= \left\langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right) - \eta_t E_{\xi_{t-1}}\left[G\left(x_{t-1}, \xi_{t-1}\right)\right]\right\rangle = 0$$

The above equation requires that η_t is independent to ξ_{t-1}. The two terms are independent due to the fact that at the t^{th} iteration, η_t is decided by ξ_0 to ξ_{t-2}. Hence, η_t can be taken out of the expectation.

4.2 Necessity of Smoothness Constant - Lemma 3 & 8

In this section, we discuss why specific knowledge on the value of smoothness constant M for the M-smooth function to be optimized is necessary.

From Lemma 8 in Xiaoyu Li et al., we have:

$$E\left[\sum_{t=1}^{T} \eta_t^2 \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right]$$
$$= E\left[\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{\eta_{t+1}^2} \left\|\boldsymbol{G\left(x_{t-1}, \xi_{t-1}\right)}\right\|^{\boldsymbol{2}} + \sum_{t=1}^{T} \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2 \left(\eta_t^2 - \eta_{t+1}^2\right)\right]$$
$$\leq \boldsymbol{2\eta^2 \ln\left(\sqrt{b_0^2 + 2T\sigma^2} + \sqrt{2}E\left[\sqrt{\sum_{t=1}^{T} \left\|\nabla f\left(x_{t-1}\right)\right\|^2}\right]\right) - 2\eta^2 ln(b_0)}$$
$$+ \tfrac{4\eta^2}{b_0^2}(1 + \ln T)\sigma^2 + \tfrac{4\eta}{b_0} E\left[\sum_{t=1}^{T} \eta_t \left\|\nabla f\left(x_{t-1}\right)\right\|^2\right]$$

With the tricks used by Xiaoyu Li et al., demonstrated in the next subsection, we can claim that the term $\boldsymbol{\sqrt{2}E\left[\sqrt{\sum_{t=1}^{T} \left\|\nabla f\left(x_{t-1}\right)\right\|^2}\right]}$ can be dropped, and $\boldsymbol{\sum_{t=1}^{T} \eta_{t+1}^2 \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2}$ is bounded by $O(ln(\sqrt{T}))$. Intuitively speaking, the term $\sum_{t=1}^{T} \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2 \left(\eta_t^2 - \eta_{t+1}^2\right)$ is the penalty caused by "borrowing" the bold term, which is from the next iteration, to bound the stochastic gradient norm squared. With some tricks, the penalty can be bounded by $\frac{4\eta^2}{b_0^2}(1+\ln T)\sigma^2 + \frac{4\eta}{b_0} E\left[\sum_{t=1}^{T} \eta_t \left\|\nabla f\left(x_{t-1}\right)\right\|^2\right]$. When Lemma 3 is applied in an attempt to bound the term $E\left[\sum_{t=1}^{T} \eta_t \left\|\nabla f\left(x_{t-1}\right)\right\|^2\right]$, scaling the same term on the RHS of Lemma 8 by $\frac{M}{2}$, we find the below inequality:

$$\left(1 - \tfrac{2\eta M}{\sqrt{b_0^2}}\right) E\left[\sum_{t=1}^{T} \eta_t \left\|\nabla f\left(x_{t-1}\right)\right\|^2\right] \leq f\left(x_0\right) - f^\star$$
$$+ M\left(\eta^2 \ln\left(\sqrt{b_0^2 + 2T\sigma^2} + \sqrt{2}E\left[\sqrt{\sum_{t=1}^{T} \left\|\nabla f\left(x_{t-1}\right)\right\|^2}\right]\right) - \tfrac{\eta^2 ln(b_0^2)}{2}\right)$$
$$+ \tfrac{2\eta M}{b_0^2}(1 + \ln T)\sigma^2.$$

Consider the case where $\left(1-\frac{2\eta M}{\sqrt{b_0^2}}\right) \leq 0$, we can see that the inequality always holds, and thus we gain no information of the term $E\left[\sum_{t=1}^{T}\eta_t \left\|\nabla f\left(x_{t-1}\right)\right\|^2\right]$. Hence we need to know the constant M to initialize η and b_0.

4.3 Bounding $E\left[\sqrt{\Delta}\right]$ with $a=2$

Here, we show how choosing the value $a=2$ works as a trick to help bound $E\left[\sqrt{\Delta}\right]$, where $\Delta := \sum_{t=1}^{T}\left\|\nabla f_{t-1}\right\|^2$ from ③ in the Standard SGD Proof Format for the proof of Xiaoyu Li et al..

From Lemma 3 and 8 in Xiaoyu Li et al.'s work, have:

$E\left[\sum_{t=1}^{T}\eta_t \left\|\nabla f_{t-1}\right\|^2\right] = O\left(\ln\left(\sqrt{T}+E\left[\sqrt{\Delta}\right]\right)\right).$

In order to bound $E\left[\sqrt{\Delta}\right]$, we utilize Holder to find the following inequality

$$E\left[\left(\left(\frac{1}{\eta_T}\right)^{\frac{a-1}{a}}\right)^a\right]^{\frac{1}{a-1}} E\left[\sum_{t=1}^{T}\eta_t \left\|\nabla f_{t-1}\right\|^2\right] \geq E\left[\Delta^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}},$$

where

$$E\left[\frac{1}{\eta_T}\right] \leq E\left[\frac{1}{\eta}\left(b_0^2+\sum_{t=1}^{T-1}\left(\left\|\boldsymbol{G}_{t-1}-\nabla f_{t-1}\right\|^2+\left|\nabla f_{t-1}\right|^2\right)\right)^{1/2}\right] = O\left(\sqrt{T}+E\left[\sqrt{\Delta}\right]\right)$$

Intuitively, when $a=2$, we have

$$E\left[\sqrt{\Delta}\right]^2 = O\left(\ln\left(\sqrt{T}+E\left[\sqrt{\Delta}\right]\right)\right) O\left(\sqrt{T}+E\left[\sqrt{\Delta}\right]\right)$$

We can then consider two cases:

1. $E\left[\sqrt{\Delta}\right] = \omega\left(\sqrt{T}\right)$: $E\left[\sqrt{\Delta}\right]^2 = O\left(E\left[\sqrt{\Delta}\right]\ln\left(E\left[\sqrt{\Delta}\right]\right)\right)$

   Which holds only if $E\left[\sqrt{\Delta}\right] = O\left(1\right) \rightarrow$ contradict.

2. Otherwise : $E\left[\sqrt{\Delta}\right]^2 = O\left(\sqrt{T}\ln\left(\sqrt{T}\right)\right) = \tilde{O}\left(\sqrt{T}\right)$

   $\rightarrow E\left[\sqrt{\Delta}\right]^2 = O\left(\sqrt{T}\right)$ By dropping logarithmic term.

Which is the goal case for $\alpha = \frac{1}{2}$, $a = 2$ in ③ of the standard SGD convergence proof format:

$$E\left[\left(\sum_{t=1}^{T}||\nabla f(x_t)||^2\right)^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}} = O(\frac{1}{T^{\alpha-1}})$$

4.4 Comparison with ADAGRAD-Norm (Rachel Ward et al.)

Finally, in order to further demonstrate the effects of the above constraints in the proof, we attempt to apply the proof techniques and constraints in Xiaoyu Li et al.'s work to the algorithm in the work of Rachel Ward et al..

First, when attempting to take η_t out of the expectation in the intermediary step of Lemma 3 that requires $E_{\xi_{t-1}}\left[\left\langle \nabla f\left(x_{t-1}\right), \eta_t\left(\nabla f\left(x_{t-1}\right) - G\left(x_{t-1}, \xi_{t-1}\right)\right)\right\rangle\right]$ $= \left\langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right) - \eta_t E_t\left[G\left(x_{t-1}, \xi_{t-1}\right)\right]\right\rangle = 0$, we find that this is infeasible. This is becauses for the t^{th} iteration, Rachel Ward et al. updates the learning rate (η_t) before the weights, leading to η_t not being independent of ξ_{t-1}.

Second, in Rachel Ward et al.'s article, η_t^* is an estimation of η_t instead of being exactly η_t. Since in the fragment $E\left[\sum_{t=1}^{T}\eta_t^*\left\|\nabla f\left(\boldsymbol{x}_{t-1}\right)\right\|^2\right] \geq E\left[\eta_T^*\Delta\right]$ of the main proof of convergence rate in Xiaoyu Li et al. requires descending η_t^* and it is hard to confirm whether the estimation is indeed descending, the proof cannot be generalized to Rachel Ward et al. trivially.

4.5 Convergence Rate Proof Sketch

We summarize Xiaoyu's work by mapping the main components of their proof to those five steps in Standard SGD Proof Format:

⑤ $|f(x_t) - f(x_{t-1}) - \langle \nabla f(x_{t-1}), x_t - x_{t-1}\rangle| \leq \frac{M}{2}\|x_t - x_{t-1}\|^2$

④ $E\left[\sum_{t=1}^{T}\eta_t\left\|\nabla f_{t-1}\right\|^2\right] \leq f_0 - f^* + \frac{M}{2}E\left[\sum_{t=1}^{T}\left\|\eta_t G_{t-1}\right\|^2\right]$

③ $E\left[\left(\sum_{t=1}^{T}||\nabla f(x_{t-1})||^2\right)^{\frac{2-1}{2}}\right]^{\frac{2}{2-1}} = O(\frac{1}{T^{\frac{1}{2}-1}})$

② $E[min_{0\leq t\leq T}||\nabla f(x_t)||^{\frac{2\cdot 2-2}{2}}]^{\frac{2}{2-1}} = O(\frac{1}{T^{\frac{1}{2}}})$

① $P([min_{0\leq t\leq T}||\nabla f(x_t)||^2 = O(\frac{1}{T^{\frac{1}{2}}})) \geq 1-\delta$

① $\xleftarrow{\text{Markov's}}$ ② $\xleftarrow{\text{Trivial}}$ ③ $\xleftarrow{\text{Knowledge of smoothness constant}}$ ④ $\xleftarrow{\text{Change of update sequence}}$ ⑤(Assumptions)

5 Algorithm - Constraint Relation: Rachel Ward et al.

5.1 On the Inner Product Term of ⑤ in SGD Proof Format

From the L-smooth definition and our goal of bounding the expectation of gradient norm squared, we want to move f related terms in the inner product to the LHS of the below inequality.

$$f_{t+1} - f_t \leq -\boldsymbol{\eta}\left\langle \boldsymbol{\nabla f_t}, \frac{\boldsymbol{G_t}}{\boldsymbol{b_{t+1}}} \right\rangle + \frac{\eta^2 M}{2b_{t+1}^2}\left\|G_t\right\|^2$$

The result is as follows, where the bold terms are from the inner product term.

$$\boldsymbol{E}\left[\frac{\boldsymbol{\eta}\left\|\boldsymbol{\nabla f_t}\right\|^{\boldsymbol{2}}}{\boldsymbol{2}\sqrt{\boldsymbol{b_t^2}+\left\|\boldsymbol{\nabla f_t}\right\|^{\boldsymbol{2}}+\boldsymbol{\sigma^2}}}\right] \leq E\left[f_t\right] - E\left[f_{t+1}\right] + \frac{\boldsymbol{4\sigma\eta}}{\boldsymbol{2}}\boldsymbol{E}\left[\frac{\left\|\boldsymbol{G_t}\right\|^{\boldsymbol{2}}}{\boldsymbol{b_{t+1}^2}}\right] + \frac{M\eta^2}{2}E\left[\frac{\left\|G_t\right\|^2}{b_{t+1}^2}\right]$$

For future references, we define $\eta_t^* = \frac{\eta}{2\sqrt{b_{t-1}^2+||\nabla f_{t-1}||^2+\sigma^2}}$.

5.2 The Necessity of L-Lipschitz Constraint

In Xiaoyu Li et al.'s work, there is a step where Holder is used to bound $\left(E\left[\Delta^{1/2}\right]\right)^2$, and (indirectly) bound η_t, which is the counterpart of the LHS on the last page in Xiaoyu Li's proof. Here, $\Delta := \sum_{t=1}^{T}\left\|\nabla f_{t-1}\right\|^2$ and $a = 2$.

$$E\left[\sum_{t=1}^{T}\eta_t\left\|\nabla f_{t-1}\right\|^2\right] \geq E\left[\eta_T\Delta\right] = E\left[\left(\left(\eta_T\Delta\right)^{\frac{a-1}{a}}\right)^{\frac{a}{a-1}}\right] \geq \frac{E\left[\Delta^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}}}{E\left[\left(\left(\frac{1}{\eta_T}\right)^{\frac{a-1}{a}}\right)^a\right]^{\frac{1}{a-1}}}$$

However, directly replacing η_t with η_t^* does not work, as the first inequality requires η_t to be decreasing, but η_t^* holds no such guarantees. Hence, Rachel Ward et al. introduce the L-Lipschitz constraint in order to bound the stochastic gradient related terms from the previous subsection.

5.3 The Relation between a and δ

$$E\left[\frac{\|\nabla f_t\|^2}{2\sqrt{b_t^2+\|\nabla f_t\|^2+\sigma^2}}\right] \geq \frac{E\left[\left(\|\nabla f_t\|^2\right)^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}}}{2\left(E\left[\sqrt{b_t^2+\|\nabla f_t\|^2+\sigma^2}^{a-1}\right]\right)^{\frac{1}{a-1}}} \geq \frac{\left(E\|\nabla f_t\|^{\frac{2a-2}{a}}\right)^{\frac{a}{a-1}}}{2\sqrt{E\left[b_t^2+\|\nabla f_t\|^2+\sigma^2\right]}}$$

In the fragment of their proof, moving the expectation into the squared root term requires $a \leq 3$,

as otherwise $\sqrt{b_t^2 + \|\nabla f_t\|^2 + \sigma^2}^{a-1}$ would not be concave.

On the other hand, through some omitted calculations via similar techniques to Lemma 3 and Lemma 8 from Xiaoyu Li et al. Have: $P\left(\min_{t\in[T]} \|\nabla f_t\|^2 \geq \frac{C_T}{\delta^{\frac{a}{a-1}}}\right) \leq \delta$, where C_T is a term unrelated to a but related to T. Thus, larger a results in tighter bound.

From the above, we can see why $a = 3$ is optimal.

5.4 Comparison with ADAGRAD-Norm (Xiaoyu Li et al.)

Without knowledge of M coefficient (for M-smooth), with a large enough coefficient of (2), the combined results of Lemma 3 and 8 in Xiaoyu Li et al.'s work provide no useful bound for $E\left[\sum_{t=1}^{T} \eta_t \|\nabla f_{t-1}\|^2\right]$.

(1) are bounded by the same techniques as Rachel Ward et al., but (2) are not dealt with.

Note: L-Lipschitz provides an upper bound, γ, for $\|\nabla f_t\|$.

$$E\left[\sum_{t=1}^{T} \eta_t \|\nabla f(x_{t-1})\|^2\right] \leq O\left(\ln\left(\sqrt{b_0^2 + 2T\sigma^2} + \sqrt{2}E\left[\sqrt{T\cdot\gamma^2}\right]\right)\right) \text{... (1)}$$

$$+ \frac{2\eta M}{\sqrt{b_0^2}} E\left[\sum_{t=1}^{T} \eta_t \|\nabla f(x_{t-1})\|^2\right] \text{... (2)}$$

For $\frac{2\eta M}{\sqrt{b_0^2}} > 1$, the inequality always holds, so $E\left[\sum_{t=1}^{T} \eta_t \|\nabla f(x_{t-1})\|^2\right]$ is not bounded, and adding the Lipschitz constraint is of no use.

5.5 Convergence Rate Proof Sketch

We summarize Rachel Ward's work by mapping the main components of their proof to those five steps in Standard SGD Proof Format:

⑤ $|f(x_t) - f(x_{t-1}) - \langle \nabla f(x_{t-1}), x_t - x_{t-1} \rangle| \leq \frac{M}{2} \|x_t - x_{t-1}\|^2$

④ $\sum_{t=1}^{T} E\left[\frac{\eta \|\nabla f_{t-1}\|^2}{2\sqrt{b_{t-1}^2 + \|\nabla f_{t-1}\|^2 + \sigma^2}}\right] \leq f_0 - f^* + \sum_{t=1}^{T}\left(\frac{4\sigma\eta + \eta^2 M}{2} E\left[\frac{\|G_{t-1}\|^2}{b_t^2}\right]\right)$

③ $E\left[\left(\sum_{t=1}^{T} ||\nabla f(x_{t-1})||^2\right)^{\frac{3-1}{3}}\right]^{\frac{3}{3-1}} \leq E\left[\left(\sum_{t=1}^{T} \left(||\nabla f(x_{t-1})||^2\right)^{\frac{3-1}{3}}\right)\right]^{\frac{3}{3-1}} = O(\frac{1}{T^{\frac{1}{2}-1}})$

② $E[min_{0\leq t\leq T}||\nabla f(x_t)||^{\frac{2\cdot 3-2}{3}}]^{\frac{3}{3-1}} = O(\frac{1}{T^{\frac{1}{2}}})$

① $P([min_{0\leq t\leq T}||\nabla f(x_{t-1})||^2 = O(\frac{1}{T^{\frac{1}{2}}})) \geq 1 - \delta$

① $\xleftarrow{\text{Markov's}}$ ② $\xleftarrow{\text{Trivial}}$ ③ $\xleftarrow{\text{L-Lipschitz}}$ ④ $\xleftarrow{\text{Unbiased and bounded variance of stochastic part}}$ ⑤(Assumptions)

6 Conclusion

To summarize, we analyzed two articles that propose variants of AdaGrad-Norm. We unified the notation of the articles, extracted a five-step procedure for proving the convergence rate, identified the reasons for different update sequences, summarized the usage of constraints and the constant 'a' in the proofs, and pointed out the bottleneck of exchanging their constraints to derive a similar convergence rate.

7 Future Work and Potential Extensions

In main proof of the convergence rate of Xiaoyu Li's work, we claimed that it is non-trivial to prove the Holder-related inequality

$$E\left[\sum_{t=1}^{T} \eta_t^* \|\nabla f_{t-1}\|^2\right] \geq \frac{E\left[\Delta^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}}}{E\left[\left(\left(\frac{1}{\eta_T}\right)^{\frac{a-1}{a}}\right)^a\right]^{\frac{1}{a-1}}}$$

for cases where η_t^* is not decreasing. However, we have no concrete proof that it cannot hold for such cases. Hence, we can do one of the following in the future.

1. Prove that the inequality does not hold when η_t^* is not decreasing.
2. Drop the Lipschitz constraint for Rachel Ward et al..

8 References

1. Quoc Tran-Dinh. Sublinear Convergence Rates of Extragradient-Type Methods: A Survey on Classical and Recent Developments.

2. Xiaoyu Li, Francesco Orabona. On the Convergence of Stochastic Gradient Descent. with Adaptive Stepsizes

3. Rachel Ward, Xiaoxia Wu, Léon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes.

4. Yen-Huan Li. Optimization algorithms Lecture 3.

5. Chih-Jen Lin. Optimization Methods for Deep Learning: Convergence of stochastic gradient methods.

9 Appendix

9.1 On Lemma 3 in "Xiaoyu Li, Francesco Orabona. On the Convergence of Stochastic Gradient Descent. with Adaptive Stepsizes"

9.1.1 Description of Lemma 3

Assume f is M-smooth and $E[G(x_{t-1}, \xi_{t-1})] = \nabla f(x_{t-1})$. Then, the iterates of SGD with stepsizes $\eta_t \in R^d$ satisfy the following inequality.

$$E\left[\sum_{t=1}^{T} \left\langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right)\right\rangle\right] \leq f\left(x_{t-1}\right) - f^* + \frac{M}{2} E\left[\sum_{t=1}^{T} \left\|\eta_t G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right]$$

9.1.2 Proof

From the definition of M-smooth, we have $|f(y)-f(x)-\langle \nabla f(x), y-x\rangle| \leq \frac{M}{2}\|y-x\|^2$ Thus,

$$\begin{aligned} f\left(x_t\right) &\leq f\left(x_{t-1}\right) + \left\langle \nabla f\left(x_{t-1}\right), x_t - x_{t-1}\right\rangle + \frac{M}{2}\left\|x_t - x_{t-1}\right\|^2 \\ &= f\left(x_{t-1}\right) + \left\langle \nabla f\left(x_{t-1}\right), \eta_t\left(\nabla f\left(x_{t-1}\right) - G\left(x_{t-1}, \xi_{t-1}\right)\right)\right\rangle \\ &- \left\langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right)\right\rangle + \frac{M}{2}\left\|\eta_t G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2. \end{aligned}$$

Taking the conditional expectation with respect to $\xi_0, \cdots, \xi_{t-2}$, we have

$$\begin{aligned} & E_{\xi_{t-1}}\left[\left\langle \nabla f\left(x_{t-1}\right), \eta_t\left(\nabla f\left(x_{t-1}\right)-G\left(x_{t-1}, \xi_{t-1}\right)\right)\right\rangle\right] \\ = & \left\langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right)-\eta_t E_t\left[G\left(x_{t-1}, \xi_{t-1}\right)\right]\right\rangle=0 . \end{aligned}$$

Hence, from the law of total expectation, we have

$$E\left[\left\langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right)\right\rangle\right] \leq E\left[f\left(x_{t-1}\right)-f\left(x_t\right)+\tfrac{M}{2}\left\|\eta_t g\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right] .$$

Summing over $t=1$ to T and lower bounding $f\left(x_T\right)$ with $f^{\star}$, we have the stated bound.

9.2 On Lemma 8 in "Xiaoyu Li, Francesco Orabona. On the Convergence of Stochastic Gradient Descent. with Adaptive Stepsizes"

9.2.1 Description of Lemma 8

Assume f is M-smooth, $E[G(x_{t-1}, \xi_{t-1})] = \nabla f(x_{t-1})$ and the stochastic gradient satisfies

$E\left[\exp\left(\|\nabla f(x) - G(x, \xi)\|^2/\sigma^2\right)\right] \leq \exp(1), \forall x$ (slightly stronger constraint).

Then,

$$E\left[\sum_{t=1}^{T} \eta_t^2\left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right] \leq K+\frac{4\eta^2}{b_0^2}(1+\ln T)\sigma^2+\frac{4\eta}{b_0} E\left[\sum_{t=1}^{T} \eta_t\left\|\nabla f\left(x_{t-1}\right)\right\|^2\right] ,$$

where $K=2\eta^2 \ln\left(\sqrt{b_0^2+2T\sigma^2}+\sqrt{2} E\left[\sqrt{\sum_{t=1}^{T}\left\|\nabla f\left(x_{t-1}\right)\right\|^2}\right]\right)-2\eta^2 ln(b_0)$

9.2.2 Proof

Lemma 10 states: If $x > 0, \eta > 0$, then $\ln(\frac{1}{x}) \geq \eta\left(1 - x^{\frac{1}{\eta}}\right)$.

$$E\left[\sum_{t=1}^{T} \eta_{t+1}^2 \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right] = E\left[\sum_{t=1}^{T} \frac{\eta^2 \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2}{\left(b_0^2 + \sum_{i=1}^{t} \left\|G\left(x_{i-1}, \xi_{i-1}\right)\right\|^2\right)}\right]$$

$$\leq 2\eta^2 E\left[\ln\left(\sqrt{b_0^2 + \sum_{t=1}^{T} \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2}\right)\right] - \eta^2 \ln\left(b_0^2\right)$$

$$\leq 2\eta^2 \ln\left(\sqrt{b_0^2 + 2T\sigma^2} + \sqrt{2}E\left[\sqrt{\sum_{t=1}^{T} \left\|\nabla f\left(x_{t-1}\right)\right\|^2}\right]\right) - \eta^2 \ln\left(b_0^2\right)$$

Where in the first inequality we used Lemma 10 and in the third we used Jensen's inequality. Putting things together, we have

$$E\left[\sum_{t=1}^{T} \eta_t^2 \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right]$$

$$= E\left[\sum_{t=1}^{T} \eta_{t+1}^2 \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2 + \sum_{t=1}^{T} \left\|G\left(x_{t-1}, \xi_{t-1}\right)\right\|^2 \left(\eta_t^2 - \eta_{t+1}^2\right)\right]$$

$$\leq 2\eta^2 \ln\left(\sqrt{b_0^2 + 2T\sigma^2} + \sqrt{2}E\left[\sqrt{\sum_{t=1}^{T} \left\|\nabla f\left(x_{t-1}\right)\right\|^2}\right]\right) - \eta^2 \ln\left(b_0^2\right)$$

$$+ \frac{4\eta^2}{b_0^2}(1 + \ln T)\sigma^2 + \frac{4\eta}{b_0^{2\frac{1}{2}}} E\left[\sum_{t=1}^{T} \eta_t \left\|\nabla f\left(x_{t-1}\right)\right\|^2\right]$$

9.3 On the Use of Markov's Inequality

Markov's inequality states that $P(X \geq y) \leq \frac{E[X]}{y}$,if X is a non-negative random variable and $y > 0$. We use this inequality to prove the equivalence of the expectation form and probability form of convergence bound in our studies.

$\rightarrow P(min_{0 \leq t \leq T} ||\nabla f(x_t)||^{\frac{2a-2}{a}} \geq \frac{E[(min_{0 \leq t \leq T} ||\nabla f(x_t)||^{\frac{2a-2}{a}}]}{\delta}) \leq \delta$

$$\rightarrow P(min_{0\leq t\leq T}||\nabla f(x_t)||^2 \leq \frac{E\Big[(min_{0\leq t\leq T}||\nabla f(x_t)||^{\frac{2a-2}{a}}\Big]^{\frac{a}{a-1}}}{\delta^{\frac{a}{a-1}}}) \geq 1-\delta$$

$$\rightarrow H(\delta, T, D^*) = \frac{E\Big[(min_{0\leq t\leq T}||\nabla f(x_t)||^{\frac{2a-2}{a}}\Big]^{\frac{a}{a-1}}}{\delta^{\frac{a}{a-1}}}$$

9.4 On the Use of Holder's Inequality

Holder's inequality states that, $E\left[X^p\right]^{\frac{1}{p}} \cdot E\left[Y^q\right]^{\frac{1}{q}} \geq E\left[XY\right]$,given $\frac{1}{p}+\frac{1}{q}=1 \wedge p, q, X, Y > 0$.

In our studies, by choosing $p=\frac{a}{a-1}$, $q=a$, $X=A^{\frac{a-1}{a}}$, $Y=B^{\frac{a-1}{a}}$, $a>1$, we have the inequality:

$$E\left[\left(A^{\frac{a-1}{a}}\right)^{\frac{a}{a-1}}\right]^{\frac{a-1}{a}} \cdot E\left[\left(B^{\frac{a-1}{a}}\right)^{a}\right]^{\frac{1}{a}} \geq E\left[(AB)^{\frac{a-1}{a}}\right]$$