Progress Report 3: A Study on Convergence Results of Stochastic Gradient Methods

B09902055 Weiping Li, B09902073 Chun-Neng Chu

2023.6.6

Problem Statement and Motivation

While surveying several stochastic gradient based methods, the slight differences in the algorithms and constraints / assumptions that resulted in similar convergence complexities in the following articles stood out to us:

1. On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes - Xiaoyu Li et al.
2. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes - Rachel Ward et al.

Hence, we try to understand the purpose and necessity of the differences by studying the proofs of both papers' convergence.

Outline

Differences Between Xiaoyu Li and Rachel Ward

1. Algorithm: The order of gradient / learning rate update is swapped. Xiaoyu Li updates the weights before the learning rate. Rachel Ward does the opposite.
2. Constraints: Other than some shared constraints, Xiaoyu Li additionally requires knowledge of the M-smooth coefficient. Rachel Ward requires L-Lipschitz constraint.

Standard SGD Convergence Proof Format

Intuition: We want to find a complexity bound for $||\nabla f(x_t)||^a$ in the form of $O(\frac{1}{T^\alpha})$. Hence, we use the following thought process.

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

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

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

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

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

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

Note 1: η_t is the learning rate at the t^{th} iteration; η_t^* may be η_t or the estimation of η_t; f^* is the optimal target function value.

Note 2: The increment of numbers (ⓘ) represents the thought process of proof, while the arrows are logical / arithmetic implications.

Insights from ADAGRAD-Norm (Xiaoyu Li et al.)

1. Two non-arithmetic lemmas: 3 and 8
2. Necessity of update sequence swap (Lemma 3)
3. Knowledge of M-smooth constant (Lemma 3 and 8)
4. Xiaoyu Li et al.'s Constraints Applied on Rachel Ward et al.

Note: The slides from the last progress report are included in the appendix.

Goals

1. Find the reason for the different update sequence between the two algorithms.
2. Understand why Xiaoyu Li et al. require prior knowledge of specific smoothness constant in the proof.
3. Understand why Rachel Ward et al. require Lipschitz constraint in proof.
4. Can we acquire the same convergence rate with Rachel Ward et al.'s algorithm, but using the constraints of Xiaoyu Li et al.? Vice versa?

On the Inner Product Term of ⑤

From the M-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.

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

$$= -\frac{\eta \|\nabla f_t\|^2}{b_{t+1}} + \frac{\eta \langle \nabla f_t, \nabla f_t - G_t \rangle}{b_{t+1}} + \frac{\eta^2 M \|G_t\|^2}{2b_{t+1}^2}$$

The result is as follows, where the blue terms are from the inner product term. $E\left[\frac{\eta\|\nabla f_t\|^2}{2\sqrt{b_t^2+\|\nabla f_t\|^2+\sigma^2}}\right] \leq$

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

Note: $\nabla f_t, G_t$ stand for the gradient and stochastic gradient at the t^{th} iteration. For future references, we define $\eta_t^* = \frac{\eta}{\sqrt{b_t^2+\|\nabla f_t\|^2+\sigma^2}}$.

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, $\Delta := \sum_{t=1}^{T} \|\nabla f_t\|^2$ and $a = 2$.

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

$$\geq \frac{E\left[\Delta^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}}}{E\left[\left(\left(\frac{1}{\eta_T}\right)^{\frac{a-1}{a}}\right)^{a}\right]^{\frac{1}{a-1}}}$$

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 slide.

The Significance of a in Xiaoyu Li vs Rachel Ward

- In Xiaoyu Li et al.'s work, the choice of $a = 2$ is used as a trick to bound $E\left[\sqrt{\Delta}\right]$ (from ③), which is a necessary step to drop L-Lipschitz constraint in their proof.
- In Rachel Ward et al.'s work, the choice of $a = 3$ is used to minimize δ's impact in the complexity, or in simpler terms, improve the complexity of convergence.

Bounding $E\left[\sqrt{\Delta}\right]$ with $a=2$ - Xiaoyu Li et al.

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

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

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

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

where

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

Intuitively, when $a=2$, we have

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

Bounding $E\left[\sqrt{\Delta}\right]$ with $a = 2$ - Xiaoyu Li et al.

From the last page, 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 $E\left[\sqrt{\Delta}\right] = O(1)$.
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
   $E\left[\left(\sum_{t=1}^{T} ||\nabla f(x_t)||^2\right)^{\frac{a-1}{a}}\right]^{\frac{a}{a-1}} = O\left(\frac{1}{T^{\alpha-1}}\right)$

The Relation between a and δ - Rachel Ward et al.

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

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

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: $\mathbb{P}\left(\min_{t\in[T]} \|\nabla f_t\|^2 \geq \frac{C_T}{\delta^{\frac{a}{a-1}}}\right) \leq \delta$, 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.

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

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

Note: The technicalities of Lemma 3 and 8 and their relationship with ④ of the standard SGD convergence proof format (i.e. the unanswered question from the last report) can be found in "Appendix - Correspondence between Lemma 3, 8 and SGD Proof Format ④".

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

The red terms are bound by the same techniques for Rachel Ward et al., but the blue terms are not dealt with.
Note: L-Lipschitz provides an upper bound, γ, for $\|\nabla f_t\|$.

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

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

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 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..

The Holder-related inequality

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

QA

Appendix: Difference between Algorithms

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

a.

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

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

b.

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

Appendix: Constraint Tradeoff

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

Appendix: Proof of Lemma 3

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,

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

Taking the conditional expectation with respect to $\xi_0, \cdots, \xi_{t-2}$, we have that $E_{t-1}[\langle \nabla f(x_{t-1}), \eta_t (\nabla f(x_{t-1}) - G(x_{t-1}, \xi_{t-1})) \rangle]$
$= \langle \nabla f(x_{t-1}), \eta_t \nabla f(x_{t-1}) - \eta_t E_t[G(x_{t-1}, \xi_{t-1})] \rangle = 0.$

Appendix: Proof of Lemma 3

Hence, from the law of total expectation, we have
$E\left[\langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right)\rangle\right] \leq$
$E\left[f\left(x_{t-1}\right) - f\left(x_t\right) + \frac{M}{2}\left\|\eta_t g\left(x_{t-1}, \xi_{t-1}\right)\right\|^2\right]$. Summing over $t = 1$ to T and lower bounding $f\left(x_T\right)$ with $f^\star$, we have the stated bound.

Appendix: Proof of Lemma 8

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

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

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

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

Appendix: Proof of Lemma 8

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

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

Appendix

1. L-Lipschitz gradient implies L-smooth:
   (a) f is L-smooth if f is continuously differentiable and $\forall x, y \in dom\ f,\ \langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L||y - x||^2$
   (b) ∇f is L-Lipschitz iff $\forall x, y \in dom\ f,\ ||\nabla f(y) - \nabla f(x)|| \leq L||y - x||$
   (c) Cauchy–Schwarz inequality states that $\forall u, v \in$ an inner product space, $|\langle u, v \rangle| \leq ||u||||v||$.

   By (a), (b), (c) can write
   $\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq ||\nabla f(y) - \nabla f(x)||||y - x|| \leq L||y - x||^2$
2. Convergence rate - expectation form to probability form:
   Markov's inequality: $P(X \geq a) \leq \frac{E[X]}{a}$,if X is a non-negative random variable and $a > 0$
   $\rightarrow P(min_{0 \leq t \leq N}||\nabla f(x_t)||^2 \geq \frac{E[(min_{0 \leq t \leq N}||\nabla f(x_t)||^2]}{\delta}) \leq \delta$
   $\rightarrow P(min_{0 \leq t \leq N}||\nabla f(x_t)||^2 \leq \frac{E[(min_{0 \leq t \leq N}||\nabla f(x_t)||^2]}{\delta}) \geq 1 - \delta$
   $\rightarrow H(\delta, N, D^*) = \frac{E[(min_{0 \leq t \leq N}||\nabla f(x_t)||^2]}{\delta}$

Xiaoyu Li et al. - Non-arithmetic Segments: Lemma 3

When surveying Xiaoyu Li et al.'s article, we noticed that there are only two non-arithmetic lemmas (3, 8).

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

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

Xiaoyu Li et al. - Non-arithmetic Segments: Lemma 8

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

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

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

Xiaoyu Li et al. - Changing Update Sequence for Lemma 3

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

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

This 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.

Xiaoyu Li et al. - Smoothness Constant from Lemma 3 & 8

In the next three slides, we demonstrate why concrete knowledge on the value of smoothness constant M is necessary.

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

Xiaoyu Li et al. - Smoothness Constant from Lemma 3 & 8

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

Xiaoyu Li et al. - Smoothness Constant from Lemma 3 & 8

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

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.

What Happens when Xiaoyu Li et al.'s Constraints Applied on Rachel Ward et al.

1. Since for the t^{th} iteration, Rachel Ward et al. updates the learning rate (η_t) before the weights, we cannot take η_t out of the expectation in the intermediary step that requires $E_{\xi_{t-1}}\left[\langle \nabla f\left(x_{t-1}\right), \eta_t\left(\nabla f\left(x_{t-1}\right) - G\left(x_{t-1}, \xi_{t-1}\right)\right)\rangle\right] = \langle \nabla f\left(x_{t-1}\right), \eta_t \nabla f\left(x_{t-1}\right) - \eta_t E_t\left[G\left(x_{t-1}, \xi_{t-1}\right)\right]\rangle = 0$

2. In Rachel Ward et al.'s article, η_t^* is an estimation of η_t instead of exactly η_t. Since theorem 4 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.

Appendix - Correspondence between Lemma 3, 8 and SGD Proof Format

SGD Proof Format (4) is essentially Lemma 3. The two combined provide a bound for the blue term and is a intermediary step between (3) and (4). Lemma 3:

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

Lemma 8:

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

where

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

Appendix - Correspondence between Lemma 3, 8 and SGD Proof Format

The result is as follows (the intermediary inequality).

$$\left(1-\frac{2\eta M}{\sqrt{b_0^2}}\right) E\left[\sum_{t=1}^{T}\eta_t \left\|\nabla f_{t-1}\right\|^2\right] \leq f\left(x_0\right)-f^{\star}$$

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

$$+\frac{2\eta M}{b_0^2}(1+\ln T)\sigma^2.$$

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.