Progress Report 2: A Study on Convergence Results of Stochastic Gradient Methods

B09902055 Weiping Li, B09902073 Chun-Neng Chu

2023.5.16

Outline

1. Past Progress
   a. Difference between Algorithms
   b. Constraint Tradeoff
2. Goals
3. Standard SGD Convergence Proof Format
4. Insights from ADAGRAD-Norm (Xiaoyu Li et al.)
   a. Non-arithmetic Lemmas
   b. Necessity of Changing Update Sequence - Lemma 3
   c. Necessity of Smoothness Constant - Lemma 3 & 8
   d. Comparison with ADAGRAD-Norm (Rachel Ward et al.)
5. Future Plans
6. QA
7. Appendix
8. References

Difference between Algorithms: Update Sequence

a.

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 $\xi_{t-1}, G_{t-1} = G(x_{t-1}, \xi_{t-1})$
  $x_t \leftarrow x_{t-1} - \frac{\eta}{b_{t-1}} G_{t-1}$
  $b_t^2 \leftarrow b_{t-1}^2 + ||G_{t-1}||^2$
end for

b.

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 $\xi_{t-1}, G_{t-1} = G(x_{t-1}, \xi_{t-1})$
  $b_t^2 \leftarrow b_{t-1}^2 + ||G_{t-1}||^2$
  $x_t \leftarrow x_{t-1} - \frac{\eta}{b_t} G_{t-1}$
end for

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	✓	✓
$E\left[\left\|\nabla f\left(x_t\right)-G\left(x_t,\xi_t\right)\right\|^2\right] \leq \sigma^2$	✓	✓
$E_\xi[G(x,\xi)]=\nabla f(x)$	✓	✓
$f>-\infty$	✓	✓
know smoothness constant M	✓	
L-Lipschitz		✓

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?

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

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

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

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: 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.

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

When surveying Xiaoyu Li et al.'s article, we noticed that only two segments in the proof require 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 the next three slides, we demonstrate why concrete knowledge on the value of smoothness constant M is necessary.

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

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

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

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

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

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

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

Future Plans

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?

QA

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,

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

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

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 $\ln(\frac{1}{x}) \geq \eta\left(1 - x^{\frac{1}{\eta}}\right)$.

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

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

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

Appendix: Proof of Lemma 8

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

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

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

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.