AdaGrad stepsizes : sharp convergence over nonconvex landscapes, from any initialization

Rachel Ward, Xiaoxia Wu and Léon Bottou
(2018)

Presented by Jiin Seo

June 3, 2019

Outline

Outline

1. Introduction

- Theoretical guarantees for the convergence of AdaGrad for smooth, nonconvex functions
- Convergence rate of AdaGrad-Norm

$$\begin{cases} \mathcal{O}(log(T)/\sqrt{T}) & \text{(stochastic setting).} \\ \text{optimal } \mathcal{O}(1/T) & \text{(non-stochastic setting).} \end{cases}$$

- Strong robustness of AdaGrad-Norm to the hyper-parameters (η and b_0)

1. Introduction

Problem setting

Minimize a differentiable non-convex function $F : \mathbb{R}^d \to \mathbb{R}$ via SGD.

- Stochastic Gradient Descent (SGD).
  Starting from $x_0 \in \mathbb{R}^d$ and η_0 ; SGD iterates until convergence

$$x_{t+1} \leftarrow x_t - \eta_t G(x_t), \quad \eta_t > 0$$

- $G(x_t)$: stochastic gradient
  ($\mathbb{E}|G(x_t)| = \nabla F(x_t)$ and having bounded variance)
- $F(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$: Loss ftn $\Rightarrow$ $(Full\ gradient) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x)$
- $G_t(x) = \nabla f_{i_t}(x), i_t \sim Unif\{1, 2, \cdots, m\} \Rightarrow$ (efficient!)

1. Introduction

Notation

- $||\cdot||$: l_2-norm
- $[T] := \{0, 1, 2, \cdots, T\}$
- A function $F : \mathbb{R}^d \to \mathbb{R}$ has L-Lipschitz smooth gradient if

$$||\nabla F(x) - \nabla F(y)|| \leq L||x - y|| \quad \forall x, y \in \mathbb{R}^d$$

- If $L > 0$ is the smallest number s.t. the above is satisfied, we refer to L as the smoothness constant for F and we write $F \in \mathbb{C}^1_L$.

1. Introduction

Coordinate version of AdaGrad (Lafond et al., 2017)

It updates an entire vector of per-coefficient stepsizes.
d-scalar parameters $b_t(k)(k = 1, 2, \cdots, d)$

- Coordinate version
  1. At the t-th iteration,

$$x_{t+1}(k) \leftarrow x_t(k) - \eta \frac{G_t(k)}{b_{t+1}(k)} \quad (k = 1, 2, \cdots, d) \ (\eta > 0).$$

$$b_{t+1}(k)^2 = \begin{cases} b_t(k)^2 + (\nabla F(x_t))_k^2, & \text{(noiseless setting).} \\ b_t(k)^2 + (G_t(k))^2, & \text{(noisy gradient setting).} \end{cases}$$

1. Introduction

Norm version of AdaGrad (AdaGrad-Norm)

AdaGrad-Norm updates only a single (scalar) stepsize according to the sum of squared gradient norms observed so far.

- AdaGrad-Norm
  1. Initialize a single scalar $b_0 > 0$
  2. At the t-th iteration, observe the r.v. G_t s.t. $\mathbb{E}[G_t] = \nabla F(x_t)$ and iterate

$$x_{t+1} \leftarrow x_t - \eta \frac{G(x_t)}{b_{t+1}} \quad \textit{with} \quad b_{t+1}^2 = b_t^2 + ||G(x_t)||^2 \quad , (\eta > 0)$$

1. Introduction

Previous work

- Theoretical guarantees of convergence for AdaGrad in the setting of online convex optimization(Duchi et al., 2011)
- Guarantees of convergence in the non-convex setting(Wu et al., 2018) $\rightarrow$ only for the batch setting

Future work

- Convergence guarantees for AdaGrad-Norm over smooth, nonconvex functions, in both the stochastic and deterministic settings.

Outline

2. AdaGrad-Norm Convergence

Algorithm 1 AdaGrad-Norm

1: **Input :** Initialize $x_0 \in \mathbb{R}^d, b_t > 0, \eta > 0$ and the total iterations T
2: **for** $t = 1, 2, \cdots, T$ **do**
3: Generate ξ_{t-1} and $G_{t-1} = G(x_{t-1}, \xi_{t-1})$
4: $b_t^2 \leftarrow b_{t-1}^2 + ||G_{t-1}||^2$
5: $x_t \leftarrow x_{t-1} - \frac{\eta}{b_t} G_{t-1}$
6: **end for**

At the kth iteration, we observe a stochastic gradient $G(x_k, \xi_k) = G_k$ and $\mathbb{E}_{\xi_k}[G(x_k, \xi_k)] = \nabla F(x_k)$ is UE of $\nabla F(x_t)$.

[Assumptions]

1. The random vectors $\xi_k \perp\!\!\!\perp \xi_l$ and $\xi_k \perp\!\!\!\perp x_k$ $(k, l = 0, 1, 2, \cdots)$
2. $\mathbb{E}_{\xi_k}[||G(x_k, \xi_k) - \nabla F(x_k)||^2] \leq \sigma^2$
3. $||\nabla F(x)||^2 \leq \gamma^2$ uniformly.

2. AdaGrad-Norm Convergence

Theorem 2.1 (AdaGrad-Norm: convergence in stochastic setting)

Suppose $F \in \mathbb{C}_L^1$ and $F^* = inf_x F(x) > -\infty$. Suppose that the r.v.s $G_l, (l \geq 0)$, satisfy the above assumptions. Then with probability $1 - \delta$,

$$\min_{l \in [T-1]} ||\nabla F(x_l)||^2 \leq min\{(\frac{2b_0}{T} + \frac{2\sqrt{2}(\gamma + \sigma)}{\sqrt{T}})\frac{\mathcal{Q}}{\delta^{3/2}}, (\frac{8\mathcal{Q}}{\delta} + 2b_0)\frac{4\mathcal{Q}}{T\delta} + \frac{8\mathcal{Q}\sigma}{\delta^{3/2}\sqrt{T}}\}$$

where

$$\mathcal{Q} = \frac{F(x_0) - F^*}{\eta} + \frac{4\sigma + \eta L}{2} log(\frac{20T(\gamma^2 + \sigma^2)}{b_0^2} + 10).$$

2. AdaGrad-Norm Convergence

Theorem 2.1 (AdaGrad-Norm: convergence in stochastic setting)

- AdaGrad-Norm converges for any $\eta > 0$ and starting from any value of $b_0 > 0$.
- Good strategy for setting hyperparameters :
  Given knowledge of F^*, set $\eta = F(x_0) - F^*$ and $b_0 > 0$ to be very small.
- With a priori knowledge of L and σ^2,

$$\eta = \min\left\{\frac{1}{L}, \frac{1}{\sigma\sqrt{T}}\right\}, \quad j = 0, 1, \ldots, T-1$$

then with probability $1 - \delta$

$$\min_{\ell \in [T-1]} \|\nabla F(x_\ell)\|^2 \leq \frac{2L(F(x_0) - F^*)}{T\delta} + \frac{(L + 2(F(x_0) - F^*))\sigma}{\delta\sqrt{T}}.$$

2. AdaGrad-Norm Convergence

Theorem 2.2 (AdaGrad-Norm: convergence in deterministic setting)

Suppose $F \in \mathbb{C}_L^1$ and $F^* = inf_x F(x) > -\infty$. Consider AdaGrad-Norm in deterministic setting with following update,

$$x_t = x_{t-1} - \frac{\eta}{b_t}\nabla F(x_{t-1}), \quad b_t^2 = b_{t-1}^2 + ||\nabla F(x_{t-1})||^2$$

then $\min\limits_{t\in[T]} ||\nabla F(x_t)||^2 \leq \varepsilon$ after

(1) $T = 1 + \left\lceil \frac{1}{\varepsilon}\left(\frac{4(F(x_0)-F^*)^2}{\eta^2} + \frac{2b_0(F(x_0)-F^*)}{\eta}\right)\right\rceil$ if $b_0 \geq \eta L$

(2) $T = 1 +$
$\left\lceil \frac{1}{\varepsilon}\left(2L\left(F\left(x_0\right) - F^*\right) + \left(\frac{2(F(x_0)-F^*)}{\eta} + \eta L C_{b_0}\right)^2 + (\eta L)^2\left(1 + C_{b_0}\right) - b_0^2\right)\right\rceil$

if $b_0 < \eta L$. Here $C_{b_0} = 1 + 2\log\left(\frac{\eta L}{b_0}\right)$.

2. AdaGrad-Norm Convergence

Theorem 2.2 (AdaGrad-Norm: convergence in deterministic setting)

- AdaGrad-Norm convergence holds for any choice of parameters b_0 and η.
- Good strategy for setting hyperparameters :
  Given knowledge of L and F^*, set $\eta = F(x_0) - F^*$ and $b_0 = \eta L$.

2. AdaGrad-Norm Convergence

Lemma 2.1

Suppose that $F \in \mathbb{C}_L^1$ and $F^* = inf_x F(x) > -\infty$. Consider gradient descent with constant stepsize, $x_{t+1} = x_t - \frac{\nabla F(x_t)}{b}$.
If $b \geq L$, then $\min_{t \in [T-1]} ||\nabla F(x_t)||^2 \leq \varepsilon$ after at most a number of steps

$$T = \frac{2b(F(x_0) - F^*)}{\varepsilon}$$

Alternatively, if $b \leq \frac{L}{2}$, then convergence is not guaranteed at all - gradient descent can oscillate or diverge.

Outline

3. Proof of Theorem 2.1

Theorem 2.1 (AdaGrad-Norm: convergence in stochastic setting)

Suppose $F \in \mathbb{C}_L^1$ and $F^* = inf_x F(x) > -\infty$. Suppose that the r.v. $G_l, l \geq 0$, satisfy the above assumptions. Then with probability $1-\delta$,

$$\min_{l \in [T-1]} ||\nabla F(x_l)||^2 \leq min\{(\frac{2b_0}{T} + \frac{2\sqrt{2}(\gamma+\sigma)}{\sqrt{N}})\frac{\mathcal{Q}}{\delta^{3/2}}, (\frac{8\mathcal{Q}}{\delta} + 2b_0)\frac{4\mathcal{Q}}{T\delta} + \frac{8\mathcal{Q}\sigma}{\delta^{3/2}\sqrt{T}}\}$$

where

$$\mathcal{Q} = \frac{F(x_0) - F^*}{\eta} + \frac{4\sigma + \eta L}{2} log(\frac{20T(\gamma^2 + \sigma^2)}{b_0^2} + 10).$$

3. Proof of Theorem 2.1

Lemma 3.1 (Descent Lemma)

Let $F \in \mathbb{C}_L^1$. Then,

$$F(x) \leq F(y) + \langle \nabla F(y), x - y \rangle + \frac{L}{2}||x - y||^2.$$

Lemma 3.2

For any non-negative $a_1, \cdots, a_T$, and $a_1 \geq 1$, we have

$$\sum_{l=1}^{T} \frac{a_l}{\sum_{i=1}^{l} a_i} \leq log(\sum_{i=1}^{T} a_i) + 1.$$

3. Proof of Theorem 2.1

Proof

Let $F_t = F(x_t)$ and $\nabla F_t = \nabla F(x_t)$. By Lemma 3.1, for $t \geq 0$,

$$\begin{aligned}\frac{F_{t+1} - F_t}{\eta} &\leq - \left\langle \nabla F_t, \frac{G_t}{b_{t+1}} \right\rangle + \frac{\eta L}{2b_{t+1}^2} \|G_t\|^2 \\ &= -\frac{\|\nabla F_t\|^2}{b_{t+1}} + \frac{\langle \nabla F_t, \nabla F_t - G_t \rangle}{b_{t+1}} + \frac{\eta L \|G_t\|^2}{2b_{t+1}^2}\end{aligned}$$

Since b_{t+1} and G_t are correlated and thus for the condi. expectation

$$\mathbb{E}_{\xi_j}\left[\frac{\langle \nabla F_t, \nabla F_t - G_t \rangle}{b_{t+1}}\right] \neq \frac{\mathbb{E}_{\xi_j}\left[\langle \nabla F_t, \nabla F_t - G_t \rangle\right]}{b_{t+1}} = \frac{1}{b_{t+1}} \cdot 0$$

3. Proof of Theorem 2.1

Proof

We use the estimate $\frac{1}{\sqrt{b_t^2+||\nabla F_t^2||+\sigma^2}}$ as a surrogate for $\mathbb{E}_{\xi_t}[\frac{1}{b_{t+1}}]$ to proceed. Condition on $\xi_1, \cdots, \xi_{t-1}$ and take expectation w.r.t ξ_t,

$$0 = \frac{\mathbb{E}_{\xi_t}\left[\langle \nabla F_t, \nabla F_t - G_t\rangle\right]}{\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}} = \mathbb{E}_{\xi_t}\left[\frac{\langle \nabla F_t, \nabla F_t - G_t\rangle}{\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}}\right]$$

$$\begin{aligned}
&\frac{\mathbb{E}_{\xi_t}[F_{t+1}] - F_t}{\eta} \\
&\le \mathbb{E}_{\xi_t}\left[\frac{\langle \nabla F_t, \nabla F_t - G_t\rangle}{b_{t+1}} - \frac{\langle \nabla F_t, \nabla F_t - G_t\rangle}{\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}}\right] - \mathbb{E}_{\xi_t}\left[\frac{\|\nabla F_t\|^2}{b_{t+1}}\right] + \mathbb{E}_{\xi_t}\left[\frac{L\eta\|G_t\|^2}{2b_{t+1}^2}\right] \\
&= \mathbb{E}_{\xi_t}\left[\left(\frac{1}{\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}} - \frac{1}{b_{t+1}}\right)\langle \nabla F_t, G_t\rangle\right] - \frac{\|\nabla F_t\|^2}{\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}} + \frac{\eta L}{2}\mathbb{E}_{\xi_t}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right]
\end{aligned} \tag{1}$$

3. Proof of Theorem 2.1

Proof

$$
\begin{aligned}
&\frac{1}{\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}} - \frac{1}{b_{t+1}} \\
&= \frac{(\|G_t\| - \|\nabla F_t\|)(\|G_t\| + \|\nabla F_t\|) - \sigma^2}{b_{t+1}\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}\left(\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2} + b_{t+1}\right)} \\
&\leq \frac{\|G_t\| - \|\nabla F_t\| \,|}{b_{t+1}\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}} + \frac{\sigma}{b_{t+1}\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}}
\end{aligned}
$$

$$
\mathbb{E}_{\xi_t}\left[\left(\frac{1}{\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}} - \frac{1}{b_{t+1}}\right)\langle \nabla F_t, G_t\rangle\right]
$$

$$
\leq \mathbb{E}_{\xi_t}\left[\frac{\|G_t\| - \|\nabla F_t\| \, \|G_t\| \, \|\nabla F_t\|}{b_{t+1}\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}}\right] + \mathbb{E}_{\xi_t}\left[\frac{\sigma \, \|G_t\| \, \nabla F_t\|}{b_{t+1}\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}}\right] \tag{2}
$$

3. Proof of Theorem 2.1

Proof

By applying the inequality $ab \le \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$ with $\lambda = \frac{2\sigma^2}{\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}}$, $a = \frac{\|G_t\|}{b_{t+1}}$, and $b = \frac{\|G_t\|-\|\nabla F_t\|\|\nabla F_t\|}{b_t^2+\|\nabla F_t\||}$, the first term of the RHS in (2) can be bounded as

$$
\begin{aligned}
&\mathbb{E}_{\xi_t}\left[\frac{\|G_t\| - \|\nabla F_t\|\,\|G_t\|\,\|\nabla F_t\|}{b_{t+1}\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}}\right] \\
&\le \frac{\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}\,\|\nabla F_t\|^2\,\mathbb{E}_{\xi_t}\left[(\|G_t\| - \|\nabla F_t\|)^2\right]}{b_t^2 + \|\nabla F_t\|^2 + \sigma^2} \\
&\quad + \frac{\sigma^2}{\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}}\mathbb{E}_{\xi_t}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right] \\
&\le \frac{\|\nabla F_t\|^2}{4\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}} + \sigma\mathbb{E}_{\xi_t}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right].
\end{aligned} \tag{3}
$$

3. Proof of Theorem 2.1

Proof

Similarly, applying the inequality $ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$ with $\lambda = \frac{2}{\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}}$, $a = \frac{\sigma\|G_t\|}{b_{t+1}}$, and $b = \frac{\|\nabla F_t\|}{\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}}$, the second term of the RHS in (2) is bounded by

$$\mathbb{E}_{\xi_t}\left[\frac{\sigma\left\|\nabla F_t\right\|\left\|G_t\right\|}{b_{t+1}\sqrt{b_t^2+\left\|\nabla F_t\right\|^2+\sigma^2}}\right] \leq \sigma\mathbb{E}_{\xi_t}\left[\frac{\left\|G_t\right\|^2}{b_{t+1}^2}\right] + \frac{\left\|\nabla F_t\right\|^2}{4\sqrt{b_t^2+\left\|\nabla F_t\right\|^2+\sigma^2}}. \tag{4}$$

3. Proof of Theorem 2.1

Proof

Thus, putting inequalities (3) and (4) back into (2) gives

$$\mathbb{E}_{\xi_t}\left[\left(\frac{1}{\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}}-\frac{1}{b_{t+1}}\right)\langle\nabla F_t, G_t\rangle\right] \leq 2\sigma\mathbb{E}_{\xi_t}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right]+\frac{\|\nabla F_t\|^2}{2\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}}.$$

and, therefore, back to (1),

$$\frac{\mathbb{E}_{\xi_t}[F_{t+1}]-F_t}{\eta} \leq \frac{\eta L}{2}\mathbb{E}_{\xi_t}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right]+2\sigma\mathbb{E}_{\xi_t}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right]-\frac{\|\nabla F_t\|^2}{2\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}}.$$

Rearranging,

$$\frac{\|\nabla F_t\|^2}{2\sqrt{b_t^2+\|\nabla F_t\|^2+\sigma^2}} \leq \frac{F_t-\mathbb{E}_{\xi_t}[F_{t+1}]}{\eta}+\frac{4\sigma+\eta L}{2}\mathbb{E}_{\xi_t}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right]$$

3. Proof of Theorem 2.1

Proof

We take the expectation w.r.t. $\xi_{t-1}, \xi_{t-2}, \ldots, \xi_1$, and arrive at the recursion(Law of total expectation)

$$\mathbb{E}\left[\frac{\|\nabla F_t\|^2}{2\sqrt{b_t^2 + \|\nabla F_t\|^2 + \sigma^2}}\right] \le \frac{\mathbb{E}[F_t] - \mathbb{E}[F_{t+1}]}{\eta} + \frac{4\sigma + \eta L}{2}\mathbb{E}\left[\frac{\|G_t\|^2}{b_{t+1}^2}\right]$$

Taking $t = T$ and summing up from $k = 0$ to $k = T - 1$

$$\begin{aligned} \sum_{k=0}^{T-1}\mathbb{E}\left[\frac{\|\nabla F_k\|^2}{2\sqrt{b_k^2 + \|\nabla F_k\|^2 + \sigma^2]}}\right] & \\ \le \frac{F_0 - F^*}{\eta} + \frac{4\sigma + \eta L}{2}\mathbb{E}\sum_{k=0}^{T-1}\left[\frac{\|G_k\|^2}{b_{k+1}^2}\right] & \quad (5) \\ \le \frac{F_0 - F^*}{\eta} + \frac{4\sigma + \eta L}{2}\log\left(10 + \frac{20T\left(\sigma^2 + \gamma^2\right)}{b_0^2}\right) & \end{aligned}$$

where the second inequality we apply Lemma (3.2)

3. Proof of Theorem 2.1

Proof

and then Jensen's inequality to bound the summation:

$$\begin{aligned}\mathbb{E}\sum_{k=0}^{T-1}\left[\frac{\|G_k\|^2}{b_{k+1}^2}\right] &\leq \mathbb{E}\left[1+\log\left(1+\sum_{k=0}^{T-1}\|G_k\|^2/b_0^2\right)\right] \\ &\leq \log\left(10+\frac{20T\left(\sigma^2+\gamma^2\right)}{b_0^2}\right).\end{aligned} \tag{6}$$

since

$$\begin{aligned}\mathbb{E}\left[b_k^2-b_{k-1}^2\right] &\leq \mathbb{E}\left[\|G_k\|^2\right] \\ &\leq 2\mathbb{E}\left[\|G_k-\nabla F_k\|^2\right]+2\mathbb{E}\left[\|\nabla F_k\|^2\right] \\ &\leq 2\sigma^2+2\gamma^2\end{aligned} \tag{7}$$

3. Proof of Theorem 2.1

Proof of the 1st bound for Theorem 2.1

For the term on LHS in equation (5), we apply $H\ddot{o}lder's$ inequality,

$$\frac{\mathbb{E}|XY|}{\left(\mathbb{E}|Y|^3\right)^{\frac{1}{3}}} \leq \left(\mathbb{E}|X|^{\frac{3}{2}}\right)^{\frac{2}{3}}$$

with $X = \left(\frac{\|\nabla F_k\|^2}{\sqrt{b_k^2+\|\nabla F_k\|^2+\sigma^2}}\right)^{\frac{2}{3}}$ and $Y = \left(\sqrt{b_k^2 + \|\nabla F_k\|^2 + \sigma^2}\right)^{\frac{2}{3}}$.

$$\mathbb{E}\left[\frac{\|\nabla F_k\|^2}{2\sqrt{b_k^2 + \|\nabla F_k\|^2 + \sigma^2}}\right] \geq \frac{\left(\mathbb{E}\,\|\nabla F_k\|^{\frac{4}{3}}\right)^{\frac{3}{2}}}{2\sqrt{\mathbb{E}\left[b_k^2 + \|\nabla F_k\|^2 + \sigma^2\right]}}$$

$$\geq \frac{\left(\mathbb{E}\,\|\nabla F_k\|^{\frac{4}{3}}\right)^{\frac{3}{2}}}{2\sqrt{b_0^2 + 2(k+1)\left(\gamma^2 + \sigma^2\right)}}$$

where the last inequality is due to inequality (7).

3. Proof of Theorem 2.1

Proof of the 1st bound for Theorem 2.1

Thus (5) arrives at the inequality

$$\frac{T \min_{k\in[T-1]} \left(\mathbb{E}\left[\|\nabla F_k\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}}}{2\sqrt{b_0^2 + 2T\left(\gamma^2+\sigma^2\right)}} \leq \frac{F_0 - F^*}{\eta} + \frac{4\sigma + \eta L}{2}\left(\log\left(1 + \frac{2T\left(\sigma^2+\gamma^2\right)}{b_0^2}\right) + 1\right).$$

Multiplying by $\frac{2b_0 + 2\sqrt{2T}(\gamma+\sigma)}{T}$, the above inequality gives

$$\min_{k\in[T-1]} \left(\mathbb{E}\left[\|\nabla F_k\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \leq \underbrace{\left(\frac{2b_0}{T} + \frac{2\sqrt{2}(\gamma+\sigma)}{\sqrt{T}}\right) C_F}_{C_T}$$

$$C_F = \frac{F_0 - F^*}{\eta} + \frac{4\sigma + \eta L}{2}\log\left(\frac{20T\left(\sigma^2+\gamma^2\right)}{b_0^2} + 10\right).$$

3. Proof of Theorem 2.1

Proof of the 1st bound for Theorem 2.1

Finally, the bound is obtained by Markov's Inequality:

$$\begin{aligned}
\mathbb{P}\left(\min_{k\in[T-1]} \|\nabla F_k\|^2 \geq \frac{C_T}{\delta^{3/2}}\right) &= \mathbb{P}\left(\min_{k\in[T-1]} \left(\|\nabla F_k\|^2\right)^{2/3} \geq \left(\frac{C_T}{\delta^{3/2}}\right)^{2/3}\right) \\
&\leq \delta \frac{\mathbb{E}\left[\min_{k\in[T-1]} \|\nabla F_k\|^{4/3}\right]}{C_T^{2/3}} \\
&\leq \delta
\end{aligned}$$

where in the second step Jensen's inequality is applied to the concave function $\phi(x) = \min_k h_k(x)$.

Proof of the 2nd bound for Theorem 2.1

First, observe with probability $1-\delta'$ that

$$\sum_{i=0}^{T-1} \|\nabla F_i - G_i\|^2 \le \frac{T\sigma^2}{\delta'}$$

Let $Z = \sum_{k=0}^{T-1} \|\nabla F_k\|^2$, then

$$\begin{aligned}
b_{T-1}^2 + \|\nabla F_{T-1}\|^2 + \sigma^2 &= b_0^2 + \sum_{i=0}^{T-2} \|G_i\|^2 + \|\nabla F_{T-1}\|^2 + \sigma^2 \\
&\le b_0^2 + 2\sum_{i=0}^{T-1} \|\nabla F_i\|^2 + 2\sum_{i=0}^{T-2} \|\nabla F_i - G_i\|^2 + \sigma^2 \\
&\le b_0^2 + 2Z + 2T\frac{\sigma^2}{\delta'}
\end{aligned}$$

3. Proof of Theorem 2.1

Proof of the 2nd bound for Theorem 2.1

In addition, from inequality (5), i.e.,

$$\mathbb{E}\left[\frac{\sum_{k=0}^{T-1}\|\nabla F_k\|^2}{2\sqrt{b_{T-1}^2+\|\nabla F_{T-1}\|^2+\sigma^2}}\right] \leq \frac{F_0-F^*}{\eta}+\frac{4\sigma+\eta L}{2}\log\left(10+\frac{20T\left(\sigma^2+\gamma^2\right)}{b_0^2}\right) \triangleq C_F$$

we have with probability $1-\hat{\delta}-\delta'$ that

$$\frac{C_F}{\hat{\delta}} \geq \frac{\sum_{k=0}^{T-1}\|\nabla F_k\|^2}{2\sqrt{b_{T-1}^2+\|\nabla F_{T-1}\|^2+\sigma^2}} \geq \frac{Z}{2\sqrt{b_0^2+2Z+2T\sigma^2/\delta'}}$$

3. Proof of Theorem 2.1

Proof of the 2nd bound for Theorem 2.1

That is equivalent to solve the following quadratic equation

$$Z^2 - \frac{8C_F^2}{\hat{\delta}^2} Z - \frac{4C_F^2}{\hat{\delta}^2}\left(b_0^2 + \frac{2T\sigma^2}{\delta'}\right) \leq 0$$

which gives

$$\begin{aligned} Z &\leq \frac{4C_F^2}{\hat{\delta}^2} + \sqrt{\frac{16C_F^4}{\hat{\delta}^4} + \frac{4C_F^2}{\hat{\delta}^2}\left(b_0^2 + \frac{2T\sigma^2}{\delta'}\right)} \\ &\leq \frac{8C_F^2}{\hat{\delta}^2} + \frac{2C_F}{\hat{\delta}}\left(b_0 + \frac{\sqrt{2T}\sigma}{\sqrt{\delta'}}\right) \end{aligned}$$

Let $\hat{\delta} = \delta' = \frac{\delta}{2}$. Replacing Z with $\sum_{k=0}^{T-1} \|\nabla F_k\|^2$ and dividing both side with T we have with probability $1-\delta$

$$\min_{k\in[T-1]} \|\nabla F_k\|^2 \leq \frac{4C_F}{T\delta}\left(\frac{8C_F}{\delta} + 2b_0\right) + \frac{8\sigma C_F}{\delta^{3/2}\sqrt{T}}$$

Outline

4. Proof of Theorem 2.2

Theorem 2.2 (AdaGrad-Norm: convergence in deterministic setting)

Suppose $F \in \mathbb{C}_L^1$ and $F^* = inf_x F(x) > -\infty$. Consider AdaGrad-Norm in deterministic setting with following update,

$$x_t = x_{t-1} - \frac{\eta}{b_t}\nabla F(x_{t-1}), \quad b_t^2 = b_{t-1}^2 + ||\nabla F(x_{t-1})||^2$$

then $\min_{t\in[T]} ||\nabla F(x_t)||^2 \leq \varepsilon$ after

(1) $T = 1 + \left\lceil \frac{1}{\varepsilon}\left(\frac{4(F(x_0)-F^*)^2}{\eta^2} + \frac{2b_0(F(x_0)-F^*)}{\eta}\right)\right\rceil$ if $b_0 \geq \eta L$

(2) $T = 1 +$

$\left\lceil \frac{1}{\varepsilon}\left(2L\left(F\left(x_0\right) - F^*\right) + \left(\frac{2(F(x_0)-F^*)}{\eta} + \eta L C_{b_0}\right)^2 + (\eta L)^2\left(1 + C_{b_0}\right) - b_0^2\right)\right\rceil$

if $b_0 < \eta L$. Here $C_{b_0} = 1 + 2\log\left(\frac{\eta L}{b_0}\right)$.

4. Proof of Theorem 2.2

Lemma 4.1

Fix $\varepsilon \in (0, 1]$ and $C > 0$. For any non-negative $a_0, a_1, \ldots,$ the dynamical system

$$b_0 > 0; \quad b_{t+1}^2 = b_t^2 + a_t$$

has the property that after $T = \left\lceil \frac{C^2 - b_0^2}{\varepsilon} \right\rceil + 1$ iterations, either min $_{k=0:T-1} a_k \leq \varepsilon$, or $b_T \geq \eta L$.

$\Rightarrow$After an initial number of steps $T = \left\lceil \frac{(\eta L)^2 - b_0^2}{\varepsilon} \right\rceil + 1$, either we have already reached a point x_k s.t. $\|\nabla F(x_k)\|^2 \leq \varepsilon$, or else $b_T \geq \eta L$

4. Proof of Theorem 2.2

Lemma 4.2

Suppose $F \in \mathbb{C}_L^1$ and $F^* = \inf_x F(x) > -\infty$. Denote by $k_0 \geq 1$ the first index such that $b_{k_0} \geq \eta L$. Then for all $b_k < \eta L, k = 0, 1, \ldots, k_0 - 1$

$$F_{k_0-1} - F^* \leq F_0 - F^* + \frac{\eta^2 L}{2}\left(1 + 2\log\left(\frac{b_{k_0-1}}{b_0}\right)\right)$$

$\Rightarrow \{F(x_k)\}_{k=0}^{\infty}$ is a bounded sequence for any value of $b_0 > 0$

4. Proof of Theorem 2.2

Proof

By Lemma 4.1, if $\min_{k\in[T-1]} \|\nabla F(x_k)\|^2 \leq \varepsilon$ is not satisfied after $T = \left\lceil \frac{(\eta L)^2 - b_0^2}{\varepsilon} \right\rceil + 1$ steps, then there exits a first index $1 \leq k_0 \leq T$ s.t. $\frac{b_{k_0}}{\eta} > L$. By Lemma 3.1, for $j \geq 0$

$$
\begin{aligned}
F_{k_0+j} &\leq F_{k_0+j-1} + \langle \nabla F_{k_0+j-1}, (x_{k_0+j} - x_{k_0+j-1}) \rangle + \frac{L}{2} ||(x_{k_0+j} - x_{k_0+j-1})||^2 \\
&= F_{k_0+j-1} - \frac{\eta}{b_{k_0+j}} \left(1 - \frac{\eta L}{2b_{k_0+j}}\right) \|\nabla F_{k_0+j-1}\|^2 \\
&\leq F_{k_0-1} - \sum_{\ell=0}^{j} \frac{\eta}{2b_{k_0+\ell}} \|\nabla F_{k_0+\ell-1}\|^2 \\
&\leq F_{k_0-1} - \frac{\eta}{2b_j} \sum_{\ell=0}^{j} \|\nabla F_{k_0+\ell-1}\|^2 .
\end{aligned}
$$

4. Proof of Theorem 2.2

Proof

Let $Z = \sum_{k=k_0-1}^{M-1} \|\nabla F_k\|^2$, it follows that

$$\frac{2\left(F_{k_0-1} - F^*\right)}{\eta} \geq \frac{2\left(F_0 - F_M\right)}{\eta} \geq \frac{\sum_{k=k_0-1}^{M-1} \|\nabla F_k\|^2}{b_M} \geq \frac{Z}{\sqrt{Z + b_{k_0-1}^2}}.$$

Solving the quadratic inequality for Z,

$$\sum_{k=k_0-1}^{M-1} \|\nabla F_k\|^2 \leq \frac{4\left(F_{k_0-1} - F^*\right)^2}{\eta^2} + \frac{2\left(F_{k_0-1} - F^*\right) b_{k_0-1}}{\eta}$$

If $k_0 = 1$, the stated result holds by multiplying both side by $\frac{1}{M}$.
Otherwise, $k_0 > 1$ From Lemma 4.2, we have

$$F_{k_0-1} - F^* \leq F_0 - F^* + \frac{\eta^2 L}{2}\left(1 + 2\log\left(\frac{\eta L}{b_0}\right)\right)$$

4. Proof of Theorem 2.2

Proof

Replacing $F_{k_0-1} - F^*$ in (15) by above bound, we have

$$\sum_{k=k_0-1}^{M-1} \|\nabla F_k\|^2$$

$$\leq \left(\frac{2(F_0 - F^*)}{\eta} + \eta L \left(1 + 2\log\left(\frac{\eta L}{b_0}\right)\right)\right)^2$$

$$+ 2L(F_0 - F^*) + (\eta L)^2 \left(1 + 2\log\left(\frac{\eta L}{b_0}\right)\right) = C_M$$

Thus, we are assured that

$$\min_{k=0:T+M-1} \|\nabla F_k\|^2 \leq \varepsilon$$

where $T \leq \frac{L^2 - b_0^2}{\varepsilon}$ and $M = \frac{C_M}{\varepsilon}$. ■

4. Proof of Theorem 2.2

Proof of Lemma 4.1

If $b_0 \geq \eta C$, we are done. Else $b_0 < C$. Let T be the smallest integer such that $T \geq \frac{C^2 - b_0^2}{\varepsilon}$. Suppose $b_T < C$. Then

$$C^2 > b_T^2 = b_0^2 + \sum_{k=0}^{T-1} a_k > b_0^2 + T \min_{k \in [T-1]} a_k \quad \Rightarrow \quad \min_{k \in [T-1]} a_k \leq \frac{C^2 - b_0^2}{T}$$

Hence, for $T \geq \frac{C^2 - b_0^2}{\varepsilon_0}$, $\min_{k \in [N-1]} a_k \leq \varepsilon$. Suppose $\min_{k \in [T-1]} a_k > \epsilon$, then from above inequalities we have $b_T > C$. ■

4. Proof of Theorem 2.2

Proof of Lemma 4.2

Suppose $k_0 \geq 1$ is the first index such that $b_{k_0} \geq \eta L$. By Lemma 3.1, for $j \leq k_0 - 1$

$$F_{j+1} \leq F_j - \frac{\eta}{b_{j+1}} \left(1 - \frac{\eta L}{2b_{j+1}}\right) \|\nabla F_j\|^2$$

$$\leq F_j + \frac{\eta^2 L}{2b_{j+1}^2} \|\nabla F_j\|^2 \leq F_0 + \sum_{\ell=0}^{j} \frac{\eta^2 L}{2b_{\ell+1}^2} \|\nabla F_\ell\|^2$$

$$\Rightarrow F_{k_0-1} - F_0 \leq \frac{\eta^2 L}{2} \sum_{i=0}^{k_0-2} \frac{\|\nabla F_i\|^2}{b_{i+1}^2} \leq \frac{\eta^2 L}{2} \sum_{i=0}^{k_0-2} \frac{(\|\nabla F_i\| / b_0)^2}{\sum_{\ell=0}^{i} (\|\nabla F_\ell\| / b_0)^2 + 1}$$

$$\leq \frac{\eta^2 L}{2} \left(1 + \log\left(1 + \sum_{\ell=0}^{k_0-2} \frac{\|\nabla F_\ell\|^2}{b_0^2}\right)\right) \qquad \text{by Lemma 3.2}$$

$$\leq \frac{\eta^2 L}{2} \left(1 + \log\left(\frac{b_{k_0-1}^2}{b_0^2}\right)\right) \qquad \blacksquare$$