Supplementary Material for
AdaGrad Stepsizes: Sharp Convergence Over Nonconvex Landscapes

A. The Proof of Second Bound in Theorem 2.2

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

$$\sum_{i=0}^{N-1} \|\nabla F_i - G_i\|^2 \leq \frac{N\sigma}{\delta'}.$$

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

$$\begin{aligned}
& b_{N-1}^2 + \|\nabla F_{N-1}\|^2 + \sigma^2 \\
=& b_0^2 + \sum_{i=0}^{N-2} \|G_i\|^2 + \|\nabla F_{N-1}\|^2 + \sigma^2 \\
\leq& b_0^2 + 2\sum_{i=0}^{N-1} \|\nabla F_i\|^2 + 2\sum_{i=0}^{N-2} \|\nabla F_i - G_i\|^2 + \sigma^2 \\
\leq& b_0^2 + 2Z + 2N\frac{\sigma^2}{\delta'}
\end{aligned}$$

In addition, from equality (10), i,e,

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

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

$$\begin{aligned}
\frac{\mathcal{Q}}{\hat{\delta}} &\geq \frac{\sum_{k=0}^{N-1}\|\nabla F_k\|^2}{2\sqrt{b_{N-1}^2 + \|\nabla F_{N-1}\|^2 + \sigma^2}} \\
&\geq \frac{Z}{2\sqrt{b_0^2 + 2Z + 2N\sigma^2/\delta'}}
\end{aligned}$$

That is equivalent to solve the following quadratic equation

$$Z^2 - \frac{8\mathcal{Q}^2}{\hat{\delta}^2}Z - \frac{4\mathcal{Q}^2}{\hat{\delta}^2}\left(b_0^2 + \frac{2N\sigma^2}{\delta'}\right) \leq 0$$

which gives

$$\begin{aligned}
Z &\leq \frac{4\mathcal{Q}^2}{\hat{\delta}^2} + \sqrt{\frac{16\mathcal{Q}^4}{\hat{\delta}^4} + \frac{4\mathcal{Q}^2}{\hat{\delta}^2}\left(b_0^2 + \frac{2N\sigma^2}{\delta'}\right)} \\
&\leq \frac{8\mathcal{Q}^2}{\hat{\delta}^2} + \frac{2\mathcal{Q}}{\hat{\delta}}\left(b_0 + \frac{\sqrt{2N}\sigma}{\sqrt{\delta'}}\right)
\end{aligned}$$

Let $\hat{\delta} = \delta' = \frac{\delta}{2}$. Replacing Z with $\sum_{k=0}^{N-1}\|\nabla F_k\|^2$ and dividing both side with N we have with probability $1-\delta$

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

A. Tables

Table 1: Statistics of data sets. DIM is the dimension of a sample

DATASET	TRAIN	TEST	CLASSES	DIM
MNIST	60,000	10,000	10	28×28
CIFAR-10	50,000	10,000	10	32×32
IMAGENET	1,281,167	50,000	1000	VARIOUS

Table 2: Architecture for five-layer neural network (LeNet)

LAYER TYPE	CHANNELS	OUT DIMENSION
5×5 CONV RELU	6	28
2×2 MAX POOL, STR.2	6	14
5×5 CONV RELU	16	10
2×2 MAX POOL, STR.2	6	5
FC RELU	N/A	120
FC RELU	N/A	84
FC RELU	N/A	10

B. Implementing the Algorithm in a Neural Network

In this section, we give the details for implementing our algorithm in a neural network. In the standard neural network architecture, the computation of each neuron consists of an elementwise nonlinearity of a linear transform of input features or output of previous layer:

$$y = \phi(\langle w, x \rangle + b), \tag{11}$$

where w is the d-dimensional weight vector, b is a scalar bias term, x,y are respectively a d-dimensional vector of input features (or output of previous layer) and the output of current neuron, $\phi(\cdot)$ denotes an elementwise nonlinearity.

For fully connected layer, the stochastic gradient G in Algorithm 1 represents the gradient of the current neuron (see the green curve, Figure 5). Thus, when implementing our algorithm in PyTorch, AdaGrad-Norm is one learning rate associated to one neuron for fully connected layer, while SGD has one learning rate for all neurons.

For convolutional layer, the stochastic gradient G in Algorithms 1 represents the gradient of each channel in the neuron. For instance, there are 6 learning rates for the first layer in the LeNet architecture (Table 1). Thus, AdaGrad Norm is one learning rate associated to one channel for convolutional layer .

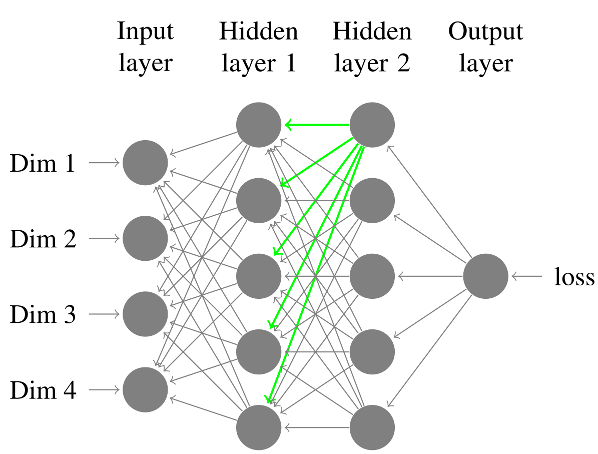

Figure 5: An example of backproporgation of two hidden layers. Green edges represent the stochastic gradient G in Algorithm 1

C. Proof of Theorem 2.2

We will use the following lemma to argue that after an initial number of steps $N = \lceil \frac{(\eta L)^2 - b_0^2}{\varepsilon} \rceil + 1$, either we have already reached a point x_k such that $\|\nabla F(x_k)\|^2 \leq \varepsilon$, or else $b_N \geq \eta L$.

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

$$b_0 > 0; \qquad b_{j+1}^2 = b_j^2 + a_j$$

has the property that after $N = \lceil \frac{(\eta L)^2 - b_0^2}{\varepsilon} \rceil + 1$ iterations, either $\min_{k=0:N-1} a_k \leq \varepsilon$, or $b_N \geq \eta L$.

Proof. If $b_0 \geq \eta L$, we are done. Else, let N be the smallest integer such that $N \geq \frac{(\eta L)^2 - b_0^2}{\varepsilon}$ and suppose $b_N < \eta L$. Then

$$(\eta L)^2 > b_N^2 = b_0^2 + \sum_{k=0}^{N-1} a_k.$$

which implies $\sum_{k=0}^{N-1} a_k \leq (\eta L)^2 - b_0^2$ and hence, for $N \geq \frac{(\eta L)^2 - b_0^2}{\varepsilon}$,

$$\min_{k=0:N-1} a_k \leq \frac{1}{N} \sum_{k=0}^{N-1} a_k \leq \frac{(\eta L)^2 - b_0^2}{N} \leq \varepsilon.$$

□

The following Lemma C.2 guarantees that the sequence $b_0, b_1, \ldots$ converges to a finite limit $b_{\max} > 0$ and that $b_{\max}$ cannot be much larger than $2L + C$ where C depends on initialization.

Lemma C.2. *Suppose $F \in C_L^1$ 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 $k \geq k_0$,*

$$b_k \leq b_{k_0-1} + 2(F(x_{k_0-1}) - F^*)/\eta \tag{12}$$

and moreover,

$$F(x_{k_0-1}) - F^* \leq F(x_0) - F^* + \frac{\eta^2 L}{2} \left(1 + 2 \log \frac{b_{k_0-1}}{b_0} \right). \tag{13}$$

Proof. Suppose $k_0 \geq 1$ is the first index such that $b_{k_0} \geq \eta L$. Then $b_j \geq \eta L$ for all $j \geq k_0$, and by Lemma 3.1, for $j \geq 0$,

$$\begin{aligned} F(x_{k_0+j}) &\leq F(x_{k_0+j-1}) - \frac{\eta}{b_{k_0+j}}(1 - \frac{\eta L}{2b_{k_0+j}})\|\nabla F(x_{k_0+j-1})\|^2 \\ &\leq F(x_{k_0+j-1}) - \frac{\eta}{2b_{k_0+j}}\|\nabla F(x_{k_0+j-1})\|^2 \\ &\leq F(x_{k_0-1}) - \sum_{\ell=0}^{j} \frac{\eta}{2b_{k_0+\ell}}\|\nabla F(x_{k_0+\ell-1})\|^2. \end{aligned} \tag{14}$$

Taking $j \to \infty$,

$$\sum_{\ell=0}^{\infty} \frac{\|\nabla F(x_{k_0+\ell-1})\|^2}{b_{k_0+\ell}} \leq 2(F(x_{k_0-1}) - F^*)/\eta.$$

Since the AdaGrad update can be equivalently written as

$$b_j = b_{j-1} + \frac{\|\nabla F(x_{j-1})\|^2}{b_j + b_{j-1}},$$

we find that

$$b_{k_0+j} \leq b_{k_0-1} + \sum_{\ell=0}^{j} \frac{\|\nabla F(x_{k_0+\ell-1})\|^2}{b_{k_0+\ell}} \leq b_{k_0-1} + 2(F(x_{k_0-1}) - F^*)/\eta \tag{15}$$

As for the upper bound of $F(x_{k_0-1})$, we invoke the Descent Lemma again, and have

$$\begin{aligned} F(x_{k_0-1}) - F(x_0) &\leq \frac{\eta^2 L}{2} \sum_{i=0}^{k_0-2} \frac{\|\nabla F(x_i)\|^2}{b_{i+1}^2} \\ &\leq \frac{\eta^2 L}{2} \sum_{i=0}^{k_0-2} \frac{(\|\nabla F(x_i)\|/b_0)^2}{\sum_{\ell=0}^{i}(\|\nabla F(x_\ell)\|/b_0)^2 + 1} \\ &\leq \frac{\eta^2 L}{2} \left(1 + \log\left(1 + \sum_{\ell=0}^{k_0-2} \frac{\|\nabla F(x_\ell)\|^2}{b_0^2}\right)\right) \\ &\leq \frac{\eta^2 L}{2} \left(1 + \log\left(\frac{b_{k_0-1}^2}{b_0^2}\right)\right) \end{aligned} \tag{16}$$

where third step uses Lemma 3.2. □

C.1. Proof of Theorem 2.2

Proof. By Lemma C.1, if $\min_{k=0:N-1} \|\nabla F(x_k)\|^2 \leq \varepsilon$ is not satisfied after $N = \lceil \frac{(\eta L)^2 - b_0^2}{\varepsilon} \rceil + 1$ steps, then there is a first index $k_0 \leq N$ such that $b_{k_0} > \eta L$. By Lemma C.2, for all $k \geq k_0$,

$$b_k \leq b_{k_0-1} + 2(F(x_{k_0-1}) - F^*)/\eta.$$

If $k_0 = 1$, it follows from (14) that

$$F(x_M) \leq F(x_0) - \frac{\eta \sum_{k=0}^{M-1} \|\nabla F(x_k)\|^2}{2(b_0 + 2(F(x_0) - F^*)/\eta)} \tag{17}$$

and thus the stated result holds straightforwardly.

Otherwise, if $k_0 > 1$, then set

$$b_{\max} = b_{k_0-1} + 2(F(x_{k_0-1}) - F^*)/\eta. \tag{18}$$

By Lemma 3.1, for any $M \geq 1$,

$$\begin{aligned} F(x_{k_0+M}) &\leq F(x_{k_0+M-1}) - \frac{\eta}{2b_{k_0+M}} \|\nabla F(x_{k_0+M-1})\|^2 \\ &\leq F(x_{k_0+M-1}) - \frac{\eta}{2b_{\max}} \|\nabla F(x_{k_0+M-1})\|^2 \\ &\leq F(x_{k_0-1}) - \frac{\eta}{2b_{\max}} \sum_{\ell=0}^{M-1} \|\nabla F(x_{k_0+\ell})\|^2. \end{aligned}$$

Thus,

$$\begin{aligned} \min_{\ell=0:M-1} \|\nabla F(x_{k_0+\ell})\|^2 &\leq \frac{1}{M} \sum_{\ell=0}^{M-1} \|\nabla F(x_{k_0+\ell})\|^2 \\ &\leq \frac{2b_{\max}(F(x_{k_0-1}) - F^*)}{\eta M} \\ &= \frac{2(\eta b_{k_0-1} + 2(F(x_{k_0-1}) - F^*))(F(x_{k_0-1}) - F^*)}{\eta^2 M} \\ &\leq \frac{4\left(F(x_{k_0-1}) - F^* + \frac{\eta b_{k_0-1}}{4}\right)^2}{\eta^2 M} \end{aligned} \tag{19}$$

By Lemma C.2, we have

$$b_{k_0-1} \leq \eta L, \quad \text{and} \quad F(x_{k_0-1}) - F^* \leq F(x_0) - F^* + \frac{\eta^2 L}{2}\left(1 + 2\log\frac{\eta L}{b_0}\right).$$

Thus, once

$$M \geq \frac{4\left((F(x_0) - F^*)/\eta + \left(\frac{3}{4} + \log\frac{\eta L}{b_0}\right)\eta L\right)^2}{\varepsilon},$$

we are assured that

$$\min_{k=0:N+M-1} \|\nabla F(x_k)\|^2 \leq \varepsilon$$

where $N \leq \frac{L^2 - b_0^2}{\varepsilon}$. □