Understanding Gradient Clipping In Incremental Gradient Methods Supplementary Materials

1 MISSING PROOFS

In this section, we give the detailed proof of the theorems on the convergence of SGD and IGC with gradient clipping. We first recall the necessary assumptions.

- **(A1)**. Assume that $f(x)$ is bounded below, that is,

$$f(x) \geq f^*, \quad \text{for any } x \in \mathbb{R}^n. \tag{1}$$

- **(A2).** (Relaxed Smoothness). There exists nonnegative scalars L_0 and L_1, such that for any $x \in \mathbb{R}^n$,

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|. \tag{2}$$

- **(A3).** All $\cos\theta_{ki} \geq 0$, and there exists some constant $c_1 > 0$ such that

$$\frac{1}{m}\sum_{i=1}^{m} \cos\theta_{ki} \geq c_1. \tag{3}$$

- **(A4).** There exists some constant $c_2 > 0$ such that

$$\frac{m_g}{M_g} \geq c_2, \tag{4}$$

  where $m_g = \min_i \|g_{ki}\|$ and $M_g = \max_i \|g_{ki}\|$.

- **(A5).** Each $f_i(x)$ also satisfies the relaxed smoothness assumption (A2), that is,

$$\|\nabla^2 f_i(x)\| \leq L_{0i} + L_{1i}\|\nabla f_i(x)\|, \quad i = 1, \cdots, m. \tag{5}$$

SGD with constant stepsize α and gradient clipping threshold η iterates as

$$x_{k+1} = x_k - \alpha \mathcal{C}(\nabla f_{i_k}(x_k); \eta), \quad k = 0, 1, \cdots, \tag{6}$$

and IGC with constant stepsize α and gradient clipping threshold η iterates as

$$x_{k,i} = x_{k,i-1} - \alpha \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta), \quad k = 0, 1, \cdots, i = 1, \cdots, m. \tag{7}$$

Write $x_k = x_{k,0} = x_{k-1,m}$, then (7) gives

$$x_{k+1} = x_k - \alpha \sum_{i=1}^{m} \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta), \quad k = 0, 1, \cdots. \tag{8}$$

1.1 Proof of Theorem 3.1

Theorem 3.1. *Assume that (A1)-(A4) hold. If*

$$\alpha \leq \frac{c_1}{4\eta L_1}, \tag{9}$$

then the iterates $\{x_k\}$ *generated by* (6) *satisfy*

$$\frac{1}{T+1}\sum_{k=0}^{T}\min\left\{\eta c_1, c_2\|\nabla f(x_k)\|\right\}\|\nabla f(x_k)\| \leq \frac{2(f(x_0)-f^*)}{(T+1)\alpha} + 5\alpha\eta^2 L_0 + \frac{\alpha\eta^3 L_1 c_1}{c_2}. \tag{10}$$

We will need the following Lemma from Zhang et al. (2020).

Lemma 1.1 (Lemma 9 in Zhang et al. (2020).)**.** *Assume that* (2) *holds, then for any* y *such that* $\|y-x\| \leq \frac{1}{L_1}$*, we have*

$$\|\nabla f(y)\| \leq 4(L_0/L_1 + \|\nabla f(x)\|). \tag{11}$$

Proof of Theorem 3.1. Recall that the SGD with constant stepsize iterates as

$$x_{k+1} = x_k - \alpha g_k := x_k - \alpha\mathcal{C}(\nabla f_{i_k}(x_k);\eta),$$

where it holds that $\|g_k\| \leq \eta$ from the definition of the clipping function. Using Taylor's theorem, we have

$$f(x_{k+1}) \leq f(x_k) - \alpha(g_k, \nabla f_k) + \frac{\|x_{k+1}-x_k\|^2}{2}\int_0^1 \|\nabla^2 f(\gamma(t))\|dt,$$

where $\gamma(t) = x_k + t(x_{k+1}-x_k)$. Using (2), (9) and (11) in Lemma 1.1, the inequality above becomes

$$f(x_{k+1}) \leq f(x_k) - \alpha(g_k, \nabla f_k) + \frac{\alpha^2\eta^2}{2}(5L_0 + 4L_1\|\nabla f_k\|).$$

Taking expectations give

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha\mathbb{E}[(g_k, \nabla f_k)] + \frac{\alpha^2\eta^2}{2}(5L_0 + 4L_1\|\nabla f_k\|) \\
&\leq f(x_k) - \alpha\min\left\{\eta c_1, c_2\|\nabla f_k\|\right\}\|\nabla f_k\| + \frac{\alpha^2\eta^2}{2}(5L_0 + 4L_1\|\nabla f_k\|) \\
&\leq f(x_k) - \alpha\min\left\{\eta c_1, c_2\|\nabla f_k\|\right\}\|\nabla f_k\| + \frac{5\alpha^2\eta^2 L_0}{2} + \frac{\alpha^2\eta^3 L_1 c_1}{2c_2} \\
&\quad + \frac{2\alpha^2\eta L_1}{c_1}\min\left\{\eta c_1, c_2\|\nabla f_k\|\right\}\|\nabla f_k\|,
\end{aligned}$$

where the last inequality follows from

$$\|\nabla f_k\| \leq \frac{\eta c_1}{4c_2} + \frac{c_2}{\eta c_1}\|\nabla f_k\|^2 \leq \frac{\eta c_1}{4c_2} + \frac{1}{\eta c_1}\min\{\eta c_1\|\nabla f_k\|, c_2\|\nabla f_k\|^2\}. \tag{12}$$

So if $\alpha \leq \frac{c_1}{4\eta L_1}$, then

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \frac{\alpha}{2}\min\left\{\eta c_1, c_2\|\nabla f_k\|\right\}\|\nabla f_k\| + \frac{5\alpha^2\eta^2 L_0}{2} + \frac{\alpha^2\eta^3 L_1 c_1}{2c_2}. \tag{13}$$

Summing from $k=0$ to T and rearranging (13) lead to

$$\frac{1}{T+1}\sum_{k=0}^{T}\min\left\{\eta c_1, c_2\|\nabla f_k\|\right\}\|\nabla f_k\| \leq \frac{2(f(x_0)-f^*)}{(T+1)\alpha} + 5\alpha\eta^2 L_0 + \frac{\alpha\eta^3 L_1 c_1}{c_2},$$

which completes the proof. □

1.2 Proof of Lemma 3.1

Lemma 3.1. *Assume that (A1)-(A5) hold. If*

$$\alpha \leq \min\left\{\frac{1}{5mL_{0i}}, \frac{1}{8m\eta L_{1i}}\right\}, \tag{14}$$

then it holds that

$$\left|\left(\sum_{i=1}^{m}\mathcal{C}(\nabla f_i(x_{k,i-1});\eta) - \mathcal{C}(\nabla f_i(x_k);\eta), \nabla f_k\right)\right| \leq \alpha\eta m^2 G\|\nabla f_k\|, \tag{15}$$

where

$$G = 5\max_{1\leq i\leq m} L_{0i} + 16\eta \max_{1\leq i\leq m} L_{1i}. \tag{16}$$

To prove Lemma 3.1, we need the following lemma.

Lemma 1.2. *If $\|z_1 - z_2\| \leq \alpha\eta C + \alpha\eta D\|z_1\|$, where C and D are positive constants, then when*

$$\alpha \leq \min\left\{\frac{1}{C}, \frac{1}{2\eta D}\right\}, \tag{17}$$

it holds that $\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| \leq \alpha\eta C + 4\alpha\eta^2 D$.

Proof. Recall that $\mathcal{C}(z;\eta) = \min\{1, \frac{\eta}{\|z\|}\}z$. We consider the following four different cases.

1. When $\|z_1\| < \eta, \|z_2\| < \eta$, $\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| = \|z_1 - z_2\| \leq \alpha\eta C + \alpha\eta D\|z_1\| \leq \alpha\eta C + \alpha\eta^2 D \leq \alpha\eta C + 4\alpha\eta^2 D$.
2. When $\|z_1\| \leq \eta, \|z_2\| \geq \eta$, it holds that $\mathcal{C}(z_1;\eta) = z_1, \mathcal{C}(z_2;\eta) = \frac{\eta}{\|z_2\|}z_2$. Denote $\tilde{z}_2 = \frac{\eta}{\|z_2\|}z_2$, and consider the function
$$h(a) = \|z_1 - a\tilde{z}_2\|^2 = \|z_1\|^2 - 2a(z_1, \tilde{z}_2) + a^2\|\tilde{z}_2\|^2, \tag{18}$$
then $h'(a) = -2(z_1, \tilde{z}_2) + 2a\|\tilde{z}_2\|^2 \geq 0$ for $a \geq 1$, as $|(z_1, \tilde{z}_2)| \leq \|z_1\|\|\tilde{z}_2\| \leq \|\tilde{z}_2\|^2$, and hence
$$\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| = \sqrt{h(1)} \leq \sqrt{h\left(\frac{\|z_2\|}{\eta}\right)} = \|z_1 - z_2\| \leq \alpha\eta C + \alpha\eta D\|z_1\| \leq \alpha\eta C + \alpha\eta^2 D \leq \alpha\eta C + 4\alpha\eta^2 D.$$
3. When $\|z_1\| \geq \eta, \|z_2\| \leq \eta$, similar as the proof in Case 2, we have
$$\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| \leq \|z_1 - z_2\| \leq \alpha\eta C + \alpha\eta D\|z_1\|.$$
Note $\eta \geq \|z_2\| \geq \|z_1\| - \|z_1 - z_2\| \geq (1 - \alpha\eta D)\|z_1\| - \alpha\eta C$, which implies that $\|z_1\| \leq \frac{1+\alpha C}{1-\alpha\eta D}\eta \leq 4\eta$ following from (17). And hence $\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| \leq \alpha\eta C + 4\alpha\eta^2 D$.
4. When $\|z_1\| > \eta, \|z_2\| > \eta$, $\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| = \left\|\frac{\eta}{\|z_1\|}z_1 - \frac{\eta}{\|z_2\|}z_2\right\| = \eta\left\|\frac{1}{\|z_1\|}z_1 - \frac{1}{\|z_2\|}z_2\right\|$. Similar as the proof in Case 2, we can show that
$$\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| \leq \frac{\eta}{\min\{\|z_1\|, \|z_2\|\}}\|z_1 - z_2\| \leq \alpha\eta C + \alpha\eta^2 D\frac{\|z_1\|}{\min\{\|z_1\|, \|z_2\|\}}.$$
If $\|z_1\| \leq \|z_2\|$, then we easily have $\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| \leq \alpha\eta C + \alpha\eta^2 D \leq \alpha\eta C + 4\alpha\eta^2 D$; while if $\|z_1\| > \|z_2\|$, it holds that
$$\|\mathcal{C}(z_1;\eta) - \mathcal{C}(z_2;\eta)\| \leq \alpha\eta C + \alpha\eta^2 D\frac{\|z_1\|}{\|z_2\|} \leq \alpha\eta C + 4\alpha\eta^2 D,$$
where the last equality uses $\frac{\|z_1\|}{\|z_2\|} \leq 4$, following from
$$\|z_1\| \leq \|z_2\| + \|z_1 - z_2\| \leq \|z_2\| + \alpha C\eta + \alpha\eta D\|z_1\| \leq \|z_2\| + \eta + \frac{1}{2}\|z_1\| \leq 2\|z_2\| + \frac{1}{2}\|z_1\|.$$

□

Now we are ready to prove Lemma 3.1.

Proof of Lemma 3.1. Noting (7), we have

$$\|\nabla f_i(x_{k,i-1}) - \nabla f_i(x_k)\| = \|\nabla^2 f_i(\gamma(t^*))\|\|x_{k,i-1} - x_{k,0}\| \leq \alpha m\eta(5L_{0i} + 4L_{1i}\|\nabla f_i(x_k)\|),$$

where $\gamma(t) = x_k + t(x_{k,i-1} - x_k)$ and t^* is some point lying in $[0, 1]$, and the inequality follows from Assumption (A5) and Lemma 1.1. Then applying Lemma 1.2, we have

$$\|\mathcal{C}(\nabla f_i(x_{k,i-1}); \eta) - \mathcal{C}(\nabla f_i(x_k)\| \leq \alpha m\eta(5L_{0i} + 16\eta L_{1i}) \leq \alpha m\eta G.$$

Hence

$$\left|\left(\sum_{i=1}^m \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta) - \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k\right)\right| \leq \sum_{i=1}^m \|\mathcal{C}(\nabla f_i(x_{k,i-1}); \eta) - \mathcal{C}(\nabla f_i(x_k)\|\|\nabla f_k\| \leq \alpha m^2\eta G\|\nabla f_k\|,$$

which completes the proof. □

1.3 Proof of Theorem 3.2

Theorem 3.2. *Assume that (A1)-(A5) hold. If*

$$\alpha \leq \min\left\{\frac{c_1}{2m(2\eta L_1 + G)}, \frac{1}{5mL_{0i}}, \frac{1}{8m\eta L_{1i}}\right\}, \tag{19}$$

then the iterates $\{x_k\}$ *generated by* (7) *and* (8) *satisfy*

$$\frac{1}{T+1}\sum_{k=0}^T \min\{\eta c_1, c_2\|\nabla f(x_k)\|\}\|\nabla f(x_k)\| \leq \frac{2(f(x_0) - f^*)}{(T+1)\alpha m} + 5\alpha m\eta^2 L_0 + \frac{\alpha m\eta^2 c_1(2\eta L_1 + G)}{2c_2}. \tag{20}$$

Proof. With $g_k = \sum_{i=1}^m \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta)$, using Taylor's theorem on (8), we have

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) - \alpha(g_k, \nabla f_k) + \frac{\|x_{k+1} - x_k\|^2}{2}\int_0^1 \|\nabla^2 f(\gamma(t))\|dt \\
&\leq f(x_k) - \alpha\left(\sum_{i=1}^m \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k\right) \\
&\qquad - \alpha\left(\sum_{i=1}^m \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta) - \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k\right) + \frac{\alpha^2 m^2\eta^2}{2}(5L_0 + 4L_1\|\nabla f_k\|) \\
&\leq f(x_k) - \alpha m\min\{\eta c_1, c_2\|\nabla f_k\|\}\|\nabla f_k\| + \frac{5\alpha^2 m^2\eta^2 L_0}{2} + \alpha^2 m^2\eta(2\eta L_1 + G)\|\nabla f_k\| \\
&\leq f(x_k) - \alpha m\min\{\eta c_1, c_2\|\nabla f_k\|\}\|\nabla f_k\| + \frac{5\alpha^2 m^2\eta^2 L_0}{2} + \frac{\alpha^2 m^2\eta^2 c_1(2\eta L_1 + G)}{4c_2} \\
&\qquad + \alpha^2\frac{m^2(2\eta L_1 + G)}{c_1}\min\{\eta c_1, c_2\|\nabla f_k\|\}\|\nabla f_k\| \\
&\leq f(x_k) - \frac{1}{2}\alpha m\min\{\eta c_1, c_2\|\nabla f_k\|\}\|\nabla f_k\| + \frac{5\alpha^2 m^2\eta^2 L_0}{2} + \frac{\alpha^2 m^2\eta^2 c_1(2\eta L_1 + G)}{4c_2},
\end{aligned}$$

where the third inequality uses Lemma 3.1, the fourth one uses (12), and the last one follows from (19). Rearranging and summing from $k = 0$ to T give arise to

$$\frac{1}{T+1}\sum_{k=0}^T \min\{\eta c_1, c_2\|\nabla f(x_k)\|\}\|\nabla f(x_k)\| \leq \frac{2(f(x_0) - f^*)}{(T+1)\alpha m} + 5\alpha m\eta^2 L_0 + \frac{\alpha m\eta^2 c_1(2\eta L_1 + G)}{2c_2},$$

which completes the proof.

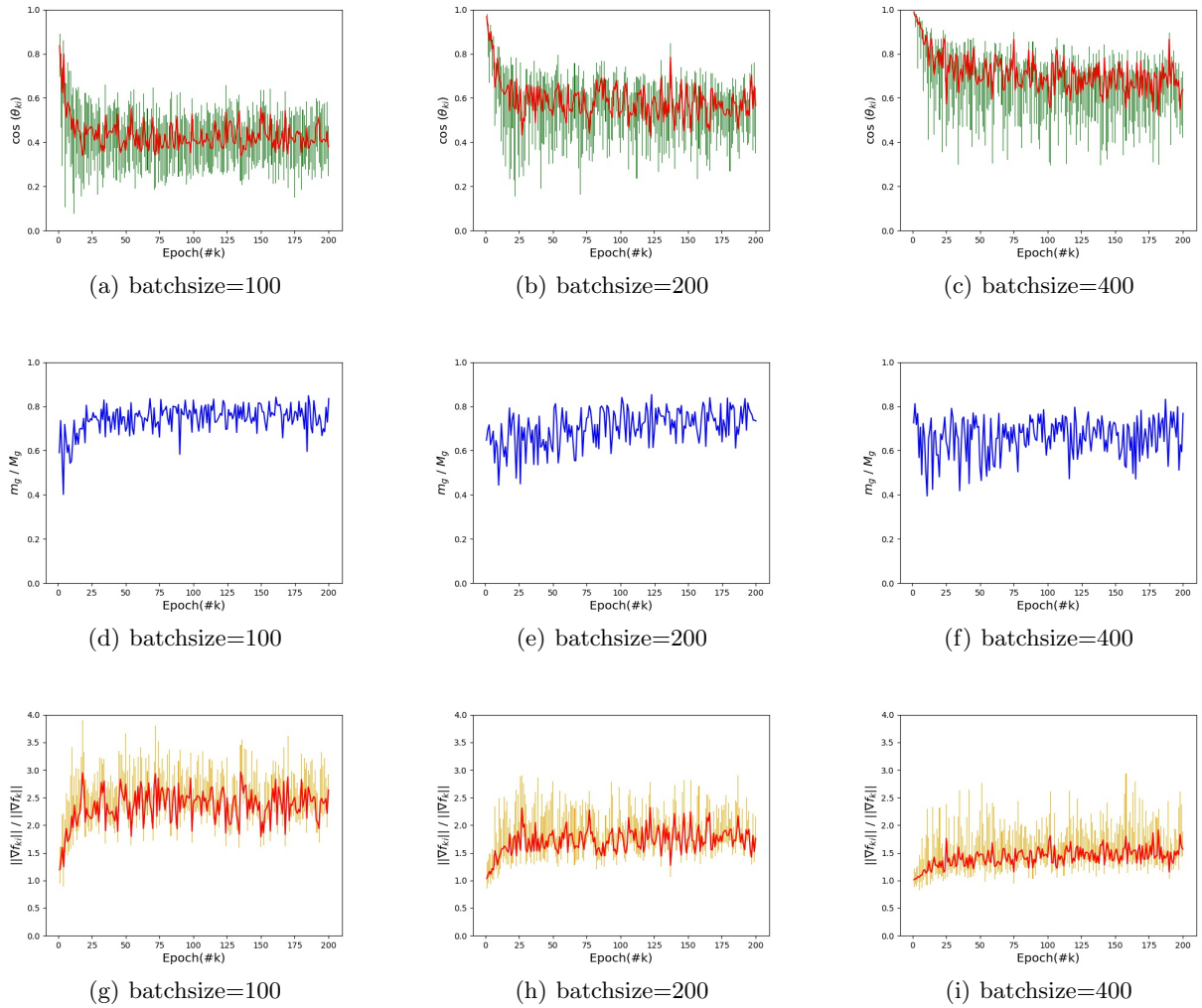
□

2 ADDITIONAL EXPERIMENTS

In this section, additional experimental results are illustrated to validate Assumptions (A3) and (A4), including neural language modelling on the popular Penn Treebank (PTB) and Word level WikiText-2 (WT2) datasets and image classification tasks on CIFAR-10 and CIFAR-100 datasets. Regarding Assumption (A3), we plot the distribution of $\cos\theta_{ki}$ with respect to the number k of epoches, where θ_{ki} is the angle between $\nabla f_i(x_k)$ and $\nabla f(x_k)$, and the mean values $\frac{1}{m}\sum_{i=1}^{m}\cos\theta_{ki}$ are shown in red. Regarding Assumption (A4), we plot m_g/M_g (i.e. $\min_i \|\nabla f_i(x_k)\|/\max_i \|\nabla f_i(x_k)\|$) with respect to the number k of epoches, and we also plot the distribution of $\|\nabla f_i(x_k)\|/\|\nabla f(x_k)\|$ with mean values in red.

2.1 Neural Language Modelling

We basically follow the settings in Merity et al. (2018), using a 1-layer LSTM language model with word embedding size of 100, hidden size of 1150 and BPTT of 50. The gradient clipping threshold and learning rate are default 0.25 and 30, respectively. We test on the widely used PTB and WT2 datasets, and the results are illustrated in the following Figures 1 and 2, respectively. The three rows, respectively, represent the distribution of $\cos\theta_{ki}$, m_g/M_g and the distribution of $\|\nabla f_i(x_k)\|/\|\nabla f(x_k)\|$ with respect to the number k of epoches (mean values in red), while the columns are results for different batchsizes.

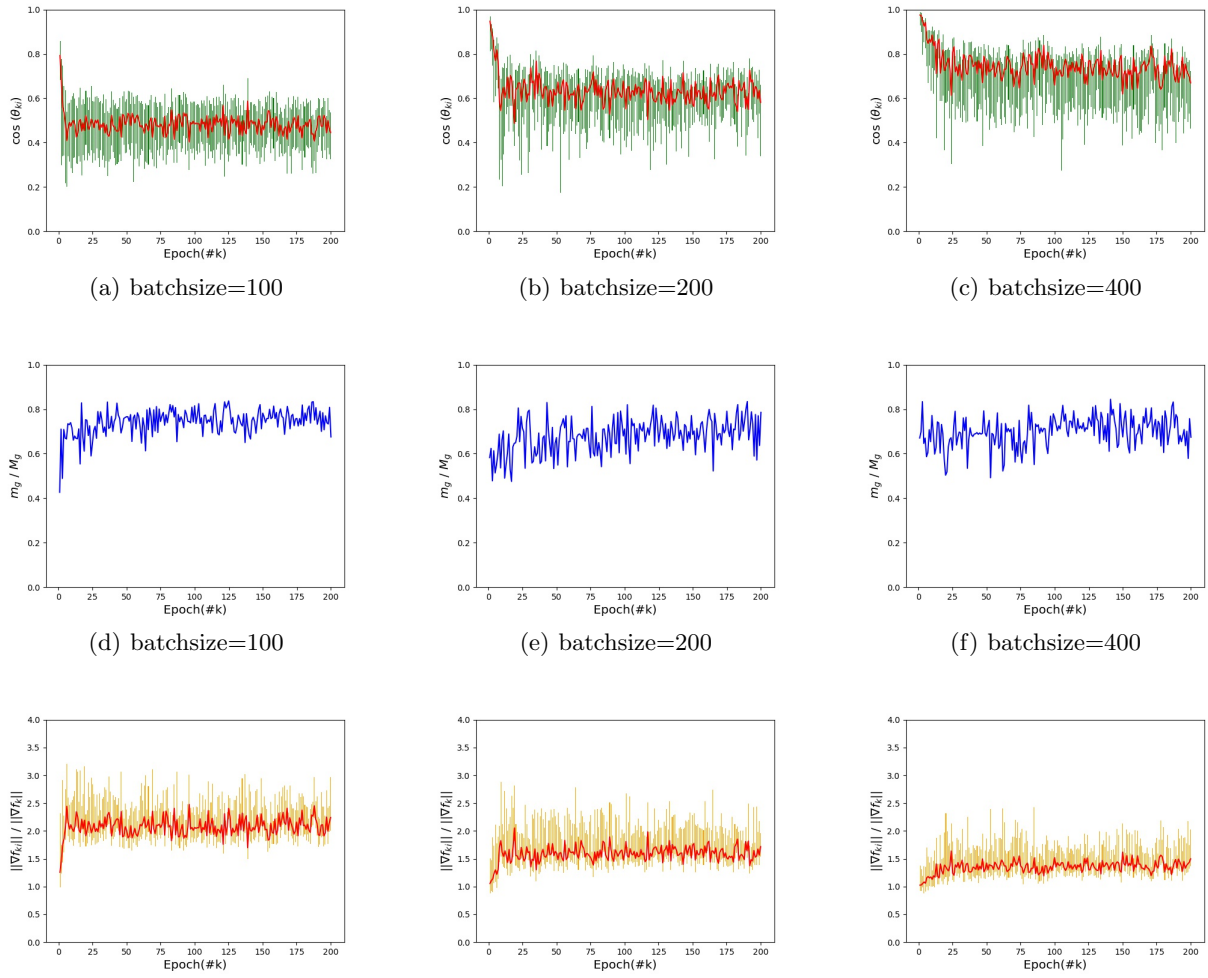


(a) batchsize=100 (b) batchsize=200 (c) batchsize=400

(d) batchsize=100 (e) batchsize=200 (f) batchsize=400

(g) batchsize=100 (h) batchsize=200 (i) batchsize=400

Figure 1: Results on the Penn Treebank (PTB) dataset: the first row for the distribution of $\cos\theta_{ki}$ (mean values in red); the second row for m_g/M_g; the third row for the distribution of $\|\nabla f_i(x_k)\|/\|\nabla f(x_k)\|$ (mean values in red); the three columns for batchsize of 100, 200 and 400, respectively.



(a) batchsize=100 (b) batchsize=200 (c) batchsize=400

(d) batchsize=100 (e) batchsize=200 (f) batchsize=400

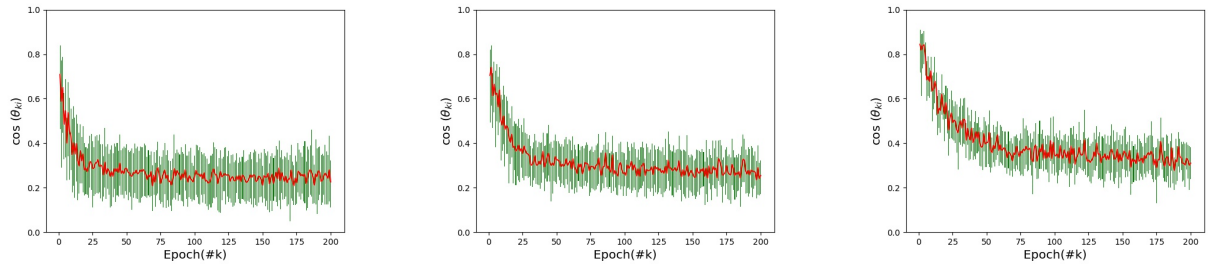
(g) batchsize=100 (h) batchsize=200 (i) batchsize=400

Figure 2: Results on the WikiText-2 (WT2) dataset: the first row for the distribution of $\cos\theta_{ki}$ (mean values in red); the second row for m_g/M_g; the third row for the distribution of $\|\nabla f_i(x_k)\|/\|\nabla f(x_k)\|$ (mean values in red); the first column for batchsize of 100; the second column for batchsize of 200; the third column for batchsize of 400.

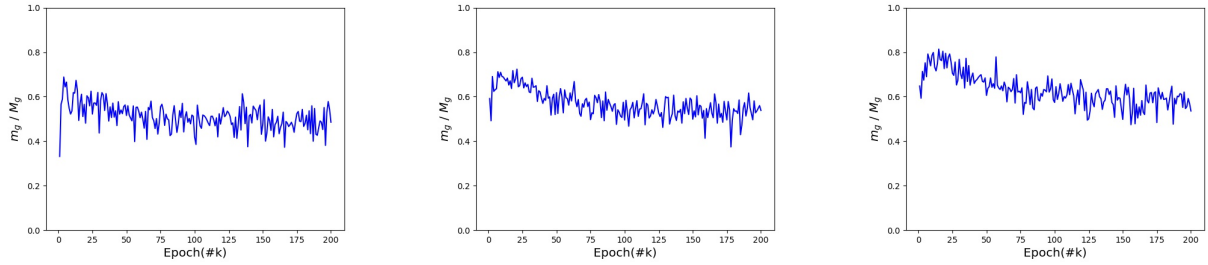
2.2 Image Classification

We run the standard ResNet18 network on the CIFAR-10 and CIFAR-100 datasets. The results are illustrated in the following Figures 3 and 4, respectively. Again, the three rows, respectively, represent the distribution of $\cos\theta_{ki}$, m_g/M_g and the distribution of $\|\nabla f_i(x_k)\|/\|\nabla f(x_k)\|$ with respect to the number k of epoches (mean values in red), while the columns are results for different batchsizes.

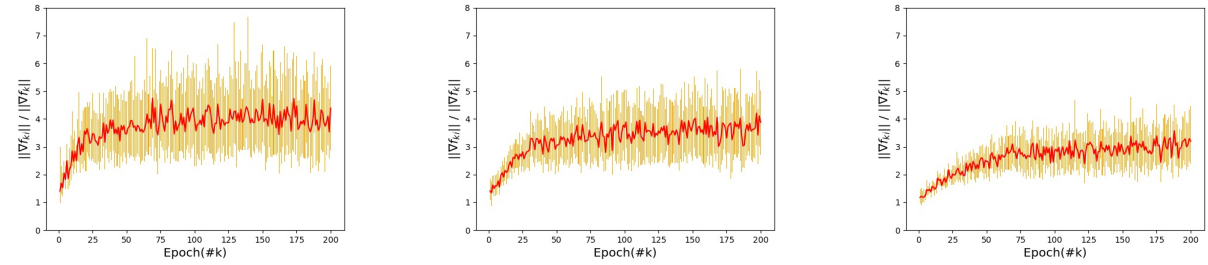
These figures show that Assumptions (A3) and (A4) are reasonable. Also note that the figures (a)-(c) for all four datasets show a tendency of increase on $\cos\theta_{ki}$ as the batchsizes grow larger, which is in line with intuition. Regarding m_g/M_g, this increasing tendency is more significant for the CIFAR-10 and CIFAR-100 datasets, as in figures (d)-(f) of Figures 3 and 4, while these values for the PTB and WT2 datasets are more similar with respect to different batchsizes.



(a) batchsize=300 (b) batchsize=500 (c) batchsize=1000

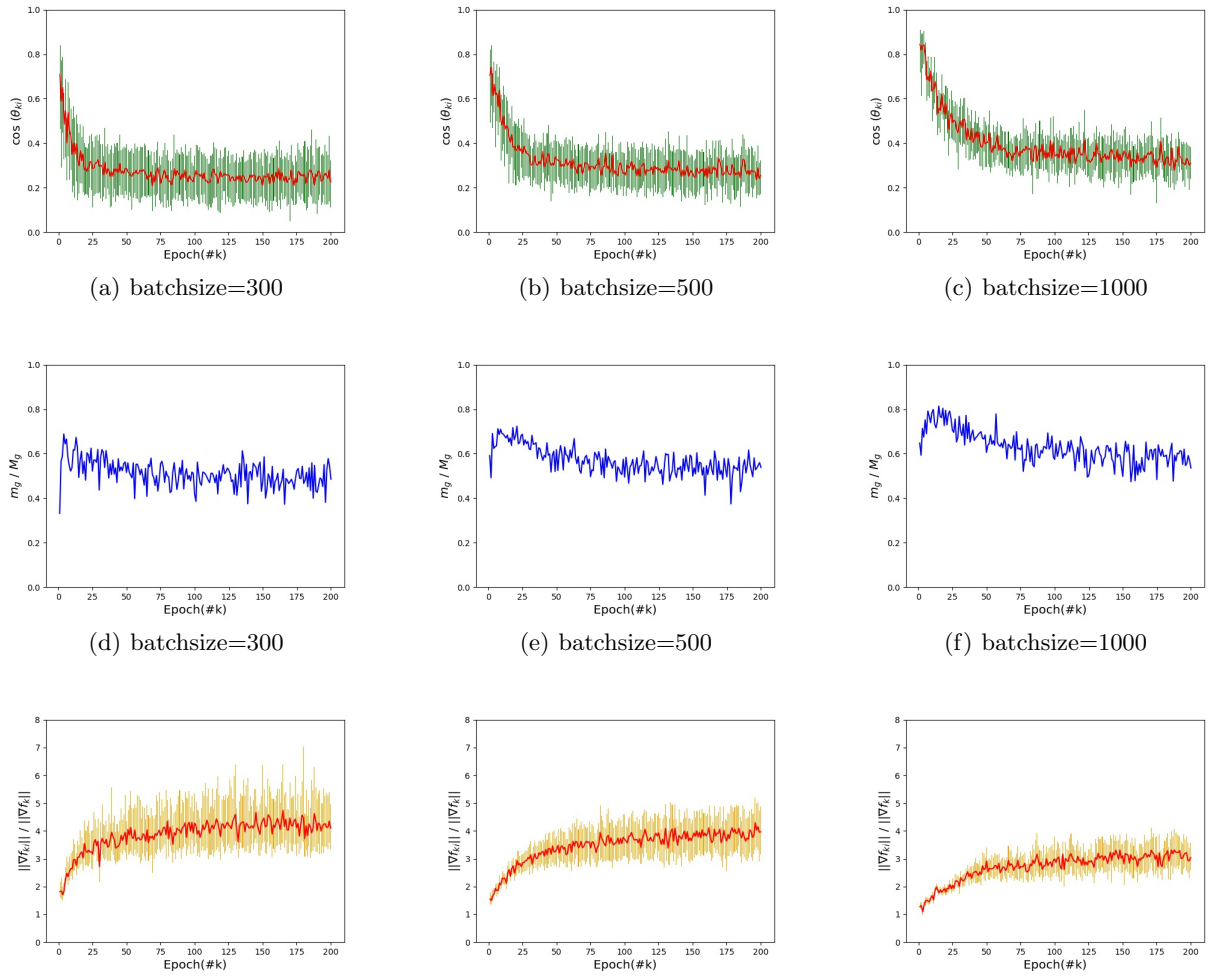


(d) batchsize=300 (e) batchsize=500 (f) batchsize=1000



(g) batchsize=300 (h) batchsize=500 (i) batchsize=1000

Figure 3: Results on the CIFAR-10 dataset: the first row for the distribution of $\cos\theta_{ki}$; the second row for m_g/M_g (mean values in red); the third row for the distribution of $\|\nabla f_i(x_k)\|/\|\nabla f(x_k)\|$ (mean values in red); the first column for batchsize of 300; the second column for batchsize of 500; the third column for batchsize of 1000.



(a) batchsize=300 (b) batchsize=500 (c) batchsize=1000

(d) batchsize=300 (e) batchsize=500 (f) batchsize=1000

(g) batchsize=300 (h) batchsize=500 (i) batchsize=1000

Figure 4: Results on the CIFAR-100 dataset: the first row for the distribution of $\cos\theta_{ki}$ (mean values in red); the second row for m_g/M_g; the third row for the distribution of $\|\nabla f_i(x_k)\|/\|\nabla f(x_k)\|$ (mean values in red); the first column for batchsize of 300; the second column for batchsize of 500; the third column for batchsize of 1000.