Understanding Gradient Clipping In Incremental Gradient Methods

Jiang Qian **Yuren Wu** **Bojin Zhuang** **Shaojun Wang** **Jing Xiao**

Ping An Technology

Abstract

We provide a theoretical analysis on how gradient clipping affects the convergence of the incremental gradient methods on minimizing an objective function that is the sum of a large number of component functions. We show that clipping on gradients of component functions leads to bias on the descent direction, which is affected by the clipping threshold, the norms of gradients of component functions, together with the angles between gradients of component functions and the full gradient. We then propose some sufficient conditions under which the increment gradient methods with gradient clipping can be shown to be convergent under the more general relaxed smoothness assumption. We also empirically observe that the angles between gradients of component functions and the full gradient generally decrease as the batchsize increases, which may help to explain why larger batchsizes generally lead to faster convergence in training deep neural networks with gradient clipping.

1 Introduction

We consider the optimization problem of the form

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are differentiable nonconvex functions, which arises in many deep learning tasks in various fields, such as computer vision, natural language processing and so on. The objective function $f(x)$ accumulates some specific losses on all training samples, which are referred to as those component functions $f_i(x)$. Practically, to fasten the speed of computation, the enormous amount of training samples are split into batches, and then each $f_i(x)$ accumulates loss over samples in each batch.

The incremental gradient method is widely used for solving (1), which can be formulated as

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k), \quad k = 0, 1, 2, \cdots, \tag{2}$$

with x_0 being the initial value, $\alpha_k > 0$ being the stepsize and index $i_k \in \{1, \cdots, m\}$. One typical choice for i_k is to uniformly draw each i_k from the set $\{1, \cdots, m\}$ with replacement, in which (2) becomes the well known stochastic gradient descent method (SGD). Another commonly used choice for i_k is the cyclic order from 1 to m periodically, in which (2) is generally reformulated as

$$\begin{aligned} x_{k,i} = x_{k,i-1} - \alpha_k \nabla f_i(x_{k,i-1}), \\ i = 1, \cdots, m, \ k = 0, 1, 2, \cdots, \end{aligned} \tag{3}$$

with $x_{0,0}$ being the initial value and $x_{k,0} = x_{k-1,m}$. For simplicity of notation, we refer to (3) as IGC later on. IGC is often used in training neural language models, where each $f_i(x)$ accumulates cross-entropy loss over words in a sub-sequence (batch) of the whole corpus. The order of $f_i(x)$ follows from the order of the words in the corpus, and hence cannot be shuffled generally. Thus the cyclic order becomes a natural choice.

To tackle with the gradient explosion problem, a widely used technique is gradient clipping (Mikolov et al., 2011a,b; Pascanu et al., 2013; Goodfellow and Bengio, 2016; Merity et al., 2018) as an intuitive approach, especially in training recurrent neural networks for neural language models, which shrinks the gradient whenever its norm exceeding some certain threshold η. We formally define the clipping function $\mathcal{C} : \mathbb{R}^n \to \mathbb{R}^n$ with threshold $\eta > 0$ as

$$\mathcal{C}(g; \eta) = \min\left\{1, \frac{\eta}{\|g\|}\right\} \cdot g. \tag{4}$$

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Then SGD with gradient clipping becomes

$$x_{k+1} = x_k - \alpha_k \mathcal{C}(\nabla f_{i_k}(x_k); \eta), \tag{5}$$

and IGC with gradient clipping becomes

$$x_{k,i} = x_{k,i-1} - \alpha_k \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta). \tag{6}$$

Recently, a theoretical explanation for the effectiveness of gradient clipping is provided in Zhang et al. (2020), under a newly proposed relaxed smoothness assumption, which is strictly weaker than the well known gradient Lipschitz smoothness condition used in literature. The Lipschitz smoothness condition assumes that the gradient smoothness is upper bounded, which may, however, not be satisfied in many deep learning tasks (Sun, 2019). The relaxed smoothness condition allow the gradient smoothness to grow with the norm of the gradient, under which Zhang et al. (2020) shows that the ordinary/stochastic gradient descent (GD/SGD) methods with gradient clipping converge arbitrarily faster than GD/SGD without gradient clipping. However, the clipping strategy in the theorem for SGD (Zhang et al., 2020, Theorem 7) is quite different from the practical strategy (4). Chen et al. (2020) theoretically study gradient clipping in SGD/private SGD, which is based on some symmetric gradient distribution assumption. The symmetric distribution assumption is suggested by empirical evaluations on private SGD, which adds a Gaussian noisy vector to the gradient on a random subsample at each iteration. However, such symmetric distribution is not observed in tasks with SGD/IGC methods, especially when large batchsizes are used.

In this paper, we propose theoretical analysis on SGD/IGC methods with gradient clipping, and provide:

- **How and when clipping works.** A key point for convergence of the SGD/IGC with gradient clipping is that $(\sum_{i=1}^m \mathcal{C}(\nabla f_i; \eta), \nabla f)$ can be lower bounded by some constant rescaling of $\|\nabla f\|$ or $\|\nabla f\|^2$. However, this might be violated with inappropriate clipping, and additional constraints should be applied to ensure this. We show that this is influenced by the clipping threshold η, the norms of ∇f_i, and the angles between each ∇f_i and ∇f. We propose some sufficient conditions theoretically, and also based on empirically observations. We also relate these conditions to assumptions in literature (Zhang et al., 2020; Chen et al., 2020).

- **Theoretical analysis of SGD with clipping.** Based on these assumptions, we prove the convergence of SGD with gradient clipping under the more general relaxed smoothness condition proposed in Zhang et al. (2020). Note that a theorem has been proposed in Zhang et al. (2020) on the convergence of SGD with clipping. However, the clipping strategy there is quite different from the practical strategy (4), and the proof there indeed implies a rather strict assumption, although not explicitly stated.

- **Theoretical analysis of IGC with clipping.** We also prove the convergence of IGC with gradient clipping. The proof for IGC exhibits more challenges as (6) introduces an additional error term on the difference between $\mathcal{C}(\nabla f_i(x_{k,i-1}))$ and $\mathcal{C}(\nabla f_i(x_{k,0}))$. We show that this error term can also be upper bounded even under the relaxed smoothness assumption and will not affect convergence. To the best of our knowledge, this is the first convergence result on IGC with gradient clipping.

- **Why larger batchsize leads to faster convergence under gradient clipping.** We also empirically observe that the angles between each ∇f_i and ∇f decrease as the batchsize increases. This observation, together with the theoretical results, may help to explain why larger batchsize generally leads to faster convergence when gradient clipping is applied at some extent.

The rest of the paper is organized as follows. Section 2 reviews some related work. Our main results are presented in Section 3: we first, in Section 3.1, study how and when gradient clipping works and propose some sufficient conditions which are intuitively reasonable and also validated by empirical experiments, and then based on these assumptions, convergence results of the SGD and IGC methods with gradient clipping are, respectively, established in Sections 3.2 and 3.3. Some conclusion remarks are drawn in Section 4.

2 Related Work

Convergence analysis of gradient based methods has always been an active research topic. There exist many classical methods and corresponding convergence results under different assumptions, e.g. see Polyak (1987); Nesterov (2004); Boyd and Vandenberghe (2008); Bertsekas (2016); Necoara et al. (2018) and references therein.

For the problem of minimizing the sum of component functions, which arises in many machine learning problems, incremental gradient method has been one of the natural choices. In each iteration step, instead of computing the full gradients (summing gra-

dients of all component functions), incremental gradient method only selects a part of component functions and uses their gradients as approximations to the full gradients. There exist several strategies for selecting the component functions. The classical stochastic gradient method (SGD) randomly but uniformly select one component function in each iteration, that is examples are picked randomly with replacement, and has been widely studied in machine learning community (Bottou, 2010; Ruder, 2016; Bottou et al., 2018). Many theoretical results on its convergence behavior exist in literature (Bertsekas, 2010; Bottou et al., 2018; Nguyen et al., 2018; Sun, 2019). Another strategy is to take the cyclic order, which is often used in training recurrent neural network based language model, as the order of the component functions follows from the order of words in corpus and hence cannot be shuffled generally. Some convergence results can be found in Bertsekas (1997); Bertsekas and Tsitsiklis (2000); Gürbüzbalaban et al. (2015a); Bertsekas (2016). Still another choice in practice (Ruder, 2016) is the cyclic order with random reshuffling, where the component functions are selected one by one in each cycle, but the order after each cycle is randomly reshuffled. This is the strategy that picks the examples with replacement. It is empirically observed in Bottou (2009) that the randomly reshuffling strategy converges much faster than the classical SGD. This phenomena is explained theoretically in HaoChen and Sra (2019); Gürbüzbalaban et al. (2015b); Jain et al. (2019); Safran and Shamir (2019).

However, neither of these results consider gradient clipping, until recently Zhang et al. (2020) provides a theoretical explanation on why gradient clipping accelerates convergence. The key contribution is a newly proposed relaxed smoothness condition, which allows the gradient smoothness to grow with the norm of the gradient. This assumption relaxes the pervasive gradient Lipschitz smoothness condition in literature. Under this relaxed smoothness condition, the ordinary GD with gradient clipping is shown to be arbitrarily faster than GD without clipping. A convergence result on SGD with gradient clipping is also proposed there. However, the clipping strategy used there is quite different from the practical clipping strategy (4), and the setting in the theorem indeed assumes that the relative difference between the unbiased stochastic gradient and the exact gradient is rather small, although it is not explicitly stated, which restricts the result in a small domain (Please refer to the last paragraph of Section 3.1 below for details). Chen et al. (2020) shows that gradient clipping might lead to bias in the gradient direction, which would prevent convergence in the worst case. Based on the symmetric gradient distribution empirically observed from tasks with private SGD, Chen et al. (2020) provide theoretical explanation why SGD/private SGD with gradient clipping are effective. However, such symmetric gradient distribution from private SGD is not observed for tasks with SGD/IGC, and the convergence result for SGD with gradient clipping is under the standard gradient Lipschitz smoothness condition, instead of the weaker relaxed smoothness condition.

3 Theoretical Analysis

In this section, we study the convergence of the SGD/IGC methods (with constant stepsize α) with gradient clipping. We first state some assumptions.

- **(A1)**. Assume that $f(x)$ is bounded below, that is,
$$f(x) \geq f^*, \quad \text{for any } x \in \mathbb{R}^n. \tag{7}$$
- **(A2).** (Relaxed Smoothness (Zhang et al., 2020)). There exists nonnegative scalars L_0 and L_1, such that for any $x \in \mathbb{R}^n$,
$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|. \tag{8}$$

When $L_1 = 0$, (8) reduces to the pervasive gradient Lipschitz smoothness condition in literature.

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

$$x_{k+1} = x_k - \alpha \mathcal{C}(\nabla f_{i_k}(x_k); \eta), \quad k = 0, 1, \cdots, \tag{9}$$

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

$$x_{k,i} = x_{k,i-1} - \alpha \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta), \quad k = 0, 1, \cdots, i = 1, \cdots, m. \tag{10}$$

Write $x_k = x_{k,0} = x_{k-1,m}$, then (10) gives

$$x_{k+1} = x_k - \alpha \sum_{i=1}^{m} \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta), \quad k = 0, 1, \cdots. \tag{11}$$

For simplicity of notation, we write (9) and (11) in the form

$$x_{k+1} = x_k - \alpha g_k, \quad k = 0, 1, \cdots, \tag{12}$$

where $g_k = \mathcal{C}(\nabla f_{i_k}(x_k); \eta)$ for SGD with gradient clipping and $g_k = \sum_{i=1}^{m} \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta)$ for IGC with gradient clipping. Using Taylor's theorem, we have

$$f(x_{k+1}) \leq f(x_k) - \alpha (g_k, \nabla f_k) + \frac{\|x_{k+1} - x_k\|^2}{2} \int_0^1 \|\nabla^2 f(\gamma(t))\| dt, \tag{13}$$

where $\nabla f_k = \nabla f(x_k)$ and $\gamma(t) = x_k + t(x_{k+1} - x_k)$. For SGD with gradient clipping,

$$\mathbb{E}[(g_k, \nabla f_k)] = \frac{1}{m}\sum_{i=1}^{m}(\mathcal{C}(\nabla f_i(x_k);\eta), \nabla f_k), \tag{14}$$

and for IGC with gradient clipping,

$$\begin{aligned}
&(g_k, \nabla f_k)\\
=&\left(\sum_{i=1}^{m}\mathcal{C}(\nabla f_i(x_{k,i-1});\eta), \nabla f_k\right)\\
=&\left(\sum_{i=1}^{m}\mathcal{C}(\nabla f_i(x_k);\eta), \nabla f_k\right)\\
&+\left(\sum_{i=1}^{m}\mathcal{C}(\nabla f_i(x_{k,i-1});\eta) - \mathcal{C}(\nabla f_i(x_k);\eta), \nabla f_k\right).
\end{aligned} \tag{15}$$

From (14) and (15), the key point to ensure convergence is whether the following condition

$$\sum_{i=1}^{m}(\mathcal{C}(\nabla f_i(x_k);\eta), \nabla f_k) \geq C_1\|\nabla f_k\| \text{ or } C_2\|\nabla f_k\|^2 \tag{16}$$

holds for some positive constants C_1, C_2, which will be discussed in the following subsection.

3.1 How and when clipping works

It should be noted that gradient clipping may not always work. Sometimes inappropriate clipping might even lead to divergence.

Example 1. Consider the example when $f_1(x) = f_2(x) = -x^2, f_3(x) = 5x^2$, and $f(x) = \frac{1}{3}\sum_{i=1}^{3} f_i(x) = x^2$, which obviously has the optimum $x^* = 0$. If we start with $x_0 = 1$ and set the clipping threshold $\eta = 2$, then $\mathcal{C}(\nabla f_1(1);2) = \mathcal{C}(\nabla f_2(1);2) = -2$ while $\mathcal{C}(\nabla f_3(1);2) = 2$, and hence $\mathbb{E}[\mathcal{C}(\nabla f_i(1);2)] = -\frac{2}{3}$, resulting in $\mathbb{E}[x_1] = x_0 - \alpha\mathbb{E}[\mathcal{C}(\nabla f_i(1);2)] = 1 + \frac{2}{3}\alpha$, even further away from the optimum x^* than x_0 since the stepsize $\alpha > 0$.

Without clipping, ∇f is dominated by ∇f_3, the correct descending direction. However, with inappropriate clipping, $\mathcal{C}(\nabla f_i)$ are all in similar magnitude, and $\mathbb{E}(\mathcal{C}(\nabla f_i))$ is then biased to the opposite direction (that is, $\mathbb{E}(\mathcal{C}(\nabla f_i), \nabla f) < 0$), leading to divergence. We will show later that, the factors on the effectiveness of clipping include the clipping threshold η, the magnitude $\|\nabla f_i\|$, and also the angles θ_{ki} between each ∇f_i and ∇f.

Denote $g_{ki} = \nabla f_i(x_k)$, then

$$\nabla f_k \triangleq \nabla f(x_k) = \frac{1}{m}\sum_{i=1}^{m} g_{ki}. \tag{17}$$

Each g_{ki} can be decomposed as

$$g_{ki} = \beta_{ki}\nabla f_k + t_{ki}, \tag{18}$$

where

$$\beta_{ki} = \frac{(g_{ki}, \nabla f_k)}{\|\nabla f_k\|^2} = \frac{\|g_{ki}\|}{\|\nabla f_k\|}\cos\theta_{ki}, \tag{19}$$

with θ_{ki} being the angle between g_{ki} and ∇f_k, and $t_{ki} \in \nabla f_k^{\perp} = \{z | (z, \nabla f_k) = 0\}$. Note that (17) and (18) imply that

$$\sum_{i=1}^{m}\beta_{ki} = m, \quad \sum_{i=1}^{m} t_{ki} = 0. \tag{20}$$

Define

$$\gamma_{ki} = \min\{1, \eta/\|g_{ki}\|\}, \tag{21}$$

then

$$\mathcal{C}(g_{ki};\eta) = \gamma_{ki}g_{ki} = \gamma_{ki}\beta_{ki}\nabla f_k + \gamma_{ki}t_{ki}, \tag{22}$$

and hence

$$\begin{aligned}
&\left(\sum_{i=1}^{m}\mathcal{C}(g_{ki};\eta), \nabla f_k\right)\\
=&\left(\sum_{i=1}^{m}(\gamma_{ki}\beta_{ki}\nabla f_k + \gamma_{ki}t_{ki}), \nabla f_k\right)\\
=&\sum_{i=1}^{m}\gamma_{ki}\beta_{ki}\|\nabla f_k\|^2,
\end{aligned} \tag{23}$$

where the first equality uses (22) and the second equality follows from the fact $t_{ki} \in \nabla f_k^{\perp}$. Hence we are to check whether $\sum_{i=1}^{m}\gamma_{ki}\beta_{ki}$ can be lower bounded by some positive scalar.

If the threshold η is sufficiently large such that $\eta \geq \|g_{ki}\|$ for all i, then $\gamma_{ki} = 1$ (implying no clipping at all), and hence $\sum_{i=1}^{m}\gamma_{ki}\beta_{ki} = \sum_{i=1}^{m}\beta_{ki} = m$, a classical result. However, when clipping does happen, determining the sign of $\sum_{i=1}^{m}\gamma_{ki}\beta_{ki}$ is rather complicated. We start with the very simple case when $m = 2$, which leads to a rather promising result.

The case when $m = 2$. The restrictions (20) on β_{ki} and t_{ki} $(i = 1, 2)$ become

$$\beta_{k1} + \beta_{k2} = 2, \quad t_{k1} = -t_{k2}.$$

The first equality implies that at least one of β_{ki} is positive. If both β_{ki} are nonnegative, $\sum_{i=1}^{2}\gamma_{ki}\beta_{ki} > 0$ holds naturally. If one of β_{ki} is negative, we can show that $\sum_{i=1}^{2}\gamma_{ki}\beta_{ki} \geq 0$ for any clipping threshold η. Without loss of generality, assume that $\beta_{k2} < 0$, then $\beta_{k1} > -\beta_{k2} > 0$, and hence

$$\begin{aligned}
\|g_{k1}\|^2 =&\beta_{k1}^2\|\nabla f_k\|^2 + \|t_{k1}\|^2\\
>&\beta_{k2}^2\|\nabla f_k\|^2 + \|t_{k2}\|^2 = \|g_{k2}\|^2.
\end{aligned}$$

We consider the following three different cases on η:

1. When $\eta \geq \|g_{k1}\| > \|g_{k2}\|$, $\gamma_{k1} = \gamma_{k2} = 1$, and then $\sum_{i=1}^{2} \gamma_{ki}\beta_{ki} = 2$.

2. When $\|g_{k1}\| > \eta \geq \|g_{k2}\|$, it holds that $\gamma_{k1} = \frac{\eta}{\sqrt{\beta_{k1}^2\|\nabla f_k\|^2 + \|t_{k1}\|^2}}$, $\gamma_{k2} = 1$, and then

$$\begin{aligned}\beta_{k1}^2\eta^2 \geq \beta_{k1}^2\|g_{k2}\|^2 &= \beta_{k1}^2(\beta_{k2}^2\|\nabla f_k\|^2 + \|t_{k2}\|^2) \\ &\geq \beta_{k2}^2(\beta_{k1}^2\|\nabla f_k\|^2 + \|t_{k1}\|^2),\end{aligned}$$

   resulting in $\gamma_{k1}^2\beta_{k1}^2 \geq \gamma_{k2}^2\beta_{k2}^2$, and hence $\sum_{i=1}^{2} \gamma_{ki}\beta_{ki} \geq 0$.

3. When $\|g_{k1}\| > \|g_{k2}\| > \eta$, it holds that $\gamma_{ki} = \frac{\eta}{\sqrt{\beta_{ki}^2\|\nabla f_k\|^2 + \|t_{ki}\|^2}}$. Then $\beta_{k1}^2(\beta_{k2}^2\|\nabla f_k\|^2 + \|t_{k2}\|^2) \geq \beta_{k2}^2(\beta_{k1}^2\|\nabla f_k\|^2 + \|t_{k1}\|^2)$ gives $\gamma_{k1}^2\beta_{k1}^2 \geq \gamma_{k2}^2\beta_{k2}^2$, and then $\sum_{i=1}^{2} \gamma_{ki}\beta_{ki} \geq 0$.

In all, when $m = 2$, $\sum_{i=1}^{m} \gamma_{ki}\beta_{ki} \geq 0$ always holds, which is rather promising, although there is still a gap compared with the desired property (16).

The above result for $m = 2$ can be applied on general $m > 2$, in the sense that

(B1). if all $g_{ki}(i = 1, \cdots, m)$ can be partitioned into pairs (say, the j-th pair consists of $g_{k,2j-1}$ and $g_{k,2j}$) such that with the decomposition similar as in (18), it holds for all j that

$$\beta_{k,2j-1} + \beta_{k,2j} > 0, \quad t_{k,2j-1} + t_{k,2j} = 0, \tag{24}$$

then after clipping with any threshold $\eta > 0$, the resulted $\sum_{i=1}^{m} \gamma_{ki}\beta_{ki}$ will always be nonnegative. Note that (B1) assumes some 'symmetric' property of g_{ki} in the sense that g_{ki} can be grouped into pairs such that the projections onto $\nabla f_k^{\perp}$ in each pair are symmetrical allocated. It can be regarded as a discrete version of the (mixture of) symmetric distribution assumption in Chen et al. (2020). We will explain this in more details later. Indeed, from the deduction above, we can see that the second condition in (24) can be relaxed to $\|t_{k,2j-1}\| = \|t_{k,2j}\|$, which is beyond the symmetric distribution assumption in Chen et al. (2020). However, the restriction $\|t_{k,2j-1}\| = \|t_{k,2j}\|$ cannot be further relaxed generally, as the deduction above heavily relies on the fact $\|t_{k1}\| = \|t_{k2}\|$. For example, if $\|t_{k1}\| \gg \|t_{k2}\|$ (resulting in $\|g_{k1}\| \gg \|g_{k2}\|$) and η is small (for example, similar as $\|g_{k2}\|$), then γ_{k1} will be rather small, while $\gamma_{k2} \approx 1$. In this case, although $\beta_{k1} + \beta_{k2} = 2 > 0$, $\gamma_{k1}\beta_{k1} + \gamma_{k2}\beta_{k2}$ would be negative if $\beta_{k2} < 0$.

More general $m > 2$**.** The case for general $m > 2$ is much more complicated. Even when all t_{ki}, projections onto $\nabla f_k^{\perp}$, are zero, under which γ_{ki} becomes $\min\{1, \eta/(|\beta_{ki}|\|\nabla f_k\|)\}$, the perturbed sum $\sum_{i=1}^{m} \gamma_{ki}\beta_{ki}$ might be negative although $\sum_{i=1}^{m} \beta_{ki} = m$. Example 1 illustrates such a counterexample for $m = 3$.

If η is small enough such that $\|g_{ki}\| \geq \eta$ for all $i = 1, \cdots, m$, then $\gamma_{ki} = \eta/\|g_{ki}\|$, and hence

$$\sum_{i=1}^{m} \gamma_{ki}\beta_{ki} = \frac{\eta}{\|\nabla f_k\|} \sum_{i=1}^{m} \cos\theta_{ki},$$

which will obviously be positive if

(B2). $\sum_{i=1}^{m} \cos\theta_{ki} > 0$.

However (B2) can only assuring $\sum_{i=1}^{m} \gamma_{ki}\beta_{ki} > 0$ for sufficiently small η. If not, $\sum_{i=1}^{m} \gamma_{ki}\beta_{ki}$ might still be negative even under (B2).

A sufficient condition. A simple sufficient condition for $\sum_{i=1}^{m} \gamma_{ki}\beta_{ki} > 0$ is $\beta_{ki} > 0$ (or equivalently $\cos\theta_{ki} > 0$) for all $i = 1, \cdots, m$, as γ_{ki} are always positive. This seems to contradict with numerical results in Chen et al. (2020). For example, Figure 3 there illustrates that the consine similarities between per-sample stochastic gradients and the true gradient is approximate symmetric around 0, with some being positive and some being negative. However, the setting there is different from ours. Instead of SGD, differentially private stochastic gradient descent (DP-SGD) is used there, where at each iteration, a gradient based on a random sample is first computed, followed by a perturbation with a noisy gradient drawn from a multivariate Gaussian distribution. In training deep neural networks, for example in computer vision and natural language processing tasks, the objective function $f(x)$ is the sum of component functions, each of which represents some kind of loss on each individual sample. The enormous samples are generally split into batches, with each $f_i(x)$ being the sum of the same loss function, now applying on the i-th batch of samples. This intuitively suggests that as the batchsize increases, $f_i(x)$ should be more similar to $f(x)$ (in the extreme case when all training samples are fed into one batch, $f_i(x)$ is exactly $f(x)$). We test ResNet18 (He et al., 2016) on CIFAR-10 dataset (image classification, Krizhevsky and Hinton (2009)) and AWD-LSTM (Merity et al., 2018) on PTB dataset (neural language model, Mikolov et al. (2010)), and find out that $\cos\theta_{ki}$ is always positive unless the batchsize is unpractically small (Results on CIFAR-10 with batchsize=1 show that around 90% of $\cos\theta_{ki}$ are positive.)

Now we assume that all $\cos\theta_{ki} \geq 0$. Combining $\beta_{ki} = (g_{ki}, \nabla f_k)/\|\nabla f_k\|^2$ and $\sum_{i=1}^{m} \beta_{ki} = m$ leads to

$$m = \sum_{i=1}^{m} \beta_{ki} = \sum_{i=1}^{m} \frac{\|g_{ki}\|}{\|\nabla f_k\|} \cos\theta_{ki}, \tag{25}$$

or equivalently,

$$\sum_{i=1}^{m} \|g_{ki}\| \cos\theta_{ki} = m\|\nabla f_k\|. \tag{26}$$

Denote $m_g = \min_i \|g_{ki}\|$ and $M_g = \max_i \|g_{ki}\|$. Then it follows from (26) that

$$m\|\nabla f_k\| \le M_g \sum_{i=1}^{m} \cos\theta_{ki},$$

and hence

$$\sum_{i=1}^{m} \cos\theta_{ki} \ge m\|\nabla f_k\|/M_g. \tag{27}$$

We can then provide lower bounds on $(\sum_{i=1}^m \mathcal{C}(g_{ki};\eta), \nabla f_k) = \sum_{i=1}^m \gamma_{ki}\beta_{ki}\|\nabla f_k\|^2$ as in (23) as follows:

1. If $\eta \ge M_g$, then $\gamma_{ki} = 1$ for all $i = 1, \cdots, m$, and

$$\left(\sum_{i=1}^{m} \mathcal{C}(g_{ki};\eta), \nabla f_k\right) = \sum_{i=1}^{m} \beta_{ki}\|\nabla f_k\|^2 = m\|\nabla f_k\|^2. \tag{28}$$

2. If $\eta \le m_g$, then $\gamma_{ki} = \eta/\|g_{ki}\|$ for all $i = 1, \cdots, m$, and

$$\left(\sum_{i=1}^{m} \mathcal{C}(g_{ki};\eta), \nabla f_k\right) = \left(\eta \sum_{i=1}^{m} \cos\theta_{ki}\right) \|\nabla f_k\|. \tag{29}$$

3. If $m_g < \eta < M_g$, then

$$\begin{aligned}
&\left(\sum_{i=1}^{m} \mathcal{C}(g_{ki};\eta), \nabla f_k\right) \\
&= \sum_{i=1}^{m} \min\{\|g_{ki}\|, \eta\}\|\nabla f_k\| \cos\theta_{ki} \\
&\ge \sum_{i=1}^{m} m_g\|\nabla f_k\| \cos\theta_{ki} \ge m\frac{m_g}{M_g}\|\nabla f_k\|^2,
\end{aligned} \tag{30}$$

where the last inequality uses (27).

Hence if we assume that

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

$$\frac{1}{m}\sum_{i=1}^{m} \cos\theta_{ki} \ge c_1. \tag{31}$$

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

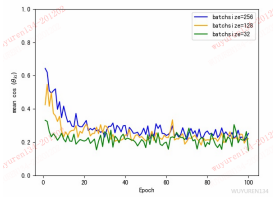
$$\frac{m_g}{M_g} \ge c_2, \tag{32}$$

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

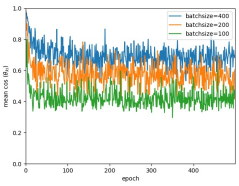
then by combining (28)-(30) we have

$$\left(\sum_{i=1}^{m} \mathcal{C}(g_{ki};\eta), \nabla f_k\right) \ge m \cdot \min\{\eta c_1, c_2\|\nabla f_k\|\} \|\nabla f_k\|. \tag{33}$$

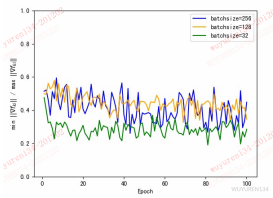
To validate Assumptions (A3) and (A4), we again test on CIFAR-10 and PTB datesets. The following Figures 1(a) and 1(b) illustrate $\frac{1}{m}\sum_{i=1}^m \cos\theta_{ki}$ in (A3) with respect to training epoch k, with different batchsizes. Figures 1(c) and 1(d) show m_g/M_g in (A4) with different batchsizes. Figures 1(a) and 1(c) are results on CIFAR-10, while Figures 1(b) and 1(d) are on PTB.



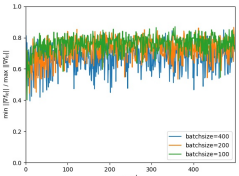
(a) A3 (CIFAR-10)



(b) A3 (PTB)



(c) A4 (CIFAR-10)



(d) A4 (PTB)

Figure 1: Results of $\frac{1}{m}\sum_{i=1}^m \cos\theta_{ki}$ in (A3) and m_g/M_g in (A4) with respect to training epoch k, with different batchsizes: the first row for $\frac{1}{m}\sum_{i=1}^m \cos\theta_{ki}$ in (A3); the second row for m_g/M_g in (A4); the first column on CIFAR-10; the second column on PTB.

These results also show that as the batchsize increases, $\cos\theta_{ki}$ tends to increase, implying that the angle between ∇f_i and ∇f tends to be smaller, which is in line with intuition. Regarding m_g/M_g, such trend is more significant on CIFAR-10.

We will show the convergence of SGD/IGC with gradient clipping under Assumptions (A1)-(A4) in the following sections.

Comparison with assumptions in literature. Theoretically verifying (16) is rather complicate,

which relies on the threshold η, norms $\|g_{ki}\|$ and angles θ_{ki}. We provide some sufficient conditions in the section as above, which can be related to existing assumptions in literature, specifically in Zhang et al. (2020); Chen et al. (2020).

The footstone in Chen et al. (2020) is the symmetric distribution of $\xi = g_{ki} - \nabla f_k$[1], assuming $p(\xi) = p(-\xi)$ with p being the probability density function. In the discrete case, this corresponds to that g_{ki} can be split into pairs such that in each pair $g_{k,2j-1} = \nabla f_k + \xi$ and $g_{k,2j} = \nabla f_k - \xi$, which satisfies (24) in (B1). Furthermore, as mentioned before, the second restriction in (24) can be relaxed to $\|t_{k,2j-1}\| = \|t_{k,2j}\|$, that is, t_{ki} is not necessarily to be symmetrically allocated.

The convergence result for SGD with gradient clipping in Zhang et al. (2020) is based on the assumption $\|g_{ki} - \nabla f_k\| \le \tau$ for some positive τ, which superficially only restricts the norms, not the angles. However, the setting in Zhang et al. (2020) [2] is

$$\alpha = \min\left\{\frac{1}{20L_0}, \frac{1}{128L_1\tau}\right\}, \tag{34}$$

$$x_{k+1} = x_k - \min\left\{\frac{1}{16L_1(\|g_{ki}\| + \tau)}, \alpha\right\} g_{ki}. \tag{35}$$

The term $\min\{\frac{1}{16L_1(\|g_{ki}\|+\tau)}, \alpha\}$ should serve as clipping, although different from the standing clipping strategy (4). Such clipping occurs only when

$$\frac{1}{16L_1(\|g_{ki}\| + \tau)} \le \alpha \le \frac{1}{128L_1\tau},$$

where $\alpha \le \frac{1}{128L_1\tau}$ follows from (34), which implies that $\tau \le \frac{1}{7}\|g_{ki}\|$. This means that if τ is large, the clipping will never happen, and (35) is just the standard SGD with constant stepsize α. In other words, (34) and (35) actually implies the assumption

$$\frac{\|g_{ki} - \nabla f_k\|}{\|g_{ki}\|} \le \frac{1}{7}, \tag{36}$$

although not explicitly stated, which is more restrictive than (A3) and (A4).

3.2 Convergence of SGD with gradient clipping

In this section, we are to show the convergence of SGD with gradient clipping (9) under Assumptions (A1)-(A4), as in the following theorem.

[1]For coincidence, we rewrite with our notations whenever necessary.

[2]Again we rewrite by using our notations, and we believe that there is a typo in the original paper.

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

$$\alpha \le \frac{c_1}{4\eta L_1}, \tag{37}$$

then the iterates $\{x_k\}$ generated by (9) satisfy

$$\begin{aligned}\frac{1}{T+1}\sum_{k=0}^{T} \min\{\eta c_1, c_2\|\nabla f(x_k)\|\}\,\|\nabla f(x_k)\| \\ \le \frac{2(f(x_0) - f^*)}{(T+1)\alpha} + 5\alpha\eta^2 L_0 + \frac{\alpha\eta^3 L_1 c_1}{c_2}.\end{aligned} \tag{38}$$

As the number of iteration T tends to infinity, the term on the right hand side of (38) converge to $5\alpha\eta^2 L_0 + \alpha\eta^3 L_1 c_1$, which tends to zero as α approaches zero. This is similar as the convergence behavior of the standard SGD method without gradient clipping (for example see Sun (2019)).

Both Zhang et al. (2020) and Chen et al. (2020) present theoretical results on the convergence of SGD with gradient clipping. However, the clipping strategy (35) in Zhang et al. (2020) is different from the practical clipping strategy (indeed, the term $\frac{1}{16L_1(\|g_{ki}\|+\tau)}$ explicitly specifies some fixed clipping threshold η and stepsize α), and the setting (34) and (35) imply the rather strict assumption (36) to activate clipping. The clipping strategy in Chen et al. (2020) takes the standard clipping form, while the convergence result is based on the symmetric distribution assumption, which is suitable for DP-SGD. For standard SGD, our empirical results show that it does not hold generally. Instead, Assumptions (A3) and (A4) are more reasonable. Furthermore, the result in Chen et al. (2020) is under the standard gradient Lipschitz smoothness assumption, which is strictly stronger than the relaxed smoothness assumption (A2) considered in Zhang et al. (2020) and this paper.

Theorem 3.1 may also help to explain why larger batchsize generally leads to faster convergence with gradient clipping at some extent, as the term on the left hand side of (38) includes c_1 and c_2, which tend to become greater as batchsizes increase. Hence $\|\nabla f(x_k)\|$ should converge faster, if regarding the right hand side term as a constant.

3.3 Convergence of IGC with gradient clipping

In this section, we consider the IGC with gradient clipping (10) or (11). Compared with (14), (15) includes an additional term

$$\left(\sum_{i=1}^{m} \mathcal{C}(\nabla f_i(x_{k,i-1}); \eta) - \mathcal{C}(\nabla f_i(x_k); \eta), \nabla f_k\right). \tag{39}$$

To bound this term, we further assume that

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

$$\|\nabla^2 f_i(x)\| \leq L_{0i} + L_{1i}\|\nabla f_i(x)\|, \quad i = 1, \cdots, m. \tag{40}$$

Then the following lemma gives an upper bound of the absolute value of the term as in (39).

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

$$\alpha \leq \min\left\{\frac{1}{5mL_{0i}}, \frac{1}{8m\eta L_{1i}}\right\}, \tag{41}$$

then it holds that

$$\begin{aligned}&\left|\left(\sum_{i=1}^{m}\mathcal{C}(\nabla f_i(x_{k,i-1});\eta) - \mathcal{C}(\nabla f_i(x_k);\eta), \nabla f_k\right)\right| \\ &\leq \alpha\eta m^2 G\|\nabla f_k\|,\end{aligned} \tag{42}$$

where

$$G = 5\max_{1\leq i\leq m} L_{0i} + 16\eta \max_{1\leq i\leq m} L_{1i}. \tag{43}$$

We are now ready to state the theorem on the convergence of the IGC method with gradient clipping as follows.

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

$$\alpha \leq \min\left\{\frac{c_1}{2m(2\eta L_1 + G)}, \frac{1}{5mL_{0i}}, \frac{1}{8m\eta L_{1i}}\right\}, \tag{44}$$

then the iterates $\{x_k\}$ *generated by* (10) *and* (11) *satisfy*

$$\begin{aligned}&\frac{1}{T+1}\sum_{k=0}^{T}\min\left\{\eta c_1, c_2\|\nabla f(x_k)\|\right\}\|\nabla f(x_k)\| \\ \leq&\frac{2(f(x_0)-f^*)}{(T+1)\alpha m} + 5\alpha m\eta^2 L_0 + \frac{\alpha m\eta^2 c_1(2\eta L_1 + G)}{2c_2}.\end{aligned} \tag{45}$$

From (45) and (38) we can see that, the IGC with gradient clipping exhibits similar convergence behavior as the SGD with gradient clipping: as T tends to infinity, the term on the right hand side of (45) converge to some $O(\alpha)$, which tends to zero as α approaches zero. To the best of our knowledge, Theorem 3.2 is the first result on the convergence of IGC with gradient clipping under the relaxed smoothness assumption, which is strictly weaker than the widely used gradient Lipschitz smoothness condition.

4 Concluding Remarks

In this paper, we manage to understand the gradient clipping in the incremental gradient method with stochastic selection (SGD) and cyclic order (IGC). As there exist examples in which gradient clipping might bias the descent direction and lead to divergence, additional assumptions should be made to ensure convergence. We show that it depends on the clipping threshold, $\|\nabla f_i\|$ and the angles between each ∇f_i and ∇f. We propose some assumptions and compare them with those in literature. Finally, based on Assumptions (A3) and (A4) induced from intuition and empirical observations, we prove the convergence of the SGD and IGC methods with gradient clipping. Adaptive gradient methods (such as AdaGrad, Adam and so on) are also popular in practice. The problem of how gradient clipping works with these methods deserves further study, which will be left as future work.

Acknowledgements

The authors are grateful to the anonymous reviewers for their valuable comments and suggestions, which have improved the quality of the paper.

References

D.P. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.

D.P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Lab. for Information and Decision Systems Report LIDS-P-2848, MIT*, 2010.

D.P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 3rd edition, 2016.

D.P. Bertsekas and J.N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. *Proceedings of the Symposium on Learning and Data Science*, 2009.

L. Bottou. Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, pages 177–186, 2010.

L. Bottou, F.E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60:223–311, 2018.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2008.

X.Y Chen, S. Wu, and M.Y. Hong. Understanding gradient clipping in private sgd: A geometric per-

spective. *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

I. Goodfellow and Y. Bengio. *Deep Learning*. MIT Press, 2016.

M. Gürbüzbalaban, A. Ozdaglary, and P. Parrilo. Convergence rate of incremental gradient and Newton methods. *arXiv preprint arXiv:1510.08562*, 2015a.

M. Gürbüzbalaban, A. Ozdaglary, and P. Parrilo. Why random reshuffling beats stochastic gradient descent. *arXiv preprint arXiv:1510.08560*, 2015b.

J.Z. HaoChen and S. Sra. Random shuffling beats SGD after finite epochs. *Proceedings of the 36th International Conference on Machine Learning*, pages 2624–2633, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

P. Jain, D. Nagaraj, and P. Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. *arXiv preprint arXiv:1903.01463*, 2019.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Technical report, Citeseer*, 2009.

S. Merity, N.S. Keskar, and R. Socher. Regularizing and optimizing LSTM language models. *International Conference on Learning Representations*, 2018.

T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. Recurrent neural network based language model. *11th Annual Conference of the International Speech Communication Association*, 2010.

T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černocký. Strategies for training large scale neural network language models. *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011a.

T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Černocký. RNNLM – recurrent neural network language modeling toolkit. *Proc. ASRU*, pages 196–201, 2011b.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publisher, The Netherlands, 2004.

L.M. Nguyen, P.H. Nguyen, M. van Dijk, P. Richtárik, K. Scheinberg, and M. Takáč. SGD and Hogwild! Convergence without the bounded gradients assumption. *Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research*, 80:3750–3758, 2018.

R. Pascanu, T. Milolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *International Conference on Machine Learning*, 2013.

B.T. Polyak. *Introduction to Optimization*. Optimization Software Inc., New York, 1987.

S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

I. Safran and O. Shamir. How good is SGD with random shuffling? *arXiv preprint arXiv:1908.00045*, 2019.

R. Sun. Optimization for deep learning: Theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

J.Z. Zhang, T.X. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *Eighth International Conference on Learning Representations (ICLR)*, 2020.