Communication-efficient Federated Learning via Quantized Clipped SGD

Ninghui Jia, Zhihao Qu, and Baoliu Ye

Hohai University, Nanjing, China
{jianinghui, quzhihao, yebl}@hhu.edu.cn

1 Proof for the Convergence Rate of QCSGD

In this part, we establish the convergence rate of Quantized Clipped SGD (QC-SGD). Key notations and their descriptions are summarized as follow:

- w: model parameter vectors
- D: the training dataset
- N: the number of workers
- η: the learning rate
- $\|\cdot\|$: the L_2 norm of vectors
- $|\cdot|$: the L_1 norm of vectors
- $F(\cdot)$: the loss function
- $\nabla F(w)$: the full gradient of loss function $F(\cdot)$
- ξ: the mini-batch of each worker's local training data
- $\mathbf{g}(w;\xi)$: the stochastic gradient respect to a mini-batch ξ

We first make the following assumptions which are widely adopted in SGD-based methods for FL and distributed machine learning [1][2]:

(1) (**L-smooth**) The objective function is Lipschitz continuous and for any $\omega_1, \omega_2 \in \mathbb{R}$, we have $\|\nabla F(\omega_1) - \nabla F(\omega_2)\| \leq L\|\omega_1 - \omega_2\|$.
(2) **Bound Value** F is bounded below by a scalar F^*, i.e., for any iteration t, $F^* \leq F(\omega_t)$. w
(3) (**Unbiased Gradient**) The stochastic gradient is unbiased for any parameter w, i.e., $\mathbb{E}_\xi[g(w,\xi)] = \nabla F(w)$.
(4) (**Bound Variance**) The variance for stochastic gradient is bounded by σ^2, i.e., for any parameter w, $\mathbb{E}_\xi[\| g(w,\xi) - \nabla F(w) \|^2] \leq \sigma^2$.

Theorem 1. *Under the assumptions (1)-(4), considering that Algorithm 1 runs with a fixed stepsize $\eta = \eta_t$ for each iteration t, when the step size satisfies that $\eta \leq \frac{1}{4LNG(1+\epsilon^2)}$, where L is Lipschitz constant, N is the number of workers, ϵ is the parameter bounding the quantization error which is the previous results from [1], and $G = \sum_{i=1}^{N} \frac{D_i^2}{D^2}$, for any integer $T > 1$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \| \nabla F(w_t) \|^2 \leq \frac{2\,|F^* - F(w_1)|}{\eta \cdot T} + 4L\eta\sigma^2 GN(1+\epsilon^2) + L\gamma^2\eta \qquad (1)$$

Proof. Since F($\cdot$) is a L-smooth objective function, so we have

$$F(w_{t+1}) - F(w_t) \leq \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \| w_{t+1} - w_t \|^2. \tag{2}$$

Due to updating rule,

$$w_{t+1} = w_t - \sum_{i=1}^{N} \frac{D_i}{D} h_t Q(\mathbf{g}(w_t, \xi_t^i), \quad where\ h_c := min\left\{ \eta_c, \frac{N\gamma\eta_c}{\|\sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)\|} \right\} \tag{3}$$

we have

$$\begin{aligned} F(w_{t+1}) - F(w_t) &\leq \left\langle \nabla F(w_t), -\sum_{i=1}^{N} \frac{D_i}{D} h_t Q(\mathbf{g}(w_t, \xi_t^i)) \right\rangle + \frac{L}{2} \| \sum_{i=1}^{N} \frac{D_i}{D} h_t Q(\mathbf{g}(w_t, \xi_t^i)) \|^2 \\ &= -h_t \left\langle \nabla F(w_t), \sum_{i=1}^{N} \frac{D_i}{D} h_t Q(\mathbf{g}(w_t, \xi_t^i)) \right\rangle + \frac{L{h_t}^2}{2} \| \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \|^2, \end{aligned} \tag{4}$$

where $\langle a, b \rangle$ means the dot product of vector a and b. For each iteration t, we consider the actual step size in the following two cases.

Case 1:$h_t = \frac{\gamma\eta_t}{\|\sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i))\|}$, so we have

$$F(w_{t+1}) - F(w_t) \leq -\frac{\gamma\eta_t}{\| \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \|} \left\langle \nabla F(w_t), \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \right\rangle + \frac{L\gamma^2\eta_t^2}{2}, \tag{5}$$

Take the expectation for both sides with respect to $\{\xi\}$ and quantization operator Q:

$$\begin{aligned} \mathbb{E}[F(w_{t+1}) - F(w_t)] &\leq -\frac{\gamma\eta_t}{\| \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \|} \left\langle \nabla F(w_t), \sum_{i=1}^{N} \frac{D_i}{D} \nabla F(w_t) \right\rangle + \frac{L\gamma^2\eta_t^2}{2} \\ &= -\frac{\gamma\eta_t}{\| \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \|} \cdot \| \nabla F(w_t) \|^2 + \frac{L\gamma^2\eta_t^2}{2} \\ &\leq \frac{L\gamma^2\eta_t^2}{2} \end{aligned} \tag{6}$$

Case 2:$h_t = \eta_t$, so we have

$$F(w_{t+1}) - F(w_t) \leq -\eta_t \left\langle \nabla F(w_t), \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \right\rangle + \frac{L\eta_t^2}{2} \| \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \|^2 \tag{7}$$

Take the expectation for both sides with respect to $\{\xi\}$ and quantization operator Q:

$$\begin{aligned}\mathbb{E}[F(w_{t+1}) - F(w_t)] &\leq -\eta_t \left\langle \nabla F(w_t), \sum_{i=1}^{N} \frac{D_i}{D} \nabla F(w_t) \right\rangle + \frac{L\eta_t^2}{2} \mathbb{E} \parallel \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \parallel^2 \\ &= -\eta_t \parallel \nabla F(w_t) \parallel^2 + \frac{L\eta_t^2}{2} \mathbb{E} \parallel \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \parallel^2,\end{aligned} \tag{8}$$

where the first inequality holds due to the unbiased stochastic gradient and the unbiased quantization, such that $\mathbb{E}[Q(g(w_t, \xi_t^i))] = \nabla F(w_t)$

Now, we have to bound $\mathbb{E} \parallel \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \parallel^2$,

$$\begin{aligned}\parallel \sum_{i=1}^{N} \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \parallel^2 &\leq N \sum_{i=1}^{N} \parallel \frac{D_i}{D} Q(\mathbf{g}(w_t, \xi_t^i)) \parallel^2 \\ &= N \sum_{i=1}^{N} \frac{D_i^2}{D^2} \parallel Q(\mathbf{g}(w_t, \xi_t^i)) \parallel^2 \\ &= N \sum_{i=1}^{N} \frac{D_i^2}{D^2} \parallel Q(\mathbf{g}(w_t, \xi_t^i)) - \mathbf{g}(w_t, \xi_t^i) + \mathbf{g}(w_t, \xi_t^i) \parallel^2 \\ &\overset{(a)}{\leq} N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (2\mathbb{E} \parallel Q(\mathbf{g}(w_t, \xi_t^i)) - \mathbf{g}(w_t, \xi_t^i) \parallel^2 + 2 \parallel \mathbf{g}(w_t, \xi_t^i) \parallel^2) \\ &\overset{(b)}{\leq} N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (2\epsilon^2 \parallel \mathbf{g}(w_t, \xi_t^i) \parallel^2 + 2 \parallel \mathbf{g}(w_t, \xi_t^i) \parallel^2) \\ &= 2N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (1+\epsilon^2) \parallel \mathbf{g}(w_t, \xi_t^i) \parallel^2 \\ &= 2N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (1+\epsilon^2) \parallel \mathbf{g}(w_t, \xi_t^i) - \nabla F(w_t) + \nabla F(w_t) \parallel^2 \\ &\overset{(c)}{\leq} 4N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (1+\epsilon^2) \parallel \mathbf{g}(w_t, \xi_t^i) - \nabla F(w_t) \parallel^2 + 4N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (1+\epsilon^2) \parallel \nabla F(w_t) \parallel^2 \\ &\overset{(d)}{\leq} 4N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (1+\epsilon^2)\sigma^2 + 4N \sum_{i=1}^{N} \frac{D_i^2}{D^2} (1+\epsilon^2) \parallel \nabla F(w_t) \parallel^2\end{aligned} \tag{9}$$

where (a) and (c) come after the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. The inequality (b) holds according to the bound of the quantization error, and (d)

follows according to the **Bound Variance** assumption. Thus,

$$\mathbb{E}[F(w_{t+1}) - F(w_t)] \leq -\eta_t \parallel \nabla F(w_t) \parallel^2 + 2L\eta_t^2 N(1+\epsilon^2)\sigma^2 \sum_{i=1}^{N} \frac{D_i^2}{D^2} + 2L\eta_t^2 N(1+\epsilon^2) \sum_{i=1}^{N} \frac{D_i^2}{D^2} \parallel \nabla F(w_t) \parallel^2 . \tag{10}$$

Let $G = \sum_{i=1}^{N} \frac{D_i^2}{D^2}$, we have

$$\mathbb{E}[F(w_{t+1}) - F(w_t)] \leq -\eta_t \parallel \nabla F(w_t) \parallel^2 + 2L\eta_t^2 NG(1+\epsilon^2)\sigma^2 + 2L\eta_t^2 N(1+\epsilon^2)G \parallel \nabla F(w_t) \parallel^2, \tag{11}$$

when $2L\eta_t^2 N(1+\epsilon^2)G \leq \frac{\eta_t}{2}$, that is $\eta_t \leq \frac{1}{4LNG(1+\epsilon^2)}$,

$$\mathbb{E}[F(w_{t+1}) - F(w_t)] \leq -\frac{\eta_t}{2} \parallel \nabla F(w_t) \parallel^2 + 2L\eta_t^2 N\sigma^2 G(1+\epsilon^2) \tag{12}$$

Jointly considering **Case 1** and **Case 2**, we have

$$\mathbb{E}[F(w_{t+1}) - F(w_t)] \leq -\frac{\eta_t}{2} \parallel \nabla F(w_t) \parallel^2 + 2L\eta_t^2 N\sigma^2 G(1+\epsilon^2) + \frac{L\gamma^2\eta_t^2}{2}. \tag{13}$$

Adding the both sides of (13) from $t = 1$ to T, we have

$$\mathbb{E}[F(w_{T+1})] - F(w_1) \leq -\frac{\eta_t}{2} \sum_{t=1}^{T} \parallel \nabla F(w_t) \parallel^2 + 2L\eta_t^2 \sigma^2 G(1+\epsilon^2)T \cdot N + \frac{L\gamma^2\eta_t^2 T}{2} \tag{14}$$

Thus,

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \parallel \nabla F(w_t) \parallel^2 &\leq \frac{2\,|F(w_{T+1}) - F(w_1)|}{\eta_t \cdot T} + 4L\eta_t \sigma^2 G(1+\epsilon^2)N + L\gamma^2\eta_t \\ &\leq \frac{2\,|F(w^*) - F(w_1)|}{\eta_t \cdot T} + 4L\eta_t \sigma^2 G(1+\epsilon^2)N + L\gamma^2\eta_t \end{aligned} \tag{15}$$

The proof of **Theorem 1** is completed.

Theorem 2. *Let $\eta = \sqrt{\frac{2|F^* - F(w_1)|}{[4L\sigma^2 GN(1+\epsilon^2) + L\gamma^2] \cdot T}}$, then for sufficient large T such that*

$$\sqrt{\frac{2\,|F^* - F(w_1)|}{[4L\sigma^2 GN(1+\epsilon^2) + L\gamma^2] \cdot T}} \leq \frac{1}{4LNG(1+\epsilon^2)} \tag{16}$$

we have $\frac{1}{T}\sum_{t=1}^{T} \parallel \nabla F(w_t) \parallel^2 \preceq O(\frac{1}{\sqrt{T}})$, where $\preceq$ denotes order inequality, i.e., less than or equal to up to a constant factor.

Proof. Let $h(\eta) = \frac{2|F^* - F(w_1)|}{\eta \cdot T} + 4L\eta\sigma^2 GN(1+\epsilon^2) + L\gamma^2\eta$, if

$$\frac{2\,|F^* - F(w_1)|}{\eta \cdot T} = 4L\eta\sigma^2 GN(1+\epsilon^2) + L\gamma^2\eta, \tag{17}$$

$h(\eta)$ reaches a minimum. According to (17), we have

$$\eta = \sqrt{\frac{2\left|F^{*} - F(w_1)\right|}{[4L\sigma^2 GN(1+\epsilon^2) + L\gamma^2] \cdot T}}. \tag{18}$$

Since $\eta \leq \frac{1}{4LNG(1+\epsilon^2)}$ in **Theorem 1**, we have

$$\sqrt{\frac{2\left|F^{*} - F(w_1)\right|}{[4L\sigma^2 GN(1+\epsilon^2) + L\gamma^2] \cdot T}} \leq \frac{1}{4LNG(1+\epsilon^2)} \tag{19}$$

Then, we have

$$T \geq \frac{32\left|F^{*} - F(w_1)\right| LN^2 G^2 (1+\epsilon^2)^2}{4\sigma^2 GN(1+\epsilon^2) + \gamma^2} \tag{20}$$

For sufficient large T,

$$\frac{1}{T}\sum_{t=1}^{T} \parallel \nabla F(w_t) \parallel^2 \leq \frac{4\sigma^2 GN(1+\epsilon^2) + \gamma^2}{16\eta LN^2 G^2 (1+\epsilon^2)^2} + 4L\eta\sigma^2 GN(1+\epsilon^2) + L\gamma^2\eta \tag{21}$$

Since (18), we have $\frac{1}{T}\sum_{t=1}^{T} \parallel \nabla F(w_t) \parallel^2 \preceq O(\frac{1}{\sqrt{T}})$

The proof of **Theorem 2** is completed.

References

1. Alistarh, D., Grubic, D., Li, J., Tomioka, R., Vojnovic, M.: QSGD: Communication-efficient SGD via Gradient Quantization and Encoding. Advances in Neural Information Processing Systems **30**, 1709–1720 (2017) 1, 1
2. Lin, Y., Han, S., Mao, H., Wang, Y., Dally, B.: Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. In: International Conference on Learning Representations (2018) 1