Proof of Quantized Federated learning

February 19, 2022

1 Assumption and Theorem

In FedAvg, the distributed optimization model across N devices is given by $\min_\theta F(\theta) = \sum_{k=1}^{N} p_k F_k(\theta) = \mathbb{E}_k[F_k(\theta)]$, where p_k is the weight of the k-th device. Suppose device k holds n_k training samples, $p_k = n_k / \sum_{i=1}^{N} n_i$. The objective function of device k can be given by $F_k(\theta) = \mathbb{E}_{x_k \sim D_k} F_k(\theta; x_k)$, where x_k is sampled from the data distribution D_k at device k. Since we are considering the non-IID data distribution across devices, we have $D_i \neq D_j$ when $i \neq j$. We apply the same stochastic quantizer Q_s for weight and gradient under different resolution Δ, where $Q_s(x, \Delta)$ is defined as

$$Q_s(x,\Delta) = \Delta \begin{cases} \left\lfloor \dfrac{x}{\Delta} \right\rfloor + 1, & p \geq \dfrac{x}{\Delta} - \left\lfloor \dfrac{x}{\Delta} \right\rfloor, \\ \left\lfloor \dfrac{x}{\Delta} \right\rfloor, & p < \dfrac{x}{\Delta} - \left\lfloor \dfrac{x}{\Delta} \right\rfloor. \end{cases} \tag{1}$$

Then the k-th device performs $E(\geq 1)$ quantized local updates before aggregation, which can be given by

$$\theta_{t+i+1}^{k} = \theta_{t+i}^{k} - \alpha_t Q_s(\nabla F_k(\theta_{t+i,q}^{k}, x_k), \Delta_g^t), i = 0, 1, \dots, E-1, \tag{2}$$

where α_t is the learning rate, $\Delta_g^t = 2^{-\delta_g^t}$ is the gradient quantization resolution, and δ_g^t is the gradient precision(bit width) in the t step. Before GEMM, the weight θ_{t+i}^k will be quantized into quantized weight $\theta_{t+i,q}^k = Q_s(\theta_{t+i}^k, \Delta_\theta^t)$ with quantization resolution $\Delta_\theta^t = 2^{-\delta_\theta^t}$, where δ_θ^t is the weight precision(bit width) in the t step. The server then aggregates the quantized local models $\theta_{t+E}^1, \dots, \theta_{t+E}^N$ to generate a new global model θ_{t+E}. Assume all devices participate in the global update, and thus we have $\theta_{t+E} := \sum_{k=1}^{N} p_k \theta_{t+E}^k$. To prove the convergence of the proposed neural quantization strategy, we make the following assumptions.

- *Assumption 1.* (L-Smooth) $F_1, \dots, F_N$ are all L-smooth, that is for all $\mathbf{v}$ and $\mathbf{w}$, $F_k(\mathbf{v}) \leq F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{L}{2} ||\mathbf{v}-\mathbf{w}||_2^2$.
- *Assumption 2.* (μ-strongly convex) $F_1, \dots, F_N$ are all L-smooth, that is for all $\mathbf{v}$ and $\mathbf{w}$, $F_k(\mathbf{v}) \geq F_k(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T \mu F_k(\mathbf{w}) + \frac{L}{2} ||\mathbf{v}-\mathbf{w}||_2^2$.
- *Assumption 3.* (bounded local gradient variance) $\mathbf{E}||\nabla F_k(\theta_t^k, x_k^t) - \nabla F_k(\theta_t^k)||^2 \leq \sigma_k^2$, for $k = 1, \dots, N$, where x_k^t is be sampled from the k-th device's local data uniformly.

- *Assumption 4.* (bounded local gradient) The expected squared norm of stochastic gradients is bounded, $\mathbf{E}||\nabla F_k(\theta_t^k, x_k^t)||^2 \leq G^2$, for $k = 1, \dots, N$.

Theorem 1 *When Assumption 1-4 hold, under static weight precision and resolution* $\delta_\theta^t = \delta_\theta$, $\Delta_\theta^t = \Delta_\theta = 2^{-\delta_\theta}$, *static gradient precision and resolution* $\delta_g^t = \delta_g$, $\Delta_g^t = \Delta_g = 2^{-\delta_g}$ *and learning rate* $\alpha_t = \frac{2}{\mu(t+\gamma)}$, *if* θ^* *is optimal,* $\mathbf{E}[F(\theta_t)] - F(\theta^*)$ *is bounded by*

$$\mathbf{E}[F(\theta_t)] - F(\theta^*) \leq \frac{A_1}{\gamma + t}\Delta_\theta^2 + \frac{A_2}{\gamma + t}\Delta_g + \frac{B}{\gamma + t}, \tag{3}$$

where d *is the dimensions of the* θ*;* $A_1 = \frac{4L^3 d}{\mu^2}$ *and* $A_2 = \frac{2(1+16(E-1)^2)\sqrt{d}LG}{\mu^2}$*;*$B = \frac{2L}{\mu^2}(C + \frac{\mu^2(\gamma+1)}{4}\mathbf{E}||\theta_1 - \theta^*||^2)$ *;* $C = \sum_{k=1}^N p_k^2\sigma_k^2 + 8(E-1)^2G^2 + 6L\Omega$ *is the federated learning term and* $\Omega = F^* - \sum_{k=1}^N p_k F_k^*$ *is difference of aggregated local minimized loss* F_k^* *and global minimized loss* F^**;* $\gamma = \max\{\frac{8L}{\mu}, E\}$*, and* $t, L, E, G, \mu, \Delta_\theta, \Delta_g, \sigma_k$ *are as defined above.*

Theorem 2 *When Assumptions 1-4 hold, under learning rate* $\alpha_t = \frac{4}{\mu(t+\gamma)}$*, dynamic weight precision and resolution,* $\delta_\theta^t = 1 - \lfloor \log_2(\mu\alpha_t) \rfloor$, $\Delta_\theta^t = 2^{-\delta_\theta^t}$ *and dynamic gradient precision and resolution,* $\delta_g^t = 2 - 2\lfloor \log_2(\mu\alpha_t) \rfloor$, $\Delta_g^t = 2^{-\delta_g^t}$*, if* θ^* *is optimal,* $\mathbf{E}[F(\theta_t)] - F(\theta^*)$ *is bounded by*

$$\mathbf{E}[F(\theta_t)] - F(\theta^*) \leq \frac{B}{\gamma + t} + \frac{16(A_1 + A_2)}{(\gamma + t)^3}, \tag{4}$$

where $\gamma = \max\{16\frac{L}{\mu}, E\}$ *and the remaining variables are as defined in Theorem 1.*

2 Proof of Theorem

2.1 Additional Notation

Set θ_t^k as the model parameter in the k-th device at the t-step and $\theta_{t,q}^k$ is its quantized version from Equation **??**. We rewrite Equation **??** as :

$$\theta_{t,q}^k = \theta_t^k - r_{k,\theta}^t \tag{5}$$

where $r_{k,\theta}^t = \theta_t^k - Q_s(\theta_t^k, \Delta_\theta^t)$ donates the weight quantization error on the t-th iteration under weight quantization resolution Δ_θ^t. We also rewrite the update of FedAvg (Equation 2) with full devices active as:

$$v_{t+1}^k = \theta_t^k - \alpha_t \nabla F_k(\theta_{t,q}^k, x_k) + \alpha_t r_{k,g}^t \tag{6}$$

$$\theta_{t+1}^k = \begin{cases} v_{t+1}^k, & t+1 \neq nE, n = 1, 2, \dots \\ \sum_{k=1}^N p_k v_{t+1}^k, & t+1 = nE, n = 1, 2, \dots. \end{cases} \tag{7}$$

where $r_{k,g}^t = \nabla F_k(\theta_{t,q}^k, x_k) - Q_s(\nabla F_k(\theta_{t,q}^k, x_k), \Delta_g^t)$ donates the gradient quantization error on the t-th iteration under gradient quantization resolution Δ_g^t. we use v_{t+1}^k as the direct result of SGD from θ_t^k. And we define $\bar{v}_t = \sum_{k=1}^N p_k v_t^k, \bar{\theta}_t = \sum_{k=1}^N p_k \theta_t^k, \bar{g}_t = \sum_{k=1}^N p_k \nabla F_k(\theta_{t,q}^k)$ and $g_t = \sum_{k=1}^N p_k \nabla F_k(\theta_{t,q}^k, x_k)$. Therefore $\bar{v}_{t+1} = \bar{\theta}_t - \alpha_t g_t + \alpha_t \sum_{k=1}^N p_k r_{k,g}^t$, $\bar{v}_{t+1} = \bar{\theta}_{t+1}$ and $\mathbf{E}[g_t] = \bar{g}_t$.

2.2 Key Lemmas

Lemma 1 *Assume Assumption 4, the gradient quantization error* $r_{k,g}^t$ *for k-th device on t-th iteration can be bounded in expectation as following:*

$$\mathbf{E}||\sum_{k=1}^N p_k r_{k,g}^t||^2 \leq \sqrt{d}G\Delta_g^t \tag{8}$$

where d *donates the dimension of* θ_t

Lemma 2 *Assume Assumption 1 and 2. If $\alpha \leq \frac{1}{4L}$, we have:*

$$\mathbf{E}||\bar{v}_{t+1}-\theta^*||^2 \leq (1-\alpha_t\mu)\mathbf{E}||\bar{\theta}_t-\theta^*||^2+6L\alpha_t^2\Omega+2\mathbf{E}\left[\sum_{k=1}^{N} p_k||\bar{\theta}_t - \theta_t^k||^2\right]+\alpha_t^2||\bar{g}_t-g_t||^2+2L^2\alpha_t^2 d(\Delta_\theta^t)^2+\alpha_t^2\mathbf{E}||\sum_{k=1}^{N} p_k r_{k,g}^t||^2 \tag{9}$$

where $\Omega = F^* - \sum_{k=1}^{N} p_k F_k^* \geq 0$

Lemma 3 *Assume Assumption 3 holds. It follows that*

$$\mathbf{E}||g_t - \bar{g}_t||^2 \leq \sum_{k=1}^{N} p_k^2\sigma_k^2 \tag{10}$$

Lemma 4 *Assume Assumption 4 holds, and α_t is non-increasing and $\alpha_t \leq 2\alpha_{t+E}$. It follows that*

$$\mathbf{E}\sum_{k=1}^{N} p_k||\bar{\theta}_t - \theta_t^k||^2 \leq 4\alpha_t^2(E-1)^2(G^2 + 2\mathbf{E}||\sum_{k=1}^{N} p_k r_{k,g}^t||^2) \tag{11}$$

2.3 The proof of Theorem 1

Let $D_t = \mathbf{E}||\bar{\theta}_t - \theta^*||$. Set static weight resolution $\Delta_\theta^t = \Delta_\theta$ and static gradient resolution $\Delta_g^t = \Delta_g$ .From Lemma 1, Lemma 2, Lemma 3 and Lemma 4, it follows that

$$D_{t+1} \leq (1-\alpha_t\mu)D_t + \alpha_t^2 A \tag{12}$$

where $A = C + 2L^2 d\Delta_\theta^2 + (1+16(E-1)^2)\sqrt{d}G\Delta_g$ and $C = \sum_{k=1}^{N} p_k^2\sigma_k^2 + 8(E-1)^2G^2 + 6L\Omega$

We set $\alpha_t = \frac{\beta}{t+\gamma}$ for $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\alpha_t \leq 2\alpha_{t+E}$ and $\alpha_1 \leq \frac{1}{4L}$. We Want to prove $D_t \leq \frac{v}{\gamma+t}$, where $v = \max\{\frac{\beta^2 A}{\beta\mu-1}, (\gamma+1)D_1\}$.

We Prove it by induction. Firstly, v ensures that it holds for $t = 1$. Assume the conclusion holds for some t, it follows that:

$$\begin{aligned} D_{t+1} &\leq (1-\alpha_t\mu)D_t + \alpha_t^2 A \\ &\leq \left(1 - \frac{\beta\mu}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 A}{(t+\gamma)^2} \\ &\leq \frac{t+\gamma-1}{(t+\gamma)^2}v + \frac{\beta^2 A}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}v \\ &\leq \frac{v}{t+\gamma+1} \end{aligned} \tag{13}$$

Then by the Assumption 1,

$$\mathbb{E}[F(\bar{\theta}_t)] - F^* \leq \frac{L}{2}D_t \leq \frac{L}{2}\frac{v}{\gamma+t} \tag{14}$$

Specifically, if we choose $\beta = \frac{2}{\mu}$, and $\gamma = \max\{8\frac{L}{\mu}, E\}$, $\alpha_t = \frac{2}{\mu(\gamma+t)}$. i.e. $\alpha_1 = \frac{2}{\mu(\gamma+1)} \leq \frac{2}{\mu\gamma} \leq \frac{1}{4L}$. And $2\alpha_{t+E} = \frac{4}{\mu(\gamma+t+E)} \geq \frac{4}{\mu(2\gamma+2t)} \geq \alpha_t$. In this case, for $t \geq 1$ we have:

$$v = \max\{\frac{\beta^2 A}{\beta\mu-1}, (\gamma+1)D_1\} \leq \frac{\beta^2 A}{\beta\mu-1} + (\gamma+1)D_1 \leq \frac{4A}{\mu^2} + (\gamma+1)D_1 \tag{15}$$

and

$$\mathbb{E}[F(\bar{\theta}_t)] - F^* \leq \frac{L}{2}\frac{v}{\gamma+t} \leq \frac{L}{2(\gamma+t)}\left(\frac{4A}{\mu^2} + (\gamma+1)D_1\right) \tag{16}$$

Thus:

$$
\begin{aligned}
\mathbb{E}[F(\bar{\theta}_t)] - F^* &\leq \frac{2L}{\gamma+t}\left(\frac{C + 2L^2 d\Delta_\theta^2 + (1+16(E-1)^2)\sqrt{d}G\Delta_g}{\mu^2} + \frac{(\gamma+1)D_1}{4}\right) \\
&\leq \frac{4L^3 d}{\mu^2(\gamma+t)}\Delta_\theta^2 + \frac{2(1+16(E-1)^2)\sqrt{d}LG}{\mu^2(\gamma+t)}\Delta_g + \frac{2L}{\mu^2(\gamma+t)}\left(C + \frac{\mu^2(\gamma+1)}{4}D_1\right) \\
&\leq \frac{A_1}{\gamma+t}\Delta_\theta^2 + \frac{A_2}{\gamma+t}\Delta_g + \frac{B_1}{\gamma+t} \qquad (17)
\end{aligned}
$$

Where $A_1 = \frac{4L^3 d}{\mu^2}$ and $A_2 = \frac{2(1+16(E-1)^2)\sqrt{d}LG}{\mu^2}$; $B_1 = \frac{2L}{\mu^2}\left(C + \frac{\mu^2(\gamma+1)}{4}D_1\right)$

2.4 The proof of Theorem 2

Let $D_t = \mathbf{E}||\bar{\theta}_t - \theta^*||$. dynamic weight precision and resolution, $\delta_\theta^t = 1 - \lfloor \log_2 \mu\alpha_t \rfloor$, $\Delta_\theta^t = 2^{-\delta_\theta^t}$ and dynamic gradient precision and resolution, $\delta_g^t = 2 - 2\lfloor \log_2 \mu\alpha_t \rfloor$, $\Delta_g^t = 2^{-\delta_g^t}$, if θ^* is optimal, $\mathbf{E}[F(\theta_t)] - F(\theta^*)$ .From Lemma 1, Lemma 2, Lemma 3 and Lemma 4, it follows that

$$
D_{t+1} \leq (1-\alpha_t\mu)D_t + \alpha_t^2 H_1(\Delta_\theta^t)^2 + \alpha_t^2 H_2 \Delta_g^t + \alpha_t^2 C \qquad (18)
$$

where $H_1 = 2L^2 d$, $H_2 = (1+16(E-1)^2)\sqrt{d}G$ and $C = \sum_{k=1}^N p_k^2\sigma_k^2 + 8(E-1)^2G^2 + 6L\Omega$

Because $\delta_\theta^t \geq 1 - log_2(\mu\alpha_t)$ and $\delta_g^t \geq 2 - 2log_2(\mu\alpha_t)$, Thus

$$
\Delta_\theta^t = 2^{-\delta_\theta^t} \leq \frac{\mu\alpha_t}{2}
$$

$$
\Delta_g^t = 2^{-\delta_g^t} \leq \frac{(\mu\alpha_t)^2}{4}
$$

Thus Equation 18 will be

$$
D_{t+1} \leq (1-\alpha_t\mu)D_t + \alpha_t^4\frac{\mu^2(H_1+H_2)}{4} + \alpha_t^2 C \qquad (19)
$$

We set $\alpha_t = \frac{\beta}{t+\gamma}$ for $\beta > \frac{3}{\mu}$ and $\gamma > 0$ such that $\alpha_t \leq 2\alpha_{t+E}$ and $\alpha_1 \leq \frac{1}{4L}$. We Want to prove $D_t \leq \frac{v}{\gamma+t} + \frac{w}{(\gamma+t)^3}$, where $v = \max\{\frac{\beta^2 C}{\beta\mu-1}, (\gamma+1)D_1\}$. $w = \frac{\beta^4\mu^2(H_1+H_2)}{4(\beta\mu-3)}$.

We Prove it by induction. Firstly, v and w ensures that it holds for $t=1$. Assume the conclusion holds for some t, it follows that:

$$
\begin{aligned}
D_{t+1} &\leq (1-\alpha_t\mu)D_t + \alpha_t^4\frac{\mu^2(H_1+H_2)}{4} + \alpha_t^2 C \\
&\leq \left(1 - \frac{\beta\mu}{t+\gamma}\right)\left(\frac{v}{t+\gamma} + \frac{w}{(\gamma+t)^3}\right) + \frac{\beta^4}{(t+\gamma)^4}\frac{\mu^2(H_1+H_2)}{4} + \frac{\beta^2}{(t+\gamma)^2}C \\
&\leq \frac{t+\gamma-1}{(t+\gamma)^2}v + \frac{t+\gamma-3}{(t+\gamma)^4}w + \left[\frac{\beta^2 C}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2}v\right] + \left[\frac{\beta^4\mu^2(H_1+H_2)}{4(t+\gamma)^4} - \frac{\beta\mu-3}{(t+\gamma)^4}w\right] \\
&\leq \frac{v}{t+\gamma+1} + \frac{w}{(t+\gamma+1)^3} \qquad (20)
\end{aligned}
$$

Specifically, if we choose $\beta = \frac{4}{\mu}$, and $\gamma = \max\{16\frac{L}{\mu}, E\}$, $\alpha_t = \frac{4}{\mu(\gamma+t)}$. i.e. $\alpha_1 = \frac{4}{\mu(\gamma+1)} \leq \frac{4}{\mu\gamma} \leq \frac{1}{4L}$. And $2\alpha_{t+E} = \frac{8}{\mu(\gamma+t+E)} \geq \frac{8}{\mu(2\gamma+2t)} \geq \alpha_t$. In this case, for $t \geq 1$ we have:

$$
v = \max\{\frac{\beta^2 C}{\beta\mu-1}, (\gamma+1)D_1\} \leq \frac{\beta^2 C}{\beta\mu-1} + (\gamma+1)D_1 \leq \frac{4C}{\mu^2} + (\gamma+1)D_1 \qquad (21)
$$

and

$$
\mathbb{E}[F(\bar{\theta}_t)] - F^* \leq \frac{L}{2}(\frac{v}{\gamma+t} + \frac{w}{(t+\gamma)^3}) \leq \frac{1}{\gamma+t}\left(\frac{2LC}{\mu^2} + \frac{(\gamma+1)LD_1}{2}\right) + \frac{1}{(\gamma+t)^3}\frac{32(H_1+H_2)L}{\mu^2} \qquad (22)
$$

Thus, if we set $A_1 = \frac{2LH_1}{\mu^2} = \frac{4L^3 d}{\mu^2}$ and $A_2 = \frac{2LH_2}{\mu^2} = \frac{2(1+16(E-1)^2)\sqrt{d}LG}{\mu^2}$; $B_2 = \frac{2L}{\mu^2}\left(C + \frac{\mu^2(\gamma+1)}{4}D_1\right)$, We will have:

$$\mathbb{E}[F(\bar{\theta}_t)] - F^* \leq \frac{B_2}{\gamma+t} + \frac{1}{(\gamma+t)^3}\left(16(A_1 + A_2)\right) \tag{23}$$