Service Delay Minimization for Federated Learning over Mobile Devices

I. Proof of Theorem

A. Additional Notation

For simplicity of notations, we denote the error of weight quantization $\boldsymbol{r}_n^k \triangleq Q_w\left(\boldsymbol{w}_n^{k+1}\right) - \boldsymbol{w}_n^{k+1}$, and the local "gradient" with weight quantization as $\widetilde{\boldsymbol{g}}_n^k \triangleq \nabla \widetilde{f}_n(\boldsymbol{w}_n^k) - \boldsymbol{r}_n^k/\eta, \forall n \in \mathcal{N}$. Since the quantization scheme Q_w we used is an unbiased scheme, $\mathbb{E}_Q\left[\boldsymbol{r}_n^k\right] = 0$.

Inspired by the iterate analysis framework in we define the following virtual sequences:

$$\boldsymbol{u}_n^{k+1} = \boldsymbol{w}_n^k - \eta \widetilde{\boldsymbol{g}}_n^k, \tag{1}$$

$$\boldsymbol{w}_n^{k+1} = \begin{cases} \boldsymbol{u}_n^{k+1}, & ((k+1) \mod H) \neq 0, \\ \boldsymbol{u}_n^{k'} - \sum_{n=1}^N p_n Q_{g,n}(\Delta_n^{k'}), & \textit{otherwise}. \end{cases} \tag{2}$$

Here, $k' = k + 1 - H$ is the last synchronization step and $\Delta_n^{k'} = \boldsymbol{u}_n^{k'} - \boldsymbol{u}_n^{k+1}$ is the differences since the last synchronization. The following short-hand notation will be found useful in the convergence analysis of the proposed FL framework:

$$\boldsymbol{u}^k = \sum_{n=1}^N p_n \boldsymbol{u}_n^k, \quad \boldsymbol{w}^k = \sum_{n=1}^N p_n \boldsymbol{w}_n^k, \tag{3}$$

$$\widetilde{\boldsymbol{g}}^k = \sum_{n=1}^N p_n \widetilde{\boldsymbol{g}}_n^k, \quad \boldsymbol{g}^k = \sum_{n=1}^N p_n \nabla F_n(\boldsymbol{w}_n^k). \tag{4}$$

Thus, $\boldsymbol{u}^{k+1} = \boldsymbol{w}^k - \eta \widetilde{\boldsymbol{g}}^k$. Note that we can only obtain $\boldsymbol{w}^{k+1}$ when $((k+1) \mod H) = 0$. Further, due to the unbiased gradient quantization scheme, Q_g, no matter whether $((k+1) \mod H) = 0$ or $((k+1) \mod H) \neq 0$, we always have $\mathbb{E}[\mathbb{E}_{Q_g}[\boldsymbol{w}^{k+1}]] = \mathbb{E}[\boldsymbol{u}^{k+1}]$.

B. Key Lemmas

Now, we give four important lemmas to convey our proof.

Lemma 1 (Bounding the weight quantization error [1]).

$$\mathbb{E}_{Q_w}\left[\left\|\boldsymbol{r}_n^k\right\|_2^2\right] \leq \eta\sqrt{d}\delta_{w,n}\tau. \tag{5}$$

Lemma 2. *According to the proposed algorithm the expected inner product between stochastic gradient and full batch gradient can be bounded with:*

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_Q\left[\langle \nabla F(\boldsymbol{w}^k), \boldsymbol{u}^{k+1} - \boldsymbol{w}^k\rangle\right]\right] \\
&\leq -\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^k)\right\|_2^2\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|_2^2\right] \\
&\quad + \frac{\eta L^2}{2}\sum_{n=1}^N p_n^2 \mathbb{E}\left[\mathbb{E}_Q\left[\left\|\boldsymbol{w}^k - \boldsymbol{w}_n^k\right\|_2^2\right]\right]
\end{aligned} \tag{6}$$

Proof:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_Q\left[\langle \nabla F(\boldsymbol{w}^k), \boldsymbol{u}^{k+1} - \boldsymbol{w}^k\rangle\right]\right] \\
&= -\eta\mathbb{E}\left[\left\langle \nabla F(\boldsymbol{w}^k), \mathbb{E}_Q[\widetilde{\boldsymbol{g}}^k]\right\rangle\right] \\
&= -\eta\mathbb{E}\left[\left\langle \nabla F(\boldsymbol{w}^k), \sum_{n=1}^N p_n \nabla F_n(\boldsymbol{w}_n^k)\right\rangle\right] \\
&\overset{(a)}{=} -\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^k)\right\|_2^2\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|_2^2\right] \\
&\quad + \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^k) - \boldsymbol{g}^k\right\|_2^2\right] \\
&\overset{(b)}{\leq} -\frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{w}^k)\right\|_2^2\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|\boldsymbol{g}^k\right\|_2^2\right] \\
&\quad + \frac{\eta L^2}{2} N \sum_{n=1}^N p_n^2 \mathbb{E}\left[\mathbb{E}_Q\left[\left\|\boldsymbol{w}^k - \boldsymbol{w}_n^k\right\|_2^2\right]\right]
\end{aligned} \tag{7}$$

where (a) is due to $-2\langle a, b\rangle = ||a||^2 + ||b||^2 - ||a-b||^2$ and $\mathbb{E}[\widetilde{\boldsymbol{g}}_n^k] = \nabla F_n(\boldsymbol{w}_n^k)$, and (b) follows from L-smoothness assumption. The proof is completed. ■

Lemma 3 (Bounding the divergence)**.**

$$\begin{aligned}
&\sum_{k=0}^{K-1}\sum_{n=1}^N p_n^2 \mathbb{E}\left[\mathbb{E}_Q\left[\left\|\boldsymbol{w}^k - \boldsymbol{w}_n^k\right\|_2^2\right]\right] \\
&\leq \frac{\eta^2 pKH\sigma^2/M + 3\eta^2 pKH^2G^2 + \eta KH\sqrt{d}\bar{\delta}_w\tau}{1 - 3\eta^2 pL^2H^2} \\
&\quad + \frac{3\eta^2 pH^2}{1 - 3\eta^2 pL^2H^2}\sum_{k=0}^{K-1}\left\|\nabla F(\boldsymbol{w}^k)\right\|_2^2,
\end{aligned} \tag{8}$$

where $p = \sum_{n=1}^N p_n^2$ *and* $\bar{\delta}_w = \sum_{n=1}^N p_n^2 \delta_{w,n}$.

Proof: Recalling that at the synchronization step where $(k' \mod H = 0)$, $\boldsymbol{w}_n^{k'} = \boldsymbol{w}^{k'}$ for all $n \in \mathcal{N}$. Therefore, for any $k \geq 0$, such that $k' \leq k \leq k' + H$, we get,

$$\begin{aligned}
A1_k &:= \sum_{n=1}^N p_n^2 \mathbb{E}\left[\mathbb{E}_Q\left[\left\|\boldsymbol{w}^k - \boldsymbol{w}_n^k\right\|_2^2\right]\right] \\
&= \sum_{n=1}^N p_n^2 \mathbb{E}\left[\mathbb{E}_Q\left[\left\|(\boldsymbol{w}^k - \boldsymbol{w}^{k'}) - (\boldsymbol{w}_n^k - \boldsymbol{w}^{k'})\right\|_2^2\right]\right] \\
&\overset{(a)}{\leq} \sum_{n=1}^N p_n^2 \mathbb{E}\left[\mathbb{E}_Q\left[\left\|\boldsymbol{w}_n^k - \boldsymbol{w}_n^{k'}\right\|_2^2\right]\right] \\
&= \sum_{n=1}^N p_n^2 \mathbb{E}\left[\left\|\sum_{i=k'}^k \left(\eta\nabla\widetilde{f}_n(\boldsymbol{w}_n^i) - \boldsymbol{r}_n^i\right)\right\|_2^2\right]
\end{aligned}$$

$$= \sum_{n=1}^{N} p_n^2 \left(\eta^2 \mathbb{E}\left[\left\| \sum_{i=k'}^{k} \nabla \widetilde{f}_n(\boldsymbol{w}_n^i) \right\|_2^2 \right] + \sum_{i=k'}^{k} \mathbb{E}_Q\left[||\boldsymbol{r}_n^i||_2^2 \right] \right)$$
$$\leq \eta^2 \sum_{n=1}^{N} p_n^2 \mathbb{E}\left[\left\| \sum_{i=k'}^{k'_H} \nabla \widetilde{f}_n(\boldsymbol{w}_n^i) \right\|_2^2 \right] + H\eta\sqrt{d} \sum_{n=1}^{N} p_n^2 \delta_{w,n} \tau, \tag{9}$$

where $k'_H = k' + H - 1$, (a) results from $\sum_{n=1}^{N} p_n(\boldsymbol{w}_n^k - \boldsymbol{w}_n^{k'}) = \boldsymbol{w}^k - \boldsymbol{w}^{k'}$, $\boldsymbol{w}_n^{k'} = \boldsymbol{w}^{r'}$, and $\mathbb{E}\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\|_2^2 \leq \mathbb{E}\|\boldsymbol{x}\|_2^2$. The last inequality follows Lemma 1.

We generalize the result from [2] to upper-bound the first term in RHS of (9), (see Theorem 3 and its proof in appendix [2] for the special case of $p_n = \frac{1}{N}$):

$$\eta^2 \sum_{n=1}^{N} p_n^2 \mathbb{E}\left[\left\| \sum_{i=k'}^{k'_H} (\nabla \widetilde{f}_n(\boldsymbol{w}_n^i) - \nabla F_n(\boldsymbol{w}_n^i) + \nabla F_n(\boldsymbol{w}_n^i)) \right\|_2^2 \right]$$
$$\leq \eta^2 pH \frac{\sigma^2}{M} + 3\eta^2 pH^2 G^2 + 3\eta^2 L^2 pH \sum_{i=k'}^{k'_H} A1_i$$
$$+ 3\eta^2 pH \sum_{i=k'}^{k'_H} \left\| \nabla F(\boldsymbol{w}^i) \right\|_2^2. \tag{10}$$

where $p = \sum_{n=1}^{N} p_n^2$.

It follows that

$$\sum_{i=0}^{K-1} A1_i \leq \eta^2 pKH \frac{\sigma^2}{M} + 3\eta^2 pKH^2 G^2 + 3\eta^2 pL^2 H^2 \sum_{i=0}^{K-1} A1_i$$
$$+ 3\eta^2 H^2 p \sum_{i=0}^{K-1} \left\| \nabla F(\boldsymbol{w}^i) \right\|_2^2 + \eta KH\sqrt{d}\bar{\delta}_w \tau. \tag{11}$$

where $\bar{\delta}_w = \sum_{n=1}^{N} p_n^2 \delta_{w,n}$.

Suppose $1 - 3\eta^2 L^2 H^2 \geq 0$, we have

$$\sum_{i=0}^{K-1} A1_i \leq \frac{\eta^2 pKH\sigma^2/M + 3\eta^2 pKH^2 G^2 + \eta KH\sqrt{d}\bar{\delta}_w \tau}{1 - 3\eta^2 pL^2 H^2}$$
$$+ \frac{3\eta^2 pH^2}{1 - 3\eta^2 pL^2 H^2} \sum_{i=0}^{K-1} \left\| \nabla F(\boldsymbol{w}^i) \right\|_2^2, \tag{12}$$

and the proof is completed. ∎

Lemma 4. *According to the proposed algorithm the expected inner product between stochastic gradient and full batch gradient can be bounded with:*

$$\frac{L}{2} \mathbb{E}\left[\mathbb{E}_Q\left[\left\| \boldsymbol{w}^{k+1} - \boldsymbol{w}^k \right\|_2^2 \right] \right]$$
$$\leq \frac{\eta^2 L}{2} \sum_{n=1}^{N} p_n^2 (\delta_{g,n} H + 1) \frac{\sigma^2}{M} + \frac{3\eta^2 L}{2} \bar{\delta}_g H^2 G^2$$
$$+ \frac{3\eta^2 L^3}{2} \bar{\delta}_g H \sum_{i=k'}^{k'_H} A1_i + \frac{3\eta^2 L}{2} H \bar{\delta}_g \sum_{i=k'}^{k'_H} \left\| \nabla F(\boldsymbol{w}^i) \right\|_2^2$$
$$+ \frac{\eta^2 L}{2} \mathbb{E}\left[\left\| \boldsymbol{g}^k \right\|_2^2 \right] + \frac{\eta L}{2} \sqrt{d}\tau \sum_{n=1}^{N} p_n^2 (\delta_{g,n} + 1) \delta_{w,n}, \tag{13}$$

where $\bar{\delta}_g = \sum_{n=1}^{N} p_n^2 \delta_{g,n}$.

Proof:

$$\frac{L}{2} \mathbb{E}\left[\mathbb{E}_Q\left[\left\| \boldsymbol{w}^{k+1} - \boldsymbol{w}^k \right\|_2^2 \right] \right]$$
$$= \frac{L}{2} \mathbb{E}\left[\mathbb{E}_Q\left[\left\| \boldsymbol{w}^{k+1} - \boldsymbol{u}^{k+1} + \boldsymbol{u}^{k+1} - \boldsymbol{w}^k \right\|_2^2 \right] \right]$$
$$= \frac{L}{2} \mathbb{E}\left[\mathbb{E}_Q\left[\left\| \sum_{n=1}^{N} p_n (Q(\widetilde{\Delta}_n^k) - \widetilde{\Delta}_n^k) \right\|_2^2 \right] + \left\| \eta \widetilde{\boldsymbol{g}}^k \right\|_2^2 \right]$$
$$\overset{(a)}{\leq} \frac{L}{2} \sum_{n=1}^{N} p_n^2 \delta_{g,n} \mathbb{E}\left[\left\| \boldsymbol{w}_n^k - \boldsymbol{w}^{k'} \right\|_2^2 \right] + \frac{\eta^2 L}{2} \mathbb{E}\left[\left\| \widetilde{\boldsymbol{g}}^k \right\|_2^2 \right]$$
$$\leq \frac{L}{2} \sum_{n=1}^{N} p_n^2 \delta_{g,n} \left(\eta^2 H \frac{\sigma^2}{M} + 3\eta^2 H^2 G^2 \right)$$
$$+ \frac{3\eta^2 L}{2} H \bar{\delta}_g \sum_{i=k'}^{k'_H} \left\| \nabla F(\boldsymbol{w}^i) \right\|_2^2 + \frac{\eta L}{2} \sqrt{d}\tau \sum_{n=1}^{N} p_n^2 \delta_{w,n}$$
$$+ \frac{3\eta^2 \bar{\delta}_g L}{2} H \sum_{i=k'}^{k'_H} L^2 A1_i + \frac{\eta L}{2} \sqrt{d}\tau \sum_{n=1}^{N} p_n^2 \delta_{g,n} \delta_{w,n}$$
$$+ \frac{\eta^2 L}{2} \left(\sum_{n=1}^{N} p_n^2 \frac{\sigma^2}{M} + \mathbb{E}\left[\left\| \sum_{n=1}^{N} p_n \nabla F_n(\boldsymbol{w}_n^k) \right\|_2^2 \right] \right)$$
$$\leq \frac{\eta^2 L}{2} \sum_{n=1}^{N} p_n^2 (\delta_{g,n} H + 1) \frac{\sigma^2}{M} + \frac{3\eta^2 L}{2} \bar{\delta}_g H^2 G^2$$
$$+ \frac{3\eta^2 L^3}{2} \bar{\delta}_g H \sum_{i=k'}^{k'_H} A1_i + \frac{3\eta^2 L}{2} H \bar{\delta}_g \sum_{i=k'}^{k'_H} \left\| \nabla F(\boldsymbol{w}^i) \right\|_2^2$$
$$+ \frac{\eta^2 L}{2} \mathbb{E}\left[\left\| \boldsymbol{g}^k \right\|_2^2 \right] + \frac{\eta L}{2} \sqrt{d}\tau \sum_{n=1}^{N} p_n^2 (\delta_{g,n} + 1) \delta_{w,n}, \tag{14}$$

where $\bar{\delta}_g = \sum_{n=1}^{N} p_n^2 \delta_{g,n}$. (a) holds due to the definition of unbiased stochastic quantization. (b) follows the similar proof as (9) and (10). The proof is completed. ∎

C. Main Results

Under the L-smooth assumption of F, we have,

$$\mathbb{E}\left[F(\boldsymbol{w}^{k+1}) - F(\boldsymbol{w}^k) \right]$$
$$\leq \mathbb{E}\left[\langle \nabla F(\boldsymbol{w}^k), \boldsymbol{w}^{k+1} - \boldsymbol{w}^k \rangle \right] + \frac{L}{2} \mathbb{E}\left[\left\| \boldsymbol{w}^{k+1} - \boldsymbol{w}^k \right\|_2^2 \right]$$
$$= \mathbb{E}\left[\langle \nabla F(\boldsymbol{w}^k), \boldsymbol{w}^{k+1} - \boldsymbol{u}^{k+1} + \boldsymbol{u}^{k+1} - \boldsymbol{w}^k \rangle \right]$$
$$+ \frac{L}{2} \mathbb{E}\left[\left\| \boldsymbol{w}^{k+1} - \boldsymbol{w}^k \right\|_2^2 \right]$$
$$\overset{(a)}{\leq} \mathbb{E}\left[\langle \nabla F(\boldsymbol{w}^k), \boldsymbol{u}^{k+1} - \boldsymbol{w}^k \rangle + \frac{L}{2} \left\| \boldsymbol{w}^{k+1} - \boldsymbol{w}^k \right\|_2^2 \right] \tag{15}$$

We use Lemma 1, 2 and 4 to upper bound the RHS of (15) and set $\eta L \leq 1$, which gets,

$$\begin{aligned}
&\mathbb{E}\left[F(\boldsymbol{w}^{k+1}) - F(\boldsymbol{w}^{k})\right] \\
&\leq -\frac{\eta}{2}\mathbb{E}\left[\left\|\triangledown F(\boldsymbol{w}^{k})\right\|_2^2\right] + \frac{\eta L^2}{2}A1_k + \frac{3\eta^2 L^3}{2}H\bar{\delta}_g\sum_{i=k'}^{k'_H} A1_i \\
&\quad + \frac{\eta^2 L}{2}\sum_{n=1}^{N} p_n^2(\delta_{g,n}H+1)\frac{\sigma^2}{M} + \frac{3\eta^2 L}{2}\bar{\delta}_g H^2 G^2 \\
&\quad + \frac{3\eta^2 L}{2}H\bar{\delta}_g\sum_{i=k'}^{k'_H}\mathbb{E}\left[\left\|\triangledown F(\boldsymbol{w}^{i})\right\|_2^2\right] \\
&\quad + \frac{\eta L}{2}\sqrt{d}\tau\sum_{n=1}^{N} p_n^2(\delta_{g,n}+1)\delta_{w,n}. \qquad (16)
\end{aligned}$$

Summing up for all K training iterations and rearranging the terms gives,

$$\begin{aligned}
&\mathbb{E}\left[F(\boldsymbol{w}^{K}) - F(\boldsymbol{w}^{0})\right] \\
&\leq -\frac{\eta C_1}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\triangledown F(\boldsymbol{w}^{k})\right\|_2^2\right] + \frac{\eta L^2 C_2}{2}\sum_{k=0}^{K-1} A1_k \\
&\quad + \frac{\eta^2 L}{2}\sum_{n=1}^{N} p_n^2(\delta_{g,n}H+1)\frac{\sigma^2}{M} + \frac{3\eta^2 LKH^2}{2}G^2 \\
&\quad + \frac{\eta L}{2}\sqrt{d}\tau\sum_{n=1}^{N} p_n^2(\delta_{g,n}+1)\delta_{w,n} \qquad (17)
\end{aligned}$$

where $C_1 = 1 - 3\eta LH\bar{\delta}_g$ and $C_2 = 1 + 3\eta L\bar{\delta}_g H$.

Plugging Lemma 3 into (17), if $C_1' = C_1 - \frac{3\eta^2 H^2(pL^2+3\eta L^3\bar{\delta}_g H)}{1-3\eta^2 pL^2H^2} \geq 0$, we have,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\triangledown F(\boldsymbol{w}^{k})\right\|_2^2\right] \\
&\leq \frac{2\mathbb{E}\left[F(\boldsymbol{w}^{0}) - F(\boldsymbol{w}^{K})\right]}{\eta C_1' K} + \frac{\eta^2 L^2 C_2 H\sigma^2}{C_1' M(1-3\eta^2L^2H^2)} \\
&+ \frac{3\eta^2L^2C_2H^2G^2}{C_1'(1-3\eta^2L^2H^2)} + \frac{\eta L^2 C_2 H\sqrt{d}\bar{\delta}_w\tau}{C_1'(1-3\eta^2L^2H^2)} + \frac{3\eta LH}{C_1'}G^2 \\
&+ \frac{\eta L\sigma^2(\bar{\delta}_g H+1)}{C_1' M} + \frac{\eta L\sqrt{d}\tau}{C_1'}\sum_{n=1}^{N} p_n^2(\delta_{g,n}+1)\delta_{w,n} \qquad (18)
\end{aligned}$$

If we set $\eta = \sqrt{MN/K}$ and

$$\eta LH\delta_g \geq \frac{\eta^2H^2(L^2+3\eta L^3\delta_g H)}{1-3\eta^2L^2H^2}, \qquad (19)$$

we can get the $1/C_1' \leq 2$. Thus,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\triangledown F(\boldsymbol{w}^{k})\right\|_2^2\right] \\
&\leq \frac{4\mathbb{E}\left[F(\boldsymbol{w}^{0}) - F(\boldsymbol{w}^{K})\right]}{\sqrt{MNK}} + \frac{2L\sigma^2(2H\bar{\delta}_g+p)}{\sqrt{MNK}} \\
&\quad + \frac{12MLH\bar{\delta}_g G^2}{\sqrt{MNK}} + 2L\sqrt{d}\tau\sum_{n=1}^{N} p_n^2(\delta_{g,n}+1)\delta_{w,n}. \qquad (20)
\end{aligned}$$

and the proof is completed.

References

[1] H. Li, S. De, Z. Xu, C. Studer, H. Samet, and T. Goldstein, "Training quantized nets: A deeper understanding," in *Advances in Neural Information Processing Systems*, Long Beach, CA, December 2017, pp. 5811–5821.

[2] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," in *Advances in Neural Information Processing Systems*. Montreal, Canada: NIPS, December 2018, pp. 2525–2536.