I. Main result for non-convex setting

Motivated by the iterate analysis framework in [1], we define the auxiliary sequences for each device $i \in \mathcal{N}$, where $\boldsymbol{v}_i^0 := \tilde{\boldsymbol{w}}_i^0$ and for all $t \geq 1$ as follows:

$$\boldsymbol{v}_i^{t+1} = \boldsymbol{v}_i^t - \gamma \grave{\nabla} f_i(\tilde{\boldsymbol{w}}_i^t), \tag{1}$$

and $\boldsymbol{v}^t = \frac{1}{N}\sum_{i=1}^N \boldsymbol{v}_i^t$. For convenience, we also define $\grave{\boldsymbol{g}}^t := \frac{1}{N}\sum_{i=1}^N \grave{\nabla} f_i(\tilde{\boldsymbol{w}}_i^t)$ and $\boldsymbol{g}^t = \frac{1}{N}\sum_{i=1}^N \nabla f_i(\tilde{\boldsymbol{w}}_i^t)$. Therefore, $\boldsymbol{g}^t = \mathbb{E}_\xi\left[\grave{\boldsymbol{g}}^t\right]$ and $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t - \gamma \grave{\boldsymbol{g}}^t$.

A. Lemmas

We present some necessary lemmas that are useful in the proof of Theorem 1. Under the $\grave{L}$-smooth assumption of F, we have,

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{v}^{t+1}) - f(\boldsymbol{v}^t)\right] &\leq \mathbb{E}\left[\langle \nabla f(\boldsymbol{v}^t), \boldsymbol{v}^{t+1} - \boldsymbol{v}^t\rangle\right] + \frac{\grave{L}}{2}\mathbb{E}\left[\left\|\boldsymbol{v}_{t+1} - \boldsymbol{v}^t\right\|_2^2\right] \\
&= -\gamma \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{w}_t), \grave{\boldsymbol{g}}^t\right\rangle\right] + \frac{\gamma^2 \grave{L}}{2}\mathbb{E}\left[\left\|\grave{\boldsymbol{g}}^t\right\|_2^2\right]
\end{aligned} \tag{2}$$

According to Assumption 3, we get $\mathbb{E}_\xi\left[\left\|\boldsymbol{g}^t - \grave{\boldsymbol{g}}^t\right\|_2^2\right] \leq \frac{1}{N^2}\sum_{i=1}^N \frac{1}{M}\tilde{G}^2 = \frac{\tilde{G}^2}{MN}$. This gives

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{v}^{t+1}) - f(\boldsymbol{v}^t)\right] &\leq -\gamma \mathbb{E}\left[\left\langle \nabla f(\boldsymbol{v}^t), \frac{1}{N}\sum_{i=1}^N \nabla f_i(\tilde{\boldsymbol{w}}_i^t)\right\rangle\right] + \gamma^2 \grave{L}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2\right] + \gamma^2\frac{\tilde{G}^2}{MN} \\
&= -\frac{\gamma}{2}\left(\left\|\nabla f(\boldsymbol{v}^t)\right\|_2^2 + \left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2 - \left\|\nabla f(\boldsymbol{v}^t) - \frac{1}{N}\sum_{i=1}^N \nabla f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2\right) \\
&\quad + \gamma^2 \grave{L}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2\right] + \gamma^2\frac{\tilde{G}^2}{MN} \\
&\leq -\frac{\gamma}{2}(1 - 2\gamma\grave{L})\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|\nabla f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2\right] + \frac{\gamma \grave{L}^2}{2}\frac{1}{N}\sum_{i=1}^N \left\|\boldsymbol{v}^t - \tilde{\boldsymbol{w}}_i^t\right\|_2^2 + \frac{\gamma^2\tilde{G}^2}{MN}
\end{aligned} \tag{3}$$

Lemma 1.

$$\left\|\boldsymbol{v}^t - \tilde{\boldsymbol{w}}_i^t\right\|_2^2 \leq \frac{\gamma^2}{N^2}\sum_{i=1}^N (\kappa_{q,s} - 1)H^2\tilde{G}^2 + \frac{\gamma^2 l}{N^2}\sum_{i=1}^N \left(\frac{\tilde{\sigma}_i^t}{h_i^t\alpha_i^t}\right)^2 \tag{4}$$

where $\tilde{\sigma}_i^t = \sqrt{(h_i^t\alpha_i^t\sigma_i^t)^2\kappa_{q,s}/\theta_s + N_o}$.

Proof. We show this along the lines of the proof of [2]. Since the proposed algorithm requires a communication after H local training iterations. Therefore, for any $t \geq 0$, there exists a $t_k \leq t$, such that $t - t_k \leq H$

According to the notations of $\boldsymbol{v}^t$ and $\tilde{\boldsymbol{w}}_i^t$, we can write

$$\begin{aligned}\left\|\boldsymbol{v}^{t+1}-\tilde{\boldsymbol{w}}_i^{t+1}\right\|_2^2 &= \left\|\boldsymbol{v}^{t+1}-\boldsymbol{w}^{t+1}+\boldsymbol{w}^{t+1}-\tilde{\boldsymbol{w}}_i^{t+1}\right\|_2^2 \\ &\overset{(a)}{=} \left\|\frac{1}{N}\sum_{i=1}^{N}\gamma\left(C_{q,s}(\sum_{j\le t_k}\tilde{\boldsymbol{g}}_i^j+\boldsymbol{\eta}_i^t)-\tilde{\boldsymbol{g}}_i^t+\frac{\boldsymbol{n}_i^t}{h_i^t\alpha_i^t}\right)\right\|_2^2 \\ &\le \frac{1}{N^2}\sum_{i=1}^{N}(\kappa_{q,s}-1)\gamma^2\left\|\sum_{j}^{j+H-1}\tilde{\boldsymbol{g}}_i^j+\boldsymbol{\eta}_i^t\right\|_2^2+\frac{\gamma^2}{N}\sum_{i=1}^{N^2}\left\|\boldsymbol{\eta}_i^t\right\|_2^2+\frac{\gamma^2}{N^2}\sum_{i=1}^{N}\left\|\frac{\boldsymbol{n}_i^t}{h_i^t\alpha_i^t}\right\|_2^2 \\ &\le \frac{1}{N^2}\sum_{i=1}^{N}\gamma^2(\kappa_{q,s}-1)H^2\tilde{G}^2+\frac{\gamma^2}{N^2}\sum_{i=1}^{N}\left(\kappa_{q,s}d\sigma_i^2+\frac{N_0}{(h_i^t\alpha_i^t)^2}\right) \end{aligned} \tag{5}$$

where (a) is according to the definition of the updating rules and virtual sequences. □

B. *Proof of Theorem 1*

By substituting (3) with (4), we have

$$\mathbb{E}\left[f(\boldsymbol{v}^{t+1})-f(\boldsymbol{v}^t)\right]\le-\frac{\gamma}{2}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\triangledown f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2\right]+\gamma^2\frac{\tilde{G}^2}{MN}+\frac{\gamma\grave{L}^2}{2}\left(\frac{1}{N}\gamma^2(\kappa_{q,s}-1)H^2\tilde{G}^2+\frac{\gamma^2 l}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\sigma}_i^t}{h_i^t\alpha_i^t}\right)^2\right) \tag{6}$$

where the standard variance of the effective noise is $\tilde{\sigma}_i^t=\sqrt{(h_i^t\alpha_i^t\sigma_i^t)^2\kappa_{q,s}d\sigma_i^2+N_0}$.

Summing up from $t=0$ to $T-1$, yields

$$\frac{\gamma}{2T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\triangledown f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2\right]\le\frac{2(\mathbb{E}\left[f(\boldsymbol{v}^0)\right]-f^\star)}{\gamma T}+\frac{2\gamma\tilde{G}^2}{MN}+\frac{\gamma^2\grave{L}^2}{N}\left(\kappa_{q,s}H^2\tilde{G}^2+\frac{l}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\sigma}_i}{h_i\alpha_i}\right)^2\right) \tag{7}$$

Let $\boldsymbol{z}^T$ be a random variable sampled from $\{\tilde{\boldsymbol{w}}_i^t\}$ with probability $\Pr[\boldsymbol{z}^T=\tilde{\boldsymbol{w}}_i^t]=1/MT$. By taking $\gamma=\frac{\mu\sqrt{N}}{\sqrt{MT}}$, where μ is a constant statifying $\frac{\mu\sqrt{N}}{\sqrt{MT}}\le\frac{1}{2\grave{L}}$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\triangledown f_i(\tilde{\boldsymbol{w}}_i^t)\right\|_2^2\right]\le\frac{2\mu(\mathbb{E}\left[f(\boldsymbol{v}^0)\right]-f^\star)}{\sqrt{NT}}+\frac{2\mu\tilde{G}^2}{M\sqrt{NT}}+\frac{\grave{L}^2\mu^2}{2T}\left(\kappa_{q,s}H^2\tilde{G}^2+\frac{l}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\sigma}_i}{h_i\alpha_i}\right)^2\right) \tag{8}$$

C. *Proof of Corollary 1*

For achieving e model convergence, Let $\bar{\sigma}=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tilde{\sigma}_i}{h_i\alpha_i}\right)^2$

$$\frac{2J+2\tilde{G}^2/M}{\sqrt{NT}}+\frac{\kappa_{q,s}H^2\tilde{G}^2\grave{L}^2+\bar{\sigma}\grave{L}^2}{2T}=e \tag{9}$$

$$eT-\frac{2J+2\tilde{G}^2/M}{\sqrt{N}}\sqrt{T}-\frac{\kappa_{q,s}H^2\tilde{G}^2\grave{L}^2+\bar{\sigma}\grave{L}^2}{2}=0 \tag{10}$$

$$T = \frac{\kappa_{q,s} H^2 \tilde{G}^2 \grave{L}^2 + \bar{\sigma} \grave{L}^2}{2e} + \frac{(J + \tilde{G}^2/M)^2}{Ne^2} + \sqrt{\frac{(J + \tilde{G}^2/M)^2}{2Ne^2} + \frac{\kappa_{q,s} H^2 \tilde{G}^2 \grave{L}^2 + \bar{\sigma} \grave{L}^2}{2e}} \tag{11}$$

$$\begin{aligned} K &= \frac{\kappa_{q,s} H^2 \tilde{G}^2 \grave{L}^2 + \bar{\sigma} \grave{L}^2}{2He} + \frac{(J + \tilde{G}^2/M)^2}{NHe^2} + \sqrt{\frac{(J + \tilde{G}^2/M)^2}{2H^2Ne^2} + \frac{\kappa_{q,s} H^2 \tilde{G}^2 \grave{L}^2 + \bar{\sigma} \grave{L}^2}{2H^2e}} \\ &= \mathcal{O}(\kappa_{q,s} H) + \mathcal{O}(\frac{1}{HN} \sum_{i=1}^{N} \left(\frac{\tilde{\sigma}_i}{h_i \alpha_i} \right)^2) \end{aligned} \tag{12}$$

D. Proof of Lemma 1

Proof. According to the definition of randomized sparsification and stochastic quantization, for $v \in \mathbb{R}^d$ it holds

$$\mathbb{E}_q \left[Q(\hat{v}) \right] = \|\hat{v}\|_2^2 \cdot \operatorname{sgn}(\hat{v}_j) \cdot \frac{|\hat{v}_j|}{\|\hat{v}\|_2^2} \frac{u}{u} = \hat{v}, \tag{13}$$

$$\mathbb{E}_s \left[(1/\theta_s) \cdot (\boldsymbol{m} \otimes v) \right] = \frac{d}{l} \left[\frac{l}{d} v_1, \cdots, \frac{l}{d} v_d \right] = v. \tag{14}$$

and

$$\begin{aligned} &\mathbb{E} \left[\|C_{q,s}(v) - v\|_2^2 \right] \\ &= \|v\|_2^2 + \mathbb{E} \left[\|C_{q,s}(v)\|_2^2 \right] - 2\mathbb{E} \left[\langle v, C_{q,s}(v) \rangle \right] \\ &= \mathbb{E} \left[\|C_{q,s}(v)\|_2^2 \right] - \|v\|_2^2 \\ &\overset{(a)}{\leq} (1 + \grave{\theta}_q) \mathbb{E}_S \left[\|S_k(v)\|_2^2 \right] - \|v\|_2^2 \\ &\overset{(b)}{=} \left((1 + \grave{\theta}_q) \frac{d}{l} - 1 \right) \|v\|_2^2 . \end{aligned} \tag{15}$$

In the second equality, we used that the unbiasedness of the compressor $\mathbb{E} \left[C_{q,s}(v) \right] = v$. (a) is due to the definition of the stochastic quantization: $\mathbb{E}[\|Q(v)\|_2^2] \leq (1 + +\grave{\theta}_q) \|v\|_2^2$ with $+\grave{\theta}_q = \min\{d/(2K^2), \sqrt{d}/K\}$. With the sparsification operator, we can treat the $S_k(v)$ as a length-l vector whose entries correspond to the l non-zero entries of v and thus the variance can be efficiently reduced by a smaller dimension ($l \leq d$). (b) is due to $\mathbb{E}_S \left[\|S_k(v)\|_2^2 \right] = \sum_{j=1}^{A} (\frac{d}{l} v_j)^2 \times \frac{l}{d} = \frac{d}{l} \|v\|_2^2$. □

REFERENCES

[1] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 695–14 706.

[2] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2021–2031.