APPENDIX FOR THE PAPER "TO TALK OR TO WORK: FLEXIBLE COMMUNICATION COMPRESSION FOR ENERGY EFFICIENT FEDERATED LEARNING OVER HETEROGENEOUS MOBILE EDGE DEVICES "

Inspired by the perturbed iterate analysis framework in [1], we first define the following auxiliary sequences for all $t \geq 0$:
1) If $t = 0$, $\tilde{\boldsymbol{w}}_m^{(t)} = \widehat{\boldsymbol{w}}_m^{(0)}$; If $t \geq 1$, $\tilde{\boldsymbol{w}}_m^{(t)} = \tilde{\boldsymbol{w}}_m^{(t-1)} - \eta \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t-1)}; \mathcal{D}_m^{(t-1)})$.
2) $\boldsymbol{q}^{(t)} \triangleq \frac{1}{M} \sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)}; \mathcal{D}_m^{(t)})$
3) $\overline{\boldsymbol{q}}^{(t)} \triangleq \mathbb{E}_{\mathcal{D}_m^{(t)}}[\boldsymbol{q}^{(t)}] = \frac{1}{M} \sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})$
4) $\tilde{\boldsymbol{w}}^{(t)} \triangleq \frac{1}{M} \sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)}; \mathcal{D}_m^{(t)}) = \tilde{\boldsymbol{w}}^{(t-1)} - \eta \boldsymbol{q}^{(t-1)}$
5) $\widehat{\boldsymbol{w}}^{(t)} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\boldsymbol{w}}_m^{(t)}$
By the smoothness of $F : \mathbb{R}^d \to \mathbb{R}$, we have

$$
\begin{aligned}
& F(\tilde{\boldsymbol{w}}^{(t+1)}) - F(\tilde{\boldsymbol{w}}^{(t)}) \\
\leq & < \nabla F(\tilde{\boldsymbol{w}}^{(t)}), \tilde{\boldsymbol{w}}^{(t+1)} - \tilde{\boldsymbol{w}}^{(t)} > + \frac{L}{2} ||\tilde{\boldsymbol{w}}^{(t+1)} - \tilde{\boldsymbol{w}}^{(t)}||^2 \\
= & -\eta < \nabla F(\tilde{\boldsymbol{w}}^{(t)}), \boldsymbol{q}^{(t)} > + \frac{\eta^2 L}{2} ||\boldsymbol{q}^{(t)}||^2 \\
\overset{(a)}{\leq} & -\eta < \nabla F(\tilde{\boldsymbol{w}}^{(t)}), \boldsymbol{q}^{(t)} > + \eta^2 L ||\boldsymbol{q}^{(t)} - \overline{\boldsymbol{q}}^{(t)}||^2 + \eta^2 L ||\overline{\boldsymbol{q}}^{(t)}||^2 \\
= & -\frac{\eta}{M} \sum_{m=1}^{M} < \nabla F(\tilde{\boldsymbol{w}}^{(t)}), \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)}; \mathcal{D}_m^{(t)}) > + \eta^2 L ||\boldsymbol{q}^{(t)} - \overline{\boldsymbol{q}}^{(t)}||^2 + \eta^2 L || \frac{1}{M} \sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)}; \mathcal{D}_m^{(t)})||^2,
\end{aligned}
\tag{1}
$$

where (a) is by Jensen's inequality. Taking expectation with respect to the sampling mini-batch $\mathcal{D}_m^{(t)}$ by each edge device at time t gives

$$
\begin{aligned}
& \mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t+1)})] - F(\tilde{\boldsymbol{w}}^{(t)}) \\
\overset{(a)}{\leq} & -\frac{\eta}{2}(||\nabla F(\tilde{\boldsymbol{w}}^{(t)})||^2 + ||\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2) + \frac{\eta}{2}||\nabla F(\tilde{\boldsymbol{w}}^{(t)}) - \frac{1}{M}\sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2 + \eta^2 L ||\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2 + \frac{\eta^2 L \sigma^2}{M b^{(t)}} \\
\overset{(b)}{\leq} & -\frac{\eta}{2M}\sum_{m=1}^{M}(||\nabla F(\tilde{\boldsymbol{w}}^{(t)})||^2 - L^2 ||\tilde{\boldsymbol{w}}^{(t)} - \widehat{\boldsymbol{w}}_m^{(t)}||^2) + \frac{2\eta^2 L - \eta}{2} ||\frac{1}{M}\sum_{m=1}^{M} \nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2 + \frac{\eta^2 L \sigma^2}{M b^{(t)}} \\
= & -\frac{\eta}{2M}\sum_{m=1}^{M}(||\nabla F(\tilde{\boldsymbol{w}}^{(t)})||^2 + L^2 ||\tilde{\boldsymbol{w}}^{(t)} - \widehat{\boldsymbol{w}}_m^{(t)}||^2) + \frac{2\eta^2 L - \eta}{2M} \sum_{m=1}^{M} ||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2 + \frac{\eta^2 L \sigma^2}{M b^{(t)}} + \frac{\eta L^2}{M} ||\tilde{\boldsymbol{w}}^{(t)} - \widehat{\boldsymbol{w}}_m^{(t)}||^2
\end{aligned}
\tag{2}
$$

where (a) follows by applying two basic inequalities $< \boldsymbol{a}, \boldsymbol{a} > \leq 1/2||\boldsymbol{a}||^2 + 1/2||\boldsymbol{b}||^2$ and $\mathbb{E}[||X||^2] = \mathbb{E}[||X - \mathbb{E}[X]||^2] + ||\mathbb{E}[X]||^2$; (a) follows from the lipschitz continuity of the gradient of local functions. The first term in (2) can be bounded in terms of $||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2$ as follows:

$$
\begin{aligned}
||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2 \leq & 2||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)}) - \nabla F(\tilde{\boldsymbol{w}}^{(t)})||^2 + 2||F(\tilde{\boldsymbol{w}}^{(t)})||^2 \\
\leq & 2L^2 ||\widehat{\boldsymbol{w}}_m^{(t)} - \tilde{\boldsymbol{w}}^{(t)}||^2 + ||F(\tilde{\boldsymbol{w}}^{(t)})||^2
\end{aligned}
\tag{3}
$$

Using $\eta \leq \frac{1}{2L}$ and rearranging the terms in (2), we have

$$
\frac{\eta}{4M}\sum_{m=1}^{M} ||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2 \leq F(\tilde{\boldsymbol{w}}^{(t)}) - \mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t+1)})] + \frac{\eta^2 L \sigma^2}{M b^{(t)}} + \frac{\eta L^2}{M} ||\tilde{\boldsymbol{w}}^{(t)} - \widehat{\boldsymbol{w}}_m^{(t)}||^2
\tag{4}
$$

Taking expectation with respect to the entire process and using the basic inequality $||\boldsymbol{a}+\boldsymbol{b}||^2 \leq 2||\boldsymbol{a}||^2+2||\boldsymbol{b}||^2$ gives

$$\begin{aligned}
&\frac{\eta}{4M}\sum_{m=1}^{M}\mathbb{E}[||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2] \\
\leq&\mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t)})]-\mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t+1)})]+\frac{\eta^2L\sigma^2}{Mb^{(t)}}+2\eta L^2\mathbb{E}[||\tilde{\boldsymbol{w}}^{(t)}-\widehat{\boldsymbol{w}}^{(t)}||^2]+\frac{2\eta L^2}{M}\sum_{m=1}^{M}\mathbb{E}[||\widehat{\boldsymbol{w}}^{(t)}-\widehat{\boldsymbol{w}}_m^{(t)}||^2] \\
\overset{(a)}{\leq}&\mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t)})]-\mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t+1)})]+\frac{2\eta^2L\sigma^2}{M\rho^tb^{(0)}}+2\eta L^2\mathbb{E}[||\tilde{\boldsymbol{w}}^{(t)}-\widehat{\boldsymbol{w}}^{(t)}||^2]+\frac{2\eta L^2}{M}\sum_{m=1}^{M}\mathbb{E}[||\widehat{\boldsymbol{w}}^{(t)}-\widehat{\boldsymbol{w}}_m^{(t)}||^2],
\end{aligned} \tag{5}$$

where (a) follows by recalling $b^{(t)}=\lfloor\rho^tb^{(0)}\rfloor$ and noting $\lfloor x\rfloor > x/2$ as long as $x\geq 2$.

Now we give three important lemmas where the first two are borrowed from [1] and the last one is proved in the following.

***Lemma** 1 (Memory [1]):* The accumulated error captures the distance between the true sequence and virtual sequence. That is

$$\widehat{\boldsymbol{w}}^{(t)}-\tilde{\boldsymbol{w}}^{(t)}=\frac{1}{M}\sum_{m-1}^{M}\boldsymbol{e}_m^{(t)} \tag{6}$$

***Lemma** 2 (Contracting Deviation of Local Sequences [1]):* The deviation of the local sequences is bounded by

$$\frac{1}{M}\sum_{m-1}^{M}\mathbb{E}[||\widehat{\boldsymbol{w}}^{(t)}-\widehat{\boldsymbol{w}}_m^{(t)}||^2]\leq\eta^2G^2H^2 \tag{7}$$

***Lemma** 3 (Bounded Memory):* For worker m who synchronizes with the server every H local iterations, we have

$$\mathbb{E}[||e_m^{(t)}||^2]\leq 4\delta_m^2\eta^2G^2H^2 \tag{8}$$

Proof: Note that Algorithm **??** average the gradients every H iterations between which the accumulated error $\boldsymbol{e}_m^{(t)}$ at any participant m and the global parameter vector $\boldsymbol{w}^{(t)}$ keep unchanged. For ease of presentation, we assume that T is an integer multiple of H. Let $\mathcal{I}_T=\{t_1,t_2,...,t_{T/H}=T\}$ be the aggregation indices satisfying $t_{i+1}-t_i=H$. For every $m\in\mathcal{M}$, we have

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{e}_m^{(t_{i+1})}||^2]=&\mathbb{E}[||\boldsymbol{e}_m^{(t_{i+1}-1)}+\boldsymbol{w}^{(t_{i+1}-1)}-\widehat{\boldsymbol{w}}_m^{(t_{i+1}-\frac{1}{2})}-\boldsymbol{g}_m^{(t_{i+1}-1)}||^2] \\
\overset{(a)}{\leq}&(1-\frac{1}{\delta_m})\mathbb{E}[||\boldsymbol{e}_m^{(t_{i+1}-1)}+\boldsymbol{w}^{(t_{i+1}-1)}-\widehat{\boldsymbol{w}}_m^{(t_{i+1}-\frac{1}{2})}||^2] \\
\overset{(b)}{=}&(1-\frac{1}{\delta_m})\mathbb{E}[||\boldsymbol{e}_m^{(t_i)}+\widehat{\boldsymbol{w}}_m^{(t_i)}-\widehat{\boldsymbol{w}}_m^{(t_{i+1}-\frac{1}{2})}||^2].
\end{aligned} \tag{9}$$

Here (a) is due to the contraction property of $\mathrm{Top}_k(\boldsymbol{x})$ operator [2], that is $\mathbb{E}||\boldsymbol{x}-\mathrm{Top}_k(\boldsymbol{x})||^2\leq(1-k/d)||\boldsymbol{x}||^2,\ \forall\boldsymbol{x}\in\mathbb{R}^d$. In (b), we use $\boldsymbol{e}^{(t_{i+1}-1)}=\boldsymbol{e}^{(t_i)}$ and $\boldsymbol{w}^{(t_{i+1}-1)}=\boldsymbol{w}^{(t_i)}=\widehat{\boldsymbol{w}}_m^{(t_i)}$ that always hold. Since the inequality $||\boldsymbol{a}+\boldsymbol{b}||^2\leq(1+\tau)||\boldsymbol{a}+(1+\frac{1}{\tau})||\boldsymbol{b}||^2$ holds for every $\tau\geq 0$, we take any $p>1$ and transform (9) as follows

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{e}_m^{(t_{i+1})}||^2]\leq&(1-\frac{1}{\delta_m})\{(1+\frac{(p-1)}{p\delta_m})\mathbb{E}[||\boldsymbol{e}_m^{(t_i)}||^2]+(1+\frac{p\delta_m}{(p-1)})\mathbb{E}[||\widehat{\boldsymbol{w}}_m^{(t_i)}-\widehat{\boldsymbol{w}}_m^{(t_{i+1}-\frac{1}{2})}||^2]\} \\
\leq&(1-\frac{1}{p\delta_m})\mathbb{E}[||\boldsymbol{e}_m^{(t_i)}||^2]+\frac{p(\delta_m^2-1)}{(p-1)\delta_m}\mathbb{E}[||\sum_{j=t_i}^{t_{i+1}-1}\eta\nabla f_m(\boldsymbol{w}_m^{(j)};\mathcal{D}_m^{(j)})||^2] \\
\overset{(a)}{\leq}&(1-\frac{1}{p\delta_m})\mathbb{E}[||\boldsymbol{e}_m^{(t_i)}||^2]+\frac{p(\delta_m^2-1)}{(p-1)\delta_m}\eta^2G^2H^2,
\end{aligned} \tag{10}$$

where (a) follows from Assumption 1. Iterating the above inequality from $i=0\rightarrow l$ where $l=T/H$ yields:

$$\mathbb{E}[||\boldsymbol{e}_m^{(t_{i+1})}||^2]\leq\frac{p(\delta_m^2-1)}{(p-1)\delta_m}\eta^2G^2H^2\sum_{j=1}^{(l)}(1-\frac{1}{p\delta_m})^{l-j}\overset{(a)}{\leq}\frac{p^2(\delta_m^2-1)}{p-1}\eta^2G^2H^2\overset{(b)}{\leq}4\delta_m^2\eta^2G^2H^2, \tag{11}$$

where (a) is by the fact that $\sum_{j=1}^{(l)}(1-1/p\delta_m)^{l-j}\leq\sum_{j\geq 0}(1-1/p\delta_m)^j=p\delta_m$, and (b) is by plugging $p=2$. Note the the right-hand-side does not depend on t, i.e., for every $t=0,1,...,T$, the following holds:

$$\mathbb{E}[||\boldsymbol{e}_m^{(t)}||^2]\leq 4\delta_m^2\eta^2G^2H^2. \tag{12}$$

■

Lemma 1 and Lemma 3 together imply:

$$\mathbb{E}[||\widehat{\boldsymbol{w}}^{(t)} - \tilde{\boldsymbol{w}}^{(t)}||^2] \leq \frac{4\eta^2 G^2 H^2}{M} \sum_{m=1}^{M} \delta_m^2. \tag{13}$$

Applying Lemma 2 and (13) into (5), we get

$$\frac{\eta}{4M} \sum_{m=1}^{M} \mathbb{E}[||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2] \leq \mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t)})] - \mathbb{E}[F(\tilde{\boldsymbol{w}}^{(t+1)})] + \frac{2\eta^2 L\sigma^2}{M\rho^t b^{(0)}} + \frac{8\eta^3 L^2 G^2 H^2}{M} \sum_{m=1}^{M} \delta_m^2 + 2\eta^3 L^2 G^2 H^2, \tag{14}$$

Recursively applying the above inequality from $t = 0$ to $t = T - 1$ yields

$$\begin{aligned} \frac{1}{4MT} \sum_{t=0}^{T-1} \sum_{m=1}^{M} \mathbb{E}[||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2] \leq & \frac{\mathbb{E}[F(\tilde{\boldsymbol{w}}^{(0)})] - F^*}{\eta T} + \frac{2\eta L\sigma^2}{MTb^{(0)}} \sum_{t=0}^{T-1} \frac{1}{\rho^t} + \frac{8\eta^2 L^2 G^2 H^2}{M} \sum_{m=1}^{M} \delta_m^2 + 2\eta^2 L^2 G^2 H^2 \\ \overset{(a)}{\leq} & \frac{\mathbb{E}[F(\tilde{\boldsymbol{w}}^{(0)})] - F^*}{\eta T} + \frac{2\eta\rho L\sigma^2}{(\rho - 1)MTb^{(0)}} + \frac{8\eta^2 L^2 G^2 H^2}{M} \sum_{m=1}^{M} \delta_m^2 + 2\eta^2 L^2 G^2 H^2, \end{aligned} \tag{15}$$

where (a) follows by simplifying the partial sum of geometric series and noting that $0 < \frac{1}{\rho} < 1$. Let $\boldsymbol{z}_T$ be a random variable sampled from $\{\widehat{\boldsymbol{w}}_m^{(t)}\}$ with probability $\Pr[\boldsymbol{z}_T = \widehat{\boldsymbol{w}}_m^{(t)}] = \frac{1}{MT}$. By taking $\delta = \sqrt{\frac{1}{M}\sum_{m=1}^{M} \delta_m^2}$ and $\eta = \frac{\theta\sqrt{M}}{\sqrt{T}}$(where θ is a constant satisfying $\frac{\theta\sqrt{M}}{\sqrt{T}} \leq \frac{1}{2L}$), we have

$$\mathbb{E}[||\boldsymbol{z}_T||^2] = \frac{1}{MT} \sum_{t=0}^{T-1} \sum_{m=1}^{M} \mathbb{E}[||\nabla f_m(\widehat{\boldsymbol{w}}_m^{(t)})||^2] \leq \frac{4(\mathbb{E}[F(\boldsymbol{w}^{(0)})] - F^*)}{\theta\sqrt{MT}} + \frac{8\rho\theta L\sigma^2}{(\rho - 1)Mb^{(0)}\sqrt{M}T^{3/2}} + (4\delta^2 + 1)\frac{8M\theta^2 L^2 G^2 H^2}{T}, \tag{16}$$

Until now we complete the proof of Theorem 1.

References

[1] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2019.

[2] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2018.