Supplementary Materials

Here, we provide the proofs of the main two theorems of this paper in Sections 7 and 8 along with the necessary lemmas and discussions. Moreover, we provide more numerical results over more complicated datasets and model parameters in Section 9.

7 Proof of Theorem 1

We first introduce some additional notations which will be used throughput the proofs.

Additional notations. For each period $k = 0, 1, \cdots, K-1$ and iteration $t = 0, 1, \cdots, \tau - 1$ we denote

$$\begin{aligned}
\mathbf{x}_{k+1} &:= \mathbf{x}_k + \frac{1}{r} \sum_{i \in \mathcal{S}_k} Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right), \\
\widehat{\mathbf{x}}_{k+1} &:= \mathbf{x}_k + \frac{1}{n} \sum_{i \in [n]} Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right), \\
\overline{\mathbf{x}}_{k,t} &:= \frac{1}{n} \sum_{i \in [n]} \mathbf{x}_{k,t}^{(i)}.
\end{aligned} \tag{18}$$

We begin the proof of Theorem 1 by noting a few key observations. Based on the above notations and the assumptions we made earlier, the optimality gap of the parameter server's model at period k, i.e. $\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2$, can be decomposed as stated in the following lemma.

Lemma 1. *Consider any period $k = 0, \cdots, K-1$ and the sequences $\{\mathbf{x}_{k+1}, \widehat{\mathbf{x}}_{k+1}, \overline{\mathbf{x}}_{k,\tau}\}$ generated by the* ***FedPAQ*** *method in Algorithm 1. If Assumption 1 holds, then*

$$\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 = \mathbb{E}\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1}\|^2 + \mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau}\right\|^2 + \mathbb{E}\left\|\overline{\mathbf{x}}_{k,\tau} - \mathbf{x}^*\right\|^2, \tag{19}$$

where the expectation is with respect to all sources of randomness.

Proof. See Section 7.1. □

In the following three lemmas, we characterize each of the terms in the right-hand side (RHS) of (19).

Lemma 2. *Consider the sequence of local updates in the* ***FedPAQ*** *method in Algorithm 1 and let Assumptions 2, 3 and 4 hold. The optimality gap for the average model at the end of period k, i.e. $\overline{\mathbf{x}}_{k,\tau}$, relates to that of the initial model of the k-th period $\mathbf{x}_k$ as follows:*

$$\begin{aligned}
\mathbb{E}\left\|\overline{\mathbf{x}}_{k,\tau} - \mathbf{x}^*\right\|^2 \leq{}& \left(1 + n\eta_k^2\right)\left(1 - \mu\eta_k\right)^\tau \mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \\
&+ \tau(\tau-1)^2 L^2 \frac{\sigma^2}{n} e\eta_k^2 + \tau^2 \frac{\sigma^2}{n}\eta_k^2 \\
&+ \tau^2(\tau-1) L^2 \sigma^2 e \eta_k^4,
\end{aligned} \tag{20}$$

for the stepsize $\eta_k \leq \min\{\mu/L^2, 1/L\tau\}$.

Proof. See Section 7.2. □

Lemma 3. *For the proposed* ***FedPAQ*** *method in Algorithm 1 with stepsize $\eta_k \leq \min\{\mu/L^2, 1/L\tau\}$ and under Assumptions 1, 2, 3 and 4, we have*

$$\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau}\right\|^2 \leq 2\frac{q}{n}\tau^2 L^2 \eta_k^2 \mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 + 2q\tau^2 \frac{\sigma^2}{n}\eta_k^2 + 2q(\tau-1)\tau^2 L^2 \frac{\sigma^2}{n} e\eta_k^4, \tag{21}$$

where $\widehat{\mathbf{x}}_{k+1}$ and $\overline{\mathbf{x}}_{k,\tau}$ are defined in (18).

Proof. See Section 7.3. □

Lemma 4. *For the proposed FedPAQ method in Algorithm 1 with stepsize $\eta_k \leq \min\{\mu/L^2, 1/L\tau\}$ and under Assumptions 1–4, we have*

$$\mathbb{E}\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1}\|^2 \leq \frac{n-r}{r(n-1)} 8(1+q)\left\{\tau^2 L^2 \eta_k^2 \mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \tau^2\sigma^2\eta_k^2 + (\tau-1)\tau^2 L^2 \sigma^2 e \eta_k^4\right\}, \tag{22}$$

where r denotes the number of nodes contributing in each period of the FedPAQ method.

Proof. See Section 7.4. □

Now that we have established the main building modules for proving Theorem 1, let us proceed with the proof by putting together the results in Lemmas 1–4. That is,

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq \mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 &\left(\left(1 + n\eta_k^2\right)(1-\mu\eta_k)^\tau + 2L^2\tau^2\eta_k^2\left(\frac{q}{n} + \frac{n-r}{r(n-1)}4(1+q)\right)\right) \\
&+ \left(1 + 2q + 8(1+q)\frac{n(n-r)}{r(n-1)}\right)\frac{\sigma^2}{n}\tau^2\eta_k^2 \\
&+ L^2\frac{\sigma^2}{n}e\tau(\tau-1)^2\eta_k^2 \\
&+ \left(n + 2q + 8(1+q)\frac{n(n-r)}{r(n-1)}\right)L^2\frac{\sigma^2}{n}e(\tau-1)\tau^2\eta_k^4
\end{aligned} \tag{23}$$

Let us set the following notations:

$$\begin{aligned}
\delta_k &:= \mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2, \\
C_0 &:= \left(1 + n\eta_k^2\right)(1-\mu\eta_k)^\tau + 2L^2\tau^2\eta_k^2\left(\frac{q}{n} + \frac{n-r}{r(n-1)}4(1+q)\right), \\
C_1 &:= \frac{16}{\mu^2}\left(1 + 2q + 8(1+q)\frac{n(n-r)}{r(n-1)}\right)\frac{\sigma^2}{n}, \\
C_2 &:= \frac{16}{\mu^2}L^2\frac{\sigma^2}{n}e, \\
C_3 &:= \frac{256}{\mu^4}\left(n + 2q + 8(1+q)\frac{n(n-r)}{r(n-1)}\right)L^2\frac{\sigma^2}{n}e.
\end{aligned} \tag{24}$$

Consider C_0, the coefficient of $\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2$ in (23). One can show that if the condition in (11) in Theorem 1 is satisfied, then we have $C_0 \leq 1 - \frac{1}{2}\mu\tau\eta_k$ (See Section 7.6). Therefore, for each period $k \geq k_0$ we have

$$\delta_{k+1} \leq \left(1 - \frac{1}{2}\mu\tau\eta_k\right)\delta_k + \frac{\mu^2}{16}C_1\tau^2\eta_k^2 + \frac{\mu^2}{16}C_2\tau(\tau-1)^2\eta_k^2 + \frac{\mu^4}{256}C_3(\tau-1)\tau^2\eta_k^4. \tag{25}$$

Now, we substitute the stepsize $\eta_k = {}^{4\mu^{-1}}/_{k\tau+1}$ in (25) which yields

$$\delta_{k+1} \leq \left(1 - \frac{2}{k+1/\tau}\right)\delta_k + C_1\frac{1}{(k+1/\tau)^2} + C_2\frac{(\tau-1)^2}{\tau}\frac{1}{(k+1/\tau)^2} + C_3\frac{\tau-1}{\tau^2}\frac{1}{(k+1/\tau)^4}. \tag{26}$$

In Lemma 5, we show the convergence analysis of such sequence. In particular, we take $k_1 = 1/\tau$, $a = C_1 + C_2(\tau-1)^2/\tau$ and $b = C_3(\tau-1)/\tau^2$ in Lemma 5 and conclude for any $k \geq k_0$ that

$$\delta_k \leq \frac{(k_0 + 1/\tau)^2}{(k+1/\tau)^2}\delta_{k_0} + C_1\frac{1}{k+1/\tau} + C_2\frac{(\tau-1)^2}{\tau}\frac{1}{k+1/\tau} + C_3\frac{\tau-1}{\tau^2}\frac{1}{(k+1/\tau)^2}. \tag{27}$$

Finally, rearranging the terms in (27) yields the desired result in Theorem 1, that is

$$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \frac{(k_0\tau+1)^2}{(k\tau+1)^2}\mathbb{E}\|\mathbf{x}_{k_0} - \mathbf{x}^*\|^2 + C_1\frac{\tau}{k\tau+1} + C_2\frac{(\tau-1)^2}{k\tau+1} + C_3\frac{\tau-1}{(k\tau+1)^2}. \tag{28}$$

7.1 Proof of Lemma 1

Let $\mathcal{F}_{k,t}$ denote the history of all sources of randomness by the t-th iteration in period k. The following expectation arguments are conditional on the history $\mathcal{F}_{k,\tau}$ which we remove in our notations for simplicity. Since the random subset of nodes $\mathcal{S}_k$ is uniformly picked from the set of all the nodes $[n]$, we can write

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_k}\mathbf{x}_{k+1} &= \mathbf{x}_k + \mathbb{E}_{\mathcal{S}_k}\frac{1}{r}\sum_{i\in\mathcal{S}_k} Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right) \\
&= \mathbf{x}_k + \sum_{\substack{\mathcal{S}\subseteq[n]\\|\mathcal{S}|=r}} \Pr\left[\mathcal{S}_k = \mathcal{S}\right]\frac{1}{r}\sum_{i\in\mathcal{S}_k} Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right) \\
&= \mathbf{x}_k + \frac{1}{\binom{n}{r}}\frac{1}{r}\binom{n-1}{r-1}\sum_{i\in[n]} Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right) \\
&= \mathbf{x}_k + \frac{1}{n}\sum_{i\in[n]} Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right) \\
&= \widehat{\mathbf{x}}_{k+1}. \qquad (29)
\end{aligned}$$

Moreover, the quantizer $Q(\cdot)$ is unbiased according to Assumption 1, which yields

$$\begin{aligned}
\mathbb{E}_Q\,\widehat{\mathbf{x}}_{k+1} &= \mathbf{x}_k + \frac{1}{n}\sum_{i\in[n]} \mathbb{E}_Q\, Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right) \\
&= \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_{k,\tau}^{(i)} \\
&= \overline{\mathbf{x}}_{k,\tau}. \qquad (30)
\end{aligned}$$

Finally, since the two randomnesses induced by the quantization and random sampling are independent, together with (29) and (30) we can conclude that:

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \mathbb{E}\left\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1} + \widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau} + \overline{\mathbf{x}}_{k,\tau} - \mathbf{x}^*\right\|^2 \\
&= \mathbb{E}\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1}\|^2 + \mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau}\right\|^2 + \mathbb{E}\left\|\overline{\mathbf{x}}_{k,\tau} - \mathbf{x}^*\right\|^2. \qquad (31)
\end{aligned}$$

7.2 Proof of Lemma 2

According to update rule in Algorithm 1, local model at node i for each iteration $t = 0, \cdots, \tau-1$ of period $k = 0, \cdots, K-1$ can be written as follows:

$$\mathbf{x}_{k,t+1}^{(i)} = \mathbf{x}_{k,t}^{(i)} - \eta_k \widetilde{\nabla} f_i\left(\mathbf{x}_{k,t}^{(i)}\right), \qquad (32)$$

where all the nodes start the period with the initial model $\mathbf{x}_{k,0}^{(i)} = \mathbf{x}_k$. In parallel, let us define another sequence of updates as follows:

$$\beta_{k,t+1} = \beta_{k,t} - \eta_k \nabla f\left(\beta_{k,t}\right), \qquad (33)$$

also starting with $\beta_{k,0} = \mathbf{x}_k$. The auxiliary sequence $\{\beta_{k,t}\}$ represents Gradient Descent updates over the global loss function f while $\mathbf{x}_{k,t}^{(i)}$ captures the sequence of SGD updates on each local node. However, both sequences are initialized with $\mathbf{x}_k$ at the beginning of each period k. To evaluate the deviation $\left\|\overline{\mathbf{x}}_{k,\tau} - \mathbf{x}^*\right\|^2$, we link the two sequences. In particular, let us define the following notations for each $k = 0, \cdots, K-1$ and $t = 0, \cdots, \tau-1$:

$$\mathbf{e}_{k,t} = \frac{1}{n}\sum_{i\in[n]} \widetilde{\nabla} f_i\left(\mathbf{x}_{k,t}^{(i)}\right) - \nabla f\left(\beta_{k,t}\right). \qquad (34)$$

One can easily observe that $\mathbb{E}\mathbf{e}_{k,0} = 0$ as $\mathbf{x}_{k,0}^{(i)} = \beta_{k,0} = \mathbf{x}_k$ and $\widetilde{\nabla} f_i$ is unbiased for ∇f. However, $\mathbb{E}\mathbf{e}_{k,t} \neq 0$ for $t \geq 1$. In other words, $\frac{1}{n}\sum_{i\in[n]} \widetilde{\nabla} f_i(\mathbf{x}_{k,t}^{(i)})$ is not unbiased for $\nabla f(\beta_{k,t})$. We also define $\mathbf{e}_k = \mathbf{e}_{k,0} + \cdots + \mathbf{e}_{k,\tau-1}$

and $\mathbf{g}_k = \nabla f(\beta_{k,0}) + \cdots + \nabla f(\beta_{k,\tau-1})$. Now, the average model obtained at the end of period k can be written as

$$\begin{aligned}\overline{\mathbf{x}}_{k,\tau} &= \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_{k,\tau}^{(i)} \\ &= \mathbf{x}_k - \eta_k \left(\frac{1}{n}\sum_{i\in[n]} \widetilde{\nabla} f_i\left(\mathbf{x}_{k,0}^{(i)}\right) + \cdots + \frac{1}{n}\sum_{i\in[n]} \widetilde{\nabla} f_i\left(\mathbf{x}_{k,\tau-1}^{(i)}\right)\right) \\ &= \mathbf{x}_k - \eta_k \left(\mathbf{g}_k + \mathbf{e}_k\right). \end{aligned} \tag{35}$$

Therefore, the optimality gap for the averaged model can be written as

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{x}}_{k,\tau} - \mathbf{x}^*\right\|^2 &= \mathbb{E}\|\mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*\|^2 - 2\eta_k\mathbb{E}\left\langle \mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*, \mathbf{e}_k\right\rangle + \eta_k^2\mathbb{E}\|\mathbf{e}_k\|^2 \\ &\le \mathbb{E}\|\mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*\|^2 \\ &\quad + n\eta_k^2\mathbb{E}\|\mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*\|^2 + \frac{1}{n}\|\mathbb{E}\mathbf{e}_k\|^2 \\ &\quad + \eta_k^2\mathbb{E}\|\mathbf{e}_k\|^2 \\ &= \left(1 + n\eta_k^2\right)\mathbb{E}\|\mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*\|^2 + \frac{1}{n}\|\mathbb{E}\mathbf{e}_k\|^2 + \eta_k^2\mathbb{E}\|\mathbf{e}_k\|^2, \end{aligned} \tag{36}$$

where we used the inequality $-2\langle \mathbf{a}, \mathbf{b}\rangle \le \alpha\|\mathbf{a}\|^2 + \alpha^{-1}\|\mathbf{b}\|^2$ for any two vectors $\mathbf{a}, \mathbf{b}$ and scalar $\alpha > 0$. In the following, we bound each of the three terms in the RHS of (36). First, consider the term $\|\mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*\|^2$ and recall the auxiliary sequence $\{\beta_{k,t}\}$ defined in (37). For every t and k we have

$$\begin{aligned}\left\|\beta_{k,t+1} - \mathbf{x}^*\right\|^2 &= \left\|\beta_{k,t} - \eta_k\nabla f(\beta_{k,t}) - \mathbf{x}^*\right\|^2 \\ &= \left\|\beta_{k,t} - \mathbf{x}^*\right\|^2 - 2\eta_k\left\langle \beta_{k,t} - \mathbf{x}^*, \nabla f(\beta_{k,t})\right\rangle + \eta_k^2\left\|\nabla f(\beta_{k,t})\right\|^2 \\ &\le \left(1 - 2\mu\eta_k + L^2\eta_k^2\right)\left\|\beta_{k,t} - \mathbf{x}^*\right\|^2 \\ &\le (1 - \mu\eta_k)\left\|\beta_{k,t} - \mathbf{x}^*\right\|^2. \end{aligned} \tag{37}$$

In the above derivations, we used the facts that f is μ-strongly convex and its gradient is L-Lipschitz (Assumptions 2 and 4). The stepsize is also picked such that $\eta_k \le \mu/L^2$. Now, conditioned on the history $\mathcal{F}_{k,0}$ and using (37) we have

$$\begin{aligned}\|\mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*\|^2 &= \left\|\beta_{k,\tau} - \mathbf{x}^*\right\|^2 \\ &\le (1 - \mu\eta_k)^\tau\left\|\beta_{k,0} - \mathbf{x}^*\right\|^2 \\ &= (1 - \mu\eta_k)^\tau\|\mathbf{x}_k - \mathbf{x}^*\|^2. \end{aligned} \tag{38}$$

Secondly, consider the term $\|\mathbb{E}\mathbf{e}_k\|^2$ in (36). By definition, we have $\mathbb{E}\mathbf{e}_k = \mathbb{E}\mathbf{e}_{k,1} + \cdots + \mathbb{E}\mathbf{e}_{k,\tau-1}$ and hence $\|\mathbb{E}\mathbf{e}_k\|^2 \le (\tau-1)\left\|\mathbb{E}\mathbf{e}_{k,1}\right\|^2 + \cdots + (\tau-1)\left\|\mathbb{E}\mathbf{e}_{k,\tau-1}\right\|^2$. The first term $\left\|\mathbb{E}\mathbf{e}_{k,1}\right\|^2$ can be bounded using Assumptions

2 and 3 as follows:

$$\begin{aligned}
\left\|\mathbb{E}\mathbf{e}_{k,1}\right\|^2 &= \left\|\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\widetilde{\nabla}f_i\left(\mathbf{x}_{k,1}^{(i)}\right)-\nabla f\left(\beta_{k,1}\right)\right\|^2\\
&= \left\|\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\nabla f\left(\mathbf{x}_{k,1}^{(i)}\right)-\nabla f\left(\beta_{k,1}\right)\right\|^2\\
&\leq \frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left\|\nabla f\left(\mathbf{x}_{k,1}^{(i)}\right)-\nabla f\left(\beta_{k,1}\right)\right\|^2\\
&\leq \frac{1}{n}L^2\sum_{i\in[n]}\mathbb{E}\left\|\mathbf{x}_{k,1}^{(i)}-\beta_{k,1}\right\|^2\\
&= \frac{1}{n}L^2\sum_{i\in[n]}\mathbb{E}\left\|\left(\mathbf{x}_{k,0}^{(i)}-\eta_k\widetilde{\nabla}f_i\left(\mathbf{x}_{k,0}^{(i)}\right)\right)-\left(\beta_{k,0}-\eta_k\nabla f\left(\beta_{k,0}\right)\right)\right\|^2\\
&= \frac{1}{n}L^2\eta_k^2\sum_{i\in[n]}\mathbb{E}\left\|\widetilde{\nabla}f_i\left(\mathbf{x}_k\right)-\nabla f\left(\mathbf{x}_k\right)\right\|^2\\
&\leq L^2\sigma^2\eta_k^2.
\end{aligned}\tag{39}$$

In general, for each $t=1\cdots,\tau-1$ we can write

$$\begin{aligned}
\left\|\mathbb{E}\mathbf{e}_{k,t}\right\|^2 &= \left\|\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\widetilde{\nabla}f_i\left(\mathbf{x}_{k,t}^{(i)}\right)-\nabla f\left(\beta_{k,t}\right)\right\|^2\\
&= \left\|\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)-\nabla f\left(\beta_{k,t}\right)\right\|^2\\
&\leq \frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)-\nabla f\left(\beta_{k,t}\right)\right\|^2\\
&\leq \frac{1}{n}L^2\sum_{i\in[n]}\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)}-\beta_{k,t}\right\|^2.
\end{aligned}\tag{40}$$

Let us denote $a_{k,t} \coloneqq \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \beta_{k,t}\right\|^2$. In the following, we will derive a recursive bound on a_t. That is,

$$
\begin{aligned}
a_{k,t} &= \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \beta_{k,t}\right\|^2 \\
&= \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\Bigg\| \left(\mathbf{x}_{k,0}^{(i)} - \eta_k \widetilde{\nabla} f_i\left(\mathbf{x}_{k,0}^{(i)}\right) - \cdots - \eta_k \widetilde{\nabla} f_i\left(\mathbf{x}_{k,t-1}^{(i)}\right)\right) \\
&\qquad\qquad - \left(\beta_{k,0} - \eta_k \nabla f\left(\beta_{k,0}\right) - \cdots - \eta_k \nabla f\left(\beta_{k,t-1}\right)\right)\Bigg\|^2 \\
&= \frac{1}{n}\eta_k^2 \sum_{i\in[n]} \mathbb{E}\left\|\widetilde{\nabla} f_i\left(\mathbf{x}_{k,0}^{(i)}\right) - \nabla f\left(\beta_{k,0}\right) + \cdots + \widetilde{\nabla} f_i\left(\mathbf{x}_{k,t-1}^{(i)}\right) - \nabla f\left(\beta_{k,t-1}\right)\right\|^2 \\
&\le \eta_k^2\sigma^2 + \frac{1}{n}\eta_k^2 \sum_{i\in[n]} \mathbb{E}\left\|\widetilde{\nabla} f_i\left(\mathbf{x}_{k,1}^{(i)}\right) - \nabla f\left(\beta_{k,1}\right) + \cdots + \widetilde{\nabla} f_i\left(\mathbf{x}_{k,t-1}^{(i)}\right) - \nabla f\left(\beta_{k,t-1}\right)\right\|^2 \\
&\le \eta_k^2\sigma^2 + \frac{1}{n}\eta_k^2 \sum_{i\in[n]} \mathbb{E}\Bigg\|\widetilde{\nabla} f_i\left(\mathbf{x}_{k,1}^{(i)}\right) - \nabla f\left(\mathbf{x}_{k,1}^{(i)}\right) + \nabla f\left(\mathbf{x}_{k,1}^{(i)}\right) - \nabla f\left(\beta_{k,1}\right) \\
&\qquad\qquad + \cdots + \widetilde{\nabla} f_i\left(\mathbf{x}_{k,t-1}^{(i)}\right) - \nabla f\left(\mathbf{x}_{k,t-1}^{(i)}\right) + \nabla f\left(\mathbf{x}_{k,t-1}^{(i)}\right) - \nabla f\left(\beta_{k,t-1}\right)\Bigg\|^2 \\
&\le t\eta_k^2\sigma^2 + \frac{1}{n}\eta_k^2 \sum_{i\in[n]} \mathbb{E}\left\|\nabla f\left(\mathbf{x}_{k,1}^{(i)}\right) - \nabla f\left(\beta_{k,1}\right) + \cdots \nabla f\left(\mathbf{x}_{k,t-1}^{(i)}\right) - \nabla f\left(\beta_{k,t-1}\right)\right\|^2 \\
&\le t\eta_k^2\sigma^2 + (t-1)L^2\eta_k^2 \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\left\|\mathbf{x}_{k,1}^{(i)} - \beta_{k,1}\right\|^2 + \cdots (t-1)L^2\eta_k^2 \frac{1}{n}\sum_{i\in[n]} \mathbb{E}\left\|\mathbf{x}_{k,t-1}^{(i)} - \beta_{k,t-1}\right\|^2 \\
&= t\eta_k^2\sigma^2 + (t-1)L^2\eta_k^2\left(a_{k,1} + \cdots + a_{k,t-1}\right) \\
&\le \tau\eta_k^2\sigma^2 + \tau L^2\eta_k^2\left(a_{k,1} + \cdots + a_{k,t-1}\right). \qquad (41)
\end{aligned}
$$

Therefore, for the sequence $\{a_{k,1}, \cdots, a_{k,\tau-1}\}$ we have shown that

$$
a_{k,t} \le \tau\eta_k^2\sigma^2 + \tau L^2\eta_k^2\left(a_{k,1} + \cdots + a_{k,t-1}\right), \qquad (42)
$$

where $a_{k,1} \le \sigma^2\eta_k^2$. We can show by induction, that such sequence satisfies the following inequality:

$$
a_{k,t} \le \tau\eta_k^2\sigma^2\left(1 + \tau L^2\eta_k^2\right)^{t-1}. \qquad (43)
$$

See Section 7.5 for the detailed proof. Therefore, we have

$$
\begin{aligned}
\left\|\mathbb{E}\mathbf{e}_k\right\|^2 &\le (\tau-1)\left\|\mathbb{E}\mathbf{e}_{k,1}\right\|^2 + \cdots + (\tau-1)\left\|\mathbb{E}\mathbf{e}_{k,\tau-1}\right\|^2 \\
&\le (\tau-1)L^2\left(a_1 + \cdots + a_{\tau-1}\right) \\
&\le \tau(\tau-1)^2L^2\sigma^2\eta_k^2\left(1 + \tau L^2\eta_k^2\right)^{\tau}. \qquad (44)
\end{aligned}
$$

Now, we use the inequality $1 + x \le e^x$ and conclude that

$$
\left\|\mathbb{E}\mathbf{e}_k\right\|^2 \le \tau(\tau-1)^2L^2\sigma^2\eta_k^2 e^{\tau^2L^2\eta_k^2}. \qquad (45)
$$

Therefore, if $\tau^2L^2\eta_k^2 \le 1$, we have

$$
\left\|\mathbb{E}\mathbf{e}_k\right\|^2 \le \tau(\tau-1)^2L^2\sigma^2 e\eta_k^2. \qquad (46)
$$

Finally, we bound the third term in (36), that is $\mathbb{E}\|\mathbf{e}_k\|^2$. Using the definition, we know that $\mathbb{E}\|\mathbf{e}_k\|^2 \leq \tau\mathbb{E}\left\|\mathbf{e}_{k,0}\right\|^2 + \cdots + \tau\mathbb{E}\left\|\mathbf{e}_{k,\tau-1}\right\|^2$. Firstly, note that

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{e}_{k,0}\right\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}\widetilde{\nabla} f_i\left(\mathbf{x}_{k,0}^{(i)}\right) - \nabla f\left(\beta_{k,0}\right)\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}\widetilde{\nabla} f_i\left(\mathbf{x}_k\right) - \nabla f\left(\mathbf{x}_k\right)\right\|^2 \\
&\leq \frac{\sigma^2}{n}.
\end{aligned} \tag{47}$$

For each $t = 1, \cdots, \tau-1$ we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{e}_{k,t}\right\|^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}\widetilde{\nabla} f_i\left(\mathbf{x}_{k,t}^{(i)}\right) - \nabla f\left(\beta_{k,t}\right)\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}\widetilde{\nabla} f_i\left(\mathbf{x}_{k,t}^{(i)}\right) - \nabla f\left(\mathbf{x}_{k,t}^{(i)}\right) + \frac{1}{n}\sum_{i\in[n]}\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right) - \nabla f\left(\beta_{k,t}\right)\right\|^2 \\
&\leq \frac{\sigma^2}{n} + L^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left\|\mathbf{x}_{k,t}^{(i)} - \beta_{k,t}\right\|^2 \\
&= \frac{\sigma^2}{n} + L^2 a_{k,t}.
\end{aligned} \tag{48}$$

Summing over $t = 0, 1, \cdots, \tau-1$ results in the following

$$\begin{aligned}
\mathbb{E}\|\mathbf{e}_k\|^2 &\leq \tau\mathbb{E}\left\|\mathbf{e}_{k,0}\right\|^2 + \cdots + \tau\mathbb{E}\left\|\mathbf{e}_{k,\tau-1}\right\|^2 \\
&\leq \tau^2\frac{\sigma^2}{n} + \tau L^2\left(a_1 + \cdots + a_{\tau-1}\right) \\
&\leq \tau^2\frac{\sigma^2}{n} + \tau^2(\tau-1)L^2\sigma^2\eta_k^2\left(1+\tau L^2\eta_k^2\right)^{\tau} \\
&\leq \tau^2\frac{\sigma^2}{n} + \tau^2(\tau-1)L^2\sigma^2 e\eta_k^2.
\end{aligned} \tag{49}$$

Now, we can put everything together and conclude Lemma 2, as follows

$$\begin{aligned}
\mathbb{E}\left\|\overline{\mathbf{x}}_{k,\tau} - \mathbf{x}^*\right\|^2 &= \left(1+n\eta_k^2\right)\mathbb{E}\|\mathbf{x}_k - \eta_k\mathbf{g}_k - \mathbf{x}^*\|^2 + \frac{1}{m}\|\mathbb{E}\mathbf{e}_k\|^2 + \eta_k^2\mathbb{E}\|\mathbf{e}_k\|^2 \\
&\leq \left(1+n\eta_k^2\right)(1-\mu\eta_k)^{\tau}\,\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \\
&\quad + \tau(\tau-1)^2L^2\frac{\sigma^2}{n}e\eta_k^2 + \tau^2\frac{\sigma^2}{n}\eta_k^2 \\
&\quad + \tau^2(\tau-1)L^2\sigma^2 e\eta_k^4.
\end{aligned} \tag{50}$$

7.3 Proof of Lemma 3

According to the notations defined on (18), we can write

$$\begin{aligned}\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1}-\overline{\mathbf{x}}_{k,\tau}\right\|^2 &= \mathbb{E}\left\|\mathbf{x}_k+\frac{1}{n}\sum_{i\in[n]}Q\left(\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right)-\frac{1}{n}\sum_{i\in[n]}\mathbf{x}_{k,\tau}^{(i)}\right\|^2\\ &= \mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}Q\left(\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right)-\left(\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right)\right\|^2\\ &= \frac{1}{n^2}\sum_{i\in[n]}\mathbb{E}\left\|Q\left(\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right)-\left(\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right)\right\|^2\\ &\leq q\frac{1}{n^2}\sum_{i\in[n]}\mathbb{E}\left\|\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right\|^2, \end{aligned} \tag{51}$$

where, we used Assumption 1. In particular, the last equality above follows from the fact that the random quatizer is unbiased and the quantizations are carried out independently in each iteration and each worker. Moreover, the last inequality in (51) simply relates the variance of the quantization to its argument. Next, we bound $\mathbb{E}\left\|\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right\|^2$ for each worker $i\in[n]$. From the update rule in Algorithm 1 we have

$$\begin{aligned}\mathbf{x}_{k,\tau}^{(i)} &= \mathbf{x}_k-\eta_k\left(\widetilde{\nabla}f_i\left(\mathbf{x}_{k,0}^{(i)}\right)+\cdots+\widetilde{\nabla}f_i\left(\mathbf{x}_{k,\tau-1}^{(i)}\right)\right)\\ &= \mathbf{x}_k-\eta_k\left(\mathbf{g}_k+\mathbf{e}_k^{(i)}\right), \end{aligned} \tag{52}$$

where we denote

$$\mathbf{e}_k^{(i)} \coloneqq \widetilde{\nabla}f_i\left(\mathbf{x}_{k,0}^{(i)}\right)-\nabla f\left(\beta_{k,0}\right)+\cdots+\widetilde{\nabla}f_i\left(\mathbf{x}_{k,\tau-1}^{(i)}\right)-\nabla f\left(\beta_{k,\tau-1}\right), \tag{53}$$

and $\mathbf{g}_k=\nabla f(\beta_{k,0})+\cdots+\nabla f(\beta_{k,\tau-1})$ as defined before. Using these notations we have

$$\begin{aligned}\mathbb{E}\left\|\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right\|^2 &= \eta_k^2\mathbb{E}\left\|\mathbf{g}_k+\mathbf{e}_k^{(i)}\right\|^2\\ &\leq 2\eta_k^2\|\mathbf{g}_k\|^2+2\eta_k^2\mathbb{E}\left\|\mathbf{e}_k^{(i)}\right\|^2. \end{aligned} \tag{54}$$

Let us first bound the first term in (54), i.e. $\|\mathbf{g}_k\|^2$. That is,

$$\begin{aligned}\|\mathbf{g}_k\|^2 &\leq \tau\left\|\nabla f(\beta_{k,0})\right\|^2+\cdots+\tau\left\|\nabla f(\beta_{k,\tau-1})\right\|^2\\ &\overset{(a)}{\leq} \tau L^2\left(\|\mathbf{x}_k-\mathbf{x}^*\|^2+\cdots+(1-\mu\eta_k)^{\tau-1}\|\mathbf{x}_k-\mathbf{x}^*\|^2\right)\\ &\leq \tau^2L^2\|\mathbf{x}_k-\mathbf{x}^*\|^2, \end{aligned} \tag{55}$$

where we used the smoothness of the loss function f (Assumption 2) and the result in (37) to derive inequality (a). To bound the second term in (54), i.e. $\mathbb{E}\left\|\mathbf{e}_k^{(i)}\right\|^2$, we can employ our result in (49) for the special case $n=1$. It yields that for $\eta_k\leq 1/L\tau$,

$$\mathbb{E}\left\|\mathbf{e}_k^{(i)}\right\|^2 \leq \tau^2\sigma^2+\tau^2(\tau-1)L^2\sigma^2e\eta_k^2. \tag{56}$$

Plugging (55) and (56) in (54) implies that

$$\mathbb{E}\left\|\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right\|^2 \leq 2\tau^2L^2\eta_k^2\mathbb{E}\|\mathbf{x}_k-\mathbf{x}^*\|^2+2\tau^2\sigma^2\eta_k^2+2(\tau-1)\tau^2L^2\sigma^2e\eta_k^4, \tag{57}$$

which together with (51) concludes Lemma 3:

$$\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1}-\overline{\mathbf{x}}_{k,\tau}\right\|^2 \leq 2\frac{q}{n}\tau^2L^2\eta_k^2\mathbb{E}\|\mathbf{x}_k-\mathbf{x}^*\|^2+2q\tau^2\frac{\sigma^2}{n}\eta_k^2+2q(\tau-1)\tau^2L^2\frac{\sigma^2}{n}e\eta_k^4. \tag{58}$$

7.4 Proof of Lemma 4

For each node $i \in [n]$ denote $\mathbf{z}_{k,\tau}^{(i)} = Q(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k)$ and $\overline{\mathbf{z}}_{k,\tau} = \frac{1}{n}\sum_{i\in[n]} \mathbf{z}_{k,\tau}^{(i)}$. Then,

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_k}\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1}\|^2 &= \mathbb{E}_{\mathcal{S}_k}\left\|\frac{1}{r}\sum_{i\in\mathcal{S}_k}\mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}\right\|^2 \\
&= \frac{1}{r^2}\mathbb{E}_{\mathcal{S}_k}\left\|\sum_{i\in[n]}\mathbb{1}\{i\in\mathcal{S}_k\}\left(\mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}\right)\right\|^2 \\
&= \frac{1}{r^2}\Bigg\{\sum_{i\in[n]}\Pr\left[i\in\mathcal{S}_k\right]\left\|\mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}\right\|^2 \\
&\quad + \sum_{i\neq j}\Pr\left[i,j\in\mathcal{S}_k\right]\left\langle \mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}, \mathbf{z}_{k,\tau}^{(j)} - \overline{\mathbf{z}}_{k,\tau}\right\rangle\Bigg\} \\
&= \frac{1}{nr}\sum_{i\in[n]}\left\|\mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}\right\|^2 \\
&\quad + \frac{r-1}{rn(n-1)}\sum_{i\neq j}\left\langle \mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}, \mathbf{z}_{k,\tau}^{(j)} - \overline{\mathbf{z}}_{k,\tau}\right\rangle \\
&= \frac{1}{r(n-1)}\left(1-\frac{r}{n}\right)\sum_{i\in[n]}\left\|\mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}\right\|^2, \qquad (59)
\end{aligned}$$

where we used the fact that $\left\|\mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}\right\|^2 + \sum_{i\neq j}\left\langle \mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}, \mathbf{z}_{k,\tau}^{(j)} - \overline{\mathbf{z}}_{k,\tau}\right\rangle = 0$. Further taking expectation with respect to the quantizer yields

$$\begin{aligned}
\sum_{i\in[n]}\mathbb{E}_Q\left\|\mathbf{z}_{k,\tau}^{(i)} - \overline{\mathbf{z}}_{k,\tau}\right\|^2 &\leq 2\sum_{i\in[n]}\mathbb{E}_Q\left\|\mathbf{z}_{k,\tau}^{(i)}\right\|^2 + 2n\mathbb{E}_Q\|\overline{\mathbf{z}}_{k,\tau}\|^2 \\
&\leq 4\sum_{i\in[n]}\mathbb{E}_Q\left\|\mathbf{z}_{k,\tau}^{(i)}\right\|^2 \\
&= 4\sum_{i\in[n]}\mathbb{E}_Q\left\|Q\left(\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right)\right\|^2 \\
&\leq 4(1+q)\sum_{i\in[n]}\left\|\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right\|^2. \qquad (60)
\end{aligned}$$

In the above derivations, we used the fact that under Assumption 1 and for any $\mathbf{x}$ we have $\mathbb{E}\left\|Q(\mathbf{x})\right\|^2 \leq (1+q)\|\mathbf{x}\|^2$. Therefore, (60) together with the equality derived in (59) yields that

$$\mathbb{E}\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1}\|^2 \leq \frac{1}{r(n-1)}\left(1-\frac{r}{n}\right)4(1+q)\sum_{i\in[n]}\mathbb{E}\left\|\mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k\right\|^2. \qquad (61)$$

Finally, we substitute the bound in (57) into (61) and conclude Lemma 4 as follows:

$$\mathbb{E}\|\mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1}\|^2 \leq \frac{n-r}{r(n-1)}8(1+q)\left\{\tau^2L^2\eta^2\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \tau^2\sigma^2\eta^2 + (\tau-1)\tau^2L^2\sigma^2e\eta^4\right\}. \qquad (62)$$

7.5 Proof of Equation (43)

Let us fix the period k and for simplicity of the notations in this proof, let us take $a_t = a_{k,t}$ and $\eta = \eta_k$. We showed that $a_t \leq \tau\eta^2\sigma^2 + \tau L^2\eta^2\left(a_1 + \cdots + a_{t-1}\right)$ for every $t = 2, \cdots, \tau-1$ and also $a_1 \leq \eta^2\sigma^2$. For $t = 1$, (43) holds. Assume that (43) holds also for $\{a_1, \cdots, a_{t-1}\}$. Now, for a_t we have

$$
\begin{aligned}
a_t &\leq \tau\eta^2\sigma^2 + \tau L^2\eta^2 \left(a_1 + \cdots + a_{t-1}\right) \\
&\leq \tau\eta^2\sigma^2 + \tau L^2\eta^2 \sum_{i=0}^{t-2} \tau\eta^2\sigma^2 \left(1 + \tau L^2\eta^2\right)^i \\
&= \tau\eta^2\sigma^2 + \tau\eta^2\sigma^2 \cdot \tau L^2\eta^2 \cdot \frac{\left(1 + \tau L^2\eta^2\right)^{t-1} - 1}{\tau L^2\eta^2} \\
&= \tau\eta^2\sigma^2 \left(1 + \tau L^2\eta^2\right)^{t-1}, \qquad (63)
\end{aligned}
$$

as desired. Therefore, (43) holds for every $t = 1, \cdots, \tau - 1$.

7.6 Discussion on stepsize η_k

Here we show that for any $k \geq k_0$ we have $C_0 \leq 1 - \frac{1}{2}\mu\tau\eta_k$, where k_0 satisfies the condition in Theorem 1, that is

$$
k_0 \geq 4 \max\left\{\frac{L}{\mu}, 4\left(\frac{B_1}{\mu^2} + 1\right), \frac{1}{\tau}, \frac{4n}{\mu^2\tau}\right\}. \qquad (64)
$$

First note that this condition on k_0 implies the following conditions on the stepsize $\eta_k = {}^{4\mu^{-1}}/_{k\tau+1}$ for $k \geq k_0$:

$$
\eta_k\tau \leq \min\left\{\frac{1}{L}, \frac{\mu}{4\left(\mu^2 + B_1\right)}\right\}, \quad \text{and} \quad \eta_k \leq \min\left\{\frac{\mu}{L^2}, \frac{\mu}{4n}\right\}, \qquad (65)
$$

Now consider the term $(1 - \mu\eta_k)^\tau$ in C_0. We have

$$
\begin{aligned}
(1 - \mu\eta_k)^\tau &= \left(1 - \frac{\mu\tau\eta_k}{\tau}\right)^\tau \\
&\leq e^{-\mu\tau\eta_k} \\
&\leq 1 - \mu\tau\eta_k + \mu^2\tau^2\eta_k^2, \qquad (66)
\end{aligned}
$$

where the first inequality follows from the assumption $\eta_k \leq {}^1/_\mu$ and the second inequality uses the fact that $e^x \leq 1 + x + x^2$ for $x \leq 0$. Therefore,

$$
\begin{aligned}
C_0 &\leq \left(1 + n\eta_k^2\right)\left(1 - \mu\tau\eta_k + \mu^2\tau^2\eta_k^2\right) + B_1\tau^2\eta_k^2 \\
&= 1 - \mu\tau\eta_k + \tau^2\eta_k^2(B_1 + \mu^2) + n\eta_k^2\left(1 - \mu\tau\eta_k + \mu^2\tau^2\eta_k^2\right). \qquad (67)
\end{aligned}
$$

Note that from the assumption $\eta_k \leq {}^1/_{L\tau}$ we have $0 \leq \mu\tau\eta_k \leq {}^\mu/_L \leq 1$. This implies that $1 - \mu\tau\eta_k + \mu^2\tau^2\eta_k^2 \leq 1$. Hence,

$$
C_0 \leq 1 - \mu\tau\eta_k + \tau^2\eta_k^2(B_1 + \mu^2) + n\eta_k^2. \qquad (68)
$$

Now from the condition $\eta_k\tau \leq {}^\mu/_{4(B_1+\mu^2)}$ we have

$$
\tau^2\eta_k^2(B_1 + \mu^2) \leq \frac{1}{4}\mu\tau\eta_k, \qquad (69)
$$

and from $\eta_k \leq {}^\mu/_{4n}$ we have

$$
n\eta_k^2 \leq \frac{1}{4}\mu\tau\eta_k, \qquad (70)
$$

sine $\tau \geq 1$. Plugging (69) and (70) in (68) yields that for any $k \geq k_0$ we have $C_0 \leq 1 - \frac{1}{2}\mu\tau\eta_k$.

7.7 Skipped lemmas and proofs

Lemma 5. *Let a non-negative sequence δ_k satisfy the following*

$$\delta_{k+1} \leq \left(1 - \frac{2}{k+k_1}\right)\delta_k + \frac{a}{(k+k_1)^2} + \frac{b}{(k+k_1)^4}, \tag{71}$$

for every $k \geq k_0$, where a, b, c, k_1 are positive reals and k_0 is a positive integer. Then for every $k \geq k_0$ we have

$$\delta_k \leq \frac{(k_0+k_1)^2}{(k+k_1)^2}\delta_{k_0} + \frac{a}{k+k_1} + \frac{b}{(k+k_1)^2}. \tag{72}$$

Proof. We prove by induction on $k \geq k_0$. The claim in (72) is trivial for $k = k_0$. Let (72) hold for $s \geq k_0$, that is

$$\delta_s \leq \frac{(k_0+k_1)^2}{(s+k_1)^2}\delta_{k_0} + \frac{a}{s+k_1} + \frac{b}{(s+k_1)^2}. \tag{73}$$

We can then write

$$\begin{aligned}
\delta_{s+1} &\leq \left(1 - \frac{2}{s+k_1}\right)\delta_s + \frac{a}{s+k_1} + \frac{b}{(s+k_1)^2} \\
&\leq \left(1 - \frac{2}{s+k_1}\right)\left(\frac{(k_0+k_1)^2}{(s+k_1)^2}\delta_{k_0} + \frac{a}{s+k_1} + \frac{b}{(s+k_1)^2}\right) + \frac{a}{(s+k_1)^2} + \frac{b}{(s+k_1)^4} \\
&= \frac{s+k_1-2}{(s+k_1)^3}(k_0+k_1)^2\delta_{k_0} + \frac{s+k_1-1}{(s+k_1)^2}a + \frac{(s+k_1-1)^2}{(s+k_1)^4}b.
\end{aligned} \tag{74}$$

Now, take $s' = s + k_1$. We have for $s' \geq 1$ that

$$\frac{s'-2}{s'^3} \leq \frac{1}{(s'+1)^2}, \qquad \frac{s'-1}{s'^2} \leq \frac{1}{s'+1}, \qquad \frac{(s'-1)^2}{s'^4} \leq \frac{1}{(s'+1)^2}. \tag{75}$$

Plugging (75) in (74) yields that the claim in (72) holds for $s+1$ and hence for any $k \geq k_0$. □

8 Proof of Theorem 2

We begin the proof of Theorem 2 by noting the following property for any smooth loss function.

Lemma 6. *Consider the sequences of updates $\{\mathbf{x}_{k+1}, \widehat{\mathbf{x}}_{k+1}, \overline{\mathbf{x}}_{k,\tau}\}$ generated by* `FedPAQ` *method in Algorithm 1. If Assumptions 1 and 2 hold, then*

$$\mathbb{E}f(\mathbf{x}_{k+1}) \leq \mathbb{E}f(\overline{\mathbf{x}}_{k,\tau}) + \frac{L}{2}\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau}\right\|^2 + \frac{L}{2}\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\right\|^2, \tag{76}$$

for any period $k = 0, \cdots, K-1$.

Proof. See Section 8.2. □

In the following three lemmas, we bound each of the three terms in the RHS of (76).

Lemma 7. *Let Assumptions 2 and 3 hold and consider the sequence of updates in* `FedPAQ` *method with stepsize η. Then, for every period $k = 0, \cdots, K-1$ we have*

$$\begin{aligned}
\mathbb{E}f(\overline{\mathbf{x}}_{k,\tau}) \leq \mathbb{E}f(\mathbf{x}_k) &- \frac{1}{2}\eta\sum_{t=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}_{k,t})\right\|^2 \\
&- \eta\left(\frac{1}{2n} - \frac{1}{2n}L\eta - \frac{1}{n}L^2\tau(\tau-1)\eta^2\right)\sum_{t=0}^{\tau-1}\sum_{i\in[n]}\mathbb{E}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2 \\
&+ \eta^2\frac{L}{2}\frac{\sigma^2}{n}\tau + \eta^3\frac{\sigma^2}{n}(n+1)\frac{\tau(\tau-1)}{2}L^2.
\end{aligned} \tag{77}$$

Proof. See Section 8.3. □

Lemma 8. *If Assumptions 1 and 3 hold, then for sequences* $\{\widehat{\mathbf{x}}_{k+1}, \overline{\mathbf{x}}_{k,\tau}\}$ *defined in* (18) *we have*

$$\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau}\right\|^2 \leq q\frac{\sigma^2}{n}\tau\eta^2 + q\frac{1}{n^2}\tau\eta^2 \sum_{i\in[n]}\sum_{t=0}^{\tau-1}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2. \tag{78}$$

Proof. See Section 8.4. □

Lemma 9. *Under Assumptions 1 and 3, for the sequence of averages* $\{\widehat{\mathbf{x}}_{k+1}\}$ *defined in* (18) *we have*

$$\mathbb{E}\|\widehat{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\|^2 \leq \frac{1}{r(n-1)}\left(1-\frac{r}{n}\right)4(1+q)\left\{n\sigma^2\tau\eta^2 + \tau\eta^2\sum_{i\in[n]}\sum_{t=0}^{\tau-1}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2\right\}. \tag{79}$$

Proof. See Section 8.5. □

After establishing the main building modules in the above lemmas, we now proceed to prove the convergence rate in Theorem 2. In particular, we combine the results in Lemmas 6–9 to derive the following recursive inequality on the expected function value on the models updated at the parameter servers, i.e. $\{\mathbf{x}_k : k = 1, \cdots, K\}$:

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_{k+1}) \leq\ & \mathbb{E}f(\mathbf{x}_k) \\
& - \frac{1}{2}\eta\sum_{t=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}_{k,t})\right\|^2 \\
& - \eta\frac{1}{2n}\left(1 - L\left(1 + \frac{1}{n}q\tau + 4\frac{n-r}{r(n-1)}(1+q)\tau\right)\eta - 2L^2\tau(\tau-1)\eta^2\right)\sum_{t=0}^{\tau-1}\sum_{i\in[n]}\mathbb{E}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2 \\
& + \eta^2\frac{L}{2}(1+q)\tau\left(\frac{\sigma^2}{m} + 4\frac{\sigma^2}{r}\frac{n-r}{n-1}\right) + \eta^3\frac{\sigma^2}{m}(m+1)\frac{\tau(\tau-1)}{2}L^2.
\end{aligned} \tag{80}$$

For sufficiently small η, such that

$$1 - L\eta - L\left(\frac{1}{n}q + 4\frac{n-r}{r(n-1)}(1+q)\right)\tau\eta - 2L^2\tau(\tau-1)\eta^2 \geq 0, \tag{81}$$

we have

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_{k+1}) \leq\ & \mathbb{E}f(\mathbf{x}_k) - \frac{1}{2}\eta\sum_{t=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}_{k,t})\right\|^2 \\
& + \eta^2\frac{L}{2}(1+q)\tau\left(\frac{\sigma^2}{n} + 4\frac{\sigma^2}{r}\frac{n-r}{n-1}\right) + \eta^3\frac{\sigma^2}{n}(n+1)\frac{\tau(\tau-1)}{2}L^2.
\end{aligned} \tag{82}$$

In Section 8.1 we show that if the stepsize is picked as $\eta = 1/L\sqrt{T}$ and the T ans τ satisfy the condition (16) in Theorem 2, then (81) also holds. Now summing (82) over $k = 0, \cdots, K-1$ and rearranging the terms yield that

$$\begin{aligned}
& \frac{1}{2}\eta\sum_{k=0}^{K-1}\sum_{t=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}_{k,t})\right\|^2 \\
& \leq f(\mathbf{x}_0) - f^* + K\eta^2\frac{L}{2}(1+q)\tau\left(\frac{\sigma^2}{n} + 4\frac{\sigma^2}{r}\frac{n-r}{n-1}\right) + K\eta^3\frac{\sigma^2}{n}(n+1)\frac{\tau(\tau-1)}{2}L^2,
\end{aligned} \tag{83}$$

or

$$\begin{aligned}
& \frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{t=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}_{k,t})\right\|^2 \\
& \leq \frac{2(f(\mathbf{x}_0) - f^*)}{\eta K\tau} + \eta L(1+q)\left(\frac{\sigma^2}{n} + 4\frac{\sigma^2}{r}\frac{n-r}{n-1}\right) + \eta^2\frac{\sigma^2}{n}(n+1)(\tau-1)L^2.
\end{aligned} \tag{84}$$

Picking the stepsize $\eta = 1/L\sqrt{T} = 1/L\sqrt{K\tau}$ results in the following convergence rate:

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{t=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}_{k,t})\right\|^2 \\ \leq \frac{2L(f(\mathbf{x}_0)-f^*)}{\sqrt{T}} + (1+q)\left(\frac{\sigma^2}{n}+\frac{\sigma^2}{r}\frac{n-r}{n-1}\right)\frac{1}{\sqrt{T}} + \frac{\sigma^2}{n}(n+1)\frac{\tau-1}{T}, \tag{85}$$

which completes the proof of Theorem 2.

8.1 Discussion on stepsize η

Here, we consider the constraint on the stepsize derived in (81) and show that if η is picked according to Theorem 2, then it also satisfies (81). First, let the stepsize satisfy $1 - L\eta \geq 0.1$. Now, if the following holds

$$L\left(\frac{1}{n}q + 4\frac{n-r}{r(n-1)}(1+q)\right)\tau\eta + 2L^2(\tau\eta)^2 \leq 0.1, \tag{86}$$

the condition in (81) also holds. It is straightforward to see when (86) holds. To do so, consider the following quadratic inequality in terms of $y = \eta\tau$:

$$2L^2y^2 + LB_2y - 0.1 \leq 0, \tag{87}$$

where

$$B_2 \coloneqq \frac{1}{n}q + 4\frac{n-r}{r(n-1)}(1+q). \tag{88}$$

We can solve the quadratic form in (87) for $y = \eta\tau$ which yields

$$\eta\tau \leq \frac{\sqrt{B_2^2+0.8}-B_2}{4L}. \tag{89}$$

This implies that if the parameter τ and the stepsize η satisfy (89) and $\eta \leq 0.9/L$, then the condition (81) is satisfied. In particular, for our pick of $\eta = 1/L\sqrt{T}$, the condition $\eta \leq 0.9/L$ holds if $T \geq 2$; and the constraint in (89) is equivalent to having

$$\tau \leq \frac{\sqrt{B_2^2+0.8}-B_2}{8}\sqrt{T}. \tag{90}$$

8.2 Proof of Lemma 6

Recall that for any L-smooth function f and variables $\mathbf{x}, \mathbf{y}$ we have

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \left\langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\right\rangle + \frac{L}{2}\left\|\mathbf{x}-\mathbf{y}\right\|^2. \tag{91}$$

Therefore, we can write

$$f(\mathbf{x}_{k+1}) = f(\widehat{\mathbf{x}}_{k+1} + \mathbf{x}_{k+1} - \widehat{\mathbf{x}}_{k+1}) \\ \leq f(\widehat{\mathbf{x}}_{k+1}) + \left\langle \nabla f(\widehat{\mathbf{x}}_{k+1}), \mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1}\right\rangle + \frac{L}{2}\left\|\mathbf{x}_{k+1}-\widehat{\mathbf{x}}_{k+1}\right\|^2. \tag{92}$$

We take expectation of both sides of (92) and since $\widehat{\mathbf{x}}_{k+1}$ is unbiased for $\mathbf{x}_{k+1}$, that is $\mathbb{E}_{\mathcal{S}_k}\mathbf{x}_{k+1} = \widehat{\mathbf{x}}_{k+1}$ (See (29)), it yields that

$$\mathbb{E}f(\mathbf{x}_{k+1}) \leq \mathbb{E}f(\widehat{\mathbf{x}}_{k+1}) + \frac{L}{2}\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1}-\mathbf{x}_{k+1}\right\|^2. \tag{93}$$

Moreover, $\widehat{\mathbf{x}}_{k+1}$ is also unbiased for $\overline{\mathbf{x}}_{k,\tau}$, i.e. $\mathbb{E}_Q\widehat{\mathbf{x}}_{k+1} = \overline{\mathbf{x}}_{k,\tau}$ (See (30)), and since f is L-smooth, we can write

$$\mathbb{E}f(\widehat{\mathbf{x}}_{k+1}) \leq \mathbb{E}f(\overline{\mathbf{x}}_{k,\tau}) + \frac{L}{2}\mathbb{E}\left\|\widehat{\mathbf{x}}_{k+1}-\overline{\mathbf{x}}_{k,\tau}\right\|^2, \tag{94}$$

which together with (93) concludes the lemma.

8.3 Proof of Lemma 7

According to the update rule in Algorithm 1, for every $t = 0, \cdots, \tau - 1$ the average model is

$$\overline{\mathbf{x}}_{k,t+1} = \overline{\mathbf{x}}_{k,t} - \eta \frac{1}{n} \sum_{i \in [n]} \widetilde{\nabla} f_i \left(\mathbf{x}_{k,t}^{(i)} \right). \tag{95}$$

Since f is L-smooth, we can write

$$f(\overline{\mathbf{x}}_{k,t+1}) \leq f(\overline{\mathbf{x}}_{k,t}) - \eta \left\langle \nabla f(\overline{\mathbf{x}}_{k,t}), \frac{1}{n} \sum_{i \in [n]} \widetilde{\nabla} f_i \left(\mathbf{x}_{k,t}^{(i)} \right) \right\rangle + \eta^2 \frac{L}{2} \left\| \frac{1}{n} \sum_{i \in [n]} \widetilde{\nabla} f_i \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2. \tag{96}$$

The inner product term above can be written in expectation as follows:

$$\begin{aligned} 2\mathbb{E} \left\langle \nabla f(\overline{\mathbf{x}}_{k,t}), \frac{1}{n} \sum_{i \in [n]} \widetilde{\nabla} f_i \left(\mathbf{x}_{k,t}^{(i)} \right) \right\rangle &= \frac{1}{n} \sum_{i \in [n]} 2\mathbb{E} \left\langle \nabla f(\overline{\mathbf{x}}_{k,t}), \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\rangle \\ &= \mathbb{E} \left\| \nabla f(\overline{\mathbf{x}}_{k,t}) \right\|^2 + \frac{1}{n} \sum_{i \in [n]} \mathbb{E} \left\| \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 \\ &\quad - \frac{1}{n} \sum_{i \in [n]} \mathbb{E} \left\| \nabla f(\overline{\mathbf{x}}_{k,t}) - \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2, \end{aligned} \tag{97}$$

where we used the identity $2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$ for any two vectors $\mathbf{a}, \mathbf{b}$. In the following, we bound each of the three terms in the RHS of (97). Starting with the third term, we use the smoothness assumption to write

$$\left\| \nabla f(\overline{\mathbf{x}}_{k,t}) - \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 \leq L^2 \left\| \overline{\mathbf{x}}_{k,t} - \mathbf{x}_{k,t}^{(i)} \right\|^2. \tag{98}$$

Moreover, local models $\mathbf{x}_{k,t}^{(i)}$ and average model $\overline{\mathbf{x}}_{k,t}$ are respectively

$$\mathbf{x}_{k,t}^{(i)} = \mathbf{x}_k - \eta \left(\widetilde{\nabla} f_i(\mathbf{x}_k) + \widetilde{\nabla} f_i \left(\mathbf{x}_{k,1}^{(i)} \right) + \cdots + \widetilde{\nabla} f_i \left(\mathbf{x}_{k,t-1}^{(i)} \right) \right), \tag{99}$$

and

$$\overline{\mathbf{x}}_{k,t} = \mathbf{x}_{k\tau} - \eta \left(\frac{1}{n} \sum_{j \in [n]} \widetilde{\nabla} f_j(\mathbf{x}_k) + \frac{1}{n} \sum_{j \in [n]} \widetilde{\nabla} f_j \left(\mathbf{x}_{k,1}^{(j)} \right) + \cdots + \frac{1}{n} \sum_{j \in [n]} \widetilde{\nabla} f_j \left(\mathbf{x}_{k,t-1}^{(j)} \right) \right). \tag{100}$$

Therefore, the expected deviation of each local model form the average model can be written as

$$\begin{aligned}
&\mathbb{E}\left\|\overline{\mathbf{x}}_{k,t}-\mathbf{x}_{k,t}^{(i)}\right\|^2\\
&\le 2\eta^2\mathbb{E}\left\|\frac{1}{n}\sum_{j\in[n]}\widetilde{\nabla}f_j(\mathbf{x}_k)+\frac{1}{n}\sum_{j\in[n]}\widetilde{\nabla}f_j\left(\mathbf{x}_{k,1}^{(j)}\right)+\cdots+\frac{1}{n}\sum_{j\in[n]}\widetilde{\nabla}f_j\left(\mathbf{x}_{k,t-1}^{(j)}\right)\right\|^2\\
&\quad+2\eta^2\mathbb{E}\left\|\widetilde{\nabla}f_i(\mathbf{x}_k)+\widetilde{\nabla}f_i\left(\mathbf{x}_{k,1}^{(i)}\right)+\cdots+\widetilde{\nabla}f_i\left(\mathbf{x}_{k,t-1}^{(i)}\right)\right\|^2\\
&\le 2\eta^2\left(t\frac{\sigma^2}{n}+\left\|\frac{1}{n}\sum_{j\in[n]}\nabla f(\mathbf{x}_k)+\frac{1}{n}\sum_{j\in[n]}\nabla f\left(\mathbf{x}_{k,1}^{(j)}\right)+\cdots+\frac{1}{n}\sum_{j\in[n]}\nabla f\left(\mathbf{x}_{k,t-1}^{(j)}\right)\right\|^2\right)\\
&\quad+2\eta^2\left(t\sigma^2+\left\|\nabla f(\mathbf{x}_k)+\nabla f\left(\mathbf{x}_{k,1}^{(i)}\right)+\cdots+\nabla f\left(\mathbf{x}_{k,t-1}^{(i)}\right)\right\|^2\right)\\
&\le 2\eta^2t\frac{\sigma^2}{n}+2\eta^2t\left(\frac{1}{n}\sum_{j\in[n]}\left\|\nabla f(\mathbf{x}_k)\right\|^2+\frac{1}{n}\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,1}^{(j)}\right)\right\|^2+\cdots+\frac{1}{n}\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,t-1}^{(j)}\right)\right\|^2\right)\\
&\quad+2\eta^2t\sigma^2+2\eta^2t\left(\left\|\nabla f(\mathbf{x}_k)\right\|^2+\left\|\nabla f\left(\mathbf{x}_{k,1}^{(i)}\right)\right\|^2+\cdots+\left\|\nabla f\left(\mathbf{x}_{k,t-1}^{(i)}\right)\right\|^2\right).
\end{aligned}\tag{101}$$

Summing (101) over all the workers $i\in[n]$ yields

$$\begin{aligned}
&\sum_{i\in[n]}\mathbb{E}\left\|\overline{\mathbf{x}}_{k,t}-\mathbf{x}_{k,t}^{(i)}\right\|^2\\
&\le 2\eta^2t\sigma^2+2\eta^2t\left(\sum_{j\in[n]}\left\|\nabla f(\mathbf{x}_k)\right\|^2+\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,1}^{(j)}\right)\right\|^2+\cdots+\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,t-1}^{(j)}\right)\right\|^2\right)\\
&\quad+2\eta^2t\sigma^2n+2\eta^2t\left(\sum_{i\in[n]}\left\|\nabla f(\mathbf{x}_k)\right\|^2+\sum_{i\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,1}^{(i)}\right)\right\|^2+\cdots+\sum_{i\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,t-1}^{(i)}\right)\right\|^2\right)\\
&=2\eta^2t\sigma^2(n+1)+4\eta^2t\left(\sum_{j\in[n]}\left\|\nabla f(\mathbf{x}_k)\right\|^2+\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,1}^{(j)}\right)\right\|^2+\cdots+\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,t-1}^{(j)}\right)\right\|^2\right).
\end{aligned}\tag{102}$$

Finally, summing (102) over $t=0,\cdots,\tau-1$ results in the following:

$$\begin{aligned}
&\sum_{t=0}^{\tau-1}\sum_{i\in[n]}\mathbb{E}\left\|\overline{\mathbf{x}}_{k,t}-\mathbf{x}_{k,t}^{(i)}\right\|^2\\
&\le 2\eta^2\sigma^2(n+1)\sum_{t=0}^{\tau-1}t+4\eta^2\sum_{t=0}^{\tau-1}t\left(\sum_{j\in[n]}\left\|\nabla f(\mathbf{x}_k)\right\|^2+\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,1}^{(j)}\right)\right\|^2+\cdots+\sum_{j\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,t-1}^{(j)}\right)\right\|^2\right)\\
&\le \eta^2\sigma^2(n+1)\tau(\tau-1)+2\eta^2\tau(\tau-1)\sum_{t=0}^{\tau-2}\sum_{i\in[n]}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2.
\end{aligned}\tag{103}$$

Next, we bound the third term in (96). Using Assumption 3 we have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}\widetilde{\nabla}f_i\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2&=\mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2+\mathbb{E}\left\|\frac{1}{n}\sum_{i\in[n]}\widetilde{\nabla}f_i\left(\mathbf{x}_{k,t}^{(i)}\right)-\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2\\
&\le\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2+\frac{\sigma^2}{n}.
\end{aligned}\tag{104}$$

Summing (104) over iterations $t = 0, \cdots, \tau - 1$ yields

$$\sum_{t=0}^{\tau-1} \eta^2 \frac{L}{2} \mathbb{E} \left\| \frac{1}{n} \sum_{i \in [n]} \widetilde{\nabla} f_i \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 \leq \eta^2 \frac{L}{2} \frac{1}{n} \sum_{t=0}^{\tau-1} \sum_{i \in [n]} \mathbb{E} \left\| \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 + \eta^2 \frac{L}{2} \frac{\sigma^2}{n} \tau. \tag{105}$$

Now we can sum (96) for $t = 0, \cdots, \tau - 1$ and use the results in (103) and (105) to conclude:

$$\begin{aligned}
\mathbb{E} f(\overline{\mathbf{x}}_{k,\tau}) \leq \mathbb{E} f(\mathbf{x}_k) &- \frac{1}{2} \eta \sum_{t=0}^{\tau-1} \mathbb{E} \left\| \nabla f(\overline{\mathbf{x}}_{k,t}) \right\|^2 \\
&- \frac{1}{2n} \eta \sum_{t=0}^{\tau-1} \sum_{i \in [n]} \mathbb{E} \left\| \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 \\
&+ \frac{1}{2n} \eta \sum_{t=0}^{\tau-1} \sum_{i \in [n]} \mathbb{E} \left\| \nabla f(\overline{\mathbf{x}}_{k,t}) - \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 \\
&+ \sum_{t=0}^{\tau-1} \eta^2 \frac{L}{2} \mathbb{E} \left\| \frac{1}{n} \sum_{i \in [n]} \widetilde{\nabla} f_i \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 \\
\leq \mathbb{E} f(\mathbf{x}_k) &- \frac{1}{2} \eta \sum_{t=0}^{\tau-1} \mathbb{E} \left\| \nabla f(\overline{\mathbf{x}}_{k,t}) \right\|^2 \\
&- \eta \left(\frac{1}{2n} - \frac{1}{2n} L \eta - \frac{1}{n} L^2 \tau (\tau - 1) \eta^2 \right) \sum_{t=0}^{\tau-1} \sum_{i \in [n]} \mathbb{E} \left\| \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2 \\
&+ \eta^2 \frac{L}{2} \frac{\sigma^2}{n} \tau + \eta^3 \frac{\sigma^2}{n} (n+1) \frac{\tau(\tau-1)}{2} L^2.
\end{aligned} \tag{106}$$

8.4 Proof of Lemma 8

According to definitions in (18) and using Assumption 1 we have

$$\mathbb{E} \left\| \widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau} \right\|^2 \leq \frac{1}{n^2} \sum_{i \in [n]} q \, \mathbb{E} \left\| \mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k \right\|^2. \tag{107}$$

Using the model update in (99) and Assumption 3, we can write

$$\begin{aligned}
\mathbb{E} \left\| \mathbf{x}_{k,\tau}^{(i)} - \mathbf{x}_k \right\|^2 &= \eta^2 \mathbb{E} \left\| \widetilde{\nabla} f_i(\mathbf{x}_k) + \widetilde{\nabla} f_i \left(\mathbf{x}_{k,1}^{(i)} \right) + \cdots + \widetilde{\nabla} f_i \left(\mathbf{x}_{k,\tau-1}^{(i)} \right) \right\|^2 \\
&= \eta^2 \mathbb{E} \left\| \widetilde{\nabla} f_i(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) + \cdots + \widetilde{\nabla} f_i \left(\mathbf{x}_{k,\tau-1}^{(i)} \right) - \nabla f \left(\mathbf{x}_{k,\tau-1}^{(i)} \right) \right\|^2 \\
&\quad + \eta^2 \left\| \nabla f(\mathbf{x}_{k\tau}) + \cdots + \nabla f \left(\mathbf{x}_{k,\tau-1}^{(i)} \right) \right\|^2 \\
&\leq \eta^2 \sigma^2 \tau + \eta^2 \tau \sum_{t=0}^{\tau-1} \left\| \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2.
\end{aligned} \tag{108}$$

Summing (108) over all workers $i \in [n]$ and using (107) yields

$$\mathbb{E} \left\| \widehat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k,\tau} \right\|^2 \leq q \frac{\sigma^2}{n} \tau \eta^2 + q \frac{1}{n^2} \tau \eta^2 \sum_{i \in [n]} \sum_{t=0}^{\tau-1} \left\| \nabla f \left(\mathbf{x}_{k,t}^{(i)} \right) \right\|^2, \tag{109}$$

as desired in Lemma 8.

8.5 Proof of Lemma 9

The steps to prove the bound in (61) for strongly convex losses in Lemma 4 can also be applied for non-convex losses. That is, we can use (61) and together with (108) conclude the following:

$$\mathbb{E}\|\widehat{\mathbf{x}}_{k+1}-\mathbf{x}_{k+1}\|^2 \leq \frac{1}{r(n-1)}\left(1-\frac{r}{n}\right)4(1+q)\sum_{i\in[n]}\left\|\mathbf{x}_{k,\tau}^{(i)}-\mathbf{x}_k\right\|^2$$
$$\leq \frac{1}{r(n-1)}\left(1-\frac{r}{n}\right)4(1+q)\left\{n\sigma^2\tau\eta^2+\tau\eta^2\sum_{i\in[n]}\sum_{t=0}^{\tau-1}\left\|\nabla f\left(\mathbf{x}_{k,t}^{(i)}\right)\right\|^2\right\}. \tag{110}$$

9 Additional Numerical Results

To further illustrate the practical performance of the proposed `FedPAQ` method, in this section we provide more numerical results using different and more complicated datasets and model parameters. The network settings, communication and computation time models remain the same as those in Section 5. The following figures demonstrate the training time corresponding to the following scenarios:

- Figure 2: Training time of a neural network with four hidden layers and more than 248K parameters over 10K samples of the CIFAR-10 dataset with 10 labels.
- Figure 3: Training time of a neural network with one hidden layer over 10K samples of the CIFAR-100 dataset with 100 labels.
- Figure 4: Training time of a neural network with one hidden layer over 10K samples of the Fashion-MNIST dataset with 10 labels.

Similar to Section 5.2, in all of the above scenarios, the data samples are uniformly distributed among $n = 50$ nodes. We also keep the communication-computation ratio and the batchsize to be $C_{\texttt{comm}}/C_{\texttt{comp}} = 1000/1$ and $B = 10$ respectively, and finely tune the stepsize for every training.

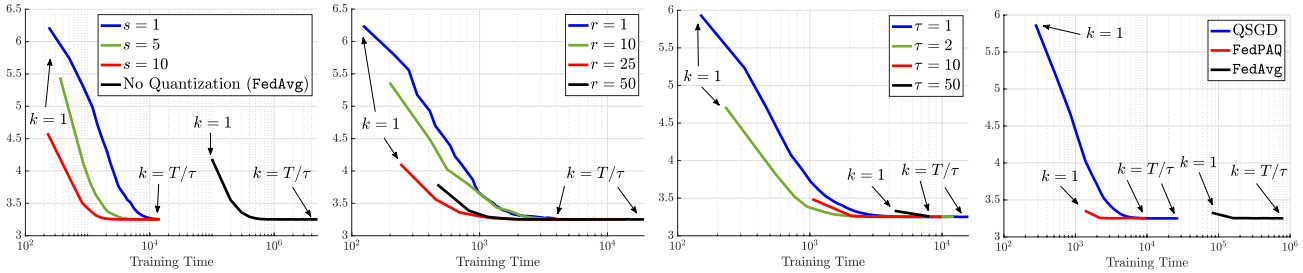

Figure 2: Training Loss vs. Training Time: Neural Network on CIFAR-10 dataset with 248K parameters.

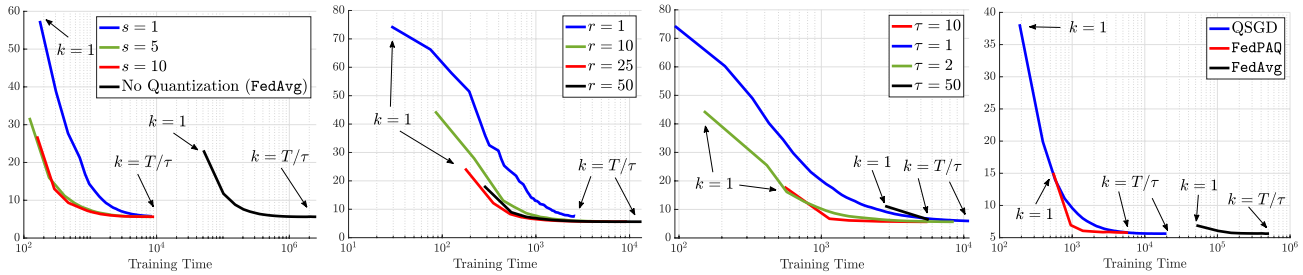

Figure 3: Training Loss vs. Training Time: Neural Network on CIFAR-100 dataset.

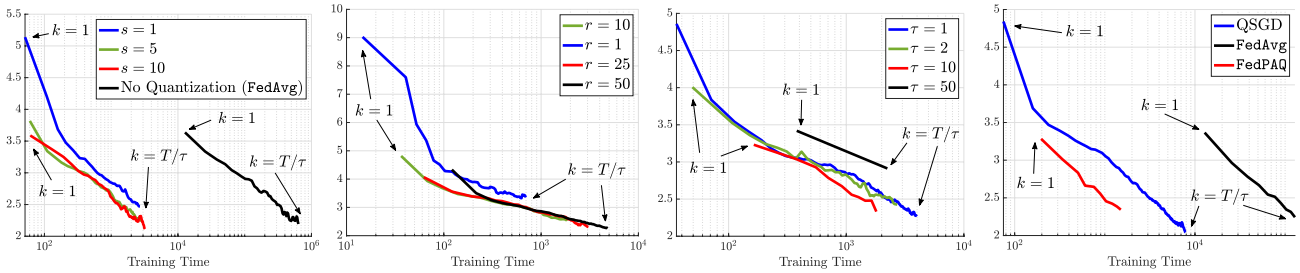

Figure 4: Training Loss vs. Training Time: Neural Network on Fashion-MNIST dataset.