Supplementary for Robust Over-The-Air Federated Learning In Heterogeneous Networks

Zubair Shaban, Nazreen Shah, Ranjitha Prasad

I. Detailed Proofs of Key Lemmas and Theorems

Proof of Lemma 1: Adding $\lambda \mathbf{I}_d$ on the LHS of $\nabla^2 f_k(\boldsymbol{\theta}) \geq -\bar{L}\mathbf{I}_d$, and using the definition of $\bar{\mu}$, we obtain

$$\nabla^2 f_k(\boldsymbol{\theta}) + \lambda \mathbf{I}_d \geq \bar{\mu}\mathbf{I}_d. \tag{1}$$

Using the expression for $h_k(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$ as given in definition 1, we have $\nabla^2 h_k(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \nabla^2 f_k(\boldsymbol{\theta}) + \lambda \mathbf{I}_d$. Substituting in the above, we have $\nabla^2 h_k(\boldsymbol{\theta}, \boldsymbol{\theta}^t) \geq \bar{\mu}\mathbf{I}_d$, which implies that $h_k(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$ is $\bar{\mu}$-strongly convex for all t.

Proof of Lemma 2: Using the reverse triangular inequality for two vectors $\boldsymbol{\theta}^t$ and $\tilde{\boldsymbol{\theta}}^t$, and Lipschitz smoothness, we have:

$$\begin{aligned}\|\nabla f(\boldsymbol{\theta}^t)\| - \|\nabla f(\tilde{\boldsymbol{\theta}}^t)\| &\leq \|\nabla f(\boldsymbol{\theta}^t) - \nabla f(\tilde{\boldsymbol{\theta}}^t)\| \\ &\leq L\|\boldsymbol{\theta}^t - \tilde{\boldsymbol{\theta}}^t\|. \end{aligned}\tag{2}$$

Since $\mathbf{w}^t = \tilde{\boldsymbol{\theta}}^t - \boldsymbol{\theta}^t$, we have $\|\nabla f(\boldsymbol{\theta}^t)\| \leq \|\nabla f(\tilde{\boldsymbol{\theta}}^t)\| + L\|\mathbf{w}^t\|$. The same result is obtained (as an approximation) by using Taylor series expansion as follows:

$$f(\boldsymbol{\theta}^t) = f(\tilde{\boldsymbol{\theta}}^t - \mathbf{w}^t) \approx f(\tilde{\boldsymbol{\theta}}^t) - \mathbf{w}^t \nabla f(\tilde{\boldsymbol{\theta}}^t). \tag{3}$$

Differentiating both sides of the above equation and considering the norm, we have

$$\begin{aligned}\|\nabla f(\boldsymbol{\theta}^t)\| &\lessapprox \|\nabla f(\tilde{\boldsymbol{\theta}}^t)\| + \|\mathbf{w}^t\|\|\nabla^2 f(\tilde{\boldsymbol{\theta}}^t)\| \\ &\leq \|\nabla f(\tilde{\boldsymbol{\theta}}^t)\| + \|\mathbf{w}^t\| L, \end{aligned}\tag{4}$$

where the last step holds by the spectral norm property, i,e., $\|\nabla^2 f(\tilde{\boldsymbol{\theta}}^t)\| \leq L$ if f satisfies $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Proof of Lemma 3: The local objective function, as given in P2 is defined as follows,

$$h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t) = f_k(\boldsymbol{\theta}_k^{t+1}) + \frac{\lambda}{2}\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\|^2. \tag{5}$$

where $\tilde{\boldsymbol{\theta}}^t$ is the available aggregated global model from the t-th aggregation epoch. From (5) in the main manuscript, we have the noisy FedAvg decoding rule given as $\tilde{\boldsymbol{\theta}}^t = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_k^t + \mathbf{w}^t$. Differentiating (5) with respect to $\boldsymbol{\theta}_k^{t+1}$, we obtain

$$\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t) = \nabla f_k(\boldsymbol{\theta}_k^{t+1}) + \lambda\left[\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\right]. \tag{6}$$

We introduce the noiseless parameter update as $\boldsymbol{\theta}^t = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\theta}_k^t$, which leads to $\tilde{\boldsymbol{\theta}}^t = \boldsymbol{\theta}^t + \mathbf{w}^t$. Considering ℓ_2 norm of both the sides of the above expression, we have

$$\|\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\| = \|\nabla f_k(\boldsymbol{\theta}_k^{t+1}) + \lambda\left(\boldsymbol{\theta}_k^{t+1} - \boldsymbol{\theta}^t\right) - \lambda\mathbf{w}^t\|. \tag{7}$$

Applying triangle inequality to the above, we obtain

$$\begin{aligned}\|\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\| &\leq \|\nabla f_k(\boldsymbol{\theta}_k^{t+1}) + \lambda\left(\boldsymbol{\theta}_k^{t+1} - \boldsymbol{\theta}^t\right)\| + \lambda\|\mathbf{w}^t\|, \\ &\leq \|\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \boldsymbol{\theta}^t)\| + \lambda\|\mathbf{w}^t\|. \end{aligned}\tag{8}$$

Using the notion of inexactness as mentioned in definition 1, we have $\|\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \boldsymbol{\theta}^t)\| \leq \gamma\|\nabla f_k(\boldsymbol{\theta}^t)\|$, the expression in (8) can be rewritten as

$$\|\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\| \leq \gamma\|\nabla f_k(\boldsymbol{\theta}^t)\| + \lambda\|\mathbf{w}^t\| \tag{9}$$

Finally, using Lemma 2, we have

$$\begin{aligned}\|\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\| &\leq \gamma\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\| + \gamma L\|\mathbf{w}^t\| + \lambda\|\mathbf{w}^t\| \\ &\leq \gamma\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\| + (\gamma L + \lambda)\|\mathbf{w}^t\| \end{aligned}\tag{10}$$

Proof of Lemma 4: From Assumption 4, we have

$$\begin{aligned}\frac{1}{p^t} &\leq \frac{1}{P}\max_k \mathbb{E}_k[\|\nabla f_{j_k^t}(\tilde{\boldsymbol{\theta}}^t)\|^2] \\ &\leq \frac{1}{P}\kappa\mathbb{E}_k\left[\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\|^2\right] \\ &\leq \frac{1}{P}\kappa B^2\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\|^2, \end{aligned}\tag{11}$$

where κ is a constant. Note that in COTAF, κ is a function of the number of SGD epochs and learning rate [1]. Further, the last inequality follows from Assumption 1.

Since in the case of full participation, $\mathbf{w}^t \sim \mathcal{N}(0, \frac{\sigma^2}{K^2 p^t}\mathbf{I}_d)$, therefore

$$\mathbb{E}_{\mathbf{w}^t}\left[\|\mathbf{w}^t\|^2\right] = \frac{d\sigma^2}{K^2 p^t}. \tag{12}$$

Then, using Lemma 4, we have

$$\mathbb{E}_{\mathbf{w}^t}\left[\|\mathbf{w}^t\|^2\right] = \frac{d\sigma^2}{K^2 p^t} \leq \frac{\kappa d\sigma^2 B^2}{K^2 P}\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\|^2. \tag{13}$$

Using Jensen's inequality, we can rewrite the above as

$$\mathbb{E}_{\mathbf{w}^t}\left[\|\mathbf{w}^t\|\right] \leq \sqrt{\mathbb{E}_{\mathbf{w}^t}\left[\|\mathbf{w}^t\|^2\right]} \leq \frac{\sqrt{\kappa}\sqrt{d}\sigma B}{K\sqrt{P}}\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \tag{14}$$

Furthermore, in partial participation case, $\mathbf{w}_p^t \sim \mathcal{N}(0, \frac{\sigma^2}{\hat{K}^2 p^t}\mathbf{I}_d)$, hence

$$\mathbb{E}_{\mathbf{w}_p^t}\left[\|\mathbf{w}_p^t\|^2\right] = \frac{d\sigma^2}{\hat{K}^2 p^t}. \tag{15}$$

Then using Lemma 4 and Jensen's inequality, we have

$$\mathbb{E}_{\mathbf{w}_p^t}\left[\|\mathbf{w}_p^t\|\right] \leq \sqrt{\mathbb{E}_{\mathbf{w}_p^t}\left[\|\mathbf{w}_p^t\|^2\right]} \leq \frac{\sqrt{\kappa}\sqrt{d}\sigma B}{\hat{K}\sqrt{P}}\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \tag{16}$$

† Indraprastha Institute of Information Technology Delhi, New Delhi
∗ Equal Contribution.

Similarly, for fading case, $\mathbf{w}_f^t \sim \mathcal{N}(0, \frac{\sigma^2}{|\mathcal{K}^t|^2 h_{min}^2 p^t}\mathbf{I}_d)$, therefore

$$\mathbb{E}_{\mathbf{w}_f^t}\left[\|\mathbf{w}_f^t\|^2\right] = \frac{d\sigma^2}{|\mathcal{K}^t|^2 h_{min}^2 p^t}. \tag{17}$$

Then using Lemma 4 and Jensen's inequality, we have

$$\mathbb{E}_{\mathbf{w}_f^t}\left[\|\mathbf{w}_f^t\|\right] \leq \sqrt{\mathbb{E}_{\mathbf{w}_f^t}\left[\|\mathbf{w}_f^t\|^2\right]} \leq \frac{\sqrt{\kappa}\sqrt{d}\sigma B}{|\mathcal{K}^t| h_{min}\sqrt{P}}\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \tag{18}$$

Proof of Theorem 1: Consider the local objective function in (15) of the main manuscript as follows,

$$h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t) = f_k(\boldsymbol{\theta}_k^{t+1}) + \frac{\lambda}{2}\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\|^2. \tag{19}$$

Denoting $\bar{\boldsymbol{\theta}}^{t+1} = \mathbb{E}_k[\boldsymbol{\theta}_k^{t+1}]$ and differentiating the above equation and taking the expectation $\mathbb{E}_k[\cdot]$, we obtain the following:

$$\begin{aligned}\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t =& \frac{-1}{\lambda}\mathbb{E}_k\left[\nabla f_k(\boldsymbol{\theta}_k^{t+1})\right] + \frac{1}{\lambda}\mathbb{E}_k\left[\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\right] \\ &+ \mathbf{w}^t,\end{aligned} \tag{20}$$

where $\mathbb{E}_k[\tilde{\boldsymbol{\theta}}^t] = \boldsymbol{\theta}^t + \mathbf{w}^t$.

From Lemma 1, we know $h_k(\cdot,\cdot)$ is $\bar{\mu}$-strongly convex. Let $\boldsymbol{\theta}_k^{*,t+1} = \arg\min_{\boldsymbol{\theta}} \nabla h_k(\boldsymbol{\theta}; \boldsymbol{\theta}^t)$. Using $\bar{\mu}$-strong convexity of $h_k(\cdot,\cdot)$ and (9) we obtain

$$\|\boldsymbol{\theta}_k^{*,t+1} - \boldsymbol{\theta}_k^{t+1}\| \leq \frac{\gamma}{\bar{\mu}}\|\nabla f_k(\boldsymbol{\theta}^t)\| + \frac{\lambda}{\bar{\mu}}\|\mathbf{w}^t\|. \tag{21}$$

Directly from $\bar{\mu}$-strong convexity of $h_k(\cdot)$ we have that

$$\|\boldsymbol{\theta}_k^{*,t+1} - \tilde{\boldsymbol{\theta}}^t\| \leq \frac{1}{\bar{\mu}}\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\|. \tag{22}$$

Combining (21) and (22) and using triangle inequality we obtain

$$\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\| \leq \frac{\gamma}{\bar{\mu}}\|\nabla f_k(\boldsymbol{\theta}^t)\| + \frac{1}{\bar{\mu}}\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\| + \frac{\lambda}{\bar{\mu}}\|\mathbf{w}^t\|. \tag{23}$$

Substituting for $\|\nabla f_k(\boldsymbol{\theta}^t)\|$ from Lemma 2, we obtain

$$\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\| \leq \frac{1+\gamma}{\bar{\mu}}\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\| + \frac{\gamma L + \lambda}{\bar{\mu}}\|\mathbf{w}^t\| \tag{24}$$

Now we bound $\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t\|$ from (20) as follows.

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t\| &= \|\mathbb{E}_k[\boldsymbol{\theta}_k^{t+1}] - \mathbb{E}_k[\tilde{\boldsymbol{\theta}}^t] + \mathbf{w}^t\| \\ &\leq \mathbb{E}_k\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\| + \|\mathbf{w}^t\|,\end{aligned} \tag{25}$$

where the last inequality is due to triangular inequality. Substituting the upper bound on $\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\|$ from (24), we obtain the following:

$$\begin{aligned}\mathbb{E}_k\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\| &\leq \frac{1+\gamma}{\bar{\mu}}\mathbb{E}_k[\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\|] + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}^t\| \\ &\leq \left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}^t\|.\end{aligned} \tag{26}$$

The last inequality is due to the bounded local dissimilarity assumption, i.e., $\mathbb{E}_k[\|\nabla f_k(\tilde{\boldsymbol{\theta}})\|] \leq \sqrt{\mathbb{E}_k\|\nabla f_k(\tilde{\boldsymbol{\theta}})\|^2} \leq \|\nabla F(\tilde{\boldsymbol{\theta}})\|B$. After substituting (26) in (25),we have

$$\begin{aligned}\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t\| &= \left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \\ &+ \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}^t\| + \|\mathbf{w}^t\| \\ &= \left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \\ &+ \left(\frac{\bar{\mu} + \gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}^t\|.\end{aligned} \tag{27}$$

We simplify (20) as follows:

$$\begin{aligned}\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t &= \frac{-1}{\lambda}\mathbb{E}_k\left[\nabla f_k(\boldsymbol{\theta}_k^{t+1})\right] + \frac{1}{\lambda}\mathbb{E}_k\left[\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\right] + \mathbf{w}^t \\ &= \frac{-1}{\lambda}\Big\{\mathbb{E}_k\left[\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\right] \\ &+ \mathbb{E}_k\left[\nabla f_k(\boldsymbol{\theta}_k^{t+1}) - \nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t) - \nabla f_k(\tilde{\boldsymbol{\theta}}^t)\right]\Big\} + \mathbf{w}^t.\end{aligned} \tag{28}$$

We define,
$\mathbf{m}^{t+1} \triangleq \mathbb{E}_k\left[\nabla f_k(\boldsymbol{\theta}_k^{t+1}) - \nabla f_k(\tilde{\boldsymbol{\theta}}^t) - \nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\right]$, which is the second term on the right hand side of the expression above. Since $\mathbb{E}_k\left[\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\right] = \nabla F(\tilde{\boldsymbol{\theta}}^t)$, we have

$$\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t = \mathbb{E}_k[\boldsymbol{\theta}_k^{t+1}] - \boldsymbol{\theta}^t = \frac{-1}{\lambda}\left(\nabla F(\tilde{\boldsymbol{\theta}}^t) + \mathbf{m}^{t+1}\right) + \mathbf{w}^t \tag{29}$$

Now we derive upper bounds for the two terms on the right hand side above. To obtain an upperbound on the norm of $\mathbf{m}^{t+1}$, we use the L-Lipschitz smoothness assumption, triangle inequality, (26) and Lemma 3 to obtain the following:

$$\begin{aligned}\|\mathbf{m}^{t+1}\| &\leq \mathbb{E}_k\left[L\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\|\right] + \mathbb{E}_k\|\nabla h_k(\boldsymbol{\theta}_k^{t+1}; \tilde{\boldsymbol{\theta}}^t)\| \\ &\leq L\left[\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}^t\|\right] \\ &+ \gamma\mathbb{E}_k[\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\|] + (\gamma L + \lambda)\|\mathbf{w}^t\|.\end{aligned} \tag{30}$$

Further, using Assumption 1 to simplify $\mathbb{E}_k[\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\|]$ in the above expression, we have

$$\begin{aligned}\|\mathbf{m}^{t+1}\| &\leq \left[LB\left(\frac{1+\gamma}{\bar{\mu}}\right) + \gamma B\right]\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \\ &+ \left[L\left(\frac{\gamma L + \lambda}{\bar{\mu}}\right) + (\gamma L + \lambda)\right]\|\mathbf{w}^t\|.\end{aligned} \tag{31}$$

Using Cauchy-Schwartz inequality, we know that $\frac{-1}{\lambda}\langle\nabla F(\tilde{\boldsymbol{\theta}}^t), \mathbf{m}^{t+1}\rangle \leq \frac{1}{\lambda}\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\|\|\mathbf{m}^{t+1}\|$. Hence, it can be shown that

$$\begin{aligned}\frac{-1}{\lambda}\langle\nabla F(\tilde{\boldsymbol{\theta}}^t), \mathbf{m}^{t+1}\rangle &\leq \frac{1}{\lambda}\left[LB\left(\frac{1+\gamma}{\bar{\mu}}\right) + \gamma B\right]\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\|^2 \\ &+ \frac{1}{\lambda}\left[L\left(\frac{\gamma L + \lambda}{\bar{\mu}}\right) + (\gamma L + \lambda)\right]\|\mathbf{w}^t\|\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\|.\end{aligned} \tag{32}$$

Using L-Lipschitz smoothness of $F(\cdot)$ and Cauchy Schwartz inequality, we have

$$F(\bar{\boldsymbol{\theta}}^{t+1}) - F(\tilde{\boldsymbol{\theta}}^{t}) \leq \langle \nabla F(\tilde{\boldsymbol{\theta}}^{t}), \bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t} \rangle - \langle \nabla F(\tilde{\boldsymbol{\theta}}^{t}), \mathbf{w}^{t} \rangle + \frac{L}{2}\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\|^2 + \frac{L}{2}\|\mathbf{w}^{t}\|^2 - L\langle \bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}, \mathbf{w}^{t} \rangle \quad (33)$$

Substituting for $\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}$ from (29), we obtain

$$F(\bar{\boldsymbol{\theta}}^{t+1}) - F(\tilde{\boldsymbol{\theta}}^{t}) \leq \nabla F(\tilde{\boldsymbol{\theta}}^{t})^{T} \left[\frac{-1}{\lambda} \left(\nabla F(\tilde{\boldsymbol{\theta}}^{t}) + \mathbf{m}^{t+1} \right) \right] + \frac{L}{2}\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\|^2 + \frac{L}{2}\|\mathbf{w}^{t}\|^2 - L\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\|\|\mathbf{w}^{t}\| \quad (34)$$

Substituting for $\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\|$ as derived in (27), we obtain

$$\begin{aligned} F(\bar{\boldsymbol{\theta}}^{t+1}) - F(\tilde{\boldsymbol{\theta}}^{t}) &\leq \tfrac{-1}{\lambda}\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2 + \tfrac{L}{2}\left\{ \left(\tfrac{1+\gamma}{\bar{\mu}} \right) B\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\| \right. \\ &+ \left. \left(\tfrac{\bar{\mu}+\gamma L+\lambda}{\bar{\mu}} \right) \|\mathbf{w}^{t}\| \right\}^2 + \tfrac{1}{\lambda}\left[LB\left(\tfrac{1+\gamma}{\bar{\mu}} \right) + \gamma B \right] \|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2 \\ &+ \tfrac{1}{\lambda}\left[L\left(\tfrac{\gamma L+\lambda}{\bar{\mu}} \right) + (\gamma L + \lambda) \right] \|\mathbf{w}^{t}\|\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\| \\ &+ \tfrac{L}{2}\|\mathbf{w}^{t}\|^2 - L\left\{ \left(\tfrac{1+\gamma}{\bar{\mu}} \right) B\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\| + \left(\tfrac{\bar{\mu}+\gamma L+\lambda}{\bar{\mu}} \right) \|\mathbf{w}^{t}\| \right\} \|\mathbf{w}^{t}\| \end{aligned} \quad (35)$$

Taking expectation $\mathbb{E}_{\mathbf{w}^t}[.]$ on both sides of the above expression, rearranging the terms and subsequently using (14), we obtain the following:

$$F(\bar{\boldsymbol{\theta}}^{t+1}) \leq F(\tilde{\boldsymbol{\theta}}^{t}) - \alpha \times \|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2, \quad (36)$$

where

$$\begin{aligned} \alpha &= \left(\rho - \mathrm{c}_1 \frac{d\sigma^2}{K^2 P} - \mathrm{c}_2 \frac{\sqrt{d}\sigma}{K\sqrt{P}} \right), \\ \rho &= \left(\frac{1}{\lambda} - \frac{\gamma B}{\lambda} - \frac{(1+\gamma)LB}{\bar{\mu}\lambda} - \frac{LB^2(1+\gamma)^2}{2\bar{\mu}^2} \right), \\ \mathrm{c}_1 &= \frac{\kappa L B^2}{2}\left(\frac{\gamma L + \lambda}{\bar{\mu}} \right)^2, \text{and,} \\ \mathrm{c}_2 &= \sqrt{\kappa}\left(\frac{LB(\gamma L+\lambda)}{\bar{\mu}\lambda} + \frac{B(\gamma L+\lambda)}{\lambda} \right. \\ &\left. + \frac{LB^2(1+\gamma)(\bar{\mu}+\gamma L+\lambda)}{\bar{\mu}^2} - \frac{LB^2(1+\gamma)}{\bar{\mu}} \right). \end{aligned} \quad (37)$$

It is important to note that if $\sigma = 0$, we get the same result as FedProx.

Proof of Corollary 1: From Theorem 1, we have

$$\alpha \times \|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2 \leq F(\tilde{\boldsymbol{\theta}}^{t}) - F(\bar{\boldsymbol{\theta}}^{t+1})$$

Now, telescoping on both sides, i.e., considering $\sum_{t=0}^{T-1} \alpha \|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2 \leq \sum_{t=0}^{T-1} \left(F(\tilde{\boldsymbol{\theta}}^{t}) - F(\tilde{\boldsymbol{\theta}}^{t+1}) \right)$ leads to the following

$$\alpha \sum_{t=0}^{T-1} \|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2 \leq F(\tilde{\boldsymbol{\theta}}^{0}) - F(\tilde{\boldsymbol{\theta}}^{T}) \quad (38)$$

Essentially, this above implies that $\frac{\alpha}{T}\sum_{t=0}^{T-1}\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2 \leq \frac{\Delta}{T} \leq \alpha\epsilon$, where $\Delta = F(\tilde{\boldsymbol{\theta}}^{0}) - F(\tilde{\boldsymbol{\theta}}^{T})$. Hence, we have $T \geq \mathcal{O}\left(\frac{\Delta}{\left(\rho - \mathrm{c}_1 \frac{d\sigma^2}{K^2 P} - \mathrm{c}_2 \frac{\sqrt{d}\sigma}{K\sqrt{P}} \right)\epsilon} \right)$, i.e., as the number of communication rounds T is increased beyond this stipulated lower bound, it is possible to obtain diminishing value of $\sum_{t=0}^{T-1}\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\|^2$, which leads to diminishing difference between $F(\tilde{\boldsymbol{\theta}}^{t})$ and $F(\bar{\boldsymbol{\theta}}^{t+1})$.

Proof of Theorem 2: We now present the proof of convergence of the NoROTA-FL algorithm when only a subset of the devices participate in the FL process, i.e., $\hat{K}$ clients are chosen randomly for federation. We use the local Lipschitz continuity of $F(\cdot)$ which states that

$$F(\tilde{\boldsymbol{\theta}}^{t+1}) \leq F(\bar{\boldsymbol{\theta}}^{t+1}) + L_0\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|, \quad (39)$$

where L_0 is the local Lipschitz constant. Considering $\mathbb{E}_{\mathcal{S}^t}[.]$ on both sides of (39), we obtain

$$\mathbb{E}[F(\tilde{\boldsymbol{\theta}}^{t+1})] \leq F(\bar{\boldsymbol{\theta}}^{t+1}) + q^t, \quad (40)$$

where $q^t = \mathbb{E}[L_0\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|]$. Evidently, we need to obtain an upperbound on the expected norm of q^t so that the expected decrease and the rate of decrease in the loss function can be quantified. Towards this, we use the bound L_0 as given in [2], i.e,

$$L_0 \leq \|\nabla F(\boldsymbol{\theta}^{t})\| + L\left(\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| + \|\tilde{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| \right) \quad (41)$$

Using the above result in $q^t = \mathbb{E}\left[L_0\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| \right]$, the upperbound on the q^t is given as

$$q^t \leq \mathbb{E}[\underbrace{\left\{ \|\nabla F(\boldsymbol{\theta}^{t})\| + L\left(\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| + \|\tilde{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| \right) \right\}}_{\geq L_0} \|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|] \quad (42)$$

Using Lemma 2 in the context of $F(\boldsymbol{\theta}^t)$, we obtain the following

$$\begin{aligned} q^t &\leq \mathbb{E}\left[\left\{ \|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\| + L\|\mathbf{w}_p^t\| + L\left(\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| + \|\tilde{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| \right) \right\} \|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| \right] \\ &\leq \left(\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\| + L\mathbb{E}\|\mathbf{w}_p^t\| + L\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| \right) \mathbb{E}\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| \\ &\quad + L\mathbb{E}\left[\|\tilde{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\|\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| \right] \\ &\underset{(1)}{\leq} \left(\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\| + L\mathbb{E}\|\mathbf{w}_p^t\| + L\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| \right) \mathbb{E}\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| \\ &\quad + L\mathbb{E}\left[\left(\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| + \|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| \right) \|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| \right] \end{aligned} \quad (43)$$

where (1) holds by the triangular inequality (applied as $\|a-b\| \leq \|a-c\| + \|c-b\|$). Rearranging the terms above, we see that

$$\begin{aligned} q^t = &\left(\|\nabla F(\tilde{\boldsymbol{\theta}}^{t})\| + L\mathbb{E}\|\mathbf{w}_p^t\| + 2L\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^{t}\| \right) \\ &\mathbb{E}\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| + L\mathbb{E}\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2. \end{aligned} \quad (44)$$

We now consider upper-bounds for individual terms in the above expression (44). First, we consider $\mathbb{E}\|\tilde{\boldsymbol{\theta}}^{t+1} -$

$\bar{\boldsymbol{\theta}}^{t+1}\| \leq \sqrt{\mathbb{E}\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2}$ and subsequently upper bound $\mathbb{E}\left[\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2\right]$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{\hat{K}}\sum_{k=1}^{\hat{K}}\boldsymbol{\theta}_k^{t+1} + \mathbf{w}_p^t - \bar{\boldsymbol{\theta}}^{t+1}\right\|^2\right] \\
&\underset{(1)}{\leq} \frac{1}{(\hat{K})^2}\sum_{k=1}^{\hat{K}}\mathbb{E}[\|\boldsymbol{\theta}_k^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2] + \|\mathbf{w}_p^t\|^2 \\
&\quad + 2\mathbb{E}\langle\boldsymbol{\theta}_k^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}, \mathbf{w}_p^t\rangle \\
&\underset{(2)}{\leq} \frac{1}{\hat{K}}\mathbb{E}_k[\|\boldsymbol{\theta}_k^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2] + \|\mathbf{w}_p^t\|^2 \\
&\underset{(3)}{\leq} \frac{1}{\hat{K}}\mathbb{E}_k[\|(\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t) - (\bar{\boldsymbol{\theta}}^{t+1} - \tilde{\boldsymbol{\theta}}^t)\|^2] + \|\mathbf{w}_p^t\|^2 \\
&\underset{(4)}{\leq} \frac{2}{\hat{K}}\mathbb{E}_k[\|(\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t)\|^2] + \|\mathbf{w}_p^t\|^2 \qquad (45)
\end{aligned}$$

where (1) follows from Jensen's inequality. (2) is derived using Lemma 4 in [3] and $\mathbb{E}_{S^t}\langle\boldsymbol{\theta}_k^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}, \mathbf{w}_p^t\rangle = 0$. We add and subtract $\tilde{\boldsymbol{\theta}}^t$ in (3) and finally we arrive at (4) since $\mathbb{E}_k\left[\boldsymbol{\theta}_k^{t+1}\right] = \bar{\boldsymbol{\theta}}^{t+1}$.

$$\begin{aligned}
\mathbb{E}_{S^t}\left[\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2\right] &\leq \frac{2}{\hat{K}}\mathbb{E}_k[\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\|^2] + \|\mathbf{w}_p^t\|^2 \\
&\underset{(5)}{\leq} \frac{2}{\hat{K}}\mathbb{E}_k\left[\left(\frac{1+\gamma}{\bar{\mu}}\right)\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}^t\|\right]^2 + \|\mathbf{w}_p^t\|^2 \\
&\underset{(6)}{\leq} \frac{2}{\hat{K}}\left[\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}^t\|\right]^2 + \|\mathbf{w}_p^t\|^2, \qquad (46)
\end{aligned}$$

where (24) and Assumption 1 yields inequalities (5) and (6) respectively. We complete the upperbound on q^t by substituting and thereafter adjusting the bounds from (26) and (46) in (44) and we get

$$\begin{aligned}
q^t \leq\ & \left[\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + L\|\mathbf{w}_p^t\| + 2L\left\{\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \right.\right. \\
& \left.\left. + \left(\frac{\bar{\mu} + \gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_p^t\|\right\}\right] \times \\
& \left[\frac{\sqrt{2}}{\sqrt{\hat{K}}}\left\{\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_p^t\|\right\} + \|\mathbf{w}_p^t\|\right] \\
& + L\left[\frac{2}{\hat{K}}\left\{\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_p^t\|\right\}^2 + \|\mathbf{w}_p^t\|^2\right] \qquad (47)
\end{aligned}$$

Now taking expectation $\mathbb{E}_{\mathbf{w}_p^t}[.]$ and using (16), we get

$$\begin{aligned}
q^t \leq & \left[\frac{B}{\sqrt{\hat{K}}\bar{\mu}}\left(1 + \gamma + \frac{(\gamma L + \lambda)\sqrt{\kappa}\sqrt{d}\sigma}{\hat{K}\sqrt{P}}\right)\right. \\
& \left(\sqrt{2} + \frac{3\sqrt{\kappa}\sqrt{2d}LB\sigma}{\hat{K}\sqrt{P}} + \frac{2LB\sqrt{\kappa}\sqrt{d}\sigma}{\sqrt{\hat{K}P}}\right) \\
& + \frac{LB^2}{\hat{K}\bar{\mu}^2}\left(1 + \gamma + \frac{(\gamma L + \lambda)\sqrt{\kappa}\sqrt{d}\sigma}{\hat{K}\sqrt{P}}\right)^2(2\sqrt{2\hat{K}} + 2) \\
& \left. + \frac{\sqrt{\kappa}\sqrt{d}\sigma B}{\hat{K}\sqrt{P}} + \frac{4LB^2\kappa d\sigma^2}{\hat{K}^2 P}\right] \times \|\nabla F(\tilde{\boldsymbol{\theta}}^t)\|^2. \qquad (48)
\end{aligned}$$

It is important to note here also that if $\sigma = 0$, we get the same result as FedProx.

Finally, we prove the theorem by substituting the bounds from (36) and (48) into (40).

A. Fading

Following the proof steps similar to Theorem 2, we have

$$\mathbb{E}_{\mathcal{K}^t}[F(\tilde{\boldsymbol{\theta}}^{t+1})] \leq F(\bar{\boldsymbol{\theta}}^{t+1}) + q^t, \qquad (49)$$

The upperbound on the q^t is given as

$$\begin{aligned}
q^t \leq\ & \mathbb{E}_{\mathcal{K}^t}\left[\underbrace{\left\{\|\nabla F(\boldsymbol{\theta}^t)\| + L\left(\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t\| + \|\tilde{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t\|\right)\right\}}_{L_0}\right. \\
& \left.\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|\right] \\
& \leq \left(\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + L\mathbb{E}_{\mathcal{K}^t}\|\mathbf{w}_f^t\| + 2L\|\bar{\boldsymbol{\theta}}^{t+1} - \boldsymbol{\theta}^t\|\right) \\
& \mathbb{E}_{\mathcal{K}^t}\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\| + L\mathbb{E}_{\mathcal{K}^t}\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2. \qquad (50)
\end{aligned}$$

Further, $\mathbb{E}_{\mathcal{K}^t}\left[\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2\right]$ can be upperbounded as

$$\begin{aligned}
\mathbb{E}_{\mathcal{K}^t}\left[\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2\right] &= \mathbb{E}_{\mathcal{K}^t}\left[\left\|\frac{1}{|\mathcal{K}^t|}\sum_{k\in\mathcal{K}^t}\boldsymbol{\theta}_k^{t+1} + \mathbf{w}_f^t - \bar{\boldsymbol{\theta}}^{t+1}\right\|^2\right] \\
&\underset{(1)}{\leq} \frac{1}{(\hat{K})^2}\sum_{k=1}^{\hat{K}}\mathbb{E}_{\mathcal{K}^t}[\|\boldsymbol{\theta}_k^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2] + \|\mathbf{w}_f^t\|^2 \\
&\underset{(2)}{\leq} \frac{1}{\hat{K}}\mathbb{E}_k[\|\boldsymbol{\theta}_k^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2] + \|\mathbf{w}_f^t\|^2 \\
&\underset{(3)}{\leq} \frac{1}{\hat{K}}\mathbb{E}_k[\|(\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t) - (\bar{\boldsymbol{\theta}}^{t+1} - \tilde{\boldsymbol{\theta}}^t)\|^2] + \|\mathbf{w}_f^t\|^2 \\
&\underset{(4)}{\leq} \frac{2}{\hat{K}}\mathbb{E}_k[\|(\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t)\|^2] + \|\mathbf{w}_f^t\|^2 \qquad (51)
\end{aligned}$$

where (1) follows from Jensen's inequality. (2) is derived using Lemma 4 in [3]. We add and subtract $\tilde{\boldsymbol{\theta}}^t$ in (3) and finally we arrive at (4) because $\mathbb{E}\left[\boldsymbol{\theta}_k^{t+1}\right] = \bar{\boldsymbol{\theta}}^{t+1}$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{K}^t}\left[\|\tilde{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^{t+1}\|^2\right] &\leq \frac{2}{\hat{K}}\mathbb{E}_k[\|\boldsymbol{\theta}_k^{t+1} - \tilde{\boldsymbol{\theta}}^t\|^2] + \|\mathbf{w}_f^t\|^2 \\
&\underset{(5)}{\leq} \frac{2}{\hat{K}}\mathbb{E}_k\left[\left(\frac{1+\gamma}{\bar{\mu}}\right)\|\nabla f_k(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_f^t\|\right]^2 + \|\mathbf{w}_f^t\|^2 \\
&\underset{(6)}{\leq} \frac{2}{\hat{K}}\left[\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_f^t\|\right]^2 + \|\mathbf{w}_f^t\|^2, \qquad (52)
\end{aligned}$$

where (24) and Assumption 1 yields (5) and (6) respectively. We complete the upperbound on q^t by substituting and thereafter adjusting the bounds from (26) and (52) in (44) and we get

$$\begin{aligned}
q^t \leq\ & \left[\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + L\|\mathbf{w}_f^t\| + 2L\left\{\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| \right.\right. \\
& \left.\left. + \left(\frac{\bar{\mu} + \gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_f^t\|\right\}\right] \times \\
& \left[\frac{\sqrt{2}}{\sqrt{\hat{K}}}\left\{\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_f^t\|\right\} + \|\mathbf{w}_f^t\|\right] \\
& + L\left[\frac{2}{\hat{K}}\left\{\left(\frac{1+\gamma}{\bar{\mu}}\right)B\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\| + \left(\frac{\gamma L + \lambda}{\bar{\mu}}\right)\|\mathbf{w}_f^t\|\right\}^2 + \|\mathbf{w}_f^t\|^2\right]. \qquad (53)
\end{aligned}$$

Now taking expectation $\mathbb{E}_{\mathbf{w}_f^t}[.]$ and using (18), we have

$$q^t \leq \left[\frac{B}{\sqrt{\hat{K}h_{min}}\bar{\mu}}\left(1+\gamma+\frac{(\gamma L+\lambda)\sqrt{\kappa}\sqrt{d}\sigma}{\hat{K}h_{min}\sqrt{P}}\right)\right.$$
$$\left(\sqrt{2}+\frac{3\sqrt{\kappa}\sqrt{2d}LB\sigma}{\hat{K}h_{min}\sqrt{P}}+\frac{2LB\sqrt{\kappa}\sqrt{d}\sigma}{\sqrt{\hat{K}h_{min}P}}\right)$$
$$+\frac{LB^2}{\hat{K}h_{min}\bar{\mu}^2}\left(1+\gamma+\frac{(\gamma L+\lambda)\sqrt{\kappa}\sqrt{d}\sigma}{\hat{K}h_{min}\sqrt{P}}\right)^2$$
$$\left.(2\sqrt{2\hat{K}h_{min}}+2)+\frac{\sqrt{\kappa}\sqrt{d}\sigma B}{\hat{K}h_{min}\sqrt{P}}+\frac{4LB^2\kappa d\sigma^2}{\hat{K}^2h_{min}^2P}\right]\times\|\nabla F(\tilde{\boldsymbol{\theta}}^t)\|^2. \tag{54}$$

Finally, we prove the theorem by substituting the bounds from (36) and (54) into (49).

B. Computations to Compute Optimal λ:

In discussions after theorem 1 and theorem 2, we alluded to the constants a_1,a_2 and a_3 and and b_1,b_2 and b_3, respectively, for optimal λ computation. The expressions to compute these constants are as given below:

$$a_1=\frac{LB^2\kappa d\sigma^2}{2K^2P},$$
$$a_2=(LB^2\gamma+LB^2+B+LB+\gamma^2L^2B^2)\frac{\sqrt{\kappa d}\sigma}{K\sqrt{P}}$$
$$+\frac{L^2B^2\gamma\kappa d\sigma^2}{K^2P}+\gamma B-1,$$
$$a_3=(\gamma L^2B^2+\gamma LB+\gamma L^2B)\frac{\sqrt{\kappa d}\sigma}{K\sqrt{P}}+\frac{\gamma^2L^3B^2\kappa d\sigma^2}{2K^2P}$$
$$+\frac{LB^2(1+\gamma)^2}{2}+(1+\gamma)B,$$
$$b_1=\frac{LB^2\kappa d\sigma^2}{2K^2P}+\frac{5\sqrt{2}LB^2\kappa d\sigma^2+\sqrt{2}B\sqrt{\kappa d}\sigma}{\hat{K}^2\sqrt{\hat{K}}P}+\frac{8LB^2\kappa d\sigma^2}{\hat{K}^2P}+B,$$
$$b_2=\frac{(LB^2\gamma+LB^2+B+LB)\sqrt{\kappa d}\sigma}{K\sqrt{P}}+\frac{L^2B^2\gamma\kappa d\sigma^2}{K^2P}+\frac{14\kappa d\sigma^2}{\hat{K}^2P}$$
$$+\left(\frac{16LB^2(1+\gamma)+\sqrt{2}\gamma LB}{\sqrt{\hat{K}}}+2(1+\gamma)LB^2\right)\frac{\sqrt{\kappa d}\sigma}{\hat{K}\sqrt{P}}$$
$$+\frac{B(1+\gamma)\sqrt{2}}{\sqrt{\hat{K}}}+\gamma B-1,$$
$$b_3=(1+\gamma)LB+\frac{LB^2(1+\gamma)^2}{2}+\frac{LB^2(1+\gamma)^2(2\sqrt{2\hat{K}}+2)}{\hat{K}}$$
$$+\frac{L^3B^2\gamma^2\kappa d\sigma^2}{K^2P}+\frac{(L^2B\gamma+B\gamma L+L^2B^2\gamma+L^2B^2\gamma^2)\sqrt{\kappa d}\sigma}{K\sqrt{P}}$$
$$+\frac{5L^3B^2\gamma^2\kappa d\sigma^2}{\hat{K}^2P}+\frac{16L^2B^2\gamma^2\sqrt{\kappa d}\sigma}{\hat{K}^2\sqrt{P}}. \tag{55}$$

REFERENCES

[1] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3796–3811, 2021.

[2] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[3] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.