APPENDIX A
PROOF OF THEOREM 1

We bring in the following assumptions from [1], [2] for analytical tractability.

Assumption 1. *The loss function is L-smooth as* $\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\| \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|$ *for arbitrary given* $\mathbf{w}_1$ *and* $\mathbf{w}_2$.

Assumption 2. *The expected squared norm of stochastic gradients for each vehicle* k *is upper-bounded by*

$$\mathbb{E}\left\|\nabla f(\mathbf{w}_k^{r,m})\right\|^2 \leq \ G^2, \forall k, \forall r, \forall m.$$

Assumption 3. *The variance of mini-batch gradients is upper-bounded by*

$$\left\|g(\mathbf{w}) - \nabla f(\mathbf{w})\right\|^2 \leq \sigma^2.$$

Assumption 4. *The divergence between local and global loss functions is bounded by*

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla f(\mathbf{w}) - \nabla F(\mathbf{w})\right\|^2 \leq \epsilon_g^2, \forall \mathbf{w}.$$

Based on the L-smoothness assumption, we have the following equation:

$$F(\mathbf{w}^{r+1}) - F(\mathbf{w}^r) \leq \langle \nabla F(\mathbf{w}^r), \mathbf{w}^{r+1} - \mathbf{w}^r \rangle + L/2\|\mathbf{w}^{r+1} - \mathbf{w}^r\|^2. \tag{A.1}$$

We give the expression of the local update of each model $\mathbf{w}_k^{m,r}$ with each modality m for vehicle k as

$$\mathbf{w}_{k,e+1}^{m,r} = \mathbf{w}_{k,e}^{m,r} - \eta \mathbf{g}_{k,e}^{m,r}, \tag{A.2}$$

here, η is the learning rate, and $g_{k,e}^{m,r}$ is the gradient descent for modality m in round r epoch $e \in \{0, \ldots, E-1\}$. We then substitute the local update with the gradient and model updates of all modalities from all vehicles and take the expectation on both sides as

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}^{r+1}) - F(\mathbf{w}^r)] \leq \mathbb{E}\langle \mathbf{w}^{r+1} - \mathbf{w}^r, \nabla F(\mathbf{w}^r)\rangle + \frac{L}{2}\mathbb{E}\|\mathbf{w}^{r+1} - \mathbf{w}^r\|^2 \\
&\leq \sum_{m=1}^{M}\mathbb{E}\langle \nabla F(\mathbf{w}^{m,r}), -\frac{1}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\eta \mathbf{g}_{k,e}^{m,r}\rangle + \sum_{m=1}^{M}\frac{L}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\eta \mathbf{g}_{k,e}^{m,r}\|^2
\end{aligned} \tag{A.3}$$

For the left side of the sum, we obtain the following expressions according to Assumptions 2 and 3:

$$\begin{aligned}
&\sum_{m=1}^{M}\mathbb{E}\langle \nabla F(\mathbf{w}^{m,r}), -\frac{1}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\eta \mathbf{g}_{k,e}^{m,r}\rangle \leq -\eta\sum_{m=1}^{M}\frac{1}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\mathbb{E}\langle \nabla F(\mathbf{w}^{m,r}), \nabla f_k(\mathbf{w}_{k,e}^{m,r})\rangle \\
&\leq -\eta\sum_{m=1}^{M}\frac{1}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\Big(\mathbb{E}\|\nabla F(\mathbf{w}^{m,r})\|^2 + \mathbb{E}\|\nabla f_k(\mathbf{w}_{k,e}^{m,r})\|^2 \\
&- \mathbb{E}\|\nabla F(\mathbf{w}^{m,r}) - \nabla F(\mathbf{w}_{k,e}^{m,r}) + \nabla F(\mathbf{w}_{k,e}^{m,r}) - \nabla f_k(\mathbf{w}_{k,e}^{m,r})\|^2\Big) \\
&\leq -\eta\sum_{m=1}^{M}\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}\|\nabla F(\mathbf{w}^{m,r})\|^2 E + \sum_{e=0}^{E-1}\mathbb{E}\|\nabla f_k(\mathbf{w}_{k,e}^{m,r})\|^2\right. \\
&\left. - 2\sum_{e=0}^{E-1}\mathbb{E}\big[\|\nabla F(\mathbf{w}^{m,r}) - \nabla F(\mathbf{w}_{k,e}^{m,r})\|^2 + \|\nabla F(\mathbf{w}_{k,e}^{m,r}) - \nabla f_k(\mathbf{w}_{k,e}^{m,r})\|^2\big]\right) \\
&\leq -\eta\sum_{m=1}^{M}\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}\|\nabla F(\mathbf{w}^{m,r})\|^2 E - 2\sum_{e=0}^{E-1}L^2\|\mathbf{w}^{m,r} - \mathbf{w}_{k,e}^{m,r}\|^2\right) + 2\eta E M\epsilon_g^2.
\end{aligned} \tag{A.4}$$

Considering the norm $\|\mathbf{w}^{m,r} - \mathbf{w}_{k,e}^{m,r}\|^2$, we get the following expressions based on the Lemma 3 in [2]:

$$E[\sum_{k=1}^{K}\|\mathbf{w}^{m,r} - \mathbf{w}_{k,e}^{m,r}\|^2] \leq 4\eta^2(E-1)^2G^2. \tag{A.5}$$

For the right side of the sum, we obtain

$$\frac{L}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\sum_{e=0}^{E-1}\eta\mathbf{g}_{k,e}^{m,r}\|^2 \leq \frac{L\eta^2E^2}{2}\delta^2. \tag{A.6}$$

Combining the above results, we obtain the following expression:

$$\mathbb{E}[F(\mathbf{w}^{r+1}) - F(\mathbf{w}^r)] \leq -\eta\frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(\mathbb{E}\|\nabla F(\mathbf{w}^{m,r})\|^2E - 8E(E-1)^2\eta^2G^2\right) + 2EM\epsilon_g^2\eta + \frac{L\eta^2E^2\delta^2M}{2} \tag{A.7}$$

Then we rearrange the above expression and add all the terms from $r = \{0, 1, \ldots, R\}$ to get the below expression

$$\frac{1}{R}\sum_{r=0}^{R-1}E\left[\|\nabla F(\mathbf{w}^{m,r})\|^2\right] \leq \frac{2\chi^m\left(F(\mathbf{w}^0) - F(\mathbf{w}^R)\right)}{\eta ER} + 2\epsilon_g^2 + 8(E-1)^2\eta^2G^2 + \frac{L\eta E\delta^2}{2}, \tag{A.8}$$

where χ^m is the contribution ratio of modality m to the training optimization during model training. Hence, Theorem 1 is proved.

References

[1] C. Feng, H. H. Yang, D. Hu, Z. Zhao, T. Q. S. Quek, and G. Min, “Mobility-aware cluster federated learning in hierarchical wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8441–8458, Oct. 2022.

[2] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of FedAvg on non-IID data,” 2020. [Online]. Available: https://arxiv.org/abs/1907.02189