Appendix A: Comparison with a Follow-Up Work

Before submitting our manuscript to INFORMS Journal on Computing, we first released a preprint with the same technical contributions in February 2020 (Yan et al. 2020). Fifteen months after our preprint, Gu et al. (2021b) published their work titled "Fast federated learning in the presence of arbitrary device unavailability". They cited and discussed our preprint in page 3 paragraph 4: "While preparing the manuscript, we were unaware of an independent work that investigated the same setup and proposed a similar algorithm called FedLaAvg". Their availability formulation and algorithm design are basically the same as ours: they required the number of rounds that a device does not participate to be bounded by a constant in their Assumption 8, which is similar to the conclusion of our Lemma 1; their proposed MIFA algorithm augments the received updates of the active devices with the stored updates of the inactive devices to perform averaging, which shares the same idea of our FedLaAvg; to avoid exhausting the server's memory, they used the trick of storing the latest gradient locally and sending the gradient difference, which is same with our FedLaAvg. We first list and discuss the minor differences which do not affect the theoretical analysis as follows. (1) Instead of first formulating on client availability and then considering which of the available clients participate, Gu et al. (2021a) directly formulated on client participation. (2) Gu et al. (2021a) required all clients to participate in the first round in their Remark 5.2, while we do not require this. (3) In addition to the general non-convex case, Gu et al. (2021a) specifically analyzed the special case of strongly convex objective functions. (4) In the numerical experiments, Gu et al. (2021a) models the availability of the clients as independent Bernoulli random trials, meaning that each client participates with equal probability in each round, which is impractical in federated learning as we have discussed. Regarding the theoretical analysis, Gu et al. (2021a) additionally required the loss function to be Hessian Lipschitz in their Assumption 5, and relied on this to derive the linear speedup in terms of N and K (cf. the second equation in Appendix C.3.1 of their supplement). Therefore, Gu et al. (2021a) is a strict follow-up of our work.

Appendix B: Divergence of a concurrent work

In this section, we show the divergence of a concurrent work (Ruan et al. 2021), which considered a different availability setting: the number of local iterations performed by client i in round r can take an arbitrary value s_r^i from $\{0,1,\cdots,C\}$, following some time-varying distribution. For the clients submitting incomplete work, i.e., $1 \le s_r^i < C$, they proposed to scale the model update by $w_i^r = \frac{C}{s_i^r} w_i$, to force equal contribution to the global model among clients. Under this framework, they proved the convergence, which, however, is tailed with a bias term $\frac{M_R}{R}$; R is the total number of communication rounds, and M_R represents the accumulated training data bias throughout the training process. Specifically, $M_R = \sum_{r=1}^{R} \mathbb{E}[z_r]$, where $z_r = 0$ if all clients contribute equally to the global model update in round r, i.e., $\frac{\mathbb{E}[w_i^r s_i^r]}{w_i}$ take the same value for all clients i, and $z_r = 1$ else. When there exists a single client i whose $s_r^i = 0$ with probability $p_i > 0$, i.e., client i is unavailable in round r, the proposed scaling technique cannot work and $z_r = p_i$. Further, if there exists multiple clients whose $s_r^i = 0$ with probability $p_i > 0$, z_r is even larger. The work (Ruan et al. 2021) claims that their method converges if M_R increases sublinearly with R.

The convergence of their proposed method requires that in most rounds of training, there are no unavailable clients. However, under intermittent client availability considered in this work, there exist unavailable clients in almost every round. E.g., in our Example 1, two clients are available alternately, and there is always one client unavailable throughout the training process, i.e., $s_i^r = 1$ with probability 1, indicating that $M_R = R$ and the method diverges. Further, the work (Ruan et al. 2021) proposed to kick out frequently unavailable clients if the evaluated training data bias introduced by keeping the clients, i.e., $\frac{M_R}{R}$, is larger than that introduced by kicking the clients out. This, however, still cannot solve the bias problem, since the bias always exists no matter whether frequently unavailable clients are kicked out or not.

Appendix C: Detailed Proof of Theorem 1

We first show that if $\gamma < 1/2$, $\mathbf{x}^{k(t_1+t_2)}$ would converge to

$$X = \frac{(1-2\gamma)^{t_2}(\mathbf{e}_1 - \mathbf{e}_2) + \mathbf{e}_2 - \mathbf{e}_1(1-2\gamma)^{t_1+t_2}}{1-(1-2\gamma)^{t_1+t_2}}. \tag{16}$$

Note that for iterations where client 1 is available, we have

$$\forall t \in \{k(t_1+t_2)+i \mid k \in \mathbb{N}, i \in \{1,\cdots,t_1\}\},\ \mathbf{x}^{t+1} = \mathbf{x}^t - 2\gamma(\mathbf{x}^t - \mathbf{e}_1),$$

where γ is the learning rate. Rearrange the equation, we have

$$\mathbf{x}^{t+1} - \mathbf{e}_1 = (1-2\gamma)(\mathbf{x}^t - \mathbf{e}_1),$$

which implies that $(\mathbf{x}^t - \mathbf{e}_1)$ is a geometric progression. Hence, we have

$$\mathbf{x}^{k(t_1+t_2)+t_1} = (1-2\gamma)^{t_1}\left(\mathbf{x}^{k(t_1+t_2)} - \mathbf{e}_1\right) + \mathbf{e}_1. \tag{17}$$

Applying the same analysis on iterations where client 2 is available, we have

$$\mathbf{x}^{(k+1)(t_1+t_2)} = (1-2\gamma)^{t_2}\left(\mathbf{x}^{k(t_1+t_2)+t_1} - \mathbf{e}_2\right) + \mathbf{e}_2. \tag{18}$$

Replacing the term $\mathbf{x}^{k(t_1+t_2)+t_1}$ by (17), we have

$$\begin{aligned}\mathbf{x}^{(k+1)(t_1+t_2)} &= (1-2\gamma)^{t_2}\left((1-2\gamma)^{t_1}\left(\mathbf{x}^{k(t_1+t_2)} - \mathbf{e}_1\right) + \mathbf{e}_1 - \mathbf{e}_2\right) + \mathbf{e}_2. \\ &= (1-2\gamma)^{t_1+t_2}\left(\mathbf{x}^{k(t_1+t_2)} - \mathbf{e}_1\right) + (1-2\gamma)^{t_2}(\mathbf{e}_1 - \mathbf{e}_2) + \mathbf{e}_2.\end{aligned} \tag{19}$$

Based on this recursion formula, we have

$$\mathbf{x}^{k(t_1+t_2)} = (1-2\gamma)^{(t_1+t_2)k}\mathbf{x}^0 + \left(1-(1-2\gamma)^{(t_1+t_2)k}\right)X.$$

Since $\gamma < \frac{1}{2}$, we have $1-2\gamma < 1$. Hence, $k \to +\infty$ implies $(1-2\gamma)^{(t_1+t_2)k} \to 0$, and we have $\lim_{k\to+\infty} \mathbf{x}^{k(t_1+t_2)} = X$. Based on L'Hopital's rule, by replacing the numerator and denominator in (16) by their respective derivative at $\gamma = 0$, we have

$$\lim_{\gamma\to 0^+} X = \lim_{\gamma\to 0^+} \frac{t_2(1-2\gamma)^{t_2-1}(e_1-e_2) - (t_1+t_2)(1-2\gamma)^{t_1+t_2-1}e_1}{-(t_1+t_2)(1-2\gamma)^{t_1+t_2-1}} = \frac{t_1\mathbf{e}_1 + t_2\mathbf{e}_2}{t_1+t_2}.$$

The global minimization objective is

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{2}(\mathbf{x}-\mathbf{e}_i)^2 + \frac{1}{2}\sum_{i=1}^{2}\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[(\xi_i - \mathbf{e}_i)^2\right],$$

and the minimum is reached when $\mathbf{x} = \mathbf{x}^* = (\mathbf{e}_1+\mathbf{e}_2)/2$. Note that $(\mathbf{e}_1+\mathbf{e}_2)/2 = (t_1\mathbf{e}_1 + t_2\mathbf{e}_2)/(t_1+t_2)$ only when $\mathbf{e}_1 = \mathbf{e}_2$ (data distributions are IID) or $t_1 = t_2$ (availability patterns are IID). Hence, FedAvg will produce arbitrarily poor-quality results without these impractical assumptions. □

Appendix D: Detailed Proof of Lemma 1

We prove the lemma by contradiction. Suppose the latest participating iteration (denoted as t) for a certain client i is more than I iterations earlier than the current iterations. Then, client i does not participate in iterations $t+1, t+2, \cdots, t+I+1$. Under Assumption 4, client i has been available for at least $\lceil N/K \rceil$ times in these iterations. We note the $\lceil N/K \rceil$ iterations as $\tau_1, \tau_2, \cdots, \tau_{\lceil N/K \rceil}$. Since i is not selected in any of these iterations, we have $T_i^{\tau_{\lceil N/K \rceil}} = T_i^t$. In the $\lceil N/K \rceil$ iterations where client i is available, $\lceil N/K \rceil K$ clients have been selected. All these clients (noted as j) are with $T_j^\tau \leq T_i^t$ for all iterations τ before it participates in the training process and $T_j^\tau > T_i^t$ for all iterations τ after participation. Hence, the $\lceil N/K \rceil K$ clients are distinct. Including client i, the system has at least $\lceil N/K \rceil K + 1$ clients. However, the system has only $N \leq \lceil N/K \rceil K < \lceil N/K \rceil K + 1$ clients. This forms a contradiction.

□

Appendix E: Detailed Proofs of the Lemmas for Theorem 2

Proof of Lemma 2 Assumptions 2 and 3 take the expectation over the randomness of one training iteration. But we care about the expectation taken over the randomness of the whole training process. This lemma builds the gap.

For the gradient, we have

$$\mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right] \overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\|\mathbf{g}_i^t\|^2 \mid \xi^{[t-1]}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\nabla F\left(\mathbf{x}^{t-1}, \xi_i^t\right)\right\|^2 \mid \mathbf{x}^{t-1}\right]\right] \overset{(b)}{\leq} \mathbb{E}\left[G^2\right] = G^2, \tag{20}$$

where (a) follows from $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$; (b) follows from Assumption 3.

For the variance, we have

$$\begin{aligned} &\mathbb{E}\left[\left\|\mathbf{g}_i^t - \nabla f_i(\mathbf{x}^{t-1})\right\|^2\right] \overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{g}_i^t - \nabla f_i(\mathbf{x}^{t-1})\right\|^2 \mid \xi^{[t-1]}\right]\right] \\ =&\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla F\left(\mathbf{x}^{t-1}; \xi_i^t\right) - \nabla f_i\left(\mathbf{x}^{t-1}\right)\right\|^2 \mid \mathbf{x}^{t-1}\right]\right] \overset{(b)}{\leq} \mathbb{E}\left[\sigma^2\right] = \sigma^2, \end{aligned} \tag{21}$$

where (a) follows from $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$; (b) follows from Assumption 2.

□

Proof of Lemma 3. This lemma follows because training data are independent across clients. Specifically, note that

$$\begin{aligned} &\mathbb{E}\left[\left\|\sum_{i=1}^N \left(\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t - 1})\right)\right\|^2\right] = \sum_{p=1}^N \sum_{q=1}^N \mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t - 1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t - 1})\right\rangle\right] \\ \overset{(a)}{=}& \sum_{p=1}^N \sum_{q=1}^N \mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t - 1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t - 1})\right\rangle | \xi^{\left[\min\{T_p^t, T_q^t\}\right]}\right]\right] \\ \overset{(b)}{=}& \sum_{i=1}^N \mathbb{E}\left[\left\|\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t - 1})\right\|^2\right], \end{aligned} \tag{22}$$

where (a) follows from $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$. Then we illustrate (b) case by case. Note that

$$\mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t - 1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t - 1})\right\rangle \mid \xi^{\left[\min\{T_p^t, T_q^t\}\right]}\right]\right]$$

is equal to $\mathbb{E}\left[\left\|\mathbf{g}_i^{T_i^t} - \nabla f_i(\mathbf{x}^{T_i^t-1})\right\|^2\right]$ when $p=q=i$. When $p\neq q$, without loss of generality, suppose $T_p^t \leq T_q^t$. Then it is equal to

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1})\right\rangle \mid \xi^{\left[T_p^t\right]}\right]\right]\\ =&\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbb{E}\left[\mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1}) \mid \xi^{\left[T_p^t\right]}\right]\right\rangle\right]\end{aligned} \tag{23}$$

because $\mathbf{g}_p^{T_p^t}$ and $f_q(\mathbf{x}^{T_p^t-1})$ are determined by $\xi^{\left[T_p^t\right]}$. When $T_p^t < T_q^t$, we have $\mathbb{E}[\mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1}) \mid \xi^{[T_p^t]}] = 0$. When $T_p^t = T_q^t$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbb{E}\left[\mathbf{g}_q^{T_q^t} - \nabla f_q(\mathbf{x}^{T_q^t-1}) \mid \xi^{\left[T_p^t\right]}\right]\right\rangle\right]\\ \overset{(a)}{=}&\mathbb{E}\left[\mathbb{E}\left[\left\langle \mathbf{g}_p^{T_p^t} - \nabla f_p(\mathbf{x}^{T_p^t-1}), \mathbf{g}_q^{T_p^t} - \nabla f_q(\mathbf{x}^{T_p^t-1})\right\rangle \mid \xi^{\left[T_p^t-1\right]}\right]\right] \overset{(b)}{=} 0,\end{aligned}$$

where (a) follows from $\mathbb{E}[\mathbb{E}[\mathbf{X}|\mathbf{Y}]] = \mathbb{E}[\mathbf{X}]$; (b) follows because $\xi_p^{T_p^t}$ and $\xi_q^{T_p^t}$ are independent, and thus the covariance of $\mathbf{g}_p^{T_p^t}$ and $\mathbf{g}_q^{T_p^t}$ is 0.

□

Proof of Lemma 4. This lemma follows the intuition that the difference of $\mathbf{x}$ in two iterations is bounded by the number of iterations between them.

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}^{t-1}) - \nabla f_i(\mathbf{x}^{t_0-1})\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{\tau=t_0}^{t-1}\left(\nabla f_i(\mathbf{x}^{\tau}) - \nabla f_i(\mathbf{x}^{\tau-1})\right)\right\|^2\right]\\ \overset{(a)}{\leq}&(t-t_0)\sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}^{\tau}) - \nabla f_i(\mathbf{x}^{\tau-1})\right\|^2\right] \overset{(b)}{\leq} (t-t_0)\,L^2\sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\left\|\mathbf{x}^{\tau} - \mathbf{x}^{\tau-1}\right\|^2\right]\\ \overset{(c)}{=}&(t-t_0)\,L^2\gamma^2\sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{j=1}^{N}\mathbf{g}_j^{T_j^{\tau}}\right\|^2\right] \overset{(d)}{\leq} (t-t_0)\,L^2\gamma^2\frac{1}{N}\sum_{\tau=t_0}^{t-1}\sum_{j=1}^{N}\mathbb{E}\left[\left\|\mathbf{g}_j^{T_j^{\tau}}\right\|^2\right]\\ \overset{(e)}{\leq}&(t-t_0)^2\,L^2\gamma^2G^2,\end{aligned} \tag{24}$$

where (a) and (d) follows from the convexity of $\|\cdot\|^2$; (b) follows from Assumption 1; (c) follows from (2); (e) follows from Lemma 2.

□

Proof of Corollary 1.

$$\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{t_0-1})\right\|^2\right] \overset{(a)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla f_i\left(\mathbf{x}^{t-1}\right) - \nabla f_i(\mathbf{x}^{t_0-1})\right\|^2\right] \overset{(b)}{\leq} (t-t_0)^2L^2\gamma^2G^2,$$

where (a) follows from the convexity of $\|\cdot\|^2$; (b) follows from Lemma 4.

□

Appendix F: Detailed Proof of Corollary 2

We first summarize the $O(\cdot)$ form of Theorem 2:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right] = O\left(\frac{\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}} + \frac{\gamma^2I^2L^2G^2}{1-2\gamma L} + \frac{B}{\gamma T}\right). \tag{25}$$

Substituting γ with $(\beta^{1/2}N^{1/4})/(2LE^{1/2}T^{1/2})$, we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right]\overset{(a)}{=}O\left(\frac{\gamma IL\left(G^2+\sigma^2\right)}{\sqrt{N}}+I^2\gamma^2L^2G^2+\frac{B}{\gamma T}\right)\\ =&O\left(\frac{I\beta^{\frac{1}{2}}\left(G^2+\sigma^2\right)}{N^{\frac{1}{4}}E^{\frac{1}{2}}T^{\frac{1}{2}}}+\frac{I^2\beta G^2N^{\frac{1}{2}}}{ET}+\frac{BLE^{\frac{1}{2}}}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}T^{\frac{1}{2}}}\right)\overset{(b)}{=}O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+BL\right)}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}T^{\frac{1}{2}}}+\frac{EG^2N^{\frac{1}{2}}}{\beta T}\right),\end{aligned}\tag{26}$$

where (a) follows because $\gamma\leq 1/(4L)$, and thus $1-2\gamma L>1/2$; (b) follows because from Lemma 1, $I=\lceil N/K\rceil E=O(E/\beta)$.

When $T\geq EN^{3/2}/\beta$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{t-1})\right\|^2\right]=O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+BL\right)}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}T^{\frac{1}{2}}}\right)\overset{(a)}{=}O\left(\frac{E^{\frac{1}{2}}}{N^{\frac{1}{4}}T^{\frac{1}{2}}}\right),\tag{27}$$

where (a) follows if we care about N, T and E, and regard other parameters (gradient norm G, variance σ, loss difference B, smooth factor L and ratio β) as constants.

□

Appendix G: Convergence on Example 1

In this section, we demonstrate that FedLaAvg converges in Example 1, where FedAvg produces an arbitrarily poor-quality result. The convergence analysis of FedLaAvg for this simple example sheds light on the analysis for the general case of non-convex optimization.

THEOREM 4. *Suppose each client computes the exact (not stochastic) gradient. In Example 1, after T iterations, FedLaAvg with the learning rate $\gamma=1/(2\sqrt{T})$ produces a solution $\hat{\mathbf{x}}$ that is within $O(1/\sqrt{T})$ range of the optimal solution $\mathbf{x}^*$: $\left(\hat{\mathbf{x}}-\mathbf{x}^*\right)^2=O(1/\sqrt{T})$, where we choose $\hat{\mathbf{x}}=\arg\min_{\mathbf{x}^t}f(\mathbf{x}^t)$ as the output.*

Proof of Theorem 4 We recall that

$$f(\mathbf{x})=\left(\mathbf{x}-\frac{\mathbf{e}_1+\mathbf{e}_2}{2}\right)^2+\frac{\left(\mathbf{e}_1-\mathbf{e}_2\right)^2}{4}+\frac{1}{2}\sum_{i=1}^{2}\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[\left(\xi_i-\mathbf{e}_i\right)^2\right],$$

where the latter two terms are not associated with the variable $\mathbf{x}$. Hence, we only need to focus on the following part of the loss function: $\hat{f}(\mathbf{x})=\left(\mathbf{x}-\mathbf{x}^*\right)^2$, where $\mathbf{x}^*=\frac{\mathbf{e}_1+\mathbf{e}_2}{2}$ is the optimal solution. Note that

$$\hat{f}(\mathbf{x}^t)-\hat{f}(\mathbf{x}^{t-1})=\left(\mathbf{x}^t-\mathbf{x}^{t-1}\right)^2+2\left(\mathbf{x}^{t-1}-\mathbf{x}^*\right)\left(\mathbf{x}^t-\mathbf{x}^{t-1}\right).\tag{28}$$

We calculate the difference of $\mathbf{x}$ between two successive iterations:

$$\mathbf{x}^t-\mathbf{x}^{t-1}=-\frac{\gamma}{2}\left(\mathbf{g}_1^{T_1^t}+\mathbf{g}_2^{T_2^t}\right)=-\gamma\left(\mathbf{x}^{T_1^t}-\mathbf{e}_1+\mathbf{x}^{T_2^t}-\mathbf{e}_2\right)=-\gamma\left(\mathbf{x}^{T_1^t}+\mathbf{x}^{T_2^t}-2\mathbf{x}^*\right),\tag{29}$$

where T_i^t is defined in Section 4.2. Hence, we have

$$\begin{aligned}&2\left(\mathbf{x}^{t-1}-\mathbf{x}^*\right)\left(\mathbf{x}^t-\mathbf{x}^{t-1}\right)=-\gamma\left(2\mathbf{x}^{t-1}-2\mathbf{x}^*\right)\left(\mathbf{x}^{T_1^t}+\mathbf{x}^{T_2^t}-2\mathbf{x}^*\right)\\ =&-\frac{\gamma}{2}\left(2\mathbf{x}^{t-1}-2\mathbf{x}^*\right)^2-\frac{\gamma}{2}\left(\mathbf{x}^{T_1^t}+\mathbf{x}^{T_2^t}-2\mathbf{x}^*\right)^2+\frac{\gamma}{2}\left(2\mathbf{x}^{t-1}-\mathbf{x}^{T_1^t}-\mathbf{x}^{T_2^t}\right)^2.\end{aligned}\tag{30}$$

Substituting (29) and (30) into (28), we have

$$\hat{f}(\mathbf{x}^t)-\hat{f}(\mathbf{x}^{t-1})\leq-2\gamma\left(\mathbf{x}^{t-1}-\mathbf{x}^*\right)^2+\frac{\gamma}{2}\left(2\mathbf{x}^{t-1}-\mathbf{x}^{T_1^t}-\mathbf{x}^{T_2^t}\right)^2,\tag{31}$$

which follows from $0<\gamma\leq 1/2$.

The algorithm starts from model parameters $\mathbf{x}^0$. When client 1 is available, $\mathbf{x}$ moves towards $\mathbf{e}_1$, and when client 2 is available, $\mathbf{x}$ moves towards $\mathbf{e}_2$. Hence, $\mathbf{x}$ is always within $G/2$ range of $\mathbf{x}^*$: $-\frac{G}{2}\leq \mathbf{x}^t-\mathbf{x}^*\leq \frac{G}{2}$, $\forall t\geq 0$, where $G=\max\{2(\mathbf{x}^0-\mathbf{x}^*),|\mathbf{e}_1-\mathbf{e}_2|\}$ is the the largest gradient norm during the training process. Substituting the inequality into (29), we have $-\gamma G\leq \mathbf{x}^t-\mathbf{x}^{t-1}\leq \gamma G$. Referring to the specific client availability model in this example, we have $t-T_i^t\leq I=\max\{t_1,t_2\}$, $i=1,2$. Therefore, when $t\geq T_i^t+2$, summing the inequality over iterations from T_i^t+1 to $t-1$, we have

$$-\gamma IG\leq \mathbf{x}^{t-1}-\mathbf{x}^{T_i^t}\leq \gamma IG,\ i=1,2. \tag{32}$$

Note that when $t=T_i^t$ or $t=T_i^t+1$, the above formula also holds.

Substituting (32) into (31), we have $\hat{f}(\mathbf{x}^t)-\hat{f}(\mathbf{x}^{t-1})\leq -2\gamma(\mathbf{x}^{t-1}-\mathbf{x}^*)^2+2\gamma^3I^2G^2$. Then, rearranging the formula, we have $(\mathbf{x}^{t-1}-\mathbf{x}^*)^2\leq \frac{1}{2\gamma}\left(\hat{f}(\mathbf{x}^{t-1})-\hat{f}(\mathbf{x}^t)\right)+\gamma^2I^2G^2$. Summing this inequality over iterations from 1 to T, we have

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}^{t-1}-\mathbf{x}^*\right)^2\leq \frac{1}{2\gamma T}\left(\hat{f}(\mathbf{x}^0)-\hat{f}(\mathbf{x}^T)\right)+\gamma^2I^2G^2\leq \frac{1}{2\gamma T}\left(\hat{f}(\mathbf{x}^0)-\hat{f}(\mathbf{x}^*)\right)+\gamma^2I^2G^2. \tag{33}$$

Substituting $\gamma=1/(2\sqrt{T})$ into (33), we have

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}^{t-1}-\mathbf{x}^*\right)^2\leq \frac{1}{\sqrt{T}}\left(\hat{f}(\mathbf{x}^0)-\hat{f}(\mathbf{x}^*)\right)+\frac{I^2G^2}{4T}=\frac{1}{\sqrt{T}}\left(f(\mathbf{x}^0)-f(\mathbf{x}^*)\right)+\frac{I^2G^2}{4T}.$$

Finally, we have $(\hat{\mathbf{x}}-\mathbf{x}^*)^2\leq \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}^{t-1}-\mathbf{x}^*)^2\leq \frac{1}{\sqrt{T}}(f(\mathbf{x}^0)-f(\mathbf{x}^*))+\frac{I^2G^2}{4T}=O\left(\frac{1}{\sqrt{T}}\right)$.

□

Appendix H: Detailed Proof of Theorem 3

To make the proof more concise, we introduce an mathematically equivalent Algorithm 3 of Algorithm 2. When t is not multiple of C, $\mathbf{x}^t$ is intermediate variable for mathematical analysis; $\mathbf{g}_i^\tau$ $(\tau\leq 0)$ is defined to avoid undefined symbols when $R_i^{r^t}=0$ in Line 13. It can be proved by induction that all variables defined in Algorithm 2 are consistent with those in Algorithm 3. With equivalence between Algorithm 2 and 3 established, we introduce the corresponding equivalent lemmas of Lemmas 1–4.

LEMMA 5. *Under Assumption 5, with $I=\lceil N/K\rceil E-1$, $\forall r,\forall i$, we have $r-R_i^r\leq I$.*

Proof of Lemma 5 Replacing t with r and T_i^t with R_i^r, the proof is the same with that of Lemma 1.

□

LEMMA 6. $\mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right]\leq G^2$, $\mathbb{E}\left[\left\|\mathbf{g}_i^t-\nabla f_i(\mathbf{x}_i^{t-1})\right\|^2\right]\leq \sigma^2,\forall i,\forall t$.

Proof of Lemma 6 Replacing $\mathbf{x}^{t-1}$ with $\mathbf{x}_i^{t-1}$, the proof is the same with Lemma 2.

□

LEMMA 7. *Corresponding lemma of Lemma 3: $\forall i,\forall t$, we have*

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(\mathbf{g}_i^{R_i^{r^t}C-r^tC+t}-\nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right)\right\|^2\right]\\ =&\sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{g}_i^{R_i^{r^t}C-r^tC+t}-\nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\|^2\right].\end{aligned}$$

Algorithm 3 An equivalent Algorithm of Algorithm 2

Input: Initial model $\mathbf{x}^0$
$\mathbf{g}_i^\tau \leftarrow \mathbf{0}$, $\forall i \in \{1,2,\cdots,N\}$, $\tau \in \{0,-1,\cdots,1-C\}$.
$R_i^0 \leftarrow \mathbf{0}$, $\forall i \in \{1,2,\cdots,N\}$.
for $t=1$ **to** RC **do**
  $r^t \leftarrow \lfloor (t-1)/C \rfloor + 1$.
  if $t-1$ is a multiple of C **then**
    $\hat{\mathcal{C}}^{r^t} \leftarrow$ the set of available clients in round r^t.
    $\hat{\mathcal{B}}^{r^t} \leftarrow K$ clients from $\hat{\mathcal{C}}^{r^t}$ with the lowest $R_i^{r^t-1}$ values.
    Update $R_i^{r^t}$ values: $R_i^{r^t} \leftarrow r^t$, $\forall i \in \hat{\mathcal{B}}^{r^t}$; $R_i^{r^t} \leftarrow R_i^{r^t-1}$, $\forall i \notin \hat{\mathcal{B}}^{r^t}$.
    $\mathbf{x}_i^{t-1} \leftarrow \mathbf{x}^{t-1}$, $\forall i \in \hat{\mathcal{B}}^{r^t}$.
  end if
  $\mathbf{g}_i^t \leftarrow \nabla F\left(\mathbf{x}_i^{t-1}; \xi_i^t\right)$, $\forall i \in \hat{\mathcal{B}}^{r^t}$.
  Update the global model parameters: $\mathbf{x}^t \leftarrow \mathbf{x}^{t-1} - \gamma \sum_{i=1}^N \mathbf{g}_i^{R_i^{r^t} C - r^t C + t}$.
  Update the local model parameters: $\mathbf{x}_i^t \leftarrow \mathbf{x}_i^{t-1} - \gamma \mathbf{g}_i^t$.
end for

Proof of Lemma 7 Replacing $\nabla f_i(\mathbf{x}^{T_i^t-1})$ with $\nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})$ and $\mathbf{g}_i^{T_i^t}$ with $\mathbf{g}_i^{R_i^{r^t}C-r^tC+t}$, the proof is the same with that of Lemma 3.

□

Note that Lemma 4 and Corollary 1 still hold. Their proof follows as well if we replace the relation $\mathbf{x}^\tau - \mathbf{x}^{\tau-1} = \sum_{j=1}^N \mathbf{g}_j^{T_j^\tau}$ with $\mathbf{x}^\tau - \mathbf{x}^{\tau-1} = \sum_{j=1}^N \mathbf{g}_j^{R_j^{r_\tau} C - r_\tau C + \tau}$.

Based on the above lemmas, we derive the proof of Theorem 3. The proof is similar to that of Theorem 2 and Corollary 2. We illustrate it in detail as follow. Fix $t \geq 1$, by Assumption 1, we have

$$\mathbb{E}\left[f(\mathbf{x}^t)\right] \leq \mathbb{E}\left[f(\mathbf{x}^{t-1})\right] + \frac{L}{2}\mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2\right] + \mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1}\right\rangle\right]. \tag{34}$$

Focus on the second term on the right. Following the procedure of (4), we omit the intermediate results and show the final bound:

$$\mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2\right] \leq \frac{2\gamma^2\sigma^2}{N} + 2\gamma^2 \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\|^2\right]. \tag{35}$$

We define $\hat{T}^t$ as $\min_i \left(R_i^{r^t}C - r^tC + t\right)$. Similar to (6)–(10), we separate the third term in (34) and derive the bound:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}), \mathbf{x}^t - \mathbf{x}^{t-1}\right\rangle\right] \\
=&-\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{g}_i^{R_i^{r^t}C-r^tC+t} - \nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right)\right\rangle\right] \\
&-\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{t-1}) - \nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\rangle\right] \\
&-\gamma\mathbb{E}\left[\left\langle \nabla f(\mathbf{x}^{\hat{T}^t-1}), \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^{R_i^{r^t}C-r^tC+t}\right\rangle\right]. \\
\leq&-\gamma^2L\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i^{R_i^{r^t}C-r^tC+t-1})\right\|^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right] \\
&+\frac{\gamma^2ICL\left(G^2+\sigma^2\right)}{2\sqrt{N}} + \frac{\gamma^3I^2C^2L^2G^2}{2\left(1-2\gamma L\right)} + \frac{\gamma^3I^2C^2L^2G^2}{2}.
\end{aligned} \tag{36}$$

Further substituting (35) and (36) into (34), we have

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{x}^t)\right] - \mathbb{E}\left[f(\mathbf{x}^{t-1})\right] \\
\leq&\frac{\gamma^2\sigma^2L}{N} + \frac{\gamma^2ICL\left(G^2+\sigma^2\right)}{2\sqrt{N}} + \frac{\gamma^3I^2C^2L^2G^2}{2\left(1-2\gamma L\right)} + \frac{\gamma^3I^2C^2L^2G^2}{2} - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right].
\end{aligned} \tag{37}$$

Rearrange the above equation and we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right] \leq&\frac{2\gamma\sigma^2L}{N} + \frac{\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}} + \frac{\gamma^2I^2C^2L^2G^2}{\left(1-2\gamma L\right)} \\
&+\gamma^2I^2C^2L^2G^2 + \frac{2}{\gamma}\left(\mathbb{E}\left[f(\mathbf{x}^{t-1})\right] - \mathbb{E}\left[f(\mathbf{x}^t)\right]\right).
\end{aligned} \tag{38}$$

Summing (38) over t from 1 to RC and dividing both sides by RC, we have

$$\begin{aligned}
\frac{1}{RC}\sum_{t=1}^{RC}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^t-1})\right\|^2\right] \leq&\frac{2\gamma\sigma^2L}{N} + \frac{\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}} \\
&+\frac{\gamma^2I^2C^2L^2G^2}{\left(1-2\gamma L\right)} + \gamma^2I^2C^2L^2G^2 + \frac{2}{\gamma RC}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right] - \mathbb{E}\left[f(\mathbf{x}^*)\right]\right),
\end{aligned} \tag{39}$$

where $\mathbf{x}^*$ is the optimal value for the objective function $f(\mathbf{x})$.

Finally, we build the gap between $\nabla f(\mathbf{x}^{r^tC-1})$ and $\nabla f(\mathbf{x}^{\hat{T}^t-1})$. Lemma 5 implies that $t - \hat{T}^t \leq IC$, and thus $r^tC - \hat{T}^t \leq (I+1)C$. Hence, we have

$$\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{r^tC-1})\right\|^2\right] \leq 2\frac{1}{C}\sum_{\tau=(r_t-1)C+1}^{r_tC}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{\hat{T}^\tau-1})\right\|^2\right], +2\gamma^2\left(I+1\right)^2C^2L^2G^2, \tag{40}$$

which follows from the convexity of $\|\cdot\|^2$and Corollary 1. Summing (40) over $t \in \{C, 2C, \cdots, RC\}$, dividing both sides by R and substituting (39), we have

$$\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right] \leq&\frac{4}{\gamma RC}\left(\mathbb{E}\left[f(\mathbf{x}^0)\right] - \mathbb{E}\left[f(\mathbf{x}^*)\right]\right) + \frac{4\gamma\sigma^2L}{N} \\
&+\frac{2\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}} + \left(\frac{2I^2}{\left(1-2\gamma L\right)} + 4I^2 + 4I + 2\right)\gamma^2C^2L^2G^2.
\end{aligned} \tag{41}$$

Then, we write the $O(\cdot)$ expression of the above equation:

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right]=O\left(\frac{\gamma ICL\left(G^2+\sigma^2\right)}{\sqrt{N}}+\frac{I^2\gamma^2C^2L^2G^2}{(1-2\gamma L)}+\frac{B}{\gamma RC}\right). \tag{42}$$

Substituting γ with $(\beta^{1/2}N^{1/4})/(2LCE^{1/2}R^{1/2})$, we have

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right]=O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2+BL\right)}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}R^{\frac{1}{2}}}+\frac{EG^2N^{\frac{1}{2}}}{\beta R}\right). \tag{43}$$

If we further choose $R>EN^{3/2}/\beta$, we have

$$\frac{1}{R}\sum_{t=1}^{R}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^{rC-1})\right\|^2\right]=O\left(\frac{E^{\frac{1}{2}}\left(G^2+\sigma^2\right)+BL}{\beta^{\frac{1}{2}}N^{\frac{1}{4}}R^{\frac{1}{2}}}\right)\overset{(a)}{=}O\left(\frac{E^{\frac{1}{2}}}{N^{\frac{1}{4}}R^{\frac{1}{2}}}\right), \tag{44}$$

where (a) follows if we care about N and R, and regard other parameters as constants.

□

Appendix I: Comparison between FedAdam and FedLaAvg

FedAdam (Reddi et al. 2021) applies the Adam optimizer (Kingma and Ba 2015) to federated learning to accelerate the convergence under full client availability. We show a comparison between FedLaAvg and FedAdam (Reddi et al. 2021) on MNIST under the default configuration here in Figure 8. Due to the strong adaptiveness of FedAdam, the global model deviates very fast from the global optimal when one group of clients are available, leading to even more serious oscillation compared with FedAvg. From Figure 8, we can see that FedAdam diverges under intermittent client availability.

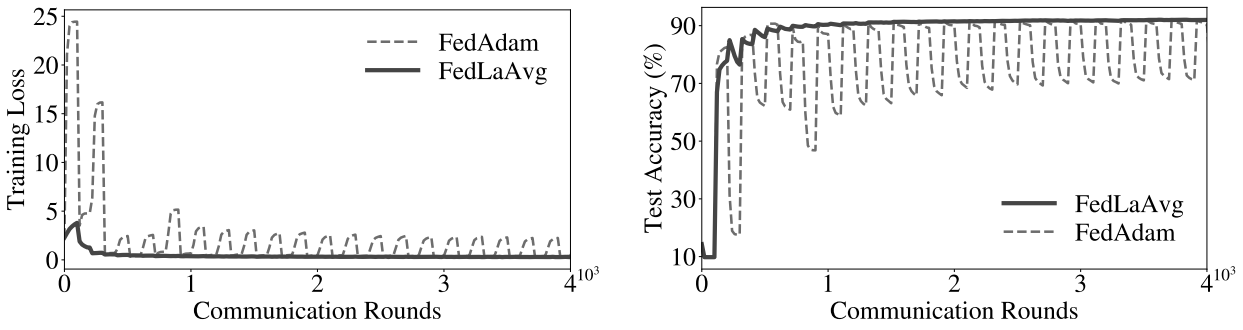

Figure 8 **Performance of FedLaAvg and FedAdam in the MNIST image classification task under the default configuration of $E=100$ and $D=1$.**