Appendix for "Communication-Learning Co-Design for Differentially Private Over-the-Air Federated Distillation"

Appendix A
Proof of Theorem 1

Proof. We first bound the expected difference of the empirical local loss function of WD i between iterations t and $t+1$, i.e.,

$$\mathbb{E}\left[F_i\left(\boldsymbol{\theta}_{i,t+1}\right)-F_i\left(\boldsymbol{\theta}_{i,t}\right)\right]=B=B_1+B_2, \tag{1}$$

where

$$\begin{aligned} B_1&=\mathbb{E}\left[F_i\left(\boldsymbol{\theta}_{i,t+1};\left\{\mathbf{q}_{t+1}^k\right\}\right)-F_i\left(\boldsymbol{\theta}_{i,t+1};\left\{\mathbf{q}_t^k\right\}\right)\right],\\ B_2&=\mathbb{E}\left[F_i\left(\boldsymbol{\theta}_{i,t+1};\left\{\mathbf{q}_t^k\right\}\right)-F_i\left(\boldsymbol{\theta}_{i,t};\left\{\mathbf{q}_t^k\right\}\right)\right]. \end{aligned} \tag{2}$$

By Assumption 1 and Eqn. (4) of the manuscript, we have

$$\begin{aligned} B_2\leq&\underbrace{-\eta_t\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)^\top\mathbb{E}\left[\nabla F_i\left(\boldsymbol{\theta}_{i,t};\left\{\hat{\mathbf{r}}_t^k\right\}\right)\right]}_{C_1}\\ &+\frac{L_1\eta_t^2}{2}\underbrace{\mathbb{E}\left[\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t};\left\{\hat{\mathbf{r}}_t^k\right\}\right)\right\|_2^2\right]}_{C_2}. \end{aligned} \tag{3}$$

According to (1)-(3) of the manuscript,

$$\begin{aligned} B_1=&\frac{\gamma}{B_i}\sum_{b=1}^{B_i}\left(\left\|G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)-\mathbf{q}_{t+1}^{v_i^b}\right\|_2^2-\left\|G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)-\mathbf{q}_t^{v_i^b}\right\|_2^2\right)\\ =&\frac{\gamma}{B_i}\sum_{b=1}^{B_i}\left(2G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)^\top\mathbf{q}_t^{v_i^b}-2G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)^\top\mathbf{q}_{t+1}^{v_i^b}+\left\|\mathbf{q}_{t+1}^{v_i^b}\right\|_2^2-\left\|\mathbf{q}_t^{v_i^b}\right\|_2^2\right)\\ =&\frac{\gamma}{B_i}\sum_{b=1}^{B_i}\left(\mathbf{q}_{t+1}^{v_i^b}+\mathbf{q}_t^{v_i^b}-2G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)\right)^\top\left(\mathbf{q}_{t+1}^{v_i^b}-\mathbf{q}_t^{v_i^b}\right)\\ \overset{(a)}{\leq}&\frac{\gamma}{B_i}\sum_{b=1}^{B_i}\left\|\mathbf{q}_{t+1}^{v_i^b}+\mathbf{q}_t^{v_i^b}-2G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)\right\|_2\left\|\mathbf{q}_{t+1}^{v_i^b}-\mathbf{q}_t^{v_i^b}\right\|_2\\ \overset{(b)}{\leq}&\frac{\gamma}{B_i}\sum_{b=1}^{B_i}\left(\left\|\mathbf{q}_{t+1}^{v_i^b}\right\|_2+\left\|\mathbf{q}_t^{v_i^b}\right\|_2+\left\|2G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)\right\|_2\right)\times\left\|\mathbf{q}_{t+1}^{v_i^b}-\mathbf{q}_t^{v_i^b}\right\|_2\\ \overset{(c)}{\leq}&\frac{\gamma}{B_i}\sum_{k=1}^{K}B_i^k 4\left\|\mathbf{q}_{t+1}^k-\mathbf{q}_t^k\right\|_2\\ =&\frac{4\gamma}{B_i}\sum_{k=1}^{K}B_i^k\left\|\frac{1}{B^k}\sum_{i=1}^{M}\sum_{b\in\mathcal{B}_i^k}\left(G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)-G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)\right)\right\|_2\\ \overset{(d)}{\leq}&\frac{4\gamma}{B_i}\sum_{k=1}^{K}\frac{B_i^k}{B^k}\sum_{i=1}^{M}\sum_{b\in\mathcal{B}_i^k}L_2\left\|\boldsymbol{\theta}_{i,t+1}-\boldsymbol{\theta}_{i,t}\right\|_2\\ =&\frac{4\gamma}{B_i}\sum_{k=1}^{K}\frac{B_i^k}{B^k}\sum_{i=1}^{M}\sum_{b\in\mathcal{B}_i^k}L_2\left\|-\eta_t\nabla F_i\left(\boldsymbol{\theta}_{i,t};\left\{\hat{\mathbf{r}}_t^k\right\}\right)\right\|_2\\ \overset{(e)}{\leq}&4\gamma L_2\eta_t S, \end{aligned} \tag{4}$$

where (a) is due to Cauchy–Schwarz Inequality. Inequality (b) follows from the triangle inequality. For the inequality (c), we first see that the l_1 norms of $\mathbf{q}_{t+1}^{v_i^b}$, $\mathbf{q}_t^{v_i^b}$ and $G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)$ are equal to 1 since they are the average of probability vectors (outputs of the softmax layer) whose entries sum to 1. Then according to the property of l_1, l_2 norms, for any vector $\mathbf{x}\in\mathbb{R}^d$, we have $\|\mathbf{x}\|_2\leq\|\mathbf{x}\|_1$, and the l_2 norms of $\mathbf{q}_{t+1}^{v_i^b}$, $\mathbf{q}_t^{v_i^b}$ and $G_{\boldsymbol{\theta}_{i,t+1}}\left(\mathbf{u}_i^b\right)$ are bounded by 1. Subsequently, we interchange the summation of sample indices b with class indices k. Inequality (d) is due to Jensen's Inequality and Assumption 2 of the manuscript. Inequality (e) follows from Assumption 3 of the manuscript.

Next, we bound C_1 and C_2. By (1) of the manuscript and chain rule, we have

$$\begin{aligned} \nabla F_i\left(\boldsymbol{\theta}_{i,t};\left\{\hat{\mathbf{r}}_t^k\right\}\right)=&\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)+\frac{2\gamma}{B_i}\sum_{b=1}^{B_i}\frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial\boldsymbol{\theta}_{i,t}}\times\left(G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)-\hat{\mathbf{r}}_t^{v_i^b}\right)\\ =&\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)+\frac{2\gamma}{B_i}\sum_{b=1}^{B_i}\frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial\boldsymbol{\theta}_{i,t}}\times\left(G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)-\hat{\mathbf{r}}_t^{v_i^b}+\mathbf{q}_t^{v_i^b}-\mathbf{q}_t^{v_i^b}\right)\\ =&\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)+\frac{2\gamma}{B_i}\sum_{b=1}^{B_i}\frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial\boldsymbol{\theta}_{i,t}}\times\left(\mathbf{q}_t^{v_i^b}-\hat{\mathbf{r}}_t^{v_i^b}\right). \end{aligned} \tag{5}$$

By plugging (5) into C_1, we get

$$\begin{aligned} C_1=&-\eta_t\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2-\frac{2\eta_t\gamma}{B_i}\sum_{b=1}^{B_i}\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)^\top\frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial\boldsymbol{\theta}_{i,t}}\times\left(\mathbf{q}_t^{v_i^b}-\mathbb{E}\left[\hat{\mathbf{r}}_t^{v_i^b}\right]\right)\\ \overset{(f)}{\leq}&-\eta_t\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2+2\gamma L_2\eta_t\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2\sum_{k=1}^{K}\frac{B_i^k}{B_i}\times\left\|\mathbf{q}_t^k-\mathbb{E}\left[\hat{\mathbf{r}}_t^k\right]\right\|_2\\ =&-\eta_t\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2+2\gamma L_2\eta_t\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2\Phi_{1,i,t}, \end{aligned} \tag{6}$$

where (f) is due to the Cauchy–Schwarz Inequality that

$$\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)^\top\frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial\boldsymbol{\theta}_{i,t}}\left(\mathbf{q}_t^{v_i^b}-\mathbb{E}\left[\hat{\mathbf{r}}_t^{v_i^b}\right]\right)\geq$$

$$- \left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \left\|\frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial \boldsymbol{\theta}_{i,t}}\right\|_2 \left\|\mathbf{q}_t^{v_i^b} - \mathbb{E}\left[\hat{\mathbf{r}}_t^{v_i^b}\right]\right\|_2. \quad (7)$$

The inequality is also due to the property of the matrix $2-$norm that for any matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ and vector $\mathbf{x} \in \mathbb{R}^N$, $\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x}\|_2$. Notice that by Assumption 2 of the manuscript, we can bound the norm of the partial derivative of the learned model function $G_{\boldsymbol{\theta}_{i,t}}(\cdot)$ by the Lipschitz constant L_2. The last equality follows from the zero mean of $\hat{\mathbf{m}}_t^k$ and the definition of $\Phi_{1,i,t}$ in (14) of the manuscript.

Similarly, with Eqn. (5), we have

$$\begin{aligned}
C_2 &= \left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2 + \frac{4\gamma}{B_i}\sum_{b=1}^{B_i} \nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)^\top \frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial \boldsymbol{\theta}_{i,t}} \Big(\mathbf{q}_t^{v_i^b} - \\
&\quad \mathbb{E}\left[\hat{\mathbf{r}}_t^{v_i^b}\right]\Big) + \mathbb{E}\left[\left\|\frac{2\gamma}{B_i}\sum_{b=1}^{B_i} \frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial \boldsymbol{\theta}_{i,t}}\left(\mathbf{q}_t^{v_i^b} - \hat{\mathbf{r}}_t^{v_i^b}\right)\right\|_2^2\right] \\
&\overset{(g)}{\leq} \left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2 + 4\gamma L_2 \left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \Phi_{1,i,t} \\
&\quad + \mathbb{E}\left[\left\|\frac{2\gamma}{B_i}\sum_{b=1}^{B_i} \frac{\partial G_{\boldsymbol{\theta}_{i,t}}\left(\mathbf{u}_i^b\right)}{\partial \boldsymbol{\theta}_{i,t}}\left(\mathbf{q}_t^{v_i^b} - \hat{\mathbf{r}}_t^{v_i^b}\right)\right\|_2^2\right] \\
&\overset{(h)}{\leq} \left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2 + 4\gamma L_2 \left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \Phi_{1,i,t} \\
&\quad + 4\gamma^2 L_2^2\left(\Phi_{1,i,t}^2 + \Phi_{2,i,t}\right),
\end{aligned} \quad (8)$$

where (g) follows from (6). The inequality (h) is due to Jensen's Inequality as well as Assumption 2 of the manuscript. By the independence and the zero mean of $\hat{\mathbf{m}}_t^k$, $\mathbb{E}\left[\sum_{k=1}^K \frac{B_i^k}{B_i}\left\|\mathbf{q}_t^{v_i^b} - \hat{\mathbf{r}}_t^{v_i^b}\right\|_2^2\right]$ can be decomposed into $\Phi_{1,i,t}^2 + \Phi_{2,i,t}$.

Plug (3), (4), (6) and (8) back to (1), we get

$$\begin{aligned}
B &\leq \left(\frac{\eta_t^2 L_1}{2} - \eta_t\right)\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2 + L_1 L_2^2 \eta_t^2\left(\Phi_{1,i,t}^2 + \Phi_{2,i,t}\right) \\
&\quad \times 2\gamma^2 + 2\gamma L_2 \eta_t\left(L_1\eta_t + 1\right)\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \Phi_{1,i,t} \\
&\quad + 4\gamma L_2 \eta_t S \\
&\leq -\frac{\eta_t}{2}\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2 + 2\gamma L_2 \eta_t\left(L_1\eta_t + 1\right)\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \\
&\quad \times \Phi_{1,i,t} + 2\gamma^2 L_1 L_2^2 \eta_t^2\left(\Phi_{1,i,t}^2 + \Phi_{2,i,t}\right) + 4\gamma L_2 \eta_t S.
\end{aligned} \quad (9)$$

The last inequality is due to $-\eta_t + \frac{\eta_t^2 L_1}{2} \leq -\frac{\eta_t}{2}$ with $\eta_t \leq \frac{1}{L_1}$. We rearrange the terms, divide both sides by $\frac{\eta_t^2}{2}$ and sum over $t = 0$ to $T-1$ to obtain

$$\begin{aligned}
\underbrace{\sum_{t=0}^{T-1}\frac{1}{\eta_t}\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2^2}_{C_3} &\leq \sum_{t=0}^{T-1}\frac{2}{\eta_t^2}\mathbb{E}\left[F_i\left(\boldsymbol{\theta}_{i,t}\right) - F_i\left(\boldsymbol{\theta}_{i,t+1}\right)\right] \\
&\quad + \sum_{t=0}^{T-1}\frac{4}{\eta_t}\gamma L_2\left(L_1\eta_t + 1\right)\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \\
&\quad \times \Phi_{1,i,t} + \sum_{t=0}^{T-1}\frac{8}{\eta_t}\gamma L_2 S \\
&\quad + \sum_{t=0}^{T-1} 4\gamma^2 L_1 L_2^2\left(\Phi_{1,i,t}^2 + \Phi_{2,i,t}\right).
\end{aligned} \quad (10)$$

According to the proof of Theorem 3.5 in [1],

$$\begin{aligned}
C_3 &\leq \frac{2}{\eta_0^2}\left[F_i\left(\boldsymbol{\theta}_{i,0}\right) + \sum_{t=1}^{T-1}\left(t - (t-1)\right)\mathbb{E}\left[F_i\left(\boldsymbol{\theta}_{i,t}\right)\right]\right] \\
&\quad + \sum_{t=0}^{T-1}\frac{8}{\eta_t}\gamma L_2 S + \sum_{t=0}^{T-1} 4\gamma^2 L_1 L_2^2\left(\Phi_{1,i,t}^2 + \Phi_{2,i,t}\right) \\
&\quad + \sum_{t=0}^{T-1}\frac{4}{\eta_t}\gamma L_2\left(L_1\eta_t + 1\right)\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \Phi_{1,i,t} \\
&\leq \frac{2T f_{i,max}}{\eta_0^2} + \sum_{t=0}^{T-1}\frac{4}{\eta_t}\gamma L_2\left(L_1\eta_t + 1\right)\left\|\nabla F_i\left(\boldsymbol{\theta}_{i,t}\right)\right\|_2 \Phi_{1,i,t} \\
&\quad + \sum_{t=0}^{T-1}\frac{8}{\eta_t}\gamma L_2 S + \sum_{t=0}^{T-1} 4\gamma^2 L_1 L_2^2\left(\Phi_{1,i,t}^2 + \Phi_{2,i,t}\right).
\end{aligned} \quad (11)$$

Recall that $\hat{\boldsymbol{\theta}}_{i,T}$ is randomly chosen from $\{\boldsymbol{\theta}_{i,t}, \forall t\}$ at all the previous iterations with probability $Pr\left\{\hat{\boldsymbol{\theta}}_{i,T} = \boldsymbol{\theta}_{i,t}\right\} = \frac{1/\eta_t}{\sum_{t=0}^{T-1} 1/\eta_t}$. We divide both sides by $\sum_{t=0}^{T-1}\frac{1}{\eta_t}$, which gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F_i\left(\hat{\boldsymbol{\theta}}_{i,T}\right)\right\|_2^2\right] &\leq \frac{3 f_{i,max}}{\eta_0\sqrt{T}} + \sum_{t=0}^{T-1}\frac{6\gamma\eta_0 L_2\left(L_1\eta_t + 1\right)}{\eta_t} \\
&\quad \times \frac{\left\|\nabla F_i(\boldsymbol{\theta}_{i,t})\right\|_2 \Phi_{1,i,t}}{T^{\frac{3}{2}}} + 8\gamma L_2 S \\
&\quad + \sum_{t=0}^{T-1} 6\eta_0\gamma^2 L_2^2 L_1\left(\frac{\Phi_{1,i,t}^2 + \Phi_{2,i,t}}{T^{\frac{3}{2}}}\right)
\end{aligned} \quad (12)$$

where the inequality holds because $\sum_{t=0}^{T-1}\frac{1}{\eta_t} \geq \frac{1}{\eta_0}\int_{t=0}^{T}\sqrt{t}\,dt = \frac{2}{3\eta_0}T^{\frac{3}{2}}$. We summarize all the terms related to the training rounds T and per-round transceiver design $\mathcal{P}_t$ in (12) into the convergence gap function $\Omega_i\left(T, \{\mathcal{P}_t\}_{t=0}^{T-1}\right)$, i.e.,

$$\begin{aligned}
\Omega_i\left(T, \{\mathcal{P}_t\}\right) &= \frac{3 f_{i,max}}{\eta_0\sqrt{T}} + \sum_{t=0}^{T-1} 6\eta_0\gamma^2 L_2^2 L_1\left(\frac{\Phi_{1,i,t}^2 + \Phi_{2,i,t}}{T^{\frac{3}{2}}}\right) \\
&\quad + \sum_{t=0}^{T-1} 6\gamma\eta_0 L_2 \frac{\left(L_1\eta_t + 1\right)\left\|\nabla F_i(\boldsymbol{\theta}_{i,t})\right\|_2 \Phi_{1,i,t}}{\eta_t T^{\frac{3}{2}}}.
\end{aligned}$$

The proof of Theorem 1 is thus completed. □

Appendix B
Proof of Lemma 1 and Theorem 2

Proof. We first derive an upper-bound of the Euclidean distance between any two points in the simplex $\Sigma_{i,t}^k$. Taking any $\mathbf{q}_1$ and $\mathbf{q}_2 \in \Sigma_{i,t}^k$, the Euclidean distance between them is calculated as $\|\mathbf{q}_1 - \mathbf{q}_2\|_2$. By expanding the expression, we have

$$\|\mathbf{q}_1 - \mathbf{q}_2\|_2 = \sqrt{\|\mathbf{q}_1\|_2 + \|\mathbf{q}_2\|_2 - 2\mathbf{q}_1^\top \mathbf{q}_2}. \quad (13)$$

By the property that the l_2 norm of a vector is upper-bounded by its l_1 norm, we have

$$\|\mathbf{q}_1\|_2 \leq \|\mathbf{q}_1\|_1 = 1. \quad (14)$$

The equality follows from the property of the probability simplex. Furthermore, since the simplex only contains vectors

with non-negative elements, we have $\mathbf{q}_1^\top \mathbf{q}_2 \geq 0$ and thus $-2\mathbf{q}_1^\top \mathbf{q}_2 \leq 0$. Combine the results above, we have

$$\|\mathbf{q}_1 - \mathbf{q}_2\|_2 = \sqrt{\|\mathbf{q}_1\|_2 + \|\mathbf{q}_2\|_2 - 2\mathbf{q}_1^\top \mathbf{q}_2} \leq \sqrt{2}. \quad (15)$$

Therefore, the Euclidean distance between any two points in this simplex is upper-bounded by $\sqrt{2}$.

Then, we prove the achievability of this upper-bound. Take any two vertices of this simplex, e.g., $(1, 0, 0, \cdots, 0)$ and $(0, 1, 0, \cdots, 0)$, where the corresponding Euclidean distance is $\sqrt{2}$. Therefore, the maximum Euclidean distance between any two points in the probability simplex $\Sigma_{i,t}^k$ is $\sqrt{2}$ for any i, t, k.

Subsequently, we analyze the maximum magnitude of the signal $\hat{\mathbf{q}}_{i,t}^k$ for each class k, i.e.,

$$\begin{aligned}\left\|\hat{\mathbf{q}}_{i,t}^k\right\|_2 &= \left\|h_{i,t}P_{1,i,t}^k\sqrt{K}\mathbf{q}_{i,t}^k\right\|_2 \leq \sqrt{K}\left|h_{i,t}P_{1,i,t}^k\right| \times \left\|\mathbf{q}_{i,t}^k\right\|_2 \\ &\leq \sqrt{K}\left|h_{i,t}P_{1,i,t}^k\right| \times \left\|\mathbf{q}_{i,t}^k\right\|_1 = \sqrt{K}\left|h_{i,t}P_{1,i,t}^k\right|. \end{aligned} \quad (16)$$

Accordingly, given that the maximum Euclidean distance between any two points in the probability simplex $\Sigma_{i,t}^k$ is $\sqrt{2}$, we can obtain the upper-bound of the l-2 sensitivity in (15) of the manuscript, i.e.,

$$\begin{aligned}\Delta_{i,t}^k &= \max_{\mathcal{B}_i, \mathcal{B}_i'} \left\|\hat{\mathbf{q}}_{i,t}^k\left(\mathcal{B}_i\right) - \hat{\mathbf{q}}_{i,t}^k\left(\mathcal{B}_i'\right)\right\|_2 \\ &= \frac{1}{B_i}\left(\sum_{b\neq b'}\left\|\hat{\mathbf{q}}_{i,t}^k\left(\mathbf{u}_i^b\right) - \hat{\mathbf{q}}_{i,t}^k\left(\mathbf{u}_i^b\right)\right\|_2\right. \\ &\quad \left. + \max_{\mathbf{u}_i^{b'}, \hat{\mathbf{u}}_i^{b'}}\left\|\hat{\mathbf{q}}_{i,t}^k\left(\mathbf{u}_i^{b'}\right) - \hat{\mathbf{q}}_{i,t}^k\left(\hat{\mathbf{u}}_i^{b'}\right)\right\|_2\right) \\ &\leq \frac{\sqrt{2K}\left|h_{i,t}P_{1,i,t}^k\right|}{B_i}, \end{aligned} \quad (17)$$

where $\mathbf{u}_i^{b'}$ and $\hat{\mathbf{u}}_i^{b'}$ are the specific altered sample in the neighboring datasets. The proof of Lemma 1 is thus completed.

According to Theorem 1 in [2], to achieve $(\varepsilon_i, \delta_i)$-DP of WD i after T training rounds, the standard deviation σ_i^k of the Gaussian DP noise imposed on the transmitted soft prediction for class k of WD i per round satisfies

$$\sigma_i^k = \frac{\Delta_{i,t}^k\sqrt{2T\ln\frac{1}{\delta_i}}}{\varepsilon_i}. \quad (18)$$

Plugging in the obtained upper-bound of the sensitivity $\Delta_{i,t}^k$ in (17), we have

$$\sum_{j=1}^M\left|h_{j,t}P_{2,j,t}^k\right|^2 + \sigma_n^2 \geq \max_{i\in\mathcal{M}} 4TK\left|h_{i,t}P_{1,i,t}^k\right|^2\rho_i, \forall k, t, \quad (19)$$

where ρ_i is defined in Theorem 2 of the manuscript. The proof of Theorem 2 is thus completed. □

Appendix C
Proof of Proposition 1

Proof. For each training round t, we see that

$$\begin{aligned}\sum_{i=1}^M \Omega_i\left(T, \{\mathcal{P}_t\}_{t=0}^{T-1}\right) &= \sum_{i=1}^M\sum_{t=0}^{T-1} A_2\frac{\Phi_{1,i,t}^2 + \Phi_{2,i,t}}{T^{\frac{3}{2}}} + \sum_{i=1}^M\sum_{t=0}^{T-1} \\ A_1\frac{(L_1\eta_t+1)\left\|\nabla F_i(\boldsymbol{\theta}_{i,t})\right\|_2\Phi_{1,i,t}}{\eta_t T^{\frac{3}{2}}} &\overset{(a)}{\geq} \sum_{i=1}^M\sum_{t=0}^{T-1} A_2\frac{\Phi_{2,i,t}}{T^{\frac{3}{2}}}, \end{aligned} \quad (20)$$

where $A_1 = 6\gamma\eta_0 L_2$ and $A_2 = 6\eta_0\gamma^2 L_2^2 L_1$ are constant terms. The equality (a) holds if $\Phi_{1,i,t} = 0, \forall i, t$, which leads to

$$\frac{h_{i,t}P_{1,i,t}^k\sqrt{K}}{\lambda_t^k} - \frac{B_i^k}{B^k} = 0, \quad \forall i, t, k. \quad (21)$$

The optimal solution of $P_{1,i,t}^k$ can be solved as

$$P_{1,i,t}^{k^*} = \frac{B_i^k\lambda_t^k\overline{h}_{i,t}}{B^k\sqrt{K}\left|h_{i,t}\right|^2}, \quad \forall i, t, k. \quad (22)$$

Then, according to the peak transmit power constraint in (9) of the manuscript, we have

$$\begin{aligned}\left(\lambda_t^k\right)^2 &= \frac{\left(B^k\right)^2 K\left|h_{i,t}P_{1,i,t}^k\right|^2}{\left(B_i^k\right)^2} \\ &\leq \frac{\left(B^k\right)^2 K\left|h_{i,t}\right|^2\left(P_i - \left|P_{2,i,t}^k\right|^2\right)}{\left(B_i^k\right)^2}, \quad \forall i, t, k. \end{aligned} \quad (23)$$

With $\left\{P_{1,i,t}^{k^*}, \forall i, k, t\right\}$ in (22), the optimization problem (P1) of the manuscript over $\left\{\left\{P_{2,i,t}^k, \forall i\right\}, \lambda_t^k, \forall t, k\right\}$ is given by

$$\begin{aligned}\text{(P2)} \quad \min_{\left\{P_{2,i,t}^k, \lambda_t^k, \forall i,t,k\right\}} \quad & \sum_{i=1}^M\sum_{t=0}^{T-1} A_2\frac{\Phi_{2,i,t}}{T^{\frac{3}{2}}} \\ \text{s.t.} \quad & \lambda_t^k \leq \Psi_{1,t}^k\left(B^k, \left\{P_i, h_{i,t}, P_{2,i,t}^k, B_i^k, \forall i\right\}\right), \\ & \forall t, k, \\ & \lambda_t^k \leq \Psi_{2,t}^k\left(B^k, \left\{h_{i,t}, P_{2,i,t}^k, B_i^k, \rho_i, \forall i\right\}\right), \\ & \forall t, k. \end{aligned} \quad (24)$$

Here

$$\begin{aligned}\Psi_{1,t}^k(\cdot) &= \min_{i\in\mathcal{M}}\frac{B^k\left|h_{i,t}\right|\sqrt{K\left(P_i - \left|P_{2,i,t}^k\right|^2\right)}}{B_i^k}, \\ \Psi_{2,t}^k(\cdot) &= B^k\sqrt{\frac{\sum_{j=1}^M\left|h_{j,t}P_{2,j,t}^k\right|^2 + \sigma_n^2}{4T\max_{i\in\mathcal{M}}\left(B_i^k\right)^2\rho_i}}. \end{aligned}$$

In the following, we solve the above problem in two cases.

- Case I: if $\sigma_n^2 \geq 4TK\left(\min_{i\in\mathcal{M}}\left|h_{i,t}\right|^2 P_i\right)\max_{i\in\mathcal{M}}\rho_i$, the optimization problem (P2) becomes

$$\begin{aligned}\text{(P2.1)} \quad \min_{\left\{P_{2,i,t}^k, \lambda_t^k, \forall i,t,k\right\}} \quad & \sum_{i=1}^M\sum_{t=0}^{T-1} A_2\frac{\Phi_{2,i,t}}{T^{\frac{3}{2}}} \\ \text{s.t.} \quad & \lambda_t^k \leq \Psi_{1,t}^k(\cdot), \quad \forall t, k. \end{aligned} \quad (25)$$

By analyzing the objective function of (25), we see that

$$\sum_{i=1}^{M}\sum_{t=0}^{T-1} A_2\frac{\Phi_{2,i,t}}{T^{\frac{3}{2}}} = \sum_{i=1}^{M}\sum_{t=0}^{T-1}\frac{A_2\sum_{k=1}^{K}\frac{B_i^k}{B_i}\mathbb{E}\left[\left\|\hat{\mathbf{m}}_t^k\right\|_2^2\right]}{\left(\lambda_t^k\right)^2 T^{\frac{3}{2}}} = \sum_{i=1}^{M}\sum_{t=0}^{T-1} A_2\sum_{k=1}^{K}\frac{B_i^k}{B_i}K\left(\frac{\sigma_n^2}{\left(\lambda_t^k\right)^2 T^{\frac{3}{2}}} + \frac{\sum_{j=1}^{M}\left|h_{j,t}P_{2,j,t}^k\right|^2}{\left(\lambda_t^k\right)^2 T^{\frac{3}{2}}}\right). \tag{26}$$

The objective function in (26) is monotonically increasing with respect to $P_{2,i,t}^k, \forall i,t,k$. Therefore, the optimal transmit power associated with artificial noises is $P_{2,i,t}^{k^*}=0, \forall i,t,k$. With the optimal $\left\{P_{2,i,t}^k, \forall i,t,k\right\}$, the objective function in (P2.1) is monotonically decreasing with respect to $\lambda_t^k, \forall t,k$, which yields

$$\lambda_t^{k^*} = \min_{i\in\mathcal{M}}\frac{B^k\sqrt{K}\left|h_{i,t}\right|\sqrt{P_i}}{B_i^k}, \quad \forall t,k. \tag{27}$$

- Case II: if $\sigma_n^2 < 4TK\left(\min_{i\in\mathcal{M}}\left|h_{i,t}\right|^2 P_i\right)\max_{i\in\mathcal{M}}\rho_i$, the optimization problem (P2) becomes

$$\text{(P2.2)} \quad \min_{\left\{P_{2,i,t}^k,\lambda_t^k,\forall i,t,k\right\}} \quad \sum_{i=1}^{M}\sum_{t=0}^{T-1} A_2\frac{\Phi_{2,i,t}}{T^{\frac{3}{2}}} \quad \text{s.t.} \quad \lambda_t^k \le \Psi_{2,t}^k\left(\cdot\right), \quad \forall t,k.$$

Notice that

$$\frac{\sum_{j=1}^{M}\left|h_{j,t}P_{2,j,t}^k\right|^2+\sigma_n^2}{\left(\lambda_t^k\right)^2 T^{\frac{3}{2}}} \overset{(b)}{\ge} \frac{4T\left(\lambda_t^k\right)^2\max_{i\in\mathcal{M}}\left(B_i^k\right)^2\rho_i}{\left(B^k\lambda_t^k\right)^2 T^{\frac{3}{2}}},$$

To achieve the equality in (b), the optimal $\left\{\left\{P_{2,i,t}^k,\forall i\right\},\lambda_t^k,\forall t,k\right\}$ satisfies the following equation,

$$\sum_{j=1}^{M}\left|h_{j,t}P_{2,j,t}^{k^*}\right|^2+\sigma_n^2 = 4T\left(\lambda_t^{k^*}\right)^2\max_{i\in\mathcal{M}}\left(\frac{B_i^k}{B^k}\right)^2\rho_i, \ \forall t,k.$$

The proof of Proposition 1 is thus completed. □

Appendix D
Proof of Proposition 2

Proof. Given the optimal transceiver design per round, we analyze the long-term optimization problem for the training rounds decision T. According to the optimal transceiver design in Proposition 1, the i-th argument of the objective function $\Omega_i\left(T\right)$ with respect to T in Problem (P1) of the manuscript is given by

$$\Omega_i\left(T\right) = \begin{cases} \frac{3f_{i,max}}{\eta_0\sqrt{T}} + A_2\sum_{t=0}^{T-1}\sum_{k=1}^{K}\frac{K\sigma_n^2 B_i^k}{B_i\left(\lambda_t^{k^*}\right)^2 T^{\frac{3}{2}}}, & \text{if} \quad T\le T_0, \\ \frac{3f_{i,max}}{\eta_0\sqrt{T}} + A_2\sum_{k=1}^{K}\frac{4\sqrt{T}KB_i^k\max_{i\in\mathcal{M}}\left(B_i^k\right)^2\rho_i}{B_i\left(B^k\right)^2}, & \text{if} \quad T> T_0, \end{cases}$$

where $T_0 = \frac{\sigma_n^2}{4K\left(\min_{i\in\mathcal{M}}\left|h_{i,t}\right|^2 P_i\right)\max_{i\in\mathcal{M}}\rho_i}$. To find the optimal number of training rounds T^*, we first discuss the optimization problems for different regions of T.

- Case I: $T\le T_0$

The optimization problem in this case is given by

$$\text{(P3.1)} \quad \min_{T} \quad \sum_{i=1}^{M}\left(\frac{3f_{i,max}}{\eta_0\sqrt{T}} + A_2\sum_{t=0}^{T-1}\sum_{k=1}^{K}\frac{K\sigma_n^2 B_i^k}{B_i\left(\lambda_t^{k^*}\right)^2 T^{\frac{3}{2}}}\right) \quad \text{s.t.} \quad T\le T_0.$$

According to (27), we have

$$\sigma_n^2 \ge 4TK\left(\min_{i\in\mathcal{M}}\frac{|h_{i,t}|^2 P_i}{\left(B_i^k\right)^2}\right)\left(\max_{i\in\mathcal{M}}\left(B_i^k\right)^2\rho_i\right) = 4T\left(\lambda_t^{k^*}\right)^2\max_{i\in\mathcal{M}}\left(\frac{B_i^k}{B^k}\right)^2\rho_i. \tag{28}$$

By (28), we can see the objective function in (P3.1) is lower-bounded, i.e.,

$$\sum_{i=1}^{M}\Omega_i\left(T\right) \ge \sum_{i=1}^{M}\left(\frac{3f_{i,max}}{\eta_0\sqrt{T}} + A_2\sum_{k=1}^{K}\frac{4\sqrt{T}KB_i^k\max_{i\in\mathcal{M}}\left(B_i^k\right)^2\rho_i}{B_i\left(B^k\right)^2}\right). \tag{29}$$

The lower-bound in (29) is exactly the expression of the objective function when $T>T_0$. Therefore, the objective value associated with the optimal T obtained by solving (P3.1) is not smaller than that obtained by minimizing the objective function when $T>T_0$.

- Case II: $T>T_0$

The optimization problem in this case is given by

$$\text{(P3.2)} \quad \min_{T} \quad \frac{3\sum_{i=1}^{M}f_{i,max}}{\eta_0\sqrt{T}} + \sqrt{T}A_2\sum_{i=1}^{M}\sum_{k=1}^{K}\frac{4KB_i^k\max_{i\in\mathcal{M}}\left(B_i^k\right)^2\rho_i}{B_i\left(B^k\right)^2} \quad \text{s.t.} \quad T>T_0.$$

We can see that the objective function of (P3.2) is in the form of $\frac{a}{\sqrt{T}}+b\sqrt{T}$ with $a,b>0$. It is a convex function and the closed-form optimal solution of the non-negative and continuous $\hat{T}$ is given by

$$\hat{T}^* = \frac{a}{b} = \frac{3\sum_{i=1}^{M}f_{i,max}}{\eta_0 A_2\sum_{i=1}^{M}\sum_{k=1}^{K}\frac{4KB_i^k\max_{i\in\mathcal{M}}\left(B_i^k\right)^2\rho_i}{B_i\left(B^k\right)^2}} = \frac{3\sum_{i=1}^{M}f_{i,max}}{4\eta_0 A_2 M\max_{i\in\mathcal{M}}\rho_i\sum_{k=1}^{K}\left(\frac{B_i^k}{B^k}\right)^2}. \tag{30}$$

Then, we round the continuous $\hat{T}^*$ to the nearest integer and the proof of Proposition 2 is thus completed.

□

Appendix E
Additional experiment details

Dataset: All the methods in the manuscript are evaluated on the image classification task over the MNIST and CIFAR-10 database. MNIST database consists of 60000 training samples

and 10000 testing images from $K = 10$ classes. CIFAR-10 database consists of 60000 32×32 color images in $K = 10$ classes. There are 50000 training images and 10000 testing images. For the non-independent and identically distributed MNIST local dataset partitioning, we allocate each wireless device (WD) with 1110 training samples from one class and 10 samples from each of the other nine classes.

Models: The considered convolutional neural network (CNN) consists of two convolution layers with the max pooling and a ReLU activation function, two fully connected layers and a softmax layer. The number of model parameters altogether is $D = 21680$. For residual network (ResNet), it comprises of 8 residual blocks, a global average pooling layer, a 512×10 fully connected layer and finally a softmax layer, with a total of $D = 11173962$ parameters.

Privacy budgets: For the FD approaches in the ResNet-CIFAR10 setting, the differential privacy (DP) requirement ε_i is drawn from the uniform distribution from 0.01 to 1, while δ_i is from 10^{-9} to 10^{-7}. For FL benchmarks, ε_i is drawn from the uniform distribution from 10^6 to 10^7, while δ_i is from 10^{-3} to 10^{-2}.

Training details: In the ResNet-CIFAR10 setting, we adopt the FedAvg algorithm instead of FedSGD used in CNN-MNIST in both of the FD and FL approaches for a better training result. The local iteration for all approaches are set to $E = 5$.

REFERENCES

[1] X. Wang, S. Magnússon, and M. Johansson, "On the convergence of step decay step-size for stochastic optimization," *Advances in Neural Information Processing Systems,* vol. 34, pp. 14226-14238, 2021.

[2] K. Wei, J. Li, M. Ding, C. Ma, H. Su, B. Zhang and H. V. Poor, "User-Level Privacy-Preserving Federated Learning: Analysis and Performance Optimization," in *IEEE Transaction on Mobile Computing,* vol. 21, no. 9, pp. 3388-3401, 1 Sept. 2022.