Roadmap of Appendix: The Appendix is organized as follows. We present theoretical proof of the robustness of distributional changes in Section A, the proof of convergence in Section B. Additional experiment results are in Section C.

A. Proof of Value Difference Upper Bound

A.1. Preliminary

Given two distributions $\mathcal{D}_s$ and $\mathcal{D}_t$ over $\mathcal{Z}$, let Π_{st} denote the collection of joint distributions over $\mathcal{Z} \times \mathcal{Z}$. In particular, for all $\pi \in \Pi_{st}$, if iid draw $(s,t) \sim \pi$, then $s \sim \mathcal{D}_s$ and $t \sim \mathcal{D}_t$. Given a metric d over $\mathcal{Z}$, the Wasserstein distance is defined as the infimum over all such $\pi \in \Pi_{st}$ of the expected distance between $(s,t) \sim \pi$.

$$W_1\left(\mathcal{D}_s, \mathcal{D}_t\right) \triangleq \inf_{\pi \in \Pi_{st}} \mathop{\mathbb{E}}_{(s,t)\sim\pi}[d(s,t)]. \tag{7}$$

A.2. Assumptions and Proofs

First, we state the assumption of Lipschitz stable, which is derived from a standard notation of deletion stability, often studied in the context of generalization [61]. Following [43], we assume our potential function is $B(k)$-Lipschitz stable.

Assumption A.1 *Let $(\mathcal{Z}, d)$ be a metric space. For potential function Γ and non-increasing $\mathcal{B} : \mathbb{N} \to [0,1]$, Γ is $\mathcal{B}$-Lipschitz stable with respect to d if for all $k \in \mathbb{N}$, $\mathcal{S} \in \mathcal{Z}^{k-1}$, and all $z, z' \in \mathcal{Z}$,*

$$\left|\Gamma(S, \{z\}) - \Gamma\left(S, \{z'\}\right)\right| \leq \mathcal{B} \cdot d\left(z, z'\right). \tag{8}$$

For the convenience of notation, for any $z \in \mathcal{Z}$ and subset $S \subseteq \mathcal{Z}$, we denote $\Delta_z\Gamma(S) = \Gamma(S\backslash\{z\}, \{z\})$. Therefore, fixing $z \in \mathcal{Z}$, we can write $\widehat{\nu}(z; \Gamma, \mathcal{D}, N)$ as $\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^N}[\Delta_z\Gamma(S)]$. Let $\pi \in \Pi_{st}$ be some coupling of $\mathcal{D}_s$ and $\mathcal{D}_t$, we reformulate this expectation as:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}_s^N}\left[\Delta_z\Gamma(S)\right] &= \mathop{\mathbb{E}}_{S\times T\sim\pi^N}\left[\Delta_z\Gamma(S)\right] \\
&= \mathop{\mathbb{E}}_{S\times T\sim\pi^N}\left[\Delta_z\Gamma(S) - \Delta_z\Gamma(T)\right] \\
&\quad + \mathop{\mathbb{E}}_{S\times T\sim\pi^N}\left[\Delta_z\Gamma(T)\right] \\
&= \mathop{\mathbb{E}}_{S\times T\sim\pi^N}\left[\Delta_z\Gamma(S) - \Delta_z\Gamma(T)\right] \\
&\quad + \mathop{\mathbb{E}}_{T\sim\mathcal{D}_t^N}\left[\Delta_z\Gamma(T)\right],
\end{aligned} \tag{9}$$

where the first and last equation follow our definition that the marginals of π are $\mathcal{D}_s$ and $\mathcal{D}_t$, and the second equation follows by the linearity of expectation.

Then we bound the first term $[\Delta_z\Gamma(S) - \Delta_z\Gamma(T)]$. By expanding the difference between $\Delta_z\Gamma(S)$ and $\Delta_z\Gamma(T)$ into a telescoping sum of N pairs of terms, we bound each pair to depend on a single draw $(s_i, t_i) \sim \pi$. For $S, T \in \mathcal{Z}^N$, and $i \in \{0, \ldots, N\}$, denote $Z_i = \left(\bigcup_{j=i+1}^{N} s_j\right) \cup \left(\bigcup_{j=1}^{i} t_j\right)$, such that $\mathcal{Z}_0 = S$ and $\mathcal{Z}_N = T$. Then we can expand the first term as:

$$\Delta_z\Gamma(S) - \Delta_z\Gamma(T) = \sum_{i=1}^{N} \Delta_z\Gamma\left(Z_{i-1}\right) - \Delta_z\Gamma\left(Z_i\right). \tag{10}$$

Since we assume Γ is $\mathcal{B}$-Lipschitz stable, we can derive the following bound:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\times T\sim\pi^N}\left[\Delta_z\Gamma(S) - \Delta_z\Gamma(T)\right] \\
&\quad= \mathop{\mathbb{E}}_{S\times T\sim\pi^N}\left[\sum_{i=1}^{N} \Delta_z\Gamma\left(Z_{i-1}\right) - \Delta_z\Gamma\left(Z_i\right)\right] \\
&\quad= \sum_{i=1}^{N} \mathop{\mathbb{E}}_{S,T\sim\pi^N}\left[\Delta_z\Gamma\left(Z_{i-1}\right) - \Delta_z\Gamma\left(Z_i\right)\right] \\
&\quad= \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\substack{(s_i,t_i)\sim\pi \\ R\in\mathcal{Z}^{N-2}}}\left[\Delta_z\Gamma\left(R \cup \{s_i\}\right) - \Delta_z\Gamma\left(R \cup \{t_i\}\right)\right] \\
&\quad\leq 2\mathcal{B} \cdot \sum_{i=1}^{N} \mathop{\mathbb{E}}_{(s_i,t_i)\sim\pi}\left[d\left(s_i, t_i\right)\right] \\
&\quad\leq 2N\mathcal{B} \mathop{\mathbb{E}}_{(s,t)\sim\pi}[d(s,t)],
\end{aligned} \tag{11}$$

where the last two inequality follow the $\mathcal{B}$-Lipschitz assumption and the fact that each draw from π is iid. Finally, we re-write the differences in values in terms of the infimum over Π_{st} to complete the bound.

$$\begin{aligned}
&\widehat{\nu}\left(z; \Gamma, \mathcal{D}_s, N\right) - \widehat{\nu}\left(z; \Gamma, \mathcal{D}_t, N\right) \\
&\quad\leq \inf_{\pi\in\Pi_{st}}\left[\mathop{\mathbb{E}}_{S\times T\sim\pi^N}\left[\Delta_z\Gamma(S) - \Delta_z\Gamma(T)\right]\right] \\
&\quad\leq 2N\mathcal{B} \inf_{\pi\in\Pi_{st}} \mathop{\mathbb{E}}_{(s,t)\sim\pi}[d(s,t)] \\
&\quad= 2N\mathcal{B} \cdot W_1\left(\mathcal{D}_s, \mathcal{D}_t\right)
\end{aligned} \tag{12}$$

B. Proof of FedCE Convergence

B.1. Preliminary

We start by setting up the basic FL training and objective. Then we give the proof of our theorem.

Let $\boldsymbol{G}_{(k,i)}$ denotes the locally accumulated stochastic gradients scaled with a factor γ. For the local client gradients and global model update, we have the following rule:

$$\begin{cases}
\boldsymbol{G}_{(k,i)} \triangleq \frac{1}{\gamma_{(k,i)}} \sum_{\lambda=0}^{\kappa_{(k,i)}-1} \gamma_{(k,i)}^{\lambda} \nabla F_i\left(\boldsymbol{w}_{(k,i)}^{\lambda}\right) \\
\boldsymbol{w}_{k+1} - \boldsymbol{w}_k = -\eta \boldsymbol{d}_k,
\end{cases} \tag{13}$$

where $\boldsymbol{d}_k \triangleq \sum_{i=1}^{N} p_i \boldsymbol{G}_{(k,i)}$, and $\kappa_{(k,i)}$ denotes the local update iterations (steps) for the client i at the k-th round. $\gamma_{(k,i)}^{\lambda}$ denotes an arbitrary scalar, where $\boldsymbol{\gamma}_{(k,i)} = [\gamma_{k,i}^{0}, \cdots, \gamma_{k,i}^{\lambda}]$, $\gamma_{(k,i)} = \|\boldsymbol{\gamma}_{(k,i)}\|$, and we assume $\sum_{i=1}^{N} \frac{p_i}{\gamma_{(k,i)}\sqrt{\kappa_{(k,i)}}} \sum_{\lambda=0}^{\kappa_{(k,i)}-1} \gamma_{(k,i)}^{\lambda} = 1$ to make sure the summation of aggregation factors is 1 for each communication round. For the global direction, we denotes it as the global gradient $\nabla\|F(\boldsymbol{w}_k)\|$. In particular, the global gradient in FL is the weighted average of all training clients, i.e., $\nabla\|F(\boldsymbol{w}_k)\| \triangleq \sum_{i=1}^{N} p_i \nabla F_i(\boldsymbol{w}_k)$, where $\nabla F_i(\boldsymbol{w}_k)$ is local gradient of $\boldsymbol{w}_k$ calculated on all training data from client i. In FL, the learning objective is to find an optimal global model $\boldsymbol{w}_K^*$ by minimizing $F(\boldsymbol{w}_K^*)$, that is:

$$\boldsymbol{w}_K^* \triangleq \arg\min F\left(\boldsymbol{w}_k\right). \tag{14}$$

In other words, the loss value of $F(\boldsymbol{w}_k)$ should decrease as training goes (k increases). For the k-th round, we have the objective of:

$$\boldsymbol{w}_{k+1}^* \triangleq \arg\min \left\{F\left(\boldsymbol{w}_{k+1}^*\right) - F\left(\boldsymbol{w}_k\right)\right\}. \tag{15}$$

By comparing Eq.(13) and Eq.(15), we have $\|\boldsymbol{d}_k\| \leq \|\nabla F(\boldsymbol{w}_k)\|$.

B.2. Assumptions

We first state the assumptions on local function smoothness and bounded gradients, which are commonly adopted in optimization literature [45–49].

Assumption B.1 *Each local objective function is Lipschitz smooth, that is, for $k \in [0, K-1]$:*

$$\|\nabla F\left(\boldsymbol{w}_{k+1}\right) - \nabla F\left(\boldsymbol{w}_k\right)\| \leq L\left\|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\right\|$$

Assumption B.2 *For any local gradient $\nabla F_i(\boldsymbol{w}_{(k,i)}^{\lambda})$ and $\lambda \in [0, \tau_{(k,i)} - 1]$, there exists $\beta_{(k,i)} \geq 0$, such that,*

$$\left\|\nabla F_i\left(\boldsymbol{w}_k\right) - \nabla F_i\left(\boldsymbol{w}_{(k,i)}^{\lambda}\right)\right\| \leq \beta_{(k,i)}\left\|\boldsymbol{w}_k - \boldsymbol{w}_{(k,i)}^{\lambda}\right\|$$

Assumption B.3 *For all local gradients, $s \in [0, \lambda]$ and $\lambda \in [1, \kappa_{(k,i)} - 1]$, there exists constants $\delta_{(k,i)} \geq 0$, such that,*

$$\left\|\sum_{s=0}^{\lambda-1} \nabla F_i\left(\boldsymbol{w}_{(k,i)}^{s}\right)\right\|^2 \leq \delta_{(k,i)} \sum_{s=0}^{\lambda-1} \|\nabla F\left(\boldsymbol{w}_k\right)\|^2$$

B.3. Proof of the convergence theorem

In this part, we show how to derive the convergence theorem. First, we start with the differences between $\boldsymbol{w}_{k+1}$ and $\boldsymbol{w}_k$. Since the global gradient is Lipschitz smooth, we have:

$$\begin{aligned}
&F\left(\boldsymbol{w}_{k+1}\right) - F\left(\boldsymbol{w}_k\right) \\
\leq &\nabla F\left(\boldsymbol{w}_k\right)\left(\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\right) + \frac{L}{2}\left\|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\right\|^2 \\
= &-\eta\left\langle \nabla F\left(\boldsymbol{w}_k\right), \boldsymbol{d}_k\right\rangle + \frac{\eta^2 L}{2}\left\|\boldsymbol{d}_k\right\|^2.
\end{aligned} \tag{16}$$

The first inequality is from Lipschitz smooth assumption and the second equation is by inserting Eq.(13). Then we reformulate the inner product term into the following form:

$$\begin{aligned}
\left\langle \nabla F\left(\boldsymbol{w}_k\right), \boldsymbol{d}_k\right\rangle = &\frac{1}{2}\left\|\nabla F\left(\boldsymbol{w}_k\right)\right\|^2 + \frac{1}{2}\left\|\boldsymbol{d}_k\right\|^2 \\
&- \frac{1}{2}\|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2.
\end{aligned} \tag{17}$$

By substituting Eq.(17) into Eq.(16), the inequation can be formulated as:

$$\begin{aligned}
&F\left(\boldsymbol{w}_{k+1}\right) - F\left(\boldsymbol{w}_k\right) \\
\leq &-\frac{1}{2}\eta\left(\|\nabla F\left(\boldsymbol{w}_k\right)\|^2 + \|\boldsymbol{d}_k\|^2 - \|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2\right) \\
&+ \frac{\eta^2 L}{2}\left\|\boldsymbol{d}_k\right\|^2 \\
= &-\frac{1}{2}\eta\|\nabla F(\boldsymbol{w}_k)\|^2 + \frac{\eta(\eta L - 1)}{2}\|\boldsymbol{d}_k\|^2 \\
&+ \frac{\eta}{2}\|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2 \\
\leq &(\frac{\eta^2 L}{2} - \eta)\|\nabla F(\boldsymbol{w}_k)\|^2 + \frac{\eta}{2}\|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2,
\end{aligned} \tag{18}$$

when $\eta L - 1 \geq 0$. The last inequality is because $\|d_k\| \leq \|\nabla F(\boldsymbol{w}_k)\|$. Next, we present how to bound the term $\|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2$.

By the definition of $\boldsymbol{d}_k$, for $i \in [1, N]$ and $k \in [0, K-1]$, we have:

$$\begin{aligned}
\|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2 &= \left\|\nabla F\left(\boldsymbol{w}_k\right) - \sum_{i=1}^{N} p_i \boldsymbol{G}_{(k,i)}\right\|^2 \\
&= \left\|\sum_{i=1}^{N} p_i\left(\nabla F_i\left(\boldsymbol{w}_k\right) - \boldsymbol{G}_{(k,i)}\right)\right\|^2 \\
&\leq \sum_{i=1}^{N} p_i \left\|\nabla F_i\left(\boldsymbol{w}_k\right) - \boldsymbol{G}_{(k,i)}\right\|^2 \\
&= \sum_{i=1}^{N} p_i \left\|\nabla F_i\left(\boldsymbol{w}_k\right) - \frac{1}{\gamma_{(k,i)}}\sum_{\lambda=0}^{\kappa_{(k,i)}-1} \gamma_{(k,i)}^{\lambda} \nabla F_i\left(\boldsymbol{w}_{(k,i)}^{\lambda}\right)\right\|^2 \\
&= \sum_{i=1}^{N} p_i \left\|\sum_{\lambda=0}^{\kappa_{(k,i)}-1} \frac{\gamma_{k,i}^{\lambda}}{\gamma_{(k,i)}}\left(\nabla F_i\left(\boldsymbol{w}_k\right) - \nabla F_i\left(\boldsymbol{w}_{(k,i)}^{\lambda}\right)\right)\right\|^2 \\
&\leq \sum_{i=1}^{N} p_i \sum_{\lambda=0}^{\kappa_{(k,i)}-1} \frac{\gamma_{(k,i)}^{\lambda}}{\gamma_{(k,i)}}\left\|\nabla F_i\left(\boldsymbol{w}_k\right) - \nabla F_i\left(\boldsymbol{w}_{(k,i)}^{\lambda}\right)\right\|^2 \\
&\leq \sum_{i=1}^{N} p_i \sum_{\lambda=0}^{\kappa_{(k,i)}-1} \frac{\beta_{(k,i)}^2 \gamma_{(k,i)}^{\lambda}}{\gamma_{(k,i)}}\left\|\boldsymbol{w}_k - \boldsymbol{w}_{(k,i)}^{\lambda}\right\|^2,
\end{aligned} \tag{19}$$

where the first and second inequality uses Jensen's Inequality and the last inequality follows our assumption B.2. For

training in FL, when local iteration $\lambda = 0$, we have $\boldsymbol{w}_k = \boldsymbol{w}^{\lambda}_{(k,i)}$, this induces $\|\boldsymbol{w}_k - \boldsymbol{w}^{\lambda}_{(k,i)}\|^2 = 0$ in Eq.(19). So we consider the differences when $\lambda \geq 1$.

$$\begin{aligned}\left\|\boldsymbol{w}_k - \boldsymbol{w}^{\lambda}_{(k,i)}\right\|^2 &= \eta^2 \left\|\sum_{s=0}^{\lambda-1} \nabla F_i\left(\boldsymbol{w}^{s}_{(k,i)}\right)\right\|^2 \\ &\leq \eta^2 \delta_{(k,i)} \sum_{s=0}^{\lambda-1} \left\|\nabla F\left(\boldsymbol{w}_k\right)\right\|^2 \\ &= \eta^2 \delta_{(k,i)} \lambda \left\|\nabla F\left(\boldsymbol{w}_k\right)\right\|^2 .\end{aligned} \tag{20}$$

The inequality here follow our assumption B.3. By inserting this equation back to Eq.(19), we obtain:

$$\begin{aligned}&\|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2 \\ &\leq \sum_{i=1}^{N} p_i \sum_{\lambda=0}^{\kappa_{(k,i)}-1} \lambda \eta^2 \frac{\beta^2_{(k,i)} \delta_{(k,i)} \gamma^{\lambda}_{(k,i)}}{\|\boldsymbol{\gamma}_{(k,i)}\|} \|\nabla F(\boldsymbol{w}_k)\|^2 \\ &= \sum_{i=1}^{N} p_i \frac{\kappa_{k,i}(\kappa_{k,i}-1)}{2} \eta^2 \beta^2_{(k,i)} \delta_{(k,i)} \frac{\|\boldsymbol{\gamma}_{(k,i)}\|_1}{\|\boldsymbol{\gamma}_{(k,i)}\|} \|\nabla F(\boldsymbol{w}_k)\|^2.\end{aligned} \tag{21}$$

For the ease of notation, we define $\rho_{k,i} = \frac{\|\boldsymbol{\gamma}_{(k,i)}\|_1}{\|\boldsymbol{\gamma}_{(k,i)}\|\sqrt{\kappa_{(k,i)}}}$ and $A_{(k,i)} = \eta\sqrt{\kappa_{(k,i)}}(\kappa_{(k,i)}-1)\beta^2_{(k,i)}\delta_{(k,i)}$. Then we have:

$$\begin{aligned}&\|\nabla F(\boldsymbol{w}_k) - \boldsymbol{d}_k\|^2 \\ &\leq \frac{\eta}{2} \sum_{i=1}^{N} p_i \rho_{(k,i)} A_{(k,i)} \|\nabla F(\boldsymbol{w}_k)\|^2 \\ &= \frac{\eta}{2} \|\nabla F(\boldsymbol{w}_k)\|^2 \sum_{i=1}^{N} p_i \rho_{(k,i)} A_{(k,i)}.\end{aligned} \tag{22}$$

After obtaining the bound of the differences between server and normalized gradient, we are now ready to derive the final result. Substituting Eq.(22) into Eq.(18), we have:

$$\begin{aligned}&F\left(\boldsymbol{w}_{k+1}\right) - F\left(\boldsymbol{w}_k\right) \\ &\leq (\frac{\eta^2 L}{2} - \eta) \|\nabla F(\boldsymbol{w}_k)\|^2 \\ &\quad + \frac{\eta^2}{4} \|\nabla F(\boldsymbol{w}_k)\|^2 \sum_{i=1}^{N} p_i \rho_{(k,i)} A_{(k,i)} \\ &= \left(\frac{\eta}{4}(2\eta L + \sum_{i=1}^{N} p_i \rho_{(k,i)} \eta A_{(k,i)}) - \eta\right) \|\nabla F(\boldsymbol{w}_k)\|^2.\end{aligned} \tag{23}$$

B.4. Proof of the convergence corollary

Here we further analyze relations between convergence and our reweighting factors to present the effects of our methods. Recall that in Eq.(18), $\eta L > 1$. We also assume the summation of aggregation factors is 1. Therefore, we can construct an inequation as below:

$$\left(\eta \sum_{i=1}^{N} p_i A^{1/2}_{(k,i)} + \eta \sum_{i=1}^{N} p_i \rho_{(k,i)} L\right) \geq 1, \tag{24}$$

where $\eta \sum_{i=1}^{N} p_i A^{1/2}_{(k,i)}$ is always positive.

Next, to ensure the model converge in Theorem 3.3, we need $\left(\frac{\eta}{4}(2\eta L + \sum_{i=1}^{N} p_i \rho_{(k,i)} \eta A_{(k,i)}) - \eta\right) \leq 0$, that is, $\left(\frac{1}{4}(2\eta L + \sum_{i=1}^{N} p_i \rho_{(k,i)} \eta A_{(k,i)})\right) \leq 1$. By inserting Eq.(24), we have:

$$\begin{aligned}&\frac{\eta}{4}\left(2L + \sum_{i=1}^{N} p_i \rho_{(k,i)} A_{(k,i)}\right) \\ &= \frac{\eta}{4} \sum_{i=1}^{N} p_i \left(2L\rho_{(k,i)} + \rho_{(k,i)} A_{(k,i)}\right) \\ &\leq \left(\eta \sum_{i=1}^{N} p_i A^{1/2}_{(k,i)} + \eta \sum_{i=1}^{N} p_i \rho_{(k,i)} L\right) \\ &= \frac{\eta}{4} \sum_{i=1}^{N} p_i \left(4(A^{1/2}_{(k,i)} + \rho_{(k,i)} L)\right).\end{aligned} \tag{25}$$

To ensure this inequality always hold, we have:

$$\begin{aligned}2L\rho_{(k,i)} + \rho_{(k,i)} A_{(k,i)} &\leq 4(A^{1/2}_{(k,i)} + \rho_{(k,i)} L) \\ \rho_{(k,i)}(A_{(k,i)} - 2L) &\leq 4A^{1/2}_{(k,i)} \\ \rho_{(k,i)} &\leq \frac{4A^{1/2}_{(k,i)}}{(A_{(k,i)} - 2L)} \\ &\quad (\text{when } A_{(k,i)} - 2L > 0)\end{aligned} \tag{26}$$

We consider the convergence case when $A_{(k,i)}$ is dominant, then we have:

$$\rho_{(k,i)} \leq \frac{4A^{1/2}_{(k,i)}}{(A_{(k,i)} - 2L)} = \mathcal{O}(\frac{1}{\sqrt{A_{(k,i)}}}). \tag{27}$$

This indicates that the model converges when $\rho_{(k,i)}$ satisfy this condition. And we are able to minimize the upper bound of $\rho_{(k,i)}$ by increasing $A_{(k,i)}$.

Recall that $A_{(k,i)} = \eta\sqrt{\kappa_{(k,i)}}(\kappa_{(k,i)} - 1)\beta^2_{(k,i)}\delta_{(k,i)}$. The items η and $\kappa_{(k,i)}$ are related with experimental settings. It is easy to understand that, if the training data are iid, increasing the learning rate η or performing more local iterations $\kappa_{(k,i)}$ improves the convergence. For non-iid data, the convergence is also affected by data distribution. If we increase learning rate or local step, the model convergence speed may be improved at an early stage. However, it may let the model trap into a local optimum or suffer large client drifts when data are heterogeneous [48].

Next we focus on terms of $\beta^2_{(k,i)}$ and $\delta_{(k,i)}$, which are related to our assumptions on the local gradients and parameters. According to Eq.(20), we have:

$$\begin{aligned}\beta^2_{(k,i)} &\geq \frac{\left\|\nabla F_i\left(\boldsymbol{w}_k\right)-\nabla F_i\left(\boldsymbol{w}^{\lambda}_{(k,i)}\right)\right\|^2}{\left\|\boldsymbol{w}_k-\boldsymbol{w}^{\lambda}_{(k,i)}\right\|^2} \\ &\geq \frac{\left\|\nabla F_i\left(\boldsymbol{w}_k\right)-\nabla F_i\left(\boldsymbol{w}^{\lambda}_{(k,i)}\right)\right\|^2}{\eta^2\delta_{(k,i)}\lambda\left\|\nabla F\left(\boldsymbol{w}_k\right)\right\|^2},\end{aligned} \tag{28}$$

which is also related to $\delta_{(k,i)}$. So we focus on discussing the relations between $\delta_{(k,i)}$ and convergence. From the Assumption B.3, we have:

$$\delta_{(k,i)} \geq \frac{\left\|\sum_{s=0}^{\lambda-1}\nabla F_i\left(\boldsymbol{w}^{s}_{(k,i)}\right)\right\|^2}{\sum_{s=0}^{\lambda-1}\left\|\nabla F\left(\boldsymbol{w}_k\right)\right\|^2}. \tag{29}$$

This term quantifies the percentage of local gradients over the global(aggregated) gradients. That is, to increase $\delta_{(k,i)}$, we need to weigh more on local gradients from client i. Since the client with boundary data or different distribution is under-represented during training, which harms the overall convergence. We need to assign higher weights to promote training on this kind of client, thus improving convergence. This well matches our contribution estimation method, i.e., allocating higher weight to clients presenting different information in gradient space or suffering high error on local data when their gradient is excluded.

C. Additional Experimental Results

In this section, we present more results of our method, including the free rider detection, discussion on client contributions, and visual comparison of segmentation results.

Free rider detection. We first present more results for the free rider detection using the prostate dataset. As discussed in the experiment section, we have combined the local-global gradients cosine similarity and local-global model error difference to detect the free rider client. Here we further present the results by using calculating the cosine similarity between local and global gradients, as shown in Fig. 7. From the figure can be observed that the similarity between local gradients from the free rider and global clients decreases lower with training goes on. The free rider client can be distinguished within 50 rounds. Interestingly, we observe that client 6 presents a high cosine similarity, except itself is the free rider. This is because client 6 has more samples than other clients, and the gradients dominate others during the aggregation. Therefore, it is critical to combine both gradients and performance, which well matches our motivation for method design.

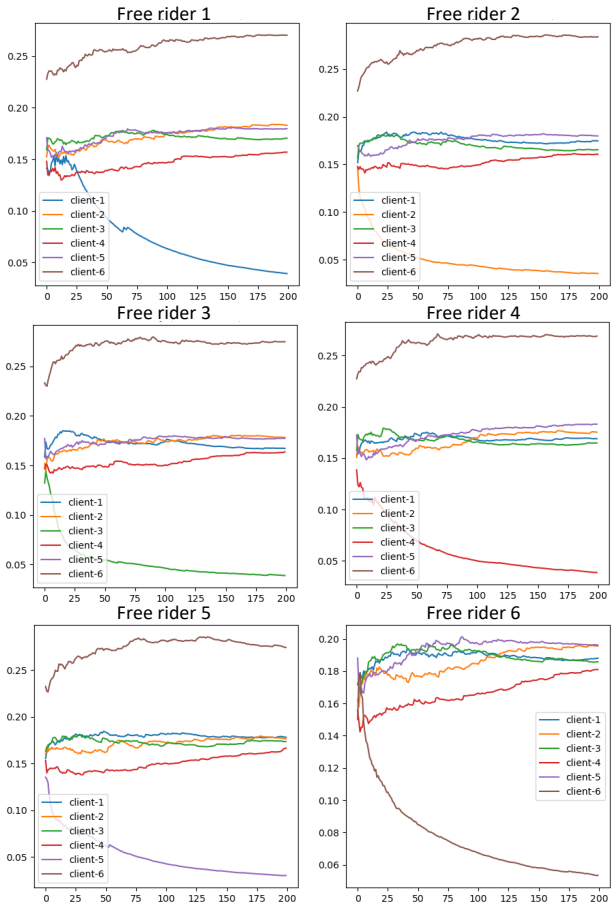

Figure 7. Free rider study by using cosine similarity between local and global gradients. X-axis denotes the communication rounds and y-axis denotes the similarity.

Client contribution quantification. We propose to quantify the client contribution by using the leave-one-out experiment, which is a popular and reliable way for data valuation [38]. It assesses how much performance we will lose if we remove a certain client. However, it would be too computationally expensive to perform in practice. We hereby calculate the leave-one-out results as a reference to quantify client contribution in the context of performance. Specifically, we run six independent federated training by removing client $i \in \{1, \ldots, 6\}$ to calculate the performance drop. Then we obtain the performance contribution by calculating the proportion of drop, i.e., a larger performance drop indicates this client has a larger performance contribution. Furthermore, in standard federated averaging algorithm [44], the sample proportion is typically used to indicate the importance (e.g., aggregation weight) of clients. So we calculate the sample contribution based on training samples. The results are shown in Table 4 and 5. From the two tables can be observed that, because the medical data collected from different sources are heterogeneous, the sample contribu-

Table 4. Client contribution quantification on the retinal fundus dataset by using performance drop with regard to leave-one-out experiments and using training sample proportions.

Client	1		2		3		4		5		6		No	
Metric	Disc	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc	Cup
Dice	86.84	74.06	88.26	74.21	88.46	73.25	87.41	73.58	82.52	70.66	89.43	74.05	89.43	75.50
Δ Dice	-2.59	-1.44	-1.17	-1.29	-0.97	-2.25	-2.02	-1.92	-6.91	-4.84	0.00	-1.45	-	
Performance Contribution	15.00%		9.50%		12.00%		15.00%		44.00%		5.50%		-	
Training Samples	50		98		47		230		80		400		-	
Sample Contribution	5.52%		10.83%		5.19%		25.41%		8.84%		44.20%		-	

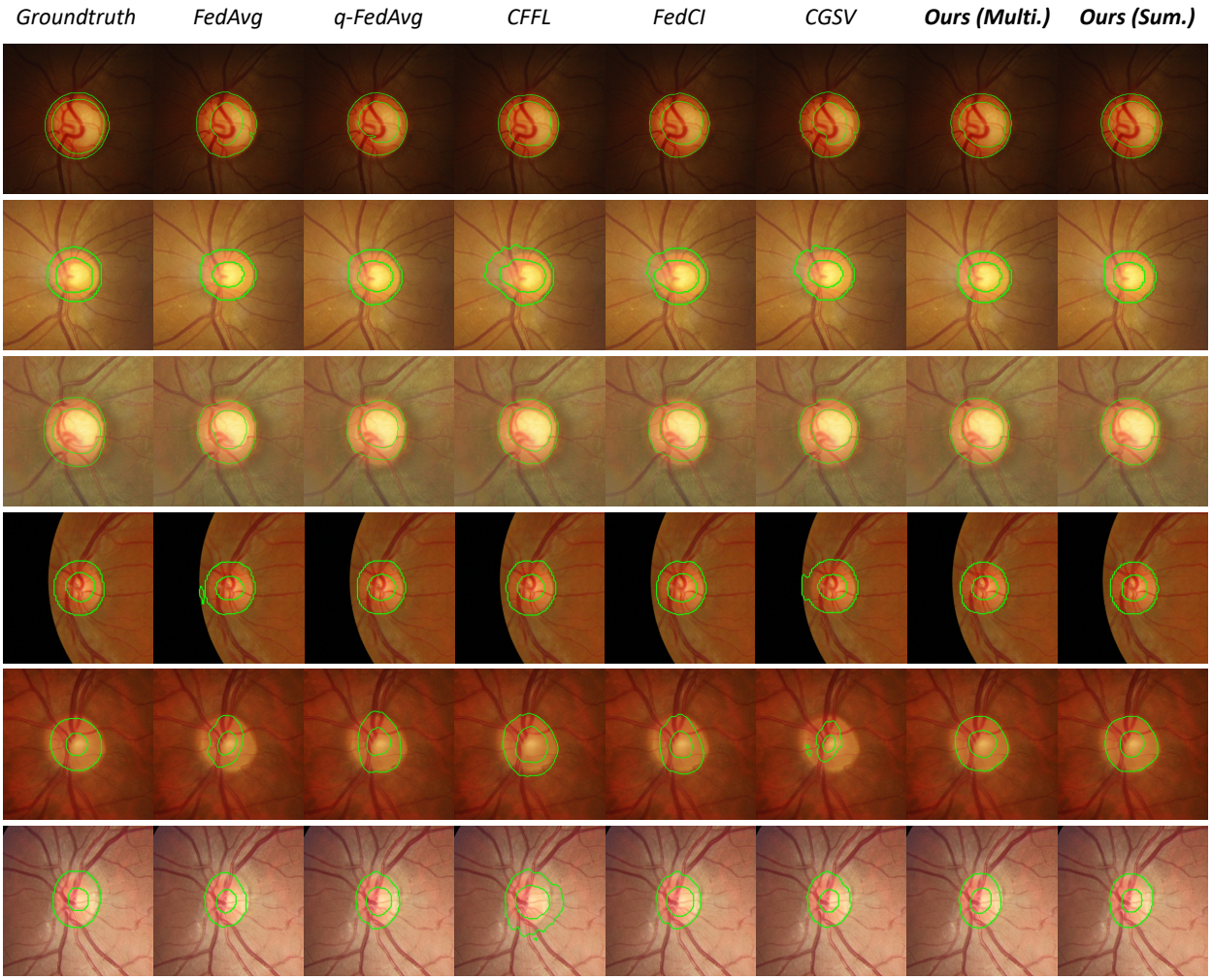

Figure 8. Qualitative comparison on the results of optic disc/cup segmentation from retinal fundus images. Each row denotes a client.

tion does not strongly correlate with performance contribution, that is, more samples from one client may not improve the overall global performance a lot. This may be because some other clients with similar data distribution play a complementary role. For example, client 6 in the retinal dataset has over 40% sample contribution, but the performance contribution is 5.5% by the leave-one-out results. Therefore, solely considering the sample number is not enough if we aim to have a global model robust to various data distributions. In our experiments, we have presented how to promote collaboration fairness by considering the client contribution, which is reflected by client performance improvements. For the final client reward or credit allocation, it is a comprehensive procedure that needs to cover multiple dif-

Table 5. Client contribution quantification on the prostate dataset by using performance drop with regard to leave-one-out experiments and using training sample proportions.

Client	1	2	3	4	5	6	No
Dice	84.90	87.95	87.91	87.97	76.67	87.53	88.32
Δ Dice	-3.43	-0.37	-0.42	-0.36	-11.65	-0.80	-
Performance Contribution	20.13%	2.19%	1.74%	2.09%	68.47%	4.68%	-
Training Samples	381	238	278	242	389	814	-
Sample Contribution	16.27%	10.16%	11.87%	10.33%	16.61%	34.76%	-

Table 6. Performance comparison using Dice score on image segmentation datasets of retinal fundus images and prostate MRI.

Task	Retinal Fundus Segmentation								Prostate MRI Segmentation							
Client	1	2	3	4	5	6	Avg.	Std.	1	2	3	4	5	6	Avg.	Std.
Standalone	86.69 ±0.32	85.51 ±1.41	86.21 ±0.69	89.91 ±0.15	79.77 ±1.59	90.98 ±0.06	86.51	3.95	91.23 ±0.40	84.59 ±0.55	87.57 ±0.86	87.37 ±0.32	86.70 ±0.05	89.25 ±0.13	87.79	2.26
FedAvg	81.34 ±3.08	85.21 ±0.15	83.28 ±1.60	88.16 ±0.45	40.81 ±6.57	90.79 ±0.46	78.27	18.66	91.10 ±0.10	84.59 ±0.44	89.02 ±0.37	89.09 ±0.75	83.87 ±0.42	**89.27** ±0.10	87.82	2.90
q-FedAvg	86.24 ±0.80	86.97 ±0.20	87.37 ±0.66	89.13 ±0.40	44.68 ±3.42	90.72 ±0.15	80.85	17.80	90.94 ±0.25	85.60 ±0.56	89.28 ±0.37	89.18 ±0.85	84.27 ±0.32	88.67 ±0.09	87.99	2.52
CFFL	85.72 ±2.17	86.29 ±1.32	86.96 ±0.58	88.62 ±1.95	41.12 ±2.35	90.16 ±0.95	79.81	19.02	91.01 ±0.67	85.49 ±0.72	89.24 ±0.39	88.98 ±0.86	82.11 ±2.20	88.17 ±0.41	87.50	3.20
FedCI	87.02 ±1.47	86.93 ±0.41	87.35 ±0.40	88.53 ±0.39	40.99 ±7.94	90.22 ±0.14	80.17	19.24	91.21 ±0.68	85.40 ±0.74	89.49 ±0.57	88.37 ±0.94	83.96 ±0.49	88.49 ±0.28	87.82	2.68
CGSV	83.46 ±1.53	85.57 ±0.15	85.47 ±0.79	88.48 ±0.71	33.79 ±2.59	**91.01** ±0.65	77.96	21.80	91.15 ±0.38	84.90 ±0.66	89.27 ±0.35	88.09 ±0.93	83.47 ±0.34	89.16 ±0.25	87.67	2.91
FedCE (Multi.)	86.73 ±1.46	**87.45** ±0.14	87.51 ±0.57	89.26 ±0.32	**57.30** ±1.32	90.25 ±0.15	**83.08**	**12.70**	**91.43** ±0.33	**85.79** ±0.55	89.21 ±0.46	89.13 ±0.59	**85.68** ±0.29	88.62 ±0.10	**88.31**	**2.22**
FedCE (Sum.)	**87.22** ±0.61	87.36 ±0.60	**87.93** ±0.56	**89.66** ±0.29	54.42 ±1.84	90.92 ±0.28	82.92	14.03	91.18 ±0.30	85.54 ±0.19	**89.59** ±0.33	**89.22** ±0.82	84.99 ±0.44	88.79 ±0.05	88.22	2.43

ferent aspects, including our studies performance, as well as more factors like the computing cost, annotation cost, data quality, etc. The study on final client rewards or monetary allocation is still an open and important question that needs to be further investigated.

Distribution shifts on two datasets In this work, we consider two types of data heterogeneity sources to cover real medical scenarios. First is feature space shift from different imaging devices/protocols and variations during the imaging process, etc. In our scenario, prostate MRI data is captured by different machines and imaging protocols, and fundus image varies with different machines, illumination conditions, field of views, etc. The retinal dataset is "less homogeneous" than the prostate dataset because of more variations in color space and field of view. Besides the feature shift, we also consider an additional special case shift, reflected by the retinal data: one of the clients has a different image setting (dual) from others (mono). This may not apply to most medical applications, hence is a "less homogeneous" data than most modalities.

Complete results with three random seeds We present the complete experiment results by reporting the mean and standard deviation of three independent runs in Table. 6. Notably, in the retinal fundus segmentation task, other compared methods exhibit a large standard deviation for the special client 5, while our method is more stable. Overall, our method yields stable results, demonstrating its reliability.

Visualization of segmentation results. We further present more qualitative segmentation results comparison on both retinal fundus dataset and prostate MRI dataset, as shown in Fig. 8 and Fig. 9. In two figures, each row denotes one sample from a specific client, and each column denotes one method. We can see the samples visually looks different, showing the data heterogeneity of medical images collected from different hospitals/sources. Compared with alternative methods, which may present a less smooth boundary or cover more or less region, our methods (i.e., the multiplication and summation versions defined in Eq. 4.) present a more complete segmentation results with more accurate boundary and segmented region.

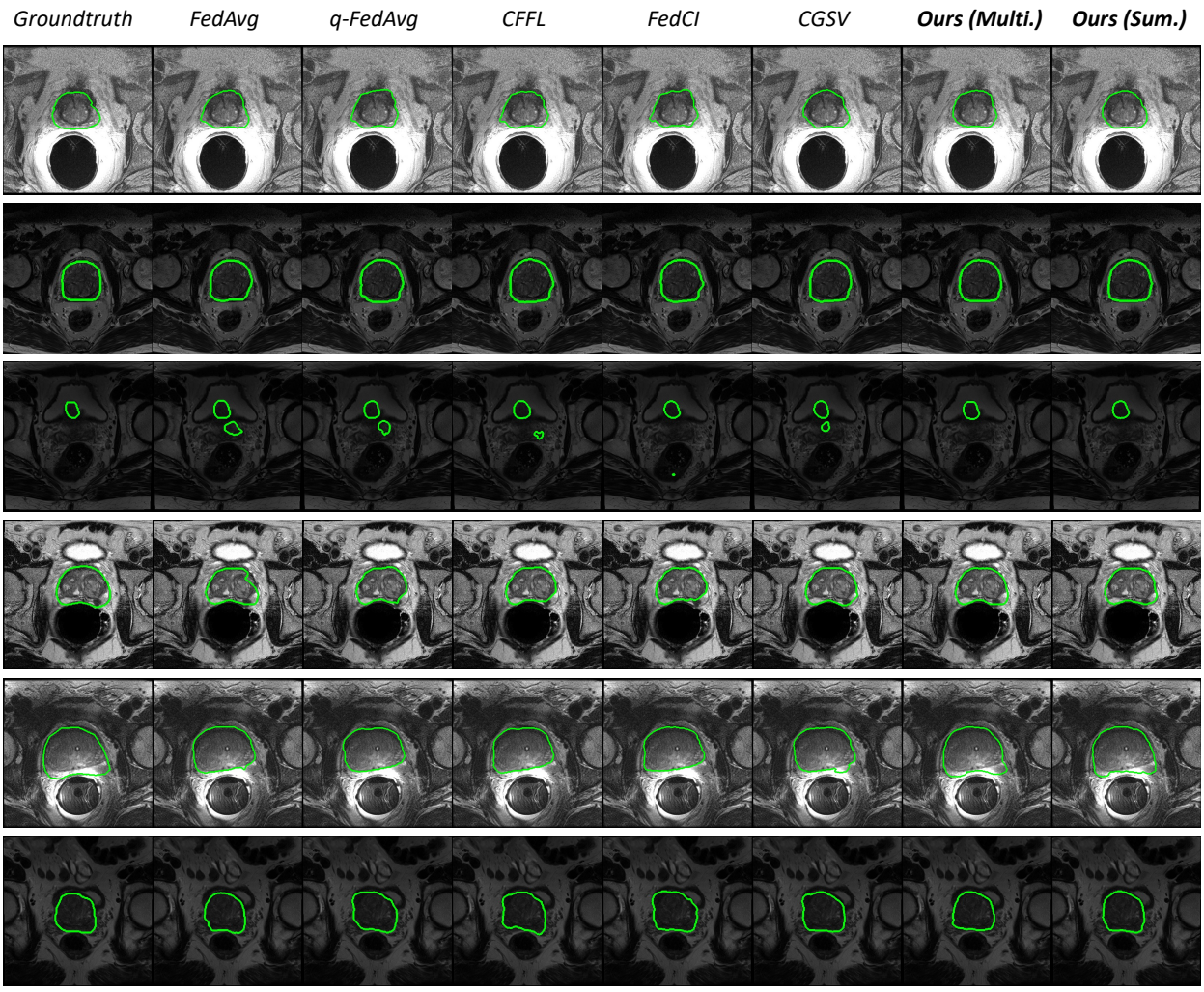

Figure 9. Qualitative comparison on the results of prostate segmentation. Each row denotes a client.