A. Theoretical Analysis

A.1. Analytical Insight into Gaussian Processes

In this section, we give a mathematical explanation about why the loss changes obey Gaussian Distributions. Our analysis based on the following assumption where we assume that the global weight update in one communication round follow a Gaussian Distribution under uniformly client selection.

Assumption 1. *In any communication round t, if the client selection $\mathbb{K}_t$ is a random variable sampled from a uniform distribution, the global model update $\Delta\mathbf{w}^t(\mathbb{K}_t) = \mathbf{w}^{t+1}(\mathbb{K}_t) - \boldsymbol{w}^t$ follows Gaussian Distribution, i.e.,*

$$\begin{aligned} &\mathbb{K}_t \sim \text{Uniform}\big(\{\mathbb{K} \subseteq \mathbb{U} : |\mathbb{K}| = C\}\big) \\ \Rightarrow &\Delta\mathbf{w}^t(\mathbb{K}_t) \sim \mathcal{N}(\Delta\boldsymbol{w}^t; -\eta_t\tilde{\boldsymbol{g}}^t, \frac{\eta_t^2\boldsymbol{B}\boldsymbol{B}^T}{C}), \end{aligned} \tag{18}$$

where $\tilde{\boldsymbol{g}}^t = \mathbb{E}_{\mathbb{k}}[\tilde{\nabla} l_{\mathbb{k}}(\boldsymbol{w}^t)]$ is the mean cumulative gradient of all the clients in $\mathbb{U}$, and $\boldsymbol{B}$ is a constant matrix.

Assumption 1 is inspired by [22] who assumes the stochastic gradients in SGD are Gaussian, and therefore the parameter update after one iteration follows a Gaussian Distribution. Note that in the FL procedure, the form in Eq. 5 is very similar to that in the SGD update. The only difference is that the average gradients within one mini-batch is replaced by the average cumulative gradients of the selected clients. Therefore, it is reasonable to make this assumption similar to [22].

To make a distinction, we use $\Delta\mathbf{w}$ without parentheses to denote a random variable w.r.t. the uniformly sampled client selection, and use $\Delta\boldsymbol{w}(\mathbb{K})$ to denote a determinate value without randomness where the client selection $\mathbb{K}$ is determined. The rule for $\Delta\mathbf{l}$ and $\Delta\boldsymbol{l}(\mathbb{K})$ in the following contents is the same.

Based on this assumption, we can easily show that the loss changes in each communication round follow a Gaussian Process under first-order approximation, with the property of Gaussian Distribution.

Corollary 1. *In any communication round t, $\forall\mathbb{S} = \{i_1, \cdots, i_{|\mathbb{S}|}\} \subseteq \mathbb{U}$, the loss changes $\Delta\mathbf{l}_{\mathbb{S}}^t = [\Delta\mathbf{l}_{i_1}^t, \cdots, \Delta\mathbf{l}_{i_{|\mathbb{S}|}}^t]^T$ follow a Multivariate Gaussian Distribution (or a Gaussian Process) under first-order approximation, i.e.,*

$$\begin{aligned} &\Delta\mathbf{l}_{\mathbb{S}}^t \sim \mathcal{N}(\Delta\boldsymbol{l}_{\mathbb{S}}^t; \boldsymbol{\mu}_{\mathbb{S}}^t, \boldsymbol{\Sigma}_{\mathbb{S}}^t), \\ &\textit{where} \\ &\boldsymbol{\mu}_{\mathbb{S}}^t = -\eta_t {\boldsymbol{G}_{\mathbb{S}}^t}^T \tilde{\boldsymbol{g}}^t; \\ &\boldsymbol{\Sigma}_{\mathbb{S}}^t = \frac{\eta_t^2}{C}{\boldsymbol{G}_{\mathbb{S}}^t}^T \boldsymbol{B}\boldsymbol{B}^T \boldsymbol{G}_{\mathbb{S}}^t; \\ &\boldsymbol{G}_{\mathbb{S}}^t = \left[\nabla l_{i_1}(\boldsymbol{w}^t), \cdots, \nabla l_{i_{|\mathbb{S}|}}(\boldsymbol{w}^t)\right]. \end{aligned} \tag{19}$$

We remove the subscript $\mathbb{S}$ to simplify the corresponding representation for the client set $\mathbb{U}$ as

$$\Delta\mathbf{l}^t \sim \mathcal{N}(\Delta\boldsymbol{l}^t; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t), \tag{20}$$

which is exactly the result in Eq. 8. And we can also obtain a mathematical reason from Eq. 19 for our choice of homogeneous linear kernel in Section 4.5, where $\boldsymbol{X}^t = \boldsymbol{B}^T\boldsymbol{G}^t$.

Remark Although an uniformly sampled client selection is required in Assumption 1 to get the loss changes to follow a GP prior, it is not necessary for the final selection to be uniformly sampled since we are predicting its loss changes with the GP posterior conditioned on the selected clients. We can view each posterior during the iterative selection process in Section 4.3 as the distribution of the loss changes w.r.t. the client selection that consists of two parts: (i) fixed selected clients in the previous iteration and (ii) uniformly sampled clients from the rest of the clients.

A.2. Proof of Lemma 1

To prove Lemma 1, we first introduce another assumption.

Assumption 2. *In any communication round t, for any client selection $\mathbb{K}$, we have*

$$Pr(\mathbb{K}|\Delta\boldsymbol{l}_{\mathbb{K}}^t(\mathbb{K})) \approx 1. \tag{21}$$

This assumption asserts that for any client selection $\mathbb{K}$, there is unlikely another client selection other than $\mathbb{K}$ which can produce the same loss changes on $\mathbb{K}$, i.e.,

$$\begin{aligned} &\forall\mathbb{K}', \mathbb{K} \subseteq \mathbb{U}, |\mathbb{K}'| = |\mathbb{K}| \\ \Rightarrow &Pr(\Delta\boldsymbol{l}_{\mathbb{K}}^t(\mathbb{K}') = \Delta\boldsymbol{l}_{\mathbb{K}}^t(\mathbb{K})|\mathbb{K}' \neq \mathbb{K}) \approx 0. \end{aligned} \tag{22}$$

We anticipate that this is a realistic assumption because of the heterogeneity between clients and the highly complexity of the neural network. When selecting different clients, the data used for training varies a lot under heterogeneous federated learning settings. This fact makes it almost impossible to produce the same neural network, and thus the same loss changes, with two different client selections. Furthermore, the selected clients usually have larger loss decreases than other clients who are not selected, because the model update is based on the mean cumulative gradient of these selected clients. The other client selection is unlikely to generate the same large loss decreases on all of them.

With Assumption 2, we can get the following corollary 2.

Corollary 2. *In any communication round t, for any client selection $\mathbb{K}$, we have*

$$Pr(\Delta\boldsymbol{l}^t(\mathbb{K})|\Delta\boldsymbol{l}_{\mathbb{K}}^t(\mathbb{K})) \approx 1. \tag{23}$$

Proof. When client selection $\mathbb{K}$ is given, we get the determinate model update $\Delta \boldsymbol{w}^t(\mathbb{K})$, thus the loss changes are known without randomness. In the other word,

$$Pr(\Delta \boldsymbol{l}^t(\mathbb{K})|\mathbb{K}) = 1 \tag{24}$$

always holds. Besides, we can extend the condition in Eq. 21 to the loss changes of all the clients and get

$$Pr(\mathbb{K}|\Delta \boldsymbol{l}^t(\mathbb{K})) \approx 1. \tag{25}$$

Combining Eq. 21, Eq. 24 and Eq. 25, we have

$$\begin{aligned} Pr(\Delta \boldsymbol{l}^t(\mathbb{K})) \approx & Pr(\Delta \boldsymbol{l}^t(\mathbb{K}), \mathbb{K}) && (26)\\ = & Pr(\mathbb{K}) && (27)\\ = & Pr(\Delta \boldsymbol{l}^t_{\mathbb{K}}(\mathbb{K}), \mathbb{K}) && (28)\\ \approx & Pr(\Delta \boldsymbol{l}^t_{\mathbb{K}}(\mathbb{K})) && (29) \end{aligned}$$

By substituting Eq. 29 into the expression of $Pr(\Delta \boldsymbol{l}^t(\mathbb{K})|\Delta \boldsymbol{l}^t_{\mathbb{K}}(\mathbb{K}))$, we get

$$\begin{aligned} Pr(\Delta \boldsymbol{l}^t(\mathbb{K})|\Delta \boldsymbol{l}^t_{\mathbb{K}}(\mathbb{K})) = & \frac{Pr(\Delta \boldsymbol{l}^t(\mathbb{K}), \Delta \boldsymbol{l}^t_{\mathbb{K}}(\mathbb{K}))}{Pr(\Delta \boldsymbol{l}^t_{\mathbb{K}}(\mathbb{K}))} && (30)\\ = & \frac{Pr(\Delta \boldsymbol{l}^t(\mathbb{K}))}{Pr(\Delta \boldsymbol{l}^t_{\mathbb{K}}(\mathbb{K}))} && (31)\\ \approx & 1. && (32) \end{aligned}$$

□

Now we are ready to prove Lemma 1.

Lemma 1. *The optimization problem in Eq. (6) is approximately equivalent to the following probabilistic form.*

$$\min_{\mathbb{K}_t} \quad \mathbb{E}_{\Delta \mathrm{l}^t|\Delta \boldsymbol{l}^t_{\mathbb{K}_t}(\mathbb{K}_t)} \Big[\sum_i p_i \Delta l_i^t \Big] = \sum_i p_i \tilde{\mu}_i^t(\Delta \boldsymbol{l}^t_{\mathbb{K}_t}(\mathbb{K}_t)), \tag{33}$$

where $\Delta \mathrm{l}^t = [\Delta \mathrm{l}_1^t, \cdots, \Delta \mathrm{l}_N^t]$ *is the loss changes of all clients in round* t*, which is a random variable w.r.t random client selection in round* t*.* $\tilde{\boldsymbol{\mu}}^t(\Delta \boldsymbol{l}^t_{\mathbb{K}_t}(\mathbb{K}_t))$ *is the posterior mean of* $\Delta \mathrm{l}^t$ *conditioned on* $\Delta \boldsymbol{l}^t_{\mathbb{K}_t}(\mathbb{K}_t) = [\Delta l_i^t(\mathbb{K}_t)]_{i \in \mathbb{K}_t}$*.*

Proof. According to Corollary 2, we can transform the optimization problem in Eq. 6 into the form in Eq. 33.

$$\begin{aligned} & \min_{\mathbb{K}_t} && \Delta L^t(\mathbb{K}_t) && (34)\\ = & \min_{\mathbb{K}_t} && \sum_i p_i \Delta l_i^t(\mathbb{K}_t) && (35)\\ \approx & \min_{\mathbb{K}_t} && Pr(\Delta \boldsymbol{l}^t(\mathbb{K}_t)|\Delta \boldsymbol{l}^t_{\mathbb{K}_t}(\mathbb{K}_t)) \sum_i p_i \Delta l_i^t(\mathbb{K}_t) && (36)\\ \approx & \min_{\mathbb{K}_t} && \mathbb{E}_{\Delta \mathrm{l}^t|\Delta \boldsymbol{l}^t_{\mathbb{K}_t}(\mathbb{K}_t)} \Big[\sum_i p_i \Delta l_i^t \Big] && (37)\\ = & \min_{\mathbb{K}_t} && \sum_i p_i \tilde{\mu}_i^t(\Delta \boldsymbol{l}^t_{\mathbb{K}_t}(\mathbb{K}_t)). && (38) \end{aligned}$$

□

A.3. Proof of Lemma 2

Lemma 2. *The selection criterion of FedCor when selecting two clients* k_1 *and* k_2 *can be written as*

$$k_1 = \arg\max_k \quad \beta^{\tau_k} \sum_i p_i \sigma_i r_{ik}, \tag{39}$$

$$k_2 = \arg\max_{k'} \quad \frac{\beta^{\tau_{k'}} \Big[\sum_i p_i \sigma_i r_{ik'} - r_{k_1 k'} \sum_i p_i \sigma_i r_{ik_1} \Big]}{\sqrt{1 - r_{k'k_1}^2}}, \tag{40}$$

where $r_{ij} = \Sigma_{i,j} / \sigma_i \sigma_j$ *is the Pearson correlation coefficient.*

Proof. We first deduce Eq. 39 for the first client k_1. By substituting the loss change estimation $\Delta \hat{l}_k$ from Eq. 9 into the criterion in Eq. 10, we can calculate the weighted sum of the posterior mean as

$$\begin{aligned} & \sum_i p_i \tilde{\mu}_i(\Delta \hat{l}_k) && (41)\\ = & \sum_i p_i \mu_i + \sum_i p_i \frac{\Sigma_{i,k}}{\sigma_k^2} (\Delta \hat{l}_k - \mu_k) && (42)\\ = & \sum_i p_i \mu_i - a \beta^{\tau_k} \sum_i p_i \sigma_i r_{ik}, && (43) \end{aligned}$$

where r_{ik} is the Pearson correlation coefficient. The first item in Eq. 43 and the factor a are constant for all k, thus the selection strategy becomes

$$k_1 = \arg\max_k \quad \beta^{\tau_k} \sum_i p_i \sigma_i r_{ik}, \tag{44}$$

which is Eq. 39.

Then we deduce Eq. 40 for selecting k_2. We can calculate the posterior covariance conditioned on $\Delta \hat{l}_{k_1}$ as

$$\begin{aligned} \tilde{\Sigma}_{i,j}(\Delta \hat{l}_{k_1}) = & \Sigma_{i,j} - \frac{\Sigma_{i,k_1} \Sigma_{k_1,j}}{\sigma_{k_1}^2} && (45)\\ = & \sigma_i \sigma_j (r_{ij} - r_{ik_1} r_{k_1 j}) && (46)\\ \tilde{\sigma}_i(\Delta \hat{l}_{k_1}) = & \sqrt{\tilde{\Sigma}_{i,i}(\Delta \hat{l}_{k_1})} && (47)\\ = & \sigma_i \sqrt{1 - r_{ik_1}^2}. && (48) \end{aligned}$$

We substitute the posterior covariance into the simplified selection criterion in Eq. 44 and get

$$\begin{aligned} & \beta^{\tau_{k'}} \sum_i p_i \frac{\tilde{\Sigma}_{i,k'}(\Delta \hat{l}_{k_1})}{\tilde{\sigma}_{k'}(\Delta \hat{l}_{k_1})} \\ = & \frac{\beta^{\tau_{k'}} \Big[\sum_i p_i \sigma_i r_{ik'} - r_{k_1 k'} \sum_i p_i \sigma_i r_{ik_1} \Big]}{\sqrt{1 - r_{k'k_1}^2}}. \end{aligned} \tag{49}$$

So we have Eq. 40:

$$k_2 = \arg\max_{k'} \frac{\beta^{\tau_{k'}}\left[\sum_i p_i\sigma_i r_{ik'} - r_{k_1k'}\sum_i p_i\sigma_i r_{ik_1}\right]}{\sqrt{1-r_{k'k_1}^2}}. \tag{50}$$

□

B. Selection Criterion and Convergence Analysis

In this section, we will analyse FedCor when selecting arbitrary number of clients. While the iterative client selection makes it obscure to analyse the convergence, we will show that we can construct a simpler proxy algorithm who can approximate the selection strategy of FedCor and there for share similar convergence characteristic. We will prove the convergence of this proxy algorithm.

B.1. Definitions

We first introduce some important definitions. In the following analysis, We denote the client selection sampled from FedCor as $\mathbb{K}_t \sim \pi$ and client selection sampled uniformly as $\mathbb{K}_t \sim \mathcal{U}$.

In the j-th iteration of FedCor, we select a client k_j to minimize the posterior mean of the loss change. Since the prior mean in each iteration is fixed, we can say that we are maximizing the decrease from prior mean $\boldsymbol{\mu}^{t,j}$ to posterior mean $\tilde{\boldsymbol{\mu}}^{t,j}$. We define the posterior gain of this iteration as the decrease from prior mean to posterior mean, namely,

$$g^{t,j}(k_j) = \sum_i p_i(\mu_i^{t,j} - \tilde{\mu}_i^{t,j}(\Delta\hat{l}_{k_j}^t)) \tag{51}$$

$$= \alpha_{k_j}^t \sum_i p_i\sigma_i^{t,j} r_{ik_j}^{t,j}. \tag{52}$$

We define $\boldsymbol{\mu}^{t,1} = \boldsymbol{\mu}^t$ and $\boldsymbol{\Sigma}^{t,1} = \boldsymbol{\Sigma}^t$. And for $j > 1$ we have

$$\boldsymbol{\mu}^{t,j} = \tilde{\boldsymbol{\mu}}^{t,j-1}(\Delta\hat{l}_{k_{j-1}}^t), \qquad \boldsymbol{\Sigma}^{t,j} = \tilde{\boldsymbol{\Sigma}}^{t,j-1}(\Delta\hat{l}_{k_{j-1}}^t). \tag{53}$$

With Lemma 2, we get

$$g^{t,j}(k_j) = \frac{g^{t,j-1}(k_j) - \frac{\alpha_{k_j}^t}{\alpha_{k_{j-1}}^t} r_{k_{j-1}k_j}^{t,j-1} g^{t,j-1}(k_{j-1})}{\sqrt{1 - {r_{k_{j-1}k_j}^{t,j-1}}^2}}. \tag{54}$$

With this notation, we can simplify our selection strategy as follows.

$$k_j^* = \arg\max_{k_j} g^{t,j}(k_j). \tag{55}$$

We further define the one-round advantage of FedCor compared with uniform sampling as follows.

$$A^t = \mathbb{E}_{\mathbb{K}_t\sim\mathcal{U}}[L(\boldsymbol{w}^{t+1}) - L(\boldsymbol{w}^t)] - \mathbb{E}_{\mathbb{K}_t\sim\pi}[L(\boldsymbol{w}^{t+1}) - L(\boldsymbol{w}^t)] \tag{56}$$

$$= \sum_{j=1}^{C} g^{t,j}(k_j^*). \tag{57}$$

The second equation directly arises from the definition of our prior distribution where $\mathbb{E}_{\mathbb{K}_t\sim\mathcal{U}}[L(\boldsymbol{w}^{t+1}) - L(\boldsymbol{w}^t)] = \sum_i \mu_i^t$.

Unfortunately, because of the iterative selection, the selection criterion of k_j depends on the previous selected clients, which makes a quantitatively analysis complicated. To bypass this difficulty, we will first point out that A^t has a lower bound that is tight in some special cases. We find that a proxy client selection strategy that maximizes this lower bound has a similar but simpler behaviour compared with FedCor, and we will also give a convergence guarantee of the proxy algorithm.

B.2. Approximation of FedCor

An important property of FedCor is that it prefers clients who have lower correlations with those selected in the previous iteration, since

$$\forall r_{k_{j-1}k_j}^{t,j-1} \in (-1,1), \frac{\partial g^{t,j}(k_j)}{\partial r_{k_{j-1}k_j}^{t,j-1}} < 0. \tag{58}$$

We further predict that FedCor tends to select clients that with $r_{k_{j-1}k_j}^{t,j-1}$ close to 0 instead of $r_{k_{j-1}k_j}^{t,j-1} < 0$ because if $r_{k_{j-1}k_j}^{t,j-1} < 0$, k_j should be far away from k_{j-1} who is closed to other clients in the embedding space, which makes k_j has low correlation with the other clients and not be selected. Therefore, we can infer that FedCor will select a group of clients who have nearly zero correlations with each other, which simplifies the expression of $g^{t,j}(k_j)$ to $g^{t,1}(k_j)$.

Based on the analysis above, we define a proxy algorithm $\tilde{\pi}$ who maximize the following objective.

$$\tilde{A}^t = \sum_{k_j\in\mathbb{K}_t}^{C} g^{t,1}(k_j) \approx \sum_{k\in\mathbb{K}_t}\sum_i p_i\Sigma_{i,k}^t, \tag{59}$$

where we further omit the difference of α_k^t and σ_k^t for different client k. We can use the client selection generated by this proxy algorithm to approximate the client selection of FedCor, and thus they share similar convergence characteristic.

In the following section, we will show that this proxy algorithm has a good property that enable it to converge to the optimal solution of the global loss L without gap, even it is a biased selection strategy.

B.3. Convergence Analysis of the Proxy Algorithm

In the following section, we denote the client selection sampled from the proxy client selection strategy as $\mathbb{K}_t \sim \tilde{\pi}$. We use $\mathbb{E}[\cdot]$ as the expectation over the mini-batch and $\mathbb{E}_{\mathbb{K}_t}[\cdot]$ as the expectation over the client selection strategy. We first give the common assumptions used in Federated Learning [4, 19].

Assumption 3. *$l_1, l_2, \cdots, l_N$ are all M-smooth: for all $\boldsymbol{v}$ and $\boldsymbol{w}$, $l_k(\boldsymbol{v}) \leq l_k(\boldsymbol{w}) + (\boldsymbol{v}-\boldsymbol{w})^T \nabla l_k(\boldsymbol{w}) + \frac{M}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2$.*

Assumption 4. *$l_1, l_2, \cdots, l_N$ are all m-strongly convex: for all $\boldsymbol{v}$ and $\boldsymbol{w}$, $l_k(\boldsymbol{v}) \geq l_k(\boldsymbol{w}) + (\boldsymbol{v}-\boldsymbol{w})^T \nabla l_k(\boldsymbol{w}) + \frac{m}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2$.*

Assumption 5. *For the mini-batch $\xi_k \in \mathbb{D}_k$ sampled uniformly on each client $k \in \mathbb{U}$, the variance of stochastic gradients is bounded: $\mathbb{E}\|\nabla l_k(\boldsymbol{w}_k, \xi_k) - \nabla l_k(\boldsymbol{w}_k)\|^2 \leq s_k^2$.*

Assumption 6. *For each client $k \in \mathbb{U}$ and any communication round t, the expected squared norm of stochastic gradients is uniformly bounded: $\mathbb{E}\|\nabla l_k(\boldsymbol{w}_k, \xi_k)\|^2 \leq G^2$.*

For concision, we omit $\mathbb{E}$ in the following content and apply an expectation over the mini-batch by default.

Now we give an important property of the proxy algorithm that will be used for proving the convergence.

Lemma 3. *In any communication round t, with Assumption 1 and Assumption 2 holds, we have*

$$\mathbb{K}_t \sim \tilde{\pi} = \arg\max_{\mathbb{K}} (\boldsymbol{B}^T \nabla L(\boldsymbol{w}^t))^T \sum_{k\in\mathbb{K}} \boldsymbol{B}^T \nabla l_k(\boldsymbol{w}^t). \tag{60}$$

Proof. In the proxy algorithm, we have

$$\mathbb{K}_t = \arg\max_{\mathbb{K}} \sum_{k\in\mathbb{K}_t} \sum_i p_i \Sigma_{i,k}^t \tag{61}$$

$$= \arg\max_{\mathbb{K}} \frac{\eta_t^2}{C} \sum_{k\in\mathbb{K}} \sum_i p_i \nabla l_i(\boldsymbol{w}^t) \boldsymbol{B}\boldsymbol{B}^T \nabla l_k(\boldsymbol{w}^t) \tag{62}$$

$$= \arg\max_{\mathbb{K}} (\boldsymbol{B}^T \nabla L(\boldsymbol{w}^t))^T \sum_{k\in\mathbb{K}} \boldsymbol{B}^T \nabla l_k(\boldsymbol{w}^t). \tag{63}$$

Eq. 62 comes from the expression of $\boldsymbol{\Sigma}^t$ in Corollary 1, and Eq. (63) arises from $L(\boldsymbol{w}^t) = \sum_i p_i l_i(\boldsymbol{w}^t)$. □

To connect this property with the convergence of the algorithm, we first define a sequence and show that the convergence of this sequence is equivalent to the convergence of the algorithm with this property. We define Sequence Δ_t as follows.

$$\Delta_t = \mathbb{E}_{\mathbb{K}_t \sim \tilde{\pi}} \|\boldsymbol{w}^t - \boldsymbol{w}_{\mathbb{K}_t}^*\|^2, \tag{64}$$

where

$$\boldsymbol{w}_{\mathbb{K}_t}^* = \arg\min_{\boldsymbol{w}} \sum_{k\in\mathbb{K}_t} l_k(\boldsymbol{w}). \tag{65}$$

We now show that if $\Delta_t \to 0$, we have $\boldsymbol{w} \to \boldsymbol{w}^*$.

Corollary 3. ***(Optimal Solution Consistency)*** *If Δ_t converges to 0, there must be $\boldsymbol{w}^t$ converges to $\boldsymbol{w}^*$.*

$$\lim_{t\to\infty} \Delta_t = 0 \Rightarrow \lim_{t\to\infty} \boldsymbol{w}^t = \boldsymbol{w}^* \tag{66}$$

Proof. With $\mathbb{K}_t \sim \tilde{\pi}$, we have

$$\lim_{t\to\infty} \Delta_t = 0 \tag{67}$$

$$\Rightarrow \lim_{t\to\infty} \boldsymbol{w}^t = \boldsymbol{w}_{\mathbb{K}_t}^* \tag{68}$$

$$\Rightarrow \lim_{t\to\infty} \sum_{k\in\mathbb{K}_t} \nabla l_k(\boldsymbol{w}^t) = \boldsymbol{0} \tag{69}$$

$$\Rightarrow \lim_{t\to\infty} (\boldsymbol{B}^T \nabla L(\boldsymbol{w}^t))^T \sum_{k\in\mathbb{K}_t} \boldsymbol{B}^T \nabla l_k(\boldsymbol{w}^t) = 0. \tag{70}$$

Since

$$\mathbb{K}_t = \arg\max_{\mathbb{K}} (\boldsymbol{B}^T \nabla L(\boldsymbol{w}^t))^T \sum_{k\in\mathbb{K}} \boldsymbol{B}^T \nabla l_k(\boldsymbol{w}^t), \tag{71}$$

If $\lim_{t\to\infty} \boldsymbol{B}^T \nabla L(\boldsymbol{w}^t) \neq \boldsymbol{0}$ or does not converge, we can say that

$$\forall \epsilon > 0, \exists \tau, \forall t > \tau, \forall \mathbb{K}, \tag{72}$$

$$(\boldsymbol{B}^T \nabla L(\boldsymbol{w}^t))^T \sum_{k\in\mathbb{K}} \boldsymbol{B}^T \nabla l_k(\boldsymbol{w}^t) \leq \epsilon, \tag{73}$$

which cannot be true since

$$\mathbb{E}_{\mathbb{K}\sim\mathcal{U}} \sum_{k\in\mathbb{K}} \nabla l_k(\boldsymbol{w}^t) = C \nabla L(\boldsymbol{w}^t). \tag{74}$$

Thus we conclude that

$$\lim_{t\to\infty} \boldsymbol{B}^T \nabla L(\boldsymbol{w}^t) = 0. \tag{75}$$

If the Gaussian Distribution in Assumption 1 is non-degenerate, we have

$$\lim_{t\to\infty} \nabla L(\boldsymbol{w}^t) = \boldsymbol{0} \Rightarrow \lim_{t\to\infty} \boldsymbol{w}^t = \boldsymbol{w}^* \tag{76}$$

□

We now only need to prove the convergence of Δ_t, which will imply the convergence of the proxy algorithm according to Corollary 3. We first introduce one extra assumption as well as two lemmas that will be used in the proof.

For convenient, we define $L_{\mathbb{K}_t}(\boldsymbol{w}) = \frac{1}{C}\sum_{k\in\mathbb{K}_t} l_k(\boldsymbol{w})$, and thus $\boldsymbol{w}_{\mathbb{K}_t}^* = \arg\min_{\boldsymbol{w}} L_{\mathbb{K}_t}(\boldsymbol{w})$. Notice that $\mathbb{K}_t \sim \tilde{\pi}$ only depends on $\boldsymbol{\Sigma}^t$, thus we can say that $\boldsymbol{w}_{\mathbb{K}_t}^*$ is given by a function of $\boldsymbol{\Sigma}^t$, i.e., $\boldsymbol{w}_{\mathbb{K}_t}^* = \Omega(\boldsymbol{\Sigma}^t)$. We further assume the smoothness of Ω:

Assumption 7. *For any t, $\mathbb{E}\|\boldsymbol{w}_{\mathbb{K}_{t+1}}^* - \boldsymbol{w}_{\mathbb{K}_t}^*\|^2 = \mathbb{E}\|\Omega(\boldsymbol{\Sigma}^{t+1}) - \Omega(\boldsymbol{\Sigma}^t)\|^2 \leq \delta \mathbb{E}\|\boldsymbol{\Sigma}^{t+1} - \boldsymbol{\Sigma}^t\|_1$, where $\|\cdot\|_1$ is the ℓ_1 norm of a vector.*

Now we introduce a lemma that bounds $\mathbb{E}\|\mathbf{\Sigma}^{t+1} - \mathbf{\Sigma}^t\|_1$.

Lemma 4. *Assume Assumption 1, Assumption 3 and Assumption 6, if $\mathbb{E}\|\boldsymbol{w}^t - \boldsymbol{w}^{t+1}\|^2 \leq q_t^2$, we have*

$$\mathbb{E}\|\mathbf{\Sigma}_{t+1} - \mathbf{\Sigma}_t\|_1 \leq \frac{bN^2}{C}[\eta_t^2(G + Mq_t)^2 - \eta_{t+1}^2 G^2], \tag{77}$$

where b is the largest eigenvalue of $\boldsymbol{B}\boldsymbol{B}^T$.

Proof. According to Assumption 1, we have

$$\Sigma_{i,j} = \frac{\eta_t^2}{C} \nabla l_i^{t\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^t. \tag{78}$$

And we can calculate

$$\begin{aligned}
&\left|\Sigma_{i,j}^{t+1} - \Sigma_{i,j}^t\right| && (79)\\
=&\Big|\frac{\eta_t^2}{C}(\nabla l_i^{t+1\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^{t+1} - \nabla l_i^{t\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^t) + \\
&\frac{\eta_{t+1}^2 - \eta_t^2}{C} \nabla l_i^{t+1\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^{t+1}\Big| && (80)\\
\leq&\frac{\eta_t^2}{C}\left|\nabla l_i^{t+1\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^{t+1} - \nabla l_i^{t\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^t\right| + \\
&\frac{\eta_t^2 - \eta_{t+1}^2}{C}\left|\nabla l_i^{t+1\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^{t+1}\right|. && (81)
\end{aligned}$$

We now bound each term in Eq. (81) separately. For the first term,

$$\begin{aligned}
&\left|\nabla l_i^{t+1\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^{t+1} - \nabla l_i^{t\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^t\right| && (82)\\
=&\Big|(\nabla l_i^{t+1} - \nabla l_i^t)^T \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^t + \\
&\nabla l_i^{t\,T} \boldsymbol{B}\boldsymbol{B}^T (\nabla l_j^{t+1} - \nabla l_j^t) + \\
&(\nabla l_i^{t+1} - \nabla l_i^t)^T \boldsymbol{B}\boldsymbol{B}^T (\nabla l_j^{t+1} - \nabla l_j^t)\Big| && (83)\\
\leq&b\Big(\|\nabla l_i^{t+1} - \nabla l_i^t\|\|\nabla l_j^t\| + \\
&\|\nabla l_j^{t+1} - \nabla l_j^t\|\|\nabla l_i^t\| + \\
&\|\nabla l_i^{t+1} - \nabla l_i^t\|\|\nabla l_j^{t+1} - \nabla l_j^t\|\Big) && (84)\\
\leq&b\Big[M\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|(\|\nabla l_j^t\| + \|\nabla l_i^t\|) + \\
&M^2\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2\Big], && (85)
\end{aligned}$$

where b is the largest eigenvalue of $\boldsymbol{B}\boldsymbol{B}^T$. For the second term,

$$\left|\nabla l_i^{t+1\,T} \boldsymbol{B}\boldsymbol{B}^T \nabla l_j^{t+1}\right| \leq b\|\nabla l_i^{t+1}\|\|\nabla l_j^{t+1}\|. \tag{86}$$

We take the expectation over both sides and with Cauchy-Schwarz inequality, we get

$$\begin{aligned}
&\mathbb{E}\left|\Sigma_{i,j}^{t+1} - \Sigma_{i,j}^t\right| && (87)\\
\leq&\frac{\eta_t^2}{C} b\Big[M\sqrt{\mathbb{E}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \mathbb{E}\|\nabla l_j^t\|^2} + \\
&M\sqrt{\mathbb{E}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \mathbb{E}\|\nabla l_i^t\|^2} + \\
&M^2\mathbb{E}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2\Big] + \\
&\frac{\eta_t^2 - \eta_{t+1}^2}{C} b\sqrt{\mathbb{E}\|\nabla l_i^{t+1}\|^2 \mathbb{E}\|\nabla l_j^{t+1}\|^2} && (88)\\
\leq&\frac{\eta_t^2 b}{C}(G^2 + 2Mq_tG + M^2q_t^2) - \frac{\eta_{t+1}^2}{C} bG^2 && (89)
\end{aligned}$$

And we have

$$\begin{aligned}
&\mathbb{E}\|\mathbf{\Sigma}_{t+1} - \mathbf{\Sigma}_t\|_1 && (90)\\
=&\sum_{i,j}^{N} \mathbb{E}\left|\Sigma_{i,j}^{t+1} - \Sigma_{i,j}^t\right| && (91)\\
\leq&\frac{bN^2}{C}[\eta_t^2(G + Mq_t)^2 - \eta_{t+1}^2 G^2] && (92)
\end{aligned}$$

□

We will also use the following lemma that is proved by [19].

Lemma 5. *Assume Assumption 3 to 6. If $\eta_t \leq \frac{1}{4M}$, with full and balanced participation in FedAvg, in any communication round t and its i-th iteration, we have*

$$\mathbb{E}\|\bar{\boldsymbol{w}}^{t,i+1} - \boldsymbol{w}^*\|^2 \leq (1 - \eta_t m)\mathbb{E}\|\bar{\boldsymbol{w}}^{t,i} - \boldsymbol{w}^*\|^2 + \eta_t^2 F, \tag{93}$$

where

$$F = \frac{1}{N}\sum_{k=1}^{N} s_k^2 + 6M\Gamma + 8(E-1)^2G^2, \tag{94}$$

$$\Gamma = L^* - \frac{1}{N}\sum_{k=1}^{N} l_k^*. \tag{95}$$

Here, $\bar{\boldsymbol{w}}^{t,i} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{w}_k^{t,i}$, and $\boldsymbol{w}_k^{t,i}$ is the local weight at the i-th iteration of communication round t. E is the total number of local training iterations. $L^* = L(\omega^*)$ and $l_k^* = l_k(\omega_k^*)$ are the optimal value of L and l_k, respectively.

Now we give the theorem of the convergence of Δ_t and prove it.

Theorem 1. *With Assumption 1 to 7 holds, with learning rate $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{m}$ and $\gamma > 0$ such that $\eta_1 \leq \min\{\frac{1}{m}, \frac{1}{4M}\} = \frac{1}{4M}$, we have*

$$\Delta_t \leq \frac{\nu}{\gamma + t}, \tag{96}$$

where

$$\nu = \max\{\frac{\beta^2(\tilde{F}+\tilde{D})}{\beta m - 1}, (\gamma+1)\Delta_1\}, \tag{97}$$

$$\tilde{F} = 2E \max_t F_t, \tag{98}$$

$$F_t = \frac{1}{C}\sum_{k\in\mathbb{K}_t} s_k^2 + 6M\Gamma_t + 8(E-1)^2G^2, \tag{99}$$

$$\Gamma_t = L^*_{\mathbb{K}_t} - \frac{1}{C}\sum_{k\in\mathbb{K}_t} l_k^*, \tag{100}$$

$$\tilde{D} = (\frac{1}{m} + \frac{1}{4M})\delta D, \tag{101}$$

$$D = \frac{bN^2}{C}(2mG^2 + 2MEG + \frac{1}{4}ME^2G^2). \tag{102}$$

Proof. For $\mathbb{K}_t \sim \tilde{\pi}(\boldsymbol{w}^t)$ and $\mathbb{K}_{t+1} \sim \tilde{\pi}(\boldsymbol{w}^{t+1})$, we have

$$\begin{aligned}
\Delta_{t+1} =& \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*_{\mathbb{K}_{t+1}}\|^2 && (103)\\
=& \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \|\boldsymbol{w}^*_{\mathbb{K}_t} - \boldsymbol{w}^*_{\mathbb{K}_{t+1}}\|^2 + \\
& 2\langle \boldsymbol{w}^{t+1} - \boldsymbol{w}^*_{\mathbb{K}_t}, \boldsymbol{w}^*_{\mathbb{K}_t} - \boldsymbol{w}^*_{\mathbb{K}_{t+1}}\rangle && (104)\\
\leq& \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \|\boldsymbol{w}^*_{\mathbb{K}_t} - \boldsymbol{w}^*_{\mathbb{K}_{t+1}}\|^2 + \\
& \eta_t m\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \\
& \frac{1}{\eta_t m}\|\boldsymbol{w}^*_{\mathbb{K}_t} - \boldsymbol{w}^*_{\mathbb{K}_{t+1}}\|^2 && (105)\\
\leq& (1+\eta_t m)\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \\
& (1+\frac{1}{\eta_t m})\delta\|\boldsymbol{\Sigma}^{t+1} - \boldsymbol{\Sigma}^t\|_1, && (106)
\end{aligned}$$

where Eq. 105 arises from AM-GM inequality and Eq. 106 arises from Assumption 7.

For the first term in Eq. 106, we can bound it by Lemma 5 as follows. The key point here is that when training in one communication round t, we can view this round a small FL process with clients in $\mathbb{K}_t$ fully participating. In this view, the global loss and the optimal global weight becomes $L_{\mathbb{K}_t}$ and $\boldsymbol{w}^*_{\mathbb{K}_t}$ instead. Thus we can apply Lemma 5 directly to bound $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2$. With $\eta_t \leq \frac{1}{4M} \leq \frac{1}{m}$, we have $\eta_t m \leq 1$ and $1+\eta_t m \leq \frac{1}{1-\eta_t m}$, and we can get

$$\begin{aligned}
& (1+\eta_t m)\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 && (107)\\
=& (1+\eta_t m)[\|\bar{\boldsymbol{w}}^{t,E} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2] && (108)\\
\leq& (1+\eta_t m)[(1-\eta_t m)\|\bar{\boldsymbol{w}}^{t,E-1} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \eta_t^2 F_t] && (109)\\
\leq& (1+\eta_t m)\{(1-\eta_t m)^2\|\bar{\boldsymbol{w}}^{t,E-2} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \\
& [1+(1-\eta_t m)]\eta_t^2 F_t\} && (110)\\
& \cdots \\
\leq& (1+\eta_t m)\Big\{(1-\eta_t m)^E\|\bar{\boldsymbol{w}}^{t,0} - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \\
& [1+(1-\eta_t m)+\cdots+(1-\eta_t m)^{E-1}]\eta_t^2 F_t\Big\} && (111)\\
\leq& (1-\eta_t m)^{E-1}\|\boldsymbol{w}^t - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + \\
& (1+\eta_t m)\frac{1-(1-\eta_t m)^E}{m}\eta_t F_t && (112)\\
\leq& (1-\eta_t m)\|\boldsymbol{w}^t - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + (1+\eta_t m)E\eta_t^2 F_t && (113)\\
\leq& (1-\eta_t m)\|\boldsymbol{w}^t - \boldsymbol{w}^*_{\mathbb{K}_t}\|^2 + 2E\eta_t^2 F_t, && (114)
\end{aligned}$$

where

$$F_t = \frac{1}{C}\sum_{k\in\mathbb{K}_t} s_k^2 + 6M\Gamma_t + 8(E-1)^2G^2, \tag{115}$$

$$\Gamma_t = L^*_{\mathbb{K}_t} - \frac{1}{C}\sum_{k\in\mathbb{K}_t} l_k^*. \tag{116}$$

Eq. 114 arises from the inequality $1-Ex \leq (1-x)^E \leq 1-x$ for $x \in [0,1]$.

We now turn to bound the second term in Eq. 106. We first find the q_t in Lemma 4.

$$\begin{aligned}
\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 =& \|\frac{1}{C}\sum_{k\in\mathbb{K}_t}\boldsymbol{w}_k^{t,E} - \boldsymbol{w}^t\|^2 && (117)\\
\leq& \frac{1}{C}\sum_{k\in\mathbb{K}_t}\|\boldsymbol{w}_k^{t,E} - \boldsymbol{w}^t\|^2 && (118)\\
=& \frac{\eta_t^2}{C}\sum_{k\in\mathbb{K}_t}\|\sum_{i=0}^{E-1}\nabla l_k(\boldsymbol{w}_k^{t,i})\|^2 && (119)\\
\leq& \frac{\eta_t^2 E}{C}\sum_{k\in\mathbb{K}_t}\sum_{i=0}^{E-1}\|\nabla l_k(\boldsymbol{w}_k^{t,i})\|^2 && (120)\\
\leq& \frac{\eta_t^2 E}{C}\sum_{k\in\mathbb{K}_t}\sum_{i=0}^{E-1} G^2 && (121)\\
\leq& \frac{\eta_t^2 E}{C}\sum_{k\in\mathbb{K}_t} EG^2 && (122)\\
=& \eta_t^2 E^2 G^2 = q_t^2, && (123)
\end{aligned}$$

where Eq. 118 and Eq. 120 comes from Jensen inequality, and Eq. 121 comes from Assumption 6. With Lemma

Lemma 4, we get

$$\|\mathbf{\Sigma}^{t+1}-\mathbf{\Sigma}^{t}\|_1 \leq \frac{bN^2}{C}[G^2(\eta_t^2-\eta_{t+1}^2)+ \tag{124}$$

$$2MEG\eta_t^3+M^2E^2G^2\eta_t^4)]. \tag{125}$$

Further with a diminishing $\eta_t=\frac{\beta}{t+\gamma}$, we have

$$\eta_t^2-\eta_{t+1}^2 =\beta^2(\frac{1}{(t+\gamma)^2}-\frac{1}{(t+1+\gamma)^2}) \tag{126}$$

$$=\beta^2\frac{2(t+\gamma)+1}{(t+\gamma)^2(t+1+\gamma)^2} \tag{127}$$

$$\leq\frac{2\beta^2}{(t+\gamma)^3} \tag{128}$$

$$=\frac{2\eta_t^3}{\beta}, \tag{129}$$

and with $\beta>\frac{1}{m}, \eta_t\leq\eta_1\leq\frac{1}{4M}$, we get

$$\|\mathbf{\Sigma}^{t+1}-\mathbf{\Sigma}^{t}\|_1 \tag{130}$$

$$\leq\frac{bN^2\eta_t^3}{C}(\frac{2G^2}{\beta}+2MEG+M^2E^2G^2\eta_t) \tag{131}$$

$$\leq\frac{bN^2\eta_t^3}{C}(2mG^2+2MEG+\frac{1}{4}ME^2G^2) \tag{132}$$

$$=\eta_t^3D, \tag{133}$$

where

$$D=\frac{bN^2}{C}(2mG^2+2MEG+\frac{1}{4}ME^2G^2). \tag{134}$$

With Eq. 114 and Eq. 133, we have

$$\Delta_{t+1} \leq(1-\eta_t m)\Delta_t+2E\eta_t^2F_t+(1+\frac{1}{\eta_t m})\eta_t^3\delta D \tag{135}$$

$$\leq(1-\eta_t m)\Delta_t+\eta_t^2(2EF_t+\frac{\delta}{m}D)+\eta_t^3\delta D \tag{136}$$

$$\leq(1-\eta_t m)\Delta_t+\eta_t^2(\tilde{F}+\tilde{D}), \tag{137}$$

where

$$\tilde{F}=2E\max_t F_t, \tag{138}$$

$$\tilde{D}=(\frac{1}{m}+\frac{1}{4M})\delta D. \tag{139}$$

Now we can use the same trick in [19] to finish the proof of convergence. With a diminishing learning rate, $\eta_t=\frac{\beta}{t+\gamma}$ for some $\beta>\frac{1}{m}$ and $\gamma>0$ such that $\eta_1\leq\min\{\frac{1}{m},\frac{1}{4M}\}=\frac{1}{4M}$, we will prove by induction that $\Delta_t\leq\frac{\nu}{\gamma+t}$, where $\nu=\max\{\frac{\beta^2(\tilde{F}+\tilde{D})}{\beta m-1},(\gamma+1)\Delta_1\}$.

With the definition of ν, we ensure that $\Delta_1\leq\frac{\nu}{\gamma+1}$. Now we assume that $\Delta_t\leq\frac{\nu}{\gamma+t}$ holds for some t, we have

$$\Delta_{t+1} \leq(1-\eta_t m)\Delta_t+\eta_t^2(\tilde{F}+\tilde{D}) \tag{140}$$

$$\leq(1-\frac{\beta m}{t+\gamma})\frac{\nu}{t+\gamma}+\frac{\beta^2(\tilde{F}+\tilde{D})}{(t+\gamma)^2} \tag{141}$$

$$=\frac{t+\gamma-1}{(t+\gamma)^2}\nu+\Big[\frac{\beta^2(\tilde{F}+\tilde{D})}{(t+\gamma)^2}-\frac{\beta m-1}{(t+\gamma)^2}\nu\Big] \tag{142}$$

$$\leq\frac{t+\gamma-1}{(t+\gamma-1)^2+2(t+\gamma)-1}\nu \tag{143}$$

$$\leq\frac{t+\gamma-1}{(t+\gamma-1)^2+2(t+\gamma-1)}\nu \tag{144}$$

$$\leq\frac{\nu}{t+\gamma+1}. \tag{145}$$

Eq. 143 also arises from the definition of ν that $\beta^2(\tilde{F}+\tilde{D})\leq(\beta m-1)\nu$. Accordingly, for all t, we have $\Delta_t\leq\frac{\nu}{\gamma+t}$ holds. □

With this result, we prove that Δ_t converges to 0 with convergence rate $\mathcal{O}(\frac{1}{T})$, and thus we can say that the proxy algorithm of FedCor converges to the global optimal with convergence rate $\mathcal{O}(\frac{1}{T})$ with Corollary 3.

C. Experiment Details

We simulate the training process of federated learning on one machine. All experiments in this paper are run on one NVIDIA 2080-Ti GPU and two Intel Xeon E5-2630 v4 CPUs. The experiments on FMNIST require around 3 hours for each seed, and the experiments on CIFAR-10 require around 10 hours for each seed.

C.1. Model Parameters

Hyperparameters in FMNIST We follow [4] to construct the neural model on FMNIST: An MLP model with two hidden layers with 64 and 30 units, respectively. Under all three heterogeneous settings, we set the local batch size $B=64$ and the number of local iterations $E=20$. The learning rate η_0 is set to 0.005 initially, and halved at the 150-th and 300-th rounds. An SGD optimizer with a weight decay of 0.0001 and no momentum is used. We allocate data to $N=100$ clients, and set the participation fraction $C=10$ for the 1SPC setting, and $C=5$ for the 2SPC and Dir settings.

Hyperparameters in CIFAR-10 We use a CNN with three convolutional layers [29] with 32, 64 and 64 kernels, respectively. And all convolution kernels are of size 3×3. Finally, the outputs of convolutional layers are fed into a fully-connected layer with 64 units. Under all three heterogeneous settings, we set the local batch size $B=50$ and the

number of local iterations $E = 40$. We use a learning rate $\eta = 0.01$ without learning rate decay, and a weight decay of 0.0003 for the SGD optimizer. The total number of clients and the client participation fraction are the same as those in FMNIST.

Hyperparameters for FedCor We set the dimension of client embedding $d = 15$ for all experiments. In Eq. (16), we set $M = 10, S = 1$ for the warm-up phase, and $M = 1, S = 1$ for the normal phase. And we set the discount factor $\gamma = \theta^{\Delta t}$ where $\theta = 0.9$ for experiments on FMNIST and $\theta = 0.99$ for experiments on CIFAR-10. In each GP update round t, we use $\boldsymbol{X}^{t-1}$ as the initialization and use an Adam optimizer [11] with learning rate 0.01 to optimize for $\boldsymbol{X}^t$. Notice that although Eq. (16) has a closed form optimal solution for $\boldsymbol{X}^t$, we still learn $\boldsymbol{X}^t$ with the gradient decent method with the initialization $\boldsymbol{X}^{t-1}$ in order to utilize the covariance stationarity and reduce the evaluation bias with small number of samples.

Hyperparameters for other baselines We use the same parameters $\alpha_1 = 0.75, \alpha_2 = 0.01$ and $\alpha_3 = 0.1$ as those in the paper [6] for Active Federated Learning. And we set $d = 2NC$ for Power-of-choice Selection Strategy, which is empirically shown to be the best value of d in a highly heterogeneous setting in the paper [4].

Note that we implement the random selection strategy as uniformly sampling clients from $\mathbb{U}$ without replacement [23], while Cho et al. [4] implement the random selection strategy as sampling clients with replacement. Thus, our implemented random selection strategy achieves better performances than their implementation.

C.2. Dirichlet Distribution for Data Partition

We follow the idea in [7] to construct the Dir heterogeneous setting, while we make some modifications to get an unbalanced non-identical data distribution.

For each client k, we sample the data distribution $\boldsymbol{q}_k \in \mathbb{R}^{10}$ from a dirichlet distribution independently, which could be formulated as

$$\boldsymbol{q}_k \sim \text{Dir}(\alpha \boldsymbol{p}), \tag{146}$$

where $\boldsymbol{p}$ is the prior label distribution and $\alpha \in \mathbb{R}_+$ is the concentration parameter of the dirichlet distribution. We group $\boldsymbol{q}_k$ of all the clients together and get a fraction matrix $Q = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_n]$. We denote the size of dataset on each client as $\boldsymbol{x} = [x_1, \cdots, x_N]^T$ and we get it from a solution of a quadratic programming:

$$\min_{\boldsymbol{x}} \quad \boldsymbol{x}^T \boldsymbol{x} \tag{147}$$

$$\text{subject to} \quad Q\boldsymbol{x} = \boldsymbol{d} \tag{148}$$

$$\boldsymbol{x} \in \mathbb{R}^N_{++}, \tag{149}$$

where $\boldsymbol{d}$ is the number of data with each label. We minimize $\|\boldsymbol{x}\|_2$ to avoid the cases where data distribution is over-concentrated on a small fraction of clients. In that case, the client selection problem might become trivial, since we can always ignore those clients with a small dataset and select those with a large dataset.

D. Extra Experimental Results

D.1. Ablation Study: Annealing Coefficient

We conduct experiments on FMNIST and CIFAR-10 with different annealing coefficient β. We setup our experiments under three heterogeneous settings as in Section 5, with different annealing coefficient β ($\beta = 0.95, 0.75, 0.5$ for FMNIST and $\beta = 0.97, 0.95, 0.9$ for CIFAR-10). We fix the GP training interval Δt to 10 for FMNIST and 50 for CIFAR-10. The test accuracy curves are shown in Figure 8. We can see that within a large range, the value of annealing coefficient only slightly influence the convergence rate as well as the final accuracy. Recalling the results of different GP training intervals Δt in Section 5.3, we can say that our method is not sensitive to the hyperparameters Δt and β.

We present the selected frequency of each client in Figure 10 and Figure 11 for FMNIST and CIFAR-10 respectively. We can see that with a smaller β, the selected frequency tends to be more "uniform". However, this does not mean that our selection strategy is equivalent to the uniformly random selection. Our sequential selection strategy introduces dependencies between selected clients as discussed in the multi-iteration insights in Section 4.3, which makes our selection strategy prefer some combinations of selected clients to others, while the uniformly random selection treats all the combinations equally. The advantage shown in Figure 8 compared to the uniformly random strategy demonstrates that selecting a good combination of clients, not only a good individual, is important.

D.2. Normality Verification

We setup experiments to show that Gaussian Distribution can model the loss changes w.r.t. uniformly sampled client selection. To verify this, in the last round of the warm-up phase, we perform the following procedure to examine the normality.

1. We uniformly sample 1000 different client selections $\{\mathbb{S}_{t,i} : i = 1, \cdots, 1000\}$ and collect the corresponding loss changes $\Delta \boldsymbol{l}^t(\mathbb{S}_{t,i}) = [\Delta l_1^t(\mathbb{S}_{t,i}), \cdots, \Delta l_N^t(\mathbb{S}_{t,i})]$ for each of them.

2. We perform PCA on $\{\Delta \boldsymbol{l}^t(\mathbb{S}_{t,i}) : i = 1, \cdots, 1000\}$ to extract the principle components.

3. We plot the histogram of each principle component and compare its distribution with the Gaussian Distribution.

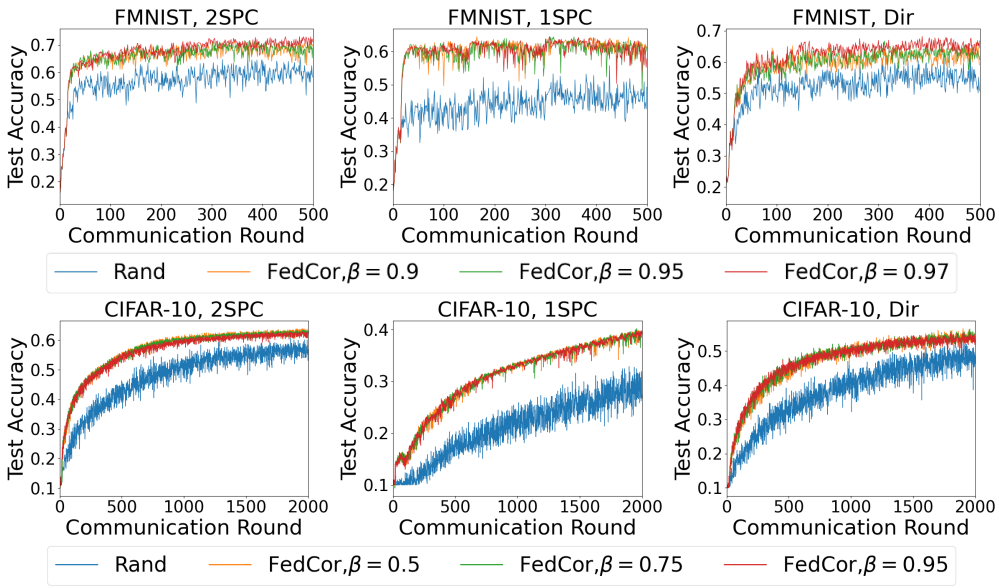

Figure 8. Test accuracy with different annealing coefficient β on FMNIST (top) and CIFAR-10 (bottom) under three heterogeneous settings (left: 2SPC; median: 1SPC; right: Dir).

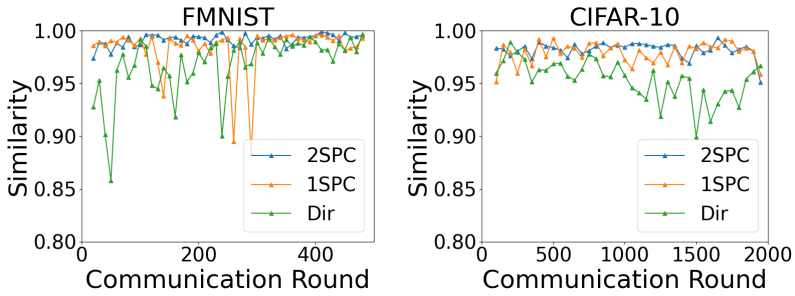

Figure 9. Verification of covariance stationarity on FMNIST and CIFAR-10.

We do not use Multivariate Normality Test directly because we find that $\mathbf{\Sigma}^t$ is always nearly singular, which makes the Multivariate Normality Test unstable. Thus, we turn to perform PCA and visualize each principle component to verify the normality.

The results of FMNIST and CIFAR-10 are shown in Figure 12 and Figure 13 respectively. The red line shows the probability density of Gaussian Distribution with the mean and variance of that principle component. We can see that in all our experiments, Gaussian Distribution can fit the distribution of the principle component well, which verifies that Lemma 1 does hold in all the experiment settings.

D.3. Covariance Stationarity Verification

We examine that assumption in Section 4.5 that the covariance keep approximately stationary during the FL training, namely,

$$\forall t, \mathbf{\Sigma}^t \approx \mathbf{\Sigma}^{t+\Delta t}. \tag{150}$$

To verify this, every Δt rounds ($\Delta t = 10$ for FMNIST and $\Delta t = 50$ for CIFAR-10), we randomly sample 1000 client selections $\mathbb{K}_i$ and collect the corresponding loss changes $\Delta \boldsymbol{l}^t(\mathbb{K}_i)$. We directly calculate the covariance matrix $\mathbf{\Sigma}^t$ with these samples $\{\Delta \boldsymbol{l}^t(\mathbb{K}_i) : i = 1, \cdots, 1000\}$. Then for each adjacent pair of covariance matrix, we calculate their cosine similarity as follows.

$$\text{similarity}(\Sigma^{\text{t}}, \Sigma^{\text{t}+\Delta\text{t}}) = \frac{\text{tr}(\Sigma^{\text{t}\,\text{T}}\Sigma^{\text{t}+\Delta\text{t}})}{\text{tr}(\Sigma^{\text{t}\,\text{T}}\Sigma^{\text{t}})\text{tr}(\Sigma^{\text{t}+\Delta\text{t}\,\text{T}}\Sigma^{\text{t}+\Delta\text{t}})} \tag{151}$$

The similarity is in range $[0, 1]$, and a larger one shows a higher similarity.

The results are shown in Figure 9. We can see that in most cases the similarity is larger than 0.9, which verifies our claim of the covariance stationarity.

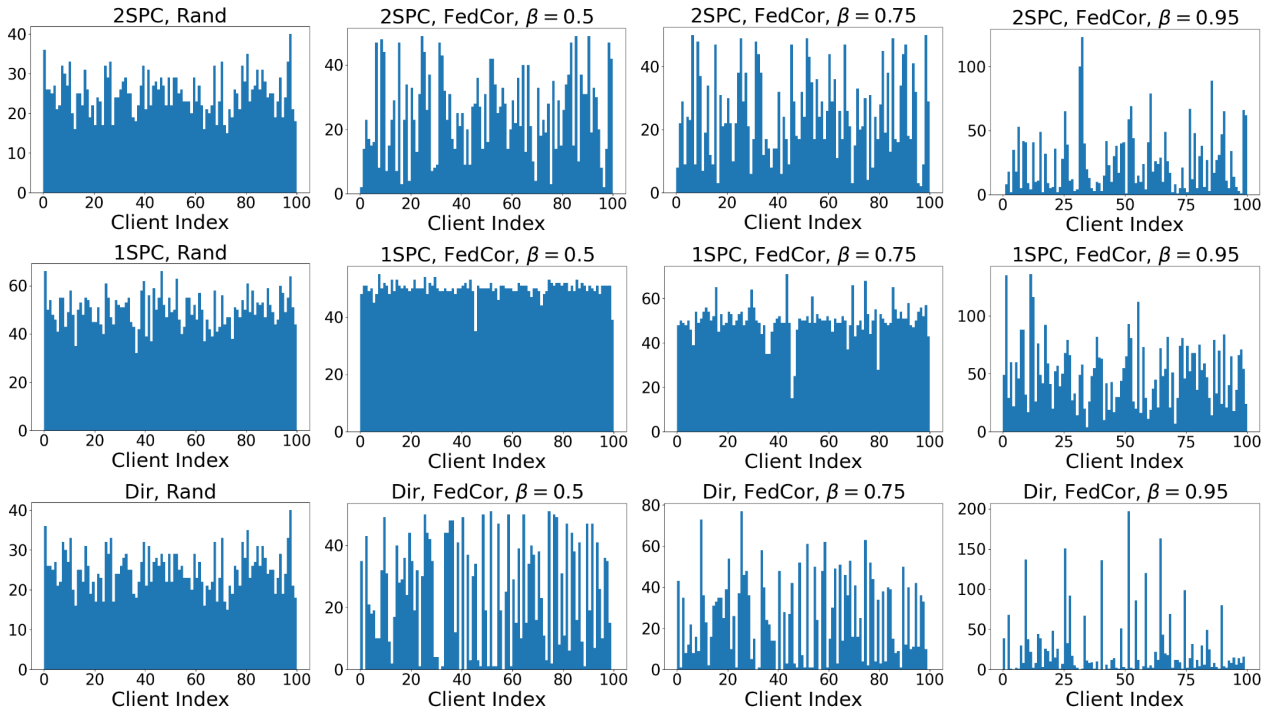

Figure 10. Selected Frequency of each client with different annealing coefficient β on FMNIST under three heterogeneous settings (top: 2SPC; median: 1SPC; bottom: Dir).

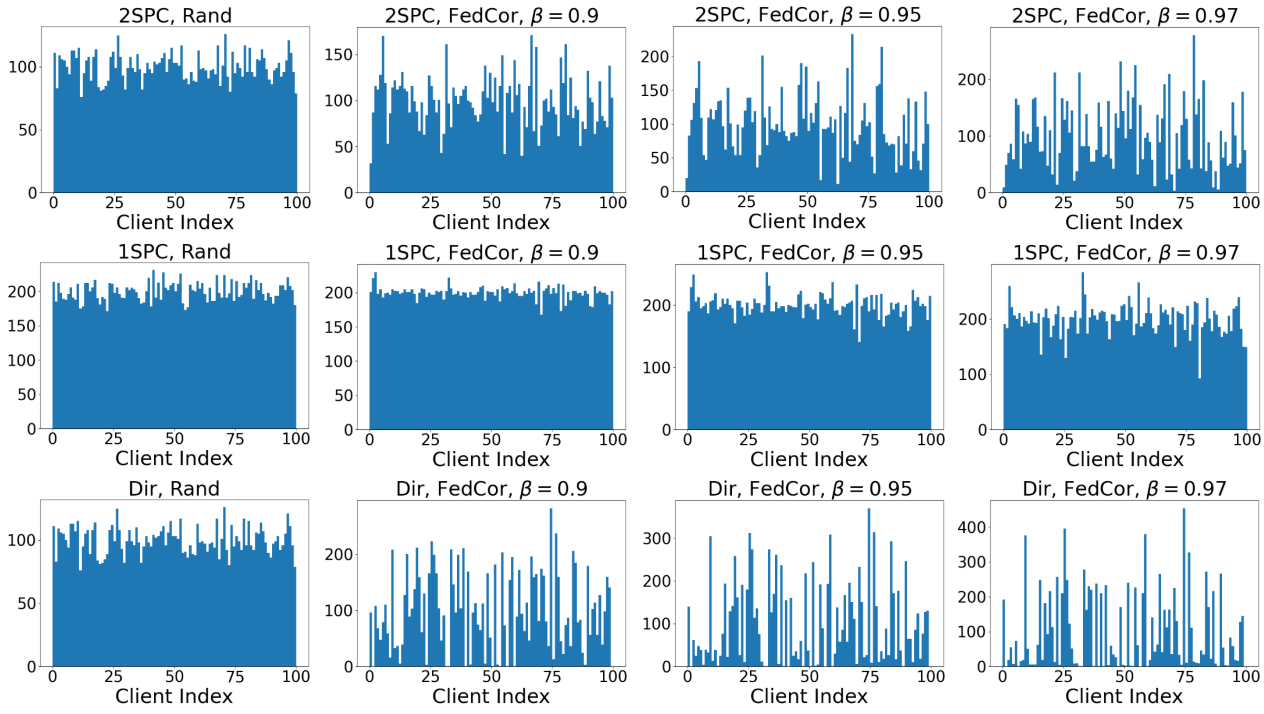

Figure 11. Selected Frequency of each client with different annealing coefficient β on CIFAR-10 under three heterogeneous settings (top: 2SPC; median: 1SPC; bottom: Dir).

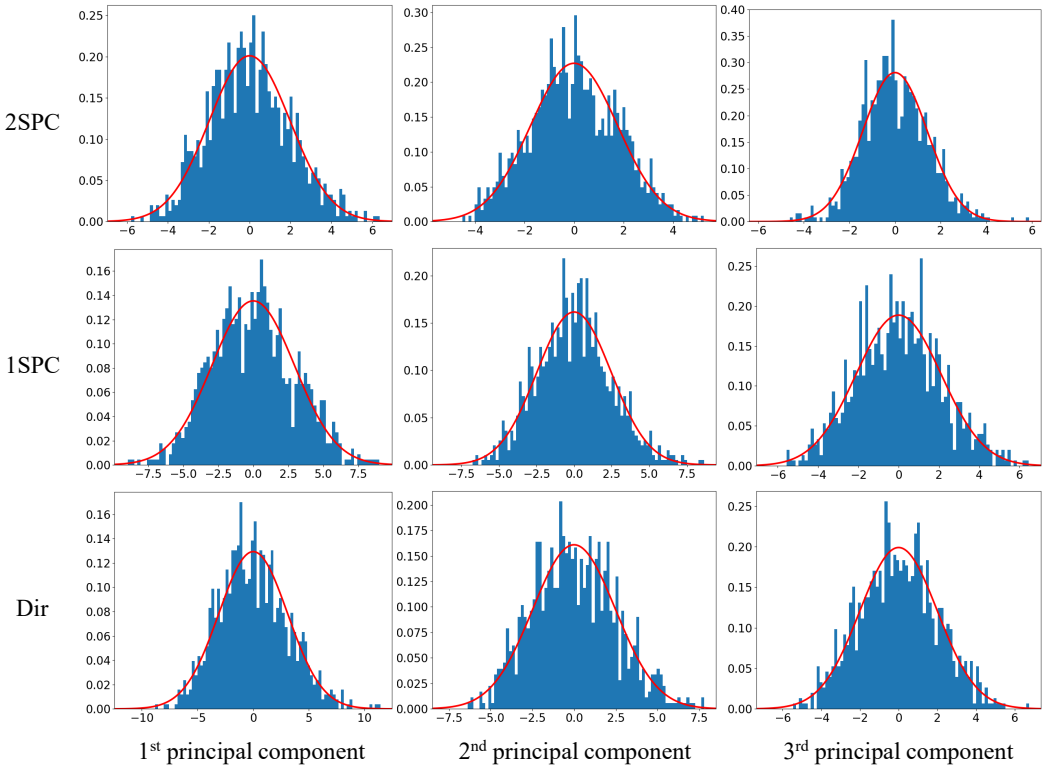

Figure 12. Normality Test on FMNIST.

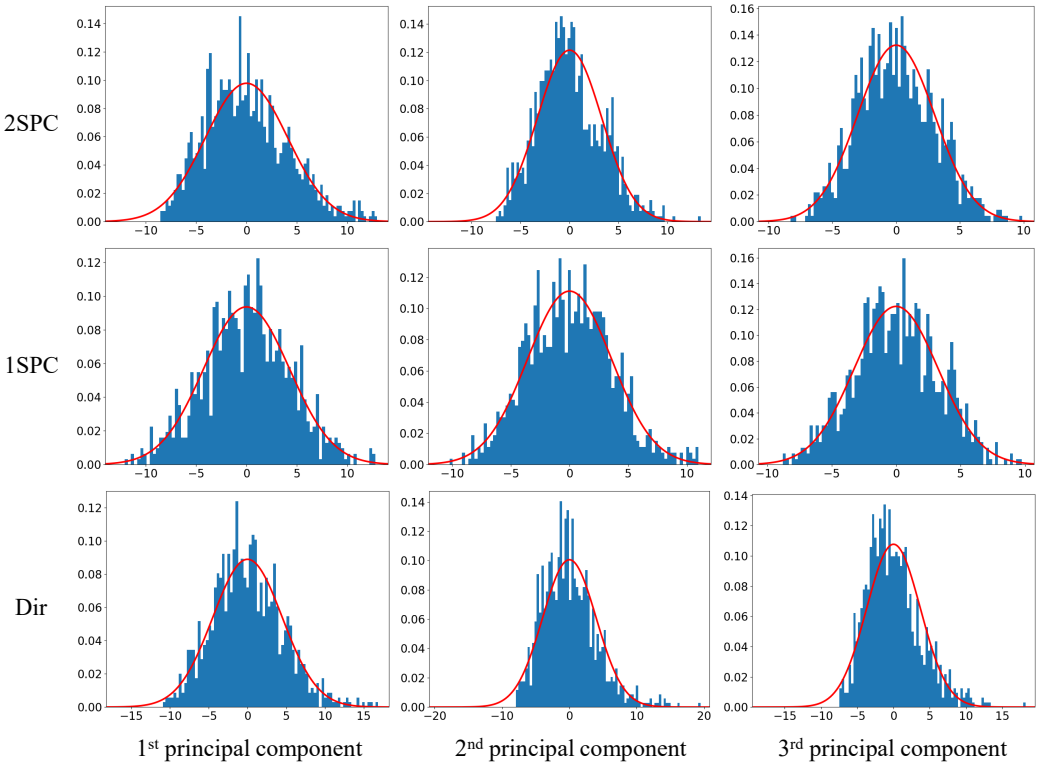

Figure 13. Normality Test on CIFAR-10.