Variance-Reduced Heterogeneous Federated Learning via Stratified Client Selection

Guangyuan Shen\textsuperscript{1,2}, Dehong Gao\textsuperscript{2}, Libin Yang\textsuperscript{1} \textsuperscript{*}, Fang Zhou\textsuperscript{1}, Duanxiao Song\textsuperscript{1}, Wei Lou\textsuperscript{3} and Shirui Pan\textsuperscript{4}

\textsuperscript{1}Department of Cyber Science and Technology, Northwestern Polytechnical University, China
\textsuperscript{2}Alibaba Group, China
\textsuperscript{3}Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
\textsuperscript{4}Department of Data Science and AI, Faculty of IT, Monash University, Australia

\{gysen, libiny, zhufang, songduanxiao\}@mail.nwpu.edu.cn, dehong.gdh@alibaba-inc.com, csweiou@comp.polyu.edu.hk, shirui.pan@monash.edu,

Abstract

Client selection strategies are widely adopted to handle the communication-efficient problem in recent studies of Federated Learning (FL). However, due to the large variance of the selected subset’s update, prior selection approaches with a limited sampling ratio cannot perform well on convergence and accuracy in heterogeneous FL. To address this problem, in this paper, we propose a novel stratified client selection scheme to reduce the variance for the pursuit of better convergence and higher accuracy. Specifically, to mitigate the impact of heterogeneity, we develop stratification based on clients’ local data distribution to derive approximate homogeneous strata for better selection in each stratum. Concentrating on a limited sampling ratio scenario, we next present an optimized sample size allocation scheme by considering the diversity of stratum’s variability, with the promise of further variance reduction. Theoretically, we elaborate the explicit relation among different selection schemes with regard to variance, under heterogeneous settings, we demonstrate the effectiveness of our selection scheme. Experimental results confirm that our approach not only allows for better performance relative to state-of-the-art methods but also is compatible with prevalent FL algorithms.\textsuperscript{1}

1 Introduction

Federated Learning (FL) is a distributed learning paradigm for training a global model from data scattered across different clients without exchanging sensitive data [Konečný et al., 2015]. During the training process, each client needs to operate their data locally and transfer the model update between the server and themselves back and forth. Such a training process may raise many challenges, with communication cost often being the critical bottleneck.

\textsuperscript{*}Contact Author
\textsuperscript{1}Code is made available at https://anonymous.4open.science/r/FL-stratified-client-selection/

Many creative methods have been proposed to address the communication cost issue. FedAvg [McMahan et al., 2017] is one of the most representative FL algorithms where the server randomly generates a subset from the entire client set instead of yielding all clients involved. The selection of subset will affect the FL convergence through the mean squared error between the subset’s update and the true model update, which is naturally equivalent to the variance of the subset’s update under unbiased selection. Generally, the variance of random selection heavily depends upon \textbf{a) sampling ratio} and \textbf{b) degree of heterogeneity}. In most FL scenarios, heterogeneity and limited sampling ratio are becoming increasingly common. However, in such a condition, current random selection schemes are probably facing performance degradation due to a large variance which directly slows the FL convergence.

Many studies try to tackle the heterogeneity problem by leveraging importance sampling in FL, where the selected probability of each client is determined proportionally to the importance of the client such as the norm of update [Chen et al., 2020], data variability [Rizk et al., 2021], and test accuracy [Mohammed et al., 2020]. However, none of these works capture the essential client correlations well, which cannot guarantee the heterogeneity reduction of the population directly. Differing from importance sampling approach, a substantial amount of other works [Fraboni et al., 2021; Muhammad et al., 2020] propose cluster sampling schemes to enable more uniform outcomes, where the random selection is applied to each cluster. Such cluster sampling schemes prefer to employ proportional allocation, namely, the sample size for each cluster is proportional to the cluster size, regardless of the diversity characteristic of the clusters’ variability. That may ultimately lead to performance degradation under a limited sampling ratio.

To deal with the above problems, we propose a novel stratified client selection scheme for heterogeneous FL, enabling variance reduction under a limited sampling ratio. Specifically, to mitigate the impact of heterogeneity, we develop stratification based on clients’ local data distribution, which is capable of deriving approximate homogeneous strata for better selection in each stratum. Concentrating on a limited sampling ratio, we make the first attempt to develop an opti-
mized sample size allocation scheme by considering the diversity of the stratum’s variability, with the promise of further variance reduction. Besides, we theoretically elaborate the explicit relation among different selection schemes and verify its advantages with regard to variance. Extensive experiments show the promising performance of the proposed scheme compared to the cutting-edge methods. Moreover, we demonstrate that our stratified client selection strategy is compatible with prevalent FL algorithms, also yielding favorable performance in heterogeneous FL. Our contributions are summarized as follows:

- We propose a novel stratified client selection scheme to mitigate the impact of heterogeneity under a limited sampling ratio, capable of deriving approximate homogeneous strata and optimizing sample size allocation, which leads to better selection.
- We elaborate the explicit relation among different selection schemes with regard to variance, which confirms the better capability of our approach for variance reduction.
- We perform extensive experiments on our approach compared with several benchmark datasets in various settings. The results demonstrate the significant improvement in terms of train convergence and final test accuracy, as well as its compatibility.

2 Related Work

**Importance sampling.** Importance sampling is a non-uniform client selection method, where the probability for each client to be chosen is proportional to their importance. Different researchers adopt different methods to measure the importance, e.g., [Rizk et al., 2021; Chen et al., 2020; Mohammed et al., 2020; Cho et al., 2020] use clients’ local data variability, norm of update, test accuracy, and local loss, respectively. Moreover, many researchers extend such importance sampling scheme to the real scenario, where the importance of each client can be measured by clients’ real history information, e.g., [Xu and Wang, 2020; AbdulRahman et al., 2020; Feng et al., 2020; Ma et al., 2021]. Further, [Nishio and Yonetani, 2019] optimizes the client selection problem under heterogeneous resources, while [Yoshida et al., 2020] optimizes under uncertain resources. Such methods can yield better selection by carefully selecting the clients to participate based on their importance. However, none of them capture the essential client correlations well, which cannot guarantee the heterogeneity reduction of the population directly.

**Cluster sampling.** [Fraboni et al., 2021; Xu et al., 2021; Muhammad et al., 2020] propose cluster sampling to enable more uniform selection, where the sample size for each cluster is proportional to the cluster size, ignoring the diversity of the stratum’s variability. That may ultimately lead to a performance degradation under a limited sampling ratio.

3 Preliminaries

Before going into more details of stratified client selection scheme in Section 4, we give a theoretical review of FedAvg. We start by giving a problem formulation of Federated Learning following [McMahan et al., 2017].

3.1 Federated Learning Problem Formulation

In Federated Learning, a total number of $N$ clients aims to jointly solve the following distributed optimization problem:

$$
\min_{\mathbf{w}} \left\{ F(\mathbf{w}) := \sum_{k=1}^{N} \omega_k F_k(\mathbf{w}) \right\},
$$

where $\omega_k$ is the weight of the $k^{th}$ client, s.t. $\omega_k \geq 0$ and $\sum_{k=1}^{N} \omega_k = 1$. Suppose the $k^{th}$ client possess $n_k$ training data: $\{x_{k,1}, x_{k,2}, \ldots, x_{k,n_k}\}$. The local objective function $F_k(\mathbf{w})$ is defined by

$$
F_k(\mathbf{w}) := \frac{1}{n_k} \sum_{j=1}^{n_k} f(\mathbf{w}; x_{k,j}),
$$

where $f(\mathbf{w}; x_{k,j})$ is a user-specified loss function made with model parameters $\mathbf{w}$.

3.2 FedAvg Description

In the $t^{th}$ train round of the FedAvg, the server broadcasts the latest global model $\mathbf{w}_t$ to all the clients. The $k^{th}$ client sets $\mathbf{w}_t^k \leftarrow \mathbf{w}_t$ and performs local training for $E$ epochs:

$$
\mathbf{w}_{t+1}^k \leftarrow \mathbf{w}_t^k - \eta_t \nabla F_k(\mathbf{w}_t^k; \xi_t^k),
$$

where $\eta_t$ is the learning rate and $\xi_t^k$ is a data sample uniformly selected from the local data. The server aggregates the local model update to produce the new global model $\mathbf{w}_{t+E}$.

$$
\mathbf{w}_{t+E} \leftarrow \sum_{k=1}^{N} \omega_k \mathbf{w}_{t+1}^k.
$$

In fact, FedAvg randomly generates a subset $S_t$ consisting of $m$ clients from the entire client set and aggregates the model update, $\{\mathbf{w}_{t+1}^1, \ldots, \mathbf{w}_{t+1}^m\}$,

$$
\mathbf{w}_{t+E} \leftarrow \sum_{k \in S_t} \sum_{i=1}^{N} m \omega_k \mathbf{w}_i^k.
$$

**Definition 3.1 (Unbiased Selection).** A client selection scheme is defined as unbiased if the expected model update of the client aggregation is equal to the global deterministic aggregation obtained when considering all the clients, i.e.,

$$
\mathbb{E}[\mathbf{w}(S_t)] = \mathbb{E} \left[ \sum_{k \in S_t} \sum_{i=1}^{N} \omega_k \mathbf{w}_i^k \right] := \sum_{k=1}^{N} \omega_k \mathbf{w}_i^k = \mathbf{w}(\mathcal{K}).
$$

For convenience, in the rest of paper, we use $\mathbf{w}(S_t)$ to denote the model update aggregated from subset $S_t$, and $\mathbf{w}(\mathcal{K})$ to denote the true model update of entire client set $\mathcal{K}$.

4 Our Approach

In Section 4.1, we formalize the partial client selection problem in FL and propose an optimization objective with two constraints. In Section 4.2, a novel client selection scheme is proposed with the promise of better train convergence and higher test accuracy.
4.1 Problem Formulation

Consider a partial participation FL framework [Li et al., 2020] where we use $w(S_t)$ the model update aggregated from a subset $S_t$ to estimate $w(K)$ the true averaged model update of the entire client set $K$. Selecting $m$ clients can be regarded as sample one client $m$ times with independent discrete probability distributions $\{P_t^i\}_{i=1}^m$. For each distribution, we have $P_t^i = \{p_k^i\}_{k=1}^N$, where $p_k^i$ denote the probability for the $k^{th}$ client to be selected in $i^{th}$ distribution $P_t$. The selection of subset $S_t$ affects the FL convergence through the mean squared error (MSE) between $w(S_t)$ and $w(K)$, i.e., $\mathbb{E}\|w(S_t) - w(K)\|^2$, which can be decomposed into:

$$\text{MSE} = \mathbb{E}\|w(S_t) - \mathbb{E}[w(S_t)]\|^2 + \mathbb{E}[\mathbb{E}[w(S_t)] - w(K)]^2. \tag{7}$$

The former in equation (7) is the variance of model update aggregated from $w(S_t)$ which goes to zero if all clients participate. The latter is the bias between the expectation of $w(S_t)$ and the true model update $w(K)$. The bias will vanish if the selection is unbiased. For better train convergence and higher test accuracy, under unbiased conditions, we seek to minimize the MSE that is equivalent to minimize the variance corresponding to the following equation:

$$\min_{p_k^i} \left\{ \mathbb{E}\|w(S_t) - w(K)\|^2 = \mathbb{V}[w(S_t)] \right\},$$

$$s.t. \sum_{k=1}^N p_k^i = 1 \text{ with } p_k^i \geq 0, \|S_t\| = qN = m, \tag{8}$$

where we use $q$ and $N$ to denote the sampling ratio and the total number of all clients, respectively.

4.2 Our Selection Scheme

The heterogeneity can have a significant impact on the performance of simple random selection in terms of variance. Specifically, the more heterogeneous the datasets, the larger the variance. However, prior selection scheme cannot reduce the heterogeneity of the population directly. To mitigate the impact of the data heterogeneity, we propose a stratified client selection scheme for the pursuit of better selection outcomes.

**Stratified Selection.** As illustrated in Figure 1, at the beginning of the entire train process, we develop stratification based on each client’s local data distribution to derive approximate homogeneous strata for better selection in each stratum. For $\forall i \neq j$, there is no overlap between $i^{th}$ and $j^{th}$ stratum, which indicates that every client must be assigned to one and only one stratum. In practice, we observed that the difference in the number of strata has a great impact on the effectiveness of the stratification. Therefore, we pick the certain number of strata with the highest stratification effectiveness measured by the Silhouette Coefficient (details of client stratification can be found in Appendix A.1). Generating a subset $S_t$ consisting of $m$ clients by stratified client selection can be regard as selecting $m$ clients using $H$ different probability distribution $\{P_t\}_{t=1}^H$ where $P_t = \{p_k^i\}_{k=1}^N$. For example, if we select a client from the $h^{th}$ stratum, probability distribution $P_h = \{p_k^h\}_{k=1}^N$ will be used, the $p_k^h$ comes as follows:

$$p_k^h = \begin{cases} 0, & \text{if } k^{th} \text{ client } \notin h^{th} \text{ stratum} \\ \frac{1}{N_h}, & \text{if } k^{th} \text{ client } \in h^{th} \text{ stratum} \end{cases} \tag{9}$$

where $N_h$ denotes the number of clients in the $h^{th}$ stratum $s.t. \sum_{h=1}^H N_h = N$. Naturally, after stratification, we assign every client into approximate homogeneous strata. As shown in Figure 1, clients in the same stratum will have similar data distribution. Different color denotes the heterogeneity of clients.

**Algorithm 1 StraSAMP (section 4)**

**Input:**
- $K$ is the entire set of clients
- $m$ is the number of clients to sample from $K$
- $S_h$ is the standard deviation of $h^{th}$ stratum
- $N_h$ is the number of clients in $h^{th}$ stratum
- $H$ is the certain number of strata

**Initialize** $G \leftarrow \text{ClientStratification}(K, H)$

1: for each stratum $G_h$ from 1 to $H$ do
2: $m_h \leftarrow \sum_{h=1}^H N_h S_h \cdot m$
3: $S \leftarrow$ the server randomly selects $m_h$ clients from each stratum $G_h$
4: end for
5: return $S$

**Output:** $S$ is a subset of $K$ consisting of $m$ clients

![Figure 1. Comparison between random and stratified selection in visualization. Different color denotes the heterogeneity of clients. Vertical arrows denote the selection process.](image-url)
Algorithm 2 FedAvg with Optimal Stratified Selection

Input:
- The $N$ clients are indexed by $k$
- $E$ is the number of local epochs
- $\eta$ is the learning rate
- $q$ is the sampling ratio
- $H$ is the certain number of strata

Initialize $w_0$
Initialize $G$ ← ClientStratification ($\mathcal{K}, H$

Server executes:
1: for each round $t = 1, 2, \cdots$ do
2: $m \leftarrow \max(qN, 1)$
3: $S_t \leftarrow$ StraSAMP($N, m, G$)
4: for each client $k \in S_t$ in parallel do
5: \hspace{1em} $w_{t+1} \leftarrow$ ClientUpdate($S_t, w_t$)
6: \hspace{1em} end for
7: $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{N}{m} w_{t+1} k$
8: end for

ClientUpdate:
1: for each local epoch $i$ from 1 to $E$ do
2: \hspace{1em} $w_{t+1} \leftarrow w_{t+1} - \eta \nabla F_k (w_{t+1}, \xi_{t+i})$
3: \hspace{1em} end for
4: $w_{t+1} \leftarrow w_{t+1} \xi_{t+i}$
5: return $w_{t+1}$

Optimal Allocation. To further improve the performance of random selection, we make the first attempt to develop an optimized sample size allocation scheme by considering both the stratum size $N_h$ and the diversity of its variability.

$$m_h = \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \cdot m,$$

where $m_h$ denotes the number of clients in subset $S_t$ from the $h$th stratum $s.t.$ $\sum_{h=1}^{H} m_h = m$, and the $S_h$ denotes standard deviation of $h$th stratum representing its variability. Considering the difficulty of evaluating the accurate numerical standard deviation of a certain stratum, we use Cohesion to reflect the variability of a certain stratum (refer to Appendix A.2).

Algorithm 1 summarizes the main steps of our stratified client selection scheme. In addition, our selection scheme is orthogonal to and compatible with the existing modified FL algorithm, e.g., [McMahan et al., 2017; Wang et al., 2020; Li et al., 2020]. Algorithm 2 summarizes the main steps of FedAvg with optimal stratified client selection.

5 Theoretical Analysis

5.1 Variance Reduction

Theorem 1 (Variance Reduction). If the population is large compared to the subset, $\frac{m}{N}, \frac{m}{n}, \frac{m}{m_n}, \frac{m}{N}$ and $\frac{1}{N}$ are negligible, then the variance of different selection schemes satisfy

$$V(w_{opt}) \leq V(w_{prop}) \leq V(w_{ran}) ,$$

where $w_{opt}, w_{prop}, w_{ran}$ denote the model update aggregated from subset $S_t$ that selected by optimal stratified, proportional stratified and random selection, respectively.

Proof sketch. We demonstrate the equation (11) below to describe the relationship between $V(w_{prop})$ and $V(w_{ran})$:

$$V(w_{ran}) = V(w_{prop}) + \sum_{h=1}^{H} \frac{N_h (w_h - w(k))}{m N}^2.$$

We have $V(w_{prop}) < V(w_{ran})$, unless $\forall h \in \{1, \cdots, H\}$ $w_h = w(k)$, i.e., each stratum have the same averaged model update with the entire population. As for the relationship between $V(w_{opt})$ and $V(w_{prop})$:

$$V(w_{prop}) = V(w_{opt}) + \sum_{h=1}^{H} \frac{N_h (S_h - \overline{S})^2}{m N}.$$

We have $V(w_{opt}) < V(w_{prop})$, unless $\forall h \in \{1, \cdots, H\}$ $S_h = \overline{S}$, i.e., each stratum never has idendtical update mean and variability in heterogeneous FL.

5.2 Convergence Behavior

Following [Li et al., 2019], we derive the theoretical guarantees regarding the convergence of FedAvg with our stratified selection scheme. To verify the faster convergence of our selection scheme, we derive the required communication round for achieving the $\epsilon$ error. We next state the assumptions used in our theorem and proof, which are common in the federated optimization literature, e.g., [Li et al., 2019]. Assume each function $F_k(k \in [N])$ is $\mu$-strongly convex and $L$-smooth. Suppose that for all $k \in [N]$ and all $t$, the variance and expectation of stochastic gradients in each client on random samples $\xi$ is bounded by $\sigma_k^2$ and $G^2$, i.e., $\mathbb{E} \left[ \left\| \nabla F_k (w_t^k, \xi) - \nabla F_k (w_t^k) \right\|^2 \right] \leq \sigma_k^2$ and $\left\| \nabla F_k^2 (w_t^k, \xi) \right\|^2 \leq G^2$, respectively. And $\Gamma$ is used to quantify the heterogeneity, where $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$.

Theorem 2 (Convergence Bound). Let Assumptions above held and $L, \mu, \sigma_k, G$ be defined therein. Consider FL with FedAvg when sampling $m$ clients, then FedAvg with stratified client selection scheme satisfies:

$$\mathbb{E} [F(w_T)] - F^* \leq O \left( \frac{\sum_{k=1}^{N} p_k \sigma_k^2 + E^2 G^2 + \gamma G^2}{\mu T} + O \left( \frac{LT}{\mu T} \right) + O \left( \frac{E^2 G^2}{m \mu T} \right) \right).$$

Remark. Theorem 2 gives the convergence bound of FedAvg with our sampling scheme. The first two terms are used to bound the FedAvg with full client participant, while the last term is used to bound the variance of the model update aggregated from the subset. If we consider an ideal full participant FL without client selection, the last term will vanish.
Corollary 2.1 (Rate of Convergence). \( T \) denotes the number of the required round for FedAvg with stratified selection to achieve an \( \epsilon \) accuracy, then we have:

\[
\frac{T}{E} \leq \mathcal{O}\left(\sum_{k=1}^{N} p_k^2 \sigma_k^2 + \frac{(k + E + E^2) G^2}{E}\right) + \mathcal{O}\left(\frac{LT}{E}\right) + \mathcal{O}\left(\frac{EG^2}{m}\right).
\]  

We defer the proof to Appendix B.1.

Remark. As the experimental result shows that the ideal full participant FedAvg achieves fast convergence confirming that the last term which is used to bound the selection variance has a significant impact on the rate of convergence. In section 5.1, we demonstrate the guaranteed capability of our selection scheme on variance reduction. Naturally, we have the proposition below.

Proposition 1. Although the selection is under a limited sampling ratio, the rate of heterogeneous FL convergence is further accelerated by our sampling scheme.

6 Experiment

In this section, we evaluate our optimal stratified client selection scheme and compare it with existing selection schemes. Our implementation is made publicly available at https://anonymous.4open.science/r/FL-stratified-client-selection/.

6.1 Experimental Setup

Baseline. We consider simple random sampling [McMahan et al., 2017], importance sampling [Chen et al., 2020] and cluster sampling [Fraboni et al., 2021] to compare with our sampling scheme.

Hyperparameters. \( N \) is the number of all clients, \( q \) is the sampling ratio, \( nSGD \) is the times of SGD running locally, \( \gamma \) is the learning rate, \( B \) is the batch size, \( T \) is the terminate round. For better comparison, we set all these hyperparameters following [Li et al., 2020].

Model. We use a fully connected network with one hidden layer of 50 nodes for MNIST. As for CIFAR-10 and FMNIST, we use the same classifier of [McMahan et al., 2017] composed of 3 convolutional and 2 fully connected layers, including dropout operation after each convolutional layer.

Datasets. We evaluate our approach on different datasets including MNIST, FMNIST [Xiao et al., 2017], CIFAR-10, and various settings.

Data Partition. We partition the whole dataset into 100 clients under IID and non-IID data distribution respectively, which is quite common in optimizing heterogeneity literature, e.g., [Wang et al., 2020]. IID data partition strategy employs the idea of random split to create a uniform federated dataset. To further simulate the impact of non-IID data distributions, we use Dirichlet distribution and "Shard" to partition the entire client set. "Shard" assigns each client the samples under only one class, while Dirichlet distribution, \( \text{Dir}(\alpha) \), gives to each client the respective partitioning across labels by changing the value of \( \alpha \). The value of \( \alpha \) can reflect the degree of heterogeneity. Specifically, the lower the value of \( \alpha \), the more heterogeneous the dataset is. The schematic table of the Dirichlet distribution is presented in Appendix A.3.

Measures. For all methods, to verify the effectiveness in terms of convergence behavior and accuracy, we report accuracy, loss (vertical axis) as a function of the number of rounds (horizontal axis).

6.2 Result and Analysis

Influence of sampling ratio and heterogeneity. We hypothesize that the performance of a random selection depends upon the sampling ratio and the heterogeneity of the population. To verify it, we evaluate the performance of FedAvg with random selection under different sampling ratios and heterogeneity settings. In Figure 2, we observe that with the decreasing of the heterogeneity and the increasing of the sampling ratio, the performance of FedAvg becoming better. Figure 2 shows that random selection can not perform well with a low sampling ratio and heterogeneous datasets, which confirms the necessity to optimize the selection scheme. For more details, please refer to Appendix C.

Effectiveness. Experiments are carried out based on both IID and non-IID partition strategies to simulate the data heterogeneity in the real scenario. Figure 3 shows the top-1 accuracy of FedAvg on MNIST and FMNIST after 200 rounds using simple random sampling, importance sampling, cluster sampling and our sampling scheme with the sampling ratio \( q = 0.1 \). The magenta dotted line refers to the final test accuracy of FedAvg on IID MNIST and FMNIST data set corresponding to 86.9% and 87.4%, respectively. As illustrated in Figure 3, our selection scheme can achieve faster convergence at the early stage of the training process and have smaller shadow areas reflecting better train stability. Table 1 shows the required rounds for different selection schemes to achieve 60% test accuracy, which also indicates the faster convergence of our stratified selection scheme.

Further, the final test accuracy of our approach on the non-IID dataset can be close to the random sampling on the IID dataset. Such results are quite reasonable since our method can reduce the heterogeneity by appropriate stratification.

![Figure 2. Illustration on how the sampling ratio \( q \) and the degree of heterogeneity \( \text{Dir}(\alpha) \) influence the final accuracy of FedAvg on MNIST. A warmer color denotes a higher accuracy.](image-url)
where we assign the clients with similar data distribution into the same stratum. Specifically, we take the average of the last 10 rounds of the test accuracy as the final accuracy after 200 rounds. Our approach achieves a final test accuracy of 83.2% on non-IID MNIST datasets, whereas random sampling, importance sampling, and cluster sampling only achieve 62.0%, 72.1%, and 74.7%, respectively. As for the non-IID FMNIST datasets, our method achieves an accuracy of 92.0%, whereas random sampling, importance sampling, and cluster sampling only achieve 75.9%, 83.7%, and 88.8%, respectively.

To demonstrate the effectiveness of our approach beyond the limited setting of MNIST and FMNIST, we conduct extensive experiments on CIFAR-10. We use Dirichlet distribution to partition the CIFAR-10 to derive datasets with different degrees of heterogeneity. Specifically, the lower the $\alpha$, the more heterogeneous the dataset is. As shown in Figure 4, with $\alpha$ decreasing, the performance of all the other sampling schemes rapidly decreases except ours. The result confirms that the more heterogeneous the dataset is, the larger the improvement achieved by our sampling scheme relative to previous approaches.

**Compatibility.** We carry out extra experiments on FedProx and FedNova, two modified FL algorithms, to verify the compatibility of our selection scheme. We consider MNIST, FMNIST, CIFAR-10 under IID and non-IID data partition. As shown in Table 2, both FedNova and FedProx achieve higher final test accuracy with our selection scheme relative to existing selection schemes. As for the convergence behavior, Figure 7-11 (refer to Appendix C) show that FedNova and FedProx with our selection scheme can achieve faster convergence and better train stability.

**7 Conclusion**

In this paper, we propose a novel stratified client selection scheme guaranteeing better train convergence and higher test accuracy. Specifically, to mitigate the impact of heterogeneity, we develop stratification based on clients’ local data distribution to derive approximate homogeneous strata for better selection in each stratum. Concentrating on a limited sampling ratio scenario, we make the first attempt to construct an optimized sample size allocation scheme by considering the stratum’s variability, with the promise of further variance reduction. Moreover, we show that the performance of our selection scheme is always better than simple random selection. It can achieve a much greater improvement depending on the heterogeneity. Experimental results demonstrate the superiority and the effectiveness of our method. In the future, the co-variance among different clients needs to be considered in our approach, since it may also lead to performance degradation in Federated Learning.
References

[AbdulRahman et al., 2020] Sawsan AbdulRahman, Hanine Tout, Azzam Mourad, and Chameseddine Talhi. Fedmecs: multicriteria client selection model for optimal iot federated learning. *IEEE Internet of Things Journal*, 8(6):4723–4735, 2020.

[Chen et al., 2020] Wenlin Chen, Samuel Horvath, and Peter Richtarik. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.

[Cho et al., 2020] Yae Jee Cho, Samarth Gupta, Gauri Joshi, and Osman Ya˘gan. Bandit-based communication-efficient client selection strategies for federated learning. In *Asilo-mar Conference on Signals, Systems, and Computers*, pages 1066–1069. IEEE, 2020.

[Feng et al., 2020] Chenyuan Feng, Yidong Wang, Zhongyuan Zhao, Tony QS Quek, and Mugen Peng. Joint optimization of data sampling and user selection for federated learning in the mobile edge computing systems. In *IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6, 2020.

[Fraboni et al., 2021] Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *International Conference on Machine Learning*, 2021.

[Koneˇcn`y et al., 2015] Jakub Koneˇcn`y, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

[Li et al., 2019] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

[Li et al., 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2020.

[Ma et al., 2021] Jiahua Ma, Xinghua Sun, Wenchao Xia, Xijun Wang, Xiang Chen, and Hongbo Zhu. Client selection based on label quantity information for federated learning. In *IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1–6, 2021.

[McMahan et al., 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Ar-cas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282, 2017.

[Mohammed et al., 2020] Ihab Mohammed, Shadha Tabatabai, Ala Al-Fuqaha, Faissal El Bouanani, Ju-naid Qadir, Basheer Qolomany, and Mohsen Guizani. Budgeted online selection of candidate iot clients to participate in federated learning. *IEEE Internet of Things Journal*, pages 5938–5952, 2020.

[Muhammad et al., 2020] Khalil Muhammad, Qin Qin Wang, Diarmuid O’Reilly-Morgan, Elias Tragos, Barry Smyth, Neil Hurley, James Geraci, and Aonghus Lawlor. Fed-fast: Going beyond average for faster training of federated recommender systems. In *SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1234–1242, 2020.

[Nishio and Yonetani, 2019] Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *IEEE International Conference on Communications (ICC)*, pages 1–7, 2019.

[Rizk et al., 2021] Elsa Rizk, Stefan Vlaski, and Ali H Sayed. Optimal importance sampling for federated learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3095–3099, 2021.

[Stich et al., 2018] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[Stich, 2018] Sebastian U Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

[Wang et al., 2020] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 2020.

[Xiao et al., 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Xu and Wang, 2020] Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2):1188–1200, 2020.

[Xu et al., 2021] Xiaohui Xu, Sijing Duan, Jinrui Zhang, Yunzhen Luo, and Deyu Zhang. Optimizing federated learning on device heterogeneity with a sampling strategy. In *IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pages 1–10, 2021.

[Yoshida et al., 2020] Naoya Yoshida, Takayuki Nishio, Masahiro Morikura, and Koji Yamamoto. Mab-based client selection for federated learning with uncertain resources in mobile networks. In *IEEE Globecom Workshops*, pages 1–6, 2020.

[Zhang et al., 2013] Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, pages 3321–3363, 2013.
A.1 Clients Stratification

Stratification is the process of dividing clients of the population into homogeneous strata before selection.

**How to determine the number of strata?** Naturally, we hope to get the stratification outcome with the best effectiveness. As a primary consideration, the number of strata has a great impact on the effectiveness of stratification. A better stratification will in turn bring better train convergence and higher test accuracy in Federated Learning. To achieve the maximized effectiveness of stratification, we use Silhouette Coefficient to quantify it under different preset numbers of strata and choose the one with the highest score as the ultimate choice. Silhouette Coefficient $s(i)$ is an evaluation indicator to measure stratification effectiveness,

$$s(i) = \begin{cases} 
1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\
0, & a(i) = b(i) \\
\frac{b(i)}{a(b(i))} - 1, & a(i) > b(i)
\end{cases} \quad (15)$$

which combines two factors of cohesion and separation for evaluation. According to the intra cohesion $a(i)$ and inter separation $b(i)$ of $X_i$ in the $h^{th}$ stratum, where $X_i$ is a vector used to represent the client features.

$$a(i) = \frac{1}{N_h} \sum_{j=1, j \neq h}^{N_h} |X_i - X_j|, \quad (16)$$

$$b(i) = \min \{b(i_1), b(i_2), \ldots, b(i_H)\}, \quad (17)$$

where $b(i_k)$ can be defined as

$$b(i_k) = \frac{1}{N_k} \sum_{j=1, j \neq h}^{N_k} |X_i - X_j|, k \in \{1,2,\ldots,H\}. \quad (18)$$

The expectation of all samples is defined as the Silhouette Coefficient:

$$\text{Silhouette Coefficient} = \frac{1}{N} \sum_{i=1}^{N} s(i). \quad (19)$$

The value of Silhouette Coefficient is between $[-1, 1]$. The higher value indicates better effectiveness of stratification, indicating that the distance between samples from the same stratum are closer and from different stratum are further.

**Stratification.** In particular, when we conduct handwritten number recognition task on MNIST, every client has their local image with different label “0” to “9”. Naturally, the number of the digit “0” to “9” will generate a 10 dimension vector, which can be regarded as the $X_i$ proposed in A.1. We can use the K-Means algorithm based on the 10 dimension vector above for the stratification of the entire set $\mathcal{K}$. Under various data distribution, all 10 dimension vector will have different features and we can get different numbers of strata with best stratification effectiveness measured by Silhouette Coefficient. Besides, there are also many alternative methods to do stratification for specific scenes in practical application. For example, in the recommendation system, we can employ the method based on client’s characteristics such as gender, age, occupation and adopt the idea of a decision tree to stratify clients to form a training data set manually. In general, the method to do stratification is an open issue and should be adapted to specific application scenes.

A.2 Measurement of Variability

Considering the difficulty of evaluating the accurate numerical standard deviation of a certain stratum, we use Cohesion, denoted as $S(h)$ to reflect the variability of a certain stratum.

$$S(h) = \frac{1}{N_h} \sum_{i=1}^{N_h} a(i) = \frac{1}{N_h^2} \sum_{i=1}^{N_h} \sum_{j=1, j \neq i}^{N_h} |X_i - X_j|. \quad (20)$$

It can be concluded that as the average distance between clients in the stratum is shorter, i.e., the closer it is to 0, the cohesion is higher. Given all of that, we can determine the variability of each stratum according to the cohesion directly. Specially, high cohesion indicates small variability among clients in one stratum. Same to the stratification, the determination of variability is also an open issue and should do some modifications according to the realistic situations.

A.3 Dirichlet Distribution Illustration

Given a series of values of the concentration parameter $\alpha$ for the Dirichlet distribution, datasets with different degrees of heterogeneity can be generated. In particular, higher values of $\alpha$ lead to a more uniform distribution, indicating that each client has an almost equally weighted combination of labels. Lower values of $\alpha$ imply weights concentrated more heavily on only one of the labels, or more extreme label membership. Table 3 is an example of Dirichlet distribution used in the experiments. As shown in the Table 3, the data distribution on each client is different, and in the case of extreme non-IID, i.e., $\alpha \to 0$, most of the data are concentrated under only one label, while the amount of others is almost zero.

B. Proof of Theorem

B.1 Proof of Theorem 2 Convergence

The full proof of the convergence guaranteed for FedAvg with simple random selection can be found in [Li et al., 2019]. We emphasize that the convergence analysis of FedAvg is NOT our contribution, we just follow [Li et al., 2019] to verify the improvement of our selection scheme in terms of the convergence rate. Following [Li et al., 2019] We first present necessary assumptions and extra notations that we used to prove the convergence of FedAvg with random client selection.

**Assumptions.** The convergence of FedAvg with random sampling scheme has been derived in [Li et al., 2019]. The proof relies on the assumptions as follows.

**Assumption 1 (L-Smooth).** $\forall \textbf{v}$ and $\textbf{w}$, $k = 1, \ldots, N$
$$F_k(\textbf{v}) \leq F_k(\textbf{w}) + \langle \textbf{v} - \textbf{w}, \nabla F_k(\textbf{w}) \rangle + \frac{\lambda}{2} \|\textbf{v} - \textbf{w}\|_2^2$$
where the $\textbf{v}$ and $\textbf{w}$ are different model parameters.

**Assumption 2 (Strongly Convex).** $\forall \textbf{v}$ and $\textbf{w}$, $k = 1, \ldots, N$
$$F_k(\textbf{v}) \geq F_k(\textbf{w}) + \langle \textbf{v} - \textbf{w}, \nabla F_k(\textbf{w}) \rangle + \frac{\lambda}{2} \|\textbf{v} - \textbf{w}\|_2^2$$
where the $\textbf{v}$ and $\textbf{w}$ are different model parameters.
Assumptions 3 and 4 have been given by [Zhang et al., 2013; Stich, 2018; Stich et al., 2018].

Assumption 3 (Bounded Variance). Let $\xi_k^t$ be sampled from the $k$th device's local data uniformly at random. The variance of stochastic gradients in each device is bounded: $\mathbb{E} \left[ \| \nabla F_k \left( w_k^t, \xi_k^t \right) - \nabla F_k \left( w_k^t \right) \|^2 \right] \leq \sigma_k^2$, $\forall k = 1, \ldots, N$.

Assumption 4 (Bounded Expectation). The expectation of stochastic gradients in squared norm is bounded by $G^2$, i.e., $\mathbb{E} \left[ \| \nabla F_k \left( w_k^t, \xi_k^t \right) \|^2 \right] \leq G^2$, $\forall k = 1, \ldots, N, t = 1, \ldots, T - 1$.

Additional Notation. We assume that FedAvg always activates all devices at the beginning of each round and then uses the parameters maintained in only a few sampled devices to produce the next-round parameter. This updating scheme is equivalent to the original. Then the update of FedAvg with partial devices active can be described as: for all $k \in [N]$, 

$$ v_{t+1}^k = w_t^k - \eta \nabla F_k \left( w_t^k, \xi_t^k \right) \tag{21} $$

$$ w_{t+1}^k = \begin{cases} v_{t+1}^k & \text{if } t+1 \notin I_E \\ S_{t+1} \text{ and average } \{v_{t+1}^i\}_{i \in S_{t+1}} & \text{if } t+1 \in I_E \end{cases} \tag{22} $$

Here, an additional variable $v_{t+1}^k$ is introduced to represent the immediate result of one step SGD update from $w_t^k$. We interpret $w_{t+1}^k$ as the parameter obtained after communication steps. Let $F^*$ and $F_k^*$ be the minimum values of $F$ and $F_k$, respectively. We use the term $\Gamma = F^* - \sum_{k=1}^N p_k F_k^*$ for quantifying the degree of non-iid. If the data are iid, then $\Gamma$ goes to zero as the number of samples grows. If the data are non-iid, then $\Gamma$ is nonzero, and its magnitude reflects the heterogeneity of the data distribution.

Theorem 3 (FedAvg Convergence). Let Assumptions 1 to 4 hold and $L, \mu, \sigma_k, G$ be defined therein. $E$ denote the number of local iteration before aggregation. Choose $\kappa = \frac{L}{\mu} \gamma = \max \{8 \kappa, E \}$ and the learning rate $\eta_t = \frac{2}{\mu (\gamma + t)}$. Then FedAvg with Random Sampling satisfies

$$ \mathbb{E} \left[ F \left( w_T \right) \right] - F^* \leq \frac{\kappa}{\gamma + T - 1} \left( \frac{2 (B + C)}{\mu} + D \right) $$

where

$$ B = \sum_{k=1}^N p_k^2 \sigma_k^2 + 6 \mu \Gamma + 8 (E - 1)^2 G^2, \quad C = \frac{4}{m} E^2 G^2 $$

$$ D = \frac{\mu}{2} \mathbb{E} \| w_1 - w^* \|^2 $$

Theorem 3 shows the convergence of FedAvg with random selection. The proof can be found in [Li et al., 2019].

Remark. [Li et al., 2019] use Lemma 1-3 to bound the error of FedAvg with full participation corresponding to the term $B$ and $D$, while using Lemma 4-5 to bound the variance resulting from the random selection. The dependency between the Lemma and Theorem 3 indicates that the proof the Lemma 4-5 is enough to maintain the convergence bound. We next show that stratified sampling also satisfies Lemma 4 and 5. As a result, FedAvg with stratified selection maintains the same convergence bound with random sampling.

Proof of Lemma 4 and 5

**Lemma 4 (Unbiased sampling scheme).** If $t+1 \in I_E$, for optimal stratified sampling with $S_t = \{i_1, \ldots, i_m\} \subset [N]$ we have

$$ \mathbb{E} \left[ w(S_t) \right] = w(K) $$

**Proof.**

$$ \mathbb{E} \left[ w(S_t) \right] = \mathbb{E}_{S_t} \sum_{k=1}^m w_{i_k} = m \mathbb{E}_{S_t} w_{i_t} = m \sum_{k=1}^N p_k w_k \tag{23} $$

\[\square\]
Lemma 5 (Bounding the variance of $w(S_t)$). For $t + 1 \in \mathcal{I}$, assume that $\eta_t$ is non-increasing and $\eta_t \leq 2\eta_{t+1}$ for all $t \geq 0$. We have the following result assuming $p_1 = p_2 = \cdots = p_{m_1} = \frac{1}{N}$, the expected difference between $\mathbf{v}_{t+1}$ and $\mathbf{w}_{t+1}$ is bounded by

$$
\mathbb{E}_{S_t} \| \mathbf{v}_{t+1} - \mathbf{w}_{t+1} \|^2 \leq \frac{4}{K} \eta_t^2 E^2 G^2
$$

Proof. Select a subset $S_{t+1}$ by stratified random selection, we have $\mathbf{w}_{t+1} = \frac{1}{m} \sum_{i=1}^m \mathbf{v}_{t+1}^{i}$. Taking expectation over $S_{t+1}$,

$$
\mathbb{E}_{S_t} \| \mathbf{w}_{t+1} - \mathbf{v}_{t+1} \|^2 = \mathbb{E}_{S_t} \frac{1}{m^2} \sum_{i=1}^m \| \mathbf{v}_{t+1}^{i} - \mathbf{v}_{t+1} \|^2
$$

$$
= \frac{1}{m} \sum_{i=1}^m \| \mathbf{v}_{t+1}^{i} - \mathbf{v}_{t+1} \|^2
$$

where the first equality follows from $\mathbf{v}_{t+1}^{i}$ are independent and unbiased. Lemma (4) verify the unbiased property of the Estimator. Since $t + 1 \in \mathcal{I}_E$, we know that the time $t_0 = t - E + 1 \in \mathcal{I}_E$ is the communication time, which implies $\{\mathbf{w}_{t_0}\}_{k=1}^N$ is identical. Then

$$
\sum_{k=1}^N p_k \left( \mathbf{v}_{t+1}^k - \mathbf{v}_{t_0} \right) = \mathbf{v}_{t+1} - \mathbf{v}_{t_0}, \text{i.e., } \sum_{k=1}^N p_k = 1
$$

The last inequality also rely on the normalization property of the probability distribution that we use to sample.

$$
\sum_{k=1}^N p_k \left( \mathbf{v}_{t+1}^k - \mathbf{v}_{t_0} \right) = \mathbf{v}_{t+1} - \mathbf{v}_{t_0}, \text{i.e., } \sum_{k=1}^N p_k = 1
$$

The last inequality also results from

$$
\mathbb{E} \| \mathbf{x} - \mathbb{E} \mathbf{x} \| ^2 \leq \mathbb{E} \| \mathbf{x} \|^2
$$

Similarly, we have

$$
\mathbb{E}_{S_t} \| \mathbf{w}_{t+1} - \mathbf{v}_{t+1} \|^2
$$

$$
\leq \frac{1}{m} \sum_{k=1}^N p_k \mathbb{E} \| \mathbf{v}_{t+1}^k - \mathbf{w}_{t_0} \|^2
$$

$$
\leq \frac{1}{m} \sum_{k=1}^N p_k \mathbb{E} \| \mathbf{v}_{t+1}^k - \mathbf{w}_{t_0} \|^2
$$

$$
\leq \frac{1}{m} \sum_{k=1}^N p_k \sum_{t_0} \mathbb{E} \| \eta_t \nabla F_k (w_t^k, \zeta_t^k) \|^2
$$

$$
\leq \frac{1}{m} E^2 \eta_t^2 G^2 \leq \frac{4}{m} \eta_t^2 E^2 G^2
$$

where in the last inequality we use the fact that $\eta_t$ is non-increasing and $\eta_t \leq 2\eta_{t+1}$.

\[ \qed \]

B.2 Proof of Theorem 1 Variance Reduction

Additional Notation. Divide the population consisting of $N$ clients into $H$ strata.

- $N_h$ denotes the number of clients in $h$th stratum, $s.t. \sum_{h=1}^H N_h = N$
- $m_h$ denotes the number of sampled clients from the $h$th stratum
- $m$ denotes the sample size, $s.t. \sum_{h=1}^H m_h = m$
- $w_h$ denotes the model update $w$ of the $i$th client in the $h$th stratum ($i = 1, \ldots, N_h; h = 1, \ldots, H$)
- $w = \sum_{i=1}^{m_h} \frac{w_h}{m}$ is the sampled averaged model update of the $h$th stratum
- $w_{st} = \frac{1}{N} \sum_{h=1}^H N_h w_h$ is the overall sampled averaged model update
- $S_h = \sum_{i=1}^{N_h} \| w_{h,i} - w_h \|^2$
- $s_h = \sum_{i=1}^{m_h} \| w_{h,i} - w_h \|^2$
- $Q_h = \frac{N_h}{N}$ is the proportion of clients in the $h$th stratum
- $q_h = \frac{m_h}{m}$ is the proportion of sampled clients in the $h$th stratum

Proof of Theorem 3. Simple Random Selection. We assume that each observation has the same variance $\sigma^2$. Then we get

$$
V(w_{ran}) = \mathbb{E} \| w - W(K) \|^2
$$

$$
= \frac{1}{m^2} \mathbb{E} \sum_{i=1}^m \| w_i - W(K) \|^2
$$

$$
= \frac{1}{m^2} \mathbb{E} \sum_{i=1}^m \| w_i - W(K) \|^2
$$

$$
+ \frac{1}{m^2} \mathbb{E} \sum_{i=1}^m \sum_{j \neq i} \| w_i - W(K) \| \| w_j - W(K) \|
$$

$$
= \frac{1}{m^2} \sum_{i=1}^m \| w_i - W(K) \|^2
$$

$$
+ \frac{1}{m^2} \sum_{i=1}^m \sum_{j \neq i} \| w_i - W(K) \| \| w_j - W(K) \|
$$

$$
= \frac{1}{m^2} \sum_{i=1}^m \sigma^2 + \frac{K}{m^2}
$$

$$
= \frac{N - 1}{Nm} \sigma^2 + \frac{K}{m^2}
$$
Where $K = \sum_{i}^{m} \sum_{j \neq i}^{m} \mathbb{E} \| w_i - \mathbf{W}(\mathbf{K}) \| \| w_j - \mathbf{W}(\mathbf{K}) \|$ assuming that each observation has variance $\sigma^2$. In order to find $K$, consider:

\[
\mathbb{E} \| w_i - \mathbf{W}(\mathbf{K}) \| \| w_j - \mathbf{W}(\mathbf{K}) \| = \frac{1}{N(N-1)} \sum_{k=1}^{N} \sum_{k \neq \ell}^{N} \| w_k - \mathbf{W}(\mathbf{K}) \| \| w_\ell - \mathbf{W}(\mathbf{K}) \| \tag{44}
\]

\[
\left\| \sum_{k=1}^{N} (w_k - \mathbf{W}(\mathbf{K})) \right\|^2 = \sum_{k=1}^{N} \| w_k - \mathbf{W}(\mathbf{K}) \|^2 + \sum_{k} \sum_{k \neq \ell} \| w_k - \mathbf{W}(\mathbf{K}) \| \| w_\ell - \mathbf{W}(\mathbf{K}) \| \tag{45}
\]

simplify it, we will get

\[
0 = (N - 1)S^2 + \sum_{k} \sum_{k \neq \ell} \| w_k - \mathbf{W}(\mathbf{K}) \| \| w_\ell - \mathbf{W}(\mathbf{K}) \| \tag{46}
\]

\[
\frac{1}{N(N-1)} \sum_{k} \sum_{k \neq \ell} \| w_k - \mathbf{W}(\mathbf{K}) \| \| w_\ell - \mathbf{W}(\mathbf{K}) \| \tag{47}
\]

\[
= \frac{1}{N(N-1)} \left[ -(N - 1)S^2 \right] \tag{48}
\]

\[
= \frac{S^2}{N} \tag{49}
\]

Thus $K = -m(m - 1) \frac{S^2}{N}$ and substitute the value of $K$, the variance of $w_{ran}$ is

\[
V(w_{ran}) = \frac{N - 1}{N m} S^2 - \frac{1}{m^2} m(m - 1) \frac{S^2}{N} \tag{50}
\]

If $N$ is infinite (large enough), we can get

\[
V(w_{ran}) = \frac{N - m}{N m} S^2 \tag{51}
\]

\[
= \left( \frac{1}{m} - \frac{1}{N} \right) S^2 \cong \frac{S^2}{m} \tag{52}
\]

**Proportional Allocation.** Under proportional allocation

\[
m_h = \frac{N_h}{m} \tag{53}
\]

\[
\text{and we have}
\]

\[
V(w_{st}) = \sum_{i=1}^{H} Q_i^2 \text{Var}(w_i) + \sum_{i \neq j}^{H} \sum_{j=1}^{m_i} Q_i Q_j \text{Cov}(w_i, w_j) \tag{54}
\]

\[
\text{Cov}(w_i, w_j) = 0, i \neq j \tag{55}
\]

\[
\text{Var}(w_i) = \frac{N_i - m_i}{N_i m_i} S_i^2 \tag{56}
\]

\[
\text{where}
\]

\[
S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (w_{ij} - W_i)^2 \tag{57}
\]

Thus

\[
V(w_{st}) = \sum_{h=1}^{H} \left( \frac{N_h - m_h}{N_h m_h} \right) Q_h^2 S_h^2 \tag{58}
\]

Therefore we can get

\[
V(w_{prop}) = \sum_{h=1}^{H} \left( \frac{N_h - m_h}{N_h} \right) \left( \frac{N_h}{N} \right)^2 \frac{S_h^2}{N_h} \tag{59}
\]

\[
= \frac{N - m}{N m} \sum_{h=1}^{H} Q_h S_h^2 \tag{60}
\]

\[
\approx \frac{\sum_{h=1}^{H} N_h S_h^2}{mN} \tag{61}
\]

**Optimal Allocation.** Under optimal allocation

\[
m_h = \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \cdot m \tag{62}
\]

We can get

\[
V(w_{opt}) = \sum_{h=1}^{H} \left( \frac{1}{m_h} - \frac{1}{N} \right) Q_h^2 S_h^2 \tag{63}
\]

\[
= \sum_{h=1}^{H} Q_h^2 S_h^2 - \sum_{h=1}^{H} \frac{Q_h^2 S_h^2}{N_h} \tag{64}
\]

\[
= \sum_{h=1}^{H} \left[ Q_h^2 S_h^2 \left( \frac{\sum_{h=1}^{H} N_h S_h}{mN S_h} \right) \right] - \sum_{h=1}^{H} \frac{Q_h^2 S_h^2}{N_h} \tag{65}
\]

\[
= \sum_{h=1}^{H} \left[ \frac{1}{m} \cdot \frac{N_h S_h}{N^2} \left( \frac{\sum_{h=1}^{H} N_h S_h}{mN S_h} \right) \right] - \sum_{h=1}^{H} \frac{Q_h^2 S_h^2}{N_h} \tag{66}
\]

\[
= \frac{1}{m} \left( \sum_{h=1}^{H} \frac{N_h S_h}{N} \right)^2 - \sum_{h=1}^{H} \frac{Q_h^2 S_h^2}{N_h} \tag{67}
\]

\[
= \frac{1}{m} \left( \sum_{h=1}^{H} Q_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^{H} Q_h S_h^2 \tag{68}
\]

\[
= \frac{1}{N^2} \left( \sum_{h=1}^{H} N_h S_h \right)^2 - \frac{1}{N^2} \sum_{h=1}^{H} N_h S_h^2 \tag{69}
\]

\[
= \frac{1}{m N^2} \left( \sum_{h=1}^{H} N_h S_h \right)^2 \tag{70}
\]
Based on all above, we have these equations below when approximations are used.

\[ V(w_{\text{ran}}) = \frac{N - m}{N m} S^2 \approx \frac{S^2}{m} \quad (74) \]

\[ V(w_{\text{prop}}) = \frac{N - m}{N} \sum_{h=1}^{H} \frac{N_h S_h^2}{m N} \approx \frac{1}{m N} \sum_{h=1}^{H} N_h S_h^2 \quad (75) \]

\[ V(w_{\text{opt}}) = \frac{1}{N^2} \left( \sum_{h=1}^{H} N_h S_h \right)^2 - \frac{1}{N^2} \sum_{h=1}^{H} N_h S_h^2 \quad (76) \]

\[ \approx \frac{1}{m N^2} \left( \sum_{h=1}^{H} N_h S_h \right)^2 \quad (77) \]

**Relationship.** In order to compare \(V(w_{\text{ran}})\) and \(V(w_{\text{prop}})\), we first attempt to express \(S^2\) as a function of \(S_h^2\).

\[(N - 1) S^2 = \sum_{h=1}^{H} \sum_{i=1}^{m_h} \|w_{h,i} - W(K)\|^2 \]
\[= \sum_{h=1}^{H} \sum_{i=1}^{m_h} \|w_{h,i} - W_h\|^2 \quad (78) \]
\[+ \sum_{h=1}^{H} N_h \|W_h - W(K)\|^2 \quad (79) \]
\[= \sum_{h=1}^{H} (N_h - 1) S_h^2 + \sum_{h=1}^{H} N_h \|W_h - W(K)\|^2 \quad (80) \]

\[\frac{N - 1}{N} S^2 = \sum_{h=1}^{H} \frac{N_h - 1}{N} S_h^2 + \sum_{h=1}^{H} \frac{N_h}{N} \|W_h - W(K)\|^2 \quad (81) \]

We assume that \(N_h\) is large enough to permit the approximation for simplification

\[\frac{N_h - 1}{N_h} \approx 1 \quad \text{and} \quad \frac{N - 1}{N} \approx 1 \quad (83) \]

Thus

\[ S^2 = \sum_{h=1}^{H} \frac{N_h}{N} S_h^2 + \sum_{h=1}^{H} \frac{N_h}{N} \|W_h - W(K)\|^2 \quad (84) \]

Therefore

\[ V(w_{\text{ran}}) = \frac{S^2}{m} = \frac{\sum_{h=1}^{H} N_h S_h^2}{m N} \quad (85) \]
\[+ \frac{\sum_{h=1}^{H} N_h \|W_h - W(K)\|^2}{m N} \quad (86) \]
\[= V(w_{\text{prop}}) + \frac{\sum_{h=1}^{H} N_h \|W_h - W(K)\|^2}{m N} \quad (87) \]

which shows that

\[V(w_{\text{prop}}) \leq V(w_{\text{ran}}) \quad (88)\]

Unless \(W_h = W(K)\) for every \(h\). By definition of optimal allocation, we must have \(V(w_{\text{opt}}) \leq V(w_{\text{prop}})\). The difference is

\[V(w_{\text{prop}}) = V(w_{\text{opt}}) + \frac{1}{m N} \sum_{h=1}^{H} N_h (S_h - \bar{S})^2 \quad (89)\]

where \(\bar{S} = \sum_{h=1}^{H} N_h S_h\) This shows that

\[V(w_{\text{opt}}) \leq V(w_{\text{prop}}) \quad (90)\]

Unless \(S_h = \bar{S}\) for every \(h\), i.e., the strata have equal variability. Therefore, we get

\[V(w_{\text{opt}}) \leq V(w_{\text{prop}}) \leq V(w_{\text{ran}}) \quad (91)\]

\[\square\]
C Additional Experiments

C.1 Influence of the Sampling Ratio on MNIST

With the sampling ratio $q$ decreasing, the performance of all sampling schemes becomes unacceptable except ours.

**Figure 5.** Impact of sampling ratio $q$ on the performance. We compare stratified sampling with importance sampling, simple random sampling on MNIST under non-IID, and set parameters $q \in \{0.1, 0.2, 0.3, 0.5, 1.0\}$, $N = 100$, $nSGD = 50$, $\eta = 0.01$, $B = 50$.

**Figure 6.** Impact of sampling ratio $q$ on the performance. We compare stratified sampling with importance sampling, simple random sampling on FMNIST under non-IID, and set parameters $q \in \{0.1, 0.2, 0.3, 0.5, 1.0\}$, $N = 100$, $nSGD = 50$, $\eta = 0.01$, $B = 50$. 
C.2 Evaluation of FedAvg and FedProx

To verify that our sampling scheme has good compatibility, we evaluate FedProx with different sampling schemes on both MNIST and FMNIST. As shown in Figure 7 and Figure 8, compared with existing sampling schemes, our sampling scheme achieves a better convergence rate and stability on the two FL algorithms under various conditions.

Figure 7. Result on the MNIST under FedAvg and FedProx with $\mu = 0.01$. We compare stratified sampling with importance sampling, cluster sampling and simple random sampling and set parameters $q = 0.1$, $N = 100$, $nSGD = 50$, $\eta = 0.01$, $B = 50$.

Figure 8. Result on the FMNIST under FedAvg and FedProx with $\mu = 0.01$. We compare stratified sampling with importance sampling, cluster sampling and simple random sampling and set parameters $q = 0.1$, $N = 100$, $nSGD = 50$, $\eta = 0.01$, $B = 50$. 
C.3 Influence of Data Heterogeneity on CIFAR-10

Figure 9 and Figure 10 show that, as the degree of heterogeneity increases, the performance of other sampling schemes becomes unacceptable. However, our sampling scheme still maintains acceptable performance and has not been greatly affected in such cases of extreme heterogeneity.

Figure 9. Impact of the Heterogeneity on the performance. We compare stratified sampling with importance sampling, simple random sampling on CIFAR-10, use a Dirichlet Distribution with $\alpha \in \{0.01, 0.001\}$ and Shard under FedAvg and FedProx with $\mu = 0.015$, and set parameters $q = 0.1$, $N = 100$, $nSGD = 80$, $\eta = 0.05$, $B = 50$.

Figure 10. Impact of the Heterogeneity on the performance. We compare stratified sampling with cluster sampling, simple random sampling on CIFAR-10, use a Dirichlet Distribution with $\alpha \in \{0.01, 0.001\}$ and Shard under FedAvg and FedProx with $\mu = 0.015$, and set parameters $q = 0.1$, $N = 100$, $nSGD = 80$, $\eta = 0.05$, $B = 50$. 
C.4 Extra Experiments on FedNova
We carry out extra experiments on FedNova, a modified FL algorithm, to further verify the compatibility of our sampling scheme. We use all datasets mentioned above and take different distributions into consideration. As illustrated in Figure 11, our sampling scheme achieves superb performance on FedNova, especially under heterogeneity.

Figure 11. Result on the MNIST, FMNIST and CIFAR10 under FedNova. We compare stratified sampling with simple random sampling, use a Dirichlet Distribution with $\alpha = 0.001$ and Shard and set parameters $q = 0.1$, $N = 100$, $nSGD = 80$, $\eta = 0.05$, $B = 50$.

C.5 Extra Result on CIFAR-10
Besides Table 2 shown in section 6, we present extra result on CIFAR-10.

Table 4: Extra Result: Final test accuracy of multiple FL algorithm with different sampling strategies on CIFAR-10 with different degrees of heterogeneity, i.e., $\alpha = 0.01$, $\alpha = 0.001$ and Shard.

| Method      | CIFAR-10 | \(\alpha = 0.01\) | \(\alpha = 0.001\) | Shard |
|-------------|----------|---------------------|---------------------|-------|
| (a) FedAvg [McMahan et al., 2017] | | | | |
| Random      | 44.3 ±0.8 | 27.2 ±1.0 | 21.8 ±1.8 |
| Importance  | 51.7 ±5.5 | 41.8 ±10.7 | 26.7 ±8.9 |
| Cluster     | 53.7 ±5.7 | 42.1 ±5.3 | 32.0 ±5.9 |
| Stratified  | 54.1 ±1.7 | 47.7 ±2.8 | 47.1 ±1.9 |
| (b) FedProx [Li et al., 2020] | | | | |
| Random      | 39.0 ±0.7 | 24.0 ±2.6 | 21.8 ±1.5 |
| Importance  | 50.6 ±6.2 | 40.9 ±8.2 | 26.2 ±8.3 |
| Cluster     | 51.7 ±5.6 | 42.7 ±5.1 | 28.7 ±5.2 |
| Stratified  | 53.0 ±1.4 | 46.4 ±1.0 | 46.3 ±1.2 |
| (c) FedNova [Wang et al., 2020] | | | | |
| Random      | 46.9 ±4.3 | 36.5 ±6.8 | 21.8 ±8.1 |
| Stratified  | 51.8 ±1.5 | 48.3 ±3.2 | 39.2 ±2.5 |