Minimizing Label Distribution Shift through Client Selection in Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

Federated learning (FL) is a distributed learning approach that trains a machine learning model using data from multiple clients without exposing each client's local data. Although standard FL assumes a uniform test label distribution, many practical applications, such as medical diagnosis, involve known or estimated test label distributions that often deviate from uniformity. Existing client selection strategies are either limited by the need to estimate clients' label distribution or restricted to uniform test label distributions, and furthermore, they lack rigorous theoretical guarantees in non-convex settings. To bridge this gap, we propose FedMLS, a novel client selection strategy that minimizes the Kullback-Leibler (KL) divergence between the test label distribution and the time-averaged label distribution of the selected clients. We establish convergence guarantees for FedMLS under test label distribution shift and extend our analysis to scenarios with loss reweighting. Moreover, we design an iterative mechanism leveraging Fully Homomorphic Encryption (FHE) to implement FedMLS without exposing clients' sensitive label statistics. Extensive experiments on the CIFAR-100, iNaturalist-User-120k, and Shakespeare datasets demonstrate that FedMLS consistently outperforms state-of-the-art baselines, achieving higher test accuracy and up to 10.7× faster convergence than FedAvg (4.6× when combined with FedIR loss reweighting).

1 Introduction

Federated learning (FL) (McMahan et al., 2017; Bonawitz et al., 2019) is a distributed learning framework where multiple devices collaboratively train a machine learning model without exposing their data. A centralized FL system comprises a central server and data-owning clients; each client maintains the privacy of its local dataset while participating in the training process. The framework finds diverse applications, such as AI-based medical analysis (Li et al., 2019) and keyboard query suggestion (Yang et al., 2018; Hard et al., 2018).

In a centralized FL system, due to limited bandwidth or hardware availability, only a subset of clients is selected for participation in each round. Standard client selection methods aim to train a model that minimizes an empirical loss under the assumption of a uniform or aggregated global label distribution (Mohri et al., 2019; Zhang et al., 2023). However, real-world deployment environments—such as regional medical clinics with distinct disease prevalence rates or local retail stores with localized product demands—frequently exhibit substantial label distribution shifts from uniform distributions or global training mixtures (Kairouz and McMahan, 2021). Standard FL models underperform in these scenarios because they optimize for a mismatched global training distribution rather than the specific test label distribution (Jiang and Lin, 2023; Bao et al., 2023).

In many institutional settings (e.g., regional demographic prevalence or store inventory ratios), the test label statistics p_{test} can often be estimated (Luo and Wu, 2022; Breitholtz et al., 2025) or explicitly specified by end-users (Lai et al., 2021; Jeong and Lee, 2024). Current approaches—ranging from local loss reweighting (Hsu et al., 2020) to client selection (Jeong and Lee, 2024)—incorporate p_{test} directly during training and demonstrate empirical performance gains under distribution shifts. However, the theoretical interplay

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

among test label distribution shifts, local loss reweighting, and client selection remains underexplored. Moreover, existing client selection mechanisms either rely on estimates of client label distributions (Jeong and Lee, 2024) or depend on privacy-preserving encryption schemes restricted to uniform test distributions (Zhang et al., 2023).

To address these challenges, we propose FedMLS, a client selection strategy that leverages a prior target test label distribution to select participating clients to minimize the KL divergence between the test label distribution and the time-averaged label distribution of the selected clients. Our approach is mathematically grounded in a novel convergence analysis that quantifies the impact of test label distribution shift and loss reweighting on non-convex optimization objectives. Moreover, we design an iterative mechanism that leverages fully homomorphic encryption (FHE), enabling label-distribution-aware client selection without exposing sensitive client label statistics.

Experiments on the CIFAR-100, iNaturalist-User-120k, and Shakespeare datasets show that FedMLS consistently outperforms state-of-the-art client selection baselines, achieving superior test accuracy across varying degrees of test label distribution shift. Additionally, FedMLS achieves up to 10.7× faster convergence than FedAvg. These results validate our theoretical findings and highlight the practical advantage of leveraging prior knowledge of the label distribution in both client selection and loss reweighting.

We make theoretical and practical contributions as follows:

1. **The FedMLS Algorithm:** We propose a client selection strategy designed to minimize the KL divergence between the test and time-averaged label distribution from selected clients across rounds. This approach ensures the training process accounts for the test label distribution shift while remaining computationally efficient without combinatorial optimization.
2. **Novel Convergence Analysis:** We provide a novel convergence analysis for FedMLS under test label distribution shift. We also derive a non-convex convergence bound for FL under arbitrary client selection and loss reweighting. To the best of our knowledge, this is the first analysis that jointly characterizes client selection and loss reweighting under test label shift.
3. **Privacy-Preserving Mechanism:** We design a privacy-preserving implementation for FedMLS to enable client selection without exposing sensitive clients' data and label information.

2 Related Work

We divide the literature review into two lines of work: (i) client selection in FL and (ii) the problem of test label distribution shift.

Early approaches select clients by uniform sampling (McMahan et al., 2017; Bonawitz et al., 2019), or the selection probability is related to the size of local datasets (Li et al., 2020). The data in the FL setting are typically non-IID (not independent and identically distributed), leading to suboptimal convergence (Zhao et al., 2018). Several heuristic client selection strategies have been proposed to improve the performance of FL algorithms in non-IID settings. For example, Oort (Lai et al., 2021) and Power-of-Choice (Cho et al., 2022) favor clients with higher training loss. The work in Ribero and Vikalo (2020) selects clients with significant weight updates, while other works in (Zeng et al., 2021; Chen et al., 2022) favor clients with bigger gradient squared norms. Although these methods improve FL performance, their test performance is not guaranteed. Other approaches (Yang et al., 2021; Duan et al., 2021; Zhang et al., 2023) that focus on the client dataset's label distribution improve performance by selecting clients to achieve a uniform label distribution. Recently, strategies have been proposed to select clients whose local distributions are close to the training distribution (Liu et al., 2023) and to align selected clients with the target distribution (Jeong and Lee, 2024). Nonetheless, these recent client selection approaches rely on estimated client label distributions and lack rigorous analytical performance guarantees.

The test label distribution shift problem has been thoroughly explored in the centralized machine learning (Lipton et al., 2018; Azizzadenesheli et al., 2019; Alexandari et al., 2020;

Šipka et al., 2022). However, this problem has only recently been considered in FL settings. The works in (Jiang and Lin, 2023; Tan et al., 2023) have shown the effect of test distribution shift on the performance of FL algorithms. The work in Mohri et al. (2019) proposes a framework for adapting the model to any mixture of client distributions by solving a minimax formulation. Other approaches (Jiang and Lin, 2023; Bao et al., 2023; Xu and Huang, 2023; Amosy et al., 2024) adapt a trained global model to the target data distributions at test time using model calibration techniques. It is worth noting that some works also consider test distribution shift at the feature level (Reisizadeh et al., 2020; Tan et al., 2023). Assuming prior knowledge of the test label distribution, existing training-stage solutions apply loss reweighting (Hsu et al., 2020) or client selection (Jeong and Lee, 2024). However, there are currently no theoretical characterizations of the joint effect of loss reweighting and client selection on model convergence bounds under test label distribution shifts.

3 Federated Learning System

We consider a learning system designed for a K-class classification problem. The system comprises a central server and N clients. Each client $n \in \{1, \dots, N\}$ possesses a local dataset $\mathcal{D}_n$ of size $|\mathcal{D}_n|$. We define the size fraction as $r_n = |\mathcal{D}_n| / \sum_{m=1}^{N} |\mathcal{D}_m|$, such that $\sum_{n=1}^{N} r_n = 1$. The local data is characterized by a label distribution $p_n = (p_{n,1}, \dots, p_{n,K})$, where $p_{n,k}$ denotes the probability that a sample in $\mathcal{D}_n$ belongs to class k.

Our objective is to minimize the test loss function $f_{test}(w) = \sum_{k=1}^{K} p_{test,k} f_{test,k}(w)$ which is bounded below by $f_{\min}$, where $p_{test} = (p_{test,1}, \dots, p_{test,K})$ represents the known and fixed target test label distribution and $f_{test,k}(w)$ is the loss function associated with class k. This objective differs from conventional FL frameworks, which assume an unknown or uniform test distribution when designing client selection strategies and analyzing convergence (Mohri et al., 2019). Zhang et al. (2023) considers a specific case with $p_{test,k} = \frac{1}{K}$ for all k.

In our considered system, the model is trained over multiple rounds $t \in \{0, 1, 2, \dots\}$. During each round t, the server selects M clients to participate, as indicated by a binary vector $S(t) = (S_1(t), \dots, S_N(t))$, where $S_n(t) = 1$ if client n is selected and 0 otherwise. The server broadcasts the global model $w(t)$ to these selected clients, who independently perform I iterations of local stochastic gradient descent (SGD) (Stich, 2018; Haddadpour et al., 2019).

Each selected client minimizes a local loss function $f_n(w) = \sum_{k=1}^{K} \lambda_{n,k} p_{n,k} f_{n,k}(w)$, where $\lambda_{n,k}$ is the weighting coefficient for class k data at client n such that $\sum_{k=1}^{K} \lambda_{n,k} p_{n,k} = 1$. This coefficient can be utilized to address label shift; for example, in Hsu et al. (2020), setting $\lambda_{n,k} = p_{test,k}/p_{n,k}$ reweighs the local loss to align with the test label distribution. This optimization produces an updated model $w_n(t, I) = w(t) - \eta \frac{1}{I} \sum_{i=0}^{I-1} g_n(w_n(t, i))$, where g_n denotes the stochastic gradient of the local loss and η_t is the learning rate. The server then aggregates models from selected clients to produce the next global model $w(t+1) = \frac{1}{M} \sum_{n:S_n(t)=1} w_n(t, I)$. The process repeats for T rounds.

This system addresses practical scenarios in which the distribution of test labels can be estimated. Later, in Section 4, we propose our client selection method, FedMLS, that aligns the label distribution of time-averaged selected clients with this target test label distribution. Notice that the system also reduces to the Federated Averaging (FedAvg) algorithm (McMahan et al., 2017) when the training and test distributions are identical, M clients are randomly selected based on $(r_n)_{n=1}^{N}$, and $\lambda_{n,k} = 1$ for all n, k.

4 FedMLS Client Selection Strategy

4.1 Label distribution shift

We first derive the label distribution resulting from rounds of client selections. Let $\bar{S}_n(T) = \frac{1}{T} \sum_{t=0}^{T-1} S_n(t)$ be the fraction of rounds in which client n is selected during the first T rounds. The time-averaged label distribution for client n and class k is $\bar{S}_n(T) p_{n,k}$. Let $\bar{S}(T) = (\bar{S}_1(T), \dots, \bar{S}_N(T))$ be the time-averaged selection of all clients. We define the label distribution shift between this time-averaged label distribution and the target test label

distribution as their Kullback–Leibler (KL) divergence:

$$D(\bar{S}(T), \lambda) = -\sum_{k=1}^{K} p_{test,k} \log \left(\frac{\frac{1}{M} \sum_{n=1}^{N} \bar{S}_n(t) \lambda_{n,k} p_{n,k}}{p_{test,k}} \right), \tag{1}$$

where $\lambda = (\lambda_1, \ldots, \lambda_N)$ and $\lambda_n = (\lambda_{n,1}, \ldots, \lambda_{n,K})$ is the vector of weighting coefficients.

This shift is the forward KL divergence and has two advantages. i) The forward KL exhibits mean-seeking (mass-covering) behavior (Bishop, 2006). Specifically, the label distribution from client selections will capture the majority classes in the test dataset while not overlooking the minority classes. ii) The forward KL is convex (Chan et al., 2022), resulting in efficient computation of selection decisions. We use this shift to design our client selection strategy.

4.2 Client selection algorithm

Our client selection strategy minimizes the shift between the target test label distribution and the time-averaged label distribution from selected clients. We formulate the selection strategy as the following stochastic optimization problem:

$$\begin{aligned} \text{Minimize} \quad & \limsup_{T \to \infty} \mathbb{E}\left[D\left(\bar{S}(T), \lambda \right) \right] \\ \text{Subject to} \quad & \liminf_{T \to \infty} \mathbb{E}\left[\bar{S}_n(T) \right] \geq cMr_n \quad \forall n \in \{1, \ldots, N\} \\ & \textstyle\sum_{n=1}^{N} S_n(t) = M \qquad\qquad \forall t \in \{0, 1, 2, \ldots\}. \end{aligned} \tag{2}$$

The problem minimizes the expected label distribution shift caused by time-averaged client selections. The first constraint enforces a minimum selection rate of cMr_n, where r_n ensures that clients with larger datasets are selected more frequently to maintain statistical representativeness, and $c \in [0, 1]$ is a hyperparameter. The last constraint ensures that exactly M clients are selected per round. Let D_λ^{opt} denote the optimal value of problem (2) for a given reweighting λ. Notice that the selection strategy minimizes a function of time averages, which attains a value no greater than the average of the function due to the shift's convexity, $D\left(\frac{1}{T} \sum_{t=0}^{T-1} S(t), \lambda \right) \leq \frac{1}{T} \sum_{t=0}^{T-1} D(S(t), \lambda)$.

We propose a client selection strategy based on the drift-plus-penalty framework (Neely, 2010) that solves the above optimization problem. In each round, the strategy solves the following problems with a trade-off parameter $V > 0$.

$$\text{(Selection problem)} \quad \begin{aligned} \text{Maximize} \quad & \textstyle\sum_{n=1}^{N} \left[A_n(t) + Z_n(t) \right] S_n(t) \\ \text{Subject to} \quad & \textstyle\sum_{n=1}^{N} S_n(t) = M \end{aligned} \tag{3}$$

$$\text{(Auxiliary problem)} \quad \begin{aligned} \text{Minimize} \quad & \textstyle\sum_{n=1}^{N} A_n(t) a_n(t) + V D(a(t), \lambda) \\ \text{Subject to} \quad & a(t) = (a_1(t), \ldots, a_N(t)) \in [0, 1]^N, \end{aligned} \tag{4}$$

where A_n and Z_n are internal states and follow the update rules initialized to 0:

$$\begin{aligned} A_n(t+1) &= \max\left[A_n(t) + a_n(t) - S_n(t), 0 \right], \quad n \in \{1, \ldots, N\} \\ Z_n(t+1) &= \max\left[Z_n(t) + cMr_n - S_n(t), 0 \right], \quad n \in \{1, \ldots, N\}. \end{aligned} \tag{5}$$

Intuitively, the auxiliary vector $a(t) \in [0, 1]^N$ is a relaxed version of the selection vector $S(t) \in \{0, 1\}^N$, achieving the minimization of the shift function in a continuous domain. The internal state A_n tracks the selections of client n across rounds to match the relaxation. The internal state Z_n tracks the selections to meet the minimum selection rate in the average sense. The decision $(S(t), a(t))$ from problems (3) and (4) ensures that the tracking works, as shown in Appendix B. Importantly, our selection strategy requires no combinatorial optimization, unlike previous works (Perazzone et al., 2022; Zhang et al., 2023).

Integrating our selection strategy into the system described in Section 3 yields Algorithm 1. Performance of the algorithm with respect to the stochastic optimization (2) is given in

Algorithm 1 FedMLS: Federated Learning with Test Label Distribution Matching

1: **Initialize:** global model $w(0)$, states $A_n(0) = Z_n(0) = 0\ \forall n$, total rounds T, and $t = 0$.
2: **for** round $t = 0$ **to** $T - 1$ **do**
3: Obtain the selection $S(t)$ from solving problems (3) and (4). Update states as (5).
4: **for** every selected client n with $S_n(t) = 1$ **do**
5: $w_n(t, 0) \leftarrow w(t)$
6: **for** local iteration $i = 0$ **to** $I - 1$ **do**
7: $w_n(t, i+1) \leftarrow w_n(t, i) - \frac{\eta}{I} g_n(w_n(t, i))$
8: **end for**
9: **end for**
10: $w(t+1) \leftarrow \frac{1}{M} \sum_{n:S_n(t)=1} w_n(t, I)$
11: **end for**

Theorem 1, proven in Appendix B. The theorem implies the shift is within N/V of the optimal D_λ^{opt}, and that the minimum selection rate is achieved asymptotically. Corollary 1, proven in Appendix B, states that the shift gets close to the optimal after a transient period.

Theorem 1. *The client selection strategy with parameter $V > 0$ in Algorithm 1 solves the optimization problem (2) with the following guarantees where $D_\lambda^{min} = \min_{a \in [0,1]^N} D(a, \lambda)$.*

$$\textit{(Optimality gap)} \quad \limsup_{T\to\infty} \mathbb{E}\left[D(\bar{S}(T), \lambda)\right] \leq D_\lambda^{opt} + N/V$$

$$\textit{(Asymptotic selection rate)} \quad \liminf_{T\to\infty} \mathbb{E}\left[\bar{S}_n(T)\right] \geq cMr_n \quad \forall n \in \{1, \ldots, N\}$$

$$\textit{(Finite-time selection rate)} \quad \mathbb{E}\left[\bar{S}_n(T)\right] \geq cMr_n - \sqrt{\frac{2N + 2V(D_\lambda^{opt} - D_\lambda^{min})}{T}}$$

Corollary 1. *For every $\epsilon > 0$, there exists a finite number of rounds $T'_{V,\lambda,\epsilon}$ such that*

$$\mathbb{E}\left[D(\bar{S}(t), \lambda)\right] < D_\lambda^{opt} + \frac{N}{V} + \epsilon \quad \textit{for all } t \geq T'_{V,\lambda,\epsilon}.$$

4.3 Privacy-preserving mechanism

While the privacy of individual clients' local data is preserved by local computation at each client (Algorithm 1, Line 7), the client selection strategy must solve the auxiliary problem (4) (Algorithm 1, Line 3). This objective is a function of the clients' label distributions, i.e., $p_{n,k}$ for all n and k. This section solves the auxiliary problem (4) based on a gradient method with a fully homomorphic encryption scheme (Gentry, 2010; Brakerski et al., 2014), such that individual label distributions are never decrypted. Note that the FHE scheme supports homomorphic addition and scalar multiplication, utilizing a public key for encryption $\mathrm{Enc}(\cdot)$ and a private key for decryption $\mathrm{Dec}(\cdot)$. For brevity, the algorithm and its implementation details are provided in Appendix E.2, and each iteration is illustrated in Figure 1.

Our central server consists of three separate, non-colluding entities: Selector, Oracle, and Service, which never share their data unless specified by our protocol. We assume a semi-honest threat model (Goldreich, 2004) in which all entities strictly adhere to the protocol but may attempt to infer private information from observed data. Since p_n is encrypted, neither the Selector nor the Service obtains p_n. The Oracle receives P masked by additive noise. Theorem 2 shows that the Service cannot uniquely reconstruct clients' label distribution from the aggregates it observes whenever two clients share a support (Appendix E.1).

Theorem 2. *Suppose the Service obtains the aggregate $P \in \mathbb{R}^{K\times O}$ over O observations, and $P = \mathbf{p}A$ where $\mathbf{p} = [p_1, \ldots, p_N] \in \mathbb{R}^{K\times N}$ is the stack of client distribution column vectors and $A = [a_1, \ldots, a_O] \in [0,1]^{N\times O}$ is the stack of auxiliary variable column vectors. If two clients $i \neq j$ hold the same set of non-zero classes, $\mathrm{supp}(p_i) = \mathrm{supp}(p_j)$, in different proportions, $p_i \neq p_j$, then the adversary's reconstruction problem admits more than one solution. Thus, $\mathbf{p}$ is not identifiable from P for any O observations.*

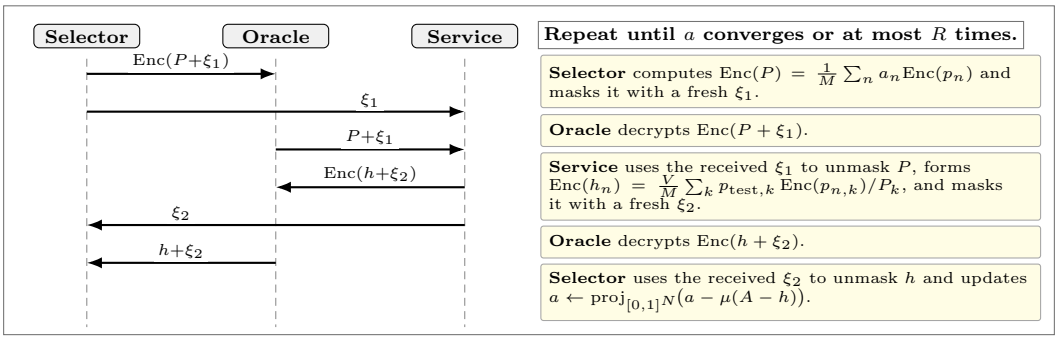


Figure 1: One inner iteration of Algorithm 2; the iteration index is omitted. Clients upload $\mathrm{Enc}(p_n)$ to the Selector and Service once at setup. The Oracle decrypts only blinded quantities, the Service observes only the aggregate P.

5 Analytical Performance

We first introduce the assumptions required for our convergence analysis.

Assumption 1 (L-smoothness)**.** *The gradients of the class-wise loss functions for clients and the test set are Lipschitz continuous. Specifically,* $\|\nabla f_{n,k}(w) - \nabla f_{n,k}(w')\| \leq L_{n,k}\|w-w'\|$ *for every* $n \in \{1,\dots,N\}$, *every* $k \in \{1,\dots,K\}$, *and* $\|\nabla f_{test}(w) - \nabla f_{test}(w')\| \leq L_{test}\|w-w'\|$ *for all* w, w'. *We define* $L_n = \sum_{k=1}^K p_{n,k}L_{n,k}$ *and* $L = \max\{\max_n L_n, L_{test}\}$.

Assumption 2 (Bounded Variance)**.** *The stochastic gradient* g_n *of client* n*'s loss function is an unbiased estimator, satisfying* $\mathbb{E}_{\xi_n}[g_n(w,\xi_n)] = \nabla f_n(w)$ *for all* w. *Furthermore, its variance is uniformly bounded by a non-negative constant* σ^2*:* $\mathbb{E}_{\xi_n}\left[\|g_n(w,\xi_n) - \nabla f_n(w)\|^2\right] \leq \sigma^2, \quad \forall w, \forall n \in \{1,\dots,N\}$.

Assumption 3 (Bounded Gradient Dissimilarity)**.** *Given non-negative weights* $(p_{test,k})_{k=1}^K$ *such that* $\sum_{k=1}^K p_{test,k} = 1$, *there exist constants* $\beta^2 \geq 1$ *and* $\gamma^2 \geq 0$ *that satisfy:* $\sum_{k=1}^K p_{test,k}\|\nabla f_{test,k}(w)\|^2 \leq \beta^2\left\|\sum_{k=1}^K p_{test,k}\nabla f_{test,k}(w)\right\|^2 + \gamma^2, \quad \forall w$.

Assumption 4 (Class-wise Gradient Similarity)**.** *The squared Euclidean distance between the gradients of the test loss and client* n*'s local loss for class* k *is bounded:* $\|\nabla f_{test,k}(w) - \nabla f_{n,k}(w)\|^2 \leq \varphi_{n,k}^2$ *for all* w, *for every* $n \in \{1,\dots,N\}$ *and every* $k \in \{1,\dots,K\}$. *We further define the global bound* $\varphi^2 = \max_{n,k}\varphi_{n,k}^2$.

While Assumptions 1, 3, and 4 are formulated relative to the test loss function, such conditions are widely adopted in the literature for local loss functions from local datasets (Wang et al., 2020; Li et al., 2020; Cho et al., 2022; Zhang et al., 2023). Similarly, Assumption 2 provides standard bounds for stochastic gradients (Wang et al., 2020; Li et al., 2020; Cho et al., 2022).

We now state the convergence of FedMLS with any loss reweighting, proven in Appendix C.

Theorem 3. *Suppose Assumptions 1 to 4 hold,* $\epsilon > 0$, *and* $H = \frac{1}{2} - \beta^2 C R_{\max}(M) > 0$. *The convergence bound of the FedMLS for* $T > T'_{V,\lambda,\epsilon}$ *communication rounds, and learning*

rate $\eta = \frac{\eta_{max}}{\sqrt{T}}$ *where* $\eta_{max} = \frac{1}{2L\sqrt{I(I-1)\left(1+\frac{8\beta^2 B}{H}\right)}}$ *is:*

$$\min_{0\le t\le T-1} \mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \le \frac{1}{HT}\left[16\eta_{max}^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I}+2\varphi^2+2B\gamma^2\right)\right] + \frac{1}{H\sqrt{T}}\left[\frac{2\eta_{max}\sigma^2 L}{MI}+\frac{2\left(f_{test}(w(0))-f_{min}\right)}{\eta_{max}}\right] + \frac{4}{H}\left[\varphi^2+\gamma^2 C\left(D_\lambda^{opt}+\frac{N}{V}+\epsilon+L_D\sqrt{M\left(1-\frac{M}{N}\right)}\right)\right].$$

where $B = \max_n \sum_{k=1}^K \frac{(\lambda_{n,k}p_{n,k})^2}{p_{test,k}}$, $C = \max_k \frac{1}{p_{test,k}^2}$, *and* L_D *is the Lipschitz constant of* D.

Theorem 3 shows that the vanishing part decays at rate $\mathcal{O}(1/\sqrt{T})$, while the non-vanishing part depends on the label distribution shift: FedMLS drives the shift to within $N/V+\epsilon$ of the optimum D_λ^{opt}, whereas the term $L_D\sqrt{M(1-M/N)}$ does not depend on the selection policy.

We further present a tighter convergence bound for a class of federated learning algorithms with any selection policy, including FedMLS, under FedIR loss reweighting technique. To the best of our knowledge, Theorem 4 is the first convergence result with the test label distribution shift and loss reweighting under any selection policy. Its proof is in Appendix D.

Theorem 4. *Suppose Assumptions 1 to 4 hold, and* $\lambda_{n,k} = \frac{p_{test,k}}{p_{n,k}}$ *where* $p_{n,k} > 0$ *for all* n, k. *The convergence bound of the federated learning algorithm with an arbitrary client selection algorithm with learning rate* $\eta = \frac{\eta_{max}}{\sqrt{T}}$ *and* $\eta_{max} = \frac{1}{2L\sqrt{I(I-1)(1+8\beta^2)}}$ *is as follows:*

$$\min_{0\le t\le T-1} \mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \le \frac{1}{T}\left[32\eta_{max}^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I}+2\varphi^2+2\gamma^2\right)\right] + \frac{1}{\sqrt{T}}\left[\frac{4\eta_{max}\sigma^2 L}{MI}+\frac{4\left(f_{test}(w(0))-f_{min}\right)}{\eta_{max}}\right]+8\varphi^2,$$

6 Experiments

We evaluate our proposed client selection strategy, FedMLS, on four datasets: (1) CNN for image classification on a non-IID partition of CIFAR-100 (Krizhevsky, 2009), (2) MobileNetV2 (Sandler et al., 2018) for image classification on iNaturalist-User-120k and Landmarks-User-160k datasets (Hsu et al., 2020), (3) LSTM for next-character prediction on the Shakespeare dataset (Caldas et al., 2018). We compare FedMLS against four baseline client selection strategies, categorized by their primary objectives. First, we include **FedAvg** (McMahan et al., 2017) and **Power-of-Choice** (Cho et al., 2022) as standard federated learning baselines. Second, we evaluate against distribution-aware strategies: **Fed-CBS** (Zhang et al., 2023), which aligns selection with a uniform label distribution, and **On-Demand FL** (Jeong and Lee, 2024), which targets a known test label distribution. Additional evaluations on the Landmarks-User-160k dataset, effects of hyperparameters, robustness analysis, and ablation study are presented in Appendix G.

Experimental setup. In each round, following Hsu et al. (2020), a server designates $M = 10$ clients for participation. Training terminates after 2,000 rounds. We report the test accuracy as the mean and standard error over 8 runs with different random seeds. We calculate the convergence speedup (Conv.) as the ratio of the number of rounds required by a baseline method to reach the final accuracy of FedAvg to the number of rounds required by FedMLS. We reweigh the local loss by applying the weighting coefficient $\lambda_{n,k} = \frac{p_{test,k}}{p_{n,k}}$ for each client n and class k. This reweighting is called **FedIR** (Hsu et al., 2020). Following Cho et al. (2022); Zhang et al. (2023), we use a Dirichlet distribution $Dir(\alpha_{test})$, where $\alpha_{test} \in \{10, 1, 0.1\}$, to simulate degrees of test label distribution shift. Further details regarding datasets, partitioning process, and experimental parameters are provided in Appendix F.

6.1 Performance of FedMLS

Test accuracy improvement under distribution shift. As summarized in Table 1, without loss reweighting, FedMLS consistently achieves the highest test accuracy across all experimental configurations. On CIFAR-100, FedMLS significantly outperforms all competing baselines across all heterogeneity levels ($p < 0.05$). Similarly, on the naturally partitioned iNaturalist-User-120k and Shakespeare datasets, FedMLS remains the top performer across all degrees of test label shift ($\alpha_{\text{test}} \in \{0.1, 1, 10\}$). Specifically, under extreme shift ($\alpha_{\text{test}} = 0.1$), FedMLS outperforms the standard FedAvg baseline by 14.63% on CIFAR-100, 2.28% on iNaturalist-User-120k, and 1.56% on Shakespeare. These results demonstrate the effectiveness of our client selection, FedMLS, in realistic, large-scale federated environments.

Convergence speedup. FedMLS consistently reaches the target accuracy faster than all other strategies on CIFAR-100 and iNaturalist-User-120k datasets, while standard baselines and distribution-aware baselines often struggle to converge to the target. On the CIFAR-100 dataset, FedMLS converges at least 2.2× faster than the most competitive baselines that achieve the target. In the naturally partitioned iNaturalist-User-120k dataset, FedMLS reaches the target up to 3.1× faster, while distribution-aware baselines (Fed-CBS and On-Demand FL) fail to reach the target.

Table 1: Test accuracy (Acc.) and convergence speedup factor (Conv.) across varying test distribution shifts (α_{test}). Results represent the mean ± standard error across independent runs. N/A in the Conv. column indicates that the method fails to reach the target accuracy. All performance improvements of FedMLS over baselines are statistically significant ($p < 0.05$). **Bold** values denote the best-performing method for each configuration.

		CIFAR-100		**iNaturalist-User-120k**		**Shakespeare**	
α_{test}	**Method**	**Acc.**	**Conv.**	**Acc.**	**Conv.**	**Acc.**	**Conv.**
10	FedAvg	21.55 ± .42	3.5×	36.08 ± .23	2.2×	38.89 ± .18	1.0×
	Power-of-Choice	20.72 ± .43	5.0×	36.08 ± .22	3.1×	38.86 ± .15	1.0×
	Fed-CBS	22.90 ± .33	5.2×	24.23 ± .12	N/A	38.16 ± .16	N/A
	On-Demand FL	10.30 ± .34	N/A	25.16 ± .38	N/A	21.46 ± .28	N/A
	FedMLS (ours)	**24.82 ± .25**	–	**38.53 ± .14**	–	**39.36 ± .10**	–
1	FedAvg	24.82 ± .32	2.4×	35.38 ± .42	2.4×	38.42 ± .21	1.1×
	Power-of-Choice	24.00 ± .60	3.4×	34.98 ± .08	2.5×	38.40 ± .14	1.1×
	Fed-CBS	24.90 ± .75	2.2×	23.31 ± .22	N/A	37.85 ± .14	N/A
	On-Demand FL	13.35 ± .25	N/A	21.06 ± .73	N/A	22.75 ± .32	N/A
	FedMLS (ours)	**27.57 ± .44**	–	**37.89 ± .07**	–	**39.68 ± .15**	–
0.1	FedAvg	21.85 ± .63	10.7×	36.66 ± .09	1.9×	35.77 ± .53	1.4×
	Power-of-Choice	20.61 ± .71	13.8×	36.91 ± .05	1.9×	35.83 ± .24	1.4×
	Fed-CBS	20.52 ± .35	8.4×	24.16 ± .18	N/A	34.76 ± .26	N/A
	On-Demand FL	34.92 ± .33	2.7×	14.47 ± .45	N/A	20.47 ± .32	N/A
	FedMLS (ours)	**36.48 ± .40**	–	**38.94 ± .16**	–	**37.33 ± .41**	–

6.2 Performance under FedIR Loss Reweighting

Effect of loss reweighting. Combining FedMLS with FedIR achieves the highest test accuracy across all evaluated settings, significantly outperforming ($p < 0.05$) both standard baselines and existing distribution-aware selection strategies. For instance, under moderate distribution shift ($\alpha_{\text{test}} = 1$), FedMLS improves accuracy over the strongest baseline by 2.62% on CIFAR-100, 1.96% on iNaturalist-User-120k, and 1.13% on Shakespeare. In addition, FedMLS demonstrates substantial convergence acceleration. Across all settings, baselines require up to 5.4× more communication rounds to match target performance, or never reach the target. These gains indicate that the distribution-aware selection in FedMLS and the loss reweighting in FedIR complement each other in mitigating performance degradation from the test label shift.

6.3 Robustness analysis

To simulate noisy or misspecified test label distributions, we perturb the true test label distribution p_{test} using additive zero-mean Gaussian noise. We inject Gaussian noise $\epsilon \sim \mathcal{N}(0, \gamma\mu)$ into the true distribution, where γ controls the noise magnitude and μ represents the mean class probability. The resulting probabilities are subsequently clipped to non-negative values and normalized.

Table 2: Test accuracy (Acc._{IR}) and convergence speedup factor (Conv.) under FedIR loss reweighting across varying test distribution shifts (α_{test}). Results represent the mean ± standard error across independent runs. N/A in the Conv. column indicates that the method fails to reach the target accuracy. All performance improvements of FedMLS over baselines are statistically significant ($p < 0.05$). **Bold** values denote the best-performing method for each configuration.

		CIFAR-100		**iNaturalist-User-120k**		**Shakespeare**	
α_{test}	**Method**	**Acc._{IR}**	**Conv.**	**Acc._{IR}**	**Conv.**	**Acc._{IR}**	**Conv.**
10	FedAvg	22.48 ± .27	1.8×	35.78 ± .10	1.8×	42.17 ± .11	1.4×
	Power-of-Choice	22.55 ± .33	1.4×	36.50 ± .36	1.8×	43.22 ± .09	1.5×
	Fed-CBS	22.75 ± .34	1.4×	23.48 ± .16	N/A	41.03 ± .13	N/A
	On-Demand FL	10.72 ± .30	N/A	24.01 ± .26	N/A	22.68 ± .26	N/A
	FedMLS (ours)	**24.95 ± .44**	–	**38.50 ± .17**	–	**44.41 ± .12**	–
1	FedAvg	29.62 ± .26	2.4×	39.18 ± .11	1.9×	46.04 ± .10	1.4×
	Power-of-Choice	29.59 ± .23	2.0×	39.77 ± .08	1.7×	47.00 ± .14	1.4×
	Fed-CBS	29.34 ± .49	2.6×	24.46 ± .08	N/A	44.95 ± .26	N/A
	On-Demand FL	13.99 ± .27	N/A	20.80 ± .34	N/A	28.95 ± .25	N/A
	FedMLS (ours)	**32.24 ± .33**	–	**41.73 ± .14**	–	**48.13 ± .24**	–
0.1	FedAvg	41.92 ± .57	4.6×	50.02 ± .22	1.6×	62.14 ± .49	1.9×
	Power-of-Choice	42.04 ± .51	5.4×	48.25 ± .03	N/A	62.38 ± .59	1.4×
	Fed-CBS	43.06 ± .24	3.4×	27.56 ± .11	N/A	61.60 ± .36	2.3×
	On-Demand FL	35.66 ± .34	N/A	14.72 ± .08	N/A	37.13 ± 1.02	N/A
	FedMLS (ours)	**44.61 ± .32**	–	**51.44 ± .12**	–	**64.36 ± .43**	–

Table 3: Test accuracy (%) under noisy and misspecified test label distributions on CIFAR-100 across varying Dirichlet parameters (α_{test}).

Method	$\alpha_{\text{test}} = 10$	$\alpha_{\text{test}} = 1$	$\alpha_{\text{test}} = 0.1$
FedAvg (no noise)	21.55 ± 0.42	24.82 ± 0.32	21.85 ± 0.63
Power-of-Choice (no noise)	20.72 ± 0.43	24.00 ± 0.60	20.61 ± 0.71
Fed-CBS (no noise)	22.90 ± 0.33	24.90 ± 0.75	20.52 ± 0.35
On-Demand FL (no noise)	10.30 ± 0.34	13.35 ± 0.25	34.92 ± 0.33
FedMLS (no noise)	24.82 ± 0.25	27.57 ± 0.44	**36.48 ± 0.40**
FedMLS ($\gamma = 0.1$)	24.54 ± 0.45	27.31 ± 0.28	36.38 ± 0.31
FedMLS ($\gamma = 0.5$)	24.36 ± 0.48	27.11 ± 0.11	29.72 ± 0.97
FedMLS ($\gamma = 1.0$)	24.21 ± 0.35	27.56 ± 0.46	28.94 ± 0.42

Results and Robustness Analysis. Table 3 demonstrates classification performance on CIFAR-100 under varying levels of test label skew (α_{test}) and noisy estimation settings. At mild noise levels ($\gamma = 0.1$), performance degradation is negligible across all Dirichlet splits ($< 0.3\%$). Even under severe Gaussian noise ($\gamma = 1.0$), accuracy drops by less than 1% for mild-to-moderate skews ($\alpha_{\text{test}} \in \{1, 10\}$), confirming that client selection remains effective despite inaccurate prior estimates. Under severe skew ($\alpha_{\text{test}} = 0.1$), however, noise with $\gamma \geq 0.5$ lowers accuracy to 28.9–29.7%, below On-Demand FL, indicating that an accurate estimate of p_{test} matters most when the shift is large.

7 Conclusion

This paper investigates the challenge of test label distribution shift in federated learning, where the target test label distribution often deviates from the training mixture or a uniform distribution. We propose FedMLS, a distribution-aware client selection strategy that minimizes the Kullback–Leibler (KL) divergence between the target test label distribution and the time-averaged distribution of selected clients. We further establish convergence bounds that quantify the impact of such shifts under client selection and loss reweighting. To ensure practical viability, we introduce a privacy-preserving mechanism based on fully homomorphic encryption (FHE), in which no individual label distributions is expose. Extensive experiments on the CIFAR-100, iNaturalist-User-120k, and Shakespeare datasets demonstrate that FedMLS consistently outperforms state-of-the-art baselines, achieving higher test accuracy across all degrees of distribution shift with up to 10.7× faster convergence than FedAvg (4.6× when combined with FedIR loss reweighting).

AI USE STATEMENT

In this work, we used generative AI tools to refine hypotheses and provide feedback on experiment methodology. We use no generative AI tools to generate synthetic data sets, formulate mathematical claims, assist in the writing of proofs, and implement methods. Presentation-wise, we used generative AI tools to create the algorithm illustration in Figure 1, and grammatically check a research paper to improve readability. We have reviewed all AI-assisted work. For example, we correct the figure, and all LLM suggestions on writing are reviewed and corrected. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, April 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html. ISSN: 2640-3498.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards Federated Learning at Scale: System Design, March 2019. URL http://arxiv.org/abs/1902.01046. arXiv:1902.01046 [cs, stat].

Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M. Jorge Cardoso, and Andrew Feng. Privacy-preserving federated brain tumour segmentation. In Heung-Il Suk, Mingxia Liu, Pingkun Yan, and Chunfeng Lian, editors, *Machine Learning in Medical Imaging*, pages 133–141, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32692-0.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *ArXiv*, abs/1812.02903, 2018. URL https://api.semanticscholar.org/CorpusID:54457251.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *ArXiv*, abs/1811.03604, 2018. URL https://api.semanticscholar.org/CorpusID:53207681.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic Federated Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4615–4625. PMLR, May 2019. URL https://proceedings.mlr.press/v97/mohri19a.html. ISSN: 2640-3498.

Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran Chen, and Hai Li. Fed-CBS: A Heterogeneity-Aware Client Sampling Mechanism for Federated Learning via Class-Imbalance Reduction. In *Proceedings of the 40th International Conference on Machine Learning*, pages 41354–41381. PMLR, July 2023. URL https://proceedings.mlr.press/v202/zhang23y.html. ISSN: 2640-3498.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Liangze Jiang and Tao Lin. Test-Time Robust Personalization for Federated Learning, June 2023. URL http://arxiv.org/abs/2205.10920. arXiv:2205.10920 [cs].

Wenxuan Bao, Tianxin Wei, Haohan Wang, and Jingrui He. Adaptive Test-Time Personalization for Federated Learning, October 2023. URL http://arxiv.org/abs/2310.18816. arXiv:2310.18816 [cs].

Jun Luo and Shandong Wu. Fedsld: Federated learning with shared label distribution for medical image classification. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. doi: 10.1109/ISBI52829.2022.9761404.

Adam Breitholtz, Edvin Listo Zec, and Fredrik D. Johansson. Overcoming label shift in targeted federated learning. In *Tiny Titans: The next wave of On-Device Learning for Foundational Models (TTODLer-FM)*, 2025. URL https://openreview.net/forum?id=ZUlrLeW82p.

Fan Lai, Xiangfeng Zhu, Harsha V. Madhyastha, and Mosharaf Chowdhury. Oort: Efficient Federated Learning via Guided Participant Selection. pages 19–35, 2021. ISBN 978-1-939133-22-9. URL https://www.usenix.org/conference/osdi21/presentation/lai.

Isu Jeong and Seulki Lee. On-demand federated learning for arbitrary target class distributions. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3421–3429. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/jeong24a.html.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated Visual Classification with Real-World Data Distribution. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 76–92, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58607-2. doi: 10.1007/978-3-030-58607-2_5.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the Convergence of FedAvg on Non-IID Data, June 2020. URL http://arxiv.org/abs/1907.02189. arXiv:1907.02189 [cs, math, stat].

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated Learning with Non-IID Data. 2018. doi: 10.48550/arXiv.1806.00582. URL http://arxiv.org/abs/1806.00582. arXiv:1806.00582 [cs, stat].

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards Understanding Biased Client Selection in Federated Learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR, May 2022. URL https://proceedings.mlr.press/v151/jee-cho22a.html. ISSN: 2640-3498.

Monica Ribero and Haris Vikalo. Communication-Efficient Federated Learning via Optimal Client Sampling, October 2020. URL http://arxiv.org/abs/2007.15197. arXiv:2007.15197 [cs, stat].

Xiao Zeng, Ming Yan, and Mi Zhang. Mercury: Efficient On-Device Distributed DNN Training via Stochastic Importance Sampling. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, SenSys '21, pages 29–41, New York, NY, USA, November 2021. Association for Computing Machinery. ISBN 978-1-4503-9097-2. doi: 10.1145/3485730.3485930. URL https://dl.acm.org/doi/10.1145/3485730.3485930.

Wenlin Chen, Samuel Horvath, and Peter Richtarik. Optimal Client Sampling for Federated Learning, August 2022. URL http://arxiv.org/abs/2010.13723. arXiv:2010.13723 [cs].

Miao Yang, Ximin Wang, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated Learning with Class Imbalance Reduction. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 2174–2178, August 2021. doi: 10.23919/EUSIPCO54536.2021.9616052. URL https://ieeexplore.ieee.org/document/9616052. ISSN: 2076-1465.

Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-Balancing Federated Learning With Global Imbalanced Data in Mobile Systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, January 2021. ISSN 1045-9219, 1558-2183, 2161-9883. doi: 10.1109/TPDS.2020.3009406. URL https://ieeexplore.ieee.org/document/9141436/.

Zichang Liu, Zhaozhuo Xu, Benjamin Coleman, and Anshumali Shrivastava. One-pass distribution sketch for measuring data heterogeneity in federated learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KMxRQO7P98.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and Correcting for Label Shift with Black Box Predictors. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3122–3130. PMLR, July 2018. URL https://proceedings.mlr.press/v80/lipton18a.html. ISSN: 2640-3498.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts, 2019.

Amr M. Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International*

Conference on Machine Learning, 2020. URL https://api.semanticscholar.org/CorpusID:220249808.

Tomáš Šipka, Milan Šulc, and Jiří Matas. The hitchhiker's guide to prior-shift adaptation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2031–2039, 2022. doi: 10.1109/WACV51458.2022.00209.

Yue Tan, Chen Chen, Weiming Zhuang, Xin Dong, Lingjuan Lyu, and Guodong Long. Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qJJmu4qsL0.

Jian Xu and Shao-Lun Huang. A joint training-calibration framework for test-time personalization with label shift in federated learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4370–4374, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615173. URL https://doi.org/10.1145/3583780.3615173.

O. Amosy, G. Eyal, and G. Chechik. Late to the party? on-demand unlabeled personalized federated learning. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2173–2182, Los Alamitos, CA, USA, jan 2024. IEEE Computer Society. doi: 10.1109/WACV57701.2024.00218. URL https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00218.

Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems*, 33:21554–21565, 2020.

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

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32, 2019.

C Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:531–537, 2006.

Alan Chan, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A Rupam Mahmood, and Martha White. Greedification operators for policy optimization: Investigating forward and reverse kl divergences. *The Journal of Machine Learning Research*, 23(1):11474–11552, 2022.

Michael Neely. *Stochastic network optimization with application to communication and queueing systems*. Morgan and Claypool, 2010.

Jake Perazzone, Shiqiang Wang, Mingyue Ji, and Kevin S Chan. Communication-efficient device scheduling for federated learning using stochastic optimization. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1449–1458. IEEE, 2022.

Craig Gentry. Computing arbitrary functions of encrypted data. *Communications of the ACM*, 53(3):97–105, 2010.

Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory (TOCT)*, 6(3):1–36, 2014.

Oded Goldreich. *Foundations of cryptography, volume 2*. Cambridge university press Cambridge, 2004.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization, July 2020. URL http://arxiv.org/abs/2007.07481. arXiv:2007.07481 [cs, stat].

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konecnỳ, H. B. McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *ArXiv*, abs/1812.01097, 2018. URL `https://api.semanticscholar.org/CorpusID:53701546`.

Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

Walter Rudin. Principles of mathematical analysis. 2021.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification, 2019.

A Preliminary Lemmas

We present the preliminary inequalities, based on Jensen's inequality, which is used throughout our analysis. In particular, it holds for any real vector $(a_1, \ldots, a_N)$ that

$$\left\| \frac{1}{N} \sum_{n=1}^{N} a_n \right\|^2 \leq \frac{1}{N} \sum_{n=1}^{n} \|a_n\|^2 \qquad \text{and} \qquad \left\| \sum_{n=1}^{N} a_n \right\|^2 \leq N \sum_{n=1}^{n} \|a_n\|^2 .$$

We then prove a preliminary lemma for the main Lemma 2.

Lemma 1. *For any scalar a_n and vector v_n for $n \in \{1, \ldots, N\}$, it holds that*

$$\left\| \sum_{n=1}^{N} a_n v_n \right\|^2 \leq \left(\sum_{n=1}^{N} a_n^2 \right) \left(\sum_{n=1}^{N} \|v_n\|^2 \right).$$

Proof. By the Triangle Inequality, we have

$$\left\| \sum_{n=1}^{N} a_n v_n \right\| \leq \sum_{n=1}^{N} \|a_n v_n\| \leq \sum_{n=1}^{N} |a_n| \, \|v_n\| .$$

Squaring both sides and applying the Cauchy-Schwarz Inequality gives

$$\begin{aligned} \left\| \sum_{n=1}^{N} a_n v_n \right\|^2 &\leq \left(\sum_{n=1}^{N} |a_n| \, \|v_n\| \right)^2 \\ &\leq \left(\sum_{n=1}^{N} |a_n|^2 \right) \left(\sum_{n=1}^{N} \|v_n\|^2 \right) \\ &= \left(\sum_{n=1}^{N} a_n^2 \right) \left(\sum_{n=1}^{N} \|v_n\|^2 \right). \end{aligned}$$

□

We now ready to prove the main lemma.

Lemma 2. *Suppose Assumptions 1 to 4 hold and $A = 2\eta^2 L^2 (I-1) I < 1$. If the learning rate is $\eta = \min \left\{ \frac{1}{2L}, \frac{1}{L\sqrt{2(I-1)I}} \right\}$, we have the following bound:*

$$\begin{aligned} &\mathbb{E}[f_{test}(w(t+1))] - \mathbb{E}[f_{test}(w(t))] \\ &\leq \left[4\eta \frac{AB}{1-A} \beta^2 + \eta \beta^2 C \mathbb{E}\left[R(S(t), \lambda \right] - \frac{1}{2}\eta \right] \mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \frac{\eta^2 \sigma^2 L}{MI} \\ &\quad + 2\eta \left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A} + 1 \right) \varphi^2 + \gamma^2 \left(\frac{2AB}{1-A} + \frac{1}{2} C \mathbb{E}\left[R(S(t), \lambda \right] \right) \right] \end{aligned}$$

where

$$R(S(t), \lambda) = \sum_{k=1}^{K} \left(\frac{1}{M} \sum_{n: S_n(t)=1} \lambda_{n,k} p_{n,k} - p_{test,k} \right)^2 ,$$

$$B = \max_n \sum_k \frac{(\lambda_{n,k} p_{n,k})^2}{p_{test,k}}, \text{ and } C = \max_k \frac{1}{p_{test,k}^2}.$$

Proof. We first note that

$$\begin{aligned} w(t+1) - w(t) &= \frac{1}{M} \sum_{n:S_n(t)=1} w_n(t,I) - w(t) \\ &= \frac{1}{M} \sum_{n:S_n(t)=1} \left[w(t) - \eta \frac{1}{I} \sum_{i=0}^{I-1} g_n(w_n(t,i)) \right] - w(t) \\ &= -\eta \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} g_n(w_n(t,i)). \end{aligned}$$

By L-smooth Assumption and Descent Lemma (Bertsekas, 2015, p. 305), we have that

$$\begin{aligned} &\mathbb{E}[f_{test}(w(t+1))] - \mathbb{E}[f_{test}(w(t))] \\ &\leq \mathbb{E}\left[\langle \nabla f_{test}(w(t)), w(t+1) - w(t) \rangle\right] + \frac{L}{2} \mathbb{E}\left[\|w(t+1) - w(t)\|^2\right] \\ &= \underbrace{-\eta \mathbb{E}\left[\left\langle \nabla f_{test}(w(t)), \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} g_n(w_n(t,i)) \right\rangle \right]}_{E_1} \\ &\quad + \underbrace{\frac{\eta^2 L}{2} \mathbb{E}\left[\left\| \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} g_n(w_n(t,i)) \right\|^2 \right]}_{E_2}. \end{aligned}$$

For the expression E_1, it follows that

$$\begin{aligned} &-\eta \mathbb{E}\left[\left\langle \nabla f_{test}(w(t)), \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} g_n(w_n(t,i)) \right\rangle \right] \\ &= -\eta \mathbb{E}\left[\left\langle \nabla f_{test}(w(t)), \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} g_n(w_n(t,i)) - \nabla f_n(w_n(t,i)) \right\rangle \right] \\ &\quad -\eta \mathbb{E}\left[\left\langle \nabla f_{test}(w(t)), \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} \nabla f_n(w_n(t,i)) \right\rangle \right] \\ &= -\eta \mathbb{E}\left[\left\langle \nabla f_{test}(w(t)), \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} \nabla f_n(w_n(t,i)) \right\rangle \right] \\ &= -\frac{1}{2}\eta \mathbb{E}\left[\|\nabla f_{test}(w(t))\|^2\right] - \frac{1}{2}\eta \mathbb{E} \left\| \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} \nabla f_n(w_n(t,i)) \right\|^2 \\ &\quad + \frac{1}{2}\eta \mathbb{E}\left[\left\| \nabla f_{test}(w(t)) - \frac{1}{MI} \sum_{n:S_n(t)=1} \sum_{i=0}^{I-1} \nabla f_n(w_n(t,i)) \right\|^2 \right], \end{aligned}$$

where the second step uses the fact that $\mathbb{E}\left[g_n(w_n(t,i))\right] = \nabla f_n(w_n(t,i))$ for all n.

For the expression E_2, we use Assumption 2 to bound the gradient variance.

$$
\begin{aligned}
&\frac{\eta^2 L}{2}\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1} g_n(w_n(t,i))\right\|^2\right] \\
&\leq \eta^2 L\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}(g_n(w_n(t,i)) - \nabla f_n(w_n(t,i)))\right\|^2\right] \\
&\quad + \eta^2 L\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right] \\
&= \frac{\eta^2 L}{(MI)^2}\left(\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\mathbb{E}\left[\|g_n(w_n(t,i)) - \nabla f_n(w_n(t,i))\|^2\right]\right) \\
&\quad + \frac{\eta^2 L}{(MI)^2}\left(\sum_{\substack{n\neq m\\ i\neq j}}\mathbb{E}\left[\langle g_n(w_n(t,i)) - \nabla f_n(w_n(t,i)), g_m(w_m(t,j)) - \nabla f_m(w(t,j))\rangle\right]\right) \\
&\quad + \eta^2 L\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right] \\
&\leq \frac{\eta^2 L}{(MI)^2}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sigma^2 + \eta^2 L\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right] \\
&= \frac{\eta^2\sigma^2 L}{MI} + \eta^2 L\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right],
\end{aligned}
$$

where the third step uses Assumption 2 and the fact that $\mathbb{E}\left[g_n(w_n(t,i))\right] = \nabla f_n(w_n(t,i))$ for every n and every i.

By substituting expressions E_1 and E_2 back, we get

$$
\begin{aligned}
&\mathbb{E}[f_{test}(w(t+1))] - \mathbb{E}[f_{test}(w(t))] \\
&\leq -\frac{1}{2}\eta\mathbb{E}\left[\|\nabla f_{test}(w(t))\|^2\right] + \eta^2\frac{\sigma^2 L}{MI} + \frac{1}{2}\eta\mathbb{E}\left[\left\|\nabla f_{test}(w(t)) - \frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right] \\
&\quad + \left(\eta^2 L - \frac{1}{2}\eta\right)\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right].
\end{aligned}
$$

If we impose $\eta \leq \frac{1}{2L}$, the bound becomes

$$
\begin{aligned}
&\mathbb{E}[f_{test}(w(t+1))] - \mathbb{E}[f_{test}(w(t))] \\
&\leq -\frac{1}{2}\eta\mathbb{E}\left[\|\nabla f_{test}(w(t))\|^2\right] + \eta^2\frac{\sigma^2 L}{MI} + \frac{1}{2}\eta\mathbb{E}\left[\left\|\nabla f_{test}(w(t)) - \frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right].
\end{aligned}
$$

We now focus on finding the upper bound of the last expression on the right-hand side above.

$$
\begin{aligned}
&\frac{1}{2}\eta\mathbb{E}\left[\left\|\nabla f_{test}(w(t)) - \frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\nabla f_n(w_n(t,i))\right\|^2\right] \\
&= \frac{1}{2}\eta\mathbb{E}\left[\left\|\sum_{k=1}^{K} p_{test,k}\nabla f_{test,k}(w(t)) - \frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\nabla f_{n,k}(w(t,i))\right\|^2\right] \\
&\leq \underbrace{\eta\mathbb{E}\left[\left\|\sum_{k=1}^{K} p_{test,k}\nabla f_{test,k}(w(t)) - \frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\nabla f_{test,k}(w(t))\right\|^2\right]}_{E_3} \\
&\quad + \underbrace{\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{test,k}(w(t)) - \nabla f_{n,k}(w(t,i)))\right\|^2\right]}_{E_4}. \qquad (6)
\end{aligned}
$$

Define $C = \max_k 1/p_{test,k}$. The expression E_3 is bounded using Lemma 1 as follows.

$$
\begin{aligned}
&\eta\mathbb{E}\left[\left\|\sum_{k=1}^{K} p_{test,k}\nabla f_{test,k}(w(t)) - \frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\nabla f_{test,k}(w(t))\right\|^2\right] \\
&= \eta\mathbb{E}\left[\left\|\sum_{k=1}^{K}\left(p_{test,k} - \frac{1}{M}\sum_{n:S_n(t)=1}\lambda_{n,k}p_{n,k}\right)\nabla f_{test,k}(w(t))\right\|^2\right] \\
&\leq \eta\mathbb{E}\left[\sum_{k=1}^{K}\left(p_{test,k} - \frac{1}{M}\sum_{n:S_n(t)=1}\lambda_{n,k}p_{n,k}\right)^2\sum_{k=1}^{K}\|\nabla f_{test,k}(w(t))\|^2\right] \\
&= \eta\mathbb{E}\left[R(S(t),\lambda)\sum_{k=1}^{K}\frac{1}{p_{test,k}}p_{test,k}\|\nabla f_{test,k}(w(t))\|^2\right] \\
&\leq \eta C\mathbb{E}\left[R(S(t),\lambda)\right]\left(\beta^2\|\nabla f_{test}(w(t))\|^2 + \gamma^2\right), \qquad (7)
\end{aligned}
$$

where $R(S(t),\lambda) = \sum_{k=1}^{K}\left(\frac{1}{M}\sum_{n:S_n(t)=1}\lambda_{n,k}p_{n,k} - p_{\text{test},k}\right)^2$. The second step uses Lemma 1, and applying Assumption 3 yields the final step.

Next, we bound the expression E_4.

$$
\begin{aligned}
&\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{test,k}(w(t))-\nabla f_{n,k}(w_n(t,i)))\right\|^2\right]\\
&\leq 2\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{test,k}(w(t))-\nabla f_{n,k}(w(t)))\right\|^2\right]\\
&\quad+2\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{n,k}(w(t))-\nabla f_{n,k}(w_n(t,i)))\right\|^2\right]\\
&\leq 2\eta\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\mathbb{E}\left[\|\nabla f_{test,k}(w(t))-\nabla f_{n,k}(w(t))\|^2\right]\\
&\quad+2\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{n,k}(w(t))-\nabla f_{n,k}(w_n(t,i)))\right\|^2\right]\\
&\leq 2\eta\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\varphi^2\\
&\quad+2\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{n,k}(w(t))-\nabla f_{n,k}(w_n(t,i)))\right\|^2\right]\\
&=2\eta\varphi^2+2\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{n,k}(w(t))-\nabla f_{n,k}(w_n(t,i)))\right\|^2\right],
\end{aligned}
$$

where the second and third steps use Jensen's Inequality and Assumption 3, respectively. We then focus on bounding the last term above as follows:

$$
\begin{aligned}
&2\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{n,k}(w(t))-\nabla f_{n,k}(w_n(t,i)))\right\|^2\right]\\
&\leq 2\eta\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\mathbb{E}\left[\|\nabla f_{n,k}(w(t))-\nabla f_{n,k}(w_n(t,i))\|^2\right]\\
&\leq 2\eta\frac{L^2}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\mathbb{E}\left[\|w(t)-w_n(t,i)\|^2\right],
\end{aligned}
$$

where the first step follows Jensen's Inequality and the second step uses Assumption 1.

We then consider the difference between local models on the right-hand side of the above bound:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|w(t)-w_n(t,i)\right\|^2\right] \\
&=\mathbb{E}\left[\left\|\eta\sum_{j=0}^{i-1}\left[g_n\left(w_n(t,j)\right)-\nabla f_n\left(w_n(t,j)\right)+\nabla f_n\left(w_n(t,j)\right)\right]\right\|^2\right] \\
&\leq 2\eta^2\mathbb{E}\left[\left\|\sum_{j=0}^{i-1}\left(g_n\left(w_n(t,j)\right)-\nabla f_n\left(w_n(t,j)\right)\right)\right\|^2\right]+2\eta^2\mathbb{E}\left[\left\|\sum_{j=0}^{i-1}\nabla f_n\left(w_n(t,j)\right)\right\|^2\right] \\
&=2\eta^2\sum_{j=0}^{i-1}\mathbb{E}\left[\left\|g_n\left(w_n(t,j)\right)-\nabla f_n\left(w_n(t,j)\right)\right\|^2\right]+2\eta^2\mathbb{E}\left[\left\|\sum_{j=0}^{i-1}\nabla f_n\left(w_n(t,j)\right)\right\|^2\right] \\
&\leq 2\eta^2\sigma^2 i+2\eta^2 i\sum_{j=0}^{i-1}\mathbb{E}\left[\left\|\nabla f_n\left(w_n(t,j)\right)\right\|^2\right] \\
&\leq 2\eta^2\sigma^2 i+2\eta^2 i\sum_{j=0}^{I-1}\mathbb{E}\left[\left\|\nabla f_n\left(w_n(t,j)\right)\right\|^2\right],
\end{aligned}
$$

where the third step is similar to E_2 bounding process. Applying Assumption 2 yields the fourth step. It follows that

$$
\begin{aligned}
\frac{1}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[\|w(t)-w_n(t,i)\|^2\right] &\leq 2\eta^2\sigma^2\sum_{i=0}^{I-1}\left(\frac{i}{I}\right)+2\eta^2\sum_{i=0}^{I-1}\left(\frac{i}{I}\right)\sum_{j=0}^{I-1}\mathbb{E}\left[\|\nabla f_n(w_n(t,j))\|^2\right] \\
&=\eta^2\sigma^2(I-1)+\eta^2(I-1)\sum_{i=0}^{I-1}\mathbb{E}\left[\|\nabla f_n(w_n(t,i))\|^2\right] \\
&\leq \eta^2\sigma^2(I-1)+2\eta^2(I-1)\sum_{i=0}^{I-1}\mathbb{E}\left[\|\nabla f_n(w(t))\|^2\right] \\
&\quad+2\eta^2(I-1)\sum_{i=0}^{I-1}\mathbb{E}\left[\|\nabla f_n(w_n(t,i))-\nabla f_n(w(t))\|^2\right] \\
&\leq \eta^2\sigma^2(I-1)+2\eta^2(I-1)I\mathbb{E}\left[\|\nabla f_n(w(t))\|^2\right] \\
&\quad+2\eta^2L^2(I-1)\sum_{i=0}^{I-1}\mathbb{E}\left[\|w(t)-w_n(t,i)\|^2\right],
\end{aligned}
$$

where the last step uses Assumption 1. After rearrangement, we get

$$
\frac{L^2}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[\|w(t)-w_n(t,i)\|^2\right]\leq\frac{\eta^2\sigma^2L^2(I-1)}{1-2\eta^2L^2(I-1)I}+\frac{2\eta^2L^2(I-1)I}{1-2\eta^2L^2(I-1)I}\mathbb{E}\left[\|\nabla f_n(w(t))\|^2\right].
$$

Define $A = 2\eta^2 L^2 (I-1)I$, and we impose that $A < 1$. The above became

$$
\begin{aligned}
&\frac{L^2}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[\|w(t) - w_n(t,i)\|^2\right] \\
&\leq \frac{\sigma^2 A}{2I(1-A)} + \frac{A}{1-A}\mathbb{E}\left[\|\nabla f_n(w(t))\|^2\right] \\
&\leq \frac{\sigma^2 A}{2I(1-A)} + \frac{2A}{1-A}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}(\nabla f_{n,k}(w(t)) - \nabla f_{test,k}(w(t)))\right\|^2\right] \\
&\quad + \frac{2A}{1-A}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\nabla f_{test,k}(w(t))\right\|^2\right] \\
&\leq \frac{\sigma^2 A}{2I(1-A)} + \frac{2A}{1-A}\mathbb{E}\left[\sum_{k=1}^{K}\lambda_{n,k}p_{n,k}\|\nabla f_{n,k}(w(t)) - \nabla f_{test,k}(w(t))\|^2\right] \\
&\quad + \frac{2A}{1-A}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\frac{\lambda_{n,k}p_{n,k}}{\sqrt{p_{test,k}}}\sqrt{p_{test,k}}\nabla f_{test,k}(w(t))\right\|^2\right] \\
&\leq \frac{\sigma^2 A}{2I(1-A)} + \frac{2A}{1-A}\varphi^2 + \frac{2A}{1-A}\mathbb{E}\left[\left\|\sum_{k=1}^{K}\frac{\lambda_{n,k}p_{n,k}}{\sqrt{p_{test,k}}}\sqrt{p_{test,k}}\nabla f_{test,k}(w(t))\right\|^2\right] \\
&\leq \frac{\sigma^2 A}{2I(1-A)} + \frac{2A}{1-A}\varphi^2 + \frac{2A}{1-A}\sum_{k=1}^{K}\left(\frac{(\lambda_{n,k}p_{n,k})^2}{p_{test,k}}\right)\mathbb{E}\left[\sum_{k=1}^{K}p_{test,k}\|\nabla f_{test,k}(w(t))\|^2\right] \\
&= \frac{\sigma^2 A}{2I(1-A)} + \frac{2A}{1-A}\varphi^2 + \frac{2AB}{1-A}\left(\beta^2\|\nabla f_{test}(w(t))\|^2 + \gamma^2\right),
\end{aligned}
$$

where the third to fifth steps use the Jensen Inequality, Assumption 4, and Lemma 1, respectively. In the last step, we define $B = \sum_{k=1}^{K}\frac{(\lambda_{n,k}p_{n,k})^2}{p_{test,k}}$ and apply Assumption 3.

The expression E_4 is finally bounded as

$$
\begin{aligned}
&\eta\mathbb{E}\left[\left\|\frac{1}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\sum_{k=1}^{K}p_{n,k}(\nabla f_{test,k}(w(t)) - \nabla f_{n,k}(w(t,i)))\right\|^2\right] \\
&\leq 2\eta\varphi^2 + 2\eta\frac{L^2}{MI}\sum_{n:S_n(t)=1}\sum_{i=0}^{I-1}\mathbb{E}\left[\|w(t) - w_n(t,i)\|^2\right] \\
&\leq 2\eta\varphi^2 + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \frac{2A}{1-A}\varphi^2 + \frac{2AB}{1-A}\left(\beta^2\|\nabla f_{test}(w(t))\|^2 + \gamma^2\right)\right]. \qquad (8)
\end{aligned}
$$

Combining results from (7) and (8) back to (6), we complete the lemma

$$
\begin{aligned}
&\mathbb{E}[f_{test}(w(t+1))] - \mathbb{E}[f_{test}(w(t))] \\
&\leq -\frac{1}{2}\eta\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \frac{\eta^2\sigma^2 L}{MI} + \eta C\mathbb{E}[R(S(t),\lambda)]\left(\beta^2\|\nabla f_{test}(w(t))\|^2 + \gamma^2\right) \\
&\quad + 2\eta\varphi^2 + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \frac{2A}{1-A}\varphi^2 + \frac{2AB}{1-A}\left(\beta^2\|\nabla f_{test}(w(t))\|^2 + \gamma^2\right)\right] \\
&= \left[4\eta\frac{AB}{1-A}\beta^2 + \eta\beta^2 C\mathbb{E}\left[R(S(t),\lambda] - \frac{1}{2}\eta\right]\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \frac{\eta^2\sigma^2 L}{MI}\right. \\
&\quad \left. + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A} + 1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + \frac{1}{2}C\mathbb{E}\left[R(S(t),\lambda]\right)\right)\right].
\end{aligned}
$$

□

B Proof of Theorem 1

Theorem 1. *The client selection strategy with parameter $V > 0$ in Algorithm 1 solves the optimization problem (2) with the following guarantees where $D_\lambda^{min} = \min_{a\in[0,1]^N} D(a,\lambda)$.*

$$\textit{(Optimality gap)} \quad \limsup_{T\to\infty} \mathbb{E}\left[D(\bar{S}(T),\lambda)\right] \leq D_\lambda^{opt} + N/V$$

$$\textit{(Asymptotic selection rate)} \quad \liminf_{T\to\infty} \mathbb{E}\left[\bar{S}_n(T)\right] \geq cMr_n \quad \forall n \in \{1,\dots,N\}$$

$$\textit{(Finite-time selection rate)} \quad \mathbb{E}\left[\bar{S}_n(T)\right] \geq cMr_n - \sqrt{\frac{2N + 2V(D_\lambda^{opt} - D_\lambda^{min})}{T}}$$

Proof. This proof is provided to make this submission self-contained. Our proof here follows the techniques in (Neely, 2010). Let $\Theta(t) = (A(t), Z(t))$ be a concatenated vector of all internal states. We define the Lyapunov function as

$$L(\Theta(t)) = \frac{1}{2}\sum_{n=1}^{N} A_n(t)^2 + \frac{1}{2}\sum_{n=1}^{N} Z_n(t)^2.$$

Moreover, the Lyapunov drift is defined by $\Delta(\Theta(t)) = \mathbb{E}\left[L(\Theta(t+1)) - L(\Theta(t))\,|\Theta(t)\right]$. Considering the update rules in (5) and using the fact that $(\max(x+y,0))^2 \leq (x+y)^2$ for any reals x and y gives

$$A_n(t+1)^2 - A_n(t)^2 \leq 1 + 2A_n(t)\left[a_n(t) - S_n(t)\right] \qquad \forall n \in \{1,\dots,N\}$$
$$Z_n(t+1)^2 - Z_n(t)^2 \leq 1 + 2Z_n(t)\left[cMr_n - S_n(t)\right] \qquad \forall n \in \{1,\dots,N\}.$$

We then bound the Lyapunov drift using the two inequalities above

$$\begin{aligned}
\Delta(\Theta(t)) &= \mathbb{E}\left[L(\Theta(t+1)) - L(\Theta(t))|\,\Theta(t)\right] \\
&= \mathbb{E}\left[\frac{1}{2}\sum_{n=1}^{N}\left(A_n(t+1)^2 - A_n(t)^2\right) + \frac{1}{2}\sum_{n=1}^{N}\left(Z_n(t+1)^2 - Z_n(t)^2\right)\middle|\,\Theta(t)\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[N + 2\sum_{n=1}^{N} A_n(t)\left[a_n(t) - S_n(t)\right] + N + 2\sum_{n=1}^{N} Z_n(t)\left[cMr_n - S_n(t)\right]\middle|\,\Theta(t)\right] \\
&= N + \sum_{n=1}^{N}\mathbb{E}\left[A_n(t)\left[a_n(t) - S_n(t)\right]|\Theta(t)\right] + \sum_{n=1}^{N}\mathbb{E}\left[Z_n(t)\left[cMr_n - S_n(t)\right]|\Theta(t)\right].
\end{aligned}$$

Adding $V\mathbb{E}\left[D(a(t),\lambda)|\Theta(t)\right]$ on both sides gives

$$\begin{aligned}
&\Delta(\Theta(t)) + V\mathbb{E}\left[D(a(t),\lambda)|\Theta(t)\right] \\
&\leq N + \sum_{n=1}^{N}\mathbb{E}\left[A_n(t)\left[a_n(t) - S_n(t)\right]|\Theta(t)\right] + \sum_{n=1}^{N}\mathbb{E}\left[Z_n(t)\left[cMr_n - S_n(t)\right]|\Theta(t)\right] \\
&\quad + V\mathbb{E}\left[D(a(t),\lambda)|\Theta(t)\right] \\
&= N + \mathbb{E}\left[VD(a(t),\lambda) + \sum_{n=1}^{N} A_n(t)a_n(t)\middle|\,\Theta(t)\right] - \mathbb{E}\left[\sum_{n=1}^{N} S_n(t)\left[A_n(t) + Z_n(t)\right]\middle|\,\Theta(t)\right] \\
&\quad + cMr_n\mathbb{E}\left[\sum_{n=1}^{N} Z_n(t)\middle|\,\Theta(t)\right].
\end{aligned}$$

It is easy to see that our selection strategy in (3) and (4) minimizes the right side of the above bound with respect to $(a(t), S(t))$ every round t. Therefore, the left side with our selection strategy is bounded above by any other (possibly randomized) strategy $(\hat{a}(t), \hat{S}(t))$ as

$$\begin{aligned}
\Delta(\Theta(t)) + V\mathbb{E}\left[D(a(t),\lambda)|\Theta(t)\right] &\leq N + V\mathbb{E}\left[D(\hat{a}(t),\lambda)|\Theta(t)\right] \\
&\quad + \sum_{n=1}^{N}\mathbb{E}\left[A_n(t)\left[\hat{a}_n(t) - \hat{S}_n(t)\right]|\Theta(t)\right] + \sum_{n=1}^{N}\mathbb{E}\left[Z_n(t)\left[cMr_n - \hat{S}_n(t)\right]|\Theta(t)\right]. \quad (9)
\end{aligned}$$

By (Neely, 2010, Thm. 4.5), for any $\varepsilon > 0$, there exists a randomized strategy (a^*, S^*) independent of $\Theta(t)$ such that,

$$\begin{aligned} \mathbb{E}\left[D(a^*, \lambda)\right] &\leq D_\lambda^{opt} - \varepsilon \\ \mathbb{E}\left[a_n^* - S_n^*\right] &\leq \varepsilon \qquad\qquad \forall n \in \{1, \ldots, N\} \\ \mathbb{E}\left[cMr_n - S_n^*\right] &\leq \varepsilon \qquad\qquad \forall n \in \{1, \ldots, N\}. \end{aligned}$$

Substituting the randomized strategy (a^*, S^*) into (9) and taking limit $\varepsilon \to 0$, we have

$$\begin{aligned} \Delta(\Theta(t)) + V\mathbb{E}\left[D(a(t), \lambda)|\Theta(t)\right] &\leq N + V(D_\lambda^{opt} - \varepsilon) + \sum_{n=1}^{N} A_n(t)\varepsilon + \sum_{n=1}^{N} Z_n(t)\varepsilon \\ &= N + VD_\lambda^{opt}. \end{aligned} \tag{10}$$

Taking the expectation on both sides and using the telescoping sum from $t = 0$ to $T - 1$ gives

$$\frac{1}{2}\mathbb{E}\left[\|\Theta(T)\|^2\right] - \frac{1}{2}\mathbb{E}\left[\|\Theta(0)\|^2\right] + V\sum_{t=0}^{T-1}\mathbb{E}\left[D(a(t), \lambda)\right] \leq NT + VD_\lambda^{opt}T. \tag{11}$$

Since all internal states are initially zero, we have $\|\Theta(0)\| = 0$. It follows from $\|\Theta(T)\|^2 \geq 0$ that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[D(a(T))\right] \leq D_\lambda^{opt} + \frac{N}{V}.$$

Recall that $\bar{a}(T) = \frac{1}{T}\sum_{t=0}^{T-1} a(t)$. Applying Jensen's inequality for the convexity of the shift function D yields the following result for all T

$$\mathbb{E}\left[D(\bar{a}(T), \lambda)\right] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[D(a(t), \lambda)\right] \leq D_\lambda^{opt} + \frac{N}{V}.$$

Taking the limsup as $T \to \infty$ gives

$$\limsup_{T\to\infty}\mathbb{E}\left[D(\bar{a}(T), \lambda)\right] \leq D_\lambda^{opt} + N/V. \tag{12}$$

The above describes the performance in terms of the time-average auxiliary vector. Next, we derive the performance from actual selections $\bar{S}(T)$.

Rearranging (11) and using the fact that $\|\Theta(0)\| = 0$ gives the following bound for every round T

$$\mathbb{E}\left[\|\Theta(T)\|^2\right] \leq 2NT + 2VD_\lambda^{opt}T - 2V\sum_{t=0}^{T-1}\mathbb{E}\left[D(a(t), \lambda)\right] \leq 2NT + 2VT(D_\lambda^{opt} - D_\lambda^{min}),$$

where $D_\lambda^{min} = \min_{a\in[0,1]^N} D(a, \lambda)$ is the minimum of the shift function. Since the variance is non-negative $\mathbb{E}\left[\|\Theta(T)\|^2\right] - \mathbb{E}\left[\|\Theta(T)\|\right]^2 \geq 0$, we have that

$$\begin{aligned} \mathbb{E}\left[\|\Theta(T)\|\right]^2 &\leq 2NT + 2VT(D_\lambda^{opt} - D_\lambda^{min}) \\ \frac{\mathbb{E}\left[\|\Theta(T)\|\right]}{T} &\leq \sqrt{\frac{2N + 2V(D_\lambda^{opt} - D_\lambda^{min})}{T}}. \end{aligned} \tag{13}$$

This implies that

$$\frac{\mathbb{E}\left[A_n(T)\right]}{T} \leq \sqrt{\frac{2N + 2V(D_\lambda^{opt} - D_\lambda^{min})}{T}} \quad \forall n \in \{1, \ldots, N\}.$$

Applying the fact that $A_n(t+1) - A_n(t) \geq a_n(t) - S_n(t)$ and telescoping sum gives $A_n(T) - A_n(0) \geq \sum_{t=0}^{T-1}\left[a_n(t) - S_n(t)\right]$. Since $A_n(0) = 0$, the above bound becomes

$$\mathbb{E}\left[\bar{a}_n(T) - \bar{S}_n(T)\right] \leq \sqrt{\frac{2N + 2V(D_\lambda^{opt} - D_\lambda^{min})}{T}}.$$

Taking the limsup as $T \to \infty$ gives

$$\limsup_{T\to\infty} \mathbb{E}\left[\bar{a}_n(T) - \bar{S}_n(T)\right] \leq 0 \quad \forall n \in \{1, \ldots, N\}.$$

Using this with the continuity and entry-wise non-increasing properties of the shift function in (12), we obtain the first part of this theorem

$$\limsup_{T\to\infty} \mathbb{E}\left[D(\bar{S}(T), \lambda)\right] \leq \limsup_{T\to\infty} \mathbb{E}\left[D(\bar{a}(T), \lambda)\right] \leq D_\lambda^{opt} + N/V.$$

For the last part, we can apply the same process for $S_n(t)$ to (13) to obtain

$$\mathbb{E}\left[cMr_n - \bar{S}_n(T)\right] \leq \sqrt{\frac{2N + 2V(D_\lambda^{opt} - D_\lambda^{min})}{T}}$$
$$\limsup_{T\to\infty} \mathbb{E}\left[cMr_n - \bar{S}_n(T)\right] \leq 0 \quad \forall n \in \{1, \ldots, N\}.$$

□

The implication of the optimality gap in Theorem 1 is that $\limsup_{T\to\infty} \mathbb{E}\left[D(\bar{S}(T), \lambda)\right]$ converges into an interval around D_λ^{opt}. We emphasize this in the following corollary that the sequence of $\mathbb{E}\left[D(\bar{S}(t), \lambda)\right]$ stays within an interval after some transient rounds.

Corollary 1. *For every $\epsilon > 0$, there exists a finite number of rounds $T'_{V,\lambda,\epsilon}$ such that*

$$\mathbb{E}\left[D(\bar{S}(t), \lambda)\right] < D_\lambda^{opt} + \frac{N}{V} + \epsilon \quad \textit{for all } t \geq T'_{V,\lambda,\epsilon}.$$

Proof. Fix $\epsilon > 0$. By Theorem 1, $\limsup_{t\to\infty} \mathbb{E}\big[D(\bar{S}(t), \lambda)\big] \leq D_\lambda^{opt} + N/V < D_\lambda^{opt} + N/V + \epsilon$. Since $D_\lambda^{opt} + N/V + \epsilon$ strictly exceeds the limit superior, there exists a finite $T'_{V,\lambda,\epsilon} \in \mathbb{N}$ such that $\mathbb{E}\big[D(\bar{S}(t), \lambda)\big] < D_\lambda^{opt} + N/V + \epsilon$ for all $t \geq T'_{V,\lambda,\epsilon}$ (Rudin, 2021, Thm. 3.17).

□

C Proof of Theorem 3

We start by stating required lemmas for proving Theorem 3.

Lemma 3. *Let v_n be a vector of component-wise client n's weighted distribution difference from p_{test} such that v_n's component is $v_{n,k} = \lambda_{n,k} p_{n,k} - p_{test,k}$ for every k. The upper bound of $R(S, \lambda) = \sum_{k=1}^K \left(\frac{1}{M} \sum_{n:S_n=1} \lambda_{n,k} p_{n,k} - p_{test,k}\right)^2$ is $R_{max}(M) = \left(\|\mu\| + \frac{\phi(N-M)}{M}\right)^2$ where $\mu = \frac{1}{N} \sum_{n=1}^N v_n$ and $\phi(M) = \max_{\sum_n S_n = M} \left\|\sum_{n:S_n=0} v_n\right\|$.*

Proof. Each v_n can be represent as $v_n = \mu + \epsilon_n$ where $\mu = \frac{1}{N} \sum_{n=1}^N v_n$ and $\sum_{n=1}^N \epsilon_n = 0$. Therefore, for any subset of selected client $S \in \{0, 1\}^N$, $\sum_{n:S_n=1} \epsilon_n = -\sum_{n:S_n=0} \epsilon_n$ and,

$$\begin{aligned}
\frac{1}{M} \sum_{n:S_n=1} v_n &= \frac{1}{M} \sum_{n:S_n=1} (\mu + \epsilon_n) \\
&= \frac{1}{M} (M\mu + \sum_{n:S_n=0} \epsilon_n) \\
&= \mu - \frac{1}{M} \sum_{n:S_n=0} \epsilon_n
\end{aligned}$$

Define $\phi(M) = \max_{|S|=M} \left\|\sum_{n:S_n=1} \epsilon_n\right\|$. We can bound the $R(S,\lambda)$ using the upper bound depending on M,

$$\begin{aligned} R(S,\lambda) &= \left\| \frac{1}{M} \sum_{n:S_n=1} v_n \right\|^2 \\ &= \left\| \mu - \frac{1}{M} \sum_{n:S_n=0} \epsilon_n \right\|^2 \\ &\leq \left(\|\mu\| + \frac{\phi(N-M)}{M} \right)^2 \\ &= R_{\max}(M) \end{aligned}$$

where we use Triangle inequality in the third derivation, and the last one follows the definition of $R_{\max}$. □

Lemma 4. *Let p_{test} and p_n for every client $n \in \{1,\dots,N\}$ be discrete probability distributions over a set $\{1,\dots,K\}$, and let $\lambda_{n,k}$ be the weighting coefficient for client n and class k such that $\sum_{k=1}^K \lambda_{n,k} p_{n,k} = 1$. The following holds for any selection vector $S(t) = (S_1(t),\dots,S_N(t))$:*

$$R\left(S(t),\lambda\right) = \sum_{k=1}^K \left(\frac{1}{M} \sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k} - p_{test,k} \right)^2 \leq 2D(S(t),\lambda).$$

Proof. We first note that $\frac{1}{M}\sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k} \geq 0$ for every $k \in \{1,\dots,K\}$ and

$$\sum_{k=1}^K \frac{1}{M} \sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k} = \frac{1}{M} \sum_{n:S_n(t)=1} \sum_{k=1}^K \lambda_{n,k} p_{n,k} = 1.$$

Therefore, the vector $\left(\frac{1}{M}\sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k}\right)_{k=1}^K$ is a discrete probability distribution. By Pinsker's inequality, we obtain

$$\begin{aligned} \sum_{k=1}^K \left(\frac{1}{M} \sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k} - p_{test,k} \right)^2 &= \sum_{k=1}^K \left| \frac{1}{M} \sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k} - p_{test,k} \right|^2 \\ &\leq \left(\sum_{k=1}^K \left| \frac{1}{M} \sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k} - p_{test,k} \right| \right)^2 \\ &\leq 2 \sum_{k=1}^K p_{test,k} \log \left(\frac{\frac{1}{M}\sum_{n:S_n(t)=1} \lambda_{n,k} p_{n,k}}{p_{test,k}} \right) \\ &= 2D(S(t),\lambda). \end{aligned}$$

□

Lemma 5. *The distribution shift D is L_D-Lipschitz continuous where $L_D = \frac{K\sqrt{N}}{M q_{min}}$ and $q_{min} = \min_{S\in\{0,1\}^N : \sum_{n=1}^N S_n = M} \frac{1}{M} \sum_{n=1}^N S_n \lambda_{n,k} p_{n,k} > 0$.*

Proof. We will show the Lipschitz continuity of D by showing that its gradient norm $\|\nabla D(S,\lambda)\|$ is bounded by L_D. We first note that the partial derivative of D can be bounded,

$$\left| \frac{\partial D}{\partial S_n} \right| = \left| -\frac{1}{M} \sum_{k=1}^K \frac{p_{test,k} \lambda_{n,k} p_{n,k}}{\frac{1}{M}\sum_{i=1}^N S_i \lambda_{i,k} p_{i,k}} \right| \leq \frac{K}{M q_{min}}$$

Consequently,

$$\|\nabla D(S,\lambda)\| \leq \sqrt{\sum_{n=1}^{N}\left(\frac{K}{Mq_{min}}\right)^2} = \frac{K\sqrt{N}}{Mq_{min}} = L_D$$

Next, we will show that D is L_D-Lipshitz continuos. By Mean Value Theorem, there exists $t_0 \in (0,1)$ and $S = S_2 + t_0(S_1 - S_2)$ such that,

$$\begin{aligned}|D(S_1,\lambda) - D(S_2,\lambda)| &= |\nabla D(S,\lambda)^\top (S_1 - S_2)| \\ &\leq \|\nabla D(S,\lambda)\| \, \|S_1 - S_2\| \\ &\leq L_D \, \|S_1 - S_2\|\end{aligned}$$

Therefore, the distribution shift D is L_D-Lipschitz continuous.

□

Theorem 3. *Suppose Assumptions 1 to 4 hold, $\epsilon > 0$, and $H = \frac{1}{2} - \beta^2 C R_{\max}(M) > 0$. The convergence bound of the FedMLS for $T > T'_{V,\lambda,\epsilon}$ communication rounds, and learning rate $\eta = \frac{\eta_{max}}{\sqrt{T}}$ where $\eta_{max} = \frac{1}{2L\sqrt{I(I-1)\left(1+\frac{8\beta^2 B}{H}\right)}}$ is:*

$$\begin{aligned}\min_{0\leq t\leq T-1} \mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \leq{}& \frac{1}{HT}\left[16\eta_{max}^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I} + 2\varphi^2 + 2B\gamma^2\right)\right] \\ &+ \frac{1}{H\sqrt{T}}\left[\frac{2\eta_{max}\sigma^2 L}{MI} + \frac{2\left(f_{test}(w(0)) - f_{min}\right)}{\eta_{max}}\right] \\ &+ \frac{4}{H}\left[\varphi^2 + \gamma^2 C\left(D_\lambda^{opt} + \frac{N}{V} + \epsilon + L_D\sqrt{M\left(1 - \frac{M}{N}\right)}\right)\right].\end{aligned}$$

where $B = \max_n \sum_{k=1}^{K} \frac{(\lambda_{n,k} p_{n,k})^2}{p_{test,k}}$, $C = \max_k \frac{1}{p_{test,k}^2}$, and L_D is the Lipschitz constant of D.

Proof. By *Lemma* 5, The distribution shift D is L_D-lipschitz continuous, that is

$$D(S(t),\lambda) \leq D(\bar{S}(T),\lambda) + L_D \left\|S(t) - \bar{S}(T)\right\|.$$

We then take average over $t = 0, \ldots, T-1$, for $T \geq T'_{V,\lambda,\epsilon}$

$$\begin{aligned}\bar{D}(S(t),\lambda) &\leq D(\bar{S}(T),\lambda) + L_D \frac{1}{T}\sum_{t=0}^{T-1}\left\|S(t) - \bar{S}(T)\right\| \\ &\leq D(\bar{S}(T),\lambda) + L_D\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\left\|S(t) - \bar{S}(T)\right\|^2} \\ &= D(\bar{S}(T),\lambda) + L_D\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\|S(t)\|^2 - \left\|\bar{S}(T)\right\|^2} \\ &< \left(D_\lambda^{opt} + \frac{N}{V} + \epsilon\right) + L_D\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\|S(t)\|^2 - \left\|\bar{S}(T)\right\|^2}\end{aligned}$$

where the second inequality uses the AM-QM inequality, and the last derivation follows Corollary 1.

Since $S(t) \in \{0,1\}^N$ and $\sum_{n=1}^{N} S_n(t) = M$, $\|S(t)\|^2 = \sum_{n=1}^{N} S_n(t)^2 = M$. Also, by Cauchy-Schwarz inequality,

$$M^2 = \left(\sum_{n=1}^{N}\bar{S}_n(T)\right)^2 \leq \left(\sum_{n=1}^{N}\bar{S}_n(T)^2\right)\left(\sum_{n=1}^{N}1^2\right) = \left\|\bar{S}(T)\right\|^2 N$$

The upper bound of $\bar{D}(S(t))$ becomes,

$$\bar{D}(S(t),\lambda) \leq \left(D_\lambda^{opt} + \frac{N}{V} + \epsilon\right) + L_D\sqrt{M\left(1-\frac{M}{N}\right)} \tag{14}$$

Taking an average over $t = 0, \ldots, T-1$ to the Lemma 2:

$$\begin{aligned}
&\frac{1}{T}\left(\mathbb{E}[f_{test}(w(T))] - f_{test}(w(0))\right) \\
&\leq \left[4\eta\beta^2\frac{AB}{1-A} - \frac{1}{2}\eta\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \eta\beta^2 C\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[R(S(t),\lambda)\right]\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \\
&\quad + \frac{\eta^2\sigma^2 L}{MI} + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + C\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[R(S(t),\lambda)\right]\right)\right] \\
&\leq \left[4\eta\beta^2\frac{AB}{1-A} - \frac{1}{2}\eta\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \eta\beta^2 C R_{\max}(M)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \\
&\quad + \frac{\eta^2\sigma^2 L}{MI} + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + C\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[R(S(t),\lambda)\right]\right)\right]
\end{aligned}$$

where the second derivation follows $R(S(t),\lambda) \leq R_{\max}(M)$, proven in Lemma 3.
We then apply the Lemma 4 and the upper bound of $\bar{D}(S(t))$ from (14), resulting in

$$\begin{aligned}
&\frac{1}{T}\left(\mathbb{E}[f_{test}(w(T))] - f_{test}(w(0))\right) \\
&\leq \left[4\eta\beta^2\frac{AB}{1-A} - \frac{1}{2}\eta + \eta\beta^2 C R_{\max}(M)\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \\
&\quad + \frac{\eta^2\sigma^2 L}{MI} + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + C\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[D(S(t),\lambda)\right]\right)\right] \\
&= \left[4\eta\beta^2\frac{AB}{1-A} - \frac{1}{2}\eta + \eta\beta^2 C R_{\max}(M)\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \\
&\quad + \frac{\eta^2\sigma^2 L}{MI} + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + C\mathbb{E}\left[\bar{D}(S(t),\lambda)\right]\right)\right] \\
&\leq \left[4\eta\beta^2\frac{AB}{1-A} - \frac{1}{2}\eta + \eta\beta^2 C R_{\max}(M)\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \frac{\eta^2\sigma^2 L}{MI} \\
&\quad + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + C\left(D_\lambda^{opt} + \frac{N}{V} + \epsilon + L_D\sqrt{M\left(1-\frac{M}{N}\right)}\right)\right)\right] \\
&= \left[4\eta\beta^2\frac{AB}{1-A} - H\eta\right]\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \frac{\eta^2\sigma^2 L}{MI} \\
&\quad + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + C\left(D_\lambda^{opt} + \frac{N}{V} + \epsilon + L_D\sqrt{M\left(1-\frac{M}{N}\right)}\right)\right)\right]
\end{aligned}$$

where $H = \frac{1}{2} - \beta^2 C R_{\max}(M)$.

If $H > 0$ and $A \leq \frac{H}{H+8\beta^2 B}$, we have $8\beta^2 AB \leq H(1-A)$. We then use $4\eta\beta^2 \frac{AB}{1-A} \leq \frac{H}{2}$ to simplify the bound to

$$\begin{aligned}
&\frac{1}{T}\left(\mathbb{E}[f_{test}(w(T))] - f_{test}(w(0))\right) \\
&\leq -\eta\left(\frac{H}{2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f_{test}(w(t))\|^2\right] + \frac{\eta^2\sigma^2 L}{MI} \\
&\quad + 2\eta\left[\frac{A}{1-A}\left(\frac{\sigma^2}{2I} + 2\varphi^2 + 2B\gamma^2\right)\right] + 2\eta\left[\varphi^2 + \gamma^2 C\left(D_\lambda^{opt} + \frac{N}{V} + \epsilon + L_D\sqrt{M\left(1-\frac{M}{N}\right)}\right)\right].
\end{aligned}$$

By rearranging terms, we get

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w_t)\|^2] &\leq \frac{2\eta\sigma^2 L}{HMI} + \frac{2\left(f_{test}(w(0)) - \mathbb{E}\left[f_{test}(w(T))\right]\right)}{\eta HT} \\
&\quad + \frac{4}{H}\left[\frac{A}{1-A}\left(\frac{\sigma^2}{2I} + 2\varphi^2 + 2B\gamma^2\right)\right] \\
&\quad + \frac{4}{H}\left[\varphi^2 + \gamma^2 C\left(D_\lambda^{opt} + \frac{N}{V} + \epsilon + L_D\sqrt{M\left(1-\frac{M}{N}\right)}\right)\right].
\end{aligned}$$

If $A \leq \frac{1}{2}$, we can simplify the bound using $\frac{A}{1-A} \leq 2A$ and obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w_t)\|^2] &\leq \frac{2\eta\sigma^2 L}{HMI} + \frac{2\left(f_{test}(w(0)) - \mathbb{E}\left[f_{test}(w(T))\right]\right)}{\eta HT} \\
&\quad + \frac{4}{H}\left[4\eta^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I} + 2\varphi^2 + 2B\gamma^2\right)\right] \\
&\quad + \frac{4}{H}\left[\varphi^2 + \gamma^2 C\left(D_\lambda^{opt} + \frac{N}{V} + \epsilon + L_D\sqrt{M\left(1-\frac{M}{N}\right)}\right)\right].
\end{aligned}$$

Note that the above bound holds when the conditions of η are:

$$\eta \leq \frac{1}{2L} \quad \text{and} \quad A = 2\eta^2 L^2 I(I-1) \leq \frac{H}{H+8\beta^2 B} \quad \text{and} \quad A \leq \frac{1}{2}.$$

We satisfy all conditions by setting $\eta_{max} = \frac{1}{2L\sqrt{I(I-1)\left(1+\frac{8\beta^2 B}{H}\right)}}$ and $\eta = \frac{\eta_{max}}{\sqrt{T}}$, since $I > 1$, $\beta^2 > 0$, and $H > 0$. We finally prove the theorem:

$$\begin{aligned}
\min_{0\leq t\leq T-1}\mathbb{E}[\|\nabla f_{test}(w_t)\|^2] &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w_t)\|^2] \\
&\leq \frac{1}{HT}\left[16\eta_{max}^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I} + 2\varphi^2 + 2B\gamma^2\right)\right] \\
&\quad + \frac{1}{H\sqrt{T}}\left[\frac{2\eta_{max}\sigma^2 L}{MI} + \frac{2\left(f_{test}(w(0)) - f_{min}\right)}{\eta_{max}}\right] \\
&\quad + \frac{4}{H}\left[\varphi^2 + \gamma^2 C\left(D_\lambda^{opt} + \frac{N}{V} + \epsilon + L_D\sqrt{M\left(1-\frac{M}{N}\right)}\right)\right].
\end{aligned}$$

□

D Proof of Theorem 4

Theorem 4. *Suppose Assumptions 1 to 4 hold, and* $\lambda_{n,k} = \frac{p_{test,k}}{p_{n,k}}$ *where* $p_{n,k} > 0$ *for all* n, k. *The convergence bound of the federated learning algorithm with an arbitrary client selection algorithm with learning rate* $\eta = \frac{\eta_{max}}{\sqrt{T}}$ *and* $\eta_{max} = \frac{1}{2L\sqrt{I(I-1)(1+8\beta^2)}}$ *is as follows:*

$$\min_{0\leq t\leq T-1} \mathbb{E}[\|\nabla f_{test}(w(t))\|^2] \leq \frac{1}{T}\left[32\eta_{max}^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I}+2\varphi^2+2\gamma^2\right)\right] + \frac{1}{\sqrt{T}}\left[\frac{4\eta_{max}\sigma^2 L}{MI} + \frac{4\left(f_{test}(w(0))-f_{min}\right)}{\eta_{max}}\right] + 8\varphi^2,$$

Proof. From Lemma 2, when $\lambda_{n,k} = \frac{p_{test,k}}{p_{n,k}}$ for all n and all k, we have $\mathbb{E}[R(S(t),\lambda)] = 0$ and $B = 1$. The bound then becomes:

$$\begin{aligned}
&\mathbb{E}[f_{test}(w(t+1))] - \mathbb{E}[f_{test}(w(t))] \\
&\leq \left[4\eta\frac{AB}{1-A}\beta^2 + \eta\beta^2 C\mathbb{E}\left[R(S(t),\lambda)\right] - \frac{1}{2}\eta\right]\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \frac{\eta^2\sigma^2 L}{MI} \\
&\quad + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\left(\frac{2AB}{1-A} + \frac{1}{2}C\mathbb{E}\left[R(S(t),\lambda\right]\right)\right] \\
&= \left[4\eta\frac{A}{1-A}\beta^2 - \frac{1}{2}\eta\right]\mathbb{E}[\|\nabla f_{test}(w(t))\|^2] + \frac{\eta^2\sigma^2 L}{MI} \\
&\quad + 2\eta\left[\frac{\sigma^2 A}{2I(1-A)} + \left(\frac{2A}{1-A}+1\right)\varphi^2 + \gamma^2\frac{2A}{1-A}\right].
\end{aligned}$$

If $A \leq \frac{1/2}{1/2+8\beta^2}$, we have $4\frac{A}{1-A}\beta^2 - \frac{1}{2} \leq -\frac{1}{4}$. We then can simplify the bound:

$$\mathbb{E}[f_{test}(w(t+1))] - \mathbb{E}[f_{test}(w(t))] \leq -\eta\frac{1}{4}\mathbb{E}\left[\|\nabla f_{test}(w(t))\|^2\right] + \frac{\eta^2\sigma^2 L}{MI} + 2\eta\left[\frac{A}{1-A}\left(\frac{\sigma^2}{2I}+2\varphi^2+2\gamma^2\right)\right] + 2\eta\varphi^2.$$

By rearranging terms, we get

$$\mathbb{E}[\|\nabla f_{test}(w_t)\|^2] \leq \frac{4\eta\sigma^2 L}{MI} + \frac{4\left(\mathbb{E}\left[f_{test}(w(t))\right] - \mathbb{E}\left[f_{test}(w(t+1))\right]\right)}{\eta} + 8\left[\frac{A}{1-A}\left(\frac{\sigma^2}{2I}+2\varphi^2+2\gamma^2\right)\right] + 8\varphi^2.$$

We then take the average over rounds. If $A \leq \frac{1}{2}$, we get

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f_{test}(w_t)\|^2] &\leq \frac{4\eta\sigma^2 L}{MI} + \frac{4\sum_{t=0}^{T-1}\left(\mathbb{E}\left[f_{test}(w(t))\right] - \mathbb{E}\left[f_{test}(w(t+1))\right]\right)}{\eta T} \\
&\quad + 8\left[\frac{A}{1-A}\left(\frac{\sigma^2}{2I}+2\varphi^2+2\gamma^2\right)\right] + 8\varphi^2 \\
&= \frac{4\eta\sigma^2 L}{MI} + \frac{4\left(f_{test}(w(0)) - f_{min}\right)}{\eta T} \\
&\quad + 8\left[\frac{A}{1-A}\left(\frac{\sigma^2}{2I}+2\varphi^2+2\gamma^2\right)\right] + 8\varphi^2 \\
&\leq \frac{4\eta\sigma^2 L}{MI} + \frac{4\left(f_{test}(w(0)) - f_{min}\right)}{\eta T} \\
&\quad + 8\left[4\eta^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I}+2\varphi^2+2\gamma^2\right)\right] + 8\varphi^2,
\end{aligned}$$

where the second derivation follows the telescoping sum, and the final derivation we use the fact that $A \leq \frac{1}{2}$. Note that the above bound holds when the conditions of η are:

$$\eta \leq \frac{1}{2L} \quad \text{and} \quad A = 2\eta^2 L^2 I(I-1) \leq \frac{1/2}{1/2 + 8\beta^2} \quad \text{and} \quad A \leq \frac{1}{2}.$$

We satisfy all conditions by setting $\eta_{max} = \frac{1}{2L\sqrt{I(I-1)(1+8\beta^2)}}$ and $\eta = \frac{\eta_{max}}{\sqrt{T}}$ and prove the theorem

$$\begin{aligned}
\min_{0\leq t\leq T-1} \mathbb{E}[\|\nabla f_{test}(w_t)\|^2] &\leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_{test}(w_t)\|^2] \\
&\leq \frac{1}{T}\left[32\eta_{max}^2 L^2 I(I-1)\left(\frac{\sigma^2}{2I} + 2\varphi^2 + 2\gamma^2\right)\right] \\
&\quad + \frac{1}{\sqrt{T}}\left[\frac{4\eta_{max}\sigma^2 L}{MI} + \frac{4\left(f_{test}(w(0)) - f_{min}\right)}{\eta_{max}}\right] + 8\varphi^2.
\end{aligned}$$

□

E Privacy-preserving mechanism

E.1 Proof of Theorem 2

Theorem 2. *Suppose the Service obtains the aggregate $P \in \mathbb{R}^{K\times O}$ over O observations, and $P = \mathbf{p}A$ where $\mathbf{p} = [p_1, \ldots, p_N] \in \mathbb{R}^{K\times N}$ is the stack of client distribution column vectors and $A = [a_1, \ldots, a_O] \in [0,1]^{N\times O}$ is the stack of auxiliary variable column vectors. If two clients $i \neq j$ hold the same set of non-zero classes,* $\mathrm{supp}(p_i) = \mathrm{supp}(p_j)$*, in different proportions, $p_i \neq p_j$, then the adversary's reconstruction problem admits more than one solution. Thus, $\mathbf{p}$ is not identifiable from P for any O observations.*

Proof. All adversary ever seen is the aggregate $P \in \mathbb{R}^{K\times O}$ from the O observations obtained so far. Both factors, $\mathbf{p} \in \mathbb{R}^{K\times N}$ and $A \in [0,1]^{N\times O}$, are not available to the adversary: the client distributions are private, and the auxiliary coefficients are known only to the Selector. The adversary attempts to produce a pair $(\mathbf{q}, B)$ belonging to

$$\mathcal{F}(P) \;=\; \Big\{(\mathbf{q}, B) \in \mathbb{R}_+^{K\times N} \times [0,1]^{N\times O} \;:\; \mathbf{q}B = P,\;\; \mathbf{1}_K^\top \mathbf{q} = \mathbf{1}_N^\top\Big\}.$$

where $\mathbf{1}_K$ is all-ones vector with dimension K. Every element of $\mathcal{F}(P)$ reproduces the observations exactly and violates nothing the adversary knows. Reconstruction of $\mathbf{p}$ succeeds if and only if $\mathcal{F}(P)$ contains a single $\mathbf{q}$.

We will show that for any solution $(\mathbf{q}, B) \in \mathcal{F}(P)$, we can always construct a solution $(\mathbf{q}^{(\varepsilon)}, B^{(\varepsilon)}) \in \mathcal{F}(P)$ using the invertible transformation matrix D_ε. Hence, $\mathcal{F}(P)$ has more than one element and $\mathbf{p}$ is not identifiable from P.

Suppose $(\mathbf{q}, B) \in \mathcal{F}(P)$ is an arbitrary solution available to the adversary, and let q_n denote the distribution it assigns to client n. Suppose there exist clients $i \neq j$ such that $\mathrm{supp}(q_i) = \mathrm{supp}(q_j) = S$ and $q_i \neq q_j$. We will construct a set of solutions $\left\{(\mathbf{q}^{(\varepsilon)}, B^{(\varepsilon)})\right\}_{\varepsilon\in(0,\delta(\mathbf{q})]}$ where $\delta(\mathbf{q}) = \min_{k\in S} \min\{q_{i,k}, q_{j,k}\}/(q_{i,k} + q_{j,k}) < \frac{1}{2}$ and show that every member of this set is in $\mathcal{F}(P)$. Since $\mathbf{q}B = (\mathbf{q}D^{-1})(DB)$ for any invertible matrix D, we define

$$D_\varepsilon \;=\; I - \varepsilon\, vv^\top, \qquad v = e_i - e_j, \qquad \varepsilon \in (0, \delta(\mathbf{q})],$$

where e_i is a vector where the i-th component is 1 and all other components are 0. This matrix coincides with the identity outside rows and columns i, j and equals $\left(\begin{smallmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{smallmatrix}\right)$ on that block. It is therefore nonnegative, its rows and its columns each sum to one, and it is invertible with

$$D_\varepsilon^{-1} \;=\; I + \frac{\varepsilon}{1-2\varepsilon}\, vv^\top,$$

as one verifies using $vv^\top vv^\top = 2\,vv^\top$. Applying it to the adversary's solution produces the candidate

$$\mathbf{q}^{(\varepsilon)} := \mathbf{q}D_\varepsilon^{-1}, \qquad B^{(\varepsilon)} := D_\varepsilon B, \qquad \text{so that} \qquad \mathbf{q}^{(\varepsilon)}B^{(\varepsilon)} = \mathbf{q}B = P.$$

It remains to check the three constraints that define $\mathcal{F}(P)$.

(i) The rows of D_ε are nonnegative and sum to one, so each row of $B^{(\varepsilon)}$ is a convex combination of rows of B: For each $o = 1, \ldots, O$, $b_i^{(\varepsilon)}(o) = (1-\varepsilon)b_i(o) + \varepsilon b_j(o)$, symmetrically for j, and $b_m^{(\varepsilon)} = b_m$ for $m \neq i, j$. Hence $B^{(\varepsilon)} \in [0,1]^{N\times O}$ for every admissible B.

(ii) The columns of D_ε sum to one, so $\mathbf{1_K}^\top D_\varepsilon = \mathbf{1_N}^\top$ and therefore $\mathbf{1_K}^\top D_\varepsilon^{-1} = \mathbf{1_N}^\top$, giving $\mathbf{1_K}^\top \mathbf{q}^{(\varepsilon)} = \mathbf{1_K}^\top \mathbf{q} D_\varepsilon^{-1} = \mathbf{1_N}^\top$.

(iii) Only the columns of clients i and j change,

$$q_i^{(\varepsilon)} = \frac{(1-\varepsilon)q_i - \varepsilon q_j}{1-2\varepsilon}, \qquad q_j^{(\varepsilon)} = \frac{(1-\varepsilon)q_j - \varepsilon q_i}{1-2\varepsilon}, \qquad q_m^{(\varepsilon)} = q_m \;\; (m \neq i, j).$$

For $k \notin S$ both entries vanish; for $k \in S$, nonnegativity of the two is equivalent to $\varepsilon \leq \min\{q_{i,k}, q_{j,k}\}/(q_{i,k} + q_{j,k})$. Hence, $\mathbf{q}^{(\varepsilon)} \geq 0$ for every $\varepsilon \in (0, \delta(\mathbf{q})]$ since $\delta(\mathbf{q}) = \min_{k\in S} \min\{q_{i,k}, q_{j,k}\}/(q_{i,k} + q_{j,k})$ and $q_{i,k}, q_{j,k} > 0$ on S.

Therefore, $(\mathbf{q}^{(\varepsilon)}, B^{(\varepsilon)}) \in \mathcal{F}(P)$ for every $\varepsilon \in (0, \delta(\mathbf{q})]$, and this solution differs from the original, since $q_i^{(\varepsilon)} - q_i = \frac{\varepsilon}{1-2\varepsilon}(q_i - q_j) \neq 0$. Thus, $\mathcal{F}(P)$ contains at least two distinct client distribution matrices, both reproducing P exactly and satisfying every constraint available to the adversary, so $\mathbf{p}$ is not determined by P.

□

E.2 Implementation details

We implement the inner loop of Algorithm 2 with the CKKS scheme using TenSEAL 0.3.18 (Microsoft SEAL) under Python 3.11.5, on a single core of a 13th-generation Intel Core i7-13700. All measurements are single-threaded.

Setup. The label distribution is encrypted twice at setup: client-major, $c_n = \mathrm{Enc}(p_n)$ over K slots, and class-major, $c_k = \mathrm{Enc}(p_{\cdot,k})$ over N slots. The Selector forms $\mathrm{Enc}(P(r))$ from the client-major ciphertexts with N plaintext-scalar multiplications, and the Service forms $\mathrm{Enc}(h(r))$ from the class-major ciphertexts with K. The masks $\xi_1(r) \in \mathbb{R}^K$ and $\xi_2(r) \in \mathbb{R}^N$ are drawn i.i.d. uniform on $[0, 10^6]$, refreshed at every inner iteration, and added to the respective ciphertexts as plaintexts.

Configurations. We sweep $N \in \{10, 50, 100, 500, 1000\}$ and $K \in \{2, 10, 50, 100\}$. Client distributions are drawn as $p_n \sim \mathrm{Dirichlet}(\alpha = 1)$, p_{test} is concentrated near the client mean, $\mu = 0.2$ and $a^{(t)}(0) = 0.5\,\mathbf{1}$. The horizon R is measured per configuration as the first r with $\|a^{(t)}(r+1) - a^{(t)}(r)\|_\infty < 10^{-6}$, capped at $R_{\max} = 64$. Computation time is taken with `time.perf_counter` as the median over 10 inner iterations and covers the homomorphic work only: encrypted-domain arithmetic and decryption.

Results. Figure 2 show that the encrypted protocol is practical at realistic scale. One inner iteration costs scale with $N + K$ and around 300 ms when $N + K = 1000$. The approximation introduced by CKKS is immaterial: the mean ℓ_∞ deviation from the exact computation stays below 1.7×10^{-7} and does not accumulate over rounds, since each iteration re-uses the original client ciphertexts and the multiplicative depth remains one.

F Experiments Setup Details

All experiments are conducted on a personal computer equipped with a single NVIDIA GeForce RTX 4080 GPU. We evaluate each client selection strategy independently, with clients performing local SGD in parallel. The algorithms are implemented using TensorFlow.

Data partitioning. The data partitioning follows a three-stage process to simulate label distribution shifts: First, the dataset is partitioned into training, validation, and test sets

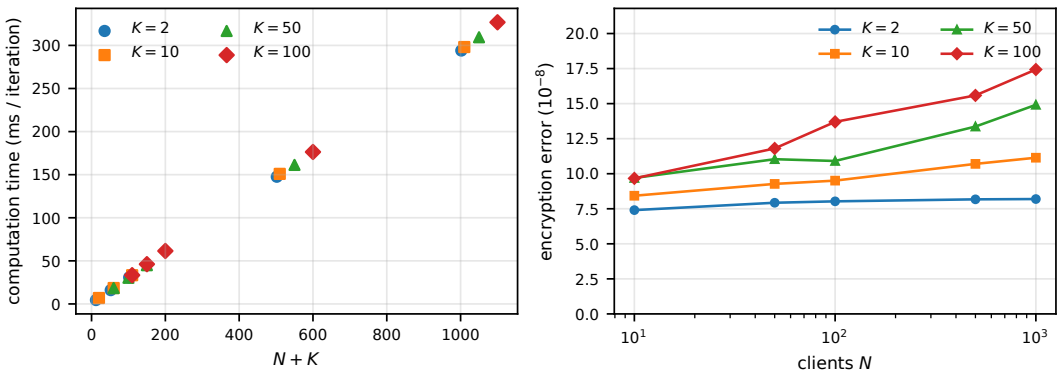


Figure 2: **Left:** per-iteration computation time against $N+K$. **Right:** mean ℓ_∞ encryption error, which grows with K as $\mathrm{Enc}(h)$ sums K noisy terms and remains below 1.8×10^{-7} throughout.

Algorithm 2 Privacy-preserving solution method for the auxiliary problem (4) in round t

Setup (once): Client n sends $\mathrm{Enc}(p_n)$ to Selector and Service, $n \in \{1, \dots, N\}$.
Initialize: $r \leftarrow 0$, $a^{(t)}(0) \leftarrow c\mathbf{1}$ with $c \in (0, 1]$.
while $a^{(t)}(r)$ has not converged and $r < R$ **do**
  Selector draws fresh $\xi_1(r) \in \mathbb{R}^K$, sends $\xi_1(r)$ to Service and sends $\mathrm{Enc}(P(r)+\xi_1(r))$ to Oracle, where $\mathrm{Enc}(P(r)) = \frac{1}{M}\sum_{n=1}^{N} a_n^{(t)}(r)\,\mathrm{Enc}(p_n)$.
  Oracle decrypts and sends $P(r) + \xi_1(r)$ to Service.
  Service recovers $P(r)$ and computes $q_k(r) \leftarrow p_{\mathrm{test},k}/P_k(r)$ for $k \in \{1, \dots, K\}$.
  Service draws fresh $\xi_2(r) \in \mathbb{R}^N$, sends $\xi_2(r)$ to Selector and sends $\mathrm{Enc}(h(r)+\xi_2(r))$ to Oracle, where $\mathrm{Enc}(h_n(r)) = \frac{V}{M}\sum_{k=1}^{K} q_k(r)\,\mathrm{Enc}(p_{n,k})$.
  Oracle decrypts and sends $h(r) + \xi_2(r)$ to Selector.
  Selector recovers $h(r)$ and sets $\nabla f_{\mathrm{aux}}(r) \leftarrow A(t) - h(r)$.
  Selector updates $a^{(t)}(r+1) \leftarrow \mathrm{proj}_{[0,1]^N}\big(a^{(t)}(r) - \mu\,\nabla f_{\mathrm{aux}}(r)\big)$, $r \leftarrow r+1$.
end while

following an 80/10/10 split. Second, following (Hsu et al., 2019; Cho et al., 2022; Zhang et al., 2023), each client $n \in \{1, \dots, 1000\}$ is assigned a label distribution p_n sampled from a Dirichlet distribution $\mathrm{Dir}(\alpha_{train})$. The concentration parameter α_{train} controls heterogeneity, where smaller values induce greater statistical heterogeneity. To make sure each client's local dataset $\mathcal{D}_n$ is characterized by its assigned p_n, we draw samples with replacement from the training dataset, ensuring each sample belongs to class k with probability $p_{n,k}$. Third, we construct centralized validation and test sets by independently sampling two probability vectors, p_{val} and p_{test}, from $\mathrm{Dir}(\alpha_{test})$. These vectors are used to subsample the validation and test sets, respectively, ensuring label distributions are not necessarily uniform and distinct from one another. The complete data statistic is provided in 4

Table 4: Federated setup for each dataset.

Dataset	Clients N	Selected per round M	Classes K
CIFAR-100	1,000	10	100
iNaturalist-User-120k	9,275	10	1,203
Landmarks-User-160k	1,262	10	2,028
Shakespeare	659	10	79

Hyperparameter tuning. To ensure a fair comparison and optimal performance, we performed a systematic grid search for all evaluated methods. Each configuration was evaluated on the centralized validation dataset, and the parameters yielding the highest validation accuracy were selected for the final experimental runs.

The tuning process encompassed both common federated learning hyperparameters and method-specific variables. The search ranges, informed by preliminary experiments to identify regions for each parameter, are summarized in Table 5.

Table 5: Hyperparameter search space for baseline methods.

Category	Hyperparameter	Search Range
General	Learning rate (η)	$\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
	Number of local iterations (I)	$\{10, 20, 50, 100, 250, 500, 1000\}$
Power-of-Choice	Candidate set size (d)	$\{2M, 5M, 10M\}$
Fed-CBS	Exploration factor (λ)	$\{0.1, 1, 10, 100, 1000\}$
FedMLS	Trade-off parameter (V)	$\{0.01, 0.1, 1, 10, 100\}$
	Minimum rate scaling parameter (c)	$\{0, 0.2, 0.5, 0.8, 1.0\}$

The optimal configurations identified through this process are as follows:

- **CIFAR-100:** We utilize a learning rate of $\eta = 0.01$ and $I = 20$ local steps. For the baselines, Power-of-Choice uses a candidate set size of $d = 2M$ and Fed-CBS uses $\lambda = 1$. For our proposed FedMLS, the trade-off parameter is set to $V = 10$ and $c = 0.8$.
- **iNaturalist-User-120k:** We utilize a learning rate of $\eta = 0.01$ and $I = 100$ local steps. Baseline configurations include $d = 2M$ for Power-of-Choice and $\lambda = 10$ for Fed-CBS. For FedMLS, the trade-off parameter is set to $V = 1$ and $c = 0.5$.
- **Landmarks-User-160k:** We set a learning rate of $\eta = 0.01$ and $I = 100$ local steps. Baseline configurations include $d = 2M$ for Power-of-Choice and $\lambda = 1$ for Fed-CBS. For FedMLS, the trade-off parameter is set to $V = 10$ and $c = 0.5$.
- **Shakespeare:** We fix a learning rate of $\eta = 0.1$ and $I = 1000$ local steps. Baseline configurations include $d = 2M$ for Power-of-Choice and $\lambda = 1$ for Fed-CBS. For FedMLS, the trade-off parameter is set to $V = 0.1$ and $c = 0.5$.

CNN on CIFAR-100. We adopt a deep convolutional neural network (CNN) consisting of two convolutional layers with max pooling, following (Zhang et al., 2023). The convolution layers have size 5×5 and the kernel size is 6 and 16, respectively. The resulting feature maps are flattened and fed into two fully connected layers with 120 and 84 units. Pixel values are normalized to the range $[0, 1]$. Following (Li et al., 2020), the local dataset size $|\mathcal{D}_n|$ follows a log-normal distribution with $\mu = 4$ and $\sigma = 2$. We set the batch size $B = 50$, and a weight decay of 5×10^{-4}. The learning rate η decay rate is set to 0.9992, consistent with (Zhang et al., 2023).

MobileNetV2 on iNaturalist-User-120k and Landmarks-User-160k Similar to (Hsu et al., 2020), we employ a pre-trained MobileNetV2 architecture (Sandler et al., 2018). During training, we apply random cropping and resizing to 299×299 pixels for iNaturalist-User-120k and 224×224 pixels for Landmarks-User-160k dataset. We set a batch size $B = 64$, a weight decay of 4×10^{-5} and the learning rate is fixed without decay.

LSTM on Shakespeare Following (Caldas et al., 2018), we adopt a long short-term memory (LSTM) of two layers of 256 units each with sequence length of 80. We set a batch size $B = 10$ and the learning rate is fixed.

Baseline implementations. We implement the baseline strategies as follows. For Power-of-Choice (pow-d) (Cho et al., 2022), we first sample a candidate set $\mathcal{A}$ of d clients with probabilities proportional to their local dataset sizes. From $\mathcal{A}$, we select the M clients with the highest training losses from their most recent local updates; for clients not previously selected, the loss is initialized to ∞. For Fed-CBS (Zhang et al., 2023), we adhere to the original study's configuration using hyperparameters $\beta_m = m, \gamma = 10$, and $L_b = 10^{-20}$. Finally, for ODFL (Jeong and Lee, 2024), we evaluate the ground-truth (GT) variant, as we assume prior knowledge of both the individual client label distributions and the test label distribution.

G Additional Experiments

G.1 Image classification with Landmarks-User-160k

To demonstrate the real-world scalability and robustness of our approach on large-scale datasets with realistic user partitioning, we evaluate FedMLS on the Google Landmarks Dataset v2 (Landmarks-User-160k) (Hsu et al., 2020). This task presents severe naturally occurring non-IID data distributions and extreme client heterogeneity. The optimal configuration for each method is provided in F

Table 6: Performance evaluation under varying test distribution shifts ($\alpha_{\text{test}} \in \{1, 0.1\}$). Results report the mean test accuracy (Acc.) expressed as a percentage, standard error (SE), and convergence speedup factor (Conv.). An entry of "N/A" in the Conv. column indicates that the method fails to reach the target accuracy. **Bold** values denote the best-performing method for each configuration.

		with FedIR		**without FedIR**	
α_{test}	**Method**	**Acc.$_{\text{IR}}$**	**Conv.**	**Acc.**	**Conv.**
1	FedAvg	29.05 ± 0.33	1.4×	28.74 ± 0.39	1.5×
	Power-of-Choice	35.64 ± 0.26	1.0×	34.95 ± 0.42	1.0×
	Fed-CBS	18.19 ± 0.49	N/A	18.31 ± 0.25	N/A
	On-Demand FL	14.11 ± 0.21	N/A	15.30 ± 0.51	N/A
	FedMLS (ours)	$\mathbf{36.49 \pm 0.26}$	–	$\mathbf{35.99 \pm 0.42}$	–
0.1	FedAvg	32.55 ± 0.44	1.6×	32.32 ± 0.31	1.5×
	Power-of-Choice	37.88 ± 0.61	1.1×	37.41 ± 0.24	1.0×
	Fed-CBS	21.01 ± 0.58	N/A	21.31 ± 0.26	N/A
	On-Demand FL	19.16 ± 0.42	N/A	17.19 ± 0.26	N/A
	FedMLS (ours)	$\mathbf{40.64 \pm 0.27}$	–	$\mathbf{39.93 \pm 0.41}$	–

Results. As shown in Table 6, FedMLS consistently outperforms all baselines across both distribution shift regimes ($\alpha_{\text{test}} \in \{1, 0.1\}$), both with and without FedIR. Under moderate distribution shift ($\alpha_{\text{test}} = 1$), baselines such as Fed-CBS and On-Demand FL fail to reach the target accuracy, receiving an "N/A" convergence rating regardless of whether FedIR is integrated. Conversely, FedMLS maintains superior generalization, achieving top test accuracies of 36.49% with FedIR and 35.99% without FedIR. Under extreme shift ($\alpha_{\text{test}} = 0.1$), the advantage of our approach expands: FedMLS achieves peak test accuracies of 40.64% (with FedIR) and 39.93% (without FedIR), outperforming Power-of-Choice (37.88% with FedIR; 37.41% without) and FedAvg (32.55% with FedIR; 32.32% without). Furthermore, FedMLS yields a 1.6× relative speedup over FedAvg under severe skew. Crucially, combining FedMLS with FedIR yields consistent gains, proving that client selection via KL-divergence minimization complements loss-reweighting schemes when mitigating label distribution skew.

G.2 Effect of the number of selected client (M).

In this section, we evaluate the FedMLS and baseline client selection method under varying degrees of test label distribution shift ($\alpha_{\text{test}} \in \{10, 1, 0.1\}$) and varying numbers of selected clients per round ($M \in \{5, 10, 20\}$) on CIFAR-100 dataset. All method apply FedIR loss reweighting technique.

As shown in Table 7, when client participation is low ($M = 5$), test accuracy remains constrained across all algorithms. However, as client participation increases to $M = 10$ and $M = 20$, FedMLS exhibits substantial performance gains. Under severe distribution shift ($\alpha_{\text{test}} = 0.1$), FedMLS achieves a peak test accuracy of 49.52% ($M = 20$), outperforming the best baseline, Fed-CBS (43.40%), by 6.12% absolute accuracy. Notably, while On-Demand FL struggles under mild shift ($\alpha_{\text{test}} = 10$, achieving only 12.49% accuracy at $M = 20$), FedMLS maintains stable generalization across all regimes (30.99%). These empirical results confirm that optimizing client selection via distribution alignment effectively mitigates performance degradation caused by test label skew.

G.3 Ablation Study

To evaluate the individual contributions of our proposed selection mechanisms and examine their compatibility with existing techniques, we conduct an ablation study on CIFAR-100 under a severe test label distribution shift ($\alpha_{\text{test}} = 0.1$). Specifically, we isolate our two proposed components—Distribution Matching and the Minimum Rate Constraint—and eval-

Table 7: Performance evaluation under varying client selection sizes ($M \in \{5, 10, 20\}$) across test label distribution shifts ($\alpha_{\text{test}} \in \{10, 1, 0.1\}$). Results report the mean test accuracy (Acc.) expressed as percentage, standard error (SE), and convergence speedup factor (Conv.). An entry of "N/A" in the Conv. column indicates that the method fails to reach the target accuracy. **Bold** values denote the best-performing method for each configuration.

		$M = 5$		$M = 10$		$M = 20$	
α_{test}	**Method**	**Acc.**	**Conv.**	**Acc.**	**Conv.**	**Acc.**	**Conv.**
10	FedAvg	20.42 ± 0.71	1.1×	22.48 ± 0.27	1.8×	28.04 ± 0.50	1.6×
	Power-of-Choice	20.85 ± 0.77	1.2×	22.55 ± 0.33	1.4×	28.03 ± 0.28	1.4×
	Fed-CBS	20.93 ± 0.81	1.2×	22.75 ± 0.34	1.4×	27.66 ± 0.27	N/A
	On-Demand FL	20.97 ± 0.76	1.1×	10.72 ± 0.30	N/A	12.49 ± 0.06	N/A
	FedMLS (ours)	$\mathbf{21.37 \pm 0.45}$	–	$\mathbf{24.95 \pm 0.44}$	–	$\mathbf{30.99 \pm 0.16}$	–
1	FedAvg	22.82 ± 1.10	1.0×	29.62 ± 0.26	2.4×	31.71 ± 0.45	2.0×
	Power-of-Choice	22.83 ± 0.97	1.0×	29.59 ± 0.23	2.0×	32.08 ± 0.37	2.0×
	Fed-CBS	22.82 ± 0.90	1.0×	29.34 ± 0.49	2.6×	31.79 ± 0.36	2.6×
	On-Demand FL	22.91 ± 0.96	1.0×	13.99 ± 0.27	N/A	15.93 ± 0.16	N/A
	FedMLS (ours)	$\mathbf{23.16 \pm 1.18}$	–	$\mathbf{32.24 \pm 0.33}$	–	$\mathbf{35.39 \pm 0.32}$	–
0.1	FedAvg	34.21 ± 1.03	2.1×	41.92 ± 0.57	4.6×	42.97 ± 0.22	6.5×
	Power-of-Choice	34.73 ± 1.20	3.0×	42.04 ± 0.51	5.4×	42.12 ± 0.85	N/A
	Fed-CBS	34.37 ± 1.01	3.0×	43.06 ± 0.24	3.4×	43.40 ± 0.67	7.2×
	On-Demand FL	34.66 ± 1.36	2.6×	35.66 ± 0.34	N/A	36.56 ± 0.52	N/A
	FedMLS (ours)	$\mathbf{37.06 \pm 1.44}$	–	$\mathbf{44.61 \pm 0.32}$	–	$\mathbf{49.52 \pm 0.34}$	–

uate them both independently and in conjunction with FedIR, an existing loss-reweighting method. Table 8 reports the test accuracy across all eight component configurations.

Table 8: Ablation study of the FedMLS components (Distribution Matching and Minimum Rate Constraint) with and without FedIR loss reweighting on CIFAR-100 ($\alpha_{\text{test}} = 0.1$). Rows without Distribution Matching set $V = 0$; rows without the Minimum Rate Constraint set $c = 0$. With both disabled, the internal states remain zero and selection is uniformly random via tie-breaking, unlike FedAvg, which samples clients in proportion to their dataset sizes r_n. FedAvg is included for reference. Results report the mean ± standard error over 8 runs. **Bold** denotes the best configuration.

FedMLS components (ours)		**External**	**Test Acc. (%)**
Distribution Matching	**Minimum Rate Constraint**	**FedIR**	
✓	✓	✓	$\mathbf{44.61 \pm 0.32}$
×	✓	✓	42.92 ± 0.34
✓	×	✓	41.48 ± 0.36
×	×	✓	32.20 ± 0.34
✓	✓	×	36.48 ± 0.40
×	✓	×	33.17 ± 0.13
✓	×	×	25.67 ± 0.27
×	×	×	21.63 ± 0.27
Reference: FedAvg		✓	41.92 ± 0.57
Reference: FedAvg		×	21.85 ± 0.63

Ablation Results and Discussion. Table 8 yields three observations.

First, FedMLS and FedIR are complementary. Without reweighting, FedMLS improves over FedAvg by 14.63% (36.48% vs. 21.85%). With FedIR, it still adds 2.69% on top of FedAvg with FedIR (44.61% vs. 41.92%), achieving the highest accuracy overall. Loss reweighting therefore does not substitute for distribution-aware selection.

Second, Distribution Matching improves accuracy beyond the Minimum Rate Constraint alone. Adding Distribution Matching to the Minimum Rate Constraint raises accuracy by 3.31% without FedIR (33.17% → 36.48%) and by 1.69% with FedIR (42.92% → 44.61%).

Third, the Minimum Rate Constraint prevents client starvation. Removing it reduces accuracy by 10.81% without FedIR (36.48% → 25.67%) and by 3.13% with FedIR (44.61% → 41.48%). Without this constraint, Distribution Matching concentrates selection on a small set of well-matched clients, reducing the diversity of participating data.

G.4 Robustness analysis

To simulate noisy or misspecified test label distributions, we perturb the true test label distribution p_{test} using two distinct noise models:

- **Additive Zero-Mean Gaussian Noise (Estimation Error):** We inject Gaussian noise $\epsilon \sim \mathcal{N}(0, \gamma\mu)$ into the true distribution, where γ controls the noise magnitude and μ represents the mean class probability.
- **Temperature Scaling (Distribution Skew):** Given a temperature parameter τ, we adjust the distribution sharpness via $p_{\text{skewed},k} = \frac{p_{\text{test},k}^{1/\tau}}{\sum_{k'} p_{\text{test},k'}^{1/\tau}}$.

The resulting probabilities are subsequently clipped to non-negative values and renormalized.

Table 9: Test accuracy (%) under noisy and misspecified test label distributions on CIFAR-100 across varying Dirichlet parameters (α_{test}).

Method	$\alpha_{\text{test}} = 10$	$\alpha_{\text{test}} = 1$	$\alpha_{\text{test}} = 0.1$
FedAvg (no noise)	21.55 ± 0.42	24.82 ± 0.32	21.85 ± 0.63
Power-of-Choice (no noise)	20.72 ± 0.43	24.00 ± 0.60	20.61 ± 0.71
Fed-CBS (no noise)	22.90 ± 0.33	24.90 ± 0.75	20.52 ± 0.35
On-Demand FL (no noise)	10.30 ± 0.34	13.35 ± 0.25	34.92 ± 0.33
FedMLS (no noise)	24.82 ± 0.25	27.57 ± 0.44	$\mathbf{36.48 \pm 0.40}$
FedMLS ($\gamma = 0.1$)	24.54 ± 0.45	27.31 ± 0.28	36.38 ± 0.31
FedMLS ($\gamma = 0.5$)	24.36 ± 0.48	27.11 ± 0.11	29.72 ± 0.97
FedMLS ($\gamma = 1.0$)	24.21 ± 0.35	27.56 ± 0.46	28.94 ± 0.42
FedMLS ($\tau = 0.5$)	$\mathbf{25.08 \pm 0.41}$	$\mathbf{28.08 \pm 0.61}$	31.15 ± 0.65
FedMLS ($\tau = 2.0$)	23.98 ± 0.29	27.02 ± 0.30	29.07 ± 0.76
FedMLS ($\tau = 5.0$)	24.24 ± 0.32	27.33 ± 0.56	27.43 ± 0.56

Results and Robustness Analysis. Table 9 demonstrates classification performance on CIFAR-100 under varying levels of test label skew (α_{test}) and noisy estimation settings. At mild noise levels ($\gamma = 0.1$), performance degradation is negligible across all Dirichlet splits ($< 0.3\%$). Even under severe Gaussian noise ($\gamma = 1.0$), accuracy drops by less than 1% for mild-to-moderate skews ($\alpha_{\text{test}} \in \{1, 10\}$), confirming that client selection remains effective despite inaccurate prior estimates. Under severe skew ($\alpha_{\text{test}} = 0.1$), however, noise with $\gamma \geq 0.5$ lowers accuracy to 28.9–29.7%, below On-Demand FL, indicating that an accurate estimate of p_{test} matters most when the shift is large.

Regarding temperature-based misspecification, low temperature scaling ($\tau = 0.5$, which sharpens the target distribution) slightly boosts performance for $\alpha_{\text{test}} = 10$ and $\alpha_{\text{test}} = 1$ to 25.08% and 28.08%, respectively. While higher temperatures ($\tau \geq 2.0$, flattening the distribution) reduce accuracy under severe skew ($\alpha_{\text{test}} = 0.1$), FedMLS maintains robust performance (27.43%–31.15%) compared to standard baselines, underscoring its stability against test label distribution shifts.