# Personalized Federated Learning with Bidirectional Communication Compression via One-Bit Random Sketching

**Jiacheng Cheng[1], Xu Zhang[2*], Guanghui Qiu[3], Yifang Zhang[2,3], Yinchuan Li[4], Kaiyuan Feng[5]**
[1]School of Automation, Northwestern Polytechnical University, Xi'an, China
[2]School of Artificial Intelligence, Xidian University, Xi'an, China
[3]Science and Technology on Electronic Information Control Laboratory, China Academy of Electronics and Information Technology, Chengdu, China
[4]Knowin AI, Shenzhen, China
[5]Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an, China
zhang.xu@xidian.edu.cn

## Abstract

Federated Learning (FL) enables collaborative training across decentralized data, but faces key challenges of bidirectional communication overhead and client-side data heterogeneity. To address communication costs while embracing data heterogeneity, we propose **pFed1BS**, a novel **p**ersonalized **fed**erated learning framework that achieves extreme communication compression through **one-b**it random **s**ketching. In personalized FL, the goal shifts from training a single global model to creating tailored models for each client. In our framework, clients transmit highly compressed one-bit sketches, and the server aggregates and broadcasts a global one-bit consensus. To enable effective personalization, we introduce a sign-based regularizer that guides local models to align with the global consensus while preserving local data characteristics. To mitigate the computational burden of random sketching, we employ the Fast Hadamard Transform for efficient projection. Theoretical analysis guarantees that our algorithm converges to a stationary neighborhood of the global potential function. Numerical simulations demonstrate that pFed1BS substantially reduces communication costs while achieving competitive performance compared to advanced communication-efficient FL algorithms.

## Introduction

Federated Learning (FL) is an increasingly popular paradigm in deep learning, designed to train machine learning models on distributed client data while preserving privacy (McMahan et al. 2017; Li et al. 2020b). Although FL performs well under ideal conditions of *i.i.d.* data and unconstrained communication, real-world deployments face two fundamental challenges inherent to its decentralized nature.

First, data across clients is typically *non-i.i.d.*, reflecting diverse user behaviors and environments, which can severely degrade the performance of a single global model. Second, the communication overhead is often prohibitive, as repeatedly transmitting high-dimensional models between a central server and numerous clients is infeasible in bandwidth-limited networks. These challenges are particularly acute in critical application domains such as the massive Internet of Things (IoT), Vehicle-to-Everything (V2X) communications, and remote sensing networks. In these settings, where devices operate under extremely constrained bandwidth, efficient communication is not merely an optimization but a fundamental necessity for the system to be viable.

To overcome the challenge of *non-i.i.d.* data, Personalized Federated Learning (PFL) (T Dinh, Tran, and Nguyen 2020; Li et al. 2020a) has been proposed. Here, it is crucial to distinguish the goal of "personalization" from merely "addressing data heterogeneity." While the latter often aims to improve a single global model, PFL explicitly seeks to produce bespoke, customized models for individual clients that capture the unique characteristics of their local data. However, most PFL methods still incur significant communication costs by transmitting full-precision, high-dimensional model parameters or updates. This leads to a more precise and challenging research question: *can we design a federated learning algorithm that not only provides high-quality personalized models for each client but also operates efficiently under extreme bidirectional communication constraints?*

To reduce this burden, communication-efficient FL (CEFL) techniques have been developed, employing methods like prototype-learning (Tan et al. 2022), sparsification (Sattler et al. 2019; Liu et al. 2023) and quantization (Reisizadeh et al. 2020; Mao et al. 2022). Among the most aggressive are one-bit compression strategies. For instance, OBDA (Zhu et al. 2020) applies symmetric one-bit quantization for bidirectional communication, OBCSAA (Fan et al. 2022) combines a one-bit compressed sensing uplink with an uncompressed downlink, and zSignFed (Tang, Wang, and Chang 2024) stabilizes sign-based compression through a noisy perturbation scheme.

While these methods achieve remarkable compression rates, they primarily focus on training a single global model and thus overlook the critical challenge of data heterogeneity. To systematically analyze the limitations of existing work and motivate our contribution, we provide a compar-

*Corresponding author.

ison of representative algorithms in Table 1. The table reveals a clear research gap: no existing framework offers both extreme, bidirectional communication efficiency and native support for personalization.

In this paper, we bridge this gap by introducing pFed1BS, a novel PFL framework designed for extreme communication constraints based on one-bit random sketching. We formulate a joint optimization problem where clients learn personalized models by minimizing a local loss augmented with a sign-based regularizer. This regularizer encourages alignment with a global consensus vector derived by the server. Critically, our framework achieves bidirectional compression: clients upload one-bit sketches of their local models, and the server broadcasts a compact one-bit consensus vector. This contrasts sharply with prior CEFL methods that either compress only the uplink or require a full-model downlink. Figure 1 provides an overview of our proposed framework.

## Main Contributions

This paper proposes **pFed1BS**, as shown in Figure 1, a personalized federated learning framework designed for settings with extreme communication constraints. Our main contributions are as follows:

- We are the first to formulate the problem of personalized learning with one-bit bidirectional communication as a principled joint optimization problem. Our framework defines two coupled objectives: a client-side objective that balances local empirical risk with a novel sign-based regularizer for global alignment, and a server-side objective for optimally aggregating the compressed client signals.
- We make our framework practical for large-scale models by introducing a highly efficient implementation of the required sketching operations. By leveraging the Fast Hadamard Transform (FHT), we reduce the complexity of the client-side sketching operation from quadratic, $\mathcal{O}(mn)$, to near-linear, $\mathcal{O}(n \log n)$, without performance degradation.
- We provide the comprehensive convergence analysis for this challenging alternating optimization scheme. We formally prove that pFed1BS converges to a stationary neighborhood of a global potential function, rigorously accounting for the interplay between personalization, local stochastic updates, and errors introduced by one-bit sketching and server aggregation.
- We conduct extensive experiments on benchmark datasets (MNIST, FMNIST, CIFAR-10, CIFAR-100, and SVHN). Our results demonstrate that pFed1BS achieves a superior trade-off between accuracy and communication. Remarkably, pFed1BS matches or exceeds the performance of state-of-the-art one-bit FL algorithms while operating at a fraction of the communication cost and, crucially, providing personalization that they lack.

## Related Works

Our work is closely related to the following topics:

**Personalized Federated Learning.** To address the challenge of heterogeneous datasets, a rich body of work has emerged in Personalized Federated Learning (PFL). These approaches can be broadly categorized into local adaptation (Li et al. 2020a; T Dinh, Tran, and Nguyen 2020; Li et al. 2021; Zhang et al. 2022), multi-task learning (Smith et al. 2017; Marfoq et al. 2021), and architecture-based methods (Arivazhagan et al. 2019; Collins et al. 2021). For instance, regularization-based methods like pFedMe (T Dinh, Tran, and Nguyen 2020) and Ditto (Li et al. 2021) learn personalized models by augmenting the local objective with a proximal term that regularizes it towards a global model. Architecture-based methods, such as FedRep (Collins et al. 2021), learn a shared feature representation while personalizing the final model layers. More recently, DisPFL (Dai et al. 2022) learns personalized sparse masks for each client. However, these advanced PFL methods typically inherit the communication bottlenecks of standard FL, as they still presuppose the transmission of full-precision, high-dimensional model parameters or updates.

**Communication-Efficient Federated Learning.** In a parallel research thrust, numerous methods have been proposed to alleviate the communication burden in FL. One prominent approach is update sparsification, where only a fraction of the model update is transmitted, using techniques like Top-k selection (Sattler et al. 2019) or identifying parameters with high-magnitude changes (Long et al. 2024). Another major direction is quantization, which reduces the numerical precision of the transmitted updates (Reisizadeh et al. 2020; Chen and Vikalo 2024; Mao et al. 2022). A critical limitation of these popular techniques, however, is that they typically compress only the uplink (client-to-server) channel, still requiring the server to broadcast a full-precision, high-dimensional model.

More aggressive strategies leverage techniques from signal processing. Some works employ Compressed Sensing (CS) to project sparse updates into a low-dimensional subspace (Li, Li, and Varshney 2021; Oh et al. 2022, 2023). Others explore one-bit quantization schemes, sometimes combined with over-the-air computation in wireless settings, to achieve extreme compression (Zhu et al. 2020; Tang, Wang, and Chang 2024; Oh et al. 2024). While highly efficient, these methods are fundamentally designed to learn a single global model and lack mechanisms to handle data heterogeneity.

pFed1BS uniquely bridges these two disparate lines of research. It is the first work to integrate a bidirectional, one-bit sketching mechanism within a principled PFL formulation, thereby explicitly and simultaneously tackling the dual challenges of communication efficiency and data heterogeneity.

# The Proposed Method

This section presents our proposed method, pFed1BS. We begin by formulating the overall optimization framework that governs the collaborative learning process by introducing a sign-based regularizer and random sketching. Subsequently, we describe the iterative algorithm, specifying distinct client-server procedures for optimizing this objective.

| Algorithm | Upload Compression | | Download Compression | | Personalization Capability |
|---|---|---|---|---|---|
| | Dim. Reduction | 1-bit Quant. | Dim. Reduction | 1-bit Quant. | |
| FedAvg (McMahan et al. 2017) | × | × | × | × | × |
| OBDA (Zhu et al. 2020) | × | ✓ | × | ✓ | × |
| OBCSAA (Fan et al. 2022) | ✓ | ✓ | × | × | × |
| zSignFed (Tang, Wang, and Chang 2024) | × | ✓ | × | × | × |
| **pFed1BS** | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparisons of communication-efficient schemes and personalization Capabilities in Federated Learning Algorithms.

Finally, to address the computational bottleneck posed by the high-dimensional random sketching in local training, we introduce an efficient sketching method based on the Fast Hadamard Transform.

## Optimization Framework

In a federated learning system with $K$ clients, our goal is to move beyond learning a single global model and instead learn a personalized model $\boldsymbol{w}_k \in \mathbb{R}^n$ for each client $k \in \{1, \dots, K\}$. The system aggregates client contributions using weights $p_k$, which are typically set based on the client's dataset size, e.g., $p_k = N_k / \sum_{i=1}^K N_i$ , where $N_k$ is the number of local samples. The local objective for each client is to minimize the expected loss over its private data distribution $\mathcal{P}_k$

$$f_k(\boldsymbol{w}_k) = \mathbb{E}_{\xi_k \sim \mathcal{P}_k}[\hat{f}_k(\boldsymbol{w}_k; \xi_k)], \quad (1)$$

where $\xi_k$ is a data sample drawn from $\mathcal{P}_k$ and $\hat{f}_k(\boldsymbol{w}_k; \xi_k)$ is the loss function for a single sample $\xi_k$. In practice, the true gradient $\nabla f_k(\boldsymbol{w}_k)$ is intractable, so we rely on stochastic gradients computed on mini-batches to approximate it.

To drastically reduce communication costs, our core technical idea is to replace the transmission of high-dimensional models $\boldsymbol{w}_k$ with low-dimensional one-bit sketches. Specifically, each client transmits only sign($\boldsymbol{\Phi}\boldsymbol{w}_k$), where $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$ is a random projection matrix.

This radical compression necessitates a new way for the server and clients to interact. To this end, we introduce a global consensus vector $\boldsymbol{v} \in \{\pm 1\}^m$, which the server aggregates from the clients' one-bit sketches. This vector $\boldsymbol{v}$ is then broadcast back to clients and incorporated into their local optimization, which acts as a target, guiding the local model $\boldsymbol{w}_k$ to produce a projection $\boldsymbol{\Phi}\boldsymbol{w}_k$ whose signs align with $\boldsymbol{v}$.

To enforce this guidance, we introduce a sign-based regularizer, denoted by $g(\boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k)$, which measures the disagreement between the signs of the projected model $\boldsymbol{\Phi}\boldsymbol{w}_k$ and the global consensus $\boldsymbol{v}$. We define this regularizer using the one-sided $\ell_1$-norm:

$$g(\boldsymbol{x}, \boldsymbol{y}) = \|[\boldsymbol{x} \odot \boldsymbol{y}]_-\|_1, \text{ where } [x_i]_- = \min(x_i, 0). \quad (2)$$

For the client-side objective, where $\boldsymbol{v}$ is a sign vector and $\boldsymbol{\Phi}\boldsymbol{w}_k$ is real-valued, this regularizer is equivalent to:

$$g(\boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k) = \frac{1}{2}(\|\boldsymbol{\Phi}\boldsymbol{w}_k\|_1 - \langle \boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k \rangle). \quad (3)$$

Furthermore, to prevent the personalized models $\boldsymbol{w}_k$ from diverging or growing unbounded during local training due to the absence of global data constraints, we introduce an $\ell_2$ penalty term.

This leads to the following regularized client objective:

$$F_k(\boldsymbol{w}_k; \boldsymbol{v}) = f_k(\boldsymbol{w}_k) + \lambda g(\boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k) + \frac{\mu}{2}\|\boldsymbol{w}_k\|_2^2, \quad (4)$$

where $\lambda$ controls the strength of the sign alignment and $\mu$ controls the norm of model parameters.

This form, however, is non-smooth due to the $\ell_1$-norm in (3), posing a challenge for gradient-based optimization. To enable gradient-based optimization, we employ a continuously differentiable approximation for the $\ell_1$-norm. A standard approach is to approximate $\|\boldsymbol{z}\|_1$ with $h_\gamma(\boldsymbol{z})$, where $h_\gamma(\boldsymbol{z}) = \frac{1}{\gamma}\sum_{i=1}^m \log(\cosh(\gamma z_i))$. So we obtain a smoothed regularizer

$$\tilde{g}(\boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k) = h_\gamma(\boldsymbol{\Phi}\boldsymbol{w}_k) - \langle \boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k \rangle, \quad (5)$$

and a smoothed client-side objective $\tilde{F}_k(\boldsymbol{w}_k; \boldsymbol{v})$

$$\begin{aligned}\tilde{F}_k(\boldsymbol{w}_k; \boldsymbol{v}) &= f_k(\boldsymbol{w}_k) + \lambda \tilde{g}(\boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k) + \frac{\mu}{2}\|\boldsymbol{w}_k\|_2^2 \\ &= f_k(\boldsymbol{w}_k) + \lambda\left(h_\gamma(\boldsymbol{\Phi}\boldsymbol{w}_k) - \langle \boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k \rangle\right) + \frac{\mu}{2}\|\boldsymbol{w}_k\|_2^2, \quad (6)\end{aligned}$$

where the factor of $\frac{1}{2}$ can be absorbed into the hyperparameter $\lambda$.

The gradient of the smoothed penalty term with respect to $\boldsymbol{w}_k$ is then given by:

$$\nabla \tilde{g}(\boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k) = \boldsymbol{\Phi}^\top\left(\tanh(\gamma\boldsymbol{\Phi}\boldsymbol{w}_k) - \boldsymbol{v}\right). \quad (7)$$

As the smoothing parameter $\gamma \to \infty$, the term $\tanh(\gamma\boldsymbol{\Phi}\boldsymbol{w}_k) \approx \text{sign}(\boldsymbol{\Phi}\boldsymbol{w}_k)$. Therefore, the gradient effectively penalizes the misalignment between the signs of the projected local model and the global consensus vector $\boldsymbol{v}$, driving the local updates towards alignment.

Having defined how each client utilizes the global vector $\boldsymbol{v}$, we now address how the server generates it. At the end of each round, the server receives a one-bit sketch $\boldsymbol{z}_k = \text{sign}(\boldsymbol{\Phi}\boldsymbol{w}_k)$ from each participating client $k$. The server's task is to aggregate these sketches into a new consensus vector $\boldsymbol{v}^{t+1}$ that best represents the collective information.

We formulate this as an optimization problem where the server seeks to find a vector $\boldsymbol{v} \in \{\pm 1\}^m$ that minimizes the total weighted disagreement with the received client sketches:

$$\min_{\boldsymbol{v} \in \{\pm 1\}^m} \sum_{k=1}^K p_k \tilde{g}(\boldsymbol{v}, \boldsymbol{z}_k). \quad (8)$$

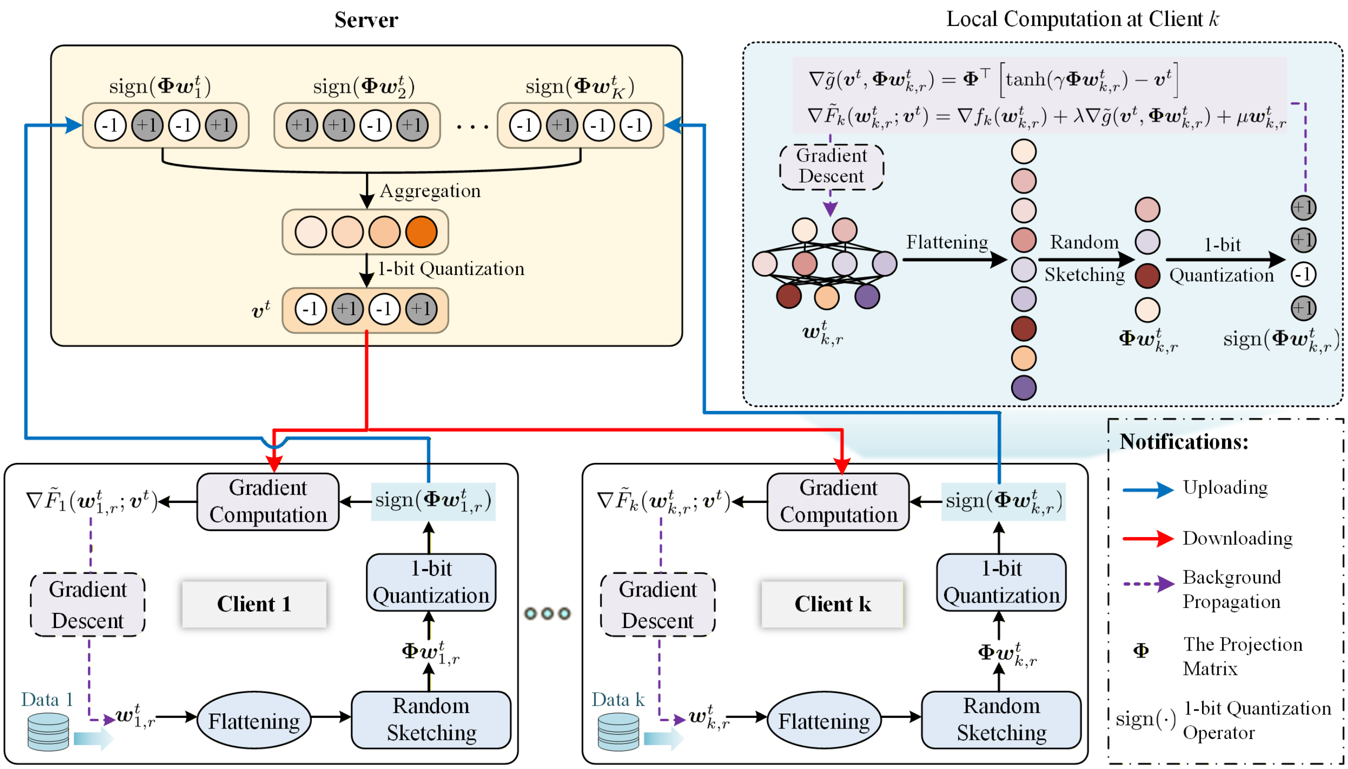


Figure 1: An overview of our proposed framework. At round $t$, each client performs a local update using a sign-based regularizer with the global one-bit vector $v^t$. Then each client projects and quantizes the updated local model to one bit vector $\text{sign}(\Phi w_k^{t+1})$, and then transmits it to the server. The server aggregates all clients' one-bit vectors to form the next global one-bit vector $v^{t+1}$, which is broadcast for the next round.

Overall, we formulate a bilevel optimization problem as follows

$$\textbf{Server:} \quad \min_{v \in \{\pm 1\}^m} \sum_{k=1}^{K} p_k \tilde{g}(v, \text{sign}(\Phi w_k^\star(v)))$$

$$\textbf{Clients:} \quad w_k^\star(v) \in \arg\min_{w_k \in \mathbb{R}^n} F_k(w_k; v) = f_k(w_k) + \lambda \tilde{g}(v, \Phi w_k) + \frac{\mu}{2}\|w_k\|_2^2, \tag{9}$$

where $w_k^\star(v)$ denotes the optimal solution of the $k$-th low-level problem for a given upper-level variable $v$.

## Algorithm

To solve the joint optimization problem defined in Eq. (9), we propose an alternating optimization scheme named pFed1BS. The core idea is to iteratively perform (i) local optimization on the client side to update the personalized models $w_k$ and (ii) global aggregation on the server side to update the consensus vector $v$. The overall procedure is presented in Algorithm 1. We describe the specifics of each component below.

At the start of round $t$, each participating client $k \in \mathcal{S}^t$ receives the global consensus vector $v^t$. The client's goal is to update its local model $w_k^t$ by approximately minimizing its smoothed objective $\tilde{F}_k(w_k; v^t)$ (from Eq. (6)). As shown in Line 16 of Algorithm 1, the client performs $R$ steps of stochastic gradient descent. For local step $r$, the update is

$$w_{k,r+1} = w_{k,r} - \eta \nabla \tilde{F}_k(w_{k,r}; v^t). \tag{10}$$

The gradient $\nabla \tilde{F}_k$ is composed of the standard task gradient and our regularization terms:

$$\nabla \tilde{F}_k(w_{k,r}^t; v^t) = \nabla f_k(w_{k,r}^t) + \lambda \Phi^\top \left[\tanh(\gamma \Phi w_{k,r}^t) - v^t\right] + \mu w_{k,r}^t. \tag{11}$$

Since the calculation of $\nabla f_k(w_{k,r}^t)$ is intractable, we use the mean over a mini-batch of data $\mathcal{B}_{k,r}$ from $\mathcal{P}_k$

$$\nabla \hat{f}_k(w_{k,r}^{t+1}; \mathcal{B}_{k,r}) = \frac{1}{|\mathcal{B}_{k,r}|} \sum_{\xi_k \in \mathcal{B}_{k,r}} \nabla \hat{f}_k(w_{k,r}^{t+1}; \xi_k). \tag{12}$$

After $R$ steps, the client transmits its new one-bit sketch $z_k^{t+1} = \text{sign}(\Phi w_k^{t+1})$ to the server. Upon receiving the sketches $\{z_k^{t+1}\}_{k \in \mathcal{S}^t}$, the server generates the next consensus vector $v^{t+1}$. As defined in Eq. (8), the server aims to solve:

$$\min_{v \in \{\pm 1\}^m} \sum_{k \in \mathcal{S}^t} p_k g(v, z_k^{t+1}). \tag{13}$$

Crucially, this discrete optimization problem admits an exact, closed-form solution. The following lemma states that the optimal aggregation is a simple weighted majority vote.

**Algorithm 1**: pFed1BS: Personalized Federated Learning via One-Bit Random Sketching

1: **Input:** Total rounds $T$, local steps $R$, learning rate $\eta$, regularization hyperparameters $\lambda$, $\mu$
2: **Server Initializes:** Model $\boldsymbol{w}^0$, random seed $I$. Broadcasts $I$ to all clients. Initializes $\boldsymbol{v}^0 = \mathbf{0}$.
3: **for** $t = 0$ to $T-1$ **do**
4:   **for** $k = 1$ to $K$ **in parallel do**
5:     $\boldsymbol{z}_k^{t+1}, \boldsymbol{w}_k^{t+1} \leftarrow \texttt{ClientUpdate}(k, \boldsymbol{w}_k^t, \boldsymbol{v}^t)$
6:   **end for**
7:   Random sample a subset of clients $\mathcal{S}^t$
8:   **Aggregate signs:** $\boldsymbol{v}^{t+1} = \text{sign}\left(\sum_{k \in \mathcal{S}^t} p_k \boldsymbol{z}_k^{t+1})\right)$
9: **end for**
10: **Function** $\texttt{ClientUpdate}(k, \boldsymbol{w}_k^t, \boldsymbol{v}^t)$:
11: **Client** $k$ **Initializes:** $\boldsymbol{w}_{k,0}^{t+1} = \boldsymbol{w}_k^t$.
12: **for** $r = 0$ to $R-1$ **do**
13:   Sample a mini-batch $\mathcal{B}_{k,r}$ from data distribution $\mathcal{P}_k$
14:   Compute gradient: $\nabla \hat{f}_k(\boldsymbol{w}_{k,r}^{t+1}; \mathcal{B}_{k,r}) = \frac{1}{|\mathcal{B}_{k,r}|} \sum_{\xi_k \in \mathcal{B}_{k,r}} \nabla \hat{f}_k(\boldsymbol{w}_{k,r}^{t+1}; \xi_k)$
15:   Compute regularization subgradient: $\nabla \tilde{g}(\boldsymbol{v}^t, \boldsymbol{\Phi} \boldsymbol{w}_{k,r}^{t+1}) = \boldsymbol{\Phi}^\top (\tanh(\gamma \boldsymbol{\Phi} \boldsymbol{w}_{k,r}^{t+1}) - \boldsymbol{v}^t)$
16:   Update local model: $\boldsymbol{w}_{k,r+1}^{t+1} \leftarrow \boldsymbol{w}_{k,r}^{t+1} - \eta(\nabla \hat{f}_k(\boldsymbol{w}_{k,r}^{t+1}; \mathcal{B}_{k,r}) + \lambda \nabla \tilde{g}(\boldsymbol{v}^t, \boldsymbol{\Phi} \boldsymbol{w}_{k,r}^{t+1}) + \mu \boldsymbol{w}_{k,r}^{t+1})$
17: **end for**
18: **return** $\text{sign}(\boldsymbol{\Phi} \boldsymbol{w}_{k,R}^{k+1})$, $\boldsymbol{w}_{k,R}^{k+1}$

**Lemma 1** (Optimal Server Aggregation)**.** *The unique minimizer of the server objective in Eq.* (13) *is given by:*

$$\boldsymbol{v}^* = \text{sign}\left(\sum_{k \in \mathcal{S}^t} p_k \boldsymbol{z}_k^{t+1}\right). \tag{14}$$

This result is a straightforward but significant application of optimization principles, ensuring that our aggregation step (Line 8 in Algorithm 1) is not a heuristic but is guaranteed to be optimal given the information available to the server.

### Efficient Projection via Fast Hadamard Transform

A naive implementation of the projection $\boldsymbol{\Phi w}$ using a dense Gaussian matrix requires $\mathcal{O}(mn)$ computation and memory, which causes a computational burden for large models ($n \gg 10^6$). To ensure scalability, we employ a structured projection based on the Subsampled Randomized Hadamard Transform (SRHT) (Zhang, Jiao, and Xu 2010), which reduces the complexity to $\mathcal{O}(n \log n)$.

**Forward Projection.** The forward projection $\boldsymbol{z} = \boldsymbol{\Phi w} \in \mathbb{R}^m$ is computed through a sequence of efficient operations, as illustrated in Figure 2 (b) (Up). Let $n$ denote the model dimension and $m \ll n$ be the target dimension. First, the input vector $\boldsymbol{w} \in \mathbb{R}^n$ is zero-padded to the next power-of-two dimension, $n' = 2^{\lceil \log_2 n \rceil}$. This can be represented as:

$$\tilde{\boldsymbol{w}} = \boldsymbol{P}_{\text{pad}} \boldsymbol{w} = \begin{bmatrix} \boldsymbol{w} \\ \mathbf{0} \end{bmatrix}, \tag{15}$$

where $\boldsymbol{P}_{\text{pad}}$ can be viewed as an $n' \times n$ matrix consisting of the identity matrix $\boldsymbol{I}_n$ stacked on top of an $(n'-n) \times n$ zero matrix, and $\tilde{\boldsymbol{w}} \in \mathbb{R}^{n'}$. This step ensures compatibility with the Fast Hadamard Transform (FHT), which requires input lengths that are powers of 2.

Then the structured random projection is performed as follows

$$\boldsymbol{\Phi w} = \boldsymbol{S}' \boldsymbol{H D} \tilde{\boldsymbol{w}}, \tag{16}$$

where $\boldsymbol{D} \in \mathbb{R}^{n' \times n'}$ is a diagonal matrix where each diagonal entry $\boldsymbol{D}_{ii}$ is an independent random variable drawn uniformly from $\{-1, +1\}$, $\boldsymbol{H} \in \mathbb{R}^{n' \times n'}$ is the normalized Walsh-Hadamard matrix, and $\boldsymbol{S}' = \sqrt{\frac{n'}{m}} \boldsymbol{S} \in \mathbb{R}^{m \times n'}$, where $S$ is a subsampling matrix that uniformly selects $m$ rows at random from the $n'$-dimensional identity matrix. This structured operator avoids forming any dense matrices. The total computational complexity is dominated by the FHT, resulting in an efficient $\mathcal{O}(n \log n)$ computation.

**Backward Computation.** In the backward pass, we compute the adjoint projection $\boldsymbol{\Phi}^\top \boldsymbol{v}$, where $\boldsymbol{v} \in \mathbb{R}^m$ denotes a low-dimensional one-bit vector.

Specifically, we first lift $\boldsymbol{v} \in \mathbb{R}^m$ to $\tilde{\boldsymbol{v}} \in \mathbb{R}^{n'}$ by zero-padding as follows

$$\tilde{\boldsymbol{v}} = \boldsymbol{S}'^\top \boldsymbol{v}. \tag{17}$$

This operation places the elements of $\boldsymbol{v}$ into the coordinates that were selected by $\boldsymbol{S}$, filling the rest with zeros. We then apply the FHT to the padded vector, and apply the same random sign flips as in the forward pass as $\boldsymbol{D H} \tilde{\boldsymbol{v}}$.

Finally, we truncate the resulting $n'$-dimensional vector back to the original dimensions. This is the action of $\boldsymbol{P}_{\text{pad}}^\top$. The operator $\boldsymbol{P}_{\text{pad}}^\top : \mathbb{R}^{n'} \to \mathbb{R}^n$ is a truncation operator, which we denote as $\boldsymbol{P}_{\text{trunc}}$. Its action is to select the first $n$ coordinates of a vector in $\mathbb{R}^{n'}$ and discard the remaining $n' - n$ coordinates.

The complete operation is formulated as

$$\boldsymbol{\Phi}^\top \boldsymbol{v} = (\boldsymbol{S}' \boldsymbol{H D P}_{\text{pad}})^\top \boldsymbol{v} = \boldsymbol{P}_{\text{pad}}^\top \boldsymbol{D}^\top \boldsymbol{H}^\top \boldsymbol{S}'^\top \boldsymbol{v}. \tag{18}$$

Given that $\boldsymbol{P}_{\text{trunc}} = \boldsymbol{P}_{\text{pad}}^\top$ and $\boldsymbol{D}$ is diagonal, this simplifies to $\boldsymbol{\Phi}^\top \boldsymbol{v} = \boldsymbol{P}_{\text{trunc}} \boldsymbol{D H}^\top \boldsymbol{S}'^\top \boldsymbol{v}$.

Similar to the forward projection, the adjoint projection is matrix-free and has a computational complexity of $\mathcal{O}(n \log n)$, making our algorithm practical and efficient.

## Theoretical Analysis

We provide a rigorous convergence guarantee for pFed1BS, showing that the algorithm converges to a neighborhood of a stationary point.

### Assumptions and Preliminaries

Our analysis relies on a set of standard assumptions common in FL literature.

**Assumption 1** ($L$-smoothness)**.** *The local objective function $f_k(\cdot)$ is L-smooth, i.e., its gradient is L-Lipschitz continuous.*

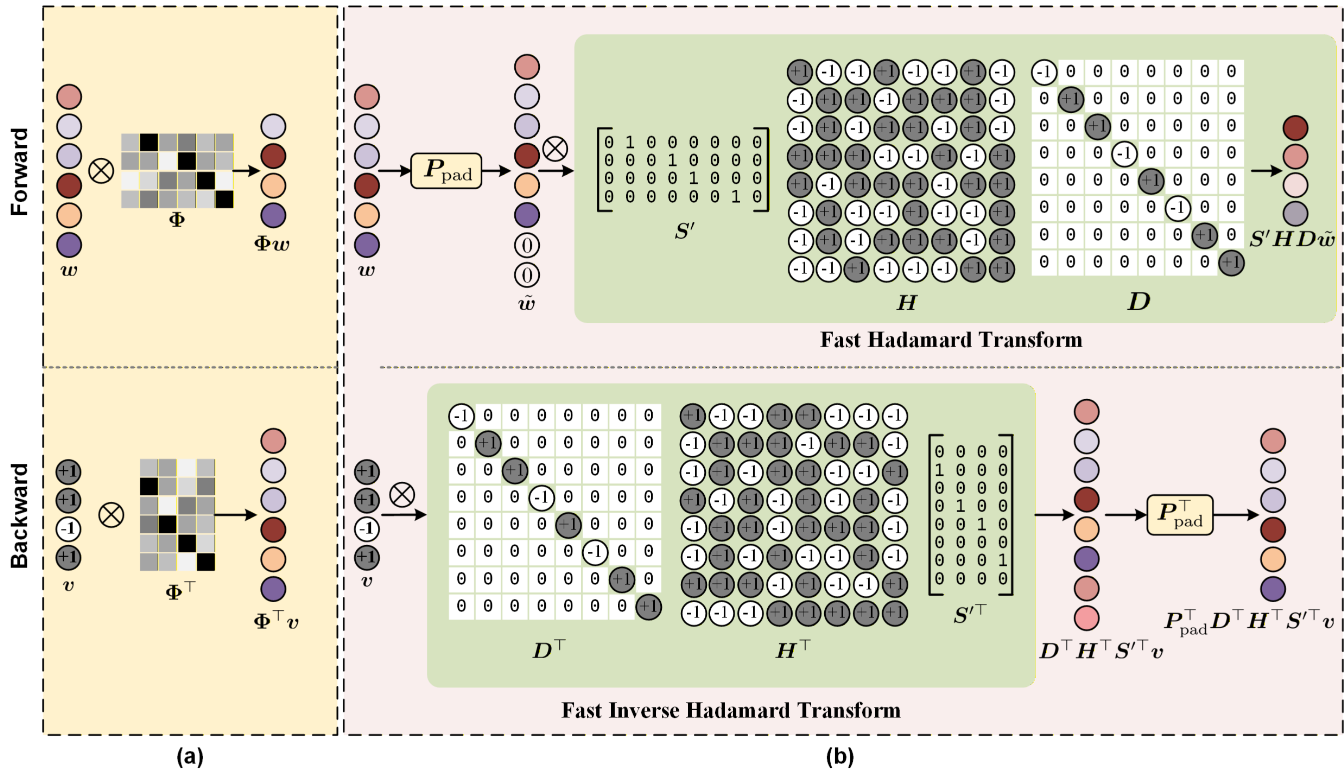


Figure 2: Comparison between (a) a dense random projection and (b) our efficient structured projection. The structured projection sequentially applies element-wise random sign flips ($\boldsymbol{D}$), a Fast Hadamard Transform ($\boldsymbol{H}$), and random subsampling ($\boldsymbol{S}$).

**Assumption 2** (Bounded Below)**.** *The global potential function $\Psi(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K; \boldsymbol{v}) = \sum_{k=1}^{K} p_k \tilde{F}_k(\boldsymbol{w}_k; \boldsymbol{v})$ is bounded below by some value $F^*$.*

**Assumption 3** (Bounded Gradient Variance)**.** *The variance of the stochastic gradients is uniformly bounded, i.e., for any client $k$ and model $\boldsymbol{w}$, $\mathbb{E}_{\mathcal{B}}\left[\|\nabla \hat{f}_k(\boldsymbol{w}; \mathcal{B}) - \nabla f_k(\boldsymbol{w})\|^2\right] \leq \sigma^2$ for some constant $\sigma^2 \geq 0$.*

**Assumption 4** (Bounded Task Gradient Variance)**.** *The stochastic gradient of the task loss $f_k$ has a bounded second moment, i.e., there exists a constant $G > 0$ such that $\mathbb{E}_{\mathcal{B}}\left[\|\nabla \hat{f}_k(\boldsymbol{w}; \mathcal{B})\|^2\right] \leq G^2$ for all $\boldsymbol{w}$ and $k$.*

Our analysis of partial client participation also relies on the following standard lemmas.

**Lemma 2** (Bounded Projection Norm)**.** *Let the projection matrix $\boldsymbol{\Phi} \in \mathcal{R}^{m \times n}$ be constructed as described in the "Efficient Projection" section, involving a normalized Hadamard matrix $\boldsymbol{H}$, a random sign matrix $\boldsymbol{D}$, a subsampling matrix $\boldsymbol{S}$ and a padding matrix $\boldsymbol{P}_{pad}$. The resulting operator has an exact spectral norm given by:*

$$\|\boldsymbol{\Phi}\| \leq \sqrt{\frac{n'}{m}}. \tag{19}$$

For our analysis, we formally define $C_\Phi = \sqrt{\frac{n'}{m}}$, where $C_\Phi = \mathcal{O}\left(\sqrt{\frac{n}{m}}\right)$.

**Lemma 3** (Client-Side Objective and Gradient)**.** *The client-side objective $\tilde{F}_k$ for a given client $k$ and server message $\boldsymbol{v}$ is defined as:*

$$\tilde{F}_k(\boldsymbol{w}_k; \boldsymbol{v}) = f_k(\boldsymbol{w}_k) + \lambda\left(h_\gamma(\boldsymbol{\Phi}\boldsymbol{w}_k) - \langle \boldsymbol{v}, \boldsymbol{\Phi}\boldsymbol{w}_k \rangle\right) + \frac{\mu}{2}\|\boldsymbol{w}_k\|_2^2, \tag{20}$$

*where $h_\gamma(\boldsymbol{z}) = \frac{1}{\gamma}\sum_{i=1}^{m} \log(\cosh(\gamma z_i))$ is a differentiable surrogate for the $\ell_1$-norm. The objective $\tilde{F}_k(\boldsymbol{w}_k)$ is differentiable with respect to $\boldsymbol{w}_k$, and its gradient is given by:*

$$\nabla \tilde{F}_k(\boldsymbol{w}_k; \boldsymbol{v}) = \nabla f_k(\boldsymbol{w}_k) + \lambda \boldsymbol{\Phi}^\top \left[\tanh(\gamma \boldsymbol{\Phi}\boldsymbol{w}_k) - \boldsymbol{v}\right] + \mu \boldsymbol{w}_k. \tag{21}$$

**Lemma 4** (Smoothness of Client Objective)**.** *Under Assumptions 1 and 1, the client-side objective $\tilde{F}_k(\boldsymbol{w}_k; \boldsymbol{v})$ is $L_F$-smooth with respect to $\boldsymbol{w}_k$, where the smoothness constant is given by:*

$$L_F = L + \lambda\gamma C_\Phi^2 + \mu. \tag{22}$$

**Lemma 5** (Bounded Model Norm)**.** *Let Assumption 1-4 hold. For a learning rate $\eta$ satisfying $\eta < \frac{1}{6\mu}$, the expected squared norm of the client weights is uniformly bounded across all rounds $t$:*

$$\mathbb{E}[\|\boldsymbol{w}_k^t\|_2^2] \leq W^2, \quad \forall t \geq 0, \tag{23}$$

*where the bound $W^2$ is defined as*

$$W^2 \triangleq \|\boldsymbol{w}_k^0\|_2^2 + \frac{C'}{(1-\alpha)(1-\alpha^R)} \tag{24}$$

*with constants $\alpha = 1 - \eta\mu(1/2 - 3\eta\mu)$ and $C'$ given by:*

$$C' \triangleq \left(\frac{2\eta}{\mu} + 3\eta^2\right) G^2 + 4\Big(\frac{\eta}{\mu} + 3\eta^2\Big)\lambda^2 C_\Phi^2 m. \tag{25}$$

**Lemma 6** (Variance of Client Sampling)**.** *Let $\{\boldsymbol{z}_k^t\}_{k=1}^K$ be the set of client sketches at round t. If a subset $\mathcal{S}^t$ of size $S$ is sampled uniformly at random without replacement, then the variance of the sample mean is bounded by:*

$$\mathbb{E}_{\mathcal{S}^t}\left[\left\|\frac{1}{S}\sum_{k\in\mathcal{S}^t} p_k \boldsymbol{z}_k^t - \bar{\boldsymbol{z}}^t\right\|_2^2\right] \leq \frac{K-S}{SK(K-1)}\sum_{k=1}^K \left\|p_k \boldsymbol{z}_k^t - \bar{\boldsymbol{z}}^t\right\|_2^2, \tag{26}$$

*where $\bar{\boldsymbol{z}}^t \triangleq \frac{1}{K}\sum_{k=1}^K p_k \boldsymbol{z}_k^t$.*

**Lemma 7** (Client-Side Objective Descent)**.** *After $R$ local steps of subgradient descent with learning rate $\eta$ on the smoothed objective $\tilde{F}_k(\cdot;\boldsymbol{v}^t)$, starting from $\boldsymbol{w}_{k,0}^{t+1} = \boldsymbol{w}_k^t$, we have*

$$\mathbb{E}\left[\tilde{F}_k(\boldsymbol{w}_{k,R}^{t+1};\boldsymbol{v}^t)\right] \leq \tilde{F}_k(\boldsymbol{w}_k^t;\boldsymbol{v}^t) + \frac{\eta^2 R L_F \sigma^2}{2} - \eta R\left(1 - \frac{\eta L_F}{2}\right)\cdot\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla\tilde{F}_k(\boldsymbol{w}_{k,r}^{t+1};\boldsymbol{v}^t)\right\|_2^2. \tag{27}$$

## Main Convergence Result

Our proof relies on several key intermediate results (proofs in the Appendix). These are combined to analyze the evolution of a global potential function:

$$\Psi^t \triangleq \sum_{k=1}^K p_k \tilde{F}_k(\boldsymbol{w}_k^t;\boldsymbol{v}^t). \tag{28}$$

Our main result bounds the time-averaged expected squared norm of the local gradients.

**Theorem 1** (Local Convergence)**.** *Under standard assumptions for federated learning analysis (detailed in the Appendix), if the learning rate satisfies $\eta \leq \frac{1}{L_F}$, after $T$ communication rounds with partial client participation, Algorithm 1 guarantees the following convergence*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\sum_{k=1}^K p_k\|\nabla\tilde{F}_k(\boldsymbol{w}_{k,r}^{t+1};\boldsymbol{v}^t)\|_2^2\right] \leq \frac{\Psi^0 - F^*}{c_1 T} + \frac{\eta^2 R L_F\sigma^2}{2c_1} + \frac{\Delta_{\max}}{c_1} + \frac{\lambda E_S}{c_1}, \tag{29}$$

*where $c_1 = \eta R(1 - \eta L_F/2)$. The term $\Psi^0$ is the initial value of the potential function in* (28)*, $F^*$ is its lower bound, $L_F = L + \lambda\gamma C_\Phi^2 + \mu$ is the smoothness constant of the client objective, $\Delta_{\max} = 2\lambda(\sqrt{m}WC_\Phi + m)$ bounds the error from the one-bit server update, and $E_S$ bounds the error from client sampling, defined as*

$$E_S = 2m\sqrt{\frac{K(K-S)}{S(K-1)}\sum_{k=1}^K p_k^2}. \tag{30}$$

**Remark 1.** *Theorem 1 shows that as $T$ grows, the average squared gradient norm converges to a neighborhood of zero at an $\mathcal{O}(1/(RT))$ rate, which means that the algorithm converges to a stationary point of the global potential function. The neighborhood (convergence error) of pFed1BS is governed by stochastic noise $\mathcal{O}(\eta L_F\sigma^2)$, communication error $\mathcal{O}(\Delta_{\max}/(\eta R))$ and client sampling error $\mathcal{O}(\lambda E_S/(\eta R))$. To bound the convergence error, the regularization parameter $\lambda$ must satisfy $\lambda = \mathcal{O}(1/n)$, which simultaneously controls $L_F$, $\Delta_{\max}$ and $E_S$.*

**Remark 2.** *Note that the client sampling error $E_S$ vanishes when $S = K$, i.e., full client participation. In this case, our convergence bound recovers the result for the full participation setting.*

# Experiments

We empirically evaluate pFed1BS against state-of-the-art baselines on several benchmarks under challenging *non-i.i.d.* conditions.

## Experimental Setup

**Datasets and Models:** Our experiments use standard image classification benchmarks: MNIST, FMNIST, CIFAR-10, CIFAR-100, and SVHN. To further assess generalizability, we also include experiments on SVHN. We use a two-layer MLP for MNIST and FMNIST, and VGG architectures for the other datasets. We simulate a highly *non-i.i.d.* environment by partitioning data among 20 clients based on labels.

**Baselines:** We compare pFed1BS against FedAvg and several one-bit CEFL methods: OBDA, OBCSAA, and zSignFed. To provide a comprehensive evaluation, we also include state-of-the-art PFL and communication-efficient baselines: EDEN (Vargaftik et al. 2022) and FedBAT(Li et al. 2024).

**Implementation Details:** We run experiments for 100-300 communication rounds with multiple local epochs. Key hyperparameters were set via a grid search to $\lambda = 0.0005$, $\mu = 0.00001$, and $\gamma = 10000$. The compression ratio is fixed at $m/n = 0.1$. All experiments are implemented in PyTorch and run on an NVIDIA RTX 3090 Ti GPU, with results averaged over 10 independent runs.

## Evaluation Metrics

We use the following metrics for evaluation:

- **Top-1 Accuracy:** We report the average Top-1 test accuracy on the held-out test set, aggregated across all clients' personalized models.
- **Communication Cost:** We define the per-round communication cost as the total number of bits transmitted between the server and all participating clients in a single

| Method | MNIST | | FMNIST | | CIFAR-10 | | CIFAR-100 | | SVHN | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. (%) | Cost (MB) | Acc. (%) | Cost (MB) | Acc. (%) | Cost (MB) | Acc. (%) | Cost (MB) | Acc. (%) | Cost (MB) |
| FedAvg | $97.21 \pm 0.48$ | 31.06 | $84.40 \pm 0.09$ | 31.06 | $87.78 \pm 1.45$ | 42.85 | $59.60 \pm 0.66$ | 2335.85 | $96.33 \pm 0.28$ | 42.85 |
| OBDA | $92.54 \pm 0.32$ | $0.97_{\downarrow 96.88\%}$ | $78.51 \pm 0.62$ | $0.97_{\downarrow 96.88\%}$ | $73.26 \pm 5.39$ | $1.34_{\downarrow 96.88\%}$ | $42.47 \pm 2.02$ | $72.95_{\downarrow 96.88\%}$ | $84.32 \pm 1.15$ | $1.34_{\downarrow 96.88\%}$ |
| OBCSAA | $92.20 \pm 0.20$ | $15.58_{\downarrow 49.84\%}$ | $80.13 \pm 0.45$ | $15.58_{\downarrow 49.84\%}$ | $83.57 \pm 0.14$ | $21.49_{\downarrow 49.84\%}$ | $48.99 \pm 0.54$ | $1171.57_{\downarrow 49.84\%}$ | $87.10 \pm 0.65$ | $21.49_{\downarrow 49.84\%}$ |
| zSignFed | $94.83 \pm 0.07$ | $16.01_{\downarrow 48.45\%}$ | $82.55 \pm 0.28$ | $16.01_{\downarrow 48.45\%}$ | $67.60 \pm 3.46$ | $22.04_{\downarrow 48.56\%}$ | $40.17 \pm 2.32$ | $1203.78_{\downarrow 48.45\%}$ | $85.33 \pm 1.05$ | $22.04_{\downarrow 48.56\%}$ |
| EDEN | $96.50 \pm 0.35$ | $12.15_{\downarrow 60.88\%}$ | $83.85 \pm 0.30$ | $12.15_{\downarrow 60.88\%}$ | $84.91 \pm 0.53$ | $22.76_{\downarrow 46.88\%}$ | $47.55 \pm 1.12$ | $1205.33_{\downarrow 48.39\%}$ | $89.01 \pm 0.48$ | $22.76_{\downarrow 46.88\%}$ |
| FedBAT | $96.42 \pm 0.41$ | $11.88_{\downarrow 61.75\%}$ | $83.70 \pm 0.35$ | $11.88_{\downarrow 61.75\%}$ | $81.20 \pm 0.95$ | $22.04_{\downarrow 48.56\%}$ | $46.89 \pm 1.25$ | $1198.71_{\downarrow 48.68\%}$ | $88.89 \pm 0.51$ | $22.04_{\downarrow 48.56\%}$ |
| **pFed1BS** | $\mathbf{97.83} \pm \mathbf{0.02}$ | $\mathbf{0.10}_{\downarrow \mathbf{99.68\%}}$ | $\mathbf{84.15} \pm \mathbf{0.21}$ | $\mathbf{0.10}_{\downarrow \mathbf{99.68\%}}$ | $\mathbf{85.21} \pm \mathbf{0.34}$ | $\mathbf{0.13}_{\downarrow \mathbf{99.69\%}}$ | $\mathbf{52.88} \pm \mathbf{0.32}$ | $\mathbf{7.30}_{\downarrow \mathbf{99.69\%}}$ | $\mathbf{95.07} \pm \mathbf{0.21}$ | $\mathbf{0.13}_{\downarrow \mathbf{99.69\%}}$ |

Table 2: Top-1 accuracy (%) and one-round communication cost (MB) of FL algorithms on various datasets under a Non-IID setting. Best results in each column are highlighted in **bold**. The row for our proposed method, pFed1BS, is highlighted with a gray background for emphasis.

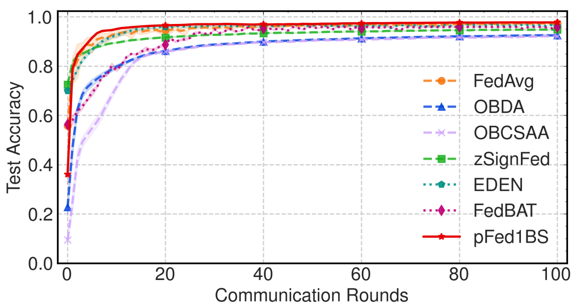


Figure 3: Test accuracy on MNIST (*non-i.i.d.*). pFed1BS achieves both faster convergence and higher final accuracy.

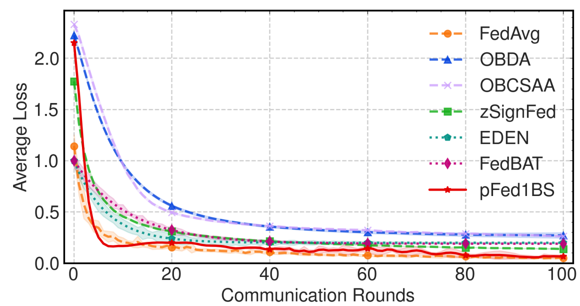


Figure 4: Average training loss on MNIST (*non-i.i.d.*). pFed1BS exhibits significantly faster convergence to a lower loss value and maintains stable behavior throughout the training process.

round. For pFed1BS, this is the sum of all uplink one-bit sketches (size $m$) and the downlink one-bit consensus vector (size $m$).

## Main Results

Table 2 presents the final test accuracy and per-round communication cost. pFed1BS establishes a new state-of-the-art for communication-constrained FL. On all datasets, it achieves accuracy that is highly competitive with or superior to all baselines, including full-precision FedAvg and advanced communication-efficient methods like OBDA, while reducing communication costs by over $96\%$. For instance, on CIFAR-10, pFed1BS achieves $85.21\%$ accuracy with only $0.13$ MB per round, whereas OBDA requires $1.34$ MB for a similar $73.26\%$ accuracy.

On more challenging datasets like CIFAR-100, the advantage is even more pronounced: one-bit baselines suffer a performance collapse, while pFed1BS maintains high accuracy, demonstrating the critical role of personalization in enabling extreme compression. The convergence plots in Figures 3 and 4 further show that pFed1BS achieves faster convergence to a better and more stable solution.

## Conclusion

In this work, we proposed pFed1BS, a novel personalized federated learning framework that successfully reconciles the dual challenges of extreme communication compression and data heterogeneity. By integrating a bidirectional one-bit sketching mechanism with a principled sign-based regularizer, our method reduces communication costs by over $99\%$ while simultaneously achieving state-of-the-art accuracy. Crucially, pFed1BS achieves this extreme compression with only a minimal trade-off in model accuracy compared to full-precision methods, while decisively outperforming other one-bit baselines that suffer a significant performance collapse in *non-i.i.d.* settings. Our work demonstrates that a carefully designed personalization strategy is the key to making extreme compression schemes viable, preventing catastrophic performance loss and establishing a new, practical frontier for deploying powerful federated models in real-world, resource-constrained environments.


## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62501432, Science and Technology on Electronic Information Control Laboratory, the Postdoctoral Fellowship Program of CPSF under Grant No. GZC20232038, and the China Postdoctoral Science Foundation under Grant No. 2024M762521.

## References


Arivazhagan, M. G.; Aggarwal, V.; Singh, A. K.; and Choudhary, S. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.

Chen, H.; and Vikalo, H. 2024. Mixed-precision quantization for federated learning on resource-constrained heterogeneous devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6138–6148.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, 2089–2099. PMLR.

Dai, R.; Shen, L.; He, F.; Tian, X.; and Tao, D. 2022. Dispfl: Towards communication-efficient personalized federated learning via decentralized sparse training. *arXiv preprint arXiv:2206.00187*.

Deng, Y.; Kamani, M. M.; and Mahdavi, M. 2020. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*.

Fallah, A.; Mokhtari, A.; and Ozdaglar, A. 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33: 3557–3568.

Fan, X.; Wang, Y.; Huo, Y.; and Tian, Z. 2022. 1-bit compressive sensing for efficient federated learning over the air. *IEEE Transactions on Wireless Communications*, 22(3): 2139–2155.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S. J.; Stich, S. U.; and Suresh, A. T. 2020. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, 5132–5143.

Li, C.; Li, G.; and Varshney, P. K. 2021. Communication-efficient federated learning based on compressed sensing. *IEEE Internet of Things Journal*, 8(20): 15531–15541.

Li, S.; Xu, W.; Wang, H.; Tang, X.; Qi, Y.; Xu, S.; Luo, W.; Li, Y.; He, X.; and Li, R. 2024. FedBAT: communication-efficient federated learning via learnable binarization. *arXiv preprint arXiv:2408.03215*.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, 6357–6368. PMLR.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020a. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2: 429–450.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020b. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

Liu, X.; Li, Y.; Wang, Q.; Zhang, X.; Shao, Y.; and Geng, Y. 2023. Sparse personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 12027–12041.

Long, Z.; Chen, Y.; Dou, H.; Zhang, Y.; and Chen, Y. 2024. FedSQ: Sparse-quantized federated learning for communication efficiency. *IEEE Transactions on Consumer Electronics*, 70(1): 4050–4061.

Mao, Y.; Zhao, Z.; Yan, G.; Liu, Y.; Lan, T.; Song, L.; and Ding, W. 2022. Communication-efficient federated learning with adaptive quantization. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4): 1–26.

Marfoq, O.; Neglia, G.; Bellet, A.; Kameni, L.; and Vidal, R. 2021. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34: 15434–15447.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282. PMLR.

Oh, J.; Lee, D.; Won, D.; Noh, W.; and Cho, S. 2024. Communication-efficient federated learning over-the-air with sparse one-bit quantization. *IEEE Transactions on Wireless Communications*.

Oh, Y.; Jeon, Y.-S.; Chen, M.; and Saad, W. 2023. FedVQCS: Federated learning via vector quantized compressed sensing. *IEEE Transactions on Wireless Communications*, 23(3): 1755–1770.

Oh, Y.; Lee, N.; Jeon, Y.-S.; and Poor, H. V. 2022. Communication-efficient federated learning via quantized compressed sensing. *IEEE Transactions on Wireless Communications*, 22(2): 1087–1100.

Reisizadeh, A.; Mokhtari, A.; Hassani, H.; Jadbabaie, A.; and Pedarsani, R. 2020. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, 2021–2031. PMLR.

Sattler, F.; Wiedemann, S.; Müller, K.-R.; and Samek, W. 2019. Robust and communication-efficient federated learning from non-iid data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9): 3400–3413.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. *Advances in Neural Information Processing Systems*, 30.

T Dinh, C.; Tran, N.; and Nguyen, J. 2020. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33: 21394–21405.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8432–8440.

Tang, Z.; Wang, Y.; and Chang, T.-H. 2024. z-signfedavg: A unified stochastic sign-based compression for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15301–15309.

Vargaftik, S.; Basat, R. B.; Portnoy, A.; Mendelson, G.; Itzhak, Y. B.; and Mitzenmacher, M. 2022. Eden: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning*, 21984–22014. PMLR.

Yu, H.; Jin, R.; and Yang, S. 2019. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *International Conference on Machine Learning*, 7184–7193. PMLR.

Zhang, G.; Jiao, S.; and Xu, X. 2010. Compressed sensing and reconstruction with semi-hadamard matrices. In *2010 2nd International Conference on Signal Processing Systems*, volume 1, V1–194. IEEE.

Zhang, X.; Li, Y.; Li, W.; Guo, K.; and Shao, Y. 2022. Personalized federated learning via variational bayesian inference. In *International Conference on Machine Learning*, 26293–26310. PMLR.

Zhu, G.; Du, Y.; Gündüz, D.; and Huang, K. 2020. One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis. *IEEE Transactions on Wireless Communications*, 20(3): 2120–2135.