

FedAlign: Differentially Private Distribution Alignment for Non-IID Federated Learning

Peng Wu* Jiapeng Zhang*[†] Yingjie Song Xiong Xiao[†] Zhuo Tang[†]
 College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
 {wp2002, zhangjp, yjsong0044, xx, ztang}@hnu.edu.cn

*These authors contributed equally. [†]Corresponding authors.

Abstract

Federated Learning (FL) enables collaborative model training without sharing raw data, but client data are often Non-Independent and Identically Distributed (Non-IID), which often slow convergence and degrade global performance. Meanwhile, privacy preservation is also a critical concern in FL. To address these two issues, we propose FedAlign, a differentially private framework that aligns local data distributions via client-side statistical moment alignment. Clients upload perturbed distribution statistics, which the server aggregates to infer global distribution characteristics and guide local alignment, thereby reducing inter-client discrepancies. Theoretical analysis and experiments show that FedAlign accelerates convergence and improves accuracy under Non-IID settings while preserving rigorous privacy guarantees.

1. Introduction

Federated learning (FL) [20] could be affected by data heterogeneities [30], which refers to the statistical differences in local data distributions among clients, such as distinct label distributions, feature spaces, and sampling biases. These skew characteristics can cause local models to deviate from the global optimum, hinder convergence, and lead to unstable or suboptimal global performance [12, 17, 18, 31], posing a fundamental challenge for the convergence of federated learning. As illustrated in Fig. 1, data heterogeneity manifests in three principal forms: label skew, quantity skew and feature skew [15].

These diverse forms of data heterogeneity collectively give rise to Non-Independent and Identically Distributed (Non-IID) characteristics among clients, which fundamentally alter the optimization dynamics of federated learning. As shown by the dashed convergence trajectory in Fig. 1, such Non-IID conditions cause local models to drift away from the global optimum. Compared with

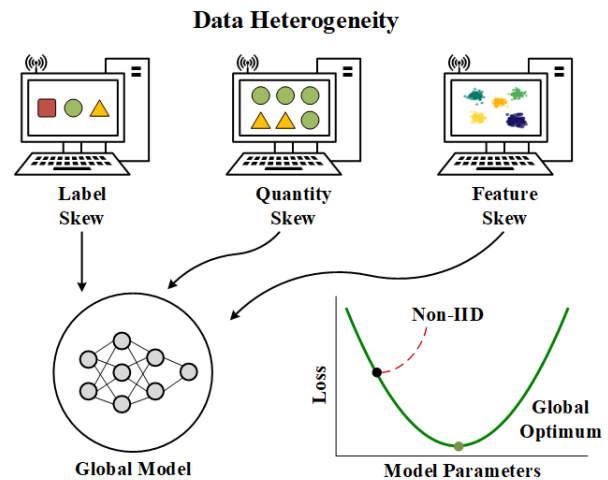


Figure 1. Illustration of data heterogeneity in federated learning.

the solid curve representing the IID scenario, the global model exhibits slower convergence and unstable optimization. This heterogeneity-induced divergence substantially degrades model performance, increases optimization variance, and poses considerable challenges to achieving stable and efficient convergence in federated learning systems.

Beyond impairing convergence, data heterogeneity also amplifies privacy vulnerabilities in federated learning [33]. The same statistical disparities that induce optimization drift make client gradients increasingly distinctive, enabling potential information leakage. Although FL avoids raw data sharing, adversaries can exploit exchanged gradients or parameters to perform model inversion [24], membership inference [22], or gradient leakage attacks [7]. These risks are further intensified under Non-IID settings, where heterogeneous data distributions enhance gradient uniqueness [37].

Differential Privacy (DP) provides a principled defense by injecting calibrated noise into gradients or updates, bounding the influence of any individual data point [4]. Variants such as Rényi DP [23], Local DP [28], Distributed

DP [27], Individualized DP [2], and Adaptive DP [14] have been developed to strengthen protection. However, DP-based FL still faces a fundamental trade-off among privacy, model utility, and convergence speed. Techniques including adaptive noise injection [8], gradient clipping [1], and model sparsification [3] have been explored to mitigate this trade-off and improve robustness.

In summary, federated learning faces intertwined challenges: data heterogeneity degrades convergence, while Non-IID conditions exacerbate privacy leakage. These coupled issues jointly threaten model stability and client confidentiality, underscoring the need for approaches that address both convergence efficiency and privacy preservation.

In this paper, we propose *FedAlign*, a federated learning framework designed for Non-IID environments to jointly mitigate data heterogeneity and privacy leakage. The core idea is to align client data distributions via statistical moments, enabling each client to match global distribution characteristics without exposing its raw data. This alignment facilitates stable convergence while preserving privacy throughout training.

Our main contributions are summarized as follows:

- We propose *FedAlign*, a privacy-preserving federated optimization framework that mitigates the adverse impact of data heterogeneity on global convergence and model performance. *FedAlign* aligns statistical distributions across clients under differential privacy constraints, enabling effective collaboration without compromising local data confidentiality.
- We theoretically model data heterogeneity through statistical moments and establish a quantitative connection between distributional divergence and gradient inconsistency. This analysis provides theoretical insight into how Non-IID data induces local model drift and slows down convergence in federated learning.
- We conduct experiments under diverse Non-IID settings which demonstrate that *FedAlign* consistently outperforms state-of-the-art methods. Ablation studies further validate that statistical moment alignment effectively reduces gradient bias, accelerates convergence, and improves model accuracy.

2. Related Work

2.1. Federated Optimization under Data Heterogeneity

Under Non-IID data distributions, discrepancies among clients local datasets are a primary factor contributing to the performance degradation of the global model. Existing studies [15] typically categorize data heterogeneity into three types of skewness:

1. Label skew, where the label distributions across clients are imbalanced.

2. Feature skew, where samples from the same class exhibit different feature distributions across clients.
3. Quantity skew, where the amount of data varies significantly among clients.

To tackle the challenges posed by Non-IID data, various federated optimization strategies have been developed. For instance, FedOPT [25] improves aggregation stability through adaptive optimizers. FedProx [16] introduces a proximal regularization term to constrain local model drift. FedSkip [6] enhances client-side optimization by intermittently skipping aggregation. Hsieh et al. [10] increase communication frequency to mitigate biased updates, albeit at the expense of training efficiency. FedFA [36] leverages global statistics to reduce feature shift, while FedProto [26] exchanges client prototypes to address distributional heterogeneity. Despite their effectiveness, these approaches generally lack theoretical validation on how inter-client distribution divergence influences the global models convergence behavior under Non-IID conditions.

In parallel, Data Augmentation (DA) has been introduced to improve local data diversity and alleviate Non-IID effects by applying transformations such as rotation, scaling, or noise injection [35]. Representative methods like FedMix [34] share averaged representations to enrich local data, while FedRDN [32] injects randomized cross-client statistics to enhance privacy. FedBR [9] generates pseudo-data to correct learning bias, and FedNova [30] normalizes updates to address objective inconsistency. FedMA [29] aligns hidden-layer elements to construct a coherent global model. While DA-based strategies enhance sample diversity and alleviate certain Non-IID effects, most rely on limited statistical information and fail to adequately address inter-client drift. Their influence on the global models convergence under heterogeneous data remains largely unexplored.

2.2. Differentially Private Federated Learning

Differentially Private Federated Learning (DP-FL) integrates differential privacy mechanisms into the FL framework, aiming to prevent information leakage while preserving model utility. Its core idea is to inject random noise into the gradients or model updates uploaded by clients, thereby constraining the influence of any individual data sample on the global model and ensuring that user privacy is protected.

For example, Fed-SMP [11] achieves client-level DP by applying model update sparsification followed by Gaussian noise injection. However, its sparsification ratio and noise scale require manual tuning, limiting scalability to large models and complex systems. McMahan et al. [21] first incorporated DP into FL in RNN-based language modeling, using gradient clipping and Gaussian noise to achieve record- or client-level DP. Yet, under Non-IID data and deep architectures, its accuracy drops significantly, and commu-

nication optimization as well as convergence analysis remain insufficient. Liu et al. [19] enhance DP-FL performance through a shuffling mechanism that combines local and central DP to improve privacy amplification. Nonetheless, its effectiveness diminishes under extreme Non-IID scenarios and relies heavily on randomized client update ordering. Erlingsson et al. [5] proposed a theoretically verifiable privacy amplification framework that upgrades local DP to approximate central DP, thereby reducing the per-client noise requirement. However, these approaches still overlook the impact of Non-IID data and lack a formal convergence analysis across iterative training rounds.

3. FedAlign Framework

3.1. Overview

The proposed framework, *FedAlign* aims to mitigate the distributional divergence caused by Non-IID data under client-level differential privacy constraints. *FedAlign* achieves this by aligning the local sampled data of each client at the level of statistical moments (mean, variance, skewness, and kurtosis) prior to training. This alignment accelerates global model convergence and improves accuracy. The framework employs the Gaussian mechanism as the differential privacy noise injection approach. Specifically, it perturbs both the uploaded local unsampled-data statistics and the local gradients. The server aggregates only the noisy statistics and noisy gradients without accessing any raw data or unperturbed local gradients.

As illustrated in Fig. 2, the workflow of the *FedAlign* framework consists of four modules:

1. Global Model Distribution: The server maintains a global model ω^t and distributes its parameters to all clients at the beginning of each communication round, serving as the initialization or reference direction for local training.
2. Global Statistical Calculation:
 - (a) *Local Unsampled Data Statistics Computation and Private Reporting*: Each client divides its local dataset into sampled dataset (used for the current local training) and unsampled dataset (used for estimating the remaining local distribution). On the unsampled dataset, the client computes local statistical moments including mean μ_k , variance σ_k^2 , skewness γ_k , and kurtosis κ_k . Gaussian noise is then added to each statistic to ensure (ϵ, δ) -DP before uploading them to the server.
 - (b) *Server-Side Private Aggregation and Global Estimation*: The server receives the noisy statistics from all clients and aggregates them using sample-size-weighted averaging to obtain global statistics μ_{global} , σ_{global}^2 , γ_{global} , and κ_{global} for subsequent alignment guidance.

3. Local Distribution Alignment: The server broadcasts the aggregated global statistics back to all clients. Each client then adjusts its local sampled data distribution to align its statistical moments with the global statistics as closely as possible. The aligned sampled data are subsequently used for local model training.
4. Local Training and Global Model Aggregation: Each client performs several local training epochs on the aligned sampled data, computes the model update or gradient g_k^t , applies gradient clipping to bound sensitivity, and adds Gaussian noise to ensure differential privacy:

$$\tilde{g}_k^t = g_k^t + \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad (1)$$

where σ denotes the noise scale. The server aggregates all noisy gradients, typically via weighted averaging, and updates the global model as:

$$\omega^{t+1} = \omega^t - \eta \sum_{k=1}^K \frac{n_k}{N} \tilde{g}_k^t, \quad (2)$$

where η is the learning rate, n_k is the local sample size of client k , and $N = \sum_{k=1}^K n_k$ is the total number of samples.

3.2. Statistical View of Non-IID Data Skew

Statistical heterogeneity among client data distributions critically affects the convergence of federated learning under Non-IID conditions. To quantify this heterogeneity, we analyze feature distributions at multiple levels: sample, client, and global.

At the sample level, each feature dimension encodes an underlying data characteristic that can be described through statistical moments. The mean reflects the central tendency and reveals potential biases in brightness, color intensity, or other intrinsic properties. The variance captures dispersion and intra-sample variability. Skewness describes distributional asymmetry, and kurtosis characterizes peakedness or tail thickness. These moments provide a quantitative basis for assessing variations across samples.

At the client level, statistics are aggregated over each participant's local dataset. Deviations from other clients quantify inter-client heterogeneity. Differences in mean, skewness, or kurtosis indicate pronounced feature skew, while variance captures heterogeneity in local feature diversity. Client-level statistics thus offer an objective measure of Non-IID feature distributions, guiding aggregation, normalization, or alignment strategies robust to client-specific biases.

Global-level reference statistics define an idealized, optimal distribution for the current Non-IID scenario, aggregating global information in a manner akin to data augmentation. Serving as a target for feature alignment, this reference enables each client to adjust its local statistics toward

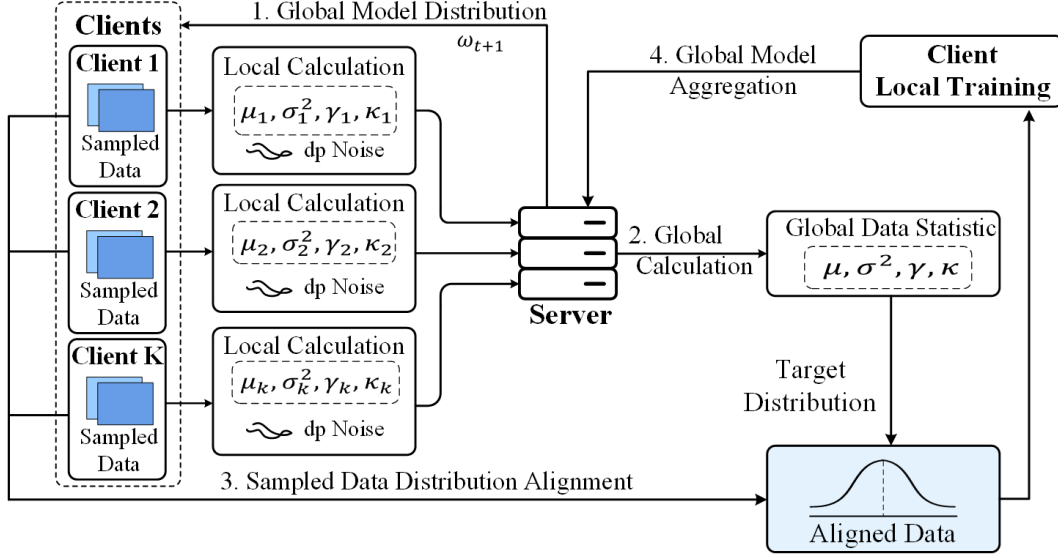


Figure 2. Overall framework of FedAlign.

the global benchmark, effectively mitigating the impact of data heterogeneity on convergence and performance.

3.3. Client-Side Private Statistic Computation

Each client k computes private statistical descriptors from its unsampled dataset $\mathcal{U}_k$. To capture the essential distributional characteristics, the first four statistical moments are computed, including the mean, variance, skewness, and kurtosis:

$$\mu_k = \frac{1}{|\mathcal{U}_k|} \sum_{\mathbf{x}_i \in \mathcal{U}_k} \mathbf{x}_i, \quad \kappa_k = \frac{1}{|\mathcal{U}_k|} \sum_{\mathbf{x}_i \in \mathcal{U}_k} \frac{(\mathbf{x}_i - \mu_k)^4}{\sigma_k^4},$$

$$\sigma_k^2 = \frac{1}{|\mathcal{U}_k|} \sum_{\mathbf{x}_i \in \mathcal{U}_k} (\mathbf{x}_i - \mu_k)^2, \quad s_k = \frac{1}{|\mathcal{U}_k|} \sum_{\mathbf{x}_i \in \mathcal{U}_k} \frac{(\mathbf{x}_i - \mu_k)^3}{\sigma_k^3}.$$

To preserve privacy, Gaussian noise is injected into each statistic:

$$\tilde{\mu}_k = \mu_k + \mathcal{N}(0, \sigma_{\text{dp}}^2), \quad \tilde{\sigma}_k^2 = \sigma_k^2 + \mathcal{N}(0, \sigma_{\text{dp}}^2), \quad (3)$$

where σ_{dp} denotes the noise scale determined by the differential privacy budget (ϵ, δ) . The noisy statistics $\{\tilde{\mu}_k, \tilde{\sigma}_k^2, \tilde{s}_k, \tilde{\kappa}_k\}$ are then transmitted to the server for aggregation.

3.4. Server-Side Statistic Aggregation

Upon receiving the noisy local statistics, the central server performs weighted aggregation to estimate the global distribution characteristics. Given the sample size n_k of client k and total number $N = \sum_{k=1}^K n_k$, the aggregated mean and variance are computed as

$$\bar{\mu} = \sum_{k=1}^K \frac{n_k}{N} \tilde{\mu}_k, \quad \bar{\sigma}^2 = \sum_{k=1}^K \frac{n_k}{N} \tilde{\sigma}_k^2. \quad (4)$$

The higher-order moments are similarly aggregated:

$$\bar{s} = \sum_{k=1}^K \frac{n_k}{N} \tilde{s}_k, \quad \bar{\kappa} = \sum_{k=1}^K \frac{n_k}{N} \tilde{\kappa}_k. \quad (5)$$

The aggregated statistics $\{\bar{\mu}, \bar{\sigma}^2, \bar{s}, \bar{\kappa}\}$ serve as the global reference for aligning client-side distributions, ensuring statistical consistency across heterogeneous local datasets.

3.5. Client-Side Distribution Alignment

Client-side data heterogeneity plays a fundamental role in determining the convergence behavior of federated learning under Non-IID data distributions. To rigorously characterize this effect, in this section, we first provide a decomposition of the deviation between each clients local gradient and the global differentially private gradient (Theorem 1), explicitly isolating contributions from the expected gradient bias, local variance, inter-client covariance, and DP-induced noise. Building upon this decomposition, we formulate a sequence of propositions that quantify the influence of the four statistical moments of the local data, namely mean, variance, skewness, and kurtosis, on gradient behavior, accounting for both intrinsic data heterogeneity and differential privacy perturbations. Leveraging these insights, Lemma 3 and Theorem 2 establish an explicit connection between client-level statistical disparities and the global convergence bound of federated learning. Collectively, this theoretical framework provides a principled bridge linking client-side data statistics to gradient dynamics and overall algorithmic convergence. Detailed proofs of all the theorems presented in this section are provided in the Appendix.

3.5.1. Statistical Decomposition of Gradient Disparity

Under the Dirichlet Non-IID data partition, the expected squared distance between client k 's local gradient and the global DP gradient admits the following decomposition:

Theorem 1 (Gradient Disparity Decomposition). *In federated learning with differential privacy and Non-IID data partitioning, the expected squared distance between client k 's local gradient and the global DP gradient satisfies,*

$$\begin{aligned} \mathbb{E}[\|\tilde{g}_k - \tilde{g}^{\text{DP}}\|^2] = & \underbrace{\|g_k - g\|^2}_{\text{Expected Gradient Bias}} + \underbrace{\text{tr}(\text{Var}(\tilde{g}_k))}_{\text{Local Variance}} \\ & - 2 \underbrace{\sum_{j=1}^K \frac{n_j}{N} \text{tr}(\text{Cov}(\tilde{g}_k, \tilde{g}_j))}_{\text{Cross-Client Covariance}} \\ & + \underbrace{\sum_{i=1}^K \sum_{j=1}^K \frac{n_i n_j}{N^2} \text{tr}(\text{Cov}(\tilde{g}_i, \tilde{g}_j))}_{\text{Global Covariance Structure}} \\ & + \underbrace{\text{tr}(\text{Var}(\eta))}_{\text{DP Noise Variance}}, \end{aligned} \quad (6)$$

where $\tilde{g}_k$ is the local stochastic gradient, $\tilde{g}^{\text{DP}}$ is the global differentially private gradient, and η is the DP noise.

3.5.2. Impact of Statistical Moments on Gradient

The four key statistical moments of client data distributions influence gradient behavior through distinct mechanisms.

Proposition 1 (Mean Disparity Impact). *The mean disparity $\Delta_{\mu,k} = \|\mu_k - \mu\|^2$ bounds the expected gradient bias:*

$$\|g_k - g\|^2 \leq \|H(w)\|_F^2 \cdot \Delta_{\mu,k}, \quad (7)$$

where $H(w)$ approximates the Hessian matrix. With DP noise $\epsilon_k \sim \mathcal{N}(0, \sigma_\mu^2 I)$, the complete bound becomes

$$\|g_k - g\|^2 \leq \|H(w)\|_F^2 \left(\mathbb{E}[\|\mu_k - \tilde{\mu}\|^2] - \sum_{j=1}^K \left(\frac{n_j}{N}\right)^2 d\sigma_\mu^2 \right), \quad (8)$$

where μ_k is the local mean, μ is the global mean, and $\tilde{\mu}$ is the DP-protected mean estimate.

Proposition 2 (Variance Impact). *The gradient variance is upper bounded by the local data variance:*

$$\text{tr}(\text{Var}(\tilde{g}_k)) \leq L^2 (\text{tr}(\Sigma_k) + \|\mu_k\|^2), \quad (9)$$

where L is the Lipschitz constant and Σ_k is the local covariance matrix.

Table 1. Summary of key notations used in this paper.

Symbol	Description
ω^t	Global model at communication round t .
$g_k^t, \tilde{g}_k^t$	Local gradient of client k , before / after DP noise injection.
η, η_l	Global / local learning rate.
τ	Number of local training epochs per round.
$F(\cdot), L$	Global loss function and its Lipschitz constant.
$K, n_k, N = \sum_k n_k$	Number of clients, local sample size, and total data size.
$\mathcal{S}_k, \mathcal{U}_k$	Sampled / unsampled subsets of client k .
$\mu_k, \sigma_k^2, \gamma_k, \kappa_k$	Local mean, variance, skewness, and kurtosis.
$\tilde{\mu}_k, \tilde{\sigma}_k^2, \tilde{\gamma}_k, \tilde{\kappa}_k$	DP-protected statistical moments.
$\mu_g, \sigma_g^2, \gamma_g, \kappa_g$	Aggregated global statistics.
Σ_k	Covariance matrix of client k 's data.
$\sigma_{\text{dp}}, \sigma_{\text{stat}}, \sigma_{\text{grad}}$	DP noise scales for statistics / gradients.
C, S_{stat}	Clipping thresholds for gradients / statistics.
ϵ, δ	DP privacy parameters.
$\Gamma_{\text{stat}}, \Delta_{\mu,k}$	Statistical disparity and mean deviation terms.
σ_{DP}^2	Variance of Gaussian DP noise.

When DP noise $\eta_k \sim \mathcal{N}(0, \sigma_{\text{DP},k}^2 I)$ is added, the bound becomes

$$\begin{aligned} \text{tr}(\text{Var}(\tilde{g}_k^{\text{DP}})) \leq & L^2 \left(\text{tr}(\Sigma_k) + \|\mu_k\|^2 \right. \\ & \left. + \sum_{j=1}^K \left(\frac{n_j}{N}\right)^2 d_\Sigma \sigma_{\text{DP},j}^2 \right), \end{aligned} \quad (10)$$

where d_Σ is the dimension of the flattened covariance matrix.

Proposition 3 (Skewness Impact on Gradient). *Let γ_k be the local skewness of client k and $\bar{\gamma} = \sum_{k=1}^K (n_k/N) \gamma_k$ the aggregated skewness. Then, under Edgeworth expansion and Gaussian DP noise $\epsilon_{\text{DP}} \sim \mathcal{N}(0, \sigma_{\text{DP}}^2)$, the deviation of the DP-protected local gradient from the global DP gradient satisfies*

$$\|g_k^{\text{DP}} - \tilde{g}^{\text{DP}}\|^2 \propto |\gamma_k - \bar{\gamma}|^2. \quad (11)$$

Proposition 4 (Kurtosis Impact). *The gradient variance is affected by the local kurtosis κ_k :*

$$\text{tr}(\text{Var}(\tilde{g}_k^{\text{DP}})) \propto \left(1 + \frac{\kappa_k - 3}{4} \right) \text{tr}(\Sigma_k) + d\sigma_{\text{DP}}^2, \quad (12)$$

where Σ_k is the local covariance, d is the gradient dimension, and σ_{DP}^2 is the variance of the Gaussian DP noise.

3.5.3. Convergence Bound with Statistical Disparity

Lemma 1 (Relation between statistical disparity and gradient variance). *For any client k , the local gradient variance satisfies*

$$\mathbb{E}\|\tilde{g}_k - g\|^2 \leq L_\mu^2 \Delta_{\mu,k} + L_\Sigma^2 \|\Sigma_k - \Sigma\|_F^2 + \mathcal{O}(\sigma_{\text{DP}}^2).$$

Building upon the gradient analysis, we establish the direct relationship between client statistical disparities and federated learning convergence.

Theorem 2 (Statistical Disparity Convergence Bound). *Under L -smoothness and non-convex assumptions, DP federated learning convergence rate satisfies*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(w_t)\|^2] \leq \frac{2\Delta_F}{\eta T} + \eta L (\sigma_{\min}^2 + \Gamma_{\text{stat}} + \sigma_{\text{DP}}^2), \quad (13)$$

where the statistical disparity term Γ_{stat} is bounded by

$$\Gamma_{\text{stat}} \leq \frac{1}{K} \sum_{k=1}^K (L_\mu^2 \Delta_{\mu,k} + L_\Sigma^2 \|\Sigma_k - \Sigma\|_F^2 + L_\gamma |\gamma_k - \gamma| + L_\kappa |\kappa_k - \kappa|) \quad (14)$$

and the DP noise variance $\sigma_{\text{DP}}^2 = 8C^2 \log(1.25/\delta)/\epsilon^2$ depends on the privacy budget (ϵ, δ) .

3.6. Private Statistic-Based Distribution Alignment

FedAlign is designed to enhance federated learning under heterogeneous and Non-IID client data distributions by aligning local client distributions to the global statistics in a differentially private manner. The overall procedure consists of server-side coordination and client-side local updates with distribution alignment, as illustrated in Algorithms 1 and 2.

On the server side (Algorithm 1), in each global round t , the server first selects a subset of clients $\mathcal{W}_t$ and broadcasts the current global model w^t to them. Each client computes DP statistics of its local data and sends them to the server. The server aggregates these statistics to form the global mean, variance, skewness, and kurtosis, which are then broadcast back to all selected clients. Subsequently, clients perform local training on their aligned data, and the server aggregates the DP-protected local updates to update the global model. This process is repeated for T rounds until convergence.

On the client side (Algorithm 2), each client first splits its local data into sampled and unsampled sets. The unsampled data is used to compute local statistical moments, which are clipped and perturbed with Gaussian noise to ensure differential privacy. After receiving aggregated global statistics from the server, the client aligns its sampled data to the global distribution. Local training is then performed

Algorithm 1 Server-Side Procedure of FedAlign

Require: Number of clients n ;

Number of selected clients per round r ;

Number of global rounds T ;

Global learning rate η ;

Total data size $N = \sum_k n_k$;

Noise multiplier for statistics σ_{stat} ;

Noise multiplier for gradients σ_{grad} .

Ensure: Final global model w^T .

1: Initialize global model $w^0 \in \mathbb{R}^d$.

2: **for** each round $t = 0, 1, \dots, T - 1$ **do**

3: Sample a subset of clients $\mathcal{W}_t \subset [n]$ uniformly at random.

4: Broadcast current global model w^t to all $k \in \mathcal{W}_t$.

5: **for** each client $k \in \mathcal{W}_t$ **in parallel do**

6: Receive DP-protected statistics $(\tilde{\mu}_k, \tilde{\sigma}_k^2, \tilde{\gamma}_k, \tilde{\kappa}_k)$.

7: **end for**

8: Compute aggregated global statistics:

$$\mu_g^t = \sum_{k \in \mathcal{W}_t} \frac{n_k}{N} \tilde{\mu}_k, \quad \sigma_g^{2,t} = \sum_{k \in \mathcal{W}_t} \frac{n_k}{N} \tilde{\sigma}_k^2,$$

$$\gamma_g^t = \sum_{k \in \mathcal{W}_t} \frac{n_k}{N} \tilde{\gamma}_k, \quad \kappa_g^t = \sum_{k \in \mathcal{W}_t} \frac{n_k}{N} \tilde{\kappa}_k.$$

9: Broadcast $(\mu_g^t, \sigma_g^{2,t}, \gamma_g^t, \kappa_g^t)$ to all $k \in \mathcal{W}_t$.

10: **for** each client $k \in \mathcal{W}_t$ **in parallel do**

11: Receive DP-protected local update $\tilde{\Delta}_k^t$.

12: **end for**

13: Aggregate client updates and update global model:

$$w^{t+1} = w^t - \eta \sum_{k \in \mathcal{W}_t} \frac{n_k}{N} \tilde{\Delta}_k^t.$$

14: **end for**

15: **return** w^T .

on the aligned data for τ epochs, producing a local gradient. This gradient is further perturbed with Gaussian noise for privacy and securely transmitted back to the server.

Overall, *FedAlign* explicitly targets the dominant sources of statistical disparity—primarily the mean and variance differences among clients—while modulating the secondary effects of skewness and kurtosis. By reducing the deviations of local updates from the global gradient, *FedAlign* improves convergence speed and enhances the final model accuracy under Non-IID and heterogeneous data conditions.

4. Experiments

4.1. Baseline Methods

We compare *FedAlign* with several federated learning algorithms:

- FedAvg [20] is the standard federated averaging method that aggregates local model updates from clients by weighted averaging according to their dataset sizes. It serves as a fundamental baseline for FL under both IID and Non-IID data.
- FedProx [16] introduces a proximal term in the local loss function to constrain each client model to remain close to the global model during updates.
- SCAFFOLD [13] mitigates client drift under Non-IID data by correcting the direction of local updates through control variates, aligning them with the ideal global update path.
- FedNova [30] normalizes the scaling of local updates to reduce bias in the aggregated global model.
- FedRDN [32] randomly inject statistics of other clients into local data of each client.

Comprehensive descriptions of the experimental environments and implementation details are presented in the Appendix.

4.2. Validation of Distribution Alignment Effectiveness

We conduct experiments on the CIFAR-10 and MNIST datasets using two representative architectures, a standard CNN and ResNet50, to evaluate the effectiveness of client-side distribution alignment in enhancing both model accuracy and convergence speed. In all experiments, the local data distributions are aligned with the global distribution by adjusting their means and variances.

As summarized in Table 2 and illustrated in Fig. 3, Fig. 4, *FedAlign* consistently outperforms all baseline methods across both datasets under various Non-IID settings. Specifically, with Dirichlet $\alpha = 0.5$ and Gaussian noise $\mathcal{N}(0, \beta)$ with $\beta = 0.05$ for data heterogeneity, *FedAlign* achieves 54.9% test accuracy on CIFAR-10, corresponding to relative improvements of 11.9%, 7.5%, 10.7%, 4.8%, and 3.9% over FedAvg, FedProx, Scaffold, FedNova, and FedRDN, respectively. The corresponding MNIST accuracy reaches 99.26%, also surpassing all baselines. Under $\beta = 0.1$, *FedAlign* attains 54.3% on CIFAR-10 and 99.12% on MNIST, again achieving the best overall performance.

Furthermore, Fig. 5 demonstrates that the superiority of *FedAlign* extends to deeper architectures such as ResNet50, yielding consistently higher accuracy on CIFAR-10 compared to all baselines. These results collectively confirm that aligning client-side distribution effectively reduces inter-client gradient bias and drift, thereby accelerating convergence and improving final model performance, which

Algorithm 2 Client-Side Procedure of FedAlign

Require: Local learning rate η_l ;

Local epoch τ ;

Statistical clipping bound S_{stat} ;

Gradient clipping threshold C .

Ensure: DP-protected local gradient $\tilde{g}_k^t$.

- 1: Receive w^t from the server.
- 2: Partition local data into sampled set S_k and unsampled set $\mathcal{U}_k$.
- 3: Compute statistical moments on $\mathcal{U}_k$:

$$(\mu_k, \sigma_k^2, \gamma_k, \kappa_k) = \text{EmpiricalMoments}(\mathcal{U}_k).$$

- 4: Clip each statistic by S_{stat} .
- 5: Add Gaussian noise for DP protection:

$$\tilde{\mu}_k = \mu_k + \mathcal{N}(0, \sigma_{\text{stat}}^2),$$

$$\tilde{\sigma}_k^2 = \sigma_k^2 + \mathcal{N}(0, \sigma_{\text{stat}}^2),$$

$$\tilde{\gamma}_k = \gamma_k + \mathcal{N}(0, \sigma_{\text{stat}}^2), \quad \tilde{\kappa}_k = \kappa_k + \mathcal{N}(0, \sigma_{\text{stat}}^2).$$

- 6: Send $(\tilde{\mu}_k, \tilde{\sigma}_k^2, \tilde{\gamma}_k, \tilde{\kappa}_k)$ to the server.
- 7: Receive aggregated global statistics $(\mu_g^t, \sigma_g^{2,t}, \gamma_g^t, \kappa_g^t)$.
- 8: Perform distribution alignment on S_k :

$$S'_k = \text{Align}(S_k; \mu_g^t, \sigma_g^{2,t}, \gamma_g^t, \kappa_g^t).$$

- 9: Initialize local model $w_{k,0}^t = w^t$.
- 10: **for** each local epoch $e = 1, \dots, \tau$ **do**
- 11: Update $w_{k,e+1}^t = w_{k,e}^t - \eta_l \nabla f_k(w_{k,e}^t; S'_k)$.
- 12: **end for**
- 13: Compute local gradient $g_k^t = \frac{w^t - w_{k,\tau}^t}{\eta_l \tau}$.
- 14: Add Gaussian noise:

$$\tilde{g}_k^t = g_k^t + \mathcal{N}(0, \sigma_{\text{grad}}^2 \mathbf{I}).$$

- 15: Encrypt and send $\tilde{g}_k^t$ to the server via secure aggregation.
-

validate the theoretical analysis presented in Section IV, highlighting the robustness of distribution alignment for federated learning with heterogeneous data.

4.3. Ablation Study

The first ablation experiment examines how the proportion of global unsampled data affects distribution alignment. Specifically, we vary the proportion of global unsampled data incorporated into the alignment process and analyze its effect on convergence speed and test accuracy. Experiments are conducted on the CIFAR-10 dataset using a CNN model to evaluate the sensitivity of *FedAlign* to this proportion.

The term proportion of global unsampled data refers to the fraction of unseen samples whose aggregated statistics are utilized for distribution alignment. A larger proportion

allows the global statistics to more accurately approximate the true data distribution, thereby providing stronger guidance for local alignment.

Let $\rho \in [0, 1]$ denote the proportion of global unsampled data incorporated into the alignment process. As ρ increases, the estimated global moments become more representative, leading to faster convergence and improved model accuracy.

As shown in Fig. 6a, increasing ρ consistently enhances test accuracy and stabilizes convergence. For example, utilizing the full set of global unsampled data yields approximately 16% higher test accuracy at 100 training rounds compared to using only 40% of the data, confirming the effectiveness of distribution alignment in *FedAlign*.

The second ablation experiment examines the contribution of aligning different statistical moments within the distribution alignment process. Four configurations are compared: aligning only the mean, aligning only the variance, aligning both mean and variance, and jointly aligning all four moments.

The results, presented in Fig. 6b, demonstrate that mean and variance plays a decisive role in improving convergence speed and model accuracy. Specifically, mean and variance dominate the expected gradient bias and inter-client drift, whereas higher-order moments primarily capture local asymmetry and tail behavior, modulating the convergence, which is consistent with the theoretical analysis. Jointly aligning all four moments further reduces inter-client statistical disparities, resulting in faster convergence and higher accuracy under Non-IID conditions.

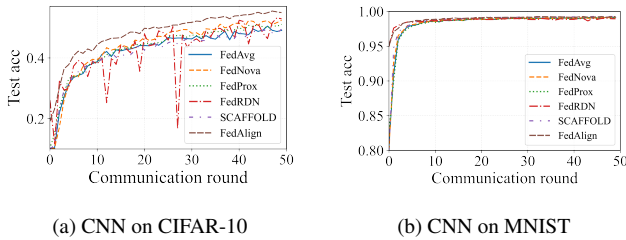


Figure 3. Test accuracy over training rounds for CNN models with $\alpha = 0.5$ and $\beta = 0.05$ on CIFAR-10 and MNIST.

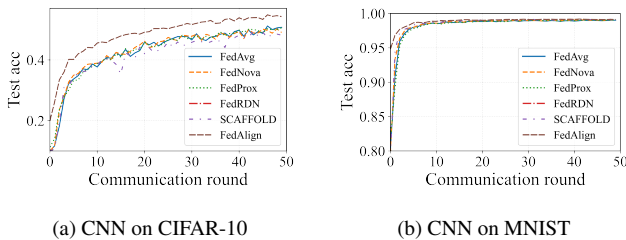


Figure 4. Test accuracy over training rounds for CNN models with $\alpha = 0.5$ and $\beta = 0.1$ on CIFAR-10 and MNIST.

Table 2. Test Accuracy (%) of Different FL Methods on CIFAR-10 and MNIST under different Non-IID settings.

Non-IID Skewness	Method	CIFAR-10	MNIST
Dirichlet $\alpha = 0.5$ Gaussian noise $\mathcal{N}(0, \beta)$ with $\beta = 0.05$	FedAvg	49.1	99.13
	FedProx	51.1	99.08
	Scaffold	49.6	99.22
	FedNova	52.4	99.04
	FedRDN	52.8	98.92
	FedAlign	54.9	99.26
Dirichlet $\alpha = 0.5$ Gaussian noise $\mathcal{N}(0, \beta)$ with $\beta = 0.1$	FedAvg	50.7	99.02
	FedProx	50.5	99.05
	Scaffold	48.1	99.12
	FedNova	49.0	99.08
	FedRDN	49.8	98.94
	FedAlign	54.3	99.12

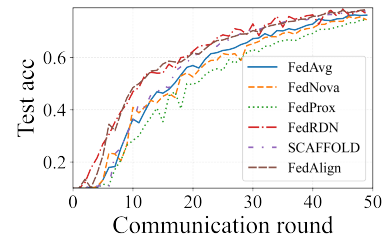


Figure 5. Test accuracy of ResNet50 on CIFAR-10.

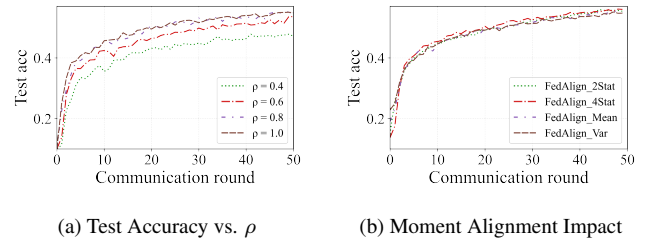


Figure 6. Ablation Study of *FedAlign* on CIFAR-10.

5. Conclusion

This paper introduces *FedAlign*, a federated learning framework designed to mitigate Non-IID data heterogeneity through moment-based distribution alignment under differential privacy constraints. Theoretical analysis and experimental evaluations demonstrate that aligning statistical moments effectively reduces gradient discrepancies, accelerates convergence, and improves global model accuracy. Ablation studies further verify the contribution of higher-order moments to effective distribution alignment. Overall, *FedAlign* provides a principled framework that enhances convergence and model accuracy while preserving privacy in federated learning.

Acknowledgment

The work is supported by the National Natural Science Foundation of China (Grant Nos. 62225205, 62302157, 62302160, 62532005), the Science and Technology Program of Changsha (kh2301011), the Major Science and Technology Research Projects of Hunan Province (Grant Nos. 2024QK2010, 2024QK2009), the Yunnan Provincial Major Science and Technology Special Plan Projects (No.202502AD080009), the Yunnan Science and Technology Talent and Platform Program (No.202605AK340003), the Natural Science Foundation of Hunan Province (Grant Nos. 2024JJ6154), and the open Project Fund of the State Key Laboratory of Cyberspace Security Defense (2024-MS-04).

References

- [1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, et al. Deep learning with differential privacy. In *Proc. CCS*, 2016.
- [2] F. Boenisch, C. Mühl, A. Dziedzic, R. Rinberg, and N. Papernot. Have it your way: Individualized privacy assignment for dp-sgd. In *Adv. Neural Inf. Process. Syst.*, 2023.
- [3] A. Cheng, P. Wang, X. S. Zhang, and J. Cheng. Differentially private federated learning with local regularization and sparsification. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10122–10131, 2022.
- [4] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4): 211–407, 2014.
- [5] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proc. 13th Annu. ACM-SIAM Symp. Discrete Algorithms*, pages 2468–2479, 2019.
- [6] Z. Fan, Y. Wang, J. Yao, L. Lyu, Y. Zhang, and Q. Tian. Fedskip: Combatting statistical heterogeneity with federated skip aggregation. In *Proc. ICDM*, pages 131–140, 2022.
- [7] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller. Inverting gradients: How easy is it to break privacy in federated learning? In *Proc. NeurIPS*, 2020.
- [8] R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.
- [9] Y. Guo, X. Tang, and T. Lin. Fedbr: Improving federated learning on heterogeneous data via local learning bias reduction. In *Proc. Int. Conf. Mach. Learn.*, pages 12034–12054, 2023.
- [10] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons. The non-iid data quagmire of decentralized machine learning. In *Proc. Int. Conf. Mach. Learn.*, pages 4387–4398, 2020.
- [11] R. Hu, Y. Guo, and Y. Gong. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy. *IEEE Trans. Mobile Comput.*, 23(8):8242–8255, 2023.
- [12] X. Hu, S. Li, and Y. Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *Proc. Int. Conf. Learn. Represent.*, 2023.
- [13] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proc. Int. Conf. Mach. Learn.*, pages 5132–5143, 2020.
- [14] S. Kiani, N. Kulkarni, A. Dziedzic, S. Draper, and F. Boenisch. Differentially private federated learning with time-adaptive privacy spending. *arXiv preprint arXiv:2502.18706*, 2025.
- [15] Q. Li, Y. Diao, Q. Chen, and B. He. Federated learning on non-iid data silos: An experimental study. In *Proc. IEEE 38th Int. Conf. Data Eng. (ICDE)*, pages 965–978, 2022.
- [16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. In *Proc. Mach. Learn. Syst.*, pages 429–450, 2020.
- [17] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *Proc. ICLR*, 2020.
- [18] J. Liu, J. Jia, T. Che, C. Huo, J. Ren, Y. Zhou, et al. Fedasmu: Efficient asynchronous federated learning with dynamic staleness-aware model update. In *Proc. AAAI Conf. Artif. Intell.*, pages 13900–13908, 2024.
- [19] R. Liu, Y. Cao, H. Chen, R. Guo, and M. Yoshikawa. Flame: Differentially private federated learning in the shuffle model. In *Proc. AAAI Conf. Artif. Intell.*, pages 8688–8696, 2021.
- [20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. Artif. Intell. Stat.*, pages 1273–1282, 2017.
- [21] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. In *Proc. ICLR*, 2018.
- [22] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *Proc. IEEE S&P*, 2019.
- [23] I. Mironov. Rényi differential privacy. In *Proc. CSF*, pages 263–275, 2017.
- [24] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proc. IEEE S&P*, 2019.
- [25] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, et al. Adaptive federated optimization. In *Proc. ICLR*, 2021.
- [26] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, et al. Fedproto: Federated prototype learning across heterogeneous clients. In *Proc. AAAI Conf. Artif. Intell.*, pages 8432–8440, 2022.
- [27] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, et al. A hybrid approach to privacy-preserving federated learning. In *Proc. ACM Workshop Artif. Intell. Secur.*, pages 1–11, 2019.
- [28] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei. Ldp-fed: Federated learning with local differential privacy. In *Proc. ACM Int. Workshop Edge Syst., Analytics and Networking*, pages 61–66, 2020.

- [29] H. Wang, M. Yurochkin, Y. Sun, et al. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.
- [30] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Adv. Neural Inf. Process. Syst.*, pages 7611–7623, 2020.
- [31] Z. Wang, X. Fan, J. Qi, and et al. Fedgs: Federated graph-based sampling with arbitrary client availability. In *Proc. AAAI Conf. Artif. Intell.*, pages 10271–10278, 2023.
- [32] Y. Yan, H. Fu, Y. Li, J. Xie, J. Ma, G. Yang, et al. A simple data augmentation for feature distribution skewed federated learning. In *Proc. CVPR*, pages 25749–25758, 2025.
- [33] X. Yin, Y. Zhu, and J. Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Comput. Surv.*, 54(6):1–36, 2021.
- [34] T. Yoon, S. Shin, S. J. Hwang, and E. Yang. Fedmix: Approximation of mixup under mean augmented federated learning. *arXiv preprint arXiv:2107.00233*, 2021.
- [35] H. Zhang, Q. Hou, T. Wu, S. Cheng, and J. Liu. Data-augmentation-based federated learning. *IEEE Internet Things J.*, 10(24):22530–22541, 2023.
- [36] T. Zhou and E. Konukoglu. Fedfa: Federated feature augmentation. *arXiv preprint arXiv:2301.12995*, 2023.
- [37] L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. In *Proc. NeurIPS*, 2019.