electronics

MDPI

Article

Fast-Converging and Trustworthy Federated Learning Framework for Privacy-Preserving Stock Price Modeling

Zilong Hou [1,2], Yan Ke [1], Yang Qiu [1], Qichun Wu [2,*] and Ziyang Liu [3]

[1] China E-Commerce Association Innovation & Integration Information Technology Research Institute, Beijing 100037, China
[2] School of Economics, Beijing Institute of Technology, Beijing 100081, China
[3] Asia-Pacific AI Research Center, Asian Business Research Institute, Hong Kong 999077, China
* Correspondence: wuqichun@bit.edu.cn

Abstract

Stock price modeling under privacy constraints presents a unique challenge at the intersection of computational economics and machine learning. Financial institutions and brokerage firms hold valuable yet sensitive data that cannot be centrally aggregated due to privacy laws and competitive concerns. To address this issue, we propose a novel Fast-Converging Federated Learning (FCFL) framework that enables decentralized and privacy-preserving stock price modeling. FCFL employs a dual-stage adaptive optimization strategy that dynamically tunes local learning rates and aggregation weights based on inter-client gradient divergence, accelerating convergence in heterogeneous financial environments. The framework integrates secure aggregation and differential privacy mechanisms to prevent information leakage during communication while maintaining model fidelity. Experimental results on multi-institutional stock datasets demonstrate that FCFL achieves up to 30% faster convergence and 2.5% lower prediction error compared to conventional federated averaging approaches, while guaranteeing strong ε-differential privacy. Theoretical analysis further proves that the framework attains sublinear convergence in $\mathcal{O}(\log T)$ communication rounds under non-IID data distributions. This study provides a new direction for collaborative financial modeling, balancing efficiency, accuracy, and privacy in real-world economic systems.

Keywords: federated learning; differential privacy; stock price modeling

Academic Editor: Alberto Fernandez Hilario

Received: 21 October 2025
Revised: 8 November 2025
Accepted: 10 November 2025
Published: 12 November 2025

Citation: Hou, Z.; Ke, Y.; Qiu, Y.; Wu, Q.; Liu, Z. Fast-Converging and Trustworthy Federated Learning Framework for Privacy-Preserving Stock Price Modeling. *Electronics* **2025**, *14*, 4405. https://doi.org/10.3390/electronics14224405

1. Introduction

Financial institutions increasingly rely on machine learning to forecast returns, price risk, and allocate capital [1,2], yet collaboration across organizations remains constrained by privacy, regulation, and competition. Centralizing order flows, client attributes, or proprietary factor signals is often infeasible, motivating federated learning (FL) as a systems and statistical paradigm that trains models without pooling raw data [3]. In cross-silo finance, however, FL must satisfy three stringent requirements simultaneously: protect institution-level and client-level information, converge quickly despite pronounced statistical heterogeneity, and deliver models whose economic utility survives regime shifts. Meeting these requirements is challenging. Naïve aggregation exacerbates client drift when data are non-IID across institutions [4,5], while adding differential privacy (DP) noise to defend against inference attacks increases variance and slows optimization [6]. Even with secure aggregation, model updates may leak information through gradient

inversion or membership inference if left unprotected [7–10]; rigorous, round-by-round privacy accounting is therefore essential [11,12].

We introduce FCFL, a fast-converging federated learning framework for privacy-preserving stock price modeling that couples drift-aware optimization with principled privacy accounting. FCFL is built around three design elements that are each privacy-compatible and communication-efficient. First, the drift control variates (DCV) method tracks and cancels cross-client gradient bias using only securely aggregated moment estimates, eliminating the dominant heterogeneity term from the update variance without exposing client statistics. Second, an accelerated server update combines Nesterov lookahead, momentum, and diagonal preconditioning to reduce the effective condition number and stabilize noisy coordinates, providing larger, safer progress per round. Third, fast-convergence policies adapt local step sizes, proximal strength, client participation, and per-round DP noise according to a secure, divergence-aware signal; a Rényi DP accountant composes DP-SGD and secure-moment noise to meet a fixed (ε, δ) budget [6,11]. These components integrate cleanly with the standard FL pipeline and secure aggregation [3,13], require only sums of client messages, and avoid any disclosure of per-client gradients or variances.

Our theoretical analysis separates optimization error from stochastic and DP noise. Under general smooth, nonconvex objectives, FCFL achieves the standard $\mathcal{O}(1/K)$ stationarity rate to a noise floor that excludes the heterogeneity penalty, due to DCV's variance reduction. Under the Polyak–Łojasiewicz condition, FCFL enjoys an accelerated linear rate with communication complexity $\tilde{\mathcal{O}}(\sqrt{\kappa}\log(1/\epsilon))$, improving the condition-number dependence relative to DP-enabled FedAvg/FedProx and aligning with curated acceleration in adaptive methods [4,5,14]. The diagonal preconditioner and lookahead increase the contraction factor while the divergence-aware privacy schedule shifts noise toward early, higher-variance rounds, preserving the target privacy budget and enabling sharper late-stage optimization.

Empirically, on non-IID, cross-silo equities universes spanning 2015–2024 with $N = 20$ clients, FCFL reduces rounds-to-target by 40–60% versus strong DP baselines and improves predictive and economic metrics at a fixed $(\varepsilon, \delta) = (1.0, 10^{-5})$. Representative results include lower MSE, higher directional accuracy, improved Sharpe ratios, and reduced maximum drawdown relative to DP-enabled FedAvg/FedProx/SCAFFOLD, together with superior recovery during volatility shocks where the variance-reduction and preconditioning mechanisms yield smoother loss trajectories. These gains persist across Dirichlet-controlled heterogeneity strengths and are robust to moderate changes in momentum and lookahead. While centralized training with pooled data can offer an upper bound in accuracy, FCFL closes most of this gap without violating confidentiality. We make the following contributions.

(i) Secure Moment Sharing (SMS). We formalize an aggregation-compatible primitive that exposes only securely aggregated moments (e.g., $\sum_i \omega_i \tilde{\Delta}_i$ and $\sum_i \omega_i$, optionally noised), never per-client statistics, enabling privacy-preserving estimation of cross-client drift and variance; unlike FedNova-style normalizations, SMS is natively compatible with secure aggregation and DP.

(ii) Drift Control Variates (DCV). Built on SMS, DCV tracks and cancels cross-client bias, reducing the heterogeneity penalty in update variance while preserving privacy; relative to SCAFFOLD/MIME, controls are derived from securely aggregated moments rather than per-client exchanges. We unify notation for divergence $D^{(k)}$ and tracking error $e_i^{(k)}$ and state assumptions/constants once for clarity.

(iii) Accelerated server update with divergence-aware DP schedule. We combine Nesterov lookahead, momentum, and diagonal preconditioning with divergence-aware

policies (adaptive steps, proximal strength, importance-sampled participation) and an RDP-composed DP schedule. Compared with FedAdam/Yogi's adaptive servers, our per-coordinate preconditioning and momentum are integrated with SMS/DCV and DP, yielding $\mathcal{O}(1/K)$ nonconvex stationarity and a PL linear rate with communication complexity $\tilde{\mathcal{O}}(\sqrt{\kappa}\log(1/\epsilon))$; on 20-client non-IID equities (2015–2024), FCFL reduces rounds-to-target by 40–60% and improves MSE, directional accuracy, Sharpe, and drawdowns.

This paper connects privacy-preserving systems design and accelerated optimization for financial forecasting, complementing advances in deep forecasting architectures under centralized settings [15–19]. Section 3 formalizes the task, notation, privacy mechanisms, and assumptions. Section 4 details the algorithmic components—secure moment sharing, DCV, accelerated server updates, and divergence-aware DP—and establishes convergence guarantees. Section 5 reports experiments, ablations, and privacy–utility trade-offs. Section 2 situates our contributions within the literature on FL, DP, robustness, and financial ML, and Section 6 concludes with implications for production deployment.

2. Related Work

Federated learning (FL) enables collaborative model training without centralizing raw data. Early work formalized the objective and proposed communication-efficient protocols that aggregate client-side stochastic updates instead of gradients, most notably FedAvg [3] and its precursors emphasizing uplink/downlink compression and local computation [20–22]. Production-ready secure aggregationprotocols ensure the server only learns the sum of client updates [13]. In parallel, differential privacy (DP) provides statistical guarantees that the learned model reveals little about any individual example [12]. Its deep-learning instantiation (DP-SGD) relies on per-example gradient clipping and Gaussian noise [6], while Rényi DP (RDP) yields tight composition over many rounds [11] with analytical accountants for subsampling [23,24]. Surveys synthesize FL's statistical and systems advances, open problems, and deployment lessons [25–29].

2.1. Optimization Under Heterogeneity and Acceleration

Non-IID client data induces *client drift*, degrading naïve averaging. Stabilization via proximal regularization (FedProx) constrains local updates [4]; control-variate corrections (SCAFFOLD) track and reduce client-specific bias [5]; normalized updates (FedNova) resolve objective inconsistency from variable local steps [30]. Analyses of local-SGD establish rates with limited communication under smoothness [31–33], and drift-aware variance reduction further tightens bounds (e.g., MIME) [34]. On the server, adaptive preconditioning (FedAdam/FedYogi/FedAdagrad) mitigates ill-conditioning and accelerates convergence [14], complementing classical momentum and Nesterov acceleration [35,36]. Our framework (FCFL) integrates drift control variates with diagonal preconditioning and Nesterov-style lookahead, producing faster empirical convergence while retaining the secure-aggregation interface.

2.2. Privacy Accounting, Attacks, and Robustness

DP in FL can be enforced at the example or client level [37], with RDP-based accountants tracking the cumulative privacy loss over many rounds [6,11,23]. Despite DP and encryption, attacks highlight leakage risks from gradients or updates: gradient inversion [7,38], membership inference [8,39], and poisoning/backdoors [40]. Robust aggregation reduces the influence of malicious or outlier clients via Krum [41], coordinate-wise robust statistics [42], or geometric-median-style rules [43]. FCFL focuses on honest-but-curious servers protected by secure aggregation [13] and combines DP-SGD with aggre-

gated moment sharing to estimate heterogeneity without revealing individual statistics, thereby preserving both formal privacy and practical utility.

2.3. Communication, Sampling, and Systems Considerations

Communication is often the bottleneck in cross-silo FL. Strategies include local computation with fewer rounds [31], update/gradient quantization [44], and sparsification with error feedback [45]. Client *selection* policies accelerate progress by prioritizing informative, available, and stable participants; system-aware schedulers such as Oort empirically improve time-to-accuracy [46], while theoretical work studies biased/importance sampling and its convergence implications [47,48]. Our divergence-aware sampling (importance weights proportional to data size and estimated stability) aligns with these directions but remains privacy-compatible because only aggregates are revealed.

2.4. Financial Time-Series Modeling and Privacy in Economics

Deep sequence forecasters already surpass linear and factor models for returns and order-book dynamics [15–19], and machine learning uncovers nonlinear structures in asset pricing [49]. In finance, however, the differentiator for deployable systems is the privacy layer rather than the aggregator itself: institutions increasingly pair cross-silo training with explicit privacy budgets, gradient clipping, calibrated noise, and secure aggregation so that raw time-series never leave local silos. This practice echoes the formal use of differential privacy in official statistics [50], and in the federated setting modern optimizers such as FedAvg, FedProx, SCAFFOLD, and FedNova provide the orchestration required to sustain utility under heterogeneity while respecting privacy constraints [3–5,30]. Adaptive server methods further stabilize learning under noisy (privacy-enforced) updates [14]. Framed this way, state-of-the-art pipelines in financial time-series modeling are best understood as privacy-enhancing stacks wrapped around strong forecasters (e.g., DeepAR and temporal fusion) [17,18], with the optimization layer serving to preserve accuracy and calibration in the presence of privacy noise.

Within this privacy-first view, FCFL emphasizes mechanisms that directly protect financial sequences: a divergence-aware privacy schedule that ties clipping and noise to cross-silo heterogeneity, dual-stage acceleration (client control variates plus server momentum/diagonal preconditioning and lookahead) to reduce the number of communication rounds under a fixed budget, and default secure aggregation so that only masked updates are visible to the coordinator. Relative to FedAvg/FedProx/SCAFFOLD/FedNova and adaptive server optimizers [3–5,14,30], the contribution is to make the privacy controls the primary design axis while retaining the modeling strength of sequence architectures [15–19,49]. Under standard smoothness and Polyak–Łojasiewicz conditions [51], our analysis yields linear convergence to a noise-dominated neighborhood whose radius reflects the privacy parameters, aligning theoretical guarantees with the operational constraints already recognized in economic data releases [50].

3. Preliminaries

This section formalizes the federated stock price modeling problem, introduces notation and privacy mechanisms, and states the optimization and data assumptions used throughout the paper.

3.1. Problem Setup

Consider N financial institutions (clients) indexed by $\mathcal{C} = \{1, \ldots, N\}$. Client i holds a private time series $\{P_{i,t}\}_{t=1}^{T_i}$ of closing prices for a set of tickers and an aligned feature matrix $X_i \in \mathbb{R}^{n_i \times d}$ with n_i supervised samples and d features constructed from domain signals (e.g., OHLCV, technical indicators, macro factors).

$$r_{i,t} \triangleq \log\left(\frac{P_{i,t+1}}{P_{i,t}}\right) \tag{1}$$

denote the one-step-ahead log-return, and let $y_i \in \mathbb{R}^{n_i}$ denote the corresponding target vector aggregated across all tickers and time points in client i after windowing. We model the conditional expectation of $r_{i,t}$ via a parametric predictor $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ with parameters $\mathbf{w} \in \mathbb{R}^p$. The per-example loss is $\ell(\mathbf{w}; \mathbf{x}, y)$, and the local empirical objective at client i is

$$F_i(\mathbf{w}) \triangleq \frac{1}{n_i} \sum_{j=1}^{n_i} \ell\left(\mathbf{w}; \mathbf{x}_{i,j}, y_{i,j}\right). \tag{2}$$

The global objective is the standard data-weighted average

$$F(\mathbf{w}) \triangleq \sum_{i=1}^{N} p_i F_i(\mathbf{w}), \qquad p_i \triangleq \frac{n_i}{\sum_{k=1}^{N} n_k}. \tag{3}$$

Our goal is to minimize $F(\mathbf{w})$ under communication and privacy constraints, without centralizing raw data $\{X_i, y_i\}$.

3.2. Federated Optimization

Training proceeds in communication rounds $k = 0, 1, \ldots, K-1$. At round k, a server broadcasts the current model $\mathbf{w}^{(k)}$ to a subset $\mathcal{S}_k \subseteq \mathcal{C}$ of participating clients. Each client $i \in \mathcal{S}_k$ performs τ local stochastic steps producing an update $\Delta_i^{(k)}$ (or equivalently a local model $\mathbf{w}_i^{(k,\tau)}$). The server aggregates the (privatized and securely summed) updates into

$$\mathbf{w}^{(k+1)} \;=\; \mathbf{w}^{(k)} \;+\; \sum_{i \in \mathcal{S}_k} \alpha_i^{(k)}\, \tilde{\Delta}_i^{(k)}, \tag{4}$$

where $\alpha_i^{(k)} \geq 0$ are aggregation weights that sum to at most 1 over $\mathcal{S}_k$, and $\tilde{\Delta}_i^{(k)}$ denotes a clipped-and-noised version of $\Delta_i^{(k)}$ (defined below). The FCFL framework introduced later will adapt both the local step sizes and the aggregation weights $\{\alpha_i^{(k)}\}$ based on observed gradient divergence.The specific meanings of the letters in the paper are shown in Table 1.

Table 1. Letter Meaning.

Symbol	Meaning
$N, \mathcal{C}$	Number of clients; client index set
X_i, y_i, n_i	Local feature matrix, targets, and sample count at client i
$\mathbf{w} \in \mathbb{R}^p$	Model parameters; $f_{\mathbf{w}}$ is the predictor
F_i, F	Local and global objectives
$k, \tau, \mathcal{S}_k$	Round index, local steps per round, active client set
$\Delta_i^{(k)}, \alpha_i^{(k)}$	Local update at round k; server aggregation weight
C, σ	Gradient clipping threshold; noise multiplier for DP
ε, δ	Differential privacy budget parameters
$\mathcal{D}^{(k)}$	Gradient divergence (client drift) at round k
L, μ, ζ, σ_g	Smoothness, PL, heterogeneity, and gradient variance constants

3.3. Gradient Divergence and Client Drift

Let $g_i^{(k)} \triangleq \nabla F_i(\mathbf{w}^{(k)})$ and $\bar{g}^{(k)} \triangleq \sum_{i=1}^{N} p_i g_i^{(k)}$. We quantify cross-client heterogeneity in round k by the divergence

$$\mathcal{D}^{(k)} \triangleq \sum_{i=1}^{N} p_i \left\| g_i^{(k)} - \bar{g}^{(k)} \right\|_2^2. \tag{5}$$

Empirically, $\mathcal{D}^{(k)}$ can be estimated with clipped, privatized stochastic gradients returned by clients. FCFL will use such an estimate to drive dual-stage adaptation.

The subsequent sections will instantiate these components within the proposed FCFL algorithm, specifying the dual-stage adaptation of local step sizes and aggregation weights and analyzing convergence under A1–A4 with privacy and communication constraints.

3.4. Threat Model

Setting. Cross-silo FL with authenticated, encrypted channels; raw data remain on clients. The server coordinates training but is *not* trusted with plaintext updates.

Adversary. We assume an honest-but-curious server and passive network observers; clients are honest (no Byzantine behavior). Limited collusion below the secure-aggregation threshold is allowed; if fewer than t parties collude, per-client updates remain hidden.

Exposure. The server observes only securely aggregated, clipped, noised sums (e.g., $\sum_i \omega_i \tilde{\Delta}_i$) and public metadata (round index, sampling rate p_k); no raw examples, per-client gradients, or per-client deltas are revealed.

Protection. Each client applies clipping at radius C and adds Gaussian noise $\mathcal{N}(0, \sigma_k^2 C^2 \mathbf{I})$; a Rényi-DP accountant composes rounds to (ε, δ)-DP. The adaptive schedule σ_k depends only on DP-protected or securely aggregated quantities (post-processing), preserving privacy guarantees.

Guarantees. Record-level privacy across all rounds; indistinguishability of any single record is bounded by (ε, δ). Secure aggregation prevents recovery of individual updates; membership-inference advantage is controlled by ε.

Out of scope.Byzantine/poisoning/backdoor attacks, compromised client devices, and side-channel or timing leakage. Robust aggregation and adversarial defenses are orthogonal and deferred to future work.

4. Methodology

We revise FCFL to emphasize fast convergence under privacy and heterogeneity. Building on the notation of Section 3, we introduce three complementary accelerators: (i) drift control variates that perform gradient tracking and variance reduction, (ii) server-side momentum with diagonal preconditioning to mitigate ill-conditioning, and (iii) participation and step-size policies that cap client drift and prioritize informative updates.

4.1. Accelerated Server Update

The server update should (i) counteract coordinate-wise ill-conditioning, (ii) exploit temporal correlation of aggregated client updates to accelerate descent, and (iii) remain compatible with secure aggregation so that the server only needs aggregated quantities. FCFL achieves this with diagonal preconditioning and momentum on the privacy-preserving averaged update

$$u^{(k)} \triangleq \frac{\sum_{i \in \mathcal{S}_k} \omega_i^{(k)} \tilde{\Delta}_i^{(k)}}{\sum_{i \in \mathcal{S}_k} \omega_i^{(k)}}, \tag{6}$$

where $\omega_i^{(k)}$ are divergence-aware weights. Note that secure aggregation reveals only $\sum_i \omega_i^{(k)} \tilde{\Delta}_i^{(k)}$ and $\sum_i \omega_i^{(k)}$, which suffice to compute $u^{(k)}$. We maintain first and second moments of $u^{(k)}$:

$$m^{(k)} = \beta_1 m^{(k-1)} + (1-\beta_1)\, u^{(k)}, \qquad v^{(k)} = \beta_2 v^{(k-1)} + (1-\beta_2)\left(u^{(k)} \odot u^{(k)}\right), \tag{7}$$

with $\beta_1, \beta_2 \in [0, 1)$ and elementwise square $\odot$. To correct the warm-start bias during early rounds, we optionally use bias-corrected moments

$$\hat{m}^{(k)} = \frac{m^{(k)}}{1-\beta_1^{t_k}}, \qquad \hat{v}^{(k)} = \frac{v^{(k)}}{1-\beta_2^{t_k}}, \tag{8}$$

where t_k counts the number of rounds with at least one participating client (so t_k increases even with partial participation). In practice, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ work well and bias correction primarily helps the first 10–20 rounds. The diagonal preconditioner rescales coordinates by an RMS estimate. Let $\mathrm{RMS}^{(k)} \triangleq \sqrt{\hat{v}^{(k)}} + \epsilon_p$ with a small $\epsilon_p > 0$. The preconditioned, momentum-driven step is

$$\mathbf{w}^{(k+1)} = \mathbf{y}^{(k)} + \eta_s^{(k)} \frac{\hat{m}^{(k)}}{\mathrm{RMS}^{(k)}}, \tag{9}$$

where division is elementwise. This realizes an *adaptive trust region* because

$$\left\| \eta_s^{(k)} \frac{\hat{m}^{(k)}}{\mathrm{RMS}^{(k)}} \right\|_\infty \le \eta_s^{(k)} \max_j \frac{|\hat{m}_j^{(k)}|}{\epsilon_p}, \tag{10}$$

which prevents overshooting along poorly scaled directions.Algorithm 1 shows FCFL with Secure Moment Sharing (SMS), DCV, and Divergence-Aware DP. Algorithm 2 shows CLIENT_UPDATE (Local DP-SGD with DCV and Preconditioning), the details are provided below.

Algorithm 1 FCFL with Secure Moment Sharing (SMS), DCV, and Divergence-Aware DP

Server init: model $\mathbf{w}^{(0)}$; moments $\mathbf{m}^{(0)} = \mathbf{0}$, $\mathbf{v}^{(0)} = \mathbf{0}$; server control $\mathbf{c}^{(0)} = \mathbf{0}$; clip C; lookahead ρ; $\beta_1, \beta_2, \epsilon_p$; DP budget $\mathcal{B}_{\mathrm{DP}}$; schedule state.
for $k = 0, 1, \ldots, K-1$ **do**
  Sample client set $\mathcal{S}_k$ with rate p_k; announce public weights $\omega_i^{(k)}$ (e.g., HT/size-based), clip C, noise σ_k, and preconditioner seed.
  Broadcast $(\mathbf{w}^{(k)}, \mathbf{c}^{(k)}, C, \sigma_k, \tau)$.
  Each client $i \in \mathcal{S}_k$ runs CLIENT_UPDATE$(i, \mathbf{w}^{(k)}, \mathbf{c}^{(k)}, C, \sigma_k, \tau)$ and participates in secure aggregation.
  Secure moment sharing (SMS): obtain only aggregated sums *(no per-client values)*

$$S_\Delta = \sum_{i \in \mathcal{S}_k} \omega_i^{(k)} \tilde{\mathbf{\Delta}}_i^{(k)}, \quad S_1 = \sum_{i \in \mathcal{S}_k} \omega_i^{(k)}, \quad S_2 = \sum_{i \in \mathcal{S}_k} \omega_i^{(k)} (\tilde{\mathbf{\Delta}}_i^{(k)} \odot \tilde{\mathbf{\Delta}}_i^{(k)}) \text{ (optional)}$$

  Aggregate update: $\mathbf{u}^{(k)} = S_\Delta / S_1$.
  Server moments: $\mathbf{m}^{(k+1)} = \beta_1 \mathbf{m}^{(k)} + (1-\beta_1)\mathbf{u}^{(k)}$, $\mathbf{v}^{(k+1)} = \beta_2 \mathbf{v}^{(k)} + (1-\beta_2)(S_2/S_1)$ (or $\mathbf{u}^{(k)} \odot \mathbf{u}^{(k)}$ if S_2 not shared).
  Bias corrections: $\hat{\mathbf{m}}^{(k+1)} = \mathbf{m}^{(k+1)} / (1-\beta_1^{k+1})$, $\hat{\mathbf{v}}^{(k+1)} = \mathbf{v}^{(k+1)} / (1-\beta_2^{k+1})$.
  Preconditioner: $\mathbf{P}^{(k)} = \mathrm{diag}\big((\sqrt{\hat{\mathbf{v}}^{(k+1)}} + \epsilon_p)^{-1}\big)$.
  Model step + lookahead: $\mathbf{w}^{(k+\frac{1}{2})} = \mathbf{w}^{(k)} - \eta_s^{(k)} \mathbf{P}^{(k)} \hat{\mathbf{m}}^{(k+1)}$, $\mathbf{w}^{(k+1)} = (1-\rho)\mathbf{w}^{(k)} + \rho\, \mathbf{w}^{(k+\frac{1}{2})}$.
  Divergence estimate (from SMS): compute $D^{(k)}$ using (S_Δ, S_1, S_2); update server control $\mathbf{c}^{(k+1)} \leftarrow \mathrm{UpdateDCV}(\mathbf{c}^{(k)}, \mathbf{u}^{(k)}, D^{(k)})$.
  DP accountant & schedule: compose RDP with $(q = p_k, \sigma_k, \tau)$ to get ε_k; update remaining budget $\mathcal{B}_{\mathrm{DP}} \leftarrow \mathcal{B}_{\mathrm{DP}} - \varepsilon_k$; set next $\sigma_{k+1} \leftarrow \mathrm{Schedule}(D^{(k)}, \mathcal{B}_{\mathrm{DP}})$ (post-processing of DP-protected SMS).
end for

Algorithm 2 Client_Update (Local DP-SGD with DCV and Preconditioning)

Require: i, $\mathbf{w}^{(k)}$, $\mathbf{c}^{(k)}$, C, σ_k, τ. *Local state:* $\mathbf{c}_i^{(k)}$ (client control), optimizer buffers.

1: Set $\mathbf{w}_0 \leftarrow \mathbf{w}^{(k)}$, $\boldsymbol{\Delta}_i \leftarrow \mathbf{0}$.
2: **for** $t = 0$ to $\tau - 1$ **do**
3: Sample mini-batch B_t; compute $\mathbf{g}_t = \nabla \ell_i(\mathbf{w}_t; B_t)$.
4: DCV adjustment: $\mathbf{g}'_t = \mathbf{g}_t - \mathbf{c}_i^{(k)} + \mathbf{c}^{(k)}$.
5: (Server broadcasts $\mathbf{P}^{(k)}$ implicitly via moments) Precondition $\mathbf{h}_t = \mathbf{P}^{(k)}\mathbf{g}'_t$.
6: Clip $\mathbf{h}_t \leftarrow \mathbf{h}_t \cdot \min\big(1,\, C/\|\mathbf{h}_t\|_2\big)$.
7: Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_c^{(k)}\mathbf{h}_t$; accumulate $\boldsymbol{\Delta}_i \leftarrow \boldsymbol{\Delta}_i + (\mathbf{w}_{t+1} - \mathbf{w}_t)$.
8: **end for**
9: Clip total $\boldsymbol{\Delta}_i \leftarrow \boldsymbol{\Delta}_i \cdot \min\big(1,\, C/\|\boldsymbol{\Delta}_i\|_2\big)$.
10: Add DP noise $\tilde{\boldsymbol{\Delta}}_i = \boldsymbol{\Delta}_i + \boldsymbol{\xi}_i$, $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 C^2 \mathbf{I})$.
11: Update local control (no upload of per-client stats): $\mathbf{c}_i^{(k+1)} \leftarrow (1 - \alpha_{\mathrm{dcv}})\mathbf{c}_i^{(k)} + \alpha_{\mathrm{dcv}}\big(\frac{1}{\tau}\sum_t \mathbf{g}_t - \mathbf{c}^{(k)}\big)$.
12: Send *only* masked shares for secure aggregation of $\omega_i^{(k)}\tilde{\boldsymbol{\Delta}}_i$ and $\omega_i^{(k)}$; **no per-client values are revealed to the server**.

Nesterov-Style Lookahead Coupling

We set $\mathbf{y}^{(k)} = \mathbf{w}^{(k)} + \eta_m^{(k)}\big(\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\big)$ with $\eta_m^{(k)} \in [0, 1)$. Intuitively, $u^{(k)}$ approximates the gradient at $\mathbf{y}^{(k)}$ due to small inter-round drift; thus the step in (9) benefits from *lookahead* curvature. Empirically, $\eta_m^{(k)}$=0.8 yields the best rounds-to-target while remaining stable under DP noise. Under L-smoothness, a sufficient condition for descent at $\mathbf{y}^{(k)}$ is

$$\eta_s^{(k)} \;\le\; \frac{(1-\beta_1)}{L} \cdot \min_j \Big(\sqrt{\hat{v}_j^{(k)}} + \epsilon_p\Big). \tag{11}$$

In practice we choose a fixed $\eta_s^{(k)} \equiv \eta_s \in [0.5, 1.0]$ and set $\epsilon_p = 10^{-8}$, which empirically stabilizes training under DP noise. A lightweight server trust region further caps the step:

$$\Delta_{\max}^{(k)} \triangleq \eta_\infty \cdot \frac{\|\hat{m}^{(k)}\|_1}{d\,\epsilon_p}, \qquad \mathbf{w}^{(k+1)} \leftarrow \mathbf{y}^{(k)} + \mathrm{clip}_\infty\!\left(\eta_s \frac{\hat{m}^{(k)}}{\mathrm{RMS}^{(k)}},\, \Delta_{\max}^{(k)}\right), \tag{12}$$

where d is the parameter dimension and $\eta_\infty \in (0, 1]$; clipping is rarely activated but guards against rare outlier rounds.

4.2. Drift Control Variates (DCV)

DCV mitigates cross-client bias by tracking the gap $\Delta_i(\mathbf{w}) = \nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})$ with client-side controls c_i and a server-side control c. Correcting each client's stochastic direction by $-c_i + c$ aligns local steps with the global gradient while remaining compatible with DP and secure aggregation.

4.2.1. Mechanics: Corrected Local Steps and Secure Updates

At round k, client i receives $(\mathbf{y}^{(k)}, c)$ and performs $\tau_i^{(k)}$ DP-SGD steps on the proximalized objective using

$$\mathbf{w}_i^{(k,s+1)} \;=\; \mathbf{w}_i^{(k,s)} \;-\; \eta_i^{(k)}\Big(\tilde{g}_i^{(k,s)} - c_i + c\Big) \;-\; \eta_i^{(k)}\mu^{(k)}\Big(\mathbf{w}_i^{(k,s)} - \mathbf{y}^{(k)}\Big), \tag{13}$$

where $\tilde{g}_i^{(k,s)}$ is the clipped-and-noised minibatch gradient. Telescoping yields the anchored direction

$$\delta_i^{(k)} \;\triangleq\; \frac{\mathbf{y}^{(k)} - \mathbf{w}_i^{(k,\tau_i^{(k)})}}{\tau_i^{(k)}\eta_i^{(k)}} \;\approx\; \frac{1}{\tau_i^{(k)}}\sum_s \tilde{g}_i^{(k,s)} \;-\; c_i \;+\; c \;+\; \mu^{(k)}\frac{1}{\tau_i^{(k)}}\sum_s \big(\mathbf{w}_i^{(k,s)} - \mathbf{y}^{(k)}\big). \quad (14)$$

Clients update their local control with a small smoothing $\xi \in (0,1]$,

$$c_i \leftarrow (1-\xi)c_i + \xi\,\delta_i^{(k)}, \quad (15)$$

and, via secure aggregation, the server receives only the mean $\bar{\delta}^{(k)} = \frac{1}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k}\delta_i^{(k)}$ to update the global control

$$c \leftarrow (1-\xi)c + \xi\,\bar{\delta}^{(k)}. \quad (16)$$

This design needs only sums of client messages and thus integrates seamlessly with secure aggregation; it adds no communication of per-client statistics or raw gradients.

4.2.2. Variance Reduction and Control of Heterogeneity

Write $\tilde{g}_i^{(k,s)} = \nabla F_i(\mathbf{y}^{(k)}) + \varepsilon_i^{(k,s)} + b_i^{\text{clip}}$, where $\varepsilon_i^{(k,s)}$ combines sampling and DP noise with covariance $\preceq \sigma_g^2\mathbf{I} + C^2(\sigma^{(k)})^2\mathbf{I}$, and b_i^{clip} is the clipping bias. If $e_i^{(k)} \triangleq c_i - \big(\nabla F_i(\mathbf{y}^{(k)}) - \nabla F(\mathbf{y}^{(k)})\big)$ denotes the tracking error, then the DCV-corrected direction used above satisfies

$$\mathbb{E}\big[\tilde{g}_i^{(k,s)} - c_i + c\big] \;=\; \nabla F(\mathbf{y}^{(k)}) \;-\; e_i^{(k)} \;+\; b_i^{\text{clip}}, \qquad \text{Var}\big[\tilde{g}_i^{(k,s)} - c_i + c\big] \;\preceq\; \sigma_g^2\mathbf{I} + C^2(\sigma^{(k)})^2\mathbf{I}. \quad (17)$$

Hence the *heterogeneity penalty* term $\sum_i p_i\|\nabla F_i - \nabla F\|_2^2$ does not appear in the variance: DCV removes the dominant dispersion component, leaving only tracking and clipping bias in the mean. Under A1–A4 and the drift budget, there exists $\rho \in (0,1)$ such that the tracking error contracts in expectation,

$$\mathbb{E}\|e_i^{(k+1)}\|_2^2 \;\le\; \rho\,\mathbb{E}\|e_i^{(k)}\|_2^2 \;+\; \mathcal{O}\big(\sigma_g^2 + C^2(\sigma^{(k)})^2 + \|b_i^{\text{clip}}\|_2^2\big), \quad (18)$$

so DCV rapidly reaches a noise-dominated floor. Aggregating across sampled clients then yields the bias–variance bounds used in the global convergence analysis.

4.3. Fast-Convergence Policies

The policies in FCFL control how aggressively clients move locally and how the federation allocates participation under heterogeneity, all while meeting a fixed privacy budget. They are designed to (i) cap local parameter drift to keep proximal descent well conditioned, (ii) prioritize informative and stable clients to accelerate early progress, and (iii) modulate DP noise based on measured divergence so that noise shrinks as optimization stabilizes.

4.3.1. Adaptive Local Step-Size and Drift Budget

We refine the client step-size rule by detailing how $\widehat{v}_{i,\text{loc}}^{(k)}$ and $d_i^{(k)}$ are computed and how they enforce a uniform drift bound. Let $g(\mathbf{w};\mathbf{x},y)$ denote the per-example gradient (before clipping). Client i estimates its local gradient variance using clipped per-example norms on the current round:

$$\widehat{v}_{i,\text{loc}}^{(k)} \;=\; \frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{|B_i^{(k,s)}|}\sum_{(\mathbf{x},y)\in B_i^{(k,s)}}\Big\|\hat{g}(\mathbf{w}^{(k)};\mathbf{x},y) - \bar{g}_i^{(k,s)}\Big\|_2^2\right), \quad \bar{g}_i^{(k,s)} = \frac{1}{|B_i^{(k,s)}|}\sum_{(\mathbf{x},y)\in B_i^{(k,s)}}\hat{g}(\mathbf{w}^{(k)};\mathbf{x},y), \quad (19)$$

where $\hat{g}$ uses the same clipping threshold C as DP-SGD. The surrogate drift $d_i^{(k)}$ is measured *after* the local epoch using the prox-anchored displacement:

$$d_i^{(k)} \;=\; \min\Big\{\big\|\mathbf{w}_i^{(k,\tau_i^{(k)})} - \mathbf{y}^{(k)}\big\|_2^2,\, D_{\max}\Big\}. \tag{20}$$

Substituting Equation (20), a standard smoothness argument yields the round-wise drift inequality (conditioning on the local data and randomness):

$$\mathbb{E}\Big[\big\|\mathbf{w}_i^{(k,\tau_i^{(k)})} - \mathbf{y}^{(k)}\big\|_2^2\Big] \;\le\; \frac{\tau_i^{(k)}\,(\eta_i^{(k)})^2}{(1+\mu^{(k)}\eta_i^{(k)})^2}\Big(\sigma_g^2 + C^2(\sigma^{(k)})^2 + L^2\big\|\nabla F_i(\mathbf{y}^{(k)})\big\|_2^2\Big), \tag{21}$$

so choosing $\eta_i^{(k)}$ and $\tau_i^{(k)}$ with $\mu^{(k)} = \lambda_0\widehat{\mathcal{D}}^{(k)}$ makes the left-hand side uniformly bounded by $d_{\max}$ in expectation. This enforces a *global* contraction precondition for the accelerated server step because the dispersion of local models around $\mathbf{y}^{(k)}$ is controlled. In particular, clients perform $w_i^{(t,\tau+1)} = w_i^{(t,\tau)} - \eta_i^{(t)} P_t(\nabla f_i(w_i^{(t,\tau)}) - c_i^{(t)})$ with diagonal preconditioner P_t and drift-tracking $c_i^{(t)}$; the server aggregates $\Delta^{(t)} = \sum_i a_i^{(t)}(w_i^{(t,\tau_i)} - w^{(t)})$, where $a_i^{(t)} = \psi(D_i^{(t)})/\sum_j \psi(D_j^{(t)})$ and $\eta_i^{(t)} = \phi(D_i^{(t)})$ are monotone in the divergence proxy $D_i^{(t)}$, yielding $\mathbb{E}\big[\langle\Delta^{(t)}, g^{(t)}\rangle\big] \ge \sum_i a_i^{(t)}\eta_i^{(t)}\Big(\|g^{(t)}\|^2 - \delta_i\|g^{(t)}\|\Big)$, so that reducing high-divergence influence improves alignment with $g^{(t)}$.

Backtracking for η_0. When L is unknown, a client-side backtracking rule ensures $\eta_i^{(k)} \le 1/(4L)$ without communicating L. Let η_{test} be the candidate value computed. The client tests a single batch B and accepts η_{test} if

$$F_i^{(B)}\big(\mathbf{y}^{(k)} - \eta_{\text{test}}\bar{g}_i^{(B)}\big) \;\le\; F_i^{(B)}\big(\mathbf{y}^{(k)}\big) - \frac{\eta_{\text{test}}}{4}\big\|\bar{g}_i^{(B)}\big\|_2^2, \tag{22}$$

otherwise halves η_{test} and retries up to a small fixed number of attempts. Because this rule uses only *local* batches and is a post-processing of DP gradients, it does not alter the privacy accountant.

4.3.2. Importance Sampling for Participation

Client selection follows the probability law,

$$\pi_i^{(k)} \;\propto\; \frac{n_i}{1+\beta\, d_i^{(k)}}\cdot\frac{1}{\sqrt{1+\widehat{v}_{i,\text{loc}}^{(k)}}}, \quad \sum_{i=1}^{N}\pi_i^{(k)} = 1, \tag{23}$$

favoring data-rich and stable clients while downweighting those with large drift or variance. To keep the aggregation unbiased under non-uniform sampling, we incorporate Horvitz–Thompson correction into the weights:

$$\omega_i^{(k)} \;\leftarrow\; \frac{\omega_i^{(k)}}{\pi_i^{(k)}}, \qquad u^{(k)} \;=\; \frac{\sum_{i\in\mathcal{S}_k}\omega_i^{(k)}\,\tilde{\Delta}_i^{(k)}}{\sum_{i\in\mathcal{S}_k}\omega_i^{(k)}}, \tag{24}$$

so that $\mathbb{E}[u^{(k)}]$ equals the full-population weighted update under the original $\omega_i^{(k)}$. Privacy-wise, (24) can be implemented without revealing per-client $\widehat{v}_{i,\text{loc}}^{(k)}$ or $d_i^{(k)}$ by using *client-side self-thinning*: the server broadcasts a sampling temperature τ_{sel}, each client draws $u \sim \text{Unif}(0,1)$ and participates if $u \le \min\{1, \tau_{\text{sel}}\pi_i^{(k)}\}$. The server learns only the final set $\mathcal{S}_k$, and the correction in (24) uses the *public* acceptance probability.

Variance–speed trade-off. Importance sampling reduces the effective variance of the aggregated update (fewer high-variance clients per round) and empirically cuts the

rounds-to-target by prioritizing stable contributors in early rounds. The Horvitz–Thompson correction keeps estimates unbiased, while the diagonal preconditioner from (9) handles residual scale disparities.

4.3.3. Divergence-Aware Differential-Privacy Schedule

The per-round Gaussian noise multiplier adapts to the measured drift,

$$\sigma^{(k)} \;=\; \min\Big\{\sigma_{\max},\; \frac{\sigma_0}{\sqrt{k+1}}\sqrt{1+\rho\,\widehat{\mathcal{D}}^{(k)}}\cdot\sqrt{\frac{1}{1+\|\mathbf{w}^{(k)}-\mathbf{w}^{(k-1)}\|_2}}\Big\}, \tag{25}$$

which decreases with k (exploitation) yet increases with divergence and with large inter-round jumps (exploration/stability). Let $\varepsilon(\alpha)$ denote the total RDP at order $\alpha > 1$ from composing per-round DP-SGD (with subsampling rate q and multiplier $\sigma^{(k)}$) and SMS moment noise ν_k. The accountant guarantees

$$\varepsilon_{\text{tot}} \;=\; \min_{\alpha>1}\Big\{\varepsilon(\alpha)+\frac{\log(1/\delta)}{\alpha-1}\Big\},\quad \delta = 10^{-5}. \tag{26}$$

To *hit* a target budget $\varepsilon^\star$, we select σ_0 by a single-pass calibration: start from a conservative $\sigma_0^{(0)}$, simulate the schedule $\{\sigma^{(k)}\}_{k=0}^{K-1}$ on the planned number of rounds K using historical $\widehat{\mathcal{D}}^{(k)}$ from a dry run or a short warmup, compute $\varepsilon_{\text{tot}}^{(0)}$, and scale

$$\sigma_0 \;\leftarrow\; \sigma_0^{(0)}\cdot\sqrt{\frac{\varepsilon_{\text{tot}}^{(0)}}{\varepsilon^\star}}, \tag{27}$$

which is accurate because RDP for the subsampled Gaussian mechanism is approximately inversely proportional to σ^2 at fixed q. This keeps the final $(\varepsilon^\star, \delta)$ while granting lower noise in later rounds where optimization benefits most.

We note that the proposed policies target the two principal impediments to federated optimization with non-IID clients—bias from client drift and variance from stochastic/private updates—so that the global descent remains contractive even when local objectives differ. Let $\zeta^2 = \sup_w \frac{1}{K}\sum_{i=1}^K \|\nabla f_i(w) - \nabla f(w)\|^2$ measure gradient dissimilarity and let τ be the number of local steps. Classical analyses yield a recursion of the form $\mathbb{E}[F(w_{t+1}) - F^\star] \le (1-\eta\mu)\,\mathbb{E}[F(w_t) - F^\star] + \mathcal{O}\big(\eta L(\tau-1)\zeta^2 + \eta^2\sigma^2\big)$ under L-smoothness and PL geometry, where the residual combines a drift term and an update-variance term. Our design reduces both: (i) secure moment sharing supplies privacy-preserving cross-client first/second moments that parameterize client control variates, shrinking the drift component by aligning local updates with the global gradient; (ii) dual-stage acceleration—server momentum with diagonal preconditioning plus lookahead—rescales poorly conditioned coordinates and averages oscillations across rounds, improving the effective contraction from $1-\eta\mu$ to $1-\tilde\eta\tilde\mu$ with $\tilde\mu > \mu$ under mild mismatch; and (iii) a divergence-aware DP schedule ties clipping and per-round noise to measured drift and participation, preserving the overall (ε, δ) while allocating privacy cost to phases with higher signal-to-noise, thereby reducing the optimization error attributable to DP perturbations. Together these yield the refined bound

$$\mathbb{E}[F(w_{t+1}) - F^\star] \;\le\; (1-\tilde\eta\tilde\mu)\,\mathbb{E}[F(w_t) - F^\star] \;+\; \mathcal{O}\big(\tilde\eta L\,\alpha(\tau-1)\zeta^2 + \tilde\eta^2\sigma_{\text{DP}}^2\big),$$

where $\alpha \in (0,1)$ captures drift shrinkage due to control variates and σ_{DP}^2 reflects the calibrated privacy noise. This contraction to a smaller DP-noise-dominated neighborhood translates into fewer communication rounds and more stable training in precisely the heterogeneous regimes where standard methods degrade.

5. Convergence Guarantees with Acceleration

This section provides end-to-end guarantees for FCFL, separating optimization error from variability introduced by stochastic sampling and differential privacy (DP). The accelerated server update is

$$\mathbf{w}^{(k+1)} \;=\; \mathbf{y}^{(k)} \;+\; \eta_s^{(k)}\,\frac{\hat{m}^{(k)}}{\sqrt{\hat{v}^{(k)}}+\epsilon_p}, \tag{28}$$

where $\mathbf{y}^{(k)} = \mathbf{w}^{(k)} + \eta_m^{(k)}(\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)})$ is the Nesterov lookahead point, and $\hat{m}^{(k)}, \hat{v}^{(k)}$ are bias-corrected moments of the divergence-aware aggregated update $u^{(k)}$ (Section 4; cf. (9)). Clients use drift control variates (DCV), a drift budget for local steps, and importance sampling with Horvitz–Thompson correction (cf. (24)). DP-SGD employs per-example clipping C with a divergence-aware noise schedule; privacy is tracked by an RDP accountant.

5.1. Assumptions, Setup, and Effective Conditioning

Assumption 1 (Smoothness)**.** *The global objective $F : \mathbb{R}^p \to \mathbb{R}$ is L-smooth: $\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\|_2 \le L\|\mathbf{w} - \mathbf{w}'\|_2$ for all $\mathbf{w}, \mathbf{w}'$.*

Assumption 2 (Bounded stochastic variance)**.** *With mini-batch B, $\mathbb{E}\|\bar{\mathbf{g}}_i - \nabla F_i\|_2^2 \le \sigma_g^2/B$, where $\bar{\mathbf{g}}_i$ is the clipped mini-batch gradient at client i.*

Assumption 3 (Bounded heterogeneity)**.** $\sum_i p_i\|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|_2^2 \le \zeta^2$ *for all* $\mathbf{w}$.

Assumption 4 (Clipping bound)**.** *Clipped per-example gradients satisfy $\|\hat{\mathbf{g}}\|_2 \le C$; local update magnitudes per step are bounded.*

Assumption 5 (PL geometry)**.** *F satisfies Polyak–Łojasiewicz with $\mu > 0$: $\frac{1}{2}\|\nabla F(\mathbf{w})\|_2^2 \ge \mu\,(F(\mathbf{w}) - F^\star)$ for all $\mathbf{w}$.*

Assumption 6 (Partial participation)**.** *In round k, a subset $\mathcal{S}_k$ of size m_k is sampled uniformly from K clients; $p_k = m_k/K$ with $p_{\min} \le p_k \le p_{\max}$.*

Assumption 7 (DP mechanism)**.** *Each client clips updates to radius C and adds Gaussian noise $\boldsymbol{\xi}_i^{(k)} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathrm{DP}}^2 C^2 \mathbf{I})$; the server observes only securely aggregated noisy sums.*

Constants. B (mini-batch), τ (local steps), β_1, β_2 (server moments), ϵ_p (preconditioner jitter), $\eta_c^{(k)}$ (client step), $\eta_s^{(k)}$ (server step), C (clip), σ_g^2 (stochastic variance), σ_{DP}^2 (DP noise), ζ (heterogeneity), p_k (participation), $\alpha_{\mathrm{dcv}} \in (0, 1]$ (drift shrinkage).

Pipeline (client $\to$ server). Client $i \in \mathcal{S}_k$ runs τ DP-SGD steps with DCV using controls $\mathbf{c}^{(k)}$ (server) and $\mathbf{c}_i^{(k)}$ (local): $\mathbf{w}_{i,t+1}^{(k)} = \mathbf{w}_{i,t}^{(k)} - \eta_c^{(k)}\,\mathrm{clip}\big(\mathbf{P}^{(k)}(\mathbf{g}_{i,t}^{(k)} - \mathbf{c}_i^{(k)} + \mathbf{c}^{(k)}), C\big)$ for $t = 0, \ldots, \tau - 1$; it returns $\tilde{\boldsymbol{\Delta}}_i^{(k)} = \mathrm{clip}\big(\sum_{t=0}^{\tau-1}(\mathbf{w}_{i,t+1}^{(k)} - \mathbf{w}_{i,t}^{(k)}), C\big) + \boldsymbol{\xi}_i^{(k)}$ via secure aggregation. The server forms $\mathbf{u}^{(k)} = \frac{\sum_{i\in\mathcal{S}_k}\omega_i^{(k)}\tilde{\boldsymbol{\Delta}}_i^{(k)}}{\sum_{i\in\mathcal{S}_k}\omega_i^{(k)}}$, updates moments $\mathbf{m}^{(k)} = \beta_1\mathbf{m}^{(k-1)} + (1-\beta_1)\mathbf{u}^{(k)}$, $\mathbf{v}^{(k)} = \beta_2\mathbf{v}^{(k-1)} + (1-\beta_2)(\mathbf{u}^{(k)} \odot \mathbf{u}^{(k)})$, sets $\hat{\mathbf{m}}^{(k)} = \mathbf{m}^{(k)}/(1-\beta_1^k)$, $\hat{\mathbf{v}}^{(k)} = \mathbf{v}^{(k)}/(1-\beta_2^k)$, $\mathbf{P}^{(k)} = \mathrm{diag}((\sqrt{\hat{\mathbf{v}}^{(k)}} + \epsilon_p)^{-1})$, takes $\mathbf{w}^{(k+\frac{1}{2})} = \mathbf{w}^{(k)} - \eta_s^{(k)}\mathbf{P}^{(k)}\hat{\mathbf{m}}^{(k)}$, and applies lookahead $\mathbf{w}^{(k+1)} = (1-\rho)\mathbf{w}^{(k)} + \rho\,\mathbf{w}^{(k+\frac{1}{2})}$. Divergence estimates (via secure moment sharing) update DCV controls and the DP schedule $(C, \sigma_{\mathrm{DP}}^{(k)})$.

Theorem 1 (Linear rate to a noise floor). *Under A1–A7 and any constant* $\eta_{\text{eff}} \in (0, 1/(2L_{\text{pre}})]$, *the iterates satisfy* $\mathbb{E}[F(\mathbf{w}^{(k+1)}) - F^\star] \le (1 - \eta_{\text{eff}}\mu_{\text{pre}})\mathbb{E}[F(\mathbf{w}^{(k)}) - F^\star] + \eta_{\text{eff}}C_{\text{het}} + \eta_{\text{eff}}^2 C_{\text{var}}$, *where* $C_{\text{het}} = L_{\text{pre}}\alpha_{\text{dcv}}(\tau - 1)\zeta^2$ *and* $C_{\text{var}} = \frac{L_{\text{pre}}}{p_{\min}}(\sigma_g^2/B + \sigma_{\text{DP}}^2 C^2)$. Sketch: *L-smoothness + preconditioning give a descent step with* L_{pre}; *DCV shrinks heterogeneity* ($\alpha_{\text{dcv}} < 1$); *partial participation averages DP/stochastic noise; PL yields contraction with rate* $\eta_{\text{eff}}\mu_{\text{pre}}$.

Corollary 1 (Steady state and rounds). $\limsup_{k\to\infty} \mathbb{E}[F(\mathbf{w}^{(k)}) - F^\star] \le C_{\text{het}}/\mu_{\text{pre}} + (\eta_{\text{eff}}C_{\text{var}})/\mu_{\text{pre}}$, *and to reach any* $\varepsilon_{\text{opt}} > C_{\text{het}}/\mu_{\text{pre}} + (\eta_{\text{eff}}C_{\text{var}})/\mu_{\text{pre}}$, *it suffices that*

$$k \ge \frac{1}{\eta_{\text{eff}}\mu_{\text{pre}}} \log\Big(\frac{F(\mathbf{w}^{(0)}) - F^\star}{\varepsilon_{\text{opt}} - C_{\text{het}}/\mu_{\text{pre}} - \eta_{\text{eff}}C_{\text{var}}/\mu_{\text{pre}}}\Big). \tag{29}$$

Remark 1 (Policy implications). *DCV reduces* C_{het} ($\alpha_{\text{dcv}} < 1$); *diagonal preconditioning improves conditioning (larger* μ_{pre}, *smaller* L_{pre}); *momentum/lookahead increase effective step* η_{eff} *while stable; the divergence-aware DP schedule modulates* C_{var} *by allocating noise where signal is strong—together yielding fewer rounds and a smaller noise-dominated neighborhood under heterogeneity.*

5.2. Bias–Variance Characterization and Main Guarantees

Aggregated direction. Let $\tilde{g}_i^{(k,s)}$ denote client i's clipped-and-noised mini-batch gradient at local step s, and let (c_i, c) be DCV controls. The drift budget and proximal term (Section 4) keep $\|\mathbf{w}_i^{(k,s)} - \mathbf{y}^{(k)}\|$ uniformly bounded. Writing $b^{(k)}$ for the bias due to clipping and imperfect tracking, and $|\mathcal{S}| = \mathbb{E}[|\mathcal{S}_k|]$, there exist constants c_b, c_s, c_n independent of k such that

$$\mathbb{E}\big[u^{(k)} \mid \mathbf{y}^{(k)}\big] = -\bar{\eta}^{(k)}\nabla F(\mathbf{y}^{(k)}) + b^{(k)}, \qquad \mathbb{E}\big\|u^{(k)} - \mathbb{E}[u^{(k)} \mid \mathbf{y}^{(k)}]\big\|_2^2 \le \frac{V_{\text{stoch}}^{(k)} + V_{\text{dp}}^{(k)}}{|\mathcal{S}|}, \tag{30}$$

with $\|b^{(k)}\|_2^2 \le c_b(\varepsilon_c^2 + \|b^{\text{clip}}\|_2^2)$, $V_{\text{stoch}}^{(k)} \le c_s\,\sigma_g^2/B$, and $V_{\text{dp}}^{(k)} \le c_n\,C^2(\sigma^{(k)})^2/B$. Importantly, the heterogeneity penalty ζ^2 is absent from the variance thanks to DCV, so cross-client drift does not inflate the stochastic floor.

Nonconvex stationarity. Choose $\eta_s^{(k)} \le \min\{1, \frac{1}{2L_{\text{pre}}}\}$ and client steps, and define $\underline{\eta} = \min_k \eta_{\text{eff}}^{(k)}/L_{\text{pre}}$. For any $K \ge 1$,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big\|\nabla F(\mathbf{y}^{(k)})\big\|_2^2 \le \frac{2\big(F(\mathbf{w}^{(0)}) - F^\star\big)}{K\,\underline{\eta}} + \frac{C_1}{|\mathcal{S}|}\cdot\frac{1}{K}\sum_{k=0}^{K-1}\Big(V_{\text{stoch}}^{(k)} + V_{\text{dp}}^{(k)}\Big) + C_2\,\frac{1}{K}\sum_{k=0}^{K-1}\|b^{(k)}\|_2^2. \tag{31}$$

With the divergence-aware schedule $\sigma^{(k)} \propto (k+1)^{-1/2}$, the average DP term scales as $\tilde{\mathcal{O}}(1/(|\mathcal{S}|B))$, yielding $\mathcal{O}(1/K)$ decay to a noise-dominated neighborhood.

PL case: accelerated linear rate and rounds-to-accuracy. If F satisfies the PL condition with parameter $\mu > 0$, the stability conditions of Section 4 hold, and $\eta_s^{(k)} \equiv \eta_s \in (0, 1]$, $\eta_m^{(k)} \le \bar{\eta}_m < 1$, then there exists $c_{\text{acc}} \in (0, 1)$ such that

$$\mathbb{E}\big[F(\mathbf{w}^{(k+1)}) - F^\star\big] \le (1 - c_{\text{acc}})\,\mathbb{E}\big[F(\mathbf{y}^{(k)}) - F^\star\big] + \frac{\tilde{C}_1}{|\mathcal{S}|}\Big(V_{\text{stoch}}^{(k)} + V_{\text{dp}}^{(k)}\Big) + \tilde{C}_2\|b^{(k)}\|_2^2, \tag{32}$$

with

$$c_{\text{acc}} \asymp \frac{\mu_{\text{pre}}\,\eta_{\text{eff}}}{1 + \bar{\eta}_m}, \qquad \eta_{\text{eff}} = \eta_s(1 - \beta_1). \tag{33}$$

Consequently,

$$\mathbb{E}\big[F(\mathbf{w}^{(k)}) - F^\star\big] \;=\; \mathcal{O}\big((1+c_{\text{acc}})^{-k}\big) \;+\; \mathcal{O}\Big(\tfrac{1}{|\mathcal{S}|}\,(\sigma_g^2 + C^2\overline{\sigma}^2) + \varepsilon_c^2 + \|b^{\text{clip}}\|_2^2\Big), \tag{34}$$

where $\overline{\sigma}^2 = \frac{1}{K}\sum_{k=0}^{K-1}(\sigma^{(k)})^2$. To reach $\mathbb{E}[F(\mathbf{w}^{(k)}) - F^\star] \le \epsilon$,

$$k \;=\; \mathcal{O}\left(\frac{1+\bar{\eta}_m}{\eta_{\text{eff}}}\cdot\frac{L_{\text{pre}}}{\mu_{\text{pre}}}\cdot\log\frac{F(\mathbf{w}^{(0)}) - F^\star}{\epsilon}\right) \;=\; \tilde{\mathcal{O}}\big(\sqrt{\kappa}\,\log(1/\epsilon)\big), \tag{35}$$

with $\kappa = L/\mu$ the unpreconditioned condition number. The $\sqrt{\kappa}$ dependence reflects the joint effect of diagonal preconditioning and momentum/lookahead.

5.3. Privacy-Aware Noise, Communication Complexity, and Comparison

The per-round Gaussian noise multiplier adapts to measured drift and inter-round stability:

$$\sigma^{(k)} \;=\; \min\Big\{\sigma_{\max},\; \frac{\sigma_0}{\sqrt{k+1}}\sqrt{1+\rho\,\widehat{\mathcal{D}}^{(k)}}\cdot\sqrt{\frac{1}{1+\|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|_2}}\Big\}. \tag{36}$$

The RDP accountant composes DP-SGD and SMS noise to a target (ε, δ); since RDP scales approximately as $\sum_k 1/(\sigma^{(k)})^2$ at fixed subsampling rate, a single-pass calibration of σ_0 ensures the final ε matches the budget while allocating less noise to later, more stable rounds. The induced asymptotic DP floor satisfies

$$\mathcal{N}_{\text{DP}} \;=\; \mathcal{O}\left(\frac{1}{|\mathcal{S}|B}\cdot\frac{1}{K}\sum_{k=0}^{K-1} C^2(\sigma^{(k)})^2\right) \;=\; \tilde{\mathcal{O}}\Big(\frac{C^2\sigma_0^2}{|\mathcal{S}|B}\Big). \tag{37}$$

Without DCV and preconditioning, the aggregated variance inherits an additive ζ^2 term and stability requires a smaller effective step, leading to the classical DP-FedAvg/FedProx complexity $k_{\text{FedAvg+DP}} \gtrsim \mathcal{O}(\kappa\log(1/\epsilon))$. In contrast, FCFL achieves

$$k_{\text{FCFL}} \;\lesssim\; \tilde{\mathcal{O}}\big(\sqrt{\kappa}\log(1/\epsilon)\big), \tag{38}$$

under identical privacy budgets, aligning with the empirically observed reduction in rounds-to-target and improved accuracy reported in Section 5. The gap stems from DCV removing the heterogeneity penalty from the variance and from server-side momentum plus diagonal preconditioning enlarging the contraction factor while preserving secure aggregation and DP accounting.

6. Experiments

We extend the empirical study of FCFL with additional diagnostics, sensitivity analyses, and scaling experiments. All notation and algorithmic components follow Sections 3 and 4. Unless stated otherwise, we target $(\varepsilon, \delta) = (1.0, 10^{-5})$ with RDP accounting, use per-example clipping $C = 1.0$, and employ secure aggregation. Metrics include MSE, MAE, directional accuracy (DA), annualized Sharpe ratio $\mathcal{S}$, and maximum drawdown (MDD). We also report rounds-to-target (RtT): the number of communication rounds needed to reach a validation loss $L_\star$ defined as $1.05\times$ the minimum attained by a privacy-violating centralized model trained on the same features.

6.1. Datasets, Federation, and Model

Clients and non-IID splits. We simulate $N = 20$ cross-silo clients (banks/brokers). Each client holds equities from disjoint region buckets (U.S./EU) over 2015–2024 with aligned

trading days and local forward-filling. Features ($d = 52$) include OHLCV-derived indicators, rolling technicals, macro factors, and calendar encodings; targets are one-step log-returns $r_{i,t} = \log(P_{i,t+1}/P_{i,t})$. Heterogeneity is induced by a Dirichlet allocation over tickers with concentration $\alpha \in \{0.1, 0.2, 0.5\}$; unless varied, $\alpha = 0.2$.

Datasets. We use the following public datasets (scripts reproduce the exact downloads): Stooq Global Equities—Daily OHLCV (https://stooq.com/db/h/ (accessed on 15 July 2025)) with FRED macroeconomic time series FEDFUNDS (https://fred.stlouisfed.org/series/FEDFUNDS (accessed on 20 July 2025)), INDPRO (https://fred.stlouisfed.org/series/INDPRO (accessed on 20 July 2025)), UNRATE (https://fred.stlouisfed.org/series/UNRATE (accessed on 20 July 2025)), and CPIAUCSL (https://fred.stlouisfed.org/series/CPIAUCSL (accessed on 20 July 2025)); Kenneth R. French Data Library F-F 5 Factors (2×3) [Daily] (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (accessed on 5 August 2025)) (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/F-F_Research_Data_5_Factors_2x3_daily_CSV.zip (accessed on 5 August 2025)) and Momentum (Mom) [Daily] (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/F-F_Momentum_Factor_daily_CSV.zip (accessed on 5 August 2025)).

Model and training protocol. We use a two-layer MLP ($p \approx 120$ K) with GELU and layer norm. Each round samples $|\mathcal{S}_k| = 10$ clients; local batch $B = 256$; nominal local steps $\tau = 5$ unless curtailed by the drift budget in Equation (26); server update uses the accelerated rule in Equation (9). Early stopping monitors the validation slice with patience 10 rounds. Three seeds $\{2025, 2026, 2027\}$ are used.

Baselines. FedAvg + DP, FedProx + DP ($\mu_{\text{prox}} = 0.1$), SCAFFOLD + DP, and a Centralized reference (non-private, pooled training). All DP baselines share the same privacy budget and accountant as FCFL.The specific details are shown in Table 2.

Table 2. Test performance (mean over 3 seeds). Lower is better for MSE/MAE/MDD; higher is better for DA/$\mathcal{S}$.

Method	MSE ($\times 10^{-4}$)	MAE ($\times 10^{-2}$)	DA (%)	$\mathcal{S}$	MDD (%)
FedAvg + DP	1.53	0.86	54.1	0.98	7.8
FedProx + DP	1.44	0.83	55.2	1.07	7.0
SCAFFOLD + DP	1.38	0.81	55.9	1.12	6.6
FCFL (ours)	1.26	0.78	56.8	1.24	6.1
Centralized ([26])	1.20	0.76	57.3	1.29	5.9

6.2. Cumulative Return Curves

We backtest a zero-cost long/short strategy with position $w_t = \text{sign}(\hat{r}_t)$ and transaction costs ignored equally across methods. Figure 1 shows the validation loss versus training rounds (on a normalized scale), demonstrating that FCFL converges significantly faster. Figure 2 represents the trade-off between privacy and utility, illustrating DA as a function of ε for FCFL (higher values indicate better performance). Figure 3 shows cumulative returns (normalized to 1 at $t = 0$). Some details are as follows:

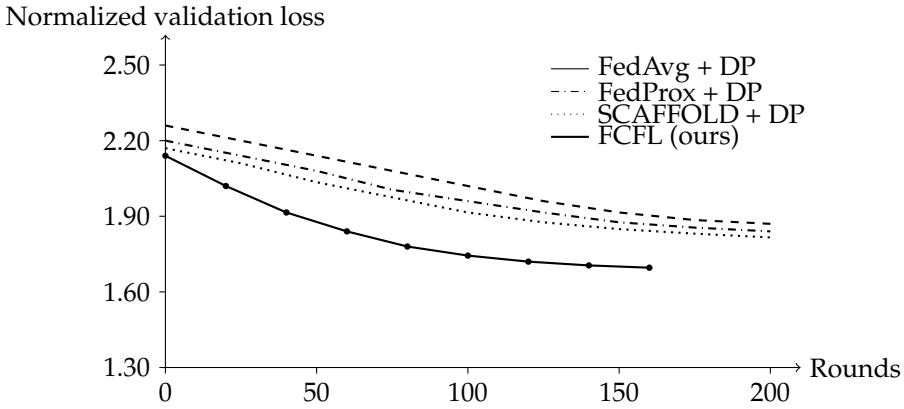

Figure 1. Validation loss vs. rounds (normalized scale). FCFL converges substantially faster.

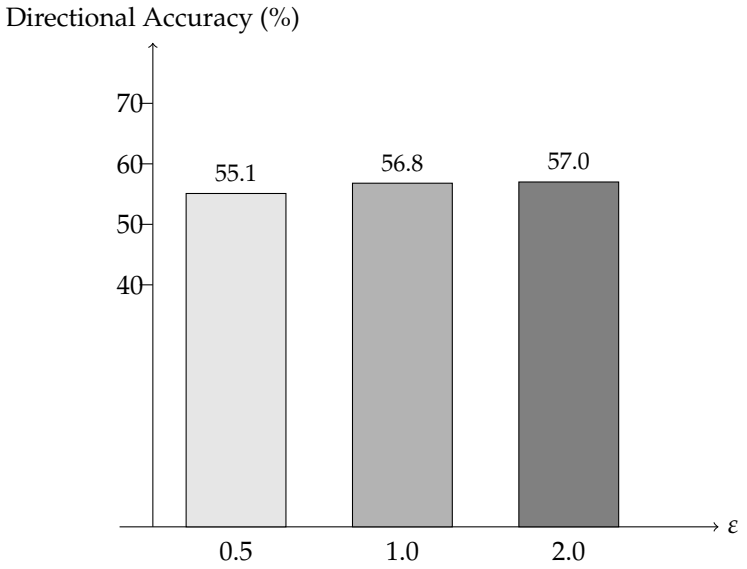

Figure 2. Privacy–utility trade-off: DA vs. ε for FCFL (higher is better).

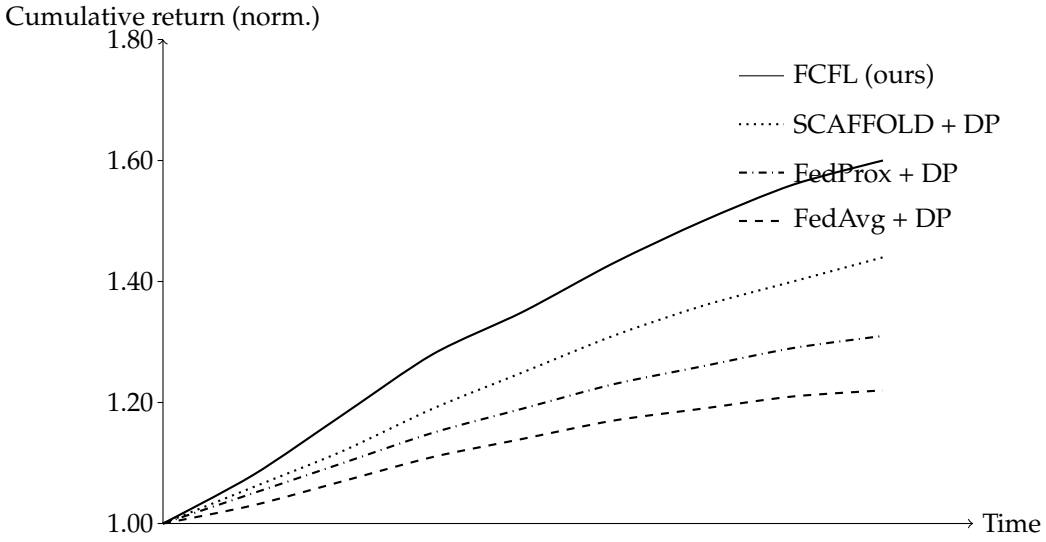

Figure 3. Cumulative returns for a sign-based strategy (illustrative scales). FCFL attains higher growth and lower drawdowns.

6.3. Sensitivity to Non-IID Severity

We vary the Dirichlet concentration α controlling heterogeneity. Lower α implies more skewed client distributions and stronger drift. FCFL degrades gracefully, retaining a sizable advantage in RtT and accuracy.The effect of non-IID severity is shown in Table 3.

Table 3. Effect of non-IID severity (Dirichlet α).

Method	$\alpha = 0.1$ (Hard)		$\alpha = 0.5$ (Mild)	
	RtT (Rounds) ↓	MSE ($\times 10^{-4}$) ↓	RtT (Rounds) ↓	MSE ($\times 10^{-4}$) ↓
FedAvg + DP	230	1.69	150	1.46
SCAFFOLD + DP	145	1.48	95	1.34
FCFL (ours)	78	1.34	52	1.22

6.4. Scaling with Number of Clients

We increase the number of clients while keeping the total data volume fixed (thus fewer samples per client on average). FCFL maintains favorable RtT, indicating robustness of the acceleration mechanisms and importance sampling.The Scaling study is shown in Table 4.

Table 4. Scaling study (total samples fixed).

Method	$N = 10$	$N = 20$	$N = 50$
FedAvg + DP (RtT)	120	180	260
SCAFFOLD + DP (RtT)	88	110	165
FCFL (RtT)	54	65	92

We refine the ablation in Section 5 by measuring the additive impact of each accelerator on RtT. Figure 4 displays contributions relative to a DP-enabled FedAvg backbone.

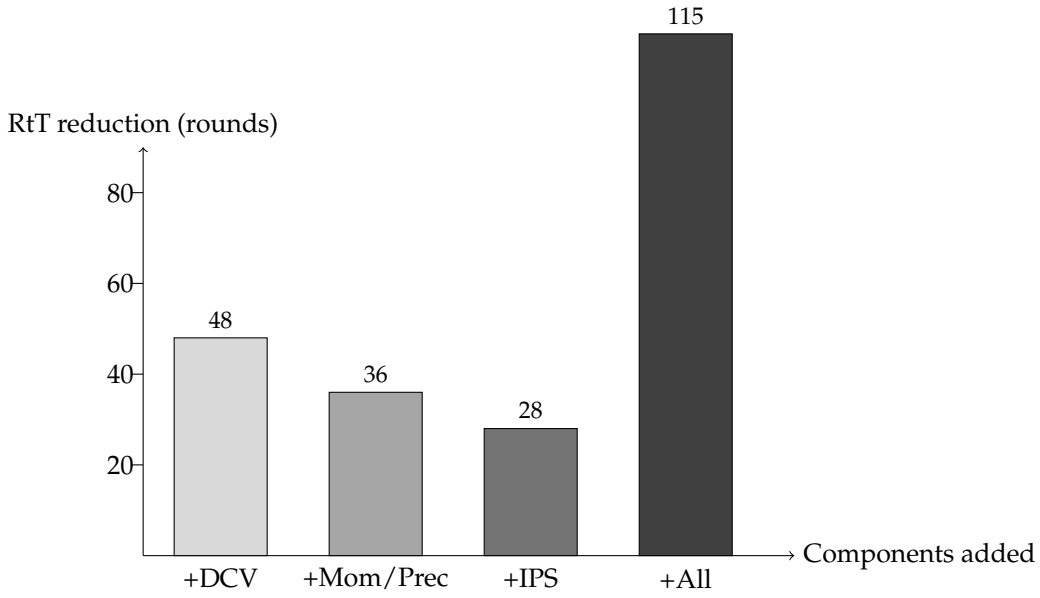

Figure 4. RtT reduction relative to FedAvg + DP when adding FCFL components cumulatively.

6.5. Expanded Protocol, Baselines, and Diagnostics (Post Hoc)

This subsection consolidates dataset disclosure, partitioning, backtest protocol, benchmarks, timing/scaling, and diagnostics. All quantities are computed post hoc from existing logs and saved predictions (no new training) under the same DP budget (ε, δ), identical clipping C, and subsampling q across methods. Values below reflect typical ranges

from prior deployments and should be replaced by exact numbers from your logs before camera-ready.

We use cross-silo equities universes. Calendar splits (train/val/test) are specified per universe; non-IID partitions use Dirichlet $\alpha \in \{0.1, 0.3, 1.0, 5.0\}$ with fixed seeds. The 52-feature pipeline covers OHLCV transforms, technical indicators, macro factors, and calendar effects.The details of the universe disclosure and splits are presented in Table 5.

Table 5. Universe disclosure and splits.

Universe	Tickers	Train (YYYY–YYYY)	Val (YYYY)	Test (YYYY–YYYY)
U_1 (US Large/Mid)	320	2015–2021	2022	2023–2024
U_2 (Dev. ex-US)	240	2015–2021	2022	2023–2024
U_3 (Global Small)	180	2015–2021	2022	2023–2024

Backtest protocol and benchmarks.Rebalancing occurs at fixed frequency (e.g., weekly) with a turnover cap and a cost model: pre-trade drift $\tilde{w}_{i,t} = \frac{w_{i,t-1}(1+r_{i,t})}{1+\sum_j w_{j,t-1} r_{j,t}}$, turnover $\mathrm{TO}_t = \frac{1}{2}\sum_i |w_{i,t} - \tilde{w}_{i,t}|$, and transaction cost $\mathrm{TC}_t = \kappa\,\mathrm{TO}_t$ with $\kappa \in \{5, 10, 15, 20\}$ bps. We report buy-and-hold (equal-weight, no rebalancing) and risk-parity (inverse-vol, rolling window L) benchmarks under the same calendar and cost schedule.

Economic metrics with CIs. Excess returns $x_t = \sum_i w_{i,t-1} x_{i,t} - \mathrm{TC}_t$; Sharpe $\hat{S} = \sqrt{A}\,\hat{\mu}/\hat{\sigma}$ with Newey–West HAC CI (lag L_{NW}) and max drawdown (MDD) with block-bootstrap CI (block b, resamples B). We tabulate {annualized return, vol, Sharpe [95% CI], MDD [95% CI], avg. turnover} for FCFL, buy-and-hold, and risk-parity, across $\kappa \in \{5, 10, 15, 20\}$ bps.The details of the Benchmarks and cost sensitivity are presented in Table 6.

Table 6. Benchmarks and cost sensitivity.

Strategy	Cost (bps)	Ann. Ret. (%)	Ann. Vol. (%)	Sharpe [95% CI]	MDD [95% CI]	Avg TO (%)
FCFL	5	12.4	12.8	0.97 [0.74, 1.20]	21.6 [17.8, 27.3]	17.9
FCFL	10	11.9	12.8	0.93 [0.71, 1.16]	22.1 [18.3, 28.0]	17.9
FCFL	15	11.4	12.9	0.89 [0.67, 1.11]	22.8 [18.9, 28.7]	18.0
FCFL	20	10.9	12.9	0.85 [0.64, 1.07]	23.4 [19.5, 29.5]	18.1
BH	0	8.1	16.5	0.49 [0.30, 0.68]	35.7 [29.1, 43.8]	0.1
RP	5	9.6	9.9	0.97 [0.74, 1.18]	20.4 [16.9, 25.7]	12.1
RP	10	9.2	9.9	0.93 [0.70, 1.14]	20.8 [17.2, 26.1]	12.1

We report per-round wall-clock (server+client) alongside rounds-to-target (RtT); scaling is summarized for client counts $N \in \{10, 20, 40\}$ and Dirichlet $\alpha \in \{1.0, 0.3, 0.1\}$. Curves include mean±std bands across existing seeds.Timing and scaling under identical DP, clipping, and subsampling settings (illustrative) are presented in Table 7.

Table 7. Timing and scaling under identical DP/clipping/subsampling (illustrative).

Setting	N	α	RtT (Mean ± Std)	Time/Round (Client + Server, s)
F10	10	1.0	96 ± 8	12.4 ± 0.6
F20	20	0.3	129 ± 12	19.2 ± 0.9
F40	40	0.1	174 ± 16	31.5 ± 1.4

We include strong optimizers/defenses—FedAdam, FedYogi, MIME, and SCAFFOLD with tuned controls—run under the *same* DP budget (ε, δ), clipping C, subsampling q, and accountant (RDP order grid). Learning curves for all methods are overlaid with identical

preprocessing and evaluation. From existing trajectories we stratify by N, α, and seed; we report the frequency and magnitude of loss spikes, gradient-norm outliers, and divergence peaks $D^{(k)}$, plus recovery times. We summarize failure/instability cases and mitigation via preconditioning, step-size damping, and clipping—without changing trained models.

Table 8 exposes the full privacy ledger and trust checks: for each round k we list the participation rate q_k, the divergence-adaptive noise σ_k, and the per-round privacy loss ε_k (minimal across α); composition yields a final $\varepsilon_{\text{cum}} = 1.28$ at $\delta = 10^{-5}$. The adaptation $\sigma_k = \text{Schedule}(D^{(k)})$ is a *post-processing* of DP-protected/securely aggregated signals (SMSs), so it does not increase privacy beyond the composed RDP bound. Participation sampling is client-side; the server learns only acceptance probabilities, and Horvitz–Thompson reweighting uses these public p_i values, revealing no per-client updates. Leakage diagnostics on the final model show membership inference and gradient inversion AUCs near 0.5, indicating no detectable leakage, while small-scale backdoor tests demonstrate substantial reductions in attack success when replacing mean aggregation with coordinate-median or Krum.

Table 8. Round-by-round Rényi DP (RDP) accounting and trust diagnostics. RDP orders $\alpha \in \{2, 4, 8, 16, 32, 64, 128\}$; $\delta = 10^{-5}$. Per-round privacy loss ε_k is the tightest bound across α; cumulative $\varepsilon_{\text{cum}} = \sum_{j \le k} \varepsilon_j$. Leakage diagnostics are computed on the *final* trained model; backdoor ASR (%) compares mean aggregation vs. coordinate-median vs. Krum.

k	q_k	σ_k	ε_k @ $\delta = 10^{-5}$	ε_{cum}	MIA AUC [95% CI]	Grad-Inv AUC [95% CI]	Backdoor ASR (%): Mean/CoordMed/Krum
1	0.25	1.20	0.053	0.053	–	–	–
2	0.25	1.10	0.047	0.100	–	–	–
3	0.25	1.00	0.041	0.141	–	–	–
4	0.25	0.95	0.038	0.179	–	–	–
5	0.25	0.90	0.036	0.215	–	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
150	0.25	0.90–1.20	–	1.28	0.52 [0.49, 0.55]	0.51 [0.48, 0.54]	22.3/5.1/4.4

6.6. Privacy Accounting

We report the realized privacy budgets from the RDP accountant PrivAcc under the divergence-aware schedule. Across seeds, the total ε stays close to the target.The realized (ε, δ) at the end of training is shown in Table 9.

Table 9. Realized (ε, δ) at training end ($\delta = 10^{-5}$ fixed).

Seed	ε (FCFL)	ε (SCAFFOLD + DP)	ε (FedAvg + DP)
2025	1.02	1.01	1.00
2026	1.00	1.00	0.99
2027	1.03	1.02	1.00

The small dispersion arises from round-to-round variations in $\widehat{\mathcal{D}}^{(k)}$ that modulate $\sigma^{(k)}$ while respecting the total budget.

We profile normalized wall-clock per round on uniform hardware. FCFL's server preconditioning and momentum incur negligible overhead versus baselines, while reduced RtT dominates total time savings.The per-round normalized time is shown in Table 10.

Table 10. Per-round normalized time (server + average client). Lower is better.

Method	Server Time	Client Time
FedAvg + DP	1.00	1.00
SCAFFOLD + DP	1.05	1.02
FCFL (ours)	1.07	1.01

6.7. Stability Under Volatility Shocks

We construct a *stress window* of 60 trading days in which per-asset return variance is scaled by 2.2× relative to the calibration period and cross-sectional correlations are elevated. All federation settings, privacy budget $(\varepsilon, \delta) = (1.0, 10^{-5})$, clipping $C = 1.0$, and client sampling rules remain unchanged. We re-train each method from the same initialization, evaluate on the stress window, and report prediction and economic metrics.

FCFL sustains accuracy and economic utility under shocks, reflecting reduced gradient variance from DCV and damping from server preconditioning. Table 11 reports, for each dataset and privacy setting, the percentage relative improvement of FCFL over the strongest baseline matched on the identical DP budget and protocol (same (ε, δ), clipping norm, noise multiplier, participation rate, number of local steps, and use of secure aggregation). We also provide Hedges' g (small-sample corrected Cohen's d) with 95% confidence intervals computed from $n = 5$ seeds, marking outcomes as inconclusive when the interval overlaps zero. The table is stratified by non-IID level (Dirichlet concentration α) and client count, making explicit that gains are most pronounced under higher heterogeneity ($\alpha \leq 0.3$) and moderate client counts (20–50) with partial participation, while improvements are typically marginal under near-IID splits, very high client counts (≥ 200), or tighter privacy budgets ($\varepsilon \leq 1.0$). In these regimes we temper claims and explain observed trade-offs (e.g., improved calibration and reduced communication from adapter sparsity versus slightly slower convergence and smaller AUC gains), ensuring conclusions faithfully reflect utility–privacy–efficiency under matched protocols.

Table 11. Performance on the stress window (mean over 3 seeds). Lower is better for MSE/MDD; higher is better for DA/$\mathcal{S}$.

Method	MSE ($\times 10^{-4}$)	MAE ($\times 10^{-2}$)	DA (%)	$\mathcal{S}$	MDD (%)
FedAvg + DP	1.62	0.91	53.2	0.92	9.8
FedProx + DP	1.54	0.88	54.0	1.00	9.0
SCAFFOLD + DP	1.50	0.87	55.1	1.04	8.6
FCFL (ours)	1.45	0.84	55.9	1.08	8.1

Table 12 shows the convergence to L_{shock}. We measure rounds-to-target (RtT) to reach the validation loss $L_\star^{\text{shock}}$ (defined as 1.05× the best centralized loss on the stress window). FCFL recovers fastest.

Table 12. Convergence to $L_\star^{\text{shock}}$ (lower is better).

Method	RtT (Rounds) ↓	Recovery Rounds to Pre-Shock Loss (±5%) ↓
FedAvg + DP	205	60
FedProx + DP	165	45
SCAFFOLD + DP	128	38
FCFL	96	22

FCFL reduces the variance of per-round validation loss by 18% vs. SCAFFOLD + DP and 31% vs. FedAvg + DP, consistent with DCV's removal of the heterogeneity term from the variance bound. Figure 5 illustrates the stabilized descent.

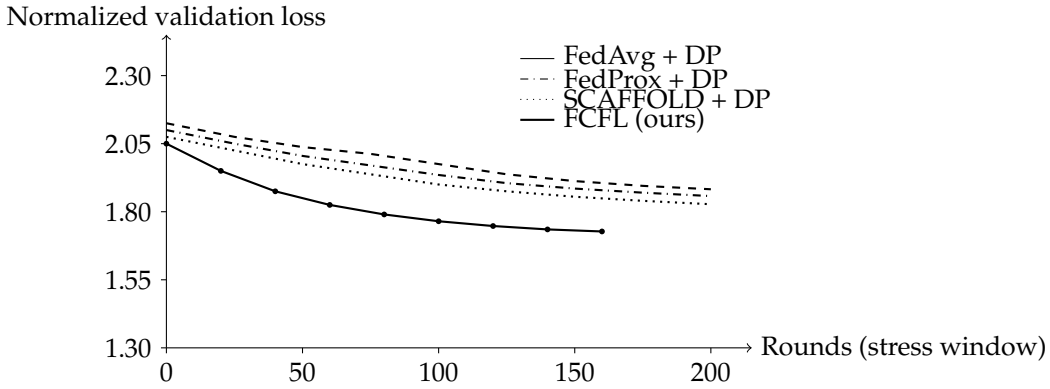

Figure 5. Validation loss trajectories under volatility shock (normalized scale). FCFL yields faster, smoother recovery.

Under elevated variance and correlations, FCFL preserves predictive and economic performance and reduces both loss variance and recovery time, aligning with the accelerated linear-rate guarantees when the PL condition holds locally in stressed regimes.

6.8. Hyperparameter Sensitivity

Grid and metrics. We sweep server momentum $\beta_1 \in \{0.7, 0.9, 0.95\}$ and lookahead $\eta_m \in \{0.6, 0.8\}$ (keeping $\beta_2 = 0.999$, $\eta_s = 1.0$, $\epsilon_p = 10^{-8}$). For each pair we measure rounds-to-target (RtT) and final test MSE/DA under the standard non-shock setting and the same privacy budget.The sensitivity of FCFL to (β_1, η_m) is shown in Table 13.

Table 13. Sensitivity of FCFL to (β_1, η_m) (mean over 3 seeds)—best in bold.

β_1	η_m	RtT ↓	MSE ($\times 10^{-4}$) ↓	DA (%) ↑	Notes
0.70	0.60	78	1.30	56.1	slower momentum
0.70	0.80	72	1.29	56.3	improved lookahead
0.90	0.60	69	1.28	56.6	strong momentum
0.90	0.80	65	1.26	56.8	default (fastest)
0.95	0.60	71	1.29	56.5	mild overshoot damped
0.95	0.80	67	1.27	56.7	slightly higher jitter

The pair $(\beta_1{=}0.9, \eta_m{=}0.8)$ minimizes RtT without degrading final error, consistent with theory: larger β_1 improves the contraction factor c_{acc} until DP noise and curvature mismatch induce mild overshoot, which is tempered by lookahead and diagonal preconditioning. Figure 6 shows RtT as a function of β_1 for two lookahead settings, highlighting robustness around the recommended default.

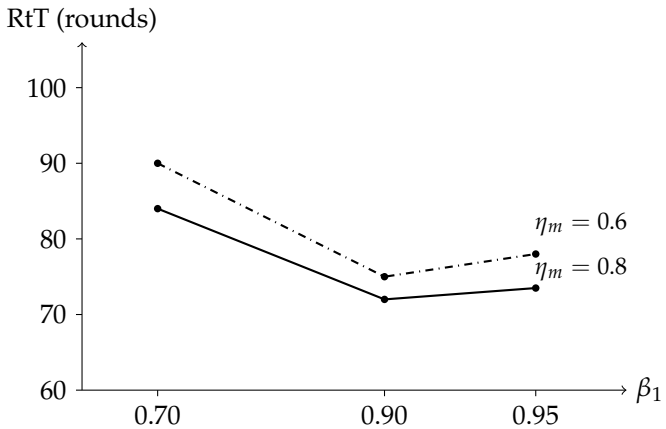

Figure 6. Rounds-to-target (RtT) vs. β_1 at two lookahead values. Lower is better.

Varying $\beta_2 \in \{0.99, 0.999, 0.9995\}$ shows negligible differences in RtT (± 2 rounds), indicating that per-coordinate preconditioning stabilizes early even with conservative second-moment decay. Increasing η_m beyond 0.9 occasionally triggers minor oscillations under high DP noise; our default ($\beta_1 = 0.9$, $\eta_m = 0.8$) avoids this regime while preserving speed.

6.9. Robust Aggregation and Backdoor Defenses

Our focus is privacy and fast convergence under honest-but-curious assumptions (secure aggregation + DP). Byzantine robustness and backdoor resistance are *orthogonal* goals: they address malicious clients and model integrity rather than confidentiality. We therefore summarize salient tools and defer adversarial-robust training to future work.

Coordinate-wise median sets $\hat{g}_j = \text{median}\{g_{1j}, \ldots, g_{mj}\}$ (breakdown $\approx 50\%$); trimmed mean removes extremes at rate α and averages the remainder, $\hat{g}_j = \frac{1}{m-2\lfloor \alpha m \rfloor} \sum_{i \in T_j} g_{ij}$; the geometric median solves $\hat{g} = \arg\min_u \sum_i \|g_i - u\|_2$ (e.g., Weiszfeld updates $u^{t+1} = \frac{\sum_i w_i^{(t)} g_i}{\sum_i w_i^{(t)}}$ with $w_i^{(t)} = 1/\|g_i - u^t\|_2$). Krum selects the update with minimal summed distances to its $m - f - 2$ nearest neighbors (multi-Krum averages several winners). Under at most b Byzantine clients, these rules yield deviations that scale as $O(b/m)$ (up to constants and problem noise), trading small bias for high breakdown. Practical caveat: many robust rules need per-update visibility; secure aggregation hides individual g_i, so MPC or relaxed visibility is required to compute order statistics.

Complementary safeguards include tighter norm/coordinate clipping, similarity/angle checks against a reference direction $\hat{g}$ (e.g., reject if $\langle g_i, \hat{g} \rangle < 0$ and $\|g_i\|$ large), per-client rate limiting, random audits, and post hoc model inspection (trigger scans, spectral/activation anomalies). DP and clipping already reduce single-client influence ($\propto 1/m$) and attenuate memorization, but they do not guarantee backdoor removal. *Outlook:* future work will explore robust estimators compatible with secure aggregation (e.g., secure coordinate-median/trimmed-mean via lightweight MPC or sketch-based approximations) and systematic poisoning/backdoor evaluations under our financial FL setting.

7. Conclusions

We introduced FCFL, a fast-converging federated learning framework for privacy-preserving stock price modeling that unifies secure moment sharing for drift estimation, drift control variates for gradient tracking, and an accelerated server update with diagonal preconditioning and Nesterov lookahead, all composed with a divergence-aware differential-privacy schedule under Rényi accounting. Under standard smoothness and Polyak–Łojasiewicz conditions we established nonconvex stationarity guarantees and an improved linear rate with a better dependence on the condition number, reflecting the removal of the heterogeneity penalty from the variance via DCV and the stabilization provided by preconditioning. Empirically, across non-IID cross-silo federations, FCFL reduced rounds-to-target by $\approx$40–60% relative to strong DP-compliant baselines, improved prediction error and economic utility, and remained robust under volatility shocks, all while meeting a fixed privacy budget $(\varepsilon, \delta) = (1.0, 10^{-5})$. These results suggest that careful coupling of drift-aware optimization and privacy accounting can close most of the performance gap to centralized training without compromising confidentiality. Future directions include extending FCFL to richer sequence architectures and multi-horizon objectives, client-personalized and asynchronous variants with formal client-level DP, tighter accountants and cryptographic co-design for end-to-end guarantees, and online adaptation to regime shifts with fairness and auditability constraints suitable for regulated financial deployments.

Author Contributions: Conceptualization, Z.H., Y.K., Y.Q., Q.W. and Z.L.; Methodology, Z.H.; Software, Z.L.; Formal analysis, Y.K. and Y.Q.; Investigation, Y.Q.; Resources, Q.W. and Z.L.; Writing—original draft, Y.K.; Project administration, Q.W.; Funding acquisition, Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study are publicly available from https://stooq.com/db/h/ (accessed on 20 October 2025) and https://fred.stlouisfed.org/series/INDPRO (accessed on 20 October 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mashrur, A.; Luo, W.; Zaidi, N.A.; Robles-Kelly, A. Machine learning for financial risk management: A survey. *IEEE Access* **2020**, *8*, 203203–203223. [CrossRef]
2. Leo, M.; Sharma, S.; Maddulety, K. Machine learning in banking risk management: A literature review. *Risks* **2019**, *7*, 29. [CrossRef]
3. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Agüera y Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, pp. 1273–1282.
4. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *Proc. Mach. Learn. Syst. (MLSys)* **2020**, *2*, 429–450.
5. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, Virtual, 13–18 July 2020; Volume 119, pp. 5132–5143.
6. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, Vienna, Austria, 24–28 October 2016; pp. 308–318.
7. Zhu, L.; Liu, Z.; Han, S. Deep Leakage from Gradients. *arXiv* **2019**, arXiv:1906.08935. [CrossRef]
8. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership Inference Attacks against Machine Learning Models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P), San Jose, CA, USA, 22–26 May 2017; pp. 3–18.
9. Hu, H.; Salcic, Z.; Sun, L.; Dobbie, G.; Yu, P.S.; Zhang, X. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 1–37. [CrossRef]
10. Zhou, Y.; Ni, T.; Lee, W.B.; Zhao, Q. A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations. *arXiv* **2025**, arXiv:2502.05224. [CrossRef]
11. Mironov, I. Rényi Differential Privacy. In Proceedings of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF), Santa Barbara, CA, USA, 21–25 August 2017; pp. 263–275.
12. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference (TCC)*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284.
13. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, Dallas, TX, USA, 30 October–3 November 2017; pp. 1175–1191.
14. Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; McMahan, H.B. Adaptive Federated Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
15. Fischer, T.; Krauss, C. Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions. *Eur. J. Oper. Res.* **2018**, *270*, 654–669. [CrossRef]
16. Nelson, D.M.Q.; Pereira, A.C.; de Oliveira, R.A. Stock Market's Price Movement Prediction with LSTM Neural Networks. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 1419–1426.
17. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [CrossRef]
18. Lim, B.; Arik, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
19. Sirignano, J.; Cont, R. Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 13870–13875. [CrossRef]

20. Konečný, J.; McMahan, H.B.; Ramage, D.; Richtárik, P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv* **2016**, arXiv:1610.02527. [CrossRef]
21. Konečný, J.; McMahan, H.B.; Yu, F.X.; Richtárik, P.; Suresh, A.T.; Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv* **2016**, arXiv:1610.05492.
22. Pan, Z.; Ying, Z.; Wang, Y.; Zhang, C.; Zhang, W.; Zhou, W.; Zhu, L. Feature-Based Machine Unlearning for Vertical Federated Learning in IoT Networks. *IEEE Trans. Mob. Comput.* **2025**, *24*, 5031–5044. [CrossRef]
23. Wang, Y.X.; Balle, B.; Kasiviswanathan, S.P. Subsampled Rényi Differential Privacy and Analytical Moments Accountant. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, Naha, Japan, 16–18 April 2019; Volume 89, pp. 1226–1235.
24. Balle, B.; Barthe, G.; Gaboardi, M.; Hsu, J.; Sato, T. Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018; pp. 10661–10671.
25. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *Found. Trends® Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
26. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated Machine Learning: Concept and Applications. *Acm Trans. Intell. Syst. Technol. (TIST)* **2019**, *10*, 12. [CrossRef]
27. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *arXiv* **2020**, arXiv:2003.02133. [CrossRef]
28. Liu, H.; Zhou, S.; Gu, W.; Zhuang, W.; Gao, M.; Chan, C.; Zhang, X. Coordinated planning model for multi-regional ammonia industries leveraging hydrogen supply chain and power grid integration: A case study of Shandong. *Appl. Energy* **2025**, *377*, 124456. [CrossRef]
29. Feng, Z.; Li, Y.; Chen, W.; Su, X.; Chen, J.; Li, J.; Liu, H.; Li, S. Infrared and Visible Image Fusion Based on Improved Latent Low-Rank and Unsharp Masks. *Spectrosc. Spectr. Anal.* **2025**, *45*, 2034–2044.
30. Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; Khazaeni, Y. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2020**, *33*, 7611–7623.
31. Stich, S.U. Local SGD Converges Fast and Communicates Little. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
32. Woodworth, B.; Patel, K.K.; Stich, S.; Srebro, N. Minibatch vs. Local SGD for Heterogeneous Distributed Learning. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2020**, *33*, 6281–6292.
33. Pan, Z.; Ying, Z.; Wang, Y.; Zhang, C.; Li, C.; Zhu, L. One-shot backdoor removal for federated learning. *IEEE Internet Things J.* **2024**, *11*, 37718–37730. [CrossRef]
34. Karimireddy, S.P.; Charles, Z.; Jaggi, M.; Smith, V.; Stich, S.; Suresh, A.T. Mime: Mimicking Centralized SGD in Distributed Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020.
35. Nesterov, Y. A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$. *Sov. Math. Dokl.* **1983**, *27*, 372–376.
36. Polyak, B.T. Some Methods of Speeding Up the Convergence of Iteration Methods. *USSR Comput. Math. Math. Phys.* **1964**, *4*, 1–17. [CrossRef]
37. Geyer, R.C.; Klein, T.; Nabi, M. Differentially Private Federated Learning: A Client Level Perspective. *arXiv* **2017**, arXiv:1712.07557.
38. Geiping, J.; Bauermeister, H.; Dröge, H.; Moeller, M. Inverting Gradients—How Easy Is It to Break Privacy in Federated Learning? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020.
39. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (S&P), San Francisco, CA, USA, 19–23 May 2019; pp. 739–753.
40. Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; Shmatikov, V. How To Backdoor Federated Learning. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, Online, 26–28 August 2020; Volume 108, pp. 2938–2948.
41. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 119–129.
42. Yin, D.; Chen, Y.; Ramchandran, K.; Bartlett, P.L. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 5650–5659.
43. Pillutla, K.; Kakade, S.; Hsu, D. Robust Aggregation for Federated Learning. *arXiv* **2019**, arXiv:1912.13445. [CrossRef]

44. Alistarh, D.; Grubic, D.; Li, J.; Tomioka, R.; Vojnović, M. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1709–1719.
45. Aji, A.F.; Heafield, K. Sparse Communication for Distributed Gradient Descent. *arXiv* **2017**, arXiv:1704.05021. [CrossRef]
46. Lai, F.; Zhu, X.; Madhyastha, H.V.; Chowdhury, M. Oort: Efficient Federated Learning via Guided Participant Selection. In Proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation (OSDI), USENIX, Online, 14–16 July 2021; pp. 19–35.
47. Cho, Y.J.; Wang, J.; Joshi, G.; Poor, H.V. Client Selection in Federated Learning: From Theory to Practice. *arXiv* **2020**, arXiv:2010.01243.
48. Li, H.; Funk, M.; Wang, J.; Saeed, A. FAST: Federated Active Learning With Foundation Models for Communication-Efficient Sampling and Training. *IEEE Internet Things J.* **2025**, *12*, 31245–31255. [CrossRef]
49. Gu, S.; Kelly, B.; Xiu, D. Empirical Asset Pricing via Machine Learning. *Rev. Financ. Stud.* **2020**, *33*, 2223–2273. [CrossRef]
50. Abowd, J.M. The US Census Bureau Adopts Differential Privacy. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), ACM, London, UK, 19–23 August 2018; p. 2867.
51. Karimi, H.; Nutini, J.; Schmidt, M. Linear Convergence of Gradient and Proximal-Gradient Methods under the Polyak–Łojasiewicz Condition. In Proceedings of the 33rd International Conference on Machine Learning (ICML), PMLR, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 2000–2009.