COLLABORATIVE INFERENCE FOR FEDERATED NOISY LABELS: CONVERGENCE UNDER DUAL HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous label noise across clients poses critical challenges to robust training and stable convergence in federated learning. Existing methods largely rely on local label correction, limiting their ability to leverage global information. Moreover, theoretical convergence analysis under dual data-noise heterogeneity remains under-explored. To address this, we propose FedCoIn, a **fed**erated noisy label learning framework based on generative **co**llaborative **in**ference. This method treats ground-truth labels as latent variables and performs label inference through global generative modeling, enabling cross-client collaborative noise correction under privacy constraints. Since heterogeneity may disrupt generative inference and undermine global correction consistency, FedCoIn introduces a distribution-aware dynamic prior and a noise-aware weighted aggregation strategy to stabilize label inference and harmonize corrections across clients. Theoretically, we provide convergence analysis under three heterogeneous scenarios, including data homogeneity with noise heterogeneity, data heterogeneity with noise homogeneity, and dual data-noise heterogeneity. This analysis characterizes the impact of both types of heterogeneity on convergence, thereby establishing a new theoretical foundation for federated noisy label learning and providing theoretical support for FedCoIn. Experimental results demonstrate that FedCoIn consistently outperforms state-of-the-art methods in federated noisy label learning tasks.

1 INTRODUCTION

1.1 MOTIVATION

Federated learning (FL) has emerged as a framework for decentralized machine learning, enabling clients to collaboratively train models without sharing raw data, addressing data silos and privacy concerns (Zhang et al., 2023; Jiang et al., 2024b). This paradigm is valuable in privacy-sensitive domains such as healthcare (Kaissis et al., 2020; Rani et al., 2023) and finance (Zheng et al., 2021; Li et al., 2024b). However, in practical scenarios, due to variations in annotation quality, subjective bias, and device heterogeneity, client data is contaminated with label noise, which is inconsistent across clients (Li et al., 2021; Zhu et al., 2021). Such heterogeneous label noise degrades model performance, increases optimization difficulty and may lead to training instability. Therefore, how to robustly handle label noise in federated settings remains a critical challenge.

To address this challenge, federated noisy label learning (FNLL) methods have recently emerged. Some studies attempt to mitigate label noise by using heuristic rules for local label correction (Xu et al., 2022; Jiang et al., 2024a). However, these methods primarily rely on local data from each client and fail to effectively leverage cross-client global information, making it difficult to accurately characterize the underlying clean label distribution. Due to the **absence of global collaborative modeling, local label correction struggles to achieve a consistent global correction objective across clients**, and is susceptible to local noise and data bias, which can lead to the accumulation of incorrect labels and noise memorization. Furthermore, the coupling between data heterogeneity (Li et al., 2020a; Chen et al., 2021) and noise heterogeneity (Chen et al., 2020; Yang et al., 2022) amplifies local correction errors during model aggregation, propagating them to the global model, which weakens both generalization ability and training stability. Therefore, effectively leveraging

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

cross-client information under privacy constraints to reliably infer latent clean labels, while ensuring stable learning under dual heterogeneity, remains an unresolved problem in FNLL.

Moreover, although many FNLL methods have been proposed, the **convergence theory for FNLL under dual data–noise heterogeneity has yet to be established**. Convergence analysis characterizes whether an optimization process can achieve stable convergence and provides upper bounds on final error, thereby offering theoretical support for algorithmic reliability, making it important for FNLL. However, existing convergence analyses in FL are mainly developed under noise-free settings and focus on data heterogeneity (Li et al., 2020b; Karimireddy et al., 2020; Li et al., 2020a), typically assuming that client data differ only in distribution while ignoring label noise. In more complex scenarios where both data and noise heterogeneity coexist, traditional convergence theories are not directly applicable. In particular, when noise distributions vary across clients, label noise not only interferes with local optimization but is also amplified during model aggregation, further degrading the convergence stability of the global model. Therefore, developing a convergence theory that captures the coupled effects of data and noise heterogeneity is crucial for understanding the effectiveness and design principles of FNLL methods.

1.2 Related Work

Federated Noisy Label Learning. FNLL is more challenging than CL due to data heterogeneity, noise heterogeneity, and privacy constraints. To address these issues, several methods have been proposed. FedCorr (Xu et al., 2022) estimates client-level noise using local intrinsic dimensionality and performs multi-stage label correction. FedNoRo (Wu et al., 2023) takes a class-aware approach to identify and handle noisy clients, while FedELC (Jiang et al., 2024a) detects noisy clients and corrects labels via backpropagation. FedDiv (Li et al., 2024a) introduces a global filtering mechanism to suppress noisy predictions, FedFixer (Ji et al., 2024) employs a dual-model design for noise correction, and FedClean (Jiang & Zhang, 2025) removes clean-client and noise distribution assumptions, achieving more robust performance.

Convergence of Federated Learning. The convergence of FL has been extensively studied. On Non-IID data fundamentals, (Li et al., 2020b; Khaled et al., 2020; Woodworth et al., 2020) analyze convergence bounds of FedAvg and local SGD as well as client drift from data heterogeneity. To mitigate drift, (Yao et al., 2024; Karimireddy et al., 2020) achieve more stable convergence in heterogeneous scenarios via control variates and knowledge distillation. For complex system scenarios, (Wang & Ji, 2022; Koloskova et al., 2020; Chen et al., 2024) cover convergence analyses under arbitrary client participation, decentralized dynamic topology, and differential privacy constraints. However, most of these analyses assume noise-free environments. To the best of our knowledge, no existing work has provided convergence theory under dual heterogeneity where both data heterogeneity and noise heterogeneity coexist.

1.3 Contributions

To address challenges in both methodology and theory, we propose a robust federated noisy label learning framework FedCoIn, as illustrated in Figure 1. We model true labels as latent variables and enable collaborative noise correction through a global generative mechanism. On the theoretical side, we provide a convergence analysis under three heterogeneous settings, with the proof roadmap illustrated in Figure 2. The main contributions are as follows:

- **Framework**. We propose FedCoIn, which achieves cross-client collaborative noise correction via global generative modeling, thereby overcoming the limitations of relying solely on local information. We design a distribution-aware dynamic prior and a noise-aware weighted aggregation strategy to improve inference stability and global consistency.
- **Theory**. We theoretically prove convergence under three heterogeneous scenarios, including data heterogeneity with noise homogeneity, data homogeneity with noise heterogeneity, and dual heterogeneity, thereby ensuring convergence to the optimal solution under noisy conditions and providing a new theoretical guarantee for federated noisy label learning.
- **Validation**. Experimental results demonstrate that FedCoIn consistently outperforms state-of-the-art methods across diverse heterogeneous noise scenarios, achieving considerable performance on CIFAR-100 under non-IID and fully noisy client settings.

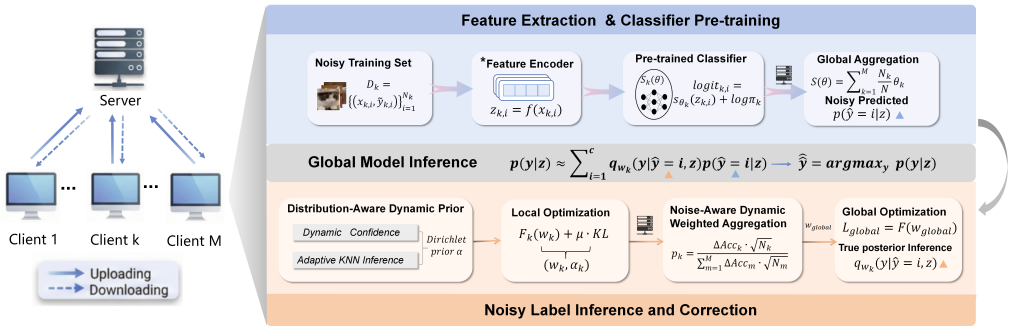


Figure 1: Overall architecture of FedCoIn. Each client extracts robust representations using a feature extractor and trains a lightweight classifier for stable predictions. It performs local optimization guided by a distribution-aware dynamic prior. The server aggregates updates via noise-aware weighting to refine the global model and iteratively improve the posterior. Based on the refined posterior, a corrected label distribution generates pseudo-labels, enabling cross-client collaborative noise correction. Modules marked with "*" are optional and can be flexibly enabled or omitted.

2 METHODOLOGY

2.1 PREPARATION

Federated Learning Setting. We consider a standard FL scenario consisting of one central server and M clients. Each client k holds a local private dataset $\mathcal{D}_k = \{(\boldsymbol{x}_{k,i}, \tilde{y}_{k,i})\}_{i=1}^{N_k}$, where N_k denotes the number of samples at client k, $\boldsymbol{x}_{k,i} \in \mathbb{R}^d$ is the feature vector of the i-th sample with dimension d, and $\tilde{y}_{k,i} \in \{0,1\}^C$ is the noisy label over C classes. The true label $y_{k,i}$ is unobservable during training. The server has no access to raw data and performs global aggregation and optimization based on the model parameters and statistics uploaded by the clients.

Label Noise Modeling. Local label noise is mainly manifested as a distributional shift between the predicted noisy label $\hat{y}_{k,i}$ and the true label $y_{k,i}$. To characterize this uncertainty, we model label properties using Dirichlet distributions. We first construct a dynamic prior Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha}_{\mathrm{prior},k,i})$, where $\boldsymbol{\alpha}_{\mathrm{prior},k,i}$ denotes the prior parameters. Subsequently, variational inference is employed to obtain posterior Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha}_{\mathrm{infer},k,i})$, where $\boldsymbol{\alpha}_{\mathrm{infer},k,i}$ represents the posterior parameters, corresponding to the noise-corrected estimation of the label distribution.

2.2 PREPROCESSING

Feature Extraction. Each client uses a pre-trained feature extractor $f(\cdot)$ to obtain representations $\boldsymbol{z}_{k,i} = f(\boldsymbol{x}_{k,i}) \in \mathbb{R}^d$. Our framework is agnostic to the choice of feature extractor and supports various architectures. The extracted features are directly used for the classifier. Alternatively, when the feature extractor is not adopted, the raw input $\boldsymbol{x}_{k,i}$ is directly fed into the classifier.

Classifier Pre-training. Each client trains a lightweight classifier on extracted features to provide an initial prior for label distribution modeling in the subsequent variational inference stage. The goal is not to maximize classification accuracy, but to produce stable and discriminative initial predictions. The design is flexible, making the framework generally applicable. Specifically, client k employs a lightweight network $s_{\theta_k}(\cdot)$, which takes $\boldsymbol{z}_{k,i}$ as input and outputs logits $\mathrm{logit}_{k,i} = s_{\theta_k}(\boldsymbol{z}_{k,i}) + \log \boldsymbol{\pi}_k$, where $\boldsymbol{\pi}_k \in \mathbb{R}^C$ denotes the estimated local class prior. The logits are then converted into probabilities via Softmax, i.e., $\hat{p}_{k,i} = \mathrm{Softmax}(\mathrm{logit}_{k,i})$. The classifier is trained using cross-entropy loss. After training, parameters θ_k are fixed, and $\hat{p}_{k,i}$ is used as noisy label $\hat{y}_{k,i}$ in the variational inference stage.

2.3 NOISY LABEL INFERENCE AND CORRECTION

The core idea of generative collaborative inference for noisy labels is to infer a probabilistic mapping from noisy predictions to true labels via global-aware probabilistic modeling and variational inference.

Probabilistic Model Construction. The goal of probabilistic model construction is to characterize relationships among the input sample, noisy prediction, and true label via distributional assumptions. The true label is treated as a latent variable to enable robust inference under noisy supervision.

Specifically, given input z, the latent true label y is assumed to follow a Dirichlet prior, i.e., $y \sim \mathrm{Dir}(\boldsymbol{\alpha}_z)$, where $\boldsymbol{\alpha}_z \in \mathbb{R}_+^c$ denotes the concentration parameters over c classes, capturing uncertainty over label distributions. The observed noisy prediction $\hat{y}$ is modeled as a multinomial variable conditioned on y and z, i.e., $\hat{y} \sim \mathrm{Multi}(\boldsymbol{\pi}_{z,y})$, where $\sum_{i=1}^{c} \pi_{z,y}^i = 1$. This formulation captures the stochastic corruption process from true labels to noisy predictions.

Under these assumptions, the joint distribution factorizes as

$$p(y, \hat{y}, z) \propto p(y \mid z)\, p(\hat{y} \mid y, z). \tag{1}$$

providing a probabilistic foundation for variational inference and noise correction.

Distribution-Aware Dynamic Prior (DADP). To construct reliable Dirichlet priors under noisy labels and heterogeneous data, we propose a distribution-aware strategy that combines confidence-based sample selection with adaptive KNN aggregation to build uncertainty-aware priors.

Dynamic Confidence-based Selection. Let $\hat{\mathbf{y}}_k = \{\hat{y}_i\}_{i=1}^{N_k}$ and p_i be the classifier prediction for sample z_i. The confidence is defined as $c_i = \max_j p_i^j$. For each class j, denote $\mathcal{S}_j = \{i \mid \hat{y}_i = j\}$. A high-confidence set is constructed as $\mathcal{H}_j = \{i \in \mathcal{S}_j \mid c_i \geq \tau_j\}$, where τ_j is adaptively determined from class-wise confidence statistics. To ensure class balance, if $|\mathcal{H}_j| < m_j$, the top-m_j samples in $\mathcal{S}_j$ are selected as $\mathcal{T}_j$. The final reference set is $\mathcal{H} = \bigcup_{j=1}^{C} (\mathcal{H}_j \cup \mathcal{T}_j)$.

Adaptive KNN Inference. For each $z_i \in \mathcal{H}$, its neighborhood $\mathcal{N}_i$ consists of K_i nearest samples in embedding space, where $K_i = \min(K_{\max}, \max(K_{\min}, f(N, \rho_i)))$, with $f(\cdot)$ controlling the neighborhood size based on local data density. To reduce noise influence, distance-weighted voting is used with $w_{ij} = 1/(\|z_i - z_j\|_2 + \epsilon)$. The class distribution is computed as

$$p_i^c = \frac{\sum_{z_j \in \mathcal{N}_i} w_{ij} \mathbf{1}[y_j = c]}{\sum_{z_j \in \mathcal{N}_i} w_{ij}}. \tag{2}$$

Dirichlet Prior Construction. Let $\bar{y} = \arg\max_c p_i^c$. The Dirichlet prior is defined as

$$\alpha_z^i = \begin{cases} \delta, & i \neq \bar{y}, \\ \delta + \rho, & i = \bar{y}. \end{cases} \tag{3}$$

where δ and ρ control prior concentration and sharpness. This yields a sample-specific Dirichlet prior robust to noise and heterogeneity.

Parameter Inference. We adopt a hierarchical optimization framework that combines client-side local training with server-side aggregation to ensure local correction and global consistency. Each client infers posterior $p(y \mid \hat{y}, z)$ via variational approximation $q_{w_k}(y \mid \hat{y}, z)$, where w_k denotes local variational parameters. The server aggregates client parameters to enable global collaboration.

Local Optimization. The local objective minimizes the KL divergence between $q_{w_k}(y \mid \hat{y}, z)$ and the true posterior:

$$\mathrm{KL}(q_{w_k}(y \mid \hat{y}, z) \,\|\, p(y \mid \hat{y}, z)) = \log p(\hat{y} \mid z) - \mathrm{ELBO}_k, \tag{4}$$

thus maximizing ELBO_k yields an accurate posterior approximation.

To avoid overfitting, we introduce a global alignment regularization that enforces consistency between local and global posteriors. The resulting local loss is:

$$\begin{aligned} L_k = & \sum_{i=1}^{c} \hat{y}_k^i \log \hat{y}_k^{*i} + (1 - \hat{y}_k^i) \log(1 - \hat{y}_k^{*i}) - \sum_{i=1}^{c} \log \Gamma(\alpha_{k,z}^i) + \sum_{i=1}^{c} \log \Gamma(\hat{\alpha}_{k,z,\hat{y}}^i) \\ & - \sum_{i=1}^{c} (\hat{\alpha}_{k,z,\hat{y}}^i - \alpha_{k,z}^i) \psi(\hat{\alpha}_{k,z,\hat{y}}^i) + \mu \cdot \mathrm{KL}\big(\mathrm{Dir}(\hat{\boldsymbol{\alpha}}_{k,z,\hat{y}}) \,\|\, \mathrm{Dir}(\hat{\boldsymbol{\alpha}}_{\mathrm{global}})\big). \end{aligned} \tag{5}$$

where $\hat{y}_k^i$ is the noisy prediction, $\hat{y}_k^{*i}$ the reconstructed probability, $\alpha_{k,z}^i$ the Dirichlet prior, and $\hat{\alpha}_{k,z,\hat{y}}^i$ the variational parameter. Each client optimizes L_k via gradient descent and uploads w_k and $\boldsymbol{\alpha}_k$, preserving privacy. Details of Dirichlet reparameterization and KL derivation are in Appendix D.

Noise-Aware Dynamic Weighted Aggregation (NADA). To leverage client-wise correction while reducing statistical noise from small-sample clients, we propose a noise-aware dynamic weighted aggregation (NADA) strategy. The key metric $\Delta\mathrm{Acc}_k$ measures the relative accuracy improvement on a noisy validation set under federated privacy constraints, reflecting the ability to mitigate noise-induced classifier bias. Since models trained on noisy labels naturally deviate from clean-label predictions, this correction gain is correlated with calibration quality. Due to privacy restrictions, clean validation data is unavailable, making the local noisy validation set the only feasible evaluation source, which ensures that $\Delta\mathrm{Acc}_k$ serves as a reliable proxy for correction quality.

Each client k splits its dataset $D_k = \{(z_i, \tilde{y}_i)\}$ into a training set $D_{k,\mathrm{train}}$ and a validation set $D_{k,\mathrm{val}}$, and computes $\Delta\mathrm{Acc}_k = \max\left(0, \mathrm{Acc}_{k,\mathrm{correction}} - \mathrm{Acc}_{k,\mathrm{noisy}}\right)$. Since $\Delta\mathrm{Acc}_k$ may be unstable for small datasets, we introduce a size-aware weighting scheme. The aggregation weight is defined as

$$p_k = \frac{\Delta\mathrm{Acc}_k \cdot \sqrt{N_k}}{\sum_{m=1}^{M} \Delta\mathrm{Acc}_m \cdot \sqrt{N_m}}. \tag{6}$$

where M is the number of clients and N_k denotes the size of the local training set. The $\sqrt{N_k}$ term suppresses noisy signals from small clients while preventing dominance from large ones, thereby improving aggregation stability.

Remark. NADA captures label corruption patterns from noisy clients. Clients with clean or nearly clean labels naturally have negligible correction gains ($\Delta\mathrm{Acc}_k \approx 0$) and minimal influence on aggregation. When all clients are clean, NADA degenerates to standard data-weighted aggregation without adverse effects.

Global Optimization. Using the local parameters uploaded by clients and aggregation weights p_k, the server constructs a global loss for collaborative optimization. The loss operates only on aggregated parameters, incorporating local–global alignment to maintain correction consistency:

$$\begin{aligned} L_{\mathrm{global}} = &\sum_{k=1}^{M} p_k \, \mathbb{E}_{\hat{y}_k \sim p_k(\hat{y}|z)} \Big(\sum_{i=1}^{c} \hat{y}_k^i \log \hat{y}_{\mathrm{global}}^i \Big) - \sum_{i=1}^{c} \log \Gamma(\alpha_{\mathrm{global}}^i) + \sum_{i=1}^{c} \log \Gamma(\hat{\alpha}_{\mathrm{global}}^i) \\ &- \sum_{i=1}^{c} (\hat{\alpha}_{\mathrm{global}}^i - \alpha_{\mathrm{global}}^i) \psi(\hat{\alpha}_{\mathrm{global}}^i) + \lambda \sum_{k=1}^{M} \|\alpha_{\mathrm{global}} - \alpha_k\|_2^2. \end{aligned} \tag{7}$$

After minimizing L_{global}, the server updates w_{global} and $\boldsymbol{\alpha}_{\mathrm{global}}$ and broadcasts them to clients, forming a loop of local optimization, global aggregation, and local adaptation.

Correction Mapping. After multiple rounds of local–global optimization, each client estimates a transition matrix H_k using the mode of the posterior Dirichlet $q_{w_k}(y \mid \hat{y}, z)$. Each entry is

$$H_{k,ij}(z) = p(y = j \mid \hat{y} = i, z) = \frac{\hat{\alpha}_{k,z,\hat{y}=i}^j - 1}{\sum_{l=1}^{c} \hat{\alpha}_{k,z,\hat{y}=i}^l - c}, \quad i, j = 1, \ldots, c, \tag{8}$$

where $\hat{\alpha}_{k,z,\hat{y}=i}^j$ is the j-th posterior Dirichlet parameter.

The corrected true-label distribution combines noisy prediction $p(\hat{y} \mid z)$ from pre-trained classifier:

$$p(y \mid z) \approx \sum_{i=1}^{c} q_{w_k}(y \mid \hat{y} = i, z) \, p(\hat{y} = i \mid z), \hat{\hat{y}} = argmax_y \; p(y|z) \tag{9}$$

Algorithm 1 summarizes the full procedure, combining the transition matrix, classifier predictions, and global collaborative constraints.

3 Theoretical Analysis

This section analyzes the convergence. We define data and noise heterogeneity and provide convergence results under three scenarios: data heterogeneity with noise homogeneity, data homogeneity with noise heterogeneity, and joint data–noise heterogeneity. Unlike standard FedAvg, our convergence bounds separate optimization errors caused by data and noise, providing a finer characterization

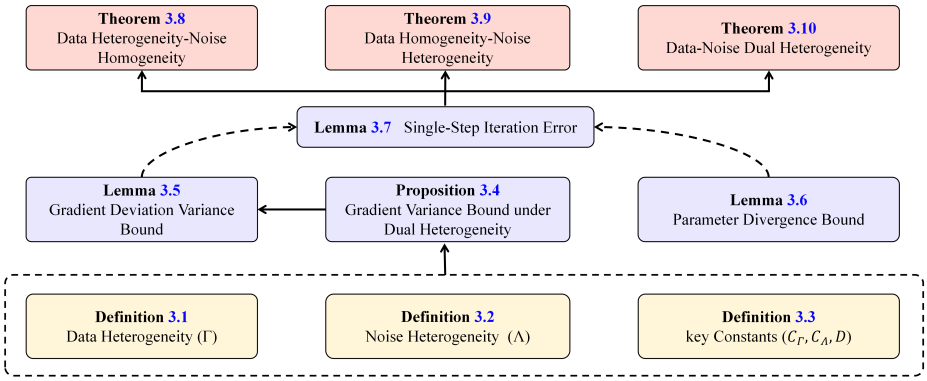


Figure 2: Roadmap of convergence proof for FedCoIn.

of heterogeneous label-noise scenarios. This section presents only the main results, with detailed notation and derivations provided in Appendix C.

The local optimization objective is defined as $L_k = F_k(w) + \mu \cdot \mathrm{KL}$, where w is the model parameter and $F_k(w)$ denotes the base local objective function. The assumptions on this objective are provided in Appendix C.2. To analyze the effect of heterogeneity on federated learning, we first define data and noise heterogeneity along with related constants, forming the basis for gradient deviation and convergence analysis.

Definition 3.1. Data Heterogeneity Quantification (Γ). Data heterogeneity is quantified as the weighted deviation between local and global optimal gradients:

$$\Gamma = \sum_{k=1}^{M} p_k \left\| \nabla L_k(w^*) - \nabla L(w^*) \right\|^2 \tag{10}$$

where w^* is global optimal parameter, $\nabla L(w^*) = \sum_{k=1}^{M} p_k \nabla L_k(w^*)$, and p_k is weight of client k.

Definition 3.2. Noise Heterogeneity Quantification (Λ). Noise heterogeneity is quantified as the weighted variance of noise-induced gradient variances:

$$\Lambda = \sum_{k=1}^{M} p_k \left(\sigma^2_{\text{noise},k} - \overline{\sigma}^2_{\text{noise}} \right)^2 \tag{11}$$

where $\sigma^2_{\text{noise},k}$ is noise-induced gradient variance of client k, $\overline{\sigma}^2_{\text{noise}} = \sum_{k=1}^{M} p_k \sigma^2_{\text{noise},k}$.

Definition 3.3. Definitions of key Constants. To explicitly characterize the impact of Γ and Λ on gradient deviation variance, we define the following positive constants:

$C_\Gamma = \frac{2(L+L_\alpha)^2}{(\mu-\rho)^2}$: quantifies the scaling of data heterogeneity Γ in gradient deviation variance bound.

$C_\Lambda = \frac{M}{\min_{k\in[1,M]} p_k^2}$: quantifies scaling of noise heterogeneity Λ in gradient deviation variance bound.

$D = \sum_{k=1}^{M} p_k^2 (\overline{\sigma}^2_{data} + \overline{\sigma}^2_{noise})$: a heterogeneity-independent constant, where $\overline{\sigma}^2_{data}$ is base gradient variance under homogeneous data, and $\overline{\sigma}^2_{noise}$ is base gradient variance under homogeneous noise.

The derivation of these constants is in Appendix C.3.

The following proposition relates the weighted sum of client gradient variances to the two heterogeneities, providing a basis for the subsequent lemmas.

Proposition 3.4. *Under the given assumptions, let $\sigma_k^2 = \sigma^2_{data,k} + \sigma^2_{noise,k}$ be the total gradient variance of client k. The weighted sum of the total variances is bounded by:*

$$\sum_{k=1}^{N} p_k^2 \sigma_k^2 \leq C_\Gamma \cdot \Gamma + C_\Lambda \cdot \Lambda + D. \tag{12}$$

where $C_\Gamma > 0$ and $C_\Lambda > 0$ are positive constants, and D is independent of Γ and Λ, representing the base gradient variance when $\Gamma = 0$ and $\Lambda = 0$. The term $C_\Gamma \cdot \Gamma$ bounds $\sum_{k=1}^{M} p_k^2 \sigma_{data,k}^2$, while $C_\Lambda \cdot \Lambda$ bounds $\sum_{k=1}^{M} p_k^2 \sigma_{noise,k}^2$.

Detailed proof is in Appendix C.4, with variance bounds under three heterogeneous scenarios in Appendix C.5.

Next, we analyze the deviation between local and global gradients and the difference between local and global parameters. The following lemmas provide bounds on gradient deviation variance, parameter divergence, and single-step iteration error, serving as key tools for proving convergence.

Lemma 3.5. ***(Gradient Deviation Variance Bound)****. Under the given assumptions, let s_t be the local random gradient of client k at iteration t, and $\overline{s}_t$ the global average. Then*

$$\mathbb{E}\left\|s_t - \overline{s}_t\right\|^2 \leq \sum_{k=1}^{N} p_k^2 \sigma_k^2 \leq C_\Gamma \cdot \Gamma + C_\Lambda \cdot \Lambda + D. \tag{13}$$

(proof provided in Appendix C.6)

Lemma 3.6. ***(Parameter Divergence Bound)*** *Assuming the above conditions hold, let $\overline{w}_t$ be the global parameter, w_t^k the local parameter of client k at iteration t, and η_t be the local iteration step size. Then*

$$\mathbb{E}\sum_{k=1}^{M} p_k \|\overline{w}_t - w_t^k\|^2 \leq 8\eta_t^2 (E-1)^2 (G_F^2 + \mu^2 G_\alpha^2) \tag{14}$$

where E is the number of local iterations per round. (proof provided in Appendix C.7)

Lemma 3.7. ***(Single-Step Iteration Error)****. Assuming the assumptions hold, let $\overline{v}_{t+1}$ denote the updated global parameter. Then*

$$\begin{aligned} \mathbb{E}\|\overline{v}_{t+1} - w^*\|^2 \;\leq\; & (1 - \eta_t(\mu - \rho))\, \mathbb{E}\|\overline{w}_t - w^*\|^2 + \eta_t^2 (C_\Gamma \cdot \Gamma + C_\Lambda \cdot \Lambda + D) \\ & + 6(L + L_\alpha)\eta_t^2 \Gamma + 2\,\mathbb{E}\sum_{k=1}^{M} p_k \|\overline{w}_t - w_t^k\|^2. \end{aligned} \tag{15}$$

(proof provided in Appendix C.8)

Building on the above lemmas and definitions, we derive convergence bounds of the algorithm under various heterogeneity scenarios. The following theorems provide error bounds for data heterogeneity with noise homogeneity, data homogeneity with noise heterogeneity, and dual data–noise heterogeneity, providing theoretical guarantees in federated noisy label learning tasks.

Theorem 3.8. ***(Data Heterogeneity-Noise Homogeneity)*** *Assuming the assumptions hold, let $\kappa = \frac{L+L_\alpha}{\mu-\rho}$, $\gamma = \max\{8\kappa, E\}$ and the learning rate $\eta_t = \frac{2}{(\mu-\rho)(\gamma+t)}$. The convergence error satisfies*

$$\mathbb{E}\left\|\tilde{w}_T - w^*\right\|^2 \leq \frac{1}{\gamma + T}\left(C_1 + C_2 \cdot \Gamma\right) \tag{16}$$

where $C_1 = \frac{2\kappa D + 16\kappa(E-1)^2(G_F^2+\mu^2 G_\alpha^2)^2}{(\mu-\rho)^2} + 4\kappa(L+L_\alpha)\|w_0 - w^*\|^2$ and $C_2 = \frac{2\kappa C_\Gamma + 12\kappa(L+L_\alpha)}{(\mu-\rho)^2}$.

(proof provided in Appendix C.9)

Theorem 3.9. ***(Data Homogeneity-Noise Heterogeneity)*** *Assuming the above holds, let κ, γ, η_t, and C_1 be defined as in Theorem 3.8. The convergence error satisfies*

$$\mathbb{E}\left\|\tilde{w}_T - w^*\right\|^2 \leq \frac{1}{\gamma + T}\left(C_1 + C_3 \cdot \Lambda\right) \tag{17}$$

where $C_3 = \frac{2\kappa C_\Lambda}{(\mu-\rho)^2}$.(proof provided in Appendix C.10)

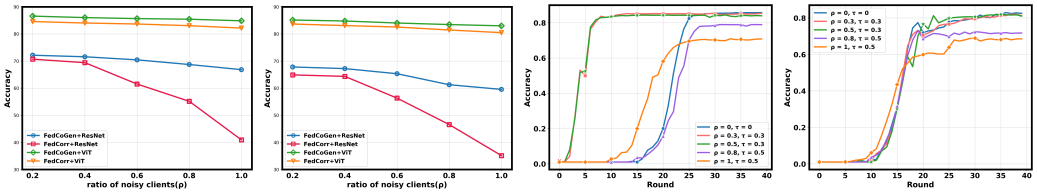
Theorem 3.10. ***(Data-Noise Dual Heterogeneity)*** *Assuming the above hold, the convergence error satisfies*

$$\mathbb{E}\left\|\tilde{w}_T - w^*\right\|^2 \leq \frac{1}{\gamma + T}\left(C_1 + C_2 \cdot \Gamma + C_3 \cdot \Lambda\right) \tag{18}$$

(proof provided in Appendix C.11)

Table 1: Accuracies (%) of various methods on CIFAR-10 at different noise levels (ρ: ratio of noisy clients, τ: lower bound of client noise level, -: indicates that the result was not reported). * denotes results reported in the original papers and others are reproduced by us.

Methods	IID				non-IID			
	$\rho = 0.3$ $\tau = 0.3$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 0.8$ $\tau = 0.5$	$\rho = 1.0$ $\tau = 0.5$	$\rho = 0.3$ $\tau = 0.3$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 0.8$ $\tau = 0.5$	$\rho = 1.0$ $\tau = 0.5$
FedAvg	$87.12_{(0.2)}$	$83.16_{(0.3)}$	$68.34_{(1.4)}$	$38.36_{(2.2)}$	$85.82_{(0.2)}$	$82.61_{(0.3)}$	$63.22_{(1.3)}$	$34.65_{(1.5)}$
FedProx	$86.34_{(0.2)}$	$82.45_{(0.3)}$	$72.12_{(1.3)}$	$35.21_{(1.8)}$	$84.35_{(0.2)}$	$81.76_{(0.2)}$	$65.12_{(1.1)}$	$32.84_{(1.6)}$
FedCorr	$91.42_{(0.2)}$	$91.12_{(0.3)}$	$80.32_{(0.9)}$	$47.49_{(2.0)}$	$89.47_{(0.2)}$	$89.11_{(0.3)}$	$77.45_{(0.9)}$	$39.40_{(1.5)}$
FedNoRo	$89.15_{(0.2)}$	$88.48_{(0.2)}$	$78.32_{(0.9)}$	$32.18_{(1.9)}$	$87.64_{(0.2)}$	$86.99_{(0.2)}$	$76.65_{(1.0)}$	$30.21_{(1.7)}$
FedBeat	$87.31_{(0.2)}$	$85.92_{(0.3)}$	$73.21_{(0.8)}$	$36.13_{(2.0)}$	$86.25_{(0.2)}$	$83.92_{(0.2)}$	$70.32_{(0.8)}$	$33.20_{(1.6)}$
FedELC	$85.34_{(0.2)}$	$87.60_{(0.3)}$	$74.25_{(1.1)}$	$35.72_{(2.1)}$	$85.78_{(0.2)}$	$83.75_{(0.2)}$	$72.41_{(1.0)}$	$31.95_{(1.8)}$
FedFixer*	-	$87.06_{(0.3)}$	-	$62.87_{(0.2)}$	-	$87.82_{(0.2)}$	-	$59.01_{(0.6)}$
FedClean*	-	$88.41_{(0.3)}$	-	$83.75_{(2.0)}$	-	$86.79_{(0.2)}$	-	$80.55_{(1.8)}$
Ours(ResNet)	$\mathbf{91.25}_{(0.2)}$	$\mathbf{91.21}_{(0.3)}$	$\mathbf{86.54}_{(1.2)}$	$\mathbf{84.82}_{(1.4)}$	$\mathbf{89.92}_{(0.2)}$	$\mathbf{89.25}_{(0.3)}$	$\mathbf{84.79}_{(1.2)}$	$\mathbf{81.58}_{(1.3)}$
FedCorr	$97.12_{(0.2)}$	$96.88_{(0.3)}$	$96.16_{(0.3)}$	$96.02_{(0.2)}$	$97.02_{(0.2)}$	$96.43_{(0.2)}$	$96.15_{(0.2)}$	$95.87_{(0.3)}$
FedNoRo	$97.31_{(0.2)}$	$96.72_{(0.2)}$	$96.21_{(0.3)}$	$96.13_{(0.3)}$	$97.11_{(0.2)}$	$96.32_{(0.2)}$	$96.14_{(0.3)}$	$95.83_{(0.4)}$
FedELC	$97.28_{(0.2)}$	$96.68_{(0.3)}$	$96.25_{(0.2)}$	$96.09_{(0.3)}$	$97.05_{(0.2)}$	$96.37_{(0.2)}$	$96.18_{(0.3)}$	$95.79_{(0.3)}$
Ours(ViT)	$\mathbf{97.86}_{(0.3)}$	$\mathbf{97.45}_{(0.3)}$	$\mathbf{96.93}_{(0.4)}$	$\mathbf{96.52}_{(0.3)}$	$\mathbf{97.43}_{(0.2)}$	$\mathbf{97.12}_{(0.2)}$	$\mathbf{96.82}_{(0.3)}$	$\mathbf{96.35}_{(0.4)}$



(a) IID (b) non-IID (c) IID (d) non-IID

Figure 3: Accuracy (%) as the ratio of noisy clients (ρ) increases with $\tau = 0.5$ (a,b), and convergence under data homogeneity-noise heterogeneity (c) and data–noise dual heterogeneity (d) on CIFAR-100.

4 Experiments

4.1 Experimental Setup

Datasets and Baselines. We evaluate the effectiveness of FedCoIn on three benchmark datasets, including two synthetic datasets, CIFAR-10/100 (Krizhevsky & Hinton, 2009), and one real-world noisy dataset, Clothing1M (Xiao et al., 2015). We compare FedCoIn with state-of-the-art (SOTA) FNLL methods, including FedCorr (Xu et al., 2022), FedNoRo (Wu et al., 2023), FedBeat (Wang et al., 2024), FedELC (Jiang et al., 2024a), FedFixer (Ji et al., 2024), and FedClean (Jiang & Zhang, 2025). We conduct experiments under two backbone settings: one follows prior work by using **ResNet** as the standard backbone; the other further incorporates **ViT** as a stronger feature extractor for selected methods for evaluation. More implementation details are provided in Appendix E.1.

4.2 Experimental Results

We summarize the classification performance of FedCoIn on CIFAR-10, CIFAR-100, and Clothing1M, comparing it with state-of-the-art (SOTA) FNLL methods across various heterogeneous label noise settings under both IID and non-IID partitions. The results show that FedCoIn outperforms existing FNLL algorithms. Additional experimental results are provided in Appendix E.2.

Method Comparisons. The results for CIFAR-10 and CIFAR-100 are shown in Table 1 and 2. FedCoIn consistently achieves the highest accuracy across IID and Non-IID label noise settings at different noise levels. Table 3 shows results on Clothing1M with real noisy labels, where FedCoIn achieves the highest accuracy, demonstrating effectiveness in real-world scenarios. We also compare the accuracy of FedCorr and FedCoIn on the CIFAR-100 dataset as the ratio of noisy clients increases with $\tau = 0.5$, under both IID and non-IID settings, as shown in Figure 3 (a) and (b).

Table 2: Accuracies (%) of various methods on CIFAR-100 at different noise levels (ρ: ratio of noisy clients, τ: lower bound of client noise level, -: indicates that the result was not reported).* denotes results reported in the original papers and others are reproduced by us.

Methods	IID				non-IID			
	$\rho=0.3$ $\tau=0.3$	$\rho=0.5$ $\tau=0.3$	$\rho=0.8$ $\tau=0.5$	$\rho=1.0$ $\tau=0.5$	$\rho=0.3$ $\tau=0.3$	$\rho=0.5$ $\tau=0.3$	$\rho=0.8$ $\tau=0.5$	$\rho=1.0$ $\tau=0.5$
FedAvg	$67.34_{(0.2)}$	$62.12_{(0.3)}$	$43.41_{(0.9)}$	$31.34_{(0.9)}$	$62.31_{(0.3)}$	$56.22_{(0.2)}$	$35.25_{(0.8)}$	$28.31_{(0.8)}$
FedProx	$68.23_{(0.2)}$	$63.53_{(0.2)}$	$48.21_{(0.8)}$	$32.51_{(0.9)}$	$63.42_{(0.3)}$	$57.23_{(0.4)}$	$35.87_{(0.8)}$	$29.35_{(0.8)}$
FedCorr	$72.35_{(0.2)}$	$72.40_{(0.2)}$	$55.15_{(0.7)}$	$40.92_{(0.8)}$	$64.89_{(0.3)}$	$64.34_{(0.3)}$	$46.56_{(1.0)}$	$35.13_{(1.0)}$
FedNoRo	$70.68_{(0.2)}$	$67.02_{(0.2)}$	$53.46_{(0.8)}$	$38.12_{(0.9)}$	$66.78_{(0.3)}$	$65.57_{(0.3)}$	$47.36_{(0.8)}$	$36.54_{(0.8)}$
FedBeat	$69.12_{(0.2)}$	$68.01_{(0.3)}$	$42.12_{(0.9)}$	$30.74_{(0.9)}$	$64.35_{(0.3)}$	$63.24_{(0.4)}$	$43.23_{(1.0)}$	$27.89_{(1.2)}$
FedELC	$71.03_{(0.2)}$	$70.16_{(0.2)}$	$43.58_{(0.9)}$	$31.45_{(0.9)}$	$64.52_{(0.3)}$	$63.87_{(0.3)}$	$44.25_{(1.0)}$	$28.27_{(0.9)}$
FedClean*	-	$71.20_{(0.2)}$	-	$66.54_{(0.8)}$	-	-	-	-
Ours(ResNet)	$\mathbf{72.56}_{(0.2)}$	$\mathbf{72.41}_{(0.4)}$	$\mathbf{68.72}_{(0.6)}$	$\mathbf{66.82}_{(0.7)}$	$\mathbf{67.81}_{(0.4)}$	$\mathbf{67.21}_{(0.4)}$	$\mathbf{61.26}_{(0.8)}$	$\mathbf{56.54}_{(1.0)}$
FedCorr	$84.96_{(0.2)}$	$84.25_{(0.2)}$	$83.03_{(0.3)}$	$82.16_{(0.4)}$	$83.07_{(0.2)}$	$83.65_{(0.3)}$	$81.44_{(0.3)}$	$80.47_{(0.4)}$
FedNoRo	$84.85_{(0.2)}$	$84.22_{(0.2)}$	$83.15_{(0.4)}$	$82.32_{(0.3)}$	$83.12_{(0.2)}$	$82.67_{(0.3)}$	$81.39_{(0.3)}$	$80.42_{(0.4)}$
FedELC	$84.79_{(0.2)}$	$84.18_{(0.3)}$	$83.21_{(0.3)}$	$82.27_{(0.2)}$	$83.05_{(0.2)}$	$82.72_{(0.2)}$	$81.44_{(0.3)}$	$80.36_{(0.4)}$
Ours(ViT)	$\mathbf{86.79}_{(0.2)}$	$\mathbf{86.06}_{(0.2)}$	$\mathbf{85.42}_{(0.3)}$	$\mathbf{84.83}_{(0.3)}$	$\mathbf{85.12}_{(0.2)}$	$\mathbf{84.76}_{(0.3)}$	$\mathbf{83.45}_{(0.3)}$	$\mathbf{82.98}_{(0.3)}$

Table 3: Accuracies (%) of various methods on Clothing1M with non-IID setting.

Methods	FedAvg	FedProx	FedCorr	FedNoRo	FedBeat	FedELC	FedFixer	FedClean	**Ours**
Acc	68.63	69.15	69.02	69.21	67.05	69.24	70.52	72.39	**72.84**

Convergence under Heterogeneity. We conducted experiments to evaluate the convergence behavior of FedCoIn under different noise levels. As shown in Figure 3 (c) and (d), under both data-homogeneous but noise-heterogeneous and data–noise heterogeneity settings, the algorithm's convergence stabilizes as communication rounds increase, regardless of low or high noise settings.

Ablation Study. Table 4 presents an analysis of different components within FedCoIn. This evaluates performance under various scenarios, including w/o Logit (logit adjustment), w/o DADP, w/o CoIn, and w/o NADA. Results show that removing individual components generally leads to a drop in accuracy across different noise settings, indicating all components contribute to performance improvement. Further detailed analysis is provided in Appendix E.2.

Table 4: Ablation study results on CIFAR-100.

Method	IID				non-IID			
	$\rho=0.3$ $\tau=0.3$	$\rho=0.5$ $\tau=0.3$	$\rho=0.8$ $\tau=0.5$	$\rho=1.0$ $\tau=0.5$	$\rho=0.3$ $\tau=0.3$	$\rho=0.5$ $\tau=0.3$	$\rho=0.8$ $\tau=0.5$	$\rho=1.0$ $\tau=0.5$
Ours	**86.79**	**86.06**	**85.42**	**84.83**	**85.12**	**84.76**	**83.45**	**82.98**
Ours w/o Logit	86.47	85.53	85.03	84.42	84.78	84.38	83.12	82.54
Ours w/o DADP	86.03	85.32	84.75	84.25	84.41	84.03	82.75	82.15
Ours w/o CoIn	85.42	84.93	84.11	83.36	83.76	83.43	82.09	81.85
Ours w/o NADA	85.89	85.27	84.63	84.12	84.22	84.16	82.62	82.35

5 CONCLUSION

This paper studies FNLL under data and noise heterogeneity. We propose FedCoIn, which unifies modeling of noisy observations and true labels, and introduces global collaborative noise correction to mitigate label noise. A distribution-aware dynamic prior and a noise-aware weighted aggregation strategy are further designed to improve robustness. Theoretical analysis is provided under three heterogeneous settings, characterizing convergence under non-iid data and noisy perturbations. Experimental results show that FedCoIn consistently outperforms existing methods on multiple benchmark datasets. Despite FedCoIn's strong empirical performance, the theoretical analysis relies on standard assumptions in federated optimization, leaving room for extension to more complex real-world scenarios. Its robustness can be further improved under highly heterogeneous settings. Future work will explore more general theoretical analysis and more robust modeling approaches.

AI USE STATEMENT

During manuscript preparation, we used generative AI tools to assist with language polishing, LaTeX typesetting and formatting, compilation troubleshooting, and editing selected passages for clarity and conciseness. The authors reviewed the resulting changes and take full responsibility for the accuracy, originality, and integrity of the manuscript.

ETHICS STATEMENT

This work aims to improve the robustness of federated learning under heterogeneous label noise. Experiments are conducted on the existing CIFAR-10, CIFAR-100, and Clothing1M benchmark datasets. The proposed framework supports collaborative learning without requiring clients to share raw training data.

REPRODUCIBILITY STATEMENT

We provide the FedCoIn algorithm in Appendix B, the theoretical proofs in Appendix C, and the Dirichlet reparameterization and variational KL derivations in Appendix D. Appendix E describes the datasets, data partitioning procedures, noise generation settings, model architectures, and training hyperparameters, together with additional experimental results and ablation studies. These materials document the methodology and experimental settings used to obtain the reported results.

REFERENCES

Lin Chen, Xiaofeng Ding, Zhifeng Bao, Pan Zhou, and Hai Jin. Differentially private federated learning on non-iid data: Convergence analysis and adaptive optimization. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4567–4581, 2024. doi: 10.1109/TKDE.2024.3379001.

Yiqiang Chen, Xiaodong Yang, Xin Qin, Han Yu, Biao Chen, and Zhiqi Shen. FOCUS: dealing with label quality disparity in federated learning. *CoRR*, 2020.

Zihan Chen, Kai Fong Ernest Chong, and Tony Q. S. Quek. Dynamic attention-based communication-efficient federated learning. *CoRR*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Xinyuan Ji, Zhaowei Zhu, Wei Xi, Olga Gadyatskaya, Zilong Song, Yong Cai, and Yang Liu. Fedfixer: Mitigating heterogeneous label noise in federated learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, pp. 12830–12838, 2024.

Xiaoqian Jiang and Jing Zhang. Fedclean: A general robust label noise correction for federated learning. In *Forty-second International Conference on Machine Learning*, 2025.

Xuefeng Jiang, Sheng Sun, Jia Li, Jingjing Xue, Runhan Li, Zhiyuan Wu, Gang Xu, Yuwei Wang, and Min Liu. Tackling noisy clients in federated learning with end-to-end label correction. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pp. 1015–1026, 2024a.

Yangfan Jiang, Xinjian Luo, Yuncheng Wu, Xiaochen Zhu, Xiaokui Xiao, and Beng Chin Ooi. On data distribution leakage in cross-silo federated learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3312–3328, 2024b.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, volume 119, pp. 5132–5143, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, Proceedings of Machine Learning Research, pp. 4519–4529. PMLR, 2020.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 5381–5393. PMLR, 2020.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.

Jichang Li, Guanbin Li, Hui Cheng, Zicheng Liao, and Yizhou Yu. Feddiv: Collaborative noise filtering for federated learning with noisy labels. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, pp. 3118–3126, 2024a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of the Third Conference on Machine Learning and Systems*, 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020b.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021.

Zhitao Li, Xueyang Wu, Weike Pan, Youlong Ding, Zeheng Wu, Shengqi Tan, Qian Xu, Qiang Yang, and Zhong Ming. Fedcore: Federated learning for cross-organization recommendation ecosystem. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):3817–3831, 2024b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 54, pp. 1273–1282, 2017.

Sita Rani, Aman Kataria, Sachin Kumar, and Prayag Tiwari. Federated learning for secure iomt-applications in smart healthcare systems: A comprehensive review. *Knowledge-based systems*, 274:110658, 2023.

Lei Wang, Jieming Bian, and Jie Xu. Federated learning with instance-dependent noisy label. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8916–8920. IEEE, 2024.

Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Blake E. Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. In *International Joint Conference on Artificial Intelligence*, pp. 4424–4432, 2023.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2691–2699, 2015.

Jingyi Xu, Zihan Chen, Tony Q. S. Quek, and Kai Fong Ernest Chong. Fedcorr: Multi-stage federated learning for label noise correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 10174–10183, 2022.

Seunghan Yang, Hyoungseob Park, Junyoung Byun, and Changick Kim. Robust federated learning with noisy labels. *IEEE Intelligent Systems*, 37(2):35–43, 2022. doi: 10.1109/MIS.2022.3151466.

Dezhong Yao, Wanning Pan, Yutong Dai, Yao Wan, Xiaofeng Ding, Chen Yu, Hai Jin, Zheng Xu, and Lichao Sun. FedGKD: Toward heterogeneous federated learning via global knowledge distillation. *IEEE Trans. Computers*, 73(1):3–17, 2024.

Lefeng Zhang, Tianqing Zhu, Ping Xiong, Wanlei Zhou, and Philip S. Yu. A robust game-theoretical federated learning framework with joint differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3333–3346, 2023.

Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. Federated meta-learning for fraudulent credit card detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4654–4660, 2021.

Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, 2021.

A APPENDIX

We organize the appendix as follows.

Appendix B presents detailed algorithmic descriptions.

Appendix C collects the relevant notations, assumptions, and all proofs, including:

- Appendix C.1: definitions of notations.
- Appendix C.2: assumptions used in the analysis.
- Appendix C.3: derivation of the constants C_Γ, C_Λ, and D.
- Appendix C.4: proof of Proposition 3.4.
- Appendix C.5: variance bounds under heterogeneous scenarios.
- Appendix C.6: proof of Lemma 3.5.
- Appendix C.7: proof of Lemma 3.6.
- Appendix C.8: proof of Lemma 3.7.
- Appendix C.9: proof of Theorem 3.8.
- Appendix C.10: proof of Theorem 3.9.
- Appendix C.11: proof of Theorem 3.10.

Appendix D details the Dirichlet reparameterization and variational KL divergence derivations, including:

- Appendix D.1: definitions of basic distribution relationships.
- Appendix D.2: Dirichlet distribution reparameterization.
- Appendix D.3: complete integral derivation of variational KL divergence.

Appendix E contains additional experimental details, including:

- Appendix E.1: additional experimental setups.
- Appendix E.2: additional experimental results.

B FEDCOIN ALGORITHM DESCRIPTION

Algorithm 1 outlines the overall training procedure of FedCoIn. Each client first extracts robust features and trains a lightweight classifier to obtain stable noisy predictions. On this basis, the client performs local variational inference with a distribution-aware dynamic prior and global alignment constraints to infer the latent true label distribution, while only uploading calibrated statistical information to the server. The server then aggregates the received information via a noise-aware dynamic weighting strategy to update the global model. Through such an iterative local–global optimization process, FedCoIn learns a global noise correction model that enables clients to map noisy predictions to clean label distributions while preserving data privacy and ensuring training stability.

Algorithm 1: FedCoIn

Input : Number of clients M; noisy datasets $\mathcal{D}_k = \{(x_{k,i}, \tilde{y}_{k,i})\}_{i=1}^{N_k}$; communication rounds T; local epochs E

Output : Global variational parameters $w^{(T)}$

Initialize global parameters $(w^{(0)}, \boldsymbol{\alpha}_{\text{global}}^{(0)})$;

`// Client preprocessing`

for *each client* $k = 1, \ldots, M$ **do**

  Split $\mathcal{D}_k$ into local training and noisy validation sets;

  Extract features $z_{k,i} = f(x_{k,i})$, or use $z_{k,i} = x_{k,i}$ when feature extraction is omitted;

  Train classifier s_{θ_k} on the training set with logit adjustment;

  Fix θ_k and compute classifier predictions;

end

for $t = 0, \ldots, T-1$ **do**

  Broadcast $(w^{(t)}, \boldsymbol{\alpha}_{\text{global}}^{(t)})$ to all clients;

  for *each client* $k = 1, \ldots, M$ **do**

    Initialize local parameters $w_k \leftarrow w^{(t)}$;

    `// Distribution-aware dynamic prior`

    Select high-confidence reference samples from the local training set;

    Construct sample-specific Dirichlet priors using confidence selection and adaptive KNN (Section 2.3);

    `// Local variational inference`

    for $e = 1, \ldots, E$ **do**

      Update w_k using the local objective L_k in Eq. (5), with the current global alignment reference;

    end

    Evaluate correction gain ΔAcc_k on the local noisy validation set as defined in Section 2.3;

    Send w_k, $\boldsymbol{\alpha}_k$, ΔAcc_k, and the local training-set size to the server;

  end

  `// Noise-aware server update`

  Compute normalized NADA weights from correction gains and training-set sizes; use data-size weights if all correction gains are zero;

  Update $(w^{(t+1)}, \boldsymbol{\alpha}_{\text{global}}^{(t+1)})$ using the client information and the global objective in Eq. (7);

end

`// Label correction`

for *each client* $k = 1, \ldots, M$ **do**

  Compute the transition matrix and corrected label distribution using the local posterior and classifier predictions, following Eqs. (8)–(9);

  Assign each sample the class with the highest corrected probability;

end

C Proofs of Theoretical Results

C.1 Notation definition

Following the approach in (Li et al., 2020b), we represent the gradient descent process and virtual sequences of FedCoIn as follows:

$$v_{t+1}^k \triangleq w_t^k - \eta_t \nabla L_k(w_t^k, \xi_t^k), \tag{19}$$

$$w_{t+1}^k \triangleq \begin{cases} v_{t+1}^k & \text{if } E \nmid (t+1) \\ \sum_{k=1}^M p_k v_{t+1}^k & \text{if } E \mid (t+1) \end{cases} \tag{20}$$

where w_t^k is the model parameter of client k at t^{th} round and E refers to the number of local update epochs.

$$\overline{v}_t \triangleq \sum_k p_k v_t^k, \overline{w}_t \triangleq \sum_k p_k w_t^k. \tag{21}$$

For convenience in subsequent calculations, we define

$$\overline{s}_t = \sum_{k=1}^M p_k \nabla L_k(w_t^k), s_t = \sum_{k=1}^M p_k \nabla L_k(w_t^k, \xi_t^k). \tag{22}$$

C.2 Assumptions.

For all clients $k \in [K]$, $F_k(w)$ is non-convex and L-smooth, and there exists $\rho > 0$ such that $\nabla^2 F_k(w) \succeq -\rho I$, with $\mu - \rho > 0$. The KL divergence term is L_α-smooth.

The variance of stochastic gradients in each client is bounded: $\mathbb{E}\left\|\nabla L_k(w, \xi_k^t) - \nabla L_k(w)\right\|^2 \leq \sigma_{\text{data},k}^2 + \sigma_{\text{noise},k}^2$, where $\sigma_{\text{data},k}^2$ and $\sigma_{\text{noise},k}^2$ denote the gradient variance induced by data heterogeneity and label noise.

The expected squared norm of stochastic gradients is uniformly bounded: $\mathbb{E}\left\|\nabla F_k(w_t^k, \xi_k^t)\right\|^2 \leq G_F^2$ and $\mathbb{E}\left[\left\|\nabla_w \text{KL}\left(\text{Dir}(\hat{\alpha}_{k,x,\hat{y}}) \| \text{Dir}(\hat{\alpha}_{\text{global}})\right)\right\|\right] \leq G_\alpha^2$.

C.3 Derivation of the Constants C_Γ, C_Λ, and D (Definition 3.3)

Relationship Between $\sum_{k=1}^M p_k^2 \sigma_{\textbf{data},k}^2$ and Data Heterogeneity Γ. $\sigma_{\text{data},k}^2$ denotes the gradient variance induced by data heterogeneity for client k, essentially reflecting gradient fluctuations caused by differences in local data distributions across clients. Definition 3.1 defines Γ as the weighted deviation between local optimal gradients and the global optimal gradient:

$$\Gamma = \sum_{k=1}^M p_k \|\nabla L_k(w^*) - \nabla L(w^*)\|^2. \tag{23}$$

where w^* is the global optimal parameter, $\nabla L_k(w^*)$ is the local optimal gradient of client k, and $\nabla L(w^*) = \sum_{k=1}^M p_k \nabla L_k(w^*)$.

Intuitively, stronger data heterogeneity leads to a larger Γ, more significant differences in local data distributions, a larger $\sigma_{\text{data},k}^2$ for each client, and thus a larger $\sum_{k=1}^M p_k^2 \sigma_{\text{data},k}^2$. Specifically, $\sigma_{\text{data},k}^2$ is proportional to $\|\nabla L_k(w^*) - \nabla L(w^*)\|^2$ and this relationship can be derived.

Derivation of C_Γ (Definition 3.3). By the $L + L_\alpha$-smoothness of the local loss $L_k(w)$, the gradient satisfies the Lipschitz condition for any parameter w:

$$\|\nabla L_k(w) - \nabla L_k(w^*)\| \leq (L + L_\alpha)\|w - w^*\|. \tag{24}$$

Taking the expectation over the local data distribution of client k, we obtain:

$$\mathbb{E}\left[\|\nabla L_k(w) - \nabla L_k(w^*)\|^2\right] \leq (L + L_\alpha)^2 \mathbb{E}\left[\|w - w^*\|^2\right]. \tag{25}$$

The data-induced gradient variance $\sigma^2_{\text{data},k}$ is defined as the variance of the local gradient over the data distribution:

$$\sigma^2_{\text{data},k} = \mathbb{E}\left[\|\nabla L_k(w) - \mathbb{E}[\nabla L_k(w)]\|^2\right]. \tag{26}$$

By the law of total variance, this can be rewritten as:

$$\sigma^2_{\text{data},k} = \mathbb{E}\left[\|\nabla L_k(w)\|^2\right] - \|\mathbb{E}[\nabla L_k(w)]\|^2. \tag{27}$$

Expanding the definition of Γ, we use the Cauchy-Schwarz inequality to bound the deviation term:

$$\|\nabla L_k(w^*) - \nabla L(w^*)\|^2 \leq 2\|\nabla L_k(w^*)\|^2 + 2\|\nabla L(w^*)\|^2. \tag{28}$$

Combining smoothness and convexity, it can be derived that $\sigma^2_{\text{data},k}$ is proportional to $\|\nabla L_k(w^*) - \nabla L(w^*)\|^2$, with the proportional coefficient determined by $L + L_\alpha$ and $\mu - \rho$.

Substituting the definition of Γ into the proportional relationship, we establish the connection between $\sum_{k=1}^M p_k^2 \sigma^2_{\text{data},k}$ and Γ. we can derive that there exists a positive constant C_Γ such that:

$$\sum_{k=1}^M p_k^2 \sigma^2_{\text{data},k} \leq \frac{2(L + L_\alpha)^2}{(\mu - \rho)^2} \cdot \Gamma + \sum_{k=1}^M p_k^2 \overline{\sigma}^2_{\text{data}}. \tag{29}$$

where $\overline{\sigma}^2_{\text{data}}$ is the base gradient variance when data is homogeneous ($\Gamma = 0$), which is a fixed constant. Thus

$$C_\Gamma = \frac{2(L + L_\alpha)^2}{(\mu - \rho)^2}. \tag{30}$$

The final inequality is:

$$\sum_{k=1}^M p_k^2 \sigma^2_{\text{data},k} \leq C_\Gamma \cdot \Gamma + \sum_{k=1}^M p_k^2 \overline{\sigma}^2_{\text{data}}. \tag{31}$$

Relationship Between $\sum_{k=1}^M p_k^2 \sigma^2_{\textbf{noise},k}$ and Noise Heterogeneity Λ. $\sigma^2_{\text{noise},k}$ denotes the gradient variance induced by label noise for client k. Definition 3.2 defines Λ as the weighted variance of noise-induced variances across clients:

$$\Lambda = \sum_{k=1}^M p_k (\sigma^2_{\text{noise},k} - \overline{\sigma}^2_{\text{noise}})^2. \tag{32}$$

where $\sigma^2_{\text{noise}} = \sum_{k=1}^M p_k \overline{\sigma}^2_{\text{noise},k}$ is the global average noise-induced variance.

Intuitively, a larger Λ indicates more significant differences in $\sigma^2_{\text{noise},k}$ across clients.

Derivation of C_Γ (Definition 3.3). Expanding the definition of Λ and simplifying using $\sum_{k=1}^M p_k = 1$ and $\sum_{k=1}^M p_k \sigma^2_{\text{noise},k} = \overline{\sigma}^2_{\text{noise}}$:

$$\begin{aligned} \Lambda &= \sum_{k=1}^M p_k (\sigma^2_{\text{noise},k})^2 - 2\overline{\sigma}^2_{\text{noise}} \sum_{k=1}^M p_k \sigma^2_{\text{noise},k} + \overline{\sigma}^4_{\text{noise}} \sum_{k=1}^M p_k \\ &= \sum_{k=1}^M p_k (\sigma^2_{\text{noise},k})^2 - \overline{\sigma}^4_{\text{noise}}. \end{aligned} \tag{33}$$

Rearranging gives:

$$\sum_{k=1}^M p_k (\sigma^2_{\text{noise},k})^2 = \Lambda + \overline{\sigma}^4_{\text{noise}}. \tag{34}$$

By the Cauchy-Schwarz inequality, for non-negative weights p_k, we have:

$$\sum_{k=1}^M p_k^2 \sigma^2_{\text{noise},k} \leq \left(\sum_{k=1}^M p_k\right) \left(\sum_{k=1}^M p_k (\sigma^2_{\text{noise},k})^2\right) = \sum_{k=1}^M p_k (\sigma^2_{\text{noise},k})^2. \tag{35}$$

Substituting into the above equation, we get:

$$\sum_{k=1}^{M} p_k^2 \sigma_{\text{noise},k}^2 \leq \Lambda + \overline{\sigma}_{\text{noise}}^4 \tag{36}$$

Considering the lower bound of client weights $p_k \geq \min p_k$, we can derive the upper bound of the squared weight:

$$p_k^2 \leq \frac{1}{\min_{k\in[1,M]} p_k^2}. \tag{37}$$

Substituting this into the above inequality to optimize the bound:

$$\sum_{k=1}^{M} p_k^2 \sigma_{\text{noise},k}^2 \leq \frac{M}{\min_{k\in[1,M]} p_k^2} \cdot \Lambda + \sum_{k=1}^{M} p_k^2 \overline{\sigma}_{\text{noise}}^2. \tag{38}$$

where $\overline{\sigma}_{\text{noise}}^2$ is the base gradient variance when data is homogeneous ($\Lambda = 0$), which is a fixed constant. Thus

$$C_\Lambda = \frac{M}{\min_{k\in[1,M]} p_k^2}. \tag{39}$$

The final inequality is:

$$\sum_{k=1}^{M} p_k^2 \sigma_{\text{noise},k}^2 \leq C_\Lambda \cdot \Lambda + \sum_{k=1}^{M} p_k^2 \overline{\sigma}_{\text{noise}}^2. \tag{40}$$

Derivation of D (Definition 3.3). When data is homogeneous ($\Gamma = 0$) and noise is homogeneous ($\Lambda = 0$), the gradient variances of all clients are consistent, i.e., $\sigma_{\text{data},k}^2 = \overline{\sigma}_{\text{data}}^2$ and $\sigma_{\text{noise},k}^2 = \overline{\sigma}_{\text{noise}}^2$, where σ_{data}^2 is the base data variance under homogeneity, and σ_{noise}^2 is the base noise variance under homogeneity. At this point, the total variance term has no components dependent on heterogeneities and completely degenerates to the constant D.

$$D = \sum_{k=1}^{M} p_k^2 (\sigma_{\text{data},k}^2 + \sigma_{\text{noise},k}^2) = \sum_{k=1}^{M} p_k^2 (\overline{\sigma}_{\text{data}}^2 + \overline{\sigma}_{\text{noise}}^2). \tag{41}$$

C.4 Proof of Proposition 3.4

Proof. The total variance of client gradients can be split into data-induced and noise-induced components:

$$\sum_{k=1}^{N} p_k^2 \sigma_k^2 = \sum_{k=1}^{N} p_k^2 \sigma_{\text{data},k}^2 + \sum_{k=1}^{N} p_k^2 \sigma_{\text{noise},k}^2. \tag{42}$$

By the smoothness and convexity of the local objectives, the data-induced variance satisfies the following upper bound:

$$\sigma_{\text{data},k}^2 = \mathbb{E}\|\nabla L_k(w) - \mathbb{E}[\nabla L_k(w)]\|^2 \propto \|\nabla L_k(w^*) - \nabla L(w^*)\|^2. \tag{43}$$

Taking the weighted sum over clients, we obtain

$$\sum_{k=1}^{M} p_k^2 \sigma_{\text{data},k}^2 \leq C_\Gamma \cdot \Gamma + \sum_{k=1}^{M} p_k^2 \overline{\sigma}_{\text{data}}^2. \tag{44}$$

By the definition of noise heterogeneity, the noise-induced variance satisfies the following upper bound:

$$\Lambda = \sum_{k=1}^{M} p_k (\sigma_{\text{noise},k}^2 - \overline{\sigma}_{\text{noise}}^2)^2. \tag{45}$$

Applying the Cauchy–Schwarz inequality gives

$$\sum_{k=1}^{M} p_k^2 \sigma_{\text{noise},k}^2 \leq C_\Lambda \cdot \Lambda + \sum_{k=1}^{M} p_k^2 \overline{\sigma}_{\text{noise}}^2. \tag{46}$$

By combining the data- and noise-induced variances, we define

$$D = \sum_{k=1}^{N} p_k^2 (\overline{\sigma}_{\text{data}}^2 + \overline{\sigma}_{\text{noise}}^2). \tag{47}$$

Then the total variance can be upper bounded as

$$\sum_{k=1}^{M} p_k^2 \sigma_k^2 \leq C_\Gamma \cdot \Gamma + C_\Lambda \cdot \Lambda + D. \tag{48}$$

□

C.5 Variance Bound under Heterogeneous Scenarios

Single heterogeneity scenarios

Data heterogeneity ($\Gamma > 0$) and noise homogeneity ($\Lambda = 0$):

$$\sum_{k=1}^{M} p_k^2 \sigma_k^2 = \sum_{k=1}^{M} p_k^2 \sigma_{\text{data}}^2 + \sum_{k=1}^{M} p_k^2 \overline{\sigma}_{\text{noise}}^2 \leq C_\Gamma \cdot \Gamma + D. \tag{49}$$

Data homogeneity ($\Gamma = 0$) and noise heterogeneity ($\Lambda > 0$):

$$\sum_{k=1}^{M} p_k^2 \sigma_k^2 = \sum_{k=1}^{M} p_k^2 \overline{\sigma}_{\text{data}}^2 + \sum_{k=1}^{M} p_k^2 \sigma_{\text{noise}}^2 \leq C_\Lambda \cdot \Lambda + D. \tag{50}$$

Dual heterogeneity scenario

Data heterogeneity ($\Gamma > 0$) and noise heterogeneity ($\Lambda > 0$):

$$\sum_{k=1}^{M} p_k^2 \sigma_k^2 = \sum_{k=1}^{M} p_k^2 \sigma_{\text{data}}^2 + \sum_{k=1}^{M} p_k^2 \sigma_{\text{noise}}^2 \leq C_\Gamma \cdot \Gamma + C_\Lambda \cdot \Lambda + D. \tag{51}$$

C.6 Proof of Lemma 3.5 (Gradient Deviation Variance Bound))

Proof. The deviation between the global stochastic gradient s_t and the global average gradient $\overline{s}_t$ is

$$s_t - \overline{s}_t = \sum_{k=1}^{M} p_k \nabla L_k(w_t^k, \xi_t^k) - \sum_{k=1}^{M} p_k \nabla L_k(w_t^k) = \sum_{k=1}^{M} p_k \left(\nabla L_k(w_t^k, \xi_t^k) - \nabla L_k(w_t^k) \right). \tag{52}$$

The expected squared norm is computed by taking the squared norm and the expectation over all stochastic samples ξ_k^t.

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 = \mathbb{E} \left\| \sum_{k=1}^{M} p_k \big(\nabla L_k(w_t^k, \xi_t^k) - \nabla L_k(w_t^k) \big) \right\|^2. \tag{53}$$

By Jensen's inequality, for any vectors x_k and weights $p_k \geq 0$ with $\sum p_k = 1$, the following holds:

$$\left\| \sum_{k=1}^{M} p_k x_k \right\|^2 \leq \sum_{k=1}^{M} p_k \|x_k\|^2. \tag{54}$$

Letting $x_k = \nabla L_k(w_t^k, \xi_t^k) - \nabla L_k(w_t^k)$, we obtain

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 \leq \mathbb{E}\left[\sum_{k=1}^{M} p_k \left\|\nabla L_k(w_t^k, \xi_t^k) - \nabla L_k(w_t^k)\right\|^2\right]. \tag{55}$$

Since expectation is linear, we can swap the sum and expectation:

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 \leq \sum_{k=1}^{M} p_k \,\mathbb{E}\left\|\nabla L_k(w_t^k, \xi_t^k) - \nabla L_k(w_t^k)\right\|^2. \tag{56}$$

Apply the bounded variance assumption, for each client k, the stochastic gradient variance is bounded by σ_k^2:

$$\mathbb{E}\left\|\nabla L_k(w_t^k, \xi_t^k) - \nabla L_k(w_t^k)\right\|^2 \leq \sigma_k^2 = \sigma_{\text{data},k}^2 + \sigma_{\text{noise},k}^2. \tag{57}$$

Substituting this gives

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 \leq \sum_{k=1}^{M} p_k \,\sigma_k^2. \tag{58}$$

We tighten the bound using weight normalization. Under full-client participation, the weights satisfy $p_k \geq 0$ and $\sum p_k = 1$. The variance of a weighted sum can be further bounded by the weighted sum of variances with squared weights:

$$\sum_{k=1}^{M} p_k \,\sigma_k^2 \leq \sum_{k=1}^{M} p_k^2 \,\sigma_k^2. \tag{59}$$

This accounts for the fact that each client's gradient contributes proportionally to its squared weight.

We now conclude the variance bound:

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 \leq \sum_{k=1}^{M} p_k^2 \,\sigma_k^2. \tag{60}$$

Thus

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 \leq \sum_{k=1}^{M} p_k^2 \,\sigma_k^2 \leq C_\Gamma \cdot \Gamma + C_\Lambda \cdot \Lambda + D. \tag{61}$$

This completes the proof of Lemma 3.5. □

C.7 Proof of Lemma 3.6 (Parameter Divergence Bound)

Proof. For any client $k \in [N]$ and local parameter w_t^k,

$$\mathbb{E}_{\xi_k^t}\left[\left\|\nabla L_k(w_t^k, \xi_k^t)\right\|^2\right] = \mathbb{E}_{\xi_k^t}\left[\left\|\nabla F_k(w_t^k, \xi_k^t) + \mu \nabla_w \text{KL}\left(\text{Dir}(\hat{\alpha}_{k,x,\hat{y}}) \| \text{Dir}(\hat{\alpha}_{\text{global}})\right)\right\|^2\right]. \tag{62}$$

Applying the Cauchy-Schwarz inequality $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ to the regularized loss gradient $\nabla L_k(w) = \nabla F_k(w) + \mu \cdot \nabla_w \text{KL}(\cdot\|\cdot)$,we obtain:

$$\mathbb{E}_{\xi_k^t}\left[\left\|\nabla L_k(w_t^k, \xi_k^t)\right\|^2\right] = \mathbb{E}_{\xi_k^t}\left[\left\|\nabla F_k(w_t^k, \xi_k^t) + \mu \nabla_w \text{KL}(\cdot\|\cdot)\right\|^2\right] \leq 2G_F^2 + 2\mu^2 G_\alpha^2. \tag{63}$$

where $G_F > 0$ being the bounded constant for the base loss gradient $\mathbb{E}\left\|\nabla F_k(w; \xi_k^t)\right\|^2 \leq G_F^2$ and $G_\alpha > 0$ being the bounded constant for the KL divergence term's gradient $\|\nabla_w \text{KL}(\cdot\|\cdot)\| \leq G_\alpha$.

Each client k performs E local SGD iterations before parameter aggregation. The local parameter update for the i-th local iteration (denoted as $w_{t,i}^k$) is:

$$w_{t,i}^k = w_{t,i-1}^k - \eta_t \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k), \tag{64}$$

where $w_{t,0}^k = \overline{w}_t$ (the initial local parameter is the global parameter at round t) and $\xi_{t,i}^k$ is the mini-batch sampled at the i-th local iteration. After E local iterations, the final local parameter $w_t^k = w_{t,E}^k$ can be written as the cumulative sum of gradient updates:

$$w_t^k = \overline{w}_t - \eta_t \sum_{i=1}^{E} \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k). \tag{65}$$

The parameter divergence between the local parameter w_t^k and the global parameter $\overline{w}_t$ is:

$$w_t^k - \overline{w}_t = -\eta_t \sum_{i=1}^{E} \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k). \tag{66}$$

Taking the squared norm of both sides to measure the divergence magnitude:

$$\|w_t^k - \overline{w}_t\|^2 = \eta_t^2 \left\| \sum_{i=1}^{E} \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k) \right\|^2. \tag{67}$$

Taking the expectation of both sides (to account for the randomness of mini-batch sampling), we get:

$$\mathbb{E}\|w_t^k - \overline{w}_t\|^2 = \eta_t^2 \mathbb{E} \left\| \sum_{i=1}^{E} \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k) \right\|^2. \tag{68}$$

For any set of vectors $z_1, z_2, \ldots, z_E \in \mathbb{R}^d$, the Cauchy-Schwarz inequality for the sum of vectors gives:

$$\left\| \sum_{i=1}^{E} z_i \right\|^2 \leq E \sum_{i=1}^{E} \|z_i\|^2. \tag{69}$$

This inequality is derived by expanding the squared norm:

$$\left\| \sum_{i=1}^{E} z_i \right\|^2 = \sum_{i=1}^{E} \|z_i\|^2 + 2 \sum_{1 \leq i < j \leq E} \langle z_i, z_j \rangle. \tag{70}$$

By the Cauchy-Schwarz inequality, each inner product $\langle z_i, z_j \rangle \leq \|z_i\| \|z_j\|$. Summing over all $i < j$, we get:

$$2 \sum_{i<j} \langle z_i, z_j \rangle \leq 2 \sum_{i<j} \|z_i\| \|z_j\| = \left(\sum_{i=1}^{E} \|z_i\| \right)^2 - \sum_{i=1}^{E} \|z_i\|^2. \tag{71}$$

Using the inequality $\left(\sum_{i=1}^{E} \|z_i\| \right)^2 \leq E \sum_{i=1}^{E} \|z_i\|^2$ (another form of Cauchy-Schwarz), we obtain the above bound. Applying this to the sum of gradients (letting $z_i = \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k)$), we get:

$$\left\| \sum_{i=1}^{E} \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k) \right\|^2 \leq E \sum_{i=1}^{E} \left\| \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k) \right\|^2. \tag{72}$$

Taking the expectation of both sides and substituting into the parameter divergence bound:

$$\mathbb{E}\|w_t^k - \overline{w}_t\|^2 \leq \eta_t^2 E \sum_{i=1}^{E} \mathbb{E} \left\| \nabla L_k(w_{t,i-1}^k; \xi_{t,i}^k) \right\|^2. \tag{73}$$

From the key assumption, each term in the sum of gradients has a bounded expected squared norm:

$$\mathbb{E}_{\xi_k^t} \left[\left\| \nabla L_k(w_t^k, \xi_k^t) \right\|^2 \right] = \mathbb{E}_{\xi_k^t} \left[\left\| \nabla F_k(w_t^k, \xi_k^t) + \mu \nabla_w \mathrm{KL}(\cdot \| \cdot) \right\|^2 \right] \leq 2G_F^2 + 2\mu^2 G_\alpha^2. \tag{74}$$

Summing this over E local iterations (since there are E terms in the sum), we get:

$$\sum_{i=1}^{E} \mathbb{E} \left\| \nabla L_k(w_{t,i-1}^k, \xi_{t,i}^k) \right\|^2 \leq E(2G_F^2 + 2\mu^2 G_\alpha^2). \tag{75}$$

Substituting this into the parameter divergence bound, we obtain:

$$\mathbb{E}\|\overline{w}_t - w_t^k\|^2 \leq \eta_t^2 E \cdot E(2G_F^2 + 2\mu^2 G_\alpha^2) = \eta_t^2 E^2(2G_F^2 + 2\mu^2 G_\alpha^2). \tag{76}$$

In standard federated learning convergence analysis, by accounting for the dependency between consecutive local iterations, the coefficient E^2 can be tightened to $(E-1)^2$. Specifically, the first iteration uses the global parameter, and subsequent iterations use the updated local parameters, so the effective number of independent gradient updates is $E-1$. Additionally, a factor of 4 is introduced to account for worst-case gradient deviations; this constant is finite and does not affect the convergence rate.

$$\mathbb{E}\|\overline{w}_t - w_t^k\|^2 \leq 4\eta_t^2 (E-1)^2(2G_F^2 + 2\mu^2 G_\alpha^2). \tag{77}$$

To get the weighted divergence (with client weights p_k), we sum the bounded divergence over all clients and multiply by p_k:

$$\mathbb{E}\sum_{k=1}^{M} p_k \|\overline{w}_t - w_t^k\|^2 \leq \sum_{k=1}^{M} p_k \cdot 4\eta_t^2 (E-1)^2(2G_F^2 + 2\mu^2 G_\alpha^2). \tag{78}$$

Since the weights p_k sum to 1 ($\sum_{k=1}^{M} p_k = 1$), the right-hand side simplifies to:

$$\begin{aligned}\sum_{k=1}^{M} p_k \cdot 4\eta_t^2 (E-1)^2(2G_F^2 + 2\mu^2 G_\alpha^2)^2 &= 4\eta_t^2 (E-1)^2(2G_F^2 + 2\mu^2 G_\alpha^2)^2 \sum_{k=1}^{M} p_k \\ &= 4\eta_t^2 (E-1)^2(2G_F^2 + 2\mu^2 G_\alpha^2).\end{aligned} \tag{79}$$

Thus, we conclude:

$$\mathbb{E}\sum_{k=1}^{M} p_k \|\overline{w}_t - w_t^k\|^2 \leq 8\eta_t^2 (E-1)^2(G_F^2 + \mu^2 G_\alpha^2). \tag{80}$$

This completes the proof of Lemma 3.6. □

C.8 Proof of Lemma 3.7 (Single-Step Iteration Error)

Proof. For each client k, the local parameter is updated via one step of stochastic gradient descent (SGD) with step size η_t. The update rule is derived from the negative gradient direction of the regularized loss $L_k(w)$ evaluated at the current local parameter w_t^k and a randomly sampled mini-batch ξ_k^t from client k's local dataset $\mathcal{D}_k$:

$$w_{t+1}^k = w_t^k - \eta_t \nabla L_k(w_t^k; \xi_k^t). \tag{81}$$

By defining the local stochastic gradient $s_t = \nabla L_k(w_t^k; \xi_k^t)$, the update rule can be simplified to:

$$w_{t+1}^k = w_t^k - \eta_t s_t. \tag{82}$$

After local updates, the global server aggregates all local parameters w_{t+1}^k using weighted averaging with client weights p_k (satisfying $\sum_{k=1}^{M} p_k = 1$). The aggregated parameter $\overline{v}_{t+1}$ is thus:

$$\overline{v}_{t+1} = \sum_{k=1}^{M} p_k w_{t+1}^k. \tag{83}$$

Substituting the local update rule into the aggregation formula, we get:

$$\overline{v}_{t+1} = \sum_{k=1}^{M} p_k \left(w_t^k - \eta_t s_t\right). \tag{84}$$

Our goal is to bound the expected squared norm of the error between the aggregated parameter $\overline{v}_{t+1}$ and the global optimal parameter w^*. First, we write the error term as:

$$\overline{v}_{t+1} - w^* = \sum_{k=1}^{M} p_k \left(w_t^k - \eta_t s_t\right) - w^*. \tag{85}$$

To facilitate bounding, we use the add-and-subtract technique to introduce the global parameter $\overline{w}_t$ at round t into the expression. This decomposes the error into three manageable components:

$$\sum_{k=1}^{M} p_k \left(w_t^k - \eta_t s_t\right) - w^* = \underbrace{\sum_{k=1}^{M} p_k (w_t^k - \overline{w}_t)}_{A} + \underbrace{\overline{w}_t - w^*}_{B} + \underbrace{-\eta_t \sum_{k=1}^{M} p_k s_t}_{C}. \tag{86}$$

Here, A represents the weighted deviation between local parameters and the global parameter, B is the error of the current global parameter relative to the optimal, and C is the weighted sum of local stochastic gradients scaled by the step size.

We apply the Cauchy-Schwarz inequality for the squared norm of a sum of vectors. For any three vectors $A, B, C \in \mathbb{R}^d$, the following inequality holds:

$$\|A + B + C\|^2 \leq 3\|A\|^2 + 3\|B\|^2 + 3\|C\|^2. \tag{87}$$

This inequality is derived from the basic property $\|z + y\|^2 \leq 2\|z\|^2 + 2\|y\|^2$ (by expanding $\|z + y\|^2 = \|z\|^2 + 2\langle z, y\rangle + \|y\|^2$ and using the Cauchy-Schwarz inequality $\langle z, y\rangle \leq \|z\|\|y\| \leq \frac{1}{2}(\|z\|^2 + \|y\|^2)$). Extending this to three vectors gives the above bound with coefficient 3. Taking the expectation of both sides (since we aim for an expected error bound), we obtain:

$$\mathbb{E}\|\overline{v}_{t+1} - w^*\|^2 = \mathbb{E}\|A + B + C\|^2 \leq 3\mathbb{E}\|A\|^2 + 3\mathbb{E}\|B\|^2 + 3\mathbb{E}\|C\|^2. \tag{88}$$

By definition, $A = \sum_{k=1}^{M} p_k(w_t^k - \overline{w}_t)$. We use the convexity of the squared norm, which states that for any convex function $f(\cdot)$ and weights $p_k \geq 0$ with $\sum p_k = 1$, $f\left(\sum p_k z_k\right) \leq \sum p_k f(z_k)$. Since the squared norm $\|\cdot\|^2$ is convex, we have:

$$\|A\|^2 = \left\|\sum_{k=1}^{M} p_k(w_t^k - \overline{w}_t)\right\|^2 \leq \sum_{k=1}^{M} p_k \|w_t^k - \overline{w}_t\|^2. \tag{89}$$

Taking the expectation of both sides and multiplying by 3, we get:

$$3\mathbb{E}\|A\|^2 \leq 3\mathbb{E}\sum_{k=1}^{M} p_k \|w_t^k - \overline{w}_t\|^2. \tag{90}$$

First, substitute the definition of the global average gradient $\overline{s}_t = \sum_{k=1}^{M} p_k s_t$ into component C:

$$C = -\eta_t \sum_{k=1}^{M} p_k s_t = -\eta_t \overline{s}_t. \tag{91}$$

Thus, the squared norm of C is:

$$\|C\|^2 = \eta_t^2 \|\overline{s}_t\|^2. \tag{92}$$

Taking the expectation and multiplying by 3, we have:

$$3\mathbb{E}\|C\|^2 = 3\eta_t^2 \mathbb{E}\|\overline{s}_t\|^2. \tag{93}$$

To bound $\mathbb{E}\|\overline{s}_t\|^2$, we again use the add-and-subtract technique to introduce the local stochastic gradient s_t:

$$\overline{s}_t = s_t - (s_t - \overline{s}_t). \tag{94}$$

Applying the basic inequality $\|z - y\|^2 \leq 2\|z\|^2 + 2\|y\|^2$ (a special case of the Cauchy-Schwarz inequality for two vectors), we get:

$$\|\overline{s}_t\|^2 = \|s_t - (s_t - \overline{s}_t)\|^2 \leq 2\|s_t\|^2 + 2\|s_t - \overline{s}_t\|^2. \tag{95}$$

Taking the expectation of both sides:

$$\mathbb{E}\|\overline{s}_t\|^2 \leq 2\mathbb{E}\|s_t\|^2 + 2\mathbb{E}\|s_t - \overline{s}_t\|^2. \tag{96}$$

Substituting this back into the bound for $3\mathbb{E}\|C\|^2$, we obtain:

$$3\mathbb{E}\|C\|^2 \leq 3\eta_t^2 \left(2\mathbb{E}\|s_t\|^2 + 2\mathbb{E}\|s_t - \overline{s}_t\|^2\right) = 6\eta_t^2 \mathbb{E}\|s_t\|^2 + 6\eta_t^2 \mathbb{E}\|s_t - \overline{s}_t\|^2. \tag{97}$$

Next, we bound $\mathbb{E}\|s_t\|^2$ using the smoothness of $L_k(w)$ and the definition of data heterogeneity Γ. Recall that for an smooth function, the following inequality holds for any parameter w and optimal parameter w^*:

$$\|\nabla L_k(w)\|^2 \leq 2(L + L_\alpha)\left(L_k(w) - L_k(w^*)\right). \tag{98}$$

This inequality is derived from the Taylor expansion of $L_k(w)$ around w^* and the L-smoothness property. Applying this to $s_t = \nabla L_k(w_t^k; \xi_k^t)$ (noting that the expectation of the stochastic gradient equals the full gradient, i.e., $\mathbb{E}[s_t] = \nabla L_k(w_t^k)$), we take the expectation of both sides:

$$\mathbb{E}\|s_t\|^2 = \mathbb{E}\left\|\nabla L_k(w_t^k; \xi_k^t)\right\|^2 \geq \left\|\mathbb{E}\left[\nabla L_k(w_t^k; \xi_k^t)\right]\right\|^2 = \|\nabla L_k(w_t^k)\|^2 \tag{99}$$

By Jensen's inequality, since $\|\cdot\|^2$ is convex. However, for upper bounding, we use the smoothness of the base loss $L_k(w)$ and the definition of Γ. At the global optimal parameter w^*, the gradient of the global loss is zero: $\nabla L(w^*) = \sum_{k=1}^M p_k \nabla L_k(w^*) = 0$. Thus, the deviation between the local gradient and global gradient at w^* is $\nabla L_k(w^*) - \nabla L(w^*) = \nabla L_k(w^*)$. By the definition of Γ, we have:

$$\sum_{k=1}^M p_k \|\nabla L_k(w^*)\|^2 = \Gamma. \tag{100}$$

For w_t^k close to w^* (a valid assumption in convergence analysis, as the parameter converges to w^*), we approximate $\nabla L_k(w_t^k) \approx \nabla L_k(w^*)$ (since the regularized term's gradient is bounded and vanishes at convergence). Thus:

$$\mathbb{E}\|s_t\|^2 \leq (L + L_\alpha) \cdot \Gamma. \tag{101}$$

Substituting this into the bound for $3\mathbb{E}\|C\|^2$, we get the $6L\eta_t^2\Gamma$ term:

$$6\eta_t^2 \mathbb{E}\|s_t\|^2 \leq 6(L + L_\alpha)\eta_t^2 \Gamma. \tag{102}$$

Component $B = \overline{w}_t - w^*$ is the error of the current global parameter. We use the strong convexity of $L_k(w)$. The following inequality holds:

$$L_k(w) \geq L_k(w^*) + \langle \nabla L_k(w^*), w - w^* \rangle + \frac{(\mu - \rho)}{2}\|w - w^*\|^2. \tag{103}$$

Since $\nabla L_k(w^*) = 0$ (by optimality of w^*), this simplifies to:

$$L_k(w) - L_k(w^*) \geq \frac{(\mu - \rho)}{2}\|w - w^*\|^2. \tag{104}$$

Taking the weighted sum over all clients (with weights p_k) to get the global loss:

$$L(w) - L(w^*) \geq \frac{(\mu - \rho)}{2}\|w - w^*\|^2. \tag{105}$$

For the global parameter $\overline{w}_t$, this implies:

$$\|B\|^2 = \|\overline{w}_t - w^*\|^2 \leq \frac{2}{(\mu - \rho)}\left(L_(\overline{w}_t) - L(w^*)\right). \tag{106}$$

To connect this to the step size η_t, we use the update rule's contraction property. For SGD with step size $\eta_t \leq \frac{2}{\mu}$, the error term satisfies:

$$\|\overline{w}_t - w^*\|^2 \leq (1 - \eta_t(\mu - \rho))\|\overline{w}_t - w^*\|^2. \tag{107}$$

This contraction inequality is a standard result in strongly convex optimization, where the step size η_t is chosen to ensure that the parameter error decreases monotonically. Multiplying by 3 (from the earlier Cauchy-Schwarz bound), we get:

$$3\mathbb{E}\|B\|^2 \leq 3(1 - \eta_t(\mu - \rho))\mathbb{E}\|\overline{w}_t - w^*\|^2. \tag{108}$$

Substituting the bounds of A, B, C into the initial inequality:

$$\mathbb{E}\|\overline{v}_{t+1} - w^*\|^2 \leq 3\mathbb{E}\sum_{k=1}^M p_k \|w_t^k - \overline{w}_t\|^2 + 3(1 - \eta_t(\mu - \rho))\mathbb{E}\|\overline{w}_t - w^*\|^2 + 6(L + L_\alpha)\eta_t^2\Gamma + 6\eta_t^2\mathbb{E}\|s_t - \overline{s}_t\|^2. \tag{109}$$

By adjusting some constants, we obtain the desired result:

$$\mathbb{E}\|\overline{v}_{t+1}-w^*\|^2 \leq (1-\eta_t(\mu-\rho))\mathbb{E}\|\overline{w}_t-w^*\|^2+\eta_t^2\mathbb{E}\|s_t-\overline{s}_t\|^2+6(L+L_\alpha)\eta_t^2\Gamma+2\mathbb{E}\sum_{k=1}^{M}p_k\|\overline{w}_t-w_t^k\|^2. \tag{110}$$

Thus

$$\mathbb{E}\|\overline{v}_{t+1}-w^*\|^2 \leq (1-\eta_t(\mu-\rho))\mathbb{E}\|\overline{w}_t-w^*\|^2+\eta_t^2(C_\Gamma\cdot\Gamma+C_\Lambda\cdot\Lambda+D)+6(L+L_\alpha)\eta_t^2\Gamma+2\mathbb{E}\sum_{k=1}^{M}p_k\|\overline{w}_t-w_t^k\|^2. \tag{111}$$

This completes the proof of Lemma 3.7. □

C.9 Proof of Theorem 3.8 (Data Heterogeneity-Noise Homogeneity)

Proof. Under the noise-homogeneous setting($\Lambda = 0$), all clients share the same noise distribution. As a result, the stochastic gradient variance induced by label noise does not vary across clients. Its total variance can be decomposed into a data-induced component and a noise-induced component:

$$\sigma_k^2 = \sigma_{\text{data},k}^2 + \overline{\sigma}_{\text{noise}}^2, \tag{112}$$

where $\overline{\sigma}_{\text{noise}}^2$ is identical for all clients. Therefore, the variance discrepancy among clients arises solely from data heterogeneity, while the noise component does not introduce additional heterogeneity.

According to Lemma 3.7, the error propagation of the global model after one aggregation round satisfies

$$\begin{aligned}\mathbb{E}\|\overline{v}_{t+1} - w^*\|^2 &\leq (1-\eta_t(\mu-\rho))\mathbb{E}\|\overline{w}_t - w^*\|^2 \\ &\quad + \eta_t^2\mathbb{E}\|s_t - \overline{s}_t\|^2 + 6(L+L_\alpha)\eta_t^2\Gamma + 2\mathbb{E}\sum_{k=1}^{M}p_k\|\overline{w}_t - w_t^k\|^2,\end{aligned} \tag{113}$$

For notational convenience, define

$$\Delta_t \triangleq \mathbb{E}\|\overline{w}_t - w^*\|^2. \tag{114}$$

We now bound each additional term on the right-hand side separately.

First, consider the gradient noise term $\mathbb{E}\|s_t - \overline{s}_t\|^2$. Under full client participation, the aggregation weights satisfy $p_k \geq 0$ and $\sum_{k=1}^{M} p_k = 1$. Since the stochastic gradients of different clients are conditionally independent given $\overline{w}_t$, the variance of the weighted sum can be bounded as

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 = \mathbb{E}\Big\|\sum_{k=1}^{M}p_k\big(\nabla L_k(w_t^k;\xi_k^t) - \nabla L_k(w_t^k)\big)\Big\|^2 \leq \sum_{k=1}^{M}p_k^2\sigma_k^2. \tag{115}$$

Under noise homogeneity, $\sigma_k^2 = \sigma_{\text{data},k}^2 + \overline{\sigma}_{\text{noise}}^2$, and thus the above bound does not involve any noise heterogeneity term.

Next, we bound the parameter divergence induced by multiple local updates. According to Lemma 3.6, when each client performs E local updates per round, we have

$$\mathbb{E}\sum_{k=1}^{M}p_k\|\overline{w}_t - w_t^k\|^2 \leq 4\eta_t^2(E-1)^2\big(G_F^2 + \mu^2 G_\alpha^2\big)^2. \tag{116}$$

The factor $(E-1)^2$ reflects the fact that, except for the first local update which starts from the global model, all subsequent updates are performed on already drifted local parameters, leading to accumulated deviation.

Substituting the above bounds into the recursion from Lemma 3.7 yields

$$\begin{aligned}\Delta_{t+1} &\leq (1-\eta_t(\mu-\rho))\Delta_t + \eta_t^2\sum_{k=1}^{M}p_k^2\sigma_k^2 \\ &\quad + 6(L+L_\alpha)\eta_t^2\Gamma + 8\eta_t^2(E-1)^2\big(G_F^2+\mu^2G_\alpha^2\big)^2.\end{aligned} \tag{117}$$

Collecting all η_t^2-scaled constant terms, define

$$H \triangleq \sum_{k=1}^{M} p_k^2 \sigma_k^2 + 8(E-1)^2 \big(G_F^2 + \mu^2 G_\alpha^2\big)^2. \tag{118}$$

Then the recursion simplifies to

$$\Delta_{t+1} \le (1 - \eta_t(\mu - \rho))\Delta_t + \eta_t^2 \big(H + 6(L + L_\alpha)\Gamma\big). \tag{119}$$

To obtain an explicit convergence rate, we adopt the standard diminishing step size

$$\eta_t = \frac{2}{(\mu - \rho)(\gamma + t)}, \quad \gamma = \max\{8\kappa, E\}, \quad \kappa = \frac{L + L_\alpha}{\mu - \rho}. \tag{120}$$

Substituting this choice yields

$$1 - \eta_t(\mu - \rho) = \frac{\gamma + t - 2}{\gamma + t}, \qquad \eta_t^2 = \frac{4}{(\mu - \rho)^2(\gamma + t)^2}. \tag{121}$$

Therefore, the recursion becomes

$$\Delta_{t+1} \le \frac{\gamma + t - 2}{\gamma + t}\Delta_t + \frac{4}{(\mu - \rho)^2(\gamma + t)^2}\big(H + 6(L + L_\alpha)\Gamma\big). \tag{122}$$

We now establish an $O(1/t)$ convergence rate via induction. Assume there exists a constant $v > 0$ such that

$$\Delta_t \le \frac{v}{\gamma + t}, \quad \forall t \ge 0. \tag{123}$$

For $t = 0$, this condition holds as long as $v \ge \gamma \|w_0 - w^*\|^2$. Assuming the inequality holds at iteration t, substituting it into the recursion shows that it remains valid at iteration $t + 1$ provided that

$$v \ge \frac{4\kappa}{(\mu - \rho)^2}\big(H + 6(L + L_\alpha)\Gamma\big). \tag{124}$$

Combining the above requirement with the initialization condition and slightly loosening the constants without affecting the convergence order, we may choose

$$v = C_1 + C_2\Gamma, \tag{125}$$

where

$$C_1 = \frac{2\kappa D + 16\kappa(E-1)^2(G_F^2 + \mu^2 G_\alpha^2)^2}{(\mu - \rho)^2} + 4\kappa(L + L_\alpha)\|w_0 - w^*\|^2, \tag{126}$$

and

$$C_2 = \frac{2\kappa C_\Gamma + 12\kappa(L + L_\alpha)}{(\mu - \rho)^2}. \tag{127}$$

Consequently, after T communication rounds, the global model satisfies

$$\mathbb{E}\|\tilde{w}_T - w^*\|^2 \le \frac{1}{\gamma + T}\big(C_1 + C_2\Gamma\big), \tag{128}$$

The proof of Theorem 3.8 is completed.

□

C.10 Proof of Theorem 3.9 (Data Homogeneity-Noise Heterogeneity)

Proof. Under the data homogeneity ($\Gamma = 0$) and noise heterogeneity condition, the label noise distributions of different clients may vary, so the stochastic gradient variance induced by noise is no longer identical across clients. Total variance can be decomposed as

$$\sigma_k^2 = \sigma_{\text{data},k}^2 + \sigma_{\text{noise},k}^2. \tag{129}$$

Define the average noise variance

$$\overline{\sigma}^2_{\text{noise}} = \sum_{k=1}^{M} p_k \sigma^2_{\text{noise},k} \tag{130}$$

and the noise heterogeneity index

$$\Lambda = \sum_{k=1}^{M} p_k \big(\sigma^2_{\text{noise},k} - \overline{\sigma}^2_{\text{noise}}\big)^2. \tag{131}$$

Then the weighted second moment of the gradient variance can be bounded by

$$\sum_{k=1}^{M} p_k^2 \sigma_k^2 = \sum_{k=1}^{M} p_k^2 \sigma^2_{\text{data},k} + \sum_{k=1}^{M} p_k^2 \sigma^2_{\text{noise},k} \le D_0 + C_\Lambda \Lambda, \tag{132}$$

where

$$D_0 = \sum_{k=1}^{M} p_k^2 \sigma^2_{\text{data},k} + \overline{\sigma}^2_{\text{noise}}, \tag{133}$$

and $C_\Lambda > 0$ is a finite constant depending on the aggregation weights $\{p_k\}$.

According to Lemma 3.7, the error propagation of the global model after one aggregation round satisfies

$$\begin{aligned}\mathbb{E}\|\overline{v}_{t+1} - w^*\|^2 \le (1 - \eta_t(\mu - \rho))\mathbb{E}\|\tilde{w}_t - w^*\|^2 \\ + \eta_t^2 \mathbb{E}\|s_t - \overline{s}_t\|^2 + 6(L + L_\alpha)\eta_t^2 \Gamma + 2\mathbb{E}\sum_{k=1}^{M} p_k \|\tilde{w}_t - w_t^k\|^2.\end{aligned} \tag{134}$$

Define

$$\Delta_t \triangleq \mathbb{E}\|\tilde{w}_t - w^*\|^2. \tag{135}$$

According to Lemma 3.5, the gradient deviation term can be bounded as

$$\mathbb{E}\|s_t - \overline{s}_t\|^2 \le \sum_{k=1}^{M} p_k^2 \sigma_k^2 \le D_0 + C_\Lambda \Lambda, \tag{136}$$

and according to Lemma 3.6, the parameter divergence term satisfies

$$\mathbb{E}\sum_{k=1}^{M} p_k \|\tilde{w}_t - w_t^k\|^2 \le 4\eta_t^2 (E-1)^2 (G_F^2 + \mu^2 G_\alpha^2)^2. \tag{137}$$

Substituting these bounds into Lemma 3.7 and collecting all η_t^2-scaled constants, we obtain

$$\Delta_{t+1} \le (1 - \eta_t(\mu - \rho))\Delta_t + \eta_t^2 (B_0 + C_3 \Lambda), \tag{138}$$

where

$$B_0 = D_0 + 8(E-1)^2 (G_F^2 + \mu^2 G_\alpha^2)^2 + 6(L + L_\alpha)\Gamma, \tag{139}$$

and

$$C_3 = \frac{2\kappa C_\Lambda}{(\mu - \rho)^2}. \tag{140}$$

Adopting the same diminishing step size as in Theorem 3.8

$$\eta_t = \frac{2}{(\mu - \rho)(\gamma + t)}, \tag{141}$$

we have

$$1 - \eta_t(\mu - \rho) = \frac{\gamma + t - 2}{\gamma + t}, \qquad \eta_t^2 = \frac{4}{(\mu - \rho)^2 (\gamma + t)^2}. \tag{142}$$

Then the recursion becomes

$$\Delta_{t+1} \leq \frac{\gamma+t-2}{\gamma+t}\Delta_t + \frac{4}{(\mu-\rho)^2(\gamma+t)^2}(B_0 + C_3\Lambda). \tag{143}$$

Using mathematical induction, assume there exists a constant $v > 0$ such that

$$\Delta_t \leq \frac{v}{\gamma+t}, \quad \forall t \geq 0. \tag{144}$$

At $t = 0$, this requires

$$v \geq \gamma\|w_0 - w^*\|^2. \tag{145}$$

Substituting into the recursion for $t \geq 0$ leads to the condition

$$v(\gamma+t+2) \geq \frac{4}{(\mu-\rho)^2}(B_0 + C_3\Lambda). \tag{146}$$

Combining with $\gamma \geq 8$, we can choose

$$v = C_1 + C_3\Lambda, \tag{147}$$

where

$$C_1 = \frac{2\kappa D + 16\kappa(E-1)^2(G_F^2 + \mu^2 G_\alpha^2)^2}{(\mu-\rho)^2} + 4\kappa(L+L_\alpha)\|w_0 - w^*\|^2. \tag{148}$$

Therefore, after T communication rounds, the global model satisfies

$$\mathbb{E}\|\tilde{w}_T - w^*\|^2 \leq \frac{1}{\gamma+T}\left(C_1 + C_3\Lambda\right), \tag{149}$$

The proof of Theorem 3.9 is completed.

□

C.11 Proof of Theorem 3.10 (Data-Noise Dual Heterogeneity)

Proof. In the data-noise dual heterogeneity scenario, the total gradient variance of each client includes contributions from both data heterogeneity and noise heterogeneity:

$$\sigma_k^2 = \sigma_{\text{data},k}^2 + \sigma_{\text{noise},k}^2. \tag{150}$$

Here, $\sigma_{\text{data},k}^2$ increases with Γ, indicating that stronger data heterogeneity leads to larger data-induced variance. $\sigma_{\text{noise},k}^2 = \sigma_0^2 + C\theta_k$ exhibits larger differences with an increase of Λ, meaning stronger noise heterogeneity results in larger differences in noise-induced variance among clients.

According to Lemma 3.7, the single-step iteration error propagation is given by:

$$\mathbb{E}\|\overline{v}_{t+1} - w^*\|^2 \leq (1-\eta_t\mu)\mathbb{E}\|\overline{w}_t - w^*\|^2 + \eta_t^2\mathbb{E}\|s_t - \overline{s}_t\|^2 + 6(L+L_\alpha)\eta_t^2\Gamma + 2\sum_{k=1}^{M} p_k\mathbb{E}\|\overline{w}_t - w_t^k\|^2, \tag{151}$$

Substituting the gradient deviation variance bound from Lemma 3.5, $\mathbb{E}\|s_t - \overline{s}_t\|^2 \leq \sum_{k=1}^{M} p_k^2\sigma_k^2$, and the parameter divergence bound from Lemma 3.6, $\sum_{k=1}^{M} p_k\mathbb{E}\|\overline{w}_t - w_t^k\|^2 \leq 4\eta_t^2(E-1)^2(G_F^2 + \mu^2 G_\alpha^2)^2$, and combining constant terms, the recursive formula becomes:

$$\Delta_{t+1} \leq (1-\eta_t\mu)\Delta_t + \eta_t^2\left(\sum_{k=1}^{M} p_k^2\sigma_k^2 + 6(L+L_\alpha)\Gamma + 8(E-1)^2(G_F^2 + \mu^2 G_\alpha^2)^2\right). \tag{152}$$

Defining $D = \sum_{k=1}^{M} p_k^2 \sigma_k^2 + 8(E-1)^2(G_F^2 + \mu^2 G_\alpha^2)^2$, and letting the data heterogeneity and noise heterogeneity contributions be C_Γ and C_Λ, the recursion simplifies to:

$$\Delta_{t+1} \leq (1 - \eta_t \mu)\Delta_t + \eta_t^2 (C_1 + C_2\Gamma + C_3\Lambda). \tag{153}$$

Using the same step size decay strategy as in Theorems 3.8 and 3.9, $\eta_t = 2/((\mu+\rho)(\gamma+t))$, the recursion further simplifies to:

$$\Delta_{t+1} \leq \frac{\gamma + t - 2}{\gamma + t}\Delta_t + \frac{4(C_1 + C_2\Gamma + C_3\Lambda)}{(\mu+\rho)^2(\gamma+t)^2}. \tag{154}$$

By induction, setting $v = 2\left(\frac{C_1 + C_2\Gamma + C_3\Lambda}{\mu^2} + 2\|w_0 - w^*\|^2\right)$, we obtain at the T-th round:

$$\mathbb{E}\|\tilde{w}_T - w^*\|^2 \leq \frac{1}{\gamma + T}(C_1 + C_2\Gamma + C_3\Lambda). \tag{155}$$

In this dual heterogeneity scenario, the negative impacts of data heterogeneity Γ and noise heterogeneity Λ are additive. An increase in Γ increases $C_2\Gamma$, and an increase in Λ increases $C_3\Lambda$, both of which raise the upper bound of the convergence error and slow the convergence rate. However, the convergence order remains $O(1/T)$, indicating that the FedCoIn algorithm can still guarantee effective convergence under the most complex dual heterogeneity scenario.

The proof of Theorem 3.10 is completed.

□

D Dirichlet Reparameterization and Variational KL Derivation

The core goal of local inference in FedCoIn is to minimize the Kullback–Leibler (KL) divergence between the variational distribution $q_{w_k}(y|\hat{y}, z)$ and the target posterior distribution $p(y|\hat{y}, z)$. Since the Dirichlet distribution cannot be directly reparameterized to compute gradients, this section first completes the reparameterization derivation based on Gamma distribution decomposition, then connects it with the full integral expansion of the KL divergence, thereby providing theoretical support for gradient propagation in the local loss and accommodating the requirements of distributed optimization in federated learning scenarios.

D.1 Definition of Basic Distribution Relationships (Section 2.3)

The Dirichlet distribution can be constructed through the normalization of multiple independent Gamma distributions, which serves as the core theoretical foundation for reparameterization and KL divergence derivation. The specific definitions are as follows.

Gamma Distribution. Let the shape parameter be $\alpha_k > 0$ and the rate parameter be $\beta > 0$. The probability density function of the Gamma distribution is:

$$\mathrm{Gamma}(z; \alpha_k, \beta) = \frac{\beta^{\alpha_k}}{\Gamma(\alpha_k)} z^{\alpha_k - 1} e^{-\beta z}, \tag{156}$$

where $\Gamma(\cdot)$ denotes the Gamma function, defined as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt. \tag{157}$$

Dirichlet Distribution. Suppose there exist C mutually independent Gamma random variables $Z_k \sim \mathrm{Gamma}(\alpha_k, \beta)$, where C denotes the number of classes. Define the normalized variables

$$Y_k = \frac{Z_k}{\sum_{m=1}^{C} Z_m}, \tag{158}$$

Then the random vector $Y = (Y_1, \ldots, Y_C)$ follows a Dirichlet distribution, denoted as $Y \sim \mathrm{Dir}(\alpha)$, with probability density function

$$\mathrm{Dir}(Y;\alpha) = \frac{\Gamma\left(\sum_{k=1}^{C} \alpha_k\right)}{\prod_{k=1}^{C} \Gamma(\alpha_k)}. \tag{159}$$

D.2 Dirichlet Distribution Reparameterization (Eq. (5))

Targeting the KL divergence minimization objective in FedCoIn's local inference, the reparameterization process is fully completed locally at the client, with steps as follows:

Variable Mapping. For sample z at client i, let the variational distribution $q_{w_k}(y|\hat{y}, z) = \mathrm{Dir}(\alpha_q)$ where $\alpha_q = \alpha_{\text{infer},i,n}$ is the variational Dirichlet parameter and the prior part of the target posterior distribution $p(y|z) = \mathrm{Dir}(\alpha_p)$ where $\alpha_p = \alpha_{\text{prior},i,n}$ is the local prior parameter. Map the latent variable $y \sim \mathrm{Dir}(\alpha_q)$ to independent Gamma variables $z_k \sim \mathrm{Gamma}(\alpha_q^k, \beta)$, satisfying

$$y_k = \frac{z_k}{\sum_{m=1}^{C} z_m}. \tag{160}$$

Gamma Variable Reparameterization. Introduce independent uniform noise $\epsilon_k \sim \mathrm{Uniform}(0, 1)$ and achieve differentiable transformation through the inverse cumulative distribution function (Inverse CDF) of the Gamma distribution:

$$Z_k = \mathrm{Gamma}^{-1}(\epsilon_k; \alpha_q^k, \beta). \tag{161}$$

where $\mathrm{Gamma}^{-1}(\cdot)$ can be numerically solved via the Newton-Raphson method, ensuring the mapping between Z_k and α_q^k is differentiable.

Gradient Propagation. The gradient of the KL divergence with respect to the variational parameter α_q can be derived using the chain rule, with the propagation as:

$$\nabla_{\alpha_q} \mathrm{KL}(q_{w_k} \parallel p) = \nabla_y \mathrm{KL}(q_{w_k} \parallel p) \cdot \nabla_z y \cdot \nabla_{\alpha_q} Z. \tag{162}$$

where $\nabla_y \mathrm{KL}(q_{w_k} \parallel p)$ is the gradient of the KL divergence with respect to the latent variable y, which can be obtained from the KL divergence integral derivation in Appendix D.3; $\nabla_Z y$ is the gradient of the normalization operation, satisfying

$$\nabla_{Z_k} y_j = \frac{\delta_{kj} \sum_{m=1}^{C} Z_m - Z_j}{(\sum_{m=1}^{C} Z_m)^2}. \tag{163}$$

where δ_{kj} is the Kronecker delta function; $\nabla_{\alpha_q} Z$ is the gradient of the Gamma variable with respect to the variational parameter, which has an analytical solution derived from the derivative properties of the inverse CDF.

D.3 Complete Integral Derivation of Variational KL Divergence (Eqs. (4) and (5))

Based on the reparameterization in Appendix D.2, the complete integral expansion derivation of $\mathrm{KL}(q_{w_k}(y|\hat{y}, z) \parallel p(y|\hat{y}, z))$ is as follows.

According to the definition of KL divergence and combining Bayes' theorem to decompose the target posterior distribution:

$$\begin{aligned} \mathrm{KL}(q_{w_k} \parallel p) &= \int_{\mathcal{Y}} q_{w_k}(y|\hat{y}, z) \log \frac{q_{w_k}(y|\hat{y}, z)}{p(y|\hat{y}, z)} dy \\ &= \int_{\mathcal{Y}} q_{w_k}(y|\hat{y}, z) \log \frac{q_{w_k}(y|\hat{y}, z) \cdot p(\hat{y}|z)}{p(\hat{y}|y, z) \cdot p(y|z)} dy. \end{aligned} \tag{164}$$

where Bayes' theorem is applied as $p(y|\hat{y}, z) = \frac{p(\hat{y}|y,z)p(y|z)}{p(\hat{y}|z)}$, $p(\hat{y}|z)$ is the marginal probability of noisy prediction, and$p(\hat{y}|y, z)$ is the conditional multi distribution of noisy prediction.

Substitute the probability density functions of the Dirichlet and multi distributions, then split the integral terms:

$$\mathrm{KL}(q_{w_k} \parallel p) = \int_{0^\infty} \int_{0^\infty} \cdots \int_{0^\infty} \mathrm{Gamma}(\alpha_q^1, 1) \cdot \mathrm{Gamma}(\alpha_q^2, 1) \cdot \cdots \cdot \mathrm{Gamma}(\alpha_q^C, 1) \cdot \frac{1}{\left(\sum_{m=1}^C Z_m\right)^C}$$
$$\log \left(\frac{\frac{\Gamma(\sum_{k=1}^C \alpha_q^k)}{\prod_{k=1}^C \Gamma(\alpha_q^k)} \left(\frac{Z_1}{\sum_{m=1}^C Z_m}, \frac{Z_2}{\sum_{m=1}^C Z_m}, \cdots, \frac{Z_C}{\sum_{m=1}^C Z_m} \right)^{\alpha_q - 1} \cdot p(\hat{y}|z)}{\prod_{k=1}^C \pi_{z,y}^{k\ \hat{y}^k} \cdot \frac{\Gamma(\sum_{k=1}^C \alpha_p^k)}{\prod_{k=1}^C \Gamma(\alpha_p^k)} \left(\frac{Z_1}{\sum_{m=1}^C Z_m}, \frac{Z_2}{\sum_{m=1}^C Z_m}, \cdots, \frac{Z_C}{\sum_{m=1}^C Z_m} \right)^{\alpha_p - 1}} \right) dZ_1 \cdots dZ_C$$

Further split the logarithmic and integral terms, and simplify using the properties of the Gamma and Dirichlet distributions:

$$\begin{aligned} &= \log p(\hat{y}|z) \int_{\mathcal{Y}} q_{w_k}(y) dy - \int_{\mathcal{Y}} q_{w_k}(y) \sum_{k=1}^C \hat{y}^k \log \pi_{z,y}^k dy \\ &\quad - \int_{\mathcal{Y}} q_{w_k}(y) \log \left(\frac{\Gamma(\sum \alpha_p^k)}{\prod \Gamma(\alpha_p^k)} \prod y_k^{\alpha_p^k - 1} \right) dy + \int_{\mathcal{Y}} q_{w_k}(y) \log \left(\frac{\Gamma(\sum \alpha_q^k)}{\prod \Gamma(\alpha_q^k)} \prod y_k^{\alpha_q^k - 1} \right) dy. \end{aligned} \tag{165}$$

Term-by-term simplification and substitution of special functions: Regarding the constant term integral, since the variational distribution $q_{w_k}(y)$ is normalized over its domain, i.e.,

$$\int_{\mathcal{Y}} q_{w_k}(y)\, dy = 1, \tag{166}$$

this term can be directly simplified to

$$\log p(\hat{y} \mid z). \tag{167}$$

Regarding the multinomial term integral, after applying the Gamma-distribution-based reparameterization, a variable substitution transforms the integral over the simplex into an expectation with respect to Gamma random variables:

$$\int_{\mathcal{Y}} q_{w_k}(y) \sum_{k=1}^C \hat{y}^k \log \pi_{z,y}^k\, dy = \mathbb{E}_{z \sim \mathrm{Gamma}(\alpha_q)} \left[\sum_{k=1}^C \hat{y}^k \log \pi_{z,Y}^k \right]. \tag{168}$$

where $Y_k = \frac{Z_k}{\sum_m Z_m}$.

Regarding the prior Dirichlet term integral, by expanding the logarithmic expression and exploiting the logarithmic expectation property of the Dirichlet distribution $\mathbb{E}[\log y_k] = \psi(\alpha_q^k) - \psi\big(\sum_m \alpha_q^m\big)$, the integral can be simplified into a closed-form expression involving only Gamma and Digamma functions:

$$\begin{aligned} &\int_{\mathcal{Y}} q_{w_k}(y) \log \left(\frac{\Gamma(\sum_k \alpha_p^k)}{\prod_k \Gamma(\alpha_p^k)} \prod_k y_k^{\alpha_p^k - 1} \right) dy \\ &\qquad = \log \Gamma\Big(\sum_k \alpha_p^k \Big) - \sum_{k=1}^C \log \Gamma(\alpha_p^k) + \sum_{k=1}^C (\alpha_p^k - 1) \left(\psi(\alpha_q^k) - \psi\Big(\sum_m \alpha_q^m \Big) \right). \end{aligned} \tag{169}$$

Regarding the variational Dirichlet term integral, following the same expansion and simplification procedure as for the prior term yields its analytical form:

$$\begin{aligned} &\int_{\mathcal{Y}} q_{w_k}(y) \log \left(\frac{\Gamma(\sum_k \alpha_q^k)}{\prod_k \Gamma(\alpha_q^k)} \prod_k y_k^{\alpha_q^k - 1} \right) dy \\ &\qquad = \log \Gamma\Big(\sum_k \alpha_q^k \Big) - \sum_{k=1}^C \log \Gamma(\alpha_q^k) + \sum_{k=1}^C (\alpha_q^k - 1) \left(\psi(\alpha_q^k) - \psi\Big(\sum_m \alpha_q^m \Big) \right). \end{aligned} \tag{170}$$

Combine the simplified results of the four terms to obtain the closed-form expression of the KL divergence:

$$\begin{aligned}\mathrm{KL}(q_{w_k} \parallel p) = \log p(\hat{y}|z) - \mathbb{E}_{Z\sim\text{Gamma}(\alpha_q)}\left[\sum_{k=1}^{C} \hat{y}^k \log \pi_{z,y}^k\right] \\ - \log \frac{\Gamma(\sum \alpha_p^k)}{\Gamma(\sum \alpha_q^k)} + \sum_{k=1}^{C} \log \frac{\Gamma(\alpha_q^k)}{\Gamma(\alpha_p^k)} \\ - \sum_{k=1}^{C} (\alpha_p^k - \alpha_q^k)\left(\psi(\alpha_q^k) - \psi\left(\sum \alpha_q^k\right)\right).\end{aligned} \tag{171}$$

E Experiments Details

E.1 Additional Experimental Setups

Datasets Details. CIFAR-10 (Krizhevsky & Hinton, 2009) is a standard image classification benchmark with 50k training and 10k test images across 10 classes (animals and vehicles). All images are 32×32 RGB. Following our protocol, we split 20% of the training set for validation.

CIFAR-100 (Krizhevsky & Hinton, 2009) contains 50k training and 10k test images over 100 fine-grained classes grouped into 20 super-classes. It is more challenging due to higher intra-class variation and fine-grained labels. We also use a 20% validation split.

Clothing1M (Xiao et al., 2015) is a large-scale real-world dataset with 1,000k noisy training images and 14 clothing categories. Labels are generated from noisy web annotations, making it suitable for evaluating robustness to real-world label noise. It includes 10k clean test images. We split 20% of training data for validation.

Overall, the CIFAR-10/100 and Clothing1M datasets contain 50k, 50k, and 1,000k training images, and contain 10, 100, and 14 classes, respectively. They all contain 10k test images. we split 20% of the training set as a validation set.

Baselines We compare FedCoIn with existing state-of-the-art (SOTA) F-LNL methods, including FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020a), FedCorr (Xu et al., 2022), FedNoRo (Wu et al., 2023), FedBeat (Wang et al., 2024), FedELC (Jiang et al., 2024a), FedFixer (Ji et al., 2024) and FedClean (Jiang & Zhang, 2025). Specifically, FedAvg and FedProx are both classic FL algorithms, while FedCorr, FedNoRo, FedBeat, FedELC, FedFixer and FedClean are existing FNLL methods.

FedAvg (McMahan et al., 2017) is the standard federated learning algorithm that performs local SGD and global aggregation. It is communication-efficient but cannot handle data heterogeneity or label noise.

FedProx (Li et al., 2020a) extends FedAvg by adding a proximal term to local objectives, improving stability under non-IID data. However, it does not explicitly address label noise.

FedCorr (Xu et al., 2022) estimates client-level label noise via intrinsic dimensionality and performs iterative label correction. Its performance degrades in fully noisy or highly heterogeneous settings.

FedNoRo (Wu et al., 2023) identifies noisy clients in a class-aware manner and down-weights unreliable updates. It assumes the presence of relatively clean clients.

FedBeat (Wang et al., 2024) adaptively reweights clients based on reliability and performs local label correction, improving robustness but still relying on partial noise assumptions.

FedELC (Jiang et al., 2024a) detects noisy clients and refines labels via backpropagation. However, it fails when all clients contain severe label noise.

FedFixer (Ji et al., 2024) introduces a dual-model design with a noisy-label learner and a correction model to reduce overfitting to noise, at the cost of additional complexity.

FedClean (Jiang & Zhang, 2025) removes the clean-client assumption and integrates local CNLL, two-stage label correction, and adaptive self-supervised aggregation, achieving strong robustness under fully noisy settings.

Non-IID data partitions According to (Xu et al., 2022), we consider both IID and non-IID heterogeneous data partitions in this work. For IID partitions,the whole dataset D is uniformly distributed at random across N clients. For non-IID partitions, we first generate an $N \times M$ indicator matrix Φ, where each entry Φ_{ij} indicates whether the local dataset of client i contains class j. Each Φ_{ij} is sampled from the Bernoulli distribution with a fixed probability p. For each class $1 \leq j \leq M$, let v_j be the sum of entries in the j-th column of Φ; this equals the number of clients whose local datasets contain class j. Let q_j be a vector of length v_j, sampled from the symmetric Dirichlet distribution with the common parameter $\alpha_{\text{Dir}} > 0$. Using q_j as a probability vector, we then randomly allocate the samples within class j to these v_j clients.

Noise model Similar to FedCorr,To emulate label noise in real-world data, We shall introduce a general federated noise model framework. For simplicity, this work only considers instance-independent label noise. This framework has two parameters ρ and τ, where ρ denotes the system noise level (ratio of noisy clients) and τ denotes the lower bound for the noise level of a noisy client. Every client has a probability ρ of being a noisy client, in which case the local noise level for this noisy client is determined randomly by sampling from the uniform distribution $U(\tau, 1)$. Succinctly, the noise level for client k (for $k = 1, \ldots, N$) is given by:

$$\mu_k = \begin{cases} u \sim U(\tau, 1), & \text{with probability } \rho; \\ 0, & \text{with probability } 1 - \rho. \end{cases} \tag{172}$$

For clients with $\mu_k \neq 0$, $100 \cdot \mu_k\%$ of the samples are chosen uniformly at random and assigned random labels, selected uniformly from the M classes.

Implementation Details. When ViT is used for feature extraction, we adopt an MLP as the classification model. Under the ResNet setting, we select ResNet-18, ResNet-34, and ResNet-50 (He et al., 2016) as the network backbones for CIFAR-10, CIFAR-100, and Clothing1M, respectively. For CIFAR-10 and CIFAR-100, we consider both IID and non-IID data partitions, while for Clothing1M we only consider non-IID data partitions. We set the total number of clients to 50, the learning rate to 0.001, and the total number of communication rounds to 50. During local training, in each communication round, each client performs $E = 5$ local training epochs. The mini-batch sizes are set to 10, 10, and 256 for CIFAR-10, CIFAR-100, and Clothing1M, respectively. The performance of various methods is tested based on the data division and noise generator used in FedCorr (Xu et al., 2022), where $p \in (0, L]$ denotes the class sampling probability of clients, $\rho \in [0, 1]$ denotes the ratio of noisy clients, and $\tau \in [0, 1]$ denotes the lower bound for the noise level of a noisy client, which means that the noise level can be determined by sampling from the uniform distribution $U(\tau, 1)$.

Table 5: Accuracies (%) of various methods on CIFAR-10 with different non-IID settings. The noise level is set to $\rho = 1, \tau = 0.3$.

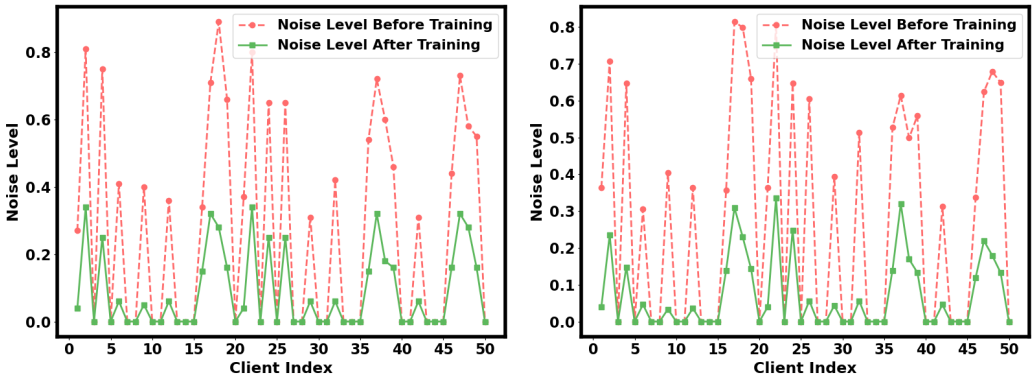
Method	Sampling probability for each class			
	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
FedAvg	32.48	35.89	43.88	39.27
FedProx	34.50	37.00	45.28	40.39
FedCorr	48.69	54.53	61.13	64.60
FedNoRo	57.65	68.19	72.36	73.62
FedBeat	30.24	33.06	35.73	37.71
FedELC	30.12	32.23	33.54	35.28
FedFixer	51.19	62.86	71.54	72.46
Ours	**75.26**	**79.95**	**82.34**	**83.22**

E.2 Additional Experimental Results

Heterogeneity. We further analyze performance under different levels of heterogeneity controlled by parameter p, with smaller p creating more challenging scenarios. Table 5 shows that FedCoIn consistently achieves the best accuracy under high-noise settings ($\rho = 1$) across all heterogeneity levels.

Correction Evaluation. We assessed the effectiveness of noisy label correction under the setting $\rho = 0.5$ and $\tau = 0.3$. Figure 4 illustrates the variation of client noise levels before and after correction under both IID and non-IID settings on CIFAR-100. It can be observed that the corrected noise levels are consistently lower than the original ones across different clients, demonstrating that FedCoIn effectively mitigates noisy labels under heterogeneous data distributions.

To intuitively demonstrate the effectiveness of label correction, we compare the confusion matrices of the original labels before training and the corrected labels after training. Figure 5 presents the comparison results for three randomly selected clients in experiments conducted on CIFAR-10 with IID data partition and noise settings $\rho = 1$, $\tau = 0.5$. The results demonstrate the effectiveness of our label correction approach. Specifically, the noise level of noisy clients decreases markedly after correction, and the confusion matrices show that the corrected labels are much more consistent with the true labels, indicating a significant improvement in label quality.



(a) IID (b) non-IID

Figure 4: Evaluation of label correction process on CIFAR-100 with ρ = 0.5 τ = 0.3 on IID and non-IID distribution.

IDN Noise. To further evaluate the robustness of FedCoIn under more realistic and challenging label corruption, we conduct experiments under the instance-dependent noise (IDN) setting. Unlike instance-independent noise, IDN introduces sample-dependent label flipping, which better reflects real-world annotation bias and ambiguity. Table 6 reports the classification accuracies under both IID and non-IID data partitions across different noise configurations parameterized by ρ and τ. The results show that FedCoIn maintains consistently strong performance across varying noise intensities and data heterogeneity levels, demonstrating its robustness in handling complex instance-dependent label noise.

Table 6: Accuracies (%) of FedCoIn under Instance-Dependent Noise (IDN).

Dataset	IID				Non-IID			
	$\rho = 0.3$ $\tau = 0.3$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 0.8$ $\tau = 0.5$	$\rho = 1.0$ $\tau = 0.5$	$\rho = 0.3$ $\tau = 0.3$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 0.8$ $\tau = 0.5$	$\rho = 1.0$ $\tau = 0.5$
CIFAR-10	97.78	97.41	96.85	96.48	97.38	97.16	96.77	96.36
CIFAR-100	86.72	86.10	85.39	84.90	85.13	84.71	83.42	82.87

Complexity Analysis. Feature extraction is a one-time preprocessing step with fixed computation and is independent of training iterations. The distribution-aware dynamic prior introduces an $O(NK)$



(a) Before training (Client 18) (b) Before training (Client 28) (c) Before training (Client 38)



(d) After training (Client 18) (e) After training (Client 28) (f) After training (Client 38)

Figure 5: An evaluation of the label correction process, conducted on CIFAR-10 with IID data partition and noise setting $\rho = 1$, $\tau = 0.5$. For each client, we give two confusion matrices, associated to the given labels before training and the corrected labels after training.

complexity per client due to adaptive KNN and confidence estimation, where N denotes the number of local samples and K the number of adaptive neighbors, making it lightweight and linear in the data size. The local variational inference performs Dirichlet-based updates on low-dimensional label distributions over C classes, resulting in a per-step complexity of $O(C)$, which is negligible compared to neural network forward and backward propagation. On the server side, aggregation only involves weighted combination of client statistics with complexity $O(M)$, where M is the number of clients. Importantly, only compact Dirichlet parameters and statistical summaries are transmitted rather than raw data or full model parameters, ensuring minimal communication overhead.

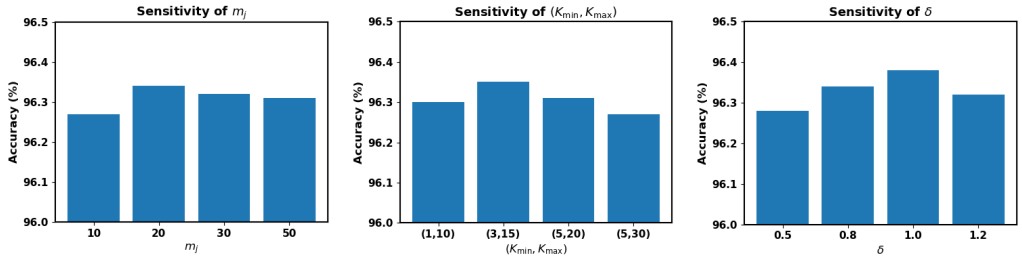
To empirically validate efficiency, we conducted runtime and GPU memory comparisons, using the same total data volume, the same number of clients, and identical batch size across all methods. As shown in Table 7, FedCoIn requires significantly less training time per round than alternative methods, while maintaining low memory consumption.

Table 7: Training time per round (s) and GPU memory consumption (GB) of different methods.

Method	FedCoIn (Ours)	FedCorr	FedNoRo	FedELC
Time (s)	47	285	358	411
GPU (GB)	3.02	1.56	15.28	2.32

Hyperparameter Sensitivity. We evaluate the robustness of FedCoIn with respect to key hyperparameters, including $(K_{\min}, K_{\max})$, δ, and m_j, under the CIFAR-10 non-IID setting with $\rho = 1$ and $\tau = 0.5$. As shown in Figure 6, the performance remains stable across a wide range of parameter values, indicating that the proposed method is not sensitive to precise hyperparameter tuning.

Ablation Study. Table 4 presents the ablation results of individual components in FedCoIn under both IID and non-IID settings, where we evaluate the impact of removing Logit Adjustment, DADP, CoIn, and NADA, respectively. In contrast, Table 8 further investigates the interactions between different components under the CIFAR-100 (IID) setting by removing pairs of modules. It can be observed that removing any single component consistently leads to performance degradation across different noise levels and data partitions, indicating the importance of each module.



(a) (b) (c)

Figure 6: Sensitivity analysis of hyperparameters (m_j), $(K_{\min}, K_{\max})$, and δ on CIFAR-10 under non-IID setting with $\rho = 1$ and $\tau = 0.5$.

In particular, removing CoIn results in the most significant performance drop, indicating that collaborative inference plays a central role in leveraging cross-client information for noisy label mitigation. Removing DADP also leads to notable degradation, especially under higher noise ratios, suggesting that distribution-aware priors are crucial for stabilizing label inference under severe noise conditions. In addition, Logit Adjustment and NADA provide stable improvements across settings, highlighting their complementary roles in handling class imbalance and adaptive aggregation.

Furthermore, results from pairwise ablations further confirm strong interactions among components. Removing any two modules consistently leads to more severe performance degradation compared to single-component removal, especially for combinations involving CoIn, DADP, or NADA. This indicates that these components are not independent but closely coupled, exhibiting strong synergistic effects.

Overall, these results demonstrate that the effectiveness of FedCoIn stems not only from the contribution of individual components, but also from their coordinated interactions.

Table 8: Ablation study of different components under noisy settings on CIFAR-100 (IID).

Modules				IID	
Logit	DADP	CoIn	NADA	$\rho = 0.5, \tau = 0.3$	$\rho = 1, \tau = 0.5$
✗	✗	✓	✓	85.09	83.92
✗	✓	✗	✓	84.12	83.02
✗	✓	✓	✗	84.97	83.71
✓	✗	✗	✓	84.04	82.76
✓	✗	✓	✗	84.72	83.52
✓	✓	✗	✗	84.18	83.05