FedHAN: Robust Federated Learning under Model Heterogeneity and Label Noise

Supplementary Appendix

I. Theoretical Analysis

We provide theoretical insights into how the reliability mechanism reduces the impact of noisy labels and why FedHAN remains stable. The analysis is conducted under simplifying assumptions, such as symmetric label noise, bounded loss functions, and smooth objectives, which are standard for tractability in nonconvex FL theory. These assumptions allow us to formalize the effect of the design, while our experiments later validate that the benefits extend beyond this restricted setting.

Lemma 1 (Noise Reduction Bound)**.** *Consider a dataset with class-symmetric label noise at rate $r \in [0, 1)$ over C classes. Each sample x receives a reliability weight*

$$w(x) = \exp\big(-\mathrm{JSD}(\sigma_x \| \sigma_{x^{\mathrm{aug}}})\big), \tag{1}$$

where σ_x and $\sigma_{x^{\mathrm{aug}}}$ are model predictions on original and augmented views. Define the weighted empirical distribution as

$$\tilde{P}_w = \frac{(1-r)\mathbb{E}[w \mid \textit{clean}]P_c + r\mathbb{E}[w \mid \textit{noisy}]P_n}{(1-r)\mathbb{E}[w \mid \textit{clean}] + r\mathbb{E}[w \mid \textit{noisy}]}, \tag{2}$$

where P_c and P_n are the clean and noisy conditional distributions. If the loss function ℓ satisfies $|\ell(\theta, y_1) - \ell(\theta, y_2)| \le B$ for all θ, y_1, y_2, then

$$\big|\mathbb{E}_{\tilde{P}_w}[\ell(\theta)] - \mathbb{E}_{P_c}[\ell(\theta)]\big| \le B \cdot r_{\textit{eff}}, \tag{3}$$

where the effective noise rate is

$$r_{\textit{eff}} = \frac{r\mathbb{E}[w \mid \textit{noisy}]}{(1-r)\mathbb{E}[w \mid \textit{clean}] + r \cdot \mathbb{E}[w \mid \textit{noisy}]}. \tag{4}$$

Furthermore, if there exists $\delta > 0$ such that $\mathbb{E}[w \mid \textit{clean}] \ge e^{\delta}\mathbb{E}[w \mid \textit{noisy}]$, then

$$r_{\textit{eff}} \le \frac{r}{(1-r)e^{\delta} + r} < r. \tag{5}$$

Proof. Let $D = (1-r)\mathbb{E}[w \mid \text{clean}] + r\mathbb{E}[w \mid \text{noisy}]$ and $\alpha = (1-r) \cdot \mathbb{E}[w \mid \text{clean}]/D$. By definition of the weighted distribution, $\tilde{P}_w = \alpha P_c + (1-\alpha)P_n$, hence

$$\mathbb{E}_{\tilde{P}_w}[\ell(\theta)] - \mathbb{E}_{P_c}[\ell(\theta)] = (1-\alpha)\big(\mathbb{E}_{P_n}[\ell(\theta)] - \mathbb{E}_{P_c}[\ell(\theta)]\big). \tag{6}$$

Under class-symmetric noise, the distributions P_c and P_n have identical feature marginals but differ only in label assignment. Given the bounded difference condition $|\ell(\theta, x, y_1) - \ell(\theta, x, y_2)| \le B$ for all x, y_1, y_2, we obtain

$$|\mathbb{E}_{P_n}[\ell(\theta)] - \mathbb{E}_{P_c}[\ell(\theta)]| \le B. \tag{7}$$

Therefore,

$$\big|\mathbb{E}_{\tilde{P}_w}[\ell(\theta)] - \mathbb{E}_{P_c}[\ell(\theta)]\big| \le B(1-\alpha), \tag{8}$$

where

$$1 - \alpha = \frac{r\mathbb{E}[w \mid \text{noisy}]}{(1-r)\mathbb{E}[w \mid \text{clean}] + r\mathbb{E}[w \mid \text{noisy}]} = r_{\text{eff}}. \tag{9}$$

For the conditional bound, if $\mathbb{E}[w \mid \text{clean}] \ge e^{\delta}\mathbb{E}[w \mid \text{noisy}]$, then

$$r_{\text{eff}} = \frac{r}{(1-r)\frac{\mathbb{E}[w|\text{clean}]}{\mathbb{E}[w|\text{noisy}]} + r} \le \frac{r}{(1-r)e^{\delta} + r} < r. \tag{10}$$

□

The separation factor δ represents the degree to which clean samples receive higher expected weights than noisy ones. When $\delta > 0$, the bound shows that the effective noise rate becomes strictly smaller than the original rate. Thus, reliability weighting helps when clean samples tend to receive larger weights on average. We estimate this separation empirically in Section III-B.

Lemma 2 (Convergence Rate of FedHAN)**.** *Assume each local objective F_i is L-smooth, stochastic gradients have bounded second moments, and reliability weights are bounded. Let g_t be FedHAN's aggregated gradient with bias $\|\mathbb{E}[g_t] - \nabla F(\theta_g^t)\| \le \epsilon_w$, where θ_g^t is the global parameter at round t. With step sizes $\eta = \Theta(1/\sqrt{TE})$ for E local epochs and $\eta_g = \Theta(1/\sqrt{TT_c})$ for T_c refinement steps, FedHAN satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla F(\theta_g^t)\|^2\big] = \mathcal{O}\Big(\tfrac{1}{\sqrt{T}} + \epsilon_w^2\Big), \tag{11}$$

where T is the number of communication rounds and the hidden constants may depend on E and T_c.

Proof. By L-smoothness,

$$F(\theta_g^{t+1}) \le F(\theta_g^t) - \eta_{\min}\langle \nabla F(\theta_g^t), g_t\rangle + \tfrac{L}{2}\eta_{\min}^2\|g_t\|^2.$$

Taking expectations and using $\|\mathbb{E}[g_t] - \nabla F(\theta_g^t)\| \le \epsilon_w$ together with $\mathbb{E}\|g_t\|^2 \le G^2$,

$$\mathbb{E}[F(\theta_g^{t+1})] \le F(\theta_g^t) - \eta_{\min}\,\mathbb{E}\big[\langle \nabla F(\theta_g^t), \mathbb{E}[g_t]\rangle\big] + \tfrac{L}{2}\eta_{\min}^2 G^2.$$

Decomposing the inner product and applying Young's inequality,

$$\mathbb{E}\big[\langle \nabla F(\theta_g^t), \mathbb{E}[g_t]\rangle\big] \ge \tfrac{1}{2}\,\mathbb{E}\big[\|\nabla F(\theta_g^t)\|^2\big] - \tfrac{1}{2}\epsilon_w^2.$$

Hence

$$\mathbb{E}[F(\theta_g^{t+1})] \le F(\theta_g^t) - \tfrac{\eta_{\min}}{2}\,\mathbb{E}\big[\|\nabla F(\theta_g^t)\|^2\big] + \tfrac{L}{2}\eta_{\min}^2 G^2 + \tfrac{\eta_{\min}}{2}\epsilon_w^2.$$

Summing over $t = 0$ to $T - 1$ and rearranging,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla F(\theta_g^t)\|^2\big] \le \frac{2\big(F(\theta_g^0) - F^*\big)}{\eta_{\min} T} + LG^2\eta_{\min} + \epsilon_w^2,$$

where the effective step size is $\eta_{\min} = \min(\eta E, \eta_g T_c)$. With $\eta = \Theta(1/\sqrt{TE})$ and $\eta_g = \Theta(1/\sqrt{TT_c})$, and treating E, T_c as constants, we obtain $\eta_{\min} = \Theta(1/\sqrt{T})$, so the bound is $\mathcal{O}(1/\sqrt{T} + \epsilon_w^2)$. □

Lemma 2 indicates that, under the stated assumptions, FedHAN achieves the standard $\mathcal{O}(1/\sqrt{T})$ federated learning convergence rate when reliability weighting effectively separates clean and noisy samples ($\epsilon_w \to 0$), while the rate degrades proportionally to ϵ_w^2 as separation quality decreases. This suggests that FedHAN's noise-robust mechanisms can preserve theoretical guarantees in ideal cases, while also emphasizing that practical performance depends critically on the effectiveness of consistency-based sample weighting.

II. Details of Datasets and Baselines

Datasets. The following datasets are used in our experiments:

- **MNIST**: A dataset of handwritten digits with 60,000 training samples and 10,000 test samples across 10 classes.
- **CIFAR-10**: A dataset consisting of 60,000 32x32 color images in 10 classes, with 50,000 training images and 10,000 test images.
- **CIFAR-100**: Similar to CIFAR-10 but with 100 classes containing 600 images each, with 500 training images and 100 test images per class.

Baselines. Below are the descriptions of the compared baselines for heterogeneous federated learning:

- **FedMD [1]**: A model-heterogeneous federated learning approach using knowledge distillation with soft labels for knowledge transfer.
- **FedClassAvg [2]**: Aggregates classifier weights from clients with heterogeneous model architectures to align decision boundaries while maintaining local feature representations.
- **RHFL [3]**: An extension of FedMD that optimizes local training and the soft label aggregation at the central server against noisy clients.
- **FedProto [4]**: Uses prototypical learning to align local and global representations among heterogeneous clients.
- **FlexiFed [5]**: Handles architecture heterogeneity among clients by aggregating common layers while preserving personalized ones, maximizing knowledge sharing.
- **Fed2PKD [6]**: Uses prototypical contrastive distillation to align local embeddings with global prototypes and semi-supervised global distillation to incorporate global data in heterogeneous FL.

III. More Experimental Results

A. Convergence under Label Noise

Convergence behavior is shown in Fig. 1. FedHAN not only achieves superior final accuracy but also converges faster and more stably, particularly on the CIFAR-10 and CIFAR-100 datasets. Under high noise ($r = 0.3$), baselines like FedMD and FedProto show unstable or plateaued trajectories, while FedHAN continues to improve. On MNIST, FedHAN converges at a similar rate to FlexiFed, though the latter reaches slightly higher plateaus.

B. Empirical Estimation of δ

To validate Lemma 1, we estimate the separation factor δ under class-symmetric label noise, where the clean and noisy subsets are known by construction. Using the reliability weights $w(x)$, we compute the empirical estimate $\hat{\delta}$ in each communication round and report the trajectories in Figure 2 for CIFAR-10 and CIFAR-100 across multiple noise rates. As shown in Figure 2(a), $\hat{\delta}$ begins near zero and steadily increases before stabilizing at a positive value. This indicates that, as training progresses, clean samples are assigned systematically larger weights than noisy ones, consistent with the condition assumed in Lemma 1. Although the separation is not perfect, the persistent positivity of $\hat{\delta}$ demonstrates that the reliability mechanism biases weights toward clean samples.

We further compute the empirical effective noise rate r_{eff} as defined in Eq. (4). Figure 2(b) shows that r_{eff} consistently falls below the original corruption level r across datasets and noise rates. These results confirm the theoretical claim that reliability weighting reduces the effective influence of mislabeled data in practice.

C. Empirical Verification of Complexity

Table I shows that FedHAN incurs slightly higher runtime than FedProto but remains comparable to RHFL, adding about 10% client-side cost due to extra computation. In terms of communication, FedHAN is comparable to RHFL: FedHAN transmits class prototypes and classifier parameters, whereas RHFL transmits logits of public datasets, so both avoid sending full model parameters. These results indicate that FedHAN achieves robustness improvements with moderate and manageable overhead.

TABLE I
Runtime (s/round, with 5 clients) and communication (MB/client/round, upload+download) overhead.

Dataset	Method	Runtime (s/round)	Comm. (MB/round/client)
CIFAR-10	RHFL [3]	221.3	0.19
	FedProto [4]	201.1	0.13
	FedHAN (Ours)	208.9	0.19
CIFAR-100	RHFL [3]	233.1	1.90
	FedProto [4]	203.2	1.25
	FedHAN (Ours)	211.9	1.91

References

[1] D. Li and J. Wang, "FedMD: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

[2] J. Jang, H. Ha, D. Jung, and S. Yoon, "FedClassAvg: Local representation learning for personalized federated learning on heterogeneous neural networks," in *Proceedings of the 51st International Conference on Parallel Processing*, 2023.

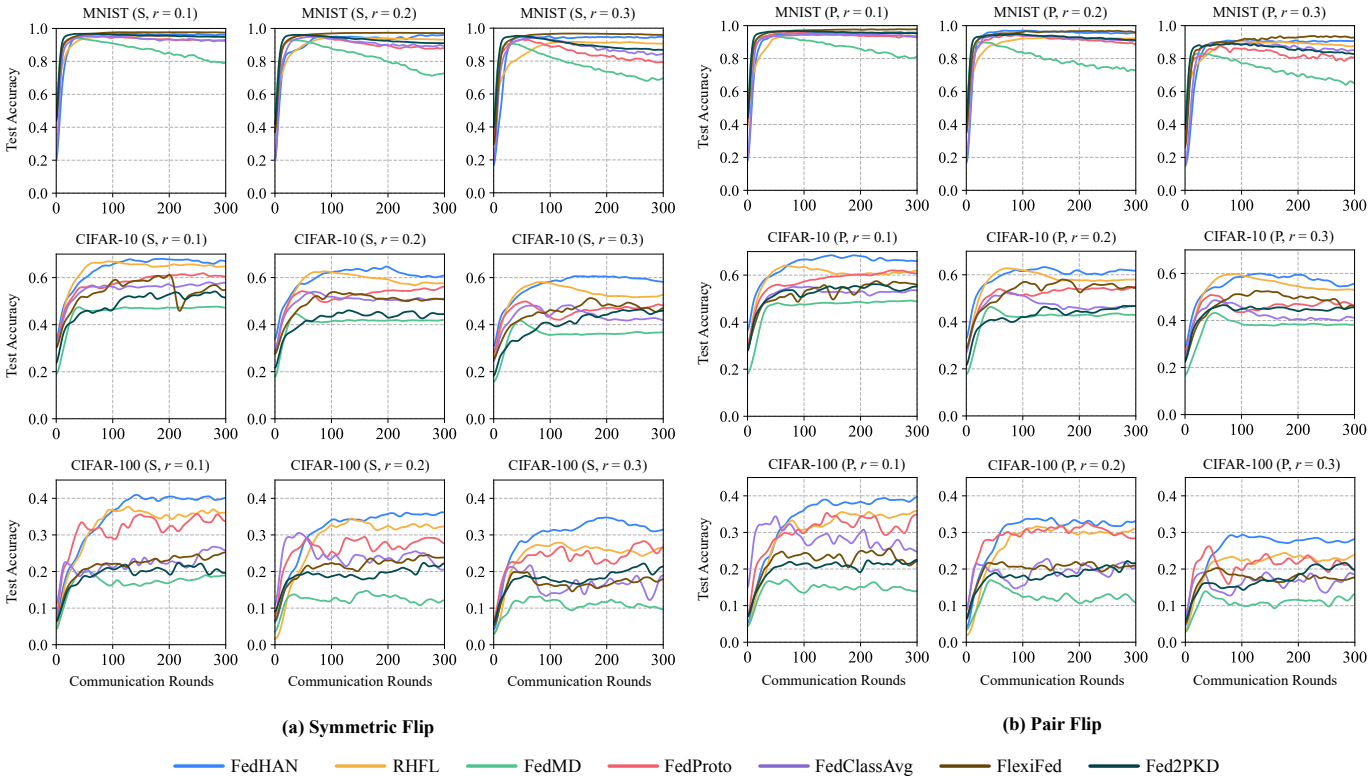

Fig. 1. Convergence of FedHAN and six baselines on MNIST, CIFAR-10, and CIFAR-100 under symmetric flip (a) and pair flip (b) with noise rates $r \in \{0.1, 0.2, 0.3\}$. FedHAN consistently achieves faster and more stable convergence across communication rounds.

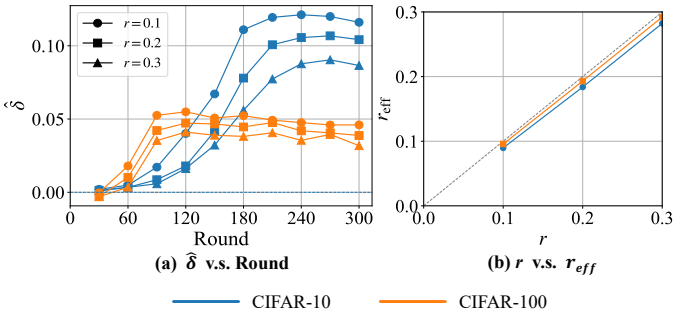

Fig. 2. Empirical validation of Lemma 1. (a) The separation factor $\hat{\delta}$ rises and stabilizes positive, indicating higher weights for clean samples. (b) The effective noise rate r_{eff} stays below r, confirming reduced label noise impact.

[3] X. Fang and M. Ye, "Robust federated learning with noisy and heterogeneous clients," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 10 072–10 081.

[4] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "FedProto: Federated prototype learning across heterogeneous clients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8432–8440.

[5] K. Wang, Q. He, F. Chen, C. Chen, F. Huang, H. Jin, and Y. Yang, "FlexiFed: Personalized federated learning for edge clients with heterogeneous model architectures," in *Proceedings of the ACM Web Conference 2023*, 2023, p. 2979–2990.

[6] Z. Xie, H. Xu, X. Gao, J. Jiang, and R. Han, "Fed2PKD: Bridging model diversity in federated learning via two-pronged knowledge distillation," in *2024 IEEE 17th International Conference on Cloud Computing*, 2024, pp. 1–11.