Federated Learning with Efficient Local Adaptation for Realized Volatility Prediction

Anonymous authors
Paper under double-blind review

Abstract

Financial markets present unique challenges for Federated Learning (FL) due to fragmented datasets, dynamic participation, and the critical need for precise and reliable predictions. Isolated local datasets often fail to capture the full spectrum of market dynamics, blocking accurate realized volatility predictions. Unlike traditional FL methods that focus on improving convergence during the training process, we propose Federated Learning with Adaptive Robustness and Efficiency for Local Adaptation (FLARE-LA), a novel framework designed to optimize predictive performance after the global training phase. FLARE-LA leverages Taylor-based local linearization and probabilistic optimization to efficiently adapt global models to local data distributions, enabling fast responsiveness to new market conditions. This adaptability ensures trained local models align with real-world scenarios, making FLARE-LA particularly suited to dynamic financial applications. Extensive experimental evaluations demonstrate FLARE-LA's superior performance, achieving a mean loss of 7.358×10^{-5}, VaR95% of 2.284×10^{-4}, and CVaR95% of 3.978×10^{-4} with an order of magnitude improvement over state-of-the-art FL algorithms. The results highlight FLARE-LA's unique ability to enhance post-FL performance, setting a new standard for FL in financial forecasting and other high-stakes, rapidly evolving domains.

1 Introduction

Predicting realized volatility is a cornerstone of financial forecasting, essential for effective risk management and informed investment strategies within the framework of deep hedging Buehler et al. (2019)Vuletić & Cont (2023)Mueller et al. (2024). However, financial markets naturally generate fragmented and asynchronous data across multiple trading venues. These platforms are unable to share data with third parties due to stringent privacy concerns, regulatory constraints, and technical challenges Kairouz et al. (2021). The data fragmentation poses significant obstacles to the accuracy and reliability of realized volatility predictions. When data is distributed across multiple platforms, the collective market understanding becomes incomplete, limiting the ability to accurately capture price movements and liquidity dynamics. Volatility prediction, which depends on comprehensive market data, is particularly vulnerable to inaccuracies and biases under these conditions. Furthermore, discrepancies in liquidity levels and pricing for the same asset across exchanges exacerbate these challenges, potentially leading to incorrect volatility estimates when one platform's data fails to reflect broader market trends Otero (2002)Madhavan (2000). Federated Learning (FL) offers a promising solution by enabling collaborative model training across distributed data sources while preserving data privacy, which addresses key privacy and regulatory concerns by ensuring data remains local to each trading platform Yang et al. (2019)Yu et al. (2020)Tan et al. (2022)Chen et al. (2023a)Meng et al. (2024).

Financial markets impose uniquely demanding requirements on FL due to the critical need for high precision, reliability, and robustness. Unlike other domains, where minor inaccuracies may be tolerable, small errors in financial forecasting can result in significant financial losses or missed opportunities Ning et al. (2023). Volatility prediction, in particular, presents a highly challenging task as it requires models that not only identify complex and rapidly evolving trends in market behavior but also provide robust and interpretable risk estimates Bergeron et al. (2021). The fragmented nature of financial data intensifies these challenges. Trading platforms operate independently, each generating datasets that reflect its unique market conditions, liquidity levels, and trading behaviors. The data decentralization introduces substantial obstacles to achieving consistent, high-quality predictions across platforms.

While data heterogeneity and dynamic participation are common challenges in FL, their impact is amplified in financial markets. The inherent variability in datasets across trading platforms leads to significant discrepancies in local and global data distributions. Additionally, trading platforms frequently enter and exit the training process due to operational constraints, creating a dynamic and unpredictable training environment. Maintaining robustness under such conditions is essential to ensuring consistent model performance across all participants. Beyond these technical challenges, the necessity for interpretable and trustworthy predictions is particularly acute in financial applications. Models must go beyond delivering accurate forecasts, which must also quantify uncertainties effectively, enabling informed decision-making in high-stakes environments. Traditional FL methods often lack the precision, adaptability, and interpretability required for such applications, limiting their practical utility in financial forecasting.

To address these challenges, we introduce Federated Learning with Adaptive Robustness and Efficiency for Local Adaptation (FLARE-LA), a cutting-edge framework designed to address the distinct requirements of financial markets. FLARE-LA utilizes Taylor-based linearization to achieve computationally efficient and accurate local adaptations, effectively aligning the global model with platform-specific datasets. Moreover, it incorporates a probabilistic mechanism that leverages the Jacobian matrix of the global model, facilitating localized optimization and delivering interpretable uncertainty quantification to enhance reliability and support informed decision-making. This integrated approach ensures robust performance in dynamic and fragmented environments while maintaining computational efficiency. By blending global insights with fine-tuned local adjustments, FLARE-LA substantially improves the accuracy of realized volatility predictions, satisfying the stringent precision and reliability demands of financial forecasting.

Although FLARE-LA is rigorously evaluated within the context of financial markets, its underlying principles, such as efficient local adaptation and uncertainty-aware predictions, are broadly applicable to a wide range of domains. The financial market serves as a challenging and representative scenario that underscores the framework's capabilities, providing valuable insights into its potential for other complex and dynamic environments. Experimental evaluations demonstrate that FLARE-LA consistently outperforms state-of-the-art baselines, achieving lower mean loss, Value at Risk (VaR95%), and Conditional Value at Risk (CVaR95%) across diverse platform distributions and participation rates. These results highlight the framework's scalability, adaptability, and robustness, making it well-suited for general federated learning tasks where data heterogeneity, dynamic participation, and computational efficiency are critical.

In the following sections, we provide an overview of related work in Section 2, positioning our contributions within the broader landscape of FL and financial forecasting. Section 3 introduces the fragmented nature of financial markets and formulates the problem addressed by our framework. In Section 4, we detail our proposed approach, highlighting the integration of probabilistic frameworks and efficient local adaptation techniques. Empirical evaluations of our method, showcasing its effectiveness and robustness across various scenarios, are detailed in Section 5. Finally, we summarize our findings, and outline potential directions for future research in Section 6.

2 Related Work

In the context of financial markets, predicting realized volatility using order book data is a challenging task due to the decentralized nature of data acquisition Banabilah et al. (2022). Order books, which capture buy and sell orders for securities, form a dynamic and fragmented data environment, often referred to as "data islands." These characteristics make FL an appealing approach for such environments Hasbrouck (2007). FL enables collaborative model training across distributed data sources while preserving data privacy, a critical requirement in financial markets.

Existing FL methods face specific challenges when applied to predicting realized volatility in financial markets, such as high data heterogeneity, rapid changes in data patterns, and the need for timely model updates. Methods like FedProx Li et al. (2020), which introduces a proximal term to the local objective function to stabilize optimization, mitigate the effects of heterogeneity by reducing the impact of local updates that deviate significantly from the global model. However, while FedProx offers stability, its proximal term may not fully capture the dynamic nature of financial data, and it can slow convergence, a critical drawback in fast-paced trading environments Arthur et al. (2018)Cantillon & Yin (2011). To address local-global mis-alignments, SCAFFOLD Karimireddy et al. (2020) incorporates control variates to correct the drift in local updates, improving alignment with the global model. However, the rapidly evolving financial landscape can still lead to misalignments that adversely affect prediction accuracy Boukherouaa et al. (2021). While effective in certain scenarios, the introduction of control variates increases computational complexity and communication overhead, posing challenges for deployment in high-frequency trading environments.

Personalized FL methods, such as FedPer Arivazhagan et al. (2019), decouple shared global parameters from client-specific local parameters to provide personalization. Despite this, the high variability and unpredictability in financial markets require frequent adjustments to personalized models, making the process resource-intensive and limiting scalability in large-scale financial networks. Similarly, LG-FedAvg Liang et al. (2020), APFL Deng et al. (2020), and pFedMe T Dinh et al. (2020) adopt approaches to balance global and local knowledge but face challenges in handling the feature and distributional heterogeneity prevalent in financial markets.

Recent innovations, including Ditto Li et al. (2021), FedRep Collins et al. (2021), and SuPerFed Hahn et al. (2022), have explored various personalization techniques, such as interpolating global and local models or applying proximity regularization. These methods highlight progress in adapting global models to client-specific data. However, they often require fine-tuning or additional computational resources for new clients, reducing scalability in highly dynamic environments like financial trading. Meng et al. (2024) explores techniques to enhance global generalization and local personalization through adaptive aggregation and dual optimization, which aligns with our goal of striking a balance between global insights and local adaptations in heterogeneous FL settings. While their work focuses on representation learning and aggregation strategies, our method introduces a Taylor-based linearization approach combined with a probabilistic framework to achieve more precise and interpretable local adaptation. Tan et al. (2022) provides insights into handling client-specific model updates using personalized layers and meta-learning, offering a solution for improving local performance in non-IID settings. This is relevant to our method in FLARE-LA, which also aims to achieve strong local adaptation but does so through efficient linearized updates and probabilistic adjustments, eliminating the need for additional network layers or meta-learning components. Chen et al. (2023a) focuses on sparse model adaptation to enhance scalability and computational efficiency in personalized FL, a strategy particularly useful in resource-constrained environments. Similarly, Yu et al. (2020) highlights the importance of localized training adjustments to address the limitations of federated aggregation in heterogeneous datasets. Both approaches emphasize the need for effective local training, which resonates with our method's focus on dynamic participation and efficient adaptation. However, our approach extends these ideas by leveraging Jacobian-based linearization and uncertainty quantification, enabling robust local updates tailored to the fragmented and rapidly changing nature of financial data.

Advances in neural network behavior further inspire solutions for FL. Research has revealed that infinitely wide deep neural networks (DNNs) exhibit behaviors similar to their Taylor expansions around initialization Chizat et al. (2019). Extensions of this analysis to finite-width DNNs demonstrate that their training dynamics resemble linear models Seleznova & Kutyniok (2022), while the inductive biases of linearized neural networks effectively summarize full network functions Maddox et al. (2021). These findings motivate the development of FL approaches that better address the unique characteristics of local trading platforms, where global models often fail to capture localized data patterns.

To address these challenges, the proposed FLARE-LA framework introduces adaptive local training mechanisms that go beyond traditional FL approaches by focusing on post-training performance optimization. FLARE-LA leverages insights from neural network linearization to enable precise and computationally efficient local adaptations, ensuring that global models are effectively refined to meet the specific needs of individual trading platforms. Unlike existing methods, FLARE-LA is designed to address the dynamic and heterogeneous nature of financial markets by rapidly adapting to new data and evolving conditions, ensuring that local models remain robust and aligned with real-world scenarios. The proposed innovative approach is a transformative solution for achieving high-precision and reliable predictions in decentralized and fragmented financial environments.

3 Fragmented Financial Markets

3.1 Background

In financial markets, trading occurs across a wide range of exchanges and platforms, resulting in highly fragmented datasets. Each platform independently maintains transaction and order book data, capturing buy and sell orders as well as their execution details. This fragmentation provides an incomplete view of market activity for any given asset, with notable variations in pricing, liquidity, and order depth across platforms Hasbrouck (2007).

The order book plays a critical role in market analysis, offering traders insights into short-term trading dynamics. By displaying order imbalances and identifying potential support and resistance levels for a stock, the order book supports informed trading decisions. Heightened market activity and uncertainty are often reflected in increased realized volatility, which arises from frequent directional price movements. Trading data, which records executed transactions, complements the order book by offering valuable insights into market dynamics, such as price trends, trading volumes, and liquidity conditions.

Predicting short-term realized volatility is essential for effective risk management and the development of trading strategies Chen et al. (2023b). By analyzing order book and trade data over fixed time intervals, traders and institutions can forecast future volatility levels, enabling improved decision-making and enhanced risk mitigation. Realized volatility predictions help market participants manage exposure, optimize portfolio allocations, and design robust trading strategies.

Extracting meaningful insights from order book data is vital for understanding market dynamics and assessing stock values. Key metrics, such as the bid-ask spread, weighted average price, and volume-related indicators, provide a wealth of information about market liquidity and potential volatility. However, the fragmented nature of financial markets poses significant challenges for comprehensive analysis, as data silos limit access to the full scope of market activity.

3.2 Problem Formulation

By leveraging diverse data sources, FL facilitates the development of a robust global model that enhances local predictions, preserving both data privacy and confidentiality, which allows trading platforms to benefit from a comprehensive understanding of market dynamics while maintaining compliance with regulatory requirements and addressing privacy concerns.

Consider a distributed dataset consisting of n data sample pairs $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ across $|E|$ trading platforms. Each data sample pair represents features extracted from order book and trading data, with $\mathbf{x}_i$ denoting the feature vector and y_i representing the corresponding label, which is the volatility. There are 363 features for each sample generated from order book and trading data, capturing essential market dynamics such as bid-ask spreads, price movements, and trading volumes.

We denote the local dataset of the c-th trading platform as P_c, which contains n_c training samples. The union of all local datasets from each trading platform, $P_1 \cup P_2 \cup \cdots \cup P_{|E|}$, covers the entire dataset, guaranteeing that each sample is assigned to exactly one trading platform's dataset. For trading platform c, the labels $\{y_i\}_{i \in P_c}$ represent the volatility levels observed in the corresponding platform's trading data. These volatility labels are used as the ground truth for training the predictive model.

We aim to develop a predictive model, represented by a deep neural network function f, which maps an input feature vector $\mathbf{x}$ to an output volatility prediction y. The model is trained using the distributed dataset across multiple trading platforms, leveraging the features extracted from order book and trading data to predict future volatility levels accurately. The local objective function for trading platform c is defined as

$$\underset{\boldsymbol{w}}{\textbf{minimize}} \quad \boldsymbol{L_c}(\boldsymbol{w}) = \frac{1}{2} \sum_{i \in P_c} \left(f(\boldsymbol{x_i}, \boldsymbol{w}) - y_i \right)^2 \quad \text{for } c = 1, \cdots, |E|, \tag{1}$$

where $\boldsymbol{w}$ represents the trainable parameters of the model. Meanwhile, the global objective function, aggregating the local objectives across all trading platforms, is given by

$$\underset{\boldsymbol{w}}{\textbf{minimize}} \quad \boldsymbol{L}(\boldsymbol{w}) = \frac{1}{|E|} \sum_{c=1}^{|E|} \boldsymbol{L_c}(\boldsymbol{w}). \tag{2}$$

Realized volatility prediction poses unique challenges due to the fragmented nature of financial data, and the rapidly evolving market conditions. Each platform's dataset captures only a localized perspective of the broader market, leading to non-IID data distributions that complicate the development of a unified model. Furthermore, the predictive task demands a model capable of capturing complex, non-linear interactions between features to provide reliable and interpretable outputs. By leveraging FL, trading platforms can train a global model that integrates diverse data sources, improving predictive accuracy while maintaining data privacy. This collaborative framework enables financial institutions to optimize their trading strategies, enhance risk management, and make informed decisions in volatile market conditions.

4 Federated Learning with Adaptive Robustness and Efficiency for Local Adaptation

In this section, we introduce our proposed approach FLARE-LA, a framework designed to address the challenges posed by heterogeneous local datasets and dynamic participation in financial markets. While the global objective function in (2) captures general patterns across all trading platforms, it may fail to fully represent the unique characteristics of each platform's local dataset, potentially resulting in suboptimal performance for individual platforms as outlined in (1). To overcome this limitation, FLARE-LA provides an innovative mechanism for trading platforms to adapt the globally trained model to their specific local data.

The FLARE-LA framework operates through a two-stage process. In the initial stage, trading platforms collaboratively train a global model while ensuring strict data privacy. This is accomplished by aggregating model updates from each platform without transmitting raw data, thereby preserving confidentiality and compliance with privacy regulations. The global model, built from the collective knowledge of all participating platforms, effectively captures shared patterns and structures across the distributed datasets. This stage provides a robust baseline for subsequent local adaptation. Moreover, FLARE-LA is designed to seamlessly integrate advanced federated learning techniques into this collaborative training stage, enhancing flexibility and adaptability to various use cases.

In the second stage, FLARE-LA introduces an innovative local adaptation mechanism that fine-tunes the globally trained model to align with the specific characteristics of each trading platform's dataset. This process leverages Taylor-based linearization and probabilistic frameworks to achieve computational efficiency and precision. By utilizing the Jacobian matrix of the global model, FLARE-LA integrates localized optimization with interpretable uncertainty quantification, enabling platforms to adapt the global model dynamically while maintaining robustness in predictive accuracy. This approach ensures that FLARE-LA excels in addressing the challenges of non-IID data distributions, enhancing model performance in diverse and fragmented environments.

By combining the strengths of FL with tailored local adaptations, FLARE-LA effectively addresses the inherent heterogeneity of financial market datasets and accommodates the dynamic participation of trading platforms. This dual-stage approach ensures that each platform benefits from the collaborative insights of FL while achieving optimal performance for its specific market conditions.

4.1 Federated Training equipped with Efficient Local Adaptation for Financial Market Dynamics

The initialization of model weights plays a pivotal role in determining the efficiency and stability of the training process. Arbitrary or poorly chosen initialization methods can hinder progress, leading to issues such as slow convergence or training stagnation Xie et al. (2017). To address these challenges, it is essential to ensure that the activation distributions maintain consistent variance as the network deepens, preventing common pitfalls like vanishing or exploding gradients.

To achieve this, the initial weights are drawn from a Gaussian distribution with a mean of zero and a standard deviation inversely proportional to the square root of the number of input units feeding into the layer, which can be expressed as

$$\boldsymbol{w^0} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{n_{\text{in}}}}\right), \tag{3}$$

where $\boldsymbol{w^0}$ denotes the initial weight vector, and n_{in} represents the number of input units in the layer. This tailored initialization ensures a balanced variance in the activation distributions across layers, fostering a smoother gradient flow and more stable training dynamics.

Federated training unfolds in a dynamically evolving environment where the participation of trading platforms fluctuates unpredictably. At each training round, a subset of trading platforms, denoted as $S^t \subseteq |E|$, is selected to participate. This dynamic subset mirrors real-world scenarios where platform availability is influenced by operational constraints, market activity, or other factors. To simulate such dynamic participation, the set S^t is sampled from a predefined distribution. In this work, we explore several distributions, including Exponential, Geometric, Gamma, and Chi-square, to capture a variety of participation patterns.

The dynamic nature of platform participation introduces additional complexity to the federated training process, as the global model must adapt to fluctuating contributions without compromising performance. By incorporating realistic participation patterns into our simulation, we ensure that the training procedure reflects the challenges of real-world financial environments, enhancing the robustness and applicability of our approach.

Upon determining the active participants for round t, the current global model $\boldsymbol{w^t}$ is distributed to the selected trading platforms in S^t. Each platform initializes its local model for the training round as

$$\{\boldsymbol{w^t_{c,0}} = \boldsymbol{w^t}\}_{c \in S^t}, \tag{4}$$

where $\boldsymbol{w^t_{c,0}}$ represents the initial local model weights for trading platform c at the onset of round t. This initialization ensures that all participating platforms begin the round with identical copies of the global model, fostering a collaborative and unified starting point in the dynamic participation environment.

During local training on trading platform c, which involves financial market data, the model undergoes iterative updates. The k-th step of this update process is defined as

$$\boldsymbol{w^t_{c,k+1}} = \boldsymbol{w^t_{c,k}} - \alpha_l \boldsymbol{\nabla L_c}(\boldsymbol{w^t_{c,k}}), \tag{5}$$

where α_l denotes the local learning rate, which is specifically tailored to the unique dynamics and characteristics of each platform's dataset. This localized learning process allows each platform to refine the model in alignment with its own market conditions.

The local training procedure continues for K iterations, resulting in a final local model given by

$$\boldsymbol{w^t_{c,K}} = \boldsymbol{w^t} - \sum_{k=1}^{K} \alpha_l \boldsymbol{\nabla L_c}(\boldsymbol{w^t_{c,k}}), \tag{6}$$

which integrates the cumulative effects of gradient-based updates performed over all local steps. This formulation highlights how each platform adapts the global model to its specific data through weighted gradient descents.

To quantify the divergence between the locally adapted model and the initial global model, we define the model discrepancy for trading platform c after K iterations as

$$\triangle \boldsymbol{w^t_c} = \boldsymbol{w^t_{c,K}} - \boldsymbol{w^t}. \tag{7}$$

This term measures the extent to which each platform's local updates diverge from the global model parameters, reflecting the influence of its unique market data on the learning process.

The aggregated local updates are used to compute the global model for the next iteration, as follows

$$\boldsymbol{w^{t+1}} \leftarrow \boldsymbol{w^t} + \frac{\alpha_g^t}{|S^t|} \sum_{c \in S^t} \triangle \boldsymbol{w^t_c}, \tag{8}$$

where α_g^t is the global learning rate for round t, and the contribution of each local model is normalized by the number of participating platforms $|S^t|$. This normalization ensures equitable integration of local updates into the global model, promoting fairness and robustness across platforms. The updated global model $\boldsymbol{w}^{\boldsymbol{t+1}}$ marks the conclusion of the t-th round of training and serves as the starting point for the next round of FL.

The above iterative process allows FLARE-LA to adaptively refine the global model by incorporating diverse contributions from participating platforms within the dynamic and heterogeneous environment of financial markets. Crucially, the federated iterations in FLARE-LA are designed to be modular, enabling the seamless integration of any advanced FL solutions. This flexibility enhances the scalability and generalization of the framework, allowing it to adapt to evolving methods and leverage state-of-the-art advancements in FL. By combining tailored local training with equitable aggregation, FLARE-LA effectively addresses the challenges of data heterogeneity and fluctuating participation rates, ensuring robust performance and broad applicability.

The global model $\boldsymbol{w}^*$, obtained after FL training, may not be fully optimized or may exhibit poor local performance due to the diverse nature of local datasets and the dynamic participation. Nonetheless, it serves as the baseline for adaptive local training. To derive the local adaptive training strategy, we consider a given neural network model function f. We can approximate f around the trained model parameters $\boldsymbol{w}^*$ using a Taylor expansion

$$f(x;\boldsymbol{w}) \approx f(x;\boldsymbol{w}^*) + J_{\boldsymbol{w}^*}(\boldsymbol{x})^T(\boldsymbol{w}-\boldsymbol{w}^*), \tag{9}$$

where $J_{\boldsymbol{w}^*}(\boldsymbol{x})$ denotes the Jacobian matrix of partial derivatives of f with respect to the model parameters at $\boldsymbol{w}^*$, with dimensions $p \times |P_c|$. This Jacobian represents the sensitivity of the output with respect to changes in the model parameters near $\boldsymbol{w}^*$.

We formulate the probabilistic model governing the output y, given input features x extracted from order book and trading data, and model parameters $\boldsymbol{w}$ as

$$p(y \,|\, x, \boldsymbol{w}) = \mathcal{N}\left(f(\boldsymbol{x};\boldsymbol{w}), \sigma_c^2\right) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(y-f(\boldsymbol{x};\boldsymbol{w}))^2}{2\sigma_c^2}}, \tag{10}$$

where σ_c^2 represents the variance associated with the Gaussian noise, capturing the inherent uncertainty and noise in the model predictions of volatility. This distribution's mean is specified by the linear approximation obtained from the Taylor expansion of f, with a variance σ_c^2.

For volatility prediction in financial markets using FL, deviations from the baseline global model $\boldsymbol{w}^*$ influence the mean prediction through the Jacobian adjustment, while the Gaussian term $\mathcal{N}(0, \sigma_c^2)$ accounts for the stochastic nature of the predictions. This framework establishes a robust basis for trading platforms to adapt and retrain the global model locally, ensuring performance optimization tailored to the unique characteristics of individual datasets.

For each trading platform c with its local dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{|P_c|}$, the likelihood function quantifies the probability of observing the given data. It incorporates both the individual variances from the Gaussian noise and the deviations of the model predictions from actual data points. This integration is captured by the model's output and its linear approximation around $\boldsymbol{w}^*$ which is formulated as

$$P_c(\boldsymbol{w}) = \frac{1}{(2\pi\sigma_c^2)^{\frac{|P_c|}{2}}} \exp\left(-\frac{1}{2\sigma_c^2}\sum_{i=1}^{|P_c|}(y_i - (f(\boldsymbol{x}_i;\boldsymbol{w}^*) + J_{\boldsymbol{w}^*}(\boldsymbol{x}_i)^T(\boldsymbol{w}-\boldsymbol{w}^*)))^2\right). \tag{11}$$

This formulation enables trading platforms to effectively assess the fit between their local data and the global model, guiding them in refining the model parameters to better capture the underlying patterns in volatility dynamics.

For rapid local adaptation within our financial market volatility prediction, we transform the likelihood function into its logarithmic form as

$$\log(P_c(\boldsymbol{w})) = -\frac{|P_c|}{2}\log(2\pi\sigma_c^2) - \frac{1}{2\sigma_c^2}\sum_{i=1}^{|P_c|}(y_i - (f(\boldsymbol{x}_i;\boldsymbol{w}^*) + J_{\boldsymbol{w}^*}(\boldsymbol{x}_i)^T(\boldsymbol{w}-\boldsymbol{w}^*)))^2, \tag{12}$$

which simplifies the expression by converting the product of probabilities into a sum of logarithms, linearizing the effects of the parameters and enhancing the tractability of the optimization problem. Notably, $-\frac{1}{2\sigma_c^2}\sum_{i=1}^{|P_c|}(y_i - f(\boldsymbol{x}_i;\boldsymbol{w}))^2$ represents the sum of squared residuals, adjusted by the inverse of the noise variance σ_c^2.

Therefore, the local adaptation process can be formulated as minimizing the following loss function

$$\hat{L}_c(\boldsymbol{w}) = \frac{1}{2\sigma_c^2}\sum_{i=1}^{|P_c|}(y_i - (f(\boldsymbol{x}_i;\boldsymbol{w}^*) + J_{\boldsymbol{w}^*}(\boldsymbol{x}_i)^T(\boldsymbol{w}-\boldsymbol{w}^*)))^2 + \frac{|P_c|}{2}\log(2\pi\sigma_c^2), \tag{13}$$

which comprises a term that evaluates the sum of squared deviations between the predicted volatility and the actual volatility, scaled by the noise variance σ_c^2, and a constant term that standardizes the loss based on the dataset size and noise level in the context of local financial market data.

We define $\boldsymbol{J_{w^*}} = \{\boldsymbol{J_{w^*}(x_i)}\}_{i=1}^{P_c}$ as the collection of Jacobian matrices of the model's predictions with respect to the features generated from order book and trading data, evaluated at $\boldsymbol{w}^*$. The sum of the outer products of these Jacobian matrices across all data points forms a symmetric matrix, which can be expressed as

$$\sum_{i=1}^{|P_c|}\boldsymbol{J_{w^*}(x_i)J_{w^*}(x_i)^T} = \boldsymbol{J_{w^*}J_{w^*}^T}, \tag{14}$$

which reflects the covariance structure of the gradients, capturing the sensitivity of the model's predictions to the features derived from the trading platforms' data. To facilitate a clearer understanding and to simplify computations in practice, this loss function can be reformulated as

$$\begin{aligned}\hat{L}_c(\boldsymbol{w}) = (\boldsymbol{w}-\boldsymbol{w}^*)^T\frac{1}{2\sigma_c^2}\boldsymbol{J_{w^*}J_{w^*}^T}(\boldsymbol{w}-\boldsymbol{w}^*) - (\boldsymbol{w}-\boldsymbol{w}^*)^T\frac{1}{\sigma_c^2}\boldsymbol{J_{w^*}(y_c - f_c)}\\ + \frac{1}{2\sigma_c^2}\boldsymbol{(y_c-f_c)^T(y_c-f_c)} + \frac{|P_c|}{2}\log(2\pi\sigma_c^2),\end{aligned} \tag{15}$$

where $\boldsymbol{f_c} = \{f(\boldsymbol{x}_i;\boldsymbol{w}^*)\}_{i=1}^{P_c}$ and $\boldsymbol{y_c} = \{y_i\}_{i=1}^{P_c}$. It quantifies the balance between the model's internal predictions and the observed deviations from the actual volatility outcomes, scaled by the noise variance, σ_c^2. This local loss function is critical for adapting the global model to better fit the specific characteristics of the local trading platform's data. The local model adaptation is achieved by setting the gradient of the designed local loss function, $\boldsymbol{\nabla\hat{L}_c(w)}$, to zero as

$$\boldsymbol{\nabla\hat{L}_c(w)} = \frac{1}{\sigma_c^2}\boldsymbol{J_{w^*}J_{w^*}^T}(\boldsymbol{w}-\boldsymbol{w}^*) - \frac{1}{\sigma_c^2}\boldsymbol{J_{w^*}(y_c-f_c)} = \boldsymbol{0}. \tag{16}$$

By solving this equation, we identify the stationary point, which is typically a minimum for a well-defined convex function

$$\boldsymbol{w} = (\boldsymbol{J_{w^*}J_{w^*}^T})^{-1}\boldsymbol{J_{w^*}(y_c-f_c)} + \boldsymbol{w}^*, \tag{17}$$

which suggests that the local model adaptation is proportional to the pseudo-inverse of the aggregated Jacobian product, adjusted by the residuals between the observed volatility and the model's predicted volatility. Importantly, the term $(\boldsymbol{J_{w^*}J_{w^*}^T})^{-1}\boldsymbol{J_{w^*}}$ only needs to be computed once, providing significant computational efficiency.

When predicting for a new data sample, $\boldsymbol{x}_i$ derived from order book and trading data, the model leverages both the learned parameters and the inherent variability in observations for making predictions by following formulation

$$\hat{y}_i = f(\boldsymbol{x}_i;\boldsymbol{w}^*) + J_{\boldsymbol{w}^*}(\boldsymbol{x}_i)^T(\boldsymbol{J_{w^*}J_{w^*}^T})^{-1}\boldsymbol{J_{w^*}(y_c-f_c)} + \mathcal{N}(0;\sigma_c^2). \tag{18}$$

This formula represents the linearized update to the model's prediction, adjusted by the newly optimized parameters, and includes a Gaussian noise term, which accounts for the inherent uncertainty in the prediction. It plays a crucial role in ensuring a realistic forecast of local volatility.

By incorporating the baseline prediction using the global model parameters $f(\boldsymbol{x}_i; \boldsymbol{w}^*)$, the adjustment to the prediction based on the local training data $J_{\boldsymbol{w}^*}(\boldsymbol{x}_i)^T(\boldsymbol{J}_{\boldsymbol{w}^*}\boldsymbol{J}_{\boldsymbol{w}^*}^{\boldsymbol{T}})^{-1}\boldsymbol{J}_{\boldsymbol{w}^*}(\boldsymbol{y}_{\boldsymbol{c}} - \boldsymbol{f}_{\boldsymbol{c}})$ and the inherent variability in the predictions, we provide an adaptive approach to predicting volatility, tailored to the unique characteristics of each trading platform's data. This approach ensures that the predictions remain both accurate and robust, even in the face of dynamic and heterogeneous market conditions. The convergence analysis is shown in Appendix A.

4.2 Analysis of the FLARE-LA Approach

FLARE-LA provides an innovative solution to the critical challenges of FL in financial markets, where high precision, robustness, and adaptability are paramount. By integrating federated training with an advanced local adaptation mechanism, FLARE-LA effectively bridges the gap between collective insights from distributed datasets and the need for platform-specific optimization. This unified approach ensures that the framework addresses the complexities of fragmented, non-IID data environments and dynamic participation rates in financial markets.

The federated training phase in FLARE-LA enables collaborative learning across decentralized trading platforms, allowing the global model to capture shared patterns and insights while maintaining data privacy and regulatory compliance. This phase establishes a robust baseline model that encapsulates market-wide trends. Importantly, the federated training iterations in this phase are modular and can incorporate any advanced federated learning solutions to enhance the scalability and generalization of the framework. By accommodating diverse federated optimization techniques, FLARE-LA ensures its adaptability to evolving FL methods and diverse application scenarios.

To complement the federated training phase, FLARE-LA introduces an advanced Taylor-based linearization strategy for computationally efficient and precise local adaptations. By leveraging the Jacobian matrix of the global model, FLARE-LA approximates complex local adjustments, enabling trading platforms to quickly tailor the global model to their unique data distributions without extensive computational overhead. Additionally, the framework integrates probabilistic modeling to capture prediction uncertainties, enhancing interpretability and reliability, which is key for high-stakes financial decision-making. This modular and extensible approach ensures that FLARE-LA remains a scalable, adaptive, and generalizable framework for FL in financial markets and beyond.

5 Experimental Evaluation

This experimental evaluation validates the efficacy and adaptability of the proposed FLARE-LA framework in addressing the challenges of FL across both domain-specific and general scenarios. Our primary focus lies in the financial domain, where the demands for high precision, robustness, and scalability are particularly pronounced. We first utilize a dataset for realized volatility prediction, consisting of order book and trade data from multiple trading platforms. These experiments aim to demonstrate FLARE-LA's ability to handle extreme data heterogeneity, dynamic participation, and the fragmented nature of financial datasets while maintaining robust predictive performance.

To further evaluate the generalizability of FLARE-LA, we extend our experiments to CIFAR10 and MNIST, two well-established datasets in FL research. These datasets allow us to test FLARE-LA's performance under non-IID data distributions, varying client participation rates, and label noise scenarios, mimicking real-world challenges. By incorporating these datasets, we provide complementary evidence of FLARE-LA's versatility and scalability, demonstrating its utility across diverse applications beyond financial forecasting.

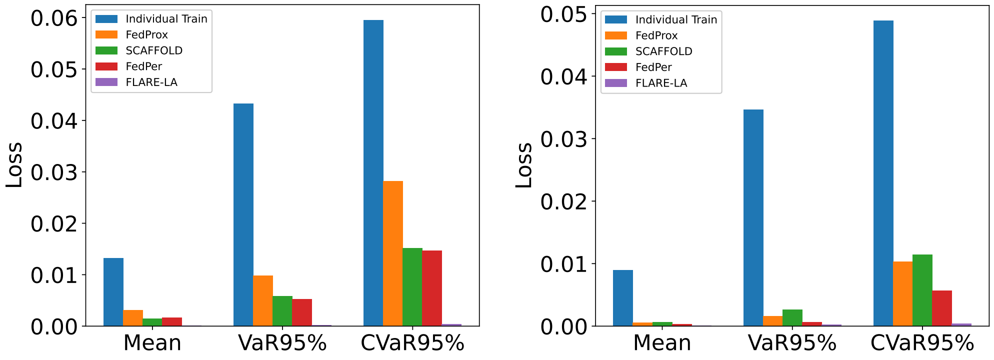
In our experiments, client participation is dynamically regulated using a participation ratio, simulating high variability in client engagement during federated training. Non-IID data distributions are modeled using a Dirichlet distribution, with the concentration parameter α controlling the heterogeneity of client data. For $\alpha \to 0$, clients primarily have data from a single class, while $\alpha \to \infty$ results in a uniform distribution of classes across clients. We evaluate the performance of FLARE-LA against several state-of-the-art FL methods, including FedProx Li et al. (2020), SCAFFOLD Karimireddy et al. (2020), FedPer Arivazhagan et al. (2019), LG-FedAvg Liang et al. (2020), pFedMe T Dinh et al. (2020), Ditto Li et al. (2021), FedRep Collins et al.

(2021), and SuPerFed Hahn et al. (2022). For local training, we utilize the ResNet model, which provides a robust architecture for handling diverse data distributions. The evaluation metrics include mean loss, Value at Risk (VaR95%), and Conditional Value at Risk (CVaR95%), offering a comprehensive assessment of model performance. These metrics underscore FLARE-LA's ability to deliver superior predictive accuracy, adapt effectively to dynamic environments, and maintain computational efficiency, even in challenging FL scenarios.

5.1 Experiments on Realized Volatility Prediction

We aim to forecast short-term volatility for stocks spanning multiple sectors Andrew Meyer (2021). The dataset comprises both order book and trade data for these stocks, aggregated into multiple time buckets. The values in the order book represent the latest snapshots of market activity, taken at one-second intervals. Each time bucket comprises order book data spanning the 600 seconds. Our experiments involve predicting the volatility for each time bucket of the stocks. There are $428,932$ samples in the entire dataset, where 107 of the stocks have data for 3830 time buckets, while 3 stocks have data for 3829 time buckets, 1 stock has data for 3820 time buckets, and another stock has data for 3815 time buckets. The entire dataset is divided into $10,000$ trading platforms based on a Dirichlet distribution-based non-IID setting Hsu et al. (2019). The Dirichlet distribution's concentration parameter, α, determines the stock distribution for each trading platform which is set to 0.5 in our experiments. Each trading platform randomly splits its data into a training set and a test set, with 20% allocated for testing. This setup allows us to estimate the performance of each FL algorithm on each trading platform's test set using its personalized model.

5.1.1 Performance Comparison



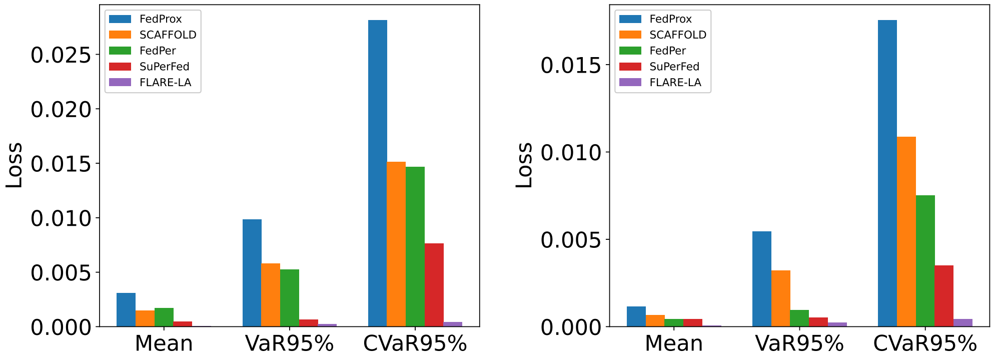
(a) 50 Epochs

(b) 200 Epochs

Figure 1: Comparison of FLARE-LA with Individual Train and other FL baselines (FedProx, SCAFFOLD, and FedPer) across different epochs, demonstrating the superiority of FLARE-LA in federated settings.

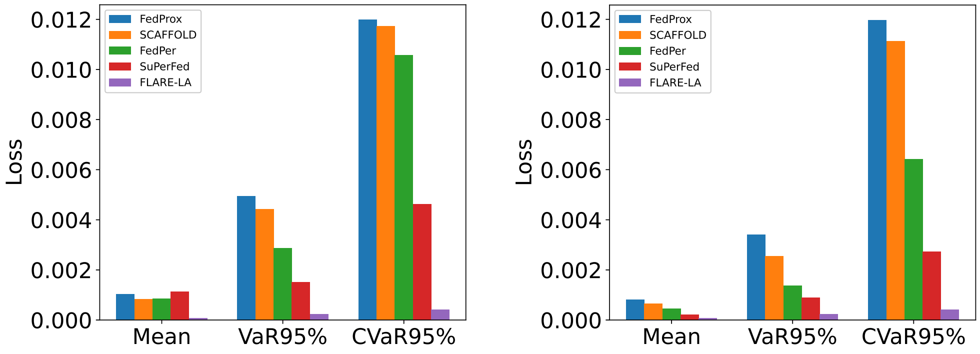
In Fig. 1, we compare the performance of FLARE-LA against Individual Train and baseline FL methods as FedProx, SCAFFOLD, and FedPer over 50 epochs as shown in Fig. 1(a) and 200 epochs as shown in Fig. 1(b). The Individual Train baseline performs training independently for each trading platform without leveraging federated collaboration. Despite identical numbers of parameter updates, FLARE-LA demonstrates significantly superior performance, emphasizing the value of FL in leveraging global insights while tailoring models to local data.

As shown in Fig. 1(a), after 50 epochs, FLARE-LA achieves a remarkably lower mean loss of 7.726×10^{-5} compared to Individual Train (0.0132), FedProx (0.0031), SCAFFOLD (0.0015), and FedPer (0.0017). In Fig. 1(b), after 200 epochs, FLARE-LA continues to outperform all baselines, maintaining its lead in terms of mean loss, VaR95%, and CVaR95%. The results highlight FLARE-LA's ability to balance global knowledge with precise local adaptation, resulting in superior performance in federated settings. This demonstrates that FLARE-LA not only accelerates convergence but also ensures higher accuracy and robustness compared to individual and baseline federated training methods.



(a) 5 Rounds, 10% Participation

(b) 20 Rounds, 10% Participation



(c) 5 Rounds, 30% Participation

(d) 20 Rounds, 30% Participation

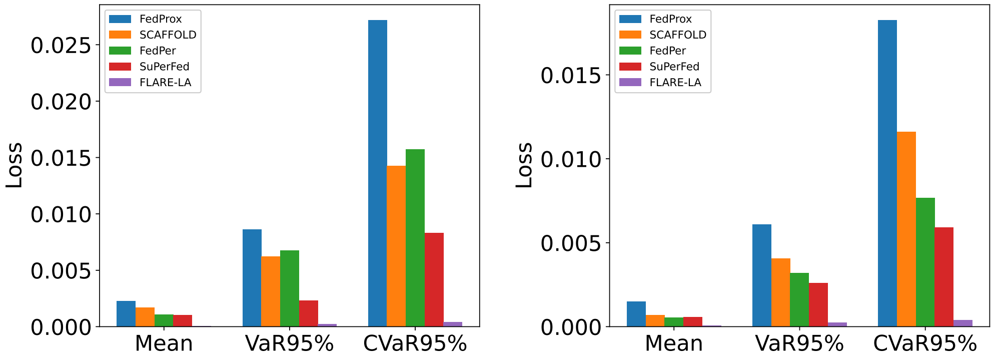
Figure 2: Performance comparison with varying participation rates and training rounds on realized volatility prediction.

Across all experimental settings as shown in Fig. 2, FLARE-LA consistently outperforms baseline methods, including FedProx, SCAFFOLD, FedPer, and SuPerFed, in terms of Mean Loss, VaR95%, and CVaR95% for realized volatility prediction tasks.

With low client participation rates as10%, FLARE-LA demonstrates exceptional robustness and precision. As shown in Fig. 2(a), after 5 training rounds, FLARE-LA achieves a Mean Loss of approximately 0.0001, compared to significantly higher values for SuPerFed (0.0008), FedPer (0.0017), SCAFFOLD (0.0031), and FedProx (0.004). The advantage becomes even more pronounced in risk-sensitive metrics such as VaR95% and CVaR95%, where FLARE-LA achieves much lower values, highlighting its ability to effectively manage tail risks even with limited trading platforms participation. These trends persist in Fig. 2(b), with 20 training rounds further consolidating FLARE-LA's dominance in all metrics.

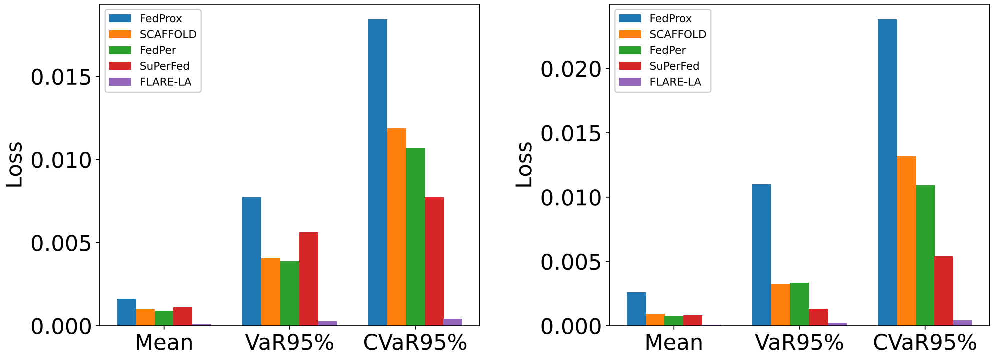
When the participation rate increases to 30% as shown in Figs. 2(c) and 2(d), the overall model performance improves across all methods. However, FLARE-LA retains a clear advantage, achieving substantially lower Mean Loss values. For instance, in Fig. 2(d), FLARE-LA reaches a Mean Loss of approximately 0.00005, outperforming SuPerFed (0.0003), FedPer (0.0005), SCAFFOLD (0.0010), and FedProx (0.002). The improvement in FLARE-LA's performance with higher participation rates underscores its ability to fully leverage the increased availability of local data while maintaining its computational efficiency and accuracy.

Furthermore, the impact of increasing the number of federated training rounds is evident. FLARE-LA demonstrates rapid convergence to low loss values within a few rounds, significantly reducing the computational burden compared to other methods. Even after just 5 training rounds, FLARE-LA achieves results comparable to or better than the baseline methods after 20 rounds, as shown in Figs. 2(b) and 2(d). This efficiency highlights FLARE-LA's capability to deliver robust performance even in scenarios with limited training rounds or participation rates.



(a) Exponential

(b) Geometric



(c) Gamma

(d) Chisquare

Figure 3: Illustrations of distributions: (a) Exponential, (b) Geometric, (c) Gamma, and (d) Chi-square.

As shown in Fig. 3, we provide the comparative analysis across various participation distributions to evaluate the efficacy of FLARE-LA in addressing the inherent challenges of FL with dynamic participation. The experiments were conducted with a 20% participation rate, 10 federated rounds, and 10 local epochs in each round. As shown in Fig. 3(a) where trading platforms are sampled from an exponential distribution, with a scale parameter of 1.0, FLARE-LA demonstrates a remarkable ability to achieve a mean loss of 7.358×10^{-5}, VaR95% of 2.284×10^{-4}, and CVaR95% of 3.978×10^{-4}, outperforming FedProx, SCAFFOLD, and FedPer by an order. As shown in Fig. 3(b), where trading platforms are sampled from a geometric distribution with a probability of success of an individual trial set at 0.35, FLARE-LA once again emerges as the top-performing algorithm. Fig. 3(c) explores the performance of algorithms when trading platforms are sampled from a Gamma distribution, with a shape parameter of 2.0 and a scale parameter of 1.0. In Fig. 3(d), where trading platforms are sampled from a chi-square distribution with the number of degrees of freedom set at 2.0, FLARE-LA continues to outshine the baseline algorithms.

By consistently delivering superior performance metrics, FLARE-LA showcases its adaptability in scenarios characterized by varying levels of data availability and participation. By consistently achieving lower mean loss, VaR95%, and CVaR95% values, FLARE-LA underscores its resilience and adaptability in optimizing federated model training across a spectrum of trading platform distributions. These results demonstrate FLARE-LA's capability in managing volatile market conditions and optimizing federated model training despite unpredictable trading platform participation patterns.

5.1.2 Computation Cost Comparison

The computation cost comparison for one round of FL is presented in Table 1, showcasing FLARE-LA's computational efficiency across varying participation rates. All experiments were conducted on an experimental

Table 1: The Computation Cost Comparison for One Round

Participation Rate	Fedprox (s)	SCAFFOLD (s)	FedPer (s)	SuPerFed (s)	FLARE-LA (s)
30%	74.39	105.9	46.18	99.63	48.16
60%	144.83	208.51	90.42	186.67	94.43

platform featuring an 8-core CPU, a 14-core GPU, and 16GB of RAM. This setup ensures consistent benchmarking across all evaluated FL methods.

At a participation rate of 30%, FLARE-LA achieves a computation time of 48.16 seconds, significantly outperforming SCAFFOLD (105.9 seconds) and SuPerFed (99.63 seconds). While FedPer demonstrates a slightly faster computation time of 46.18 seconds, its slower convergence rate necessitates more training rounds to achieve comparable results, thereby increasing the overall computational burden. FLARE-LA's superior balance between computational demands and model accuracy ensures efficient and timely model updates, even under challenging participation scenarios.

When the participation rate increases to 60%, FLARE-LA continues to excel with a computation time of 94.43 seconds, outperforming Fedprox (144.83 seconds) and SCAFFOLD (208.51 seconds) by substantial margins. Although FedPer achieves a comparable time of 90.42 seconds, FLARE-LA's faster convergence significantly reduces the total training cost, making it a more efficient and scalable solution for large-scale FL applications in financial markets.

The experimental results demonstrate FLARE-LA's robustness and adaptability in addressing challenges such as fragmented datasets and irregular client participation. By achieving competitive computation times, FLARE-LA ensures privacy-preserving collaboration and timely model updates, addressing critical requirements for FL in decentralized financial environments. Its ability to provide reliable volatility predictions is particularly valuable for effective risk management and investment decision-making in dynamic financial markets.

5.2 Experiments with CIFAR10 and MNIST

To extend our evaluation beyond the financial domain, we conducted experiments using the CIFAR10 and MNIST datasets, partitioned into 1000 clients. These datasets serve as benchmarks to demonstrate the generalizability and robustness of FLARE-LA in broader FL scenarios.

To simulate real-world challenges, we introduced artificial label noise into the training sets, employing two commonly used noise schemes: pairwise flipping Han et al. (2018) and symmetric flipping Van Rooyen et al. (2015). The pairwise flipping scheme models scenarios where labels transition to semantically similar neighboring labels with a noise ratio ϵ, while retaining the correct label with a probability of $1-\epsilon$. The symmetric flipping scheme assumes uniform mislabeling across all incorrect labels, distributing the noise ratio ϵ evenly among them, while preserving the correct label with a probability of $1-\epsilon$.

For both schemes, the test sets remain clean to ensure a fair and accurate evaluation of model performance. This setup allows us to rigorously assess FLARE-LA's ability to handle noisy labels, which is a critical capability for FL applications in dynamic and unpredictable environments. By addressing label noise effectively, FLARE-LA demonstrates its robustness and adaptability in diverse scenarios, further validating its utility across domains.

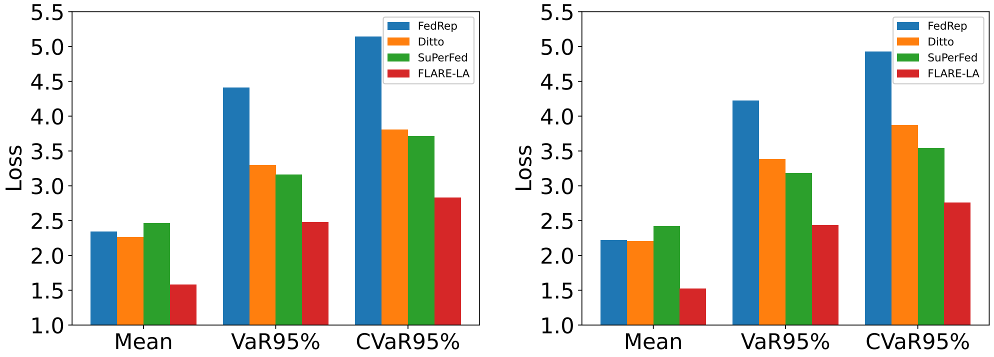
5.2.1 Convergence Analysis

The experiments on the CIFAR10 dataset assess the convergence performance of FLARE-LA in comparison to state-of-the-art FL methods, including FedRep, Ditto, and SuPerFed, across various training rounds. Fig. 4 presents the model performance in terms of Mean Loss, VaR95%, and CVaR95% after 300, 400, 500, and 600 training rounds, highlighting the consistent superiority of FLARE-LA, particularly in handling non-IID data and dynamic participation.

At 300 training rounds as shown in Fig. 4(a), FLARE-LA achieves a Mean Loss of approximately 1.2, significantly outperforming SuPerFed (2.1), Ditto (2.7), and FedRep (4.3). FLARE-LA also demonstrates a clear advantage in VaR95% and CVaR95%, showcasing its ability to effectively manage tail risks in the early training stages.

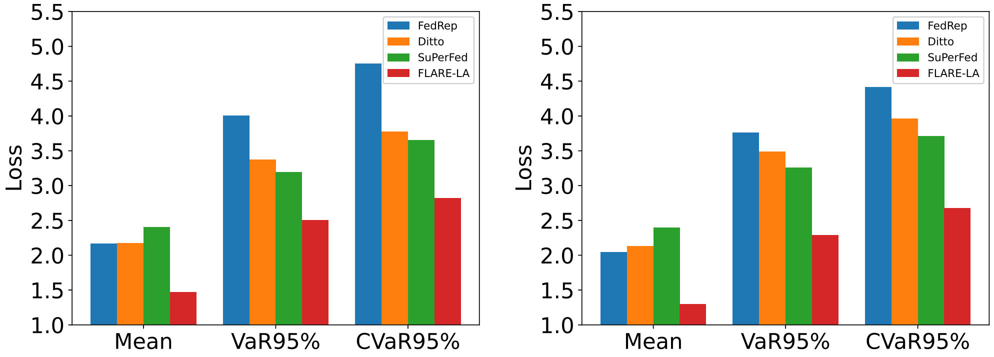
As training progresses to 400 and 500 rounds shown in Figs. 4(b) and 4(c), FLARE-LA maintains its lead across all metrics. By 600 training rounds as shown in Fig. 4(d), FLARE-LA achieves a Mean Loss of approximately 0.8, solidifying its position as the best-performing method. The consistent reduction in VaR95% and CVaR95% further emphasizes FLARE-LA's robustness and reliability in FL settings.

Compared to SuPerFed, Ditto, and FedRep, FLARE-LA demonstrates faster convergence and superior overall performance. While SuPerFed performs competitively, it lags in handling label noise and dynamic participation. Ditto, despite its strength in personalization, struggles to balance global generalization with local adaptation. FedRep, on the other hand, exhibits slower convergence due to less effective local adaptation mechanisms.



(a) 300 Rounds

(b) 400 Rounds



(c) 500 Rounds

(d) 600 Rounds

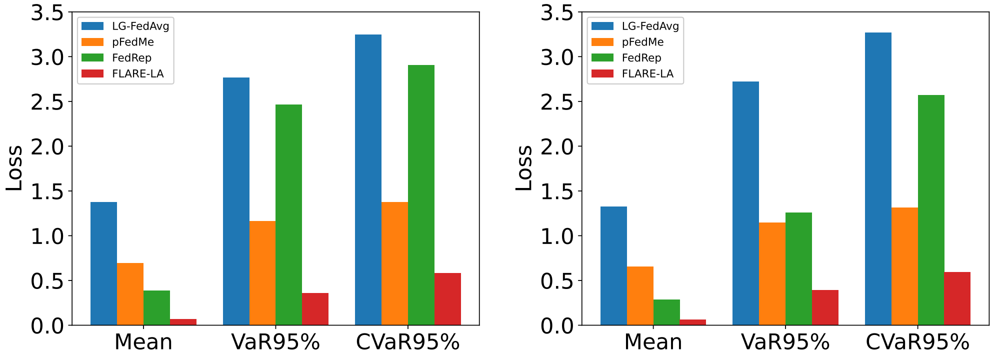
Figure 4: Convergence analysis of model performance across different federated training rounds using the CIFAR10 dataset.

Fig. 5 showcases the convergence performance of FLARE-LA compared to LG-FedAvg, pFedMe, and FedRep across various federated training rounds using the MNIST dataset. These experiments evaluate FLARE-LA's ability to handle FL challenges, such as data heterogeneity and dynamic participation, while maintaining robust convergence with MNIST dataset.

At 300 training rounds as shown in Fig. 5(a), FLARE-LA achieves the lowest Mean Loss and significantly reduced VaR95% and CVaR95% compared to the baseline methods. Specifically, FLARE-LA attains a Mean Loss of approximately 0.3, whereas LG-FedAvg and FedRep exhibit higher values around 1.0 and 0.8, respectively. Although pFedMe performs competitively, it falls short of FLARE-LA's efficiency in managing early-stage convergence, highlighting the superior adaptability of FLARE-LA to non-IID data distributions.

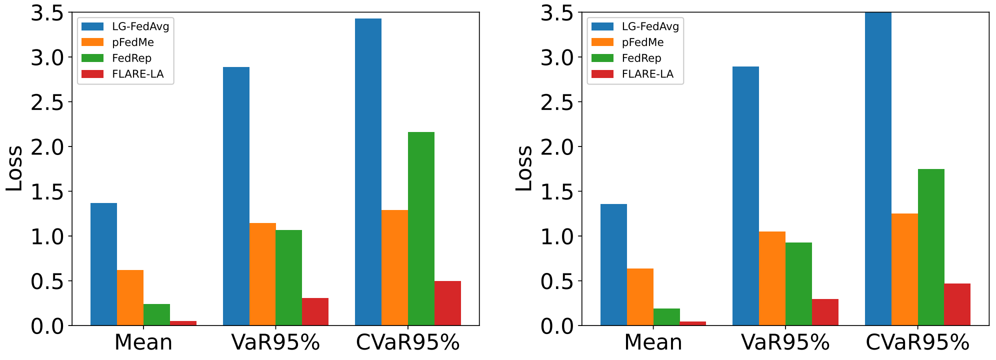
As the training progresses to 400 and 500 rounds as shown in Figs. 5(b) and 5(c), FLARE-LA's advantage becomes more pronounced. By 500 rounds, FLARE-LA achieves a Mean Loss of approximately 0.2, whereas LG-FedAvg and FedRep remain at 0.7 and 0.6, respectively. While pFedMe narrows the gap slightly, it continues to lag behind FLARE-LA, particularly in managing tail risks, as indicated by higher VaR95% and CVaR95% values. These results emphasize FLARE-LA's robustness in both accuracy and risk management, supported by its efficient local adaptation mechanism.

At 600 training rounds as shown in Fig. 5(d), FLARE-LA maintains its dominance across all metrics, achieving a Mean Loss of approximately 0.1, compared to 0.5 for FedRep and 0.6 for LG-FedAvg. The continued reduction in VaR95% and CVaR95% underscores FLARE-LA's capability to ensure precise and stable predictions, even under prolonged training scenarios. This performance demonstrates FLARE-LA's scalability and ability to sustain improvements over extended training, making it highly suitable for applications requiring prolonged federated training.



(a) 300 Rounds

(b) 400 Rounds



(c) 500 Rounds

(d) 600 Rounds

Figure 5: Convergence analysis of model performance across different federated training rounds using the MNIST dataset.

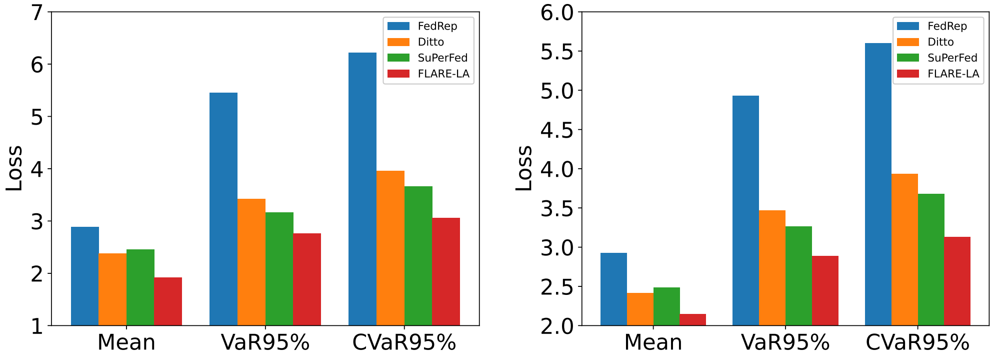
5.2.2 Analysis of Label Noise

Fig. 6 evaluates the robustness of FLARE-LA under symmetric label noise conditions, comparing its performance against FedRep, Ditto, and SuPerFed. These experiments examine noise levels ranging from 0.2 to 0.8, demonstrating the resilience of FLARE-LA in maintaining strong performance under progressively challenging noisy label scenarios.

At a noise ratio of 0.2 as shown in Fig. 6(a), FLARE-LA achieves the best performance among all methods, with a significantly lower Mean Loss, VaR95%, and CVaR95%. Specifically, FLARE-LA reports a Mean Loss of approximately 1.5, outperforming SuPerFed (2.3), Ditto (2.0), and FedRep (3.2). This demonstrates FLARE-LA's ability to handle modest label noise while preserving high predictive accuracy and robust risk estimates.

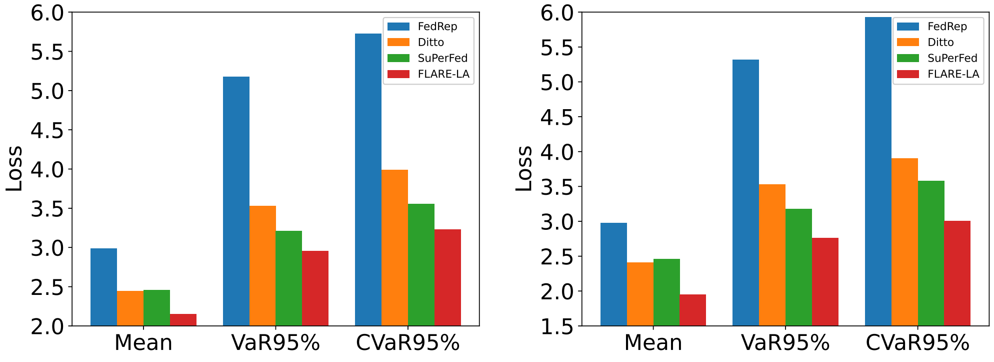
As the noise ratio increases to 0.4 as shown in Fig. 6(b), FLARE-LA continues to demonstrate a competitive edge, maintaining its superior performance across all metrics. For instance, FLARE-LA achieves a Mean Loss of approximately 2.0, while Ditto and SuPerFed report higher losses around 2.5 and 3.0, respectively. FedRep experiences further degradation, with a Mean Loss exceeding 4.0, highlighting its vulnerability to intermediate noise levels.

At higher noise ratios of 0.6 and 0.8 as shown in Figs. 6(c) and 6(d), FLARE-LA's resilience becomes even more pronounced. At a noise ratio of 0.6, FLARE-LA achieves a VaR95% of 2.8, compared to SuPerFed (3.5), Ditto (3.8), and FedRep (5.0). When the noise ratio reaches 0.8, FLARE-LA demonstrates remarkable robustness, maintaining strong performance despite challenging conditions. In contrast, the baseline methods exhibit notable performance degradation, highlighting their limited ability to handle severe label noise. This comparison underscores FLARE-LA's superior adaptability and resilience in mitigating the adverse effects of extreme noise scenarios.



(a) Noise Ratio 0.2

(b) Noise Ratio 0.4



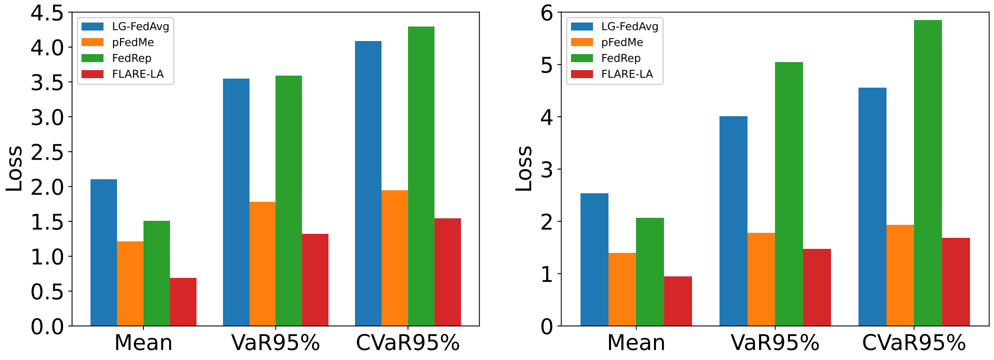
(c) Noise Ratio 0.6

(d) Noise Ratio 0.8

Figure 6: Impact of varying symmetric label noise rates on model performance during federated training with the CIFAR10 dataset.

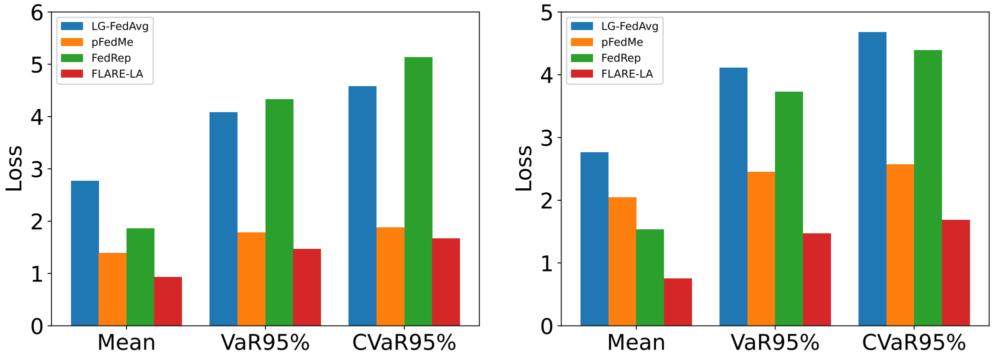
Fig. 7 evaluates the robustness of FLARE-LA under symmetric label noise conditions, comparing its performance against LG-FedAvg, pFedMe, and FedRep with the MNIST dataset. The experiments simulate varying levels of label noise, testing FLARE-LA's resilience in handling noisy labels effectively.

At a noise ratio of 0.2, FLARE-LA demonstrates clear superiority over the baseline methods, achieving robust performance and effectively managing risks even under moderate label noise. This result highlights the framework's ability to mitigate the adverse effects of noise through its probabilistic local adaptation mechanism. As the noise ratio increases to 0.4, FLARE-LA continues to maintain its competitive edge, outperforming the baselines in both predictive accuracy and risk-related metrics. Its adaptability ensures that the framework remains reliable even as noise levels increase. At higher noise levels, such as 0.6 and 0.8, the gap between FLARE-LA and the baseline methods widens significantly. FLARE-LA sustains strong performance, while the baselines exhibit considerable degradation under extreme noise conditions. These findings underscore FLARE-LA's resilience and its ability to handle challenging scenarios with noisy labels effectively using the MNIST dataset.



(a) Noise Ratio 0.2

(b) Noise Ratio 0.4



(c) Noise Ratio 0.6

(d) Noise Ratio 0.8

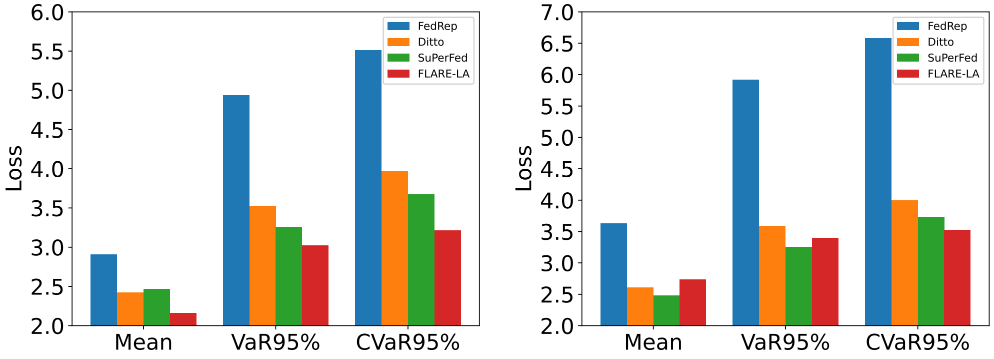
Figure 7: Impact of varying symmetric label noise rates on model performance during federated training with the MNIST dataset.

Fig. 8 presents the results of federated training on the CIFAR10 dataset under pairwise label noise conditions. The experiments compare the performance of FLARE-LA against FedRep, Ditto, and SuPerFed, with noise ratios set to 0.2 and 0.6. These results evaluate the resilience of the models in scenarios where labels are systematically flipped within similar classes, reflecting more structured and challenging noise patterns.

At a noise ratio of 0.2 as shown in Fig. 8(a), FLARE-LA achieves superior performance, with a Mean Loss of approximately 2.2, compared to approximate 2.5 for both SuPerFed and Ditto, and 3.0 for FedRep. In addition, FLARE-LA achieves substantially lower VaR95% and CVaR95% values, highlighting its ability to manage risk effectively under moderate noise conditions. The probabilistic local adaptation mechanism in FLARE-LA, which incorporates Jacobian-driven updates, enables it to leverage the structure of local data distributions, mitigating the adverse effects of mislabeled data. As the noise ratio increases to 0.6 as shown in Fig. 8(b), all methods experience performance degradation due to the increased label noise. However, FLARE-LA continues to outperform the baselines, achieving a CVaR95% of approximately 3.5, compared to 3.8 for SuPerFed, 4.1 for Ditto, and 6.6 for FedRep.

5.2.3 Analysis of Local Data Heterogeneity

Fig. 9 illustrates the impact of local data heterogeneity on model performance during federated training with the CIFAR10 dataset. The experiments evaluate two heterogeneity settings, controlled by the Dirichlet distribution's concentration parameter $\alpha = 10.0$, representing moderate non-IID data as shown in Fig. 9(a), and $\alpha = 1000.0$, representing near-IID data as shown in Fig. 9(b). The performance of FLARE-LA is compared against baseline methods, including FedRep, Ditto, and SuPerFed, across Mean Loss, VaR95%, and CVaR95%.

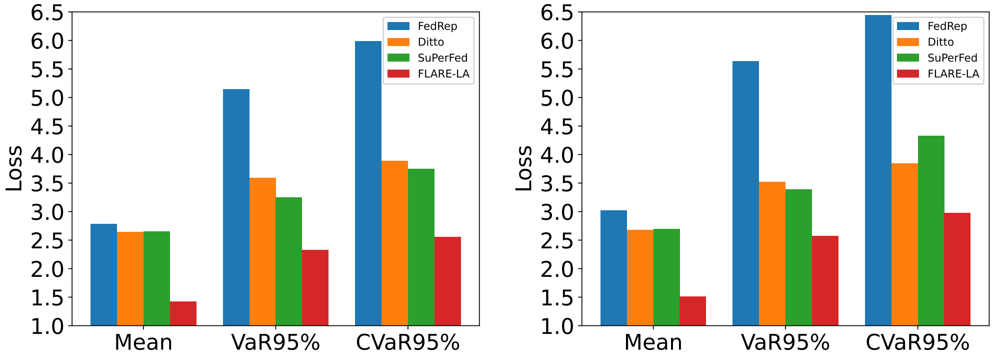


(a) Noise Ratio 0.2 (b) Noise Ratio 0.6

Figure 8: Impact of varying pairwise label noise rates on model performance during federated training with the CIFAR10 dataset.

In the moderately heterogeneous scenario with $\alpha = 10.0$, FLARE-LA significantly outperforms the baseline methods across all metrics. Specifically, FLARE-LA achieves a Mean Loss of approximately 1.2, while SuPerFed, Ditto, and FedRep report higher values above 2.5. This underscores the robustness of FLARE-LA's probabilistic local adaptation mechanism in addressing the challenges of non-IID data distributions. VaR95% and CVaR95% metrics also exhibit similar trends, with FLARE-LA consistently achieving lower values, reflecting its effectiveness in mitigating tail risks. The baseline methods, particularly SuPerFed and FedRep, demonstrate slower adaptation and less optimal performance due to their limited ability to balance local adaptation and global generalization.

In the near-IID scenario as $\alpha = 1000.0$, the reduced heterogeneity leads to improved performance across all methods. However, FLARE-LA maintains its performance advantage, achieving the lowest Mean Loss at approximately 1.5, compared to 2.6 for both SuPerFed and Ditto, and 3.0 for FedRep. VaR95% and CVaR95% metrics further highlight FLARE-LA's superior convergence and risk management capabilities. These results demonstrate FLARE-LA's adaptability to near-IID settings while continuing to outperform its competitors with the CIFAR10 dataset.



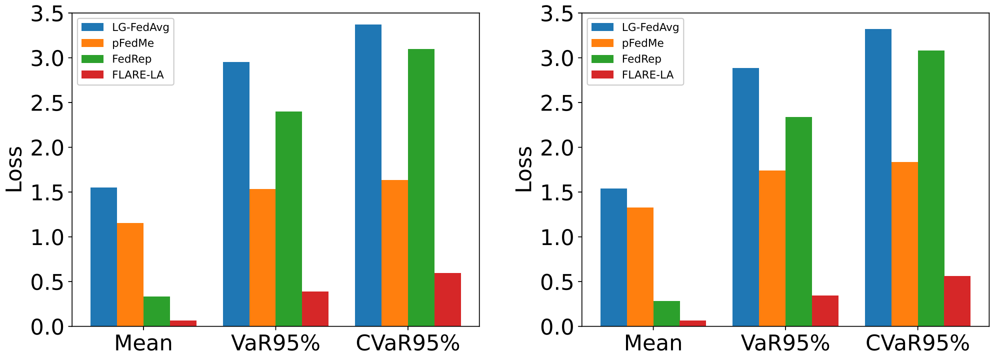
(a) $\alpha = 10.0$ (b) $\alpha = 1000.0$

Figure 9: Impact of varying degrees of local data heterogeneity on model performance during federated training with the CIFAR10 dataset.

Fig. 10 examines the impact of local data heterogeneity on model performance during federated training with the MNIST dataset. The heterogeneity is controlled using the Dirichlet distribution's concentration parameter α, where smaller values of α correspond to higher heterogeneity, and larger values indicate more homogeneous distributions across clients.

In Fig. 10(a), with $\alpha = 10.0$, representing a highly heterogeneous data setting, FLARE-LA significantly outperforms baseline methods, including LG-FedAvg, pFedMe, and FedRep. FLARE-LA achieves the lowest mean loss, demonstrating its ability to address the challenges posed by diverse data distributions across clients. Its probabilistic adaptation framework and Taylor-based linearization allow it to adapt the global model effectively to local data, ensuring robust and accurate predictions. The VaR95% and CVaR95% values further highlight FLARE-LA's superior risk management capabilities, with consistently lower values compared to the baselines. In contrast, LG-FedAvg and FedRep struggle to generalize effectively in such non-IID conditions, while pFedMe achieves competitive performance but lags behind FLARE-LA in convergence speed and accuracy.

In Fig. 10(b), with $\alpha = 1000.0$, the data distribution becomes more homogeneous across clients, resembling an IID-like setting. Although the performance gap between FLARE-LA and the baselines narrows, FLARE-LA continues to deliver the best results, particularly in terms of mean loss and risk metrics. This consistency demonstrates FLARE-LA's ability to generalize effectively even in less heterogeneous scenarios. The reduced advantage of FLARE-LA in this setting is expected, as the uniformity in data distribution reduces the need for sophisticated local adaptation. Nonetheless, FLARE-LA exhibits computational efficiency by achieving strong performance with fewer training rounds compared to the baseline methods.



(a) $\alpha = 10.0$

(b) $\alpha = 1000.0$

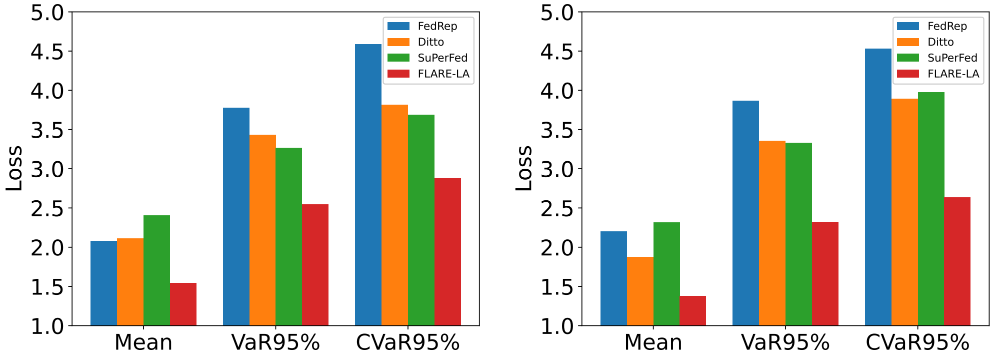
Figure 10: Impact of varying degrees of local data heterogeneity on model performance during federated training with the MNIST dataset.

5.2.4 Analysis of Client Participation Rates

Fig. 11 examines the impact of varying client participation rates on model performance during federated training with the CIFAR10 dataset. Two participation rates are evaluated, i.e., 1% and 2%, reflecting scenarios where only a small subset of clients is involved in each round of training. These experiments assess the ability of FLARE-LA to handle high dynamicity in client participation compared to FedRep, Ditto, and SuPerFed.

At the 1% participation rate as shown in Fig. 11(a), FLARE-LA demonstrates remarkable robustness and stability. Its performance significantly outpaces the baseline methods, underscoring its effectiveness in scenarios with limited client engagement. The probabilistic local adaptation mechanism in FLARE-LA allows participating clients to derive maximum benefit from the global model while effectively adapting it to their local data, even under extreme participation constraints. In contrast, FedRep, Ditto, and SuPerFed struggle to maintain stable performance, likely due to their reliance on higher client participation for effective aggregation.

With a 2% participation rate as shown in Fig. 11(b), FLARE-LA continues to outperform the baselines, achieving faster convergence and higher predictive accuracy. The increase in participation improves performance across all methods; however, FLARE-LA maintains its competitive edge. Its efficient local adaptation strategy ensures that even with a slightly larger subset of clients, the model remains robust to dynamic participation.



(a) 1% Participation Rate

(b) 2% Participation Rate

Figure 11: Impact of varying client participation rates on model performance during federated training with the CIFAR10 dataset.

Fig. 12 explores the impact of varying client participation rates on the performance of FLARE-LA, LG-FedAvg, pFedMe, and FedRep during federated training on the MNIST dataset. Four participation rates are analyzed, i.e., 0.5%, 1%, 1.5%, and 2.0%. These scenarios assess the robustness of the methods under dynamic client participation, where only a small subset of clients is involved in each training round.

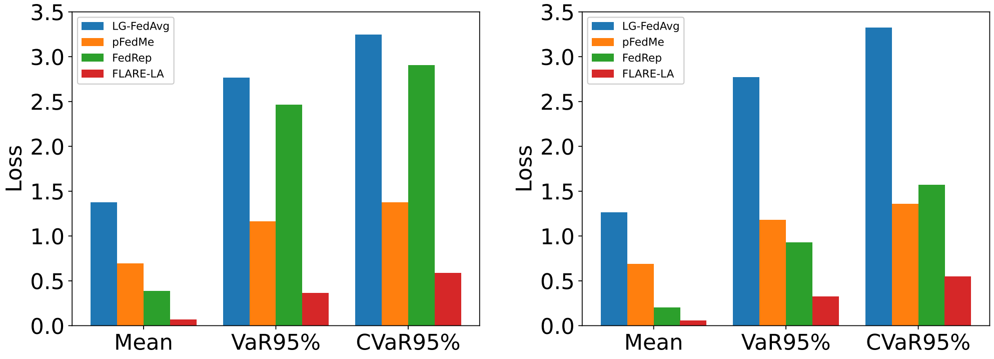
At the lowest participation rate of 0.5% as shown in Fig. 12(a), FLARE-LA achieves a mean loss of approximately 0.2, significantly outperforming LG-FedAvg (1.4), pFedMe (0.6), and FedRep (0.4). Similarly, for VaR95%, FLARE-LA records 0.2, while LG-FedAvg, pFedMe, and FedRep record 2.8, 1.2, and 2.5, respectively. In terms of CVaR95%, FLARE-LA achieves 0.5, maintaining a substantial margin over LG-FedAvg (3.2) and FedRep (3.0). These results demonstrate FLARE-LA's ability to effectively leverage limited client contributions, while the baseline methods struggle to adapt under such constraints.

As the participation rate increases to 1% as shown in Fig. 12(b), the performance of all methods improves, with FLARE-LA continuing to lead with a mean loss of 0.15, compared to LG-FedAvg (1.3), pFedMe (0.8), and FedRep (0.25). For VaR95%, FLARE-LA achieves 0.3, while LG-FedAvg and FedRep record 2.8 and 0.8, respectively. For CVaR95%, FLARE-LA remains robust at 0.5, outperforming LG-FedAvg (3.2) and pFedMe (1.3). This trend highlights FLARE-LA's superior accuracy and convergence speed as participation increases.

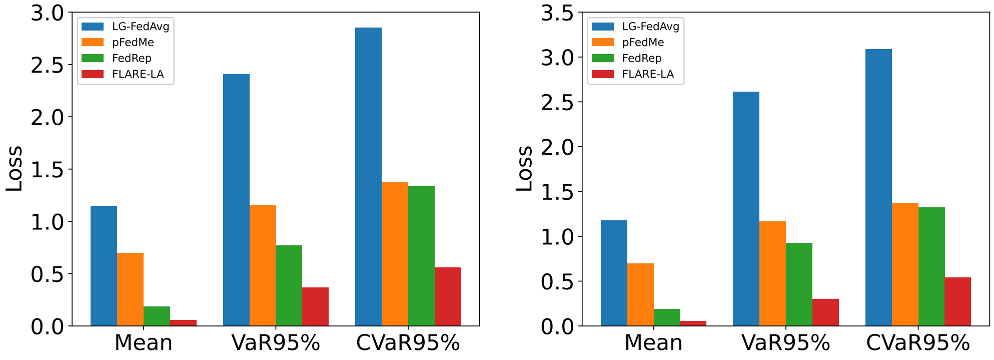
At a participation rate of 1.5% as shown in Fig. 12(c), FLARE-LA continues to demonstrate superior performance compared to LG-FedAvg, pFedMe, and FedRep. At the highest participation rate of 2.0% as shown in Fig. 12(d), FLARE-LA achieves the most favorable results among the methods evaluated. It efficiently leverages the increased client participation to deliver improved performance while maintaining robustness across key performance metrics, which highlight FLARE-LA's scalability and adaptability, underscoring its ability to outperform baseline methods consistently. Its robust local adaptation mechanism ensures effective handling of increased client engagement, maintaining strong predictive accuracy and resilience in FL scenarios.

6 Conclusions and Discussions

This work proposed FLARE-LA, a novel framework addressing the challenges of FL in diverse and dynamic environments, particularly financial markets. By integrating Taylor-based linearization for efficient local adaptation with a probabilistic mechanism leveraging the Jacobian matrix, FLARE-LA achieves precise optimization, robust performance, and interpretable uncertainty quantification. Extensive experiments demonstrated its superiority over state-of-the-art baselines, achieving higher accuracy, faster convergence, and resilience to label noise and dynamic participation. In financial applications, FLARE-LA excelled in metrics like mean loss, VaR95%, and CVaR95%, underscoring its suitability for high-stakes, heterogeneous environments. With its ability to adapt global models to local distributions, handle fragmented datasets, and ensure computational efficiency, FLARE-LA offers a scalable and versatile solution for FL challenges.



(a) 0.5% Participation Rate (b) 1% Participation Rate



(c) 1.5% Participation Rate (d) 2.0% Participation Rate

Figure 12: Impact of varying client participation rates on model performance during federated training with the MNIST dataset.

Future directions include extending the framework to domains like healthcare and IoT, integrating advanced optimization techniques, positioning FLARE-LA as a foundation for advancing FL innovations.

References

CameronOptiver IXAGPOPU Jiashen Liu Matteo Pietrobon (Optiver) OptiverMerle Sohier Dane Stefan Vallentine Andrew Meyer, BerniceOptiver. Optiver realized volatility prediction, 2021. URL `https://kaggle.com/competitions/optiver-realized-volatility-prediction`.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

W Brian Arthur, John H Holland, Blake LeBaron, Richard Palmer, and Paul Tayler. Asset pricing under endogenous expectations in an artificial stock market. In *The economy as an evolving complex system II*, pp. 15–44. CRC Press, 2018.

Syreen Banabilah, Moayad Aloqaily, Eitaa Alsayed, Nida Malik, and Yaser Jararweh. Federated learning review: Fundamentals, enabling technologies, and future applications. *Information processing & management*, 59 (6):103061, 2022.

Maxime Bergeron, Nicholas Fung, John Hull, and Zissis Poulos. Variational autoencoders: A hands-off approach to volatility. *arXiv preprint arXiv:2102.03945*, 2021.

El Bachir Boukherouaa, Mr Ghiath Shabsigh, Khaled AlAjmi, Jose Deodoro, Aquiles Farias, Ebru S Iskender, Mr Alin T Mirestean, and Rangachary Ravikumar. *Powering the digital economy: opportunities and risks of artificial intelligence in finance*. International Monetary Fund, 2021.

Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. Deep hedging. *Quantitative Finance*, 19(8): 1271–1291, 2019.

Estelle Cantillon and Pai-Ling Yin. Competition between exchanges: A research agenda. *International journal of industrial organization*, 29(3):329–336, 2011.

Daoyuan Chen, Liuyi Yao, Dawei Gao, Bolin Ding, and Yaliang Li. Efficient personalized federated learning via sparse model-adaptation. In *International Conference on Machine Learning*, pp. 5234–5256. PMLR, 2023a.

Jacky Chen, John C Hull, Zissis Poulos, Haris Rasul, Andreas Veneris, and Yuntao Wu. A variational autoencoder approach to conditional generation of possible future volatility surfaces. *Available at SSRN 4628457*, 2023b.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Seok-Ju Hahn, Minwoo Jeong, and Junghye Lee. Connecting low-loss subspace for personalized federated learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 505–515, 2022.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Joel Hasbrouck. *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press, 2007.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Meirui Jiang, Anjie Le, Xiaoxiao Li, and Qi Dou. Heterogeneous personalized federated learning by local-global updates mixing via convergence rate. In *ICLR*, 2024.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5132–5143. PMLR, 2020.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Wesley Maddox, Shuai Tang, Pablo Moreno, Andrew Gordon Wilson, and Andreas Damianou. Fast adaptation with linearized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 2737–2745. PMLR, 2021.

Ananth Madhavan. Market microstructure: A survey. *Journal of Financial Markets*, 3(3):205–258, 2000.

Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Konrad Mueller, Amira Akkari, Lukas Gonon, and Ben Wood. Fast deep hedging with second-order optimization. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 319–327, 2024.

Brian Ning, Sebastian Jaimungal, Xiaorong Zhang, and Maxime Bergeron. Arbitrage-free implied volatility surface generation with variational autoencoders. *SIAM Journal on Financial Mathematics*, 14(4):1004–1027, 2023.

Jorge Otero. High-frequency data, frequency domain inference, and volatility forecasting. *Review of Economics and Statistics*, 84(4):669–681, 2002.

Mariia Seleznova and Gitta Kutyniok. Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. In *International Conference on Machine Learning*, pp. 19522–19560. PMLR, 2022.

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

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28, 2015.

Milena Vuletić and Rama Cont. Volgan: a generative model for arbitrage-free implied volatility surfaces. *Available at SSRN*, 2023.

Di Xie, Jiang Xiong, and Shiliang Pu. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6176–6185, 2017.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. In *ACM Transactions on Intelligent Systems and Technology (TIST)*, volume 10, pp. 1–19. ACM, 2019.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

A Convergence Analysis

In each training round t, we dynamically select a subset of trading platforms $S^t \subseteq E$, where $|S^t| = S$ denotes the number of participating platforms in that round. The current global model $\boldsymbol{w^{t-1}}$ is distributed to all selected platforms. Each participating platform i initializes its local model with the received global model, i.e., $\boldsymbol{w^t_{i,0}} = \boldsymbol{w^{t-1}}$. The local models are then updated through K iterations of SGD-based on their local data. The update rule for the local parameters at iteration k is given by

$$\boldsymbol{w^t_{i,k}} = \boldsymbol{w^t_{i,k-1}} - \alpha_l \boldsymbol{\nabla L_i(w^t_{i,k-1})}, \tag{19}$$

where α_l is the local learning rate, and $\boldsymbol{\nabla L_i(w)}$ represents the stochastic gradient of the local loss function L_i at platform i. After K iterations, the final local model for platform i is

$$\boldsymbol{w^t_{i,K}} = \boldsymbol{w^{t-1}} - \sum_{k=0}^{K-1} \alpha_l \boldsymbol{\nabla L_i(w^t_{i,k})}. \tag{20}$$

We assume that $\boldsymbol{\nabla L_i(w)}$ is an unbiased stochastic gradient with variance bounded by σ^2. The global model is updated by aggregating the updates from all selected local models. The update rule for the global model with global step size α_g is

$$\boldsymbol{w^t} = \boldsymbol{w^{t-1}} + \frac{\alpha_g}{S} \sum_{i \in S^t} (\boldsymbol{w^t_{i,K}} - \boldsymbol{w^{t-1}}) = \boldsymbol{w^{t-1}} - \frac{\alpha_g}{S} \sum_{i \in S^t} \sum_{k=0}^{K-1} \alpha_l \boldsymbol{\nabla L_i(w^t_{i,k})}. \tag{21}$$

To facilitate the analysis, we define the effective step size as $\tilde{\alpha} = K\alpha_l\alpha_g$. The update applied to the server model in round t can be expressed as

$$\boldsymbol{\delta^{t-1}} = -\frac{\tilde{\alpha}}{KS} \sum_{i \in S^t} \sum_{k=0}^{K-1} \boldsymbol{\nabla L_i(w^t_{i,k})}. \tag{22}$$

The expectation of the server update, considering the participation of all platforms E, is

$$\mathbb{E}[\boldsymbol{\delta^{t-1}}] = -\frac{\tilde{\alpha}}{K|E|} \sum_{i \in E} \sum_{k=0}^{K-1} \boldsymbol{\nabla L_i(w^t_{i,k})}. \tag{23}$$

The reduction can be shown by examining the distance from the minimizer $\boldsymbol{w^*}$

$$\begin{aligned} \|\boldsymbol{w^t} - \boldsymbol{w^*}\|^2 &= \|\boldsymbol{w^{t-1}} + \boldsymbol{\delta^{t-1}} - \boldsymbol{w^*}\|^2 \\ &= \|\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\|^2 + 2\left(\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\right)^T \boldsymbol{\delta^{t-1}} + \|\boldsymbol{\delta^{t-1}}\|^2. \end{aligned} \tag{24}$$

We use $\mathbb{E}_{t-1}[\cdot]$ to denote the expectation conditioned on all the randomness generated prior to round t. Thus, we have

$$\mathbb{E}_{t-1}\left[\left(\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\right)^T \boldsymbol{\delta^{t-1}}\right] = -\frac{\tilde{\alpha}}{K|E|} \sum_{i \in E} \sum_{k=0}^{K-1} \mathbb{E}\left[\boldsymbol{\nabla L_i(w^t_{i,k})}^T \left(\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\right)\right]. \tag{25}$$

We assume the eigenvalues of the Hessian of all $\{L_i(\boldsymbol{w})\}_{i \in E}$ are bounded within (μ, β), and the quadratic upper bound and quadratic lower bound for local objective function $L_i(\boldsymbol{w^{t-1}})$ can be obtained as

$$L_i(\boldsymbol{w^{t-1}}) \leq L_i(\boldsymbol{w^t_{i,k-1}}) + \boldsymbol{\nabla L_i(w^t_{i,k-1})^T(w^{t-1} - w^t_{i,k-1})} + \frac{\beta}{2}\boldsymbol{\|w^{t-1} - w^t_{i,k-1}\|^2}, \tag{26}$$

and

$$L_i(\boldsymbol{w^*}) \geq L_i(\boldsymbol{w^t_{i,k-1}}) + \boldsymbol{\nabla L_i(w^t_{i,k-1})^T(w^* - w^t_{i,k-1})} + \frac{\mu}{2}\boldsymbol{\|w^* - w^t_{i,k-1}\|^2}. \tag{27}$$

Then, we can get

$$\boldsymbol{\nabla L_i(w^t_{i,k-1})^T(w^{t-1} - w^*)} \geq L_i(\boldsymbol{w^{t-1}}) - L_i(\boldsymbol{w^*}) + \frac{\mu}{2}\boldsymbol{\|w^* - w^t_{i,k-1}\|^2} - \frac{\beta}{2}\boldsymbol{\|w^{t-1} - w^t_{i,k-1}\|^2}. \tag{28}$$

By Triangle inequality, we have

$$\boldsymbol{\|w^* - w^t_{i,k-1}\|^2} \geq \frac{1}{2}\boldsymbol{\|w^* - w^{t-1}\|^2} - \boldsymbol{\|w^{t-1} - w^t_{i,k-1}\|^2}. \tag{29}$$

Combining with $\beta \geq \mu$, we can obtain

$$\boldsymbol{\nabla L_i(w^t_{i,k-1})^T(w^{t-1} - w^*)} \geq L_i(\boldsymbol{w^{t-1}}) - L_i(\boldsymbol{w^*}) + \frac{\mu}{4}\boldsymbol{\|w^* - w^{t-1}\|^2} - \beta\boldsymbol{\|w^{t-1} - w^t_{i,k-1}\|^2}. \tag{30}$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}_{t-1}\left[(\boldsymbol{w^{t-1}}-\boldsymbol{w}^*)^T\boldsymbol{\delta^{t-1}}\right]\\&\leq -\frac{\tilde{\alpha}}{K|E|}\sum_{i\in E}\sum_{k=0}^{K-1}\left(L_i(\boldsymbol{w^{t-1}})-L_i(\boldsymbol{w}^*)+\frac{\mu}{4}\|\boldsymbol{w^{t-1}}-\boldsymbol{w}^*\|^2-\beta\|\boldsymbol{w^t_{i,k-1}}-\boldsymbol{w^{t-1}}\|^2\right).\end{aligned}\tag{31}$$

The drift of the local model from the global model is formulated as

$$\varepsilon=\frac{1}{K|E|}\sum_{i\in E}\sum_{k=0}^{K-1}\|\boldsymbol{w^t_{i,k-1}}-\boldsymbol{w^{t-1}}\|^2,\tag{32}$$

then we obtain

$$\mathbb{E}_{t-1}\left[(\boldsymbol{w^{t-1}}-\boldsymbol{w}^*)^T\boldsymbol{\delta^{t-1}}\right]\leq -\tilde{\alpha}\left(L(\boldsymbol{w^{t-1}})-L(\boldsymbol{w}^*)+\frac{\mu}{4}\|\boldsymbol{w^{t-1}}-\boldsymbol{w}^*\|^2\right)+\tilde{\alpha}\beta\varepsilon.\tag{33}$$

For the sequence of local gradients $\boldsymbol{\{\nabla L_i(w^t_{i,k-1})\}}$ during the training procedure, the variance is defined by

$$\begin{aligned}&\mathbb{E}[\|\boldsymbol{\nabla L_i(w^t_{i,k-1})}-\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^{\boldsymbol{2}}]\\&=\mathbb{E}[\|\boldsymbol{\nabla L_i(w^t_{i,k-1})}\|^{\boldsymbol{2}}]-2\|\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^{\boldsymbol{2}}+\|\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^{\boldsymbol{2}}\\&=\mathbb{E}[\|\boldsymbol{\nabla L_i(w^t_{i,k-1})}\|^{\boldsymbol{2}}]-\|\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^{\boldsymbol{2}}.\end{aligned}\tag{34}$$

Similarly, we can get that

$$\begin{aligned}&\mathbb{E}[\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}(\boldsymbol{\nabla L_i(w^t_{i,k-1})}-\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}])\|^2]\\&=\mathbb{E}[\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\boldsymbol{\nabla L_i(w^t_{i,k-1})}\|^2]-\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^2.\end{aligned}\tag{35}$$

We assume the variance of local gradients is upper bounded by

$$\mathbb{E}[\|\boldsymbol{\nabla L_i(w^t_{i,k-1})}-\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^{\boldsymbol{2}}]\leq\gamma^2,\tag{36}$$

and by Jensen's inequality, we have that

$$\begin{aligned}&\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}(\boldsymbol{\nabla L_i(w^t_{i,k-1})}-\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}])\|^2\\&\leq\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\|\boldsymbol{\nabla L_i(w^t_{i,k-1})}-\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^2.\end{aligned}\tag{37}$$

Using the linearity of the expectation we have

$$\begin{aligned}&\mathbb{E}[\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}(\boldsymbol{\nabla L_i(w^t_{i,k-1})}-\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}])\|^2]\\&\leq\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\mathbb{E}[\|\boldsymbol{\nabla L_i(w^t_{i,k-1})}-\mathbb{E}[\boldsymbol{\nabla L_i(w^t_{i,k-1})}]\|^2].\end{aligned}\tag{38}$$

Then, we have the upper bound of $\mathbb{E}_{t-1}\left[\|\boldsymbol{\delta^{t-1}}\|^2\right]$ as

$$\mathbb{E}_{t-1}\left[\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\boldsymbol{\nabla L_i(w^t_{i,k-1})}\|^2\right]\leq\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\boldsymbol{\nabla L_i(w^t_{i,k-1})}\|^2+\frac{\tilde{\alpha}^2\gamma^2}{KS}.\tag{39}$$

By the triangle inequality, we have

$$\begin{aligned}
&\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\left(\boldsymbol{\nabla L_i(w^t_{i,k-1})} - \boldsymbol{\nabla L_i(w^{t-1})} + \boldsymbol{\nabla L_i(w^{t-1})}\right)\|^2 \\
&\leq 2\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\left(\boldsymbol{\nabla L_i(w^t_{i,k-1})} - \boldsymbol{\nabla L_i(w^{t-1})}\right)\|^2 + 2\|\frac{\tilde{\alpha}}{S}\sum_{i=1}^{S}\boldsymbol{\nabla L_i(w^{t-1})}\|^2.
\end{aligned} \tag{40}$$

By Jensen's inequality and the β-smoothness property, we have

$$\begin{aligned}
&\|\frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\left(\boldsymbol{\nabla L_i(w^t_{i,k-1})} - \boldsymbol{\nabla L_i(w^{t-1})}\right)\|^2 \\
&\leq \frac{\tilde{\alpha}}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\|\boldsymbol{\nabla L_i(w^t_{i,k-1})} - \boldsymbol{\nabla L_i(w^{t-1})}\|^2 \\
&\leq \frac{\tilde{\alpha}\beta^2}{KS}\sum_{i=1}^{S}\sum_{k=0}^{K-1}\|\boldsymbol{w^t_{i,k-1}} - \boldsymbol{w^{t-1}}\|^2.
\end{aligned} \tag{41}$$

We can also obtain

$$\begin{aligned}
\|\frac{\tilde{\alpha}}{S}\sum_{i=1}^{S}\boldsymbol{\nabla L_i(w^{t-1})}\|^2 &= \|\frac{\tilde{\alpha}}{S}\sum_{i=1}^{S}\left(\boldsymbol{\nabla L_i(w^{t-1})} - \boldsymbol{\nabla L(w^{t-1})}\right) + \boldsymbol{\nabla L(w^{t-1})}\|^2 \\
&\leq 2\|\frac{\tilde{\alpha}}{S}\sum_{i=1}^{S}\left(\boldsymbol{\nabla L_i(w^{t-1})} - \boldsymbol{\nabla L(w^{t-1})}\right)\|^2 + 2\|\tilde{\alpha}\boldsymbol{\nabla L(w^{t-1})}\|^2 \\
&\leq 2\tilde{\alpha}^2 B + 4\beta\tilde{\alpha}^2\left(L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*})\right),
\end{aligned} \tag{42}$$

by the triangle inequality, where we define the gradient dissimilarity is upper bounded by

$$\|\frac{1}{S}\sum_{i=1}^{S}\left(\boldsymbol{\nabla L_i(w^{t-1})} - \boldsymbol{\nabla L(w^{t-1})}\right)\|^2 \leq B. \tag{43}$$

We can conclude that

$$\mathbb{E}_{t-1}[\|\boldsymbol{\delta^{t-1}}\|^2] \leq 2\tilde{\alpha}\beta^2\varepsilon + 4\tilde{\alpha}^2 B + 8\beta\tilde{\alpha}^2\left(L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*})\right) + \frac{\tilde{\alpha}^2\gamma^2}{KS}, \tag{44}$$

and the improvement in one round is

$$\begin{aligned}
\mathbb{E}_{t-1}[\|\boldsymbol{w^t} - \boldsymbol{w^*}\|^2] &= \|\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\|^2 + 2\mathbb{E}_{t-1}[(\boldsymbol{w^{t-1}} - \boldsymbol{w^*})^T\boldsymbol{\delta^{t-1}}] + \mathbb{E}_{t-1}[\|\boldsymbol{\delta^{t-1}}\|^2] \\
&\leq \|\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\|^2 - 2\tilde{\alpha}\left(L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*}) + \frac{\mu}{2}\|\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\|^2\right) \\
&\quad + 2\tilde{\alpha}\beta\varepsilon + 2\tilde{\alpha}\beta^2\varepsilon + 4\tilde{\alpha}^2 B + 8\beta\tilde{\alpha}^2\left(L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*})\right) + \frac{\tilde{\alpha}^2\gamma^2}{KS} \\
&= (1 - \tilde{\alpha}\mu)\|\boldsymbol{w^{t-1}} - \boldsymbol{w^*}\|^2 + (8\beta\tilde{\alpha}^2 - 2\tilde{\alpha})\left(L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*})\right) \\
&\quad + 2\tilde{\alpha}\beta(\beta+1)\varepsilon + 4\tilde{\alpha}^2 B + \frac{\tilde{\alpha}^2\gamma^2}{KS}.
\end{aligned} \tag{45}$$

Since the local updating is stochastic, and we have defined the variance of the sampled gradient from the full local gradient as σ^2

$$\mathbb{E}\|\boldsymbol{g_i(w)} - \nabla L_i(\boldsymbol{w})\|^2 = \sigma^2 = \mathbb{E}\|\boldsymbol{g_i(w)}\|^2 - \|\nabla L_i(\boldsymbol{w})\|^2. \tag{46}$$

If we define $a = \frac{1}{K-1}$, then we can obtain the upper bound of the expectation

$$\begin{aligned}\mathbb{E}\|\boldsymbol{w_{i,k}^t} - \boldsymbol{w^{t-1}}\|^2 &\leq \left(1 + \frac{1}{K-1}\right)\mathbb{E}\|\boldsymbol{w_{i,k-1}^t} - \boldsymbol{w^{t-1}}\|^2 + K\alpha_l^2\mathbb{E}\|\boldsymbol{g_i(w_{i,k-1}^t)}\|^2 \\ &= \left(1 + \frac{1}{K-1}\right)\mathbb{E}\|\boldsymbol{w_{i,k-1}^t} - \boldsymbol{w^{t-1}}\|^2 + K\alpha_l^2\|\nabla L_i(\boldsymbol{w_{i,k-1}^t})\|^2 + K\alpha_l^2\sigma^2.\end{aligned} \tag{47}$$

Then, we want to eliminate the gradients with the local updating model $\boldsymbol{\nabla L_i(w_{i,k-1}^t)}$ by applying the inequality

$$\begin{aligned}\|\nabla L_i(\boldsymbol{w_{i,k-1}^t})\|^2 &= \|\nabla L_i(\boldsymbol{w_{i,k-1}^t}) - \nabla L_i(\boldsymbol{w^{t-1}}) + \nabla L_i(\boldsymbol{w^{t-1}})\|^2 \\ &\leq 2\|\nabla L_i(\boldsymbol{w_{i,k-1}^t}) - \nabla L_i(\boldsymbol{w^{t-1}})\|^2 + 2\|\nabla L_i(\boldsymbol{w^{t-1}})\|^2.\end{aligned} \tag{48}$$

Based on the Lipschitz continuous gradient, we have

$$\|\nabla L_i(\boldsymbol{w_{i,k-1}^t}) - \nabla L_i(\boldsymbol{w^{t-1}})\|^2 \leq \beta^2\|\boldsymbol{w_{i,k-1}^t} - \boldsymbol{w^{t-1}}\|^2, \tag{49}$$

and we can obtain

$$\begin{aligned}\mathbb{E}\|\boldsymbol{w_{i,k}^t} - \boldsymbol{w^{t-1}}\|^2 \leq &\left(1 + \frac{1}{K-1} + 2K\alpha_l^2\beta^2\right)\mathbb{E}\|\boldsymbol{w_{i,k-1}^t} - \boldsymbol{w^{t-1}}\|^2 \\ &+ 2K\alpha_l^2\|\nabla L_i(\boldsymbol{w^{t-1}})\|^2 + K\alpha_l^2\sigma^2.\end{aligned} \tag{50}$$

To upper bound the drift over K local updates, we can unroll the recursion from $\boldsymbol{w_{i,0}^t}$ to $\boldsymbol{w_{i,K-1}^t}$. Since $\boldsymbol{w_{i,0}^t = w^{t-1}}$, we can obtain

$$\mathbb{E}\|\boldsymbol{w_{i,K}^t} - \boldsymbol{w^{t-1}}\|^2 \leq \sum_{k=0}^{K-1}\left(1 + \frac{1}{K-1} + 2K\alpha_l^2\beta^2\right)^k (2K\alpha_l^2\|\nabla L_i(\boldsymbol{w^{t-1}})\|^2 + K\alpha_l^2\sigma^2). \tag{51}$$

This upper bound is a geometric series where $2K\alpha_l^2\|\nabla L_i(\boldsymbol{w^{t-1}})\|^2 + K\alpha_l^2\sigma^2$ is the coefficient, and $1 + \frac{1}{K-1} + 2K\alpha_l^2\beta^2$ is the common ratio between adjacent terms. This upper bound can also be written as

$$\sum_{k=0}^{K-1}(1 + \frac{1}{K-1} + 2K\alpha_l^2\beta^2)^k(2K\alpha_l^2\|\nabla L_i(\boldsymbol{w^{t-1}})\|^2 + K\alpha_l^2\sigma^2) = q(2K\alpha_l^2\|\nabla L_i(\boldsymbol{w^{t-1}})\|^2 + K\alpha_l^2\sigma^2). \tag{52}$$

where q is a constant with a fixed local learning rate α_l and local updating iterations K defined as

$$q = \frac{1 - \left(1 + \frac{1}{K-1} + 2K\alpha_l^2\beta^2\right)^K}{1 - \left(1 + \frac{1}{K-1} + 2K\alpha_l^2\beta^2\right)}. \tag{53}$$

Then, we come to analysis of the dynamic trading platform participation. According to the quadratic upper bound and the linear lower bound of the local objective function, we can obtain the inequality as

$$\begin{aligned}L_i(\boldsymbol{w^*}) - L_i(\boldsymbol{w}) &= L_i(\boldsymbol{w^*}) - L_i(\boldsymbol{z}) + L_i(\boldsymbol{z}) - L_i(\boldsymbol{w}) \\ &\leq \boldsymbol{\nabla L_i(w^*)^T(w^* - z) + \nabla L_i(w)^T(z - w)} + \frac{\beta}{2}\boldsymbol{\|z - w\|^2} \\ &= \boldsymbol{\nabla L_i(w^*)^T(w^* - w) + (\nabla L_i(w^*) - \nabla L_i(w))^T(w - z)} + \frac{\beta}{2}\boldsymbol{\|z - w\|^2}.\end{aligned} \tag{54}$$

We define

$$\boldsymbol{z = w} - \frac{1}{\beta}\boldsymbol{(\nabla L_i(w) - \nabla L_i(w^*))}, \tag{55}$$

and then, we have

$$\begin{aligned}&\boldsymbol{(\nabla L_i(w^*) - \nabla L_i(w))^T(w - z)} = -\frac{1}{\beta}\boldsymbol{\|\nabla L_i(w^*) - \nabla L_i(w)\|^2}, \\ &\frac{\beta}{2}\boldsymbol{\|z - w\|^2} = \frac{1}{2\beta}\boldsymbol{\|\nabla L_i(w^*) - \nabla L_i(w)\|^2},\end{aligned} \tag{56}$$

hence,

$$L_i(\boldsymbol{w^*}) - L_i(\boldsymbol{w}) \leq \boldsymbol{\nabla L_i(w^*)^T(w^* - w)} - \frac{1}{2\beta}\|\boldsymbol{\nabla L_i(w^*) - \nabla L_i(w)}\|^{\boldsymbol{2}}, \tag{57}$$

which leads to

$$L_i(\boldsymbol{w}) - L_i(\boldsymbol{w^*}) - \boldsymbol{\nabla L_i(w^*)^T(w - w^*)} \geq \frac{1}{2\beta}\|\boldsymbol{\nabla L_i(w^*) - \nabla L_i(w)}\|^{\boldsymbol{2}}. \tag{58}$$

Since

$$\frac{1}{|E|}\sum_{i\in E}(L_i(\boldsymbol{w}) - L_i(\boldsymbol{w^*})) = L(\boldsymbol{w}) - L^*, \tag{59}$$

then, we have

$$2\beta(L(\boldsymbol{w}) - L^*) \geq \frac{1}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w) - \nabla L_i(w^*)}\|^{\boldsymbol{2}}. \tag{60}$$

The bound on the local gradient can be found as

$$\begin{aligned}\frac{1}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w)}\|^{\boldsymbol{2}} &= \frac{1}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w) - \nabla L_i(w^*) + \nabla L_i(w^*)}\|^{\boldsymbol{2}}\\ &\leq \frac{2}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w) - \nabla L_i(w^*)}\|^{\boldsymbol{2}} + \frac{2}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w^*)}\|^{\boldsymbol{2}}\\ &\leq 4\beta(L(\boldsymbol{w}) - L^*) + \frac{2}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w^*)}\|^{\boldsymbol{2}}.\end{aligned} \tag{61}$$

And the upper bound of the local training drift is

$$\begin{aligned}\varepsilon &\leq \frac{1}{|E|}\sum_{i\in E} q\left(2K\alpha_l^2\|\boldsymbol{\nabla L_i(w^{t-1})}\|^{\boldsymbol{2}} + K\alpha_l^2\sigma^2\right)\\ &\leq 8qK\alpha_l^2\beta(L(\boldsymbol{w}) - L^*) + \frac{4qK\alpha_l^2}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w^*)}\|^{\boldsymbol{2}} + qK\alpha_l^2\sigma^2.\end{aligned} \tag{62}$$

The improvement in one round can be rewritten as

$$\begin{aligned}\mathbb{E}_{t-1}[\|\boldsymbol{w^t - w^*}\|^{\boldsymbol{2}}] &\leq (1-\tilde{\alpha}\mu)\|\boldsymbol{w^{t-1} - w^*}\|^{\boldsymbol{2}} + (8\beta\tilde{\alpha}^2 - 2\tilde{\alpha})(L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*}))\\ &\quad + 2\tilde{\alpha}\beta(\beta+1)\varepsilon + 4\tilde{\alpha}^2B + \frac{\tilde{\alpha}^2\gamma^2}{K|E|}\\ &\leq (1-\tilde{\alpha}\mu)\|\boldsymbol{w^{t-1} - w^*}\|^{\boldsymbol{2}} + c_3(L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*}))\\ &\quad + 2\tilde{\alpha}\beta(\beta+1)c_1 + c_2,\end{aligned} \tag{63}$$

where we define

$$\begin{aligned}c_1 &= \frac{4qK\alpha_l^2}{|E|}\sum_{i\in E}\|\boldsymbol{\nabla L_i(w^*)}\|^{\boldsymbol{2}} + qK\alpha_l^2\sigma^2,\\ c_2 &= 4\tilde{\alpha}^2B + \frac{\tilde{\alpha}^2\gamma^2}{K|E|},\\ c_3 &= 16\beta^2(\beta+1)qK\tilde{\alpha}\alpha_l^2 + 8\beta\tilde{\alpha}^2 - 2\tilde{\alpha}.\end{aligned} \tag{64}$$

Then, we can obtain the following upper bound

$$\begin{aligned}\mathbb{E}_{t-1}[L(\boldsymbol{w^{t-1}}) - L(\boldsymbol{w^*})] &\leq \mathbb{E}_{t-1}\left[\frac{1}{c_3}(1-\tilde{\alpha}\mu)\|\boldsymbol{w^{t-1} - w^*}\|^{\boldsymbol{2}} - \frac{1}{c_3}\|\boldsymbol{w^t - w^*}\|^{\boldsymbol{2}}\right]\\ &\quad + \frac{2}{c_3}\tilde{\alpha}\beta(\beta+1)c_1 + \frac{c_2}{c_3}.\end{aligned} \tag{65}$$

We assume the eigenvalues of the Hessian of $\hat{L}_i(\boldsymbol{w})$ are bounded within (μ_i, β_i), i.e.,

$$\mu_i \leq \|\frac{1}{\sigma_i^2}\boldsymbol{J_{w^*}J_{w^*}^T}\| \leq \beta_i. \tag{66}$$

We assume the local gradient w.r.t $\boldsymbol{w^*}$ is bounded by ϵ_i, i.e., $\|\boldsymbol{\nabla\hat{L}_i(w^*)}\| \leq \epsilon_i$, and $\boldsymbol{w_i^*}$ is the optimal model for trading platform i. The improvement of local adaptation in the model space can be bounded by

$$\begin{aligned}\|\boldsymbol{w} - \boldsymbol{w_i^*}\| &= \|\boldsymbol{w^*} - \boldsymbol{(\nabla^2\hat{L}_i(w^*))^{-1}\nabla\hat{L}_i(w^*)} - \boldsymbol{w_i^*}\| \\ &= \|\boldsymbol{(\nabla^2\hat{L}_i(w^*))^{-1}[\nabla\hat{L}_i(w^*) + \nabla^2\hat{L}_i(w^*)(w_i^* - w^*)]}\|.\end{aligned} \tag{67}$$

Since

$$\boldsymbol{\nabla\hat{L}_i(w_i^*) = \nabla\hat{L}_i(w^*) + \nabla^2\hat{L}_i(w^*)(w_i^* - w^*)}, \tag{68}$$

we obtain that

$$\|\boldsymbol{w} - \boldsymbol{w_i^*}\| = \|\boldsymbol{(\nabla^2\hat{L}_i(w^*))^{-1}[\nabla\hat{L}_i(w_i^*) - \nabla\hat{L}_i(w^*)] + (\nabla^2\hat{L}_i(w^*))^{-1}\nabla\hat{L}_i(w^*)}\|. \tag{69}$$

We assume the local gradient w.r.t $\boldsymbol{w^*}$ is bounded by ϵ_i. Then, we have

$$\|\boldsymbol{(\nabla^2\hat{L}_i(w^*))^{-1}\nabla\hat{L}_i(w^*)}\| \leq \|\boldsymbol{(\nabla^2\hat{L}_i(w^*))^{-1}}\|\|\boldsymbol{\nabla\hat{L}_i(w^*)}\| \leq \frac{\epsilon_i}{\mu_i}. \tag{70}$$

Furthermore, we have

$$\begin{aligned}\|\boldsymbol{(\nabla^2\hat{L}_i(w^*))^{-1}[\nabla\hat{L}_i(w_i^*) - \nabla\hat{L}_i(w^*)]}\| &\leq \|\boldsymbol{(\nabla^2\hat{L}_i(w^*))^{-1}}\|\|\boldsymbol{\nabla\hat{L}_i(w_i^*) - \nabla\hat{L}_i(w^*)}\| \\ &\leq \frac{\beta_i}{\mu_i}\|\boldsymbol{w_i^*} - \boldsymbol{w^*}\|.\end{aligned} \tag{71}$$

Therefore, we have

$$\|\boldsymbol{w} - \boldsymbol{w_i^*}\| \leq \frac{\beta_i}{\mu_i}\|\boldsymbol{w_i^*} - \boldsymbol{w^*}\| + \frac{\epsilon_i}{\mu_i}. \tag{72}$$

By approximating the global model with a local linearization w.r.t each trading platform's local dataset, the model updates are tailored to the local data distribution. This leads to more accurate predictions for each trading platform's data, reducing overall prediction error. To obtain the decrement of objective function $\hat{L}_i(\boldsymbol{w})$, we first derive the second-order Taylor expansion of $\hat{L}_i(\boldsymbol{w_i^*})$ as

$$\begin{aligned}\hat{L}_i(\boldsymbol{w}) &= \hat{L}_i(\boldsymbol{w^*}) - \boldsymbol{\nabla\hat{L}_i(w^*)^T(\nabla^2\hat{L}_i(w^*))^{-1}\nabla\hat{L}_i(w^*)} \\ &\quad + \frac{1}{2}\boldsymbol{\nabla\hat{L}_i(w^*)^T(\nabla^2\hat{L}_i(w^*))^{-1}\nabla\hat{L}_i(w^*)} \\ &= \hat{L}_i(\boldsymbol{w^*}) - \frac{1}{2}\boldsymbol{\nabla\hat{L}_i(w^*)^T(\nabla^2\hat{L}_i(w^*))^{-1}\nabla\hat{L}_i(w^*)}.\end{aligned} \tag{73}$$

Local adaptation can significantly reduce the objective function $\hat{L}_i(\boldsymbol{w})$, thereby decreasing the need for additional rounds of FL. The second-order Taylor expansion shows that the loss reduction is proportional to the squared norm of the residual $\boldsymbol{y_i - f_i}$, bounded by curvature information from the Hessian matrix as

$$\begin{aligned}\hat{L}_i(\boldsymbol{w^*}) - \hat{L}_i(\boldsymbol{w}) &= \frac{1}{2}\boldsymbol{\nabla\hat{L}_i(w^*)^T(\nabla^2\hat{L}_i(w^*))^{-1}\nabla\hat{L}_i(w^*)} \\ &\leq \|(\boldsymbol{\nabla^2\hat{L}_i(w^*)})^{-1}\| \cdot \|\boldsymbol{\nabla\hat{L}_i(w^*)}\|^2 \\ &= \sigma_i^2\|\boldsymbol{(J_{w^*}J_{w^*}^T)^{-1}}\| \cdot \|\frac{1}{\sigma_i^2}\boldsymbol{J_{w^*}(y_i - f_i)}\|^2 \\ &\leq \frac{\beta_i}{\sigma_i^2\mu_i} \cdot \|\boldsymbol{(y_i - f_i)}\|^2.\end{aligned} \tag{74}$$

By effectively reducing the local loss, each trading platform contributes more accurate volatility prediction. Consequently, fewer communication rounds are needed, making the FL process more efficient and scalable. This reduces the need for extensive FL rounds, ultimately leading to better performance in realized volatility prediction.