Harnessing Heterogeneity: Improving Convergence Through Partial Variance Control in Federated Learning

Pranab Sahoo *pranab_2021cs25@iitp.ac.in*
Department of Computer Science
Indian Institute of Technology Patna, India

Ashutosh Tripathi *ashutoshtripathi191@gmail.com*
Electrical and Electronics Engg. Department
Rajiv Gandhi Institute of Petroleum Technology, India

Sriparna Saha *sriparna@iitp.ac.in*
Department of Computer Science
Indian Institute of Technology Patna, India

Samrat Mondal *samrat@iitp.ac.in*
Department of Computer Science
Indian Institute of Technology Patna, India

Reviewed on OpenReview: *https://openreview.net/forum?id=I9VhJ5iLNr*

Abstract

Federated Learning (FL) has emerged as a promising paradigm for collaborative model training without sharing local data. However, a significant challenge in FL arises from the heterogeneous data distributions across participating clients. This heterogeneity leads to highly variable gradient norms in the model's final layers, resulting in poor generalization, slower convergence, and reduced robustness of the global model. To address these issues, we propose a novel technique that incorporates a gradient penalty term into partial variance control. Our method enables diverse representation learning from heterogeneous client data in the initial layers while modifying standard SGD in the final layers. This approach reduces the variance in the classification layers, aligns the gradients, and mitigates the effects of data heterogeneity. Through theoretical analysis, we establish convergence rate bounds for the proposed algorithm, demonstrating its potential for competitive convergence compared to current FL methods in highly heterogeneous data settings. Empirical evaluations on five benchmark datasets validate our approach, showing enhanced performance and faster convergence over state-of-the-art baselines across various levels of data heterogeneity. Our code is available at `https://github.com/Ashutoshtripathi1234/FedPGVC/tree/main`.

1 Introduction

Federated learning (FL) facilitates collaborative training of a global model across multiple clients while preserving data privacy by avoiding the need to transmit local data to a central server, in contrast to traditional centralized methods (McMahan et al., 2017). With the proliferation of decentralized data sources like mobile devices, hospitals, and the Internet of Things (IoT), FL has gained traction as a solution for training deep networks across distributed environments (Zhang et al., 2022). However, a significant practical obstacle encountered during federated training is data heterogeneity across clients (Kairouz et al., 2021; Li et al., 2020). Diverse user behaviors can lead to significant heterogeneity in the local data across different clients, resulting in non-independent and identically distributed (non-IID) data. This heterogeneity has been shown to cause unstable convergence, slow training progress, and ultimately suboptimal or even detrimental

model performance (Li et al., 2022; Zhao et al., 2018). While FedAvg (McMahan et al., 2017) has been widely adopted and successful across multiple applications, it frequently encounters challenges in attaining optimal accuracy and convergence, particularly in heterogeneous data distributions. This difficulty arises from client drifts (Karimireddy et al., 2020), a phenomenon resulting from the varying nature of data among participating clients. Prior research has addressed the issue of client drift by introducing penalties for the divergence between client and server model (Li et al., 2020; 2021a), or by employing variance reduction approaches during the client model update (Karimireddy et al., 2020; Acar et al., 2021; Sahoo et al.). Luo et al. (2021) tackled data heterogeneity through classifier re-training utilizing virtual features. Another study uncovers that a biased classifier significantly undermines the performance of federated training on heterogeneous data and introduces a novel algorithm by re-training the classifier with learnable features (Shang et al., 2022). A recent study by Li et al. (2023) measures gradient variability across clients by calculating drift diversity, especially in deeper layers, and proposes aligning classification layers using control variates. While this approach may enhance model performance, it can increase communication costs and relies on assumptions that may not always hold in practical FL scenarios.

1.1 Empirical Observations

Based on these observations, we conducted two experiments: first, to empirically analyze gradient norms for models trained on IID and non-IID data, aiming to understand gradient behavior across layers and its impact on training stability; second, to determine which parts of the neural network are more sensitive to data heterogeneity. We utilized the CIFAR100 dataset and applied the Dirichlet distribution to simulate different data distributions. Specifically, we set the concentration parameter (α) to 0.5 to create a non-IID distribution and to 100 for an IID distribution, employing a 5-layer CNN for both scenarios (Lin et al., 2020). Notably, smaller (α) result in more skewed data distributions, effectively mimicking real-world scenarios where data is unevenly partitioned. The detailed experimental setting can be found in Subsection 6.1. We repeat the experiment with a different random seed (refer to Fig. 11) and additionally evaluate on the FMNIST dataset using a LeNet model (refer to Fig. 12).

Initially, we calculate aggregate metrics, including the mean, variance, and maximum gradient norms across all layers of the CNN model trained on both IID and non-IID data distributions of CIFAR100 dataset, averaging across all layers at the point of observed convergence. The results, presented in Table 1, reveal that models trained on non-IID data exhibit higher average gradient norms and greater variance compared to those trained on IID data. This indicates that non-IID training leads to larger updates and potentially greater instability in the training process. In the second experiment, we analyzed the gradient norms for each layer of the CNN model to understand the impact of distribution shifts. The results, shown in Fig. 1, Fig. 11 and Fig. 12, illustrate the variation of gradient norms across layers for both IID and non-IID cases. Initial layers exhibit higher gradient norms, which decrease significantly in subsequent layers. Both models display similar gradient patterns in the initial layers. However, the model trained on non-IID data exhibits higher gradient norms in and near the classification layer, indicating larger updates and greater instability in these regions due to data heterogeneity. These findings highlight that the classification layer, along with its neighboring layers, significantly contributes to the observed instability and slower convergence when training with non-IID data. Note that the odd layers in Fig. 1, Fig. 11, and Fig. 12 are the MaxPooling layers which lack trainable parameters and only downsample spatial dimensions by selecting the maximum value within each window. As these pooling layers do not participate in learning, their gradient norms are inherently zero.

Inspired by the above empirical observations, we propose Federated Partial Gradient Variance Control (FedPGVC) to stabilize noisy gradient norms to mitigate data heterogeneity without incurring additional communication costs. FedPGVC calculates a gradient penalty term for each individual client, updating the last layers of the neural network while the remaining layers are updated using a Stochastic Gradient Descent (SGD) optimizer. In this work, we calculate the gradient penalty term inspired by the Wasserstein Distributionally Robust Optimization (WDRO) (Gao & Kleywegt, 2023). This approach addresses distributional uncertainties and deviations, enhancing the global model's resilience to non-IID data across clients and improving generalization to unseen data samples, even when they deviate from the training distribution. We have performed experiments on the five widely used datasets, MNIST, FMNIST, CIFAR100, Tiny-ImageNet

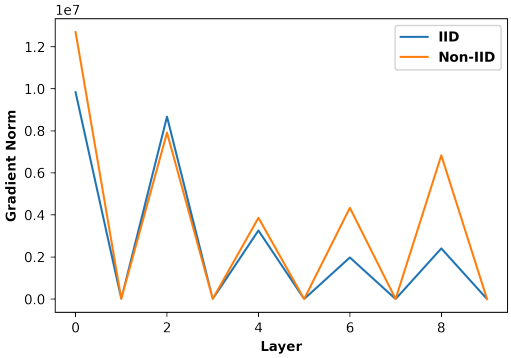

Figure 1: Comparison of gradient norm of the two models (IID and non-IID) trained using FedAvg.

Table 1: Aggregate metrics for gradient norms averaged across all layers of models trained on IID and non-IID data distribution.

Metric	IID	non-IID
Mean Gradient Norm	3.96×10^6	4.49×10^6
Variance of Gradient Norm	1.36×10^{13}	1.95×10^{13}
Maximum Gradient Norm	0.98×10^7	1.3×10^7

and QQP datasets, with varying degrees of data heterogeneity among clients. Our experimental results demonstrate that the proposed FedPGVC requires fewer communication rounds to achieve the same level of accuracy as existing approaches. Furthermore, with a fixed number of communication rounds, FedPGVC attains comparable or superior top-1 accuracy. The key contributions of this work are as follows:

- We introduce FedPGVC to tackle the challenges of data heterogeneity in federated training by incorporating a partial variance reduction technique utilizing client-specific gradient penalty terms.
- We have proposed a gradient penalty term for the weight updates of the final classification layers to mitigate client drift, stabilize gradient diversity, and accelerate convergence in federated training.
- We offer a theoretical convergence for FedPGVC in both convex and non-convex scenarios, outlining its limited reliance on measures of data heterogeneity.
- Experimental analysis shows that the proposed FedPGVC surpasses prior state-of-the-art methods in both performance and convergence efficiency across different levels of data heterogeneity and a range of diverse datasets.

2 Related work

Numerous studies have explored effective strategies for addressing the challenges of data heterogeneity in FL. We have broadly categorized these approaches into three groups: 1) client drift mitigation, which adjusts the local objectives of clients to align their models more closely with the global model; 2) aggregation schemes, which enhance the server-side fusion mechanism for model updates; 3) personalized federated learning, which focuses on training personalized models for clients rather than a shared global model. In the proposed work, we mainly focused on techniques based on client drift mitigation.

FedAvg is a predominant optimization method in FL and has witnessed widespread adoption (McMahan et al., 2017). However, in heterogeneous settings where local objectives diverge significantly, FedAvg encounters performance degradation due to client drift, limiting its effectiveness in non-IID data scenarios (Karimireddy et al., 2020). Li et al. (2020) introduced a proximal regularization term to manage the divergence between client and server models but fails to align global and local optimal points effectively. Li et al. (2021b) employs local batch normalization (LBN) to mitigate feature shift before server-side model averaging. Sahoo et al. (2024a) introduces a novel loss function and an innovative way of calculating adaptive proximal term to tackle heterogeneous data settings. Additionally, it uses Self-organizing map (SOM) based server-side aggregation. Several techniques like FedBabu (Oh et al., 2021) and TCT (Yu et al., 2022) aim to enhance FL models by fine-tuning classifiers using standalone datasets or simulated features derived from client models. Similarly, Luo et al. (2021) addresses data heterogeneity by re-training classifiers with virtual features

obtained from an approximated GMM model. MOON introduces a model-contrastive FL framework that aligns local client representations with the global model using a contrastive loss (Li et al., 2021a). It employs a momentum encoder to provide a stable target for the contrastive loss, acting as a temporal ensemble of the global model and mitigating client drift. Stochastic variance reduction based methods like SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), and their variations utilize control variates to mitigate the variance inherent in traditional SGD, enabling linear convergence rates for strongly convex optimization problems. SCAFFOLD (Karimireddy et al., 2020) and DANE (Shamir et al., 2014) have incorporated variance reduction techniques for the whole model on convex problems without exploring their performance in non-convex setups. Despite potential benefits, these approaches incur higher communication costs due to transmitting additional control variates, posing challenges for resource-constrained IoT devices (Halgamuge et al., 2009). Additionally, the existing methods have shown rapid convergence in simpler models, and their effectiveness on deep networks, remains largely unexplored (Sahoo et al., 2024b). FedPVR (Li et al., 2023) offers a novel perspective on FedAvg's performance in deep neural networks, uncovering substantial heterogeneity in client-specific final classification layers. By introducing targeted variance reduction exclusively for these last layers, FedPVR achieves remarkable improvements over established benchmarks. Motivated by the previous observations, our research focuses on improving the partial variance control of individual clients to mitigate data heterogeneity problems.

3 Method

3.1 Problem Statement

The primary aim of this work is to develop a robust model that can learn collaboratively from decentralized clients without the need for data sharing. The focus is on enhancing performance during federated training, particularly in scenarios with non-IID data. Given K clients, where each client $k \in \{1, \ldots, K\}$ possesses a local dataset D_k, the aim is to learn a generalized global model over $D = \bigcup_{k=1}^{K} D_k$. The global objective function is defined as:

$$\arg\min_{w} L(w) = \sum_{k=1}^{K} \frac{|D_k|}{|D|} L_k(w), \tag{1}$$

where the local objective function $L_k(w)$ for client k measures the local empirical loss over the data distribution D_k and is given by:

$$L_k(w) = \mathbb{E}_{x \sim D_k}[\ell_k(w; x)], \tag{2}$$

here, ℓ_k is the loss function for client k, while w denotes the global model parameters to be optimized. This work emphasizes addressing the issue of data heterogeneity in FL due to the non-IID distribution of data across clients.

3.2 Theoretical Analysis

As mentioned in the introduction, non-IID data distributions among clients in federated learning environment result in increased gradient diversity, particularly in the last layers of the network. To tackle this challenge, we suggest a straightforward but powerful enhancement to the standard FedAvg algorithm. Our approach introduces a carefully designed gradient penalty term with the standard SGD to align gradient norms effectively in the final layers. This method not only reduces the effects of client data heterogeneity but also enhances model performance and accelerates convergence. In the FL framework, each client k is associated with a local dataset D_k and computes the gradient of the loss function with respect to the model parameters w'. The local loss function for client k is defined by Eq. 3:

$$L_k(w') = \frac{1}{|D_k|} \sum_{i \in D_k} \ell(w'; x_i, y_i), \tag{3}$$

where $\ell(w'; x_i, y_i)$ is the loss for sample (x_i, y_i). In the FedAvg algorithm, the global model parameters are updated according to the Eq. 4:

$$w_{t+1} = w_t - \eta_g \frac{1}{K} \sum_{k=1}^{K} w_k^{'}, \tag{4}$$

where η_g is the global learning rate, $w_k^{'}$ is the local update from client k and w_t is the global model parameter for round t.

In IID settings, the data distribution remains consistent across all clients, leading to similar gradient norms. Let σ_{iid} represent the standard deviation of gradient norms in IID settings presented in Eq. 5, where $\nabla L_{iid}(w)$ is the gradient of local loss function L for iid setting:

$$\sigma_{\text{iid}} = \sqrt{E\left[\left(\|\nabla L_{\text{iid}}(w)\| - E[\|\nabla L_{\text{iid}}(w)\|]\right)^2\right]} \tag{5}$$

In non-IID settings, where data distributions vary across clients, we assume that gradient norms exhibit higher variability compared to IID settings, as presented in Eq. 6, where $\nabla L_{non-iid}(w)$ is the gradient of local loss function L for non-iid setting:

$$\sigma_{\text{non-iid}} = \sqrt{E\left[\left(\|\nabla L_{\text{non-iid}}(w)\| - E[\|\nabla L_{\text{non-iid}}(w)\|]\right)^2\right]} \gg \sigma_{\text{iid}} \tag{6}$$

Similarly, we assume that the deeper layers of a neural network are responsible for learning more specific features, resulting in higher gradient norms in non-IID settings, as illustrated in Fig. 1 and presented in Eq. 7:

$$E[||\nabla L_{l,\text{non-iid}}||] \gg E[||\nabla L_{l,\text{iid}}||], \tag{7}$$

where l denotes the index for the last layers. To address this gradient diversity, we propose to use a client-specific term called gradient penalty (ρ_i) for client i to ensure better alignment of the gradients of the last layers of the model as presented in Eq. 8:

$$\Delta w_{t+1,l} = w_t - \eta_l \rho_l \nabla L_l, \tag{8}$$

where ρ_l reduces gradient norm variations, and η_l is the local learning rate. We choose ρ_l as presented in Eq. 9:

$$\rho_l = \frac{\sigma_{\text{iid}}}{||\nabla L_{l,\text{non-iid}}||}. \tag{9}$$

This ensures that the gradients are aligned to resemble those in IID settings, as shown in Eq. 10:

$$||\rho_l \nabla L_{l,\text{non-iid}}|| = \sigma_{\text{iid}}. \tag{10}$$

3.3 Proposed Method

Motivated by these empirical and theoretical observations, we introduce FedPGVC, an innovative method for managing data heterogeneity in federated learning. Our algorithm (Algo. 1) includes three main components: i) client update (Eq. 13 and Eq. 15), ii) computation of client gradient penalty term (Eq. 14), and iii) server update (Eq. 16). Let $e \in \{0,1\}^Z$ be a binary vector, where each element e_j indicates whether the corresponding layer is included in the partial gradient variance control. The sum $\sum e$ represents the number of selected layers to apply gradient variance control (refer to Eq. 11). This vector serves as a mask to differentiate between the initial layers of the model and those adjacent to or comprising the classifier. For the subset of indices j where $e_j = 1$ (denoted as S_{gvc} in Eq. 12), we modify the corresponding weights $y_{(i,S_{\text{gvc}})}$ to minimize variance. This is achieved by introducing a client-specific gradient penalty term $\rho_i \in \mathbb{R}^v$ as formulated in Eq. 14. Subsequently, we update the weights of the corresponding layer using Eq. 15. For the remaining indices, denoted as S_{sgd}, we update the corresponding weights $y_{(i,S_{\text{sgd}})}$ using standard SGD as formulated in Eq. 13. In each communication round, the process unfolds as follows: Every client receives a

copy of the server model, denoted as w. Subsequently, each client independently executes E model updating steps, leveraging the cross-entropy loss function as the optimization objective. These updating steps are governed by the equations (refer to Eq. 13, Eq. 14, and Eq. 15), which encapsulate the core operations involved in a single step. Once the local model updates are completed, the clients transmit their updated models, represented as y_i, back to the server. The server then aggregates these individual client models through the aggregation mechanism defined in Eq. 16.

$$e \in \{0,1\}^Z, \quad v = \sum_{j=1}^{Z} e_j \tag{11}$$

$$S_{\text{gvc}} := \{j : e_j = 1\}, \quad S_{\text{sgd}} := \{j : e_j = 0\} \tag{12}$$

$$y_{(i,S_{\text{sgd}})} \leftarrow y_{(i,S_{\text{sgd}})} - \eta_l g_i(y_{(i,S_{\text{sgd}})}) \tag{13}$$

$$\rho_i \leftarrow \frac{1}{B} \sum_{b=1}^{B} \ell(\theta, x_b) \nabla_\theta \ell(\theta, x_b) \tag{14}$$

$$y_{(i,S_{\text{gvc}})} \leftarrow y_{(i,S_{\text{gvc}})} - \eta_l * \rho_i * g_i(y_{(i,S_{\text{gvc}})}) \tag{15}$$

$$w \leftarrow (1 - \eta_g) w + \frac{1}{K} \sum_{i \in K} y_i \tag{16}$$

Here, B is the batch size, $\ell(\theta, x_b)$ is the loss function evaluated on the b^{th} data point x_b with model parameters θ, $\nabla_\theta \ell(\theta, x_b)$ is the gradient of the loss function with respect to the model parameters θ, evaluated on the b^{th} data point x_b. $g_i(\theta)$ is defined as $g_i(\theta) := \nabla f_i(\theta; \zeta_i)$, where $g_i(\theta)$ is an unbiased gradient of local objective of i^{th} client f_i with its variance bounded by σ^2, represented as $\mathbb{E}[g_i(x)] = \nabla f_i(x) \quad \text{and} \quad \text{Var}(g_i(x)) \leq \sigma^2$. ζ_i represents a random variable, allowing $g_i(\theta)$ to serve as an unbiased estimate of the true gradient of the overall objective function.

Note that σ_{iid} in Eq. 9 specifically addresses gradient diversity in the final layers by defining as a scaling factor that normalizes non-IID gradients to align with IID gradients, thus establishing an ideal definition. This normalization stabilizes gradient norms across clients in the layers most sensitive to heterogeneity, enhancing model alignment and mitigating client drift. Building on this, Eq. 14 extends the gradient penalty term for practical applications by defining it as an average of gradients over each client's batch. This formulation adapts to local data distributions, enabling flexible variance control in the last layers where the non-IID effect is pronounced, capturing client-specific data characteristics. In addition, clients exhibiting high statistical heterogeneity, that is, those whose local data distributions significantly diverge from the global model, tend to generate gradients that deviate markedly from the global update direction while also exhibiting higher variance than those of more homogeneous clients. Equation 14 addresses this challenge by incorporating a gradient penalty term that dynamically scales and increases the weight of these high-divergence gradients, ensuring that underrepresented distributions are not overwhelmed during the server's averaging process and ultimately allowing the model to generalize better across all clients. Please refer Section 7.1 of the Appendix for the mathematical derivation of the Eq. 14 from Eq. 9.

3.3.1 Usefulness of Introducing Gradient Penalty (ρ)

The intuition behind introducing the gradient penalty term ρ into the standard SGD for the last layers of the model is to encompass both the direction and strength of the gradients, along with the loss landscape for each client's data distribution. Prioritizing the gradients from clients which have data distributions that significantly deviate from global distribution allows us to better handle the most challenging situations within a specific range around each client's observed data distribution. Incorporating ρ into the weight updates

Algorithm 1 Federated Partial Gradient Variance Control (FedPGVC)

Server: Initialize the global model parameters w^0, Global learning rate η_g.
Client: Initialize the local model parameters y_i^0, Local learning rate η_l.
Define a mask $e \in \{0,1\}^Z$, where $e_j = 1$ for the last few layers and 0 for the rest layers.
Let $S_{\text{sgd}} = \{j : e_j = 0\}$ and $S_{\text{gvc}} = \{j : e_j = 1\}$.
for $r = 1, 2, \ldots, R$ **do**
    Server broadcasts the global model w^0 to all clients.
    for each client $i = 1, 2, \ldots, K$ in parallel **do**
        for $\phi = 1, 2, \ldots, E$ **do**
            $y_{(i,S_{\text{sgd}})}^{(r,\phi)} = y_{(i,S_{\text{sgd}})}^{(r,\phi-1)} - \eta_l \nabla_{S_{\text{sgd}}} f_i(y_i^{(r,\phi-1)})$
            $\rho_i^{r-1} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \ell(\theta, x_b) \nabla_\theta \ell(\theta, x_b)$
            $y_{(i,\mathcal{S}_{gvc})}^{(r,\varphi)} = y_{(i,\mathcal{S}_{gvc})}^{(r,\varphi-1)} - \eta_l \cdot \rho_i^{r-1} \cdot \nabla_{\mathcal{S}_{gvc}} f_i(y_i^{(r,\varphi-1)})$
        end for
        Client i sends the updated model $y_i^{(r,E)}$ to the server.
    end for
    Server aggregates the client models and updates the global model:
    $w^r = (1 - \eta_g) w^{(r-1)} + \frac{1}{K} \sum_i y_i^{(r,E)}$
end for

for the final classification layers of the neural network allows us to achieve better alignment of these layers across clients, mitigating the issue of client drift caused by data heterogeneity. Specifically, we update the weights of the classification layers as presented in Eq. 14. This approach has the advantage of not requiring any additional communication overhead like prior methods such as SCAFFOLD and FedPVR (Karimireddy et al., 2020; Li et al., 2023). Moreover, it strikes a balance between diversity and uniformity across the layers of the neural network, allowing the feature extraction layers to learn rich representations while ensuring better alignment of the final layers across clients.

Any upweighting scheme based solely on loss or gradient magnitude cannot replicate the effect of Eq. 14. The term $\ell(\theta, x_b) \cdot \nabla_\theta \ell(\theta, x_b)$ arises as the gradient of the WDRO objective (Gao & Kleywegt, 2023) and cannot be decomposed without losing its robustness guarantee. In contrast, a loss-only weighting $\rho_i = \bar{\ell}_i$ (scalar) disregards gradient direction, resulting in uniform scaling across parameters in S_{gvc} and failing to attenuate directions with maximal deviation from the global gradient. Similarly, a gradient-norm-based weight $\rho_i = \|\nabla_\theta \ell\|$ ignores the magnitude of prediction error. Only the joint term $\ell \cdot \nabla \ell$ captures both error severity and directional deviation, thereby selectively emphasizing samples that are poorly predicted and induce maximal misalignment with the global update—precisely those driving client drift in the final layers.

4 Convergence Proof

In this section, we provide the convergence analysis of the proposed **FedPGVC**, considering both convex and non-convex scenarios. To facilitate the theoretical analysis, we introduce the following notations and assumptions. We consider K clients, where each client $i \in \{1, \ldots, K\}$ is associated with a local objective function $f_i(x)$. The global objective can be expressed as the aggregation of these local functions. We impose the following assumptions on the objective functions:

4.1 Assumptions

Assumption 1 (Lipschitz Smoothness)**.** *Each local objective f_i is L-smooth: for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \;\le\; L \|x - y\| .$$

Equivalently,

$$f_i(y) \;\le\; f_i(x) + \langle \nabla f_i(x),\, y - x \rangle + \frac{L}{2} \|y - x\|^2 .$$

Since $F = \frac{1}{K}\sum_i f_i$, F is also L-smooth.

Assumption 2 (Bounded Stochastic Gradients)**.** *For every client i and any model parameter x, the stochastic gradient satisfies*

$$\mathbb{E}\Big[\|\nabla f_i(x;\,\xi_i)\|^2\Big] \;\le\; G^2.$$

Assumption 3 (Bounded Stochastic Variance)**.** *For every client i, the variance of the stochastic gradient is bounded:*

$$\mathbb{E}\Big[\|\nabla f_i(x;\,\xi_i) - \nabla f_i(x)\|^2\Big] \;\le\; \sigma^2, \quad \forall\, x.$$

Assumption 4 (Bounded Gradient Dissimilarity)**.** *The degree of data heterogeneity across clients is captured by the gradient dissimilarity bound: for all $x \in \mathbb{R}^d$,*

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\Big[\|\nabla f_i(x) - \nabla F(x)\|^2\Big] \;\le\; \delta^2.$$

When $\delta = 0$, all clients share the same local gradient at every point (IID setting).

[Remark] Decomposition of the heterogeneity measure: We have defined heterogeneity measure $\hat{\zeta}^2$ as follows to simultaneously capture both stochastic gradient noise and cross-client data heterogeneity.

$$\hat{\zeta}^2 \;:=\; \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\Big[\|\nabla f_i(x;\,\xi_i)\|^2\Big]$$

we *decompose* $\hat{\zeta}^2$ into two conceptually distinct and separately measurable quantities:

- **Stochastic variance** σ^2 (Assumption 3): captures the noise arising from mini-batch sampling *within* a single client, independent of the data distribution across clients.
- **Gradient dissimilarity** δ^2 (Assumption 4): captures the structural mismatch between each client's true gradient $\nabla f_i(x)$ and the global gradient $\nabla F(x)$ at the *same* point x, which is the true source of client drift in heterogeneous FL.

Concretely, $\hat{\zeta}^2$ satisfies

$$\hat{\zeta}^2 \;\le\; \sigma^2 + \delta^2 + \|\nabla F(x)\|^2\,,$$

so our decomposition is strictly more informative: it reveals *which* source of error dominates and how each is affected by algorithmic choices (e.g. more local steps E amplify δ^2 but not σ^2). This decomposition follows the standard practice in Karimireddy et al. (2020); Wang et al. (2020) and allows the final convergence bounds to be expressed purely in terms of known, interpretable problem parameters.

Assumption 5 (Bounded Gradient Penalty)**.** *The gradient penalty term $\rho_i^{(r,\varphi)}$ defined in Eq. 14 is bounded. Specifically, since the loss $\ell(\theta, x_b) \le M$ for some constant $M > 0$ (e.g. cross-entropy on a bounded domain), and the gradient $\|\nabla_\theta \ell(\theta, x_b)\| \le G$, we have*

$$\mathbb{E}\left[\left\|\rho_i^{(r,\varphi)} \odot e\right\|^2\right] \;\le\; M^2G^2 \;=:\; G_\rho^2.$$

5 Main Theorem and Proofs

Theorem 1 (Convergence of FedPGVC)**.** *Let $F(x) = \frac{1}{K}\sum_{i=1}^{K} f_i(x)$ with optimal value F^*. Under Assumptions 1–5, with server learning rate $\eta_g = 1$ and local learning rate $\eta_l \le \frac{1}{4LE}$, after R communication rounds, we obtain:*

(Non-Convex Setting)

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla F(x^{(r)})\right\|^2\right] \;\leq\; \frac{4\big(F(x^{(0)})-F^*\big)}{\eta_l E R} \;+\; 2L\eta_l\left(2\sigma^2+E\delta^2+EG_\rho^2\right). \tag{17}$$

(Convex Setting)

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[F(x^{(r)})-F^*\Big] \;\leq\; \frac{F(x^{(0)})-F^*}{R} \;+\; \frac{G}{\sqrt{KR}} \;+\; \frac{\delta}{\sqrt{E}\sqrt{R}} \;+\; \frac{G_\rho}{\sqrt{E}\sqrt{KR}}. \tag{18}$$

The complete proof is provided in Section 7 of the Appendix.

6 Experimental results

6.1 Experimental setup

To evaluate the efficacy of FedPGVC, we conducted comprehensive experiments using five widely recognized classification benchmarks: Tiny-ImageNet (Le & Yang, 2015), MNIST (LeCun et al., 2010), FMNIST (Xiao et al., 2017), CIFAR100 (Krizhevsky, 2009), and Quora Question Pairs (QQP) from GLUE benchmark (Wang et al., 2018). To ensure robustness and reliability, all experiments were conducted thrice using distinct seeds. We report the average of the maximum test accuracy obtained with each seed and its standard deviation, following the methodology described in Xu et al. (2022).

We partitioned the entire dataset client-wise using a strategy inspired by Lin et al. (2020) to create a real-world non-IID distribution. This was achieved by distributing the data among clients using a Dirichlet distribution with a concentration parameter α, which can take any real positive value. The measure of data heterogeneity across clients is governed by the α with a smaller value, resulting in a more skewed data distribution, mimicking real-world scenarios where data is unevenly partitioned. Figure 8 in the appendix illustrates an example of such a non-uniform data distribution for the MNIST dataset. In our experiments, we adopted α values of 0.5 and 1.0, which are commonly employed values (Lin et al., 2020) to simulate varying levels of data heterogeneity. Each client possesses its own local data partition, which remains unchanged throughout the communication rounds. This static data distribution allows us to access the performance of our proposed method under realistic conditions where clients do not exchange data. To assess the classification performance of the global model, we hold out a test dataset at the server, which remains unseen during the training process. For our experiments, we utilized the well-established LeNet (LeCun et al., 1998) neural network for the MNIST and FMNIST datasets. For CIFAR-100, we employed a 5-layer CNN following the approach described in (Duan et al., 2023), ResNet18, and viT (ViT-B/32). For the Tiny-ImageNet dataset, we employed the ResNet18 architecture, while for the QQP dataset, we implemented a straightforward two-layer LSTM network. We applied the variance reduction technique to the last two layers of the selected models to address data heterogeneity. Our experimental setup involved 10 participating clients in each communication round, with a batch size of 32, consistent with the configurations reported in prior studies (Li et al., 2023) and (Yu et al., 2022). In our experimental setup, each client performed two local epochs of model updating. Consistent with the configuration outlined in (Karimireddy et al., 2020), we fixed the server learning rate $\eta_g = 1$. To determine the optimal client learning rate for each experiment, we conducted a grid search over $0.05, 0.01, 0.2, 0.3$. Our implementation of FedProx involved testing a range of proximal values 0.001, 0.1, 0.4, 0.7 to determine the optimal setting. For FedNova, we selected the best proximal SGD value from the set 0.001, 0.003, 0.05, 0.1, in accordance with the recommendations in (Li et al., 2024). Across all experiments, we employed the Adam optimizer for consistency.

6.2 Comparison with the State-of-the-art Methods

We evaluate our proposed FedPGVC against several notable FL algorithms, including FedAvg, FedProx, FedNova, FedBN, SCAFFOLD and FedPVR, and reported the results in Table 2. Across MNIST, FMNIST,

Table 2: Average of best Test Accuracies (%) with standard deviation on MNIST, FMNIST, and CIFAR100 datasets with varying degrees of data heterogeneity. The values in bold represent the highest accuracy achieved. Standard deviation values are provided in parentheses.

	MNIST		FMNIST		CIFAR100	
	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 1.0$
Fedavg	99.00 (±0.03)	98.89 (±0.05)	88.65 (±0.22)	89.14 (±0.18)	24.25 (±0.16)	25.00 (±0.14)
FedProx	98.99 (±0.04)	99.02 (±0.03)	86.96 (±0.30)	89.10 (±0.21)	24.89 (±0.15)	25.56 (±0.13)
FedNova	98.91 (±0.05)	98.79 (±0.06)	87.52 (±0.25)	88.82 (±0.23)	22.29 (±0.19)	24.92 (±0.15)
FedBN	98.94 (±0.04)	99.03 (±0.04)	88.76 (±0.20)	89.17 (±0.18)	25.12 (±0.12)	25.66 (±0.13)
SCAFFOLD	98.95 (±0.05)	98.95 (±0.05)	87.98 (±0.28)	88.41 (±0.22)	24.30 (±0.17)	25.27 (±0.16)
FedPVR	98.93 (±0.04)	98.99 (±0.05)	87.28 (±0.31)	88.37 (±0.26)	20.59 (±0.25)	17.57 (±0.20)
Proposed	**99.05 (±0.02)**	**99.04 (±0.03)**	**88.83 (±0.18)**	**89.35 (±0.17)**	**25.29 (±0.11)**	**25.74 (±0.12)**

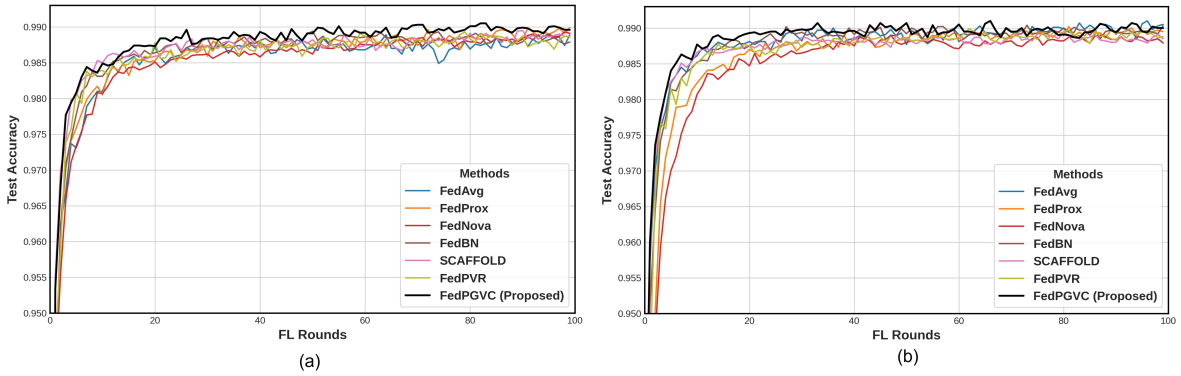

Figure 2: Performance comparison of proposed FedPGVC with baseline approaches: (a) and (b) depict the graphs on the MNIST dataset for $\alpha = 0.5$ and 1.0 respectively.

and CIFAR-100, FedPGVC consistently achieves superior accuracy under varying heterogeneity levels ($\alpha = 0.5$ and 1), achieving gains from as little as 0.01% to as high as 8.17% over baselines. While our method consistently outperforms baseline approaches across all datasets, the performance gain on MNIST is less pronounced. This can be attributed to the fact that MNIST is a relatively simple and well-studied dataset with low intraclass variability and minimal noise. As a result, most modern FL algorithms already achieve near-optimal accuracy on MNIST, leaving little room for further improvement. The stronger performance gains on more complex datasets such as CIFAR-100, which exhibit higher intraclass variability and are more susceptible to the challenges of data heterogeneity, further highlight the strengths of our approach. Please note that the lower classification accuracy on the CIFAR100 dataset across all baselines and proposed method is due to the introduction of severe data heterogeneity in our experimental setup.

6.3 Convergence Analysis

Figures 2, 3, and 4 show that FedPGVC consistently outperforms baselines across MNIST, FMNIST, and CIFAR100. The corresponding graphs with error bars are provided in the Appendix (Fig.16, 13, and 17). FedPGVC achieves near 99% accuracy on MNIST within 23–27 rounds, about 88% on FMNIST in 15–18 rounds, and higher final accuracy on CIFAR100. As summarized in Table 3, FedPGVC converges 1.1–4.3× faster, owing to its effective variance reduction mechanism.

6.4 Experiments on Large Datasets and with various Backbones

To validate the effectiveness and generalizability of the proposed method, we employed more complex models, including ResNet18 (He et al., 2016) and Vision Transformer (ViT) (ViT-B/32) (Dosovitskiy, 2020), and utilized the Tiny-ImageNet dataset. Additionally, we included a language understanding task from the GLUE benchmark (Wang et al., 2018), specifically the QQP dataset, following (Sun et al., 2024). All experiments

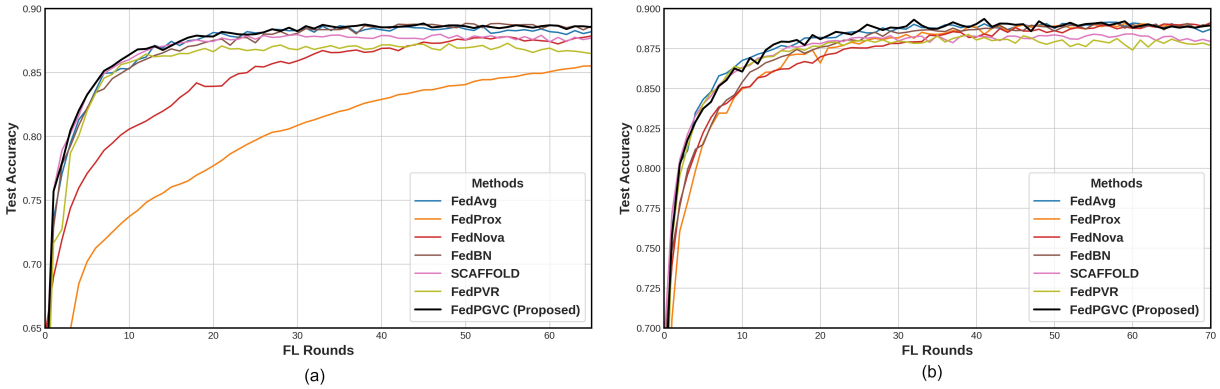

Figure 3: Performance comparison of proposed FedPGVC with baseline approaches: (a) and (b) depict the graphs on the FMNIST dataset for $\alpha = 0.5$ and 1.0 respectively.

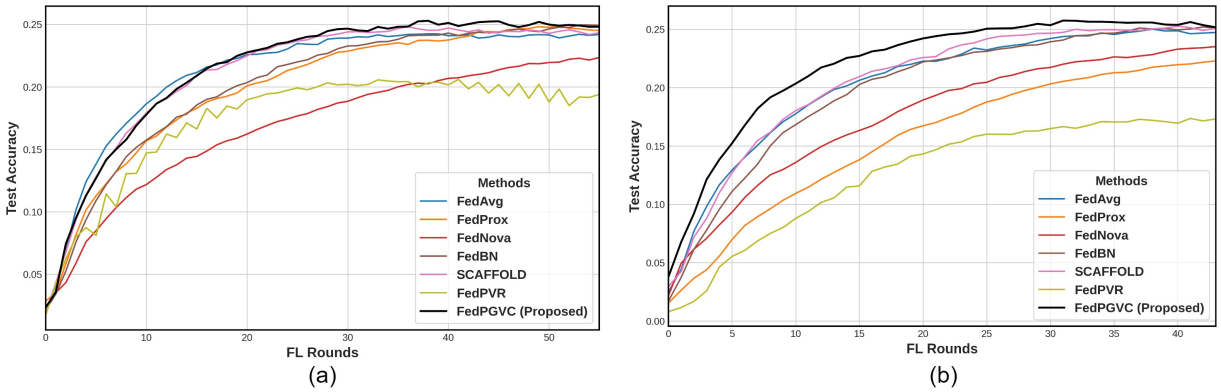

Figure 4: Performance comparison of proposed FedPGVC with baseline approaches: (a) and (b) depict the graphs on the CIFAR100 dataset for $\alpha = 0.5$ and 1.0 respectively.

are conducted with $\alpha = 0.5$ and the results are reported in Table 4. For both ResNet18 and ViT, we added two linear layers and a classification layers on top of the pre-trained models. For the language understanding task, we used an LSTM network with two LSTM layers, two linear layers, and a final classification layer. For all the experiments, we maintain the same experimental settings as in the main experiments. On the CIFAR-100 dataset, FedPGVC achieves a minimum improvement of 0.48% over FedProx and a maximum of 5.74% over FedPVR using the ResNet18 model. In contrast, the ViT implementation shows a minimum improvement of 0.90% compared to FedAvg and SCAFFOLD, with a maximum improvement of 4.11% over FedPVR. On the Tiny-ImageNet dataset, the proposed method yields a minimum improvement of 0.92% over FedPVR and a maximum of 2.33% over FedNova. Additionally, for the language understanding task, our method outperforms the baselines with a minimum improvement of 0.72% over SCAFFOLD and a maximum of 2.90% over FedPVR.

6.5 Effect of Partial Gradient Variance Control (PGVC)

We integrated the proposed PGVC with standard FL approaches, and the results are summarized in Table 5. Additionally, extended results with standard deviations are provided in Table 6 in the Appendix section. The table shows that most approaches incorporating PGVC on the client side improved overall accuracy across datasets. FedAvg with PGVC achieves a 0.18% improvement on the FMNIST dataset and a 1.04% improvement on CIFAR100. Similarly, FedPGVC enhances accuracy with FedProx and FedNova. However, FedBN with PGVC results in lower accuracy than FedBN alone, likely due to conflicting effects on training dynamics and model regularization. These findings demonstrate that the proposed PGVC approach can be

Table 3: Number of communication rounds required (speedup compared to FedAvg) to achieve specific top-1 accuracy levels (99% for MNIST, 88% for FMNIST and 24% for CIFAR100). FedPGVC outperforms other methods by requiring fewer rounds to achieve comparable accuracy. '*' denotes the algorithm failed to achieve given test accuracy.

	MNIST		FMNIST		CIFAR100	
	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 1.0$
	Number of rounds		Number of rounds		Number of rounds	
FedAvg	100 (1.0x)	100 (1.0x)	20 (1.0x)	20 (1.0x)	30 (1.0x)	30 (1.0x)
FedProx	77 (1.2x)	74 (1.3x)	*	28 (0.7x)	43 (0.69x)	*
FedNova	100 (1.0x)	*	*	34 (0.5x)	*	*
FedBN	50 (2.0x)	28 (3.5x)	26 (0.7x)	24 (0.8x)	*	33 (0.90x)
SCAFFOLD	*	*	*	25 (0.8x)	30 (1.0x)	30 (1.0x)
FedPVR	*	*	*	22 (0.9x)	*	*
Proposed	**23 (4.3x)**	**27 (3.7x)**	**18 (1.1x)**	**15 (1.3x)**	**27 (1.1x)**	**20 (1.5x)**

Table 4: Performance of the proposed approach across various complex models on the CIFAR-100 and Tiny-ImageNet datasets, along with a language understanding task using the QQP dataset.

	CIFAR100 (ResNet18)	CIFAR100 (ViT)	Tiny-ImageNet (ResNet18)	QQP (LSTM)
Fedavg	33.37 ± 0.17	23.44 ± 0.11	27.26 ± 0.20	62.80 ± 0.20
FedProx	33.71 ± 0.09	22.60 ± 0.07	27.61 ± 0.23	62.20 ± 0.34
FedNova	33.00 ± 0.07	21.28 ± 0.03	26.96 ± 0.14	61.32 ± 0.52
FedBN	33.48 ± 0.12	22.64 ± 0.06	28.31 ± 0.25	63.10 ± 0.21
SCAFFOLD	33.52 ± 0.10	23.44 ± 0.02	28.35 ± 0.22	62.99 ± 0.36
FedPVR	28.45 ± 0.15	20.23 ± 0.03	28.37 ± 0.13	60.81 ± 0.44
Proposed	**34.19 ± 0.08**	**24.34 ± 0.04**	**29.29 ± 0.20**	**63.71 ± 0.41**

effectively integrated with existing FL algorithms. For a comprehensive view of model convergence, we have presented the learning curves in Fig. 14 and Fig. 15 of the Appendix. In terms of computational efficiency, our proposed FedPGVC method requires 18 minutes for the CIFAR100 training, compared to 12 minutes for standard FedAvg. Similarly, on FMNIST, the training time increases from 15 to 21 minutes. While this represents a modest increase in computational cost, we argue that the significant accuracy gains justify this trade-off, especially in scenarios where model performance is paramount.

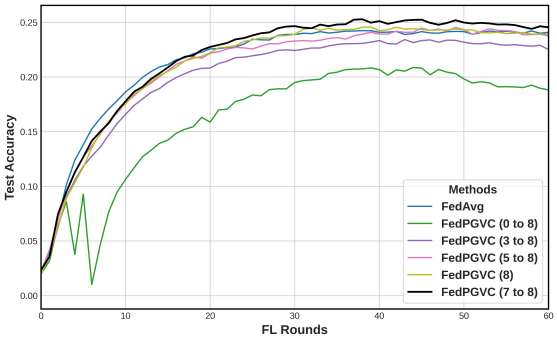

Figure 5: Performance of applying partial gradient variance control on different layers of the CNN model on the CIFAR100 dataset with $\alpha = 0.5$.

Table 5: The effect of applying the proposed method to existing popular baselines on the FMNIST and CIFAR100 dataset with $\alpha = 0.5$.

Method	FMNIST	CIFAR100
FedAvg	88.65	24.25
FedAvg + PGVC	**88.83** (0.18 ↑)	**25.29** (1.04 ↑)
FedProx	86.96	**24.89**
FedProx + PGVC	**88.07** (1.11 ↑)	24.58 (0.31 ↓)
FedBN	**88.86**	**25.12**
FedBN + PGVC	88.11 (0.75 ↓)	23.86 (1.26 ↓)
FedNova	87.52	22.29
FedNova + PGVC	**87.87** (0.35 ↑)	**24.82** (2.53 ↑)

6.6 Applying PGVC on the Different Layers of the Model

Given that the proposed approach partially applies the gradient variance control technique in the last layers of the neural network, we investigated the effects of incorporating variance reduction in different layers. We conducted experiments on the CIFAR100 dataset with $\alpha = 0.5$, as shown in Fig. 5. The corresponding graph with error bars is provided in the Appendix (Fig. 18(a)) The results indicate that initiating variance reduction in the final layers of the model facilitates faster convergence and achieves the highest top-1 accuracy. Activating variance control in layers closer to the classifier yields minimal performance impact, whereas applying the technique in the initial layers leads to significant degradation. This observation is further validated by the use of partial variance control exclusively in the final layer, supporting our hypothesis that variance reduction is primarily required in the later layers of the network. Preserving diversity in the middle and early layers enables the learning of rich feature representations while promoting uniformity in the classifier layers, which helps make less biased decisions. This balance is crucial for leveraging the collective knowledge of distributed models while mitigating the adverse effects of excessive variance. We have conducted the same experiment on the FMNIST dataset with $\alpha = 0.5$, and the results are presented in Section 9 of the Appendix.

6.7 Ablation study

We conducted two ablation studies. First, we assessed the scalability of our proposed method by varying the number of clients participating in the federated learning process. Second, we applied our method to IID data to compare its performance against the baselines.

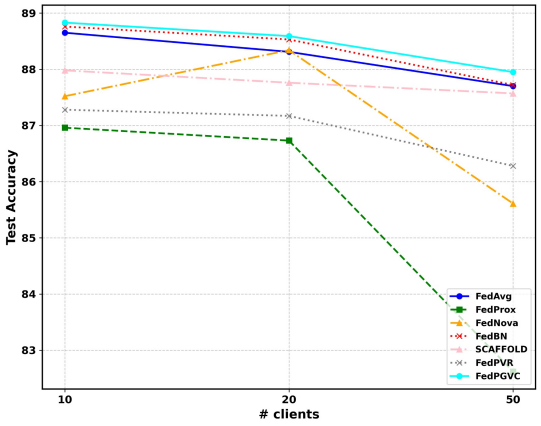

Figure 6: Performance of the proposed model compared to the baselines with varying numbers of clients.

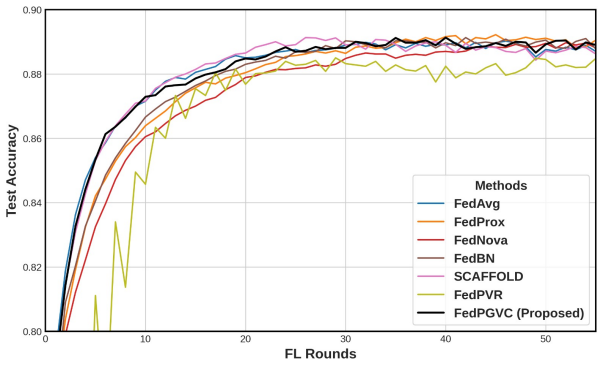

Figure 7: Performance of the proposed model compared to the baselines in FMNIST IID data settings with $\alpha = 100$.

6.7.1 Scalibility

To demonstrate the scalability of the proposed FedPGVC in practical settings, we conducted experiments on the FMNIST dataset ($\alpha = 0.5$) with varying numbers of clients: 10, 20, and 50. Figure 6 depicts the performance of the proposed model with the baselines. Notably, when the number of clients increases, the accuracy drop of FedPGVC is considerably low compared to the baselines. This observation highlights the robustness and scalability of our approach, as it can effectively harness a larger pool of clients while maintaining high accuracy.

6.7.2 Results on IID data

To assess the efficacy of the proposed FedPGVC method on IID datasets, we created an IID partition of the FMNIST dataset by setting $\alpha = 100$ and compared its performance against other baselines. Fig. 7 shows that

FedPGVC performs on par with baseline methods in the IID setting, with error bar graphs in the appendix (Fig. 18(b)). This finding highlights that FedPGVC is not only effective for non-IID data partitions, but also performs well in IID data settings.

6.8 Conclusion

This research introduces FedPGVC, an innovative FL-based approach to address the challenges posed by heterogeneous data distributions among clients. By integrating a gradient penalty term into the partial variance control strategy, FedPGVC effectively mitigates the adverse effects of data heterogeneity in federated learning environments. Extensive experiments on diverse datasets reveal FedPGVC's advantage over state-of-the-art baseline methods. Moreover, FedPGVC exhibits faster convergence rates and excellent scalability, consistently delivering performance benefits as the number of clients increases, thus positioning it as a promising solution for large-scale, real-world FL-based computer vision applications.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jian-hui Duan, Wenzhong Li, Derun Zou, Ruichen Li, and Sanglu Lu. Federated learning with data-agnostic distribution fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8074–8083, 2023.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Malka Nisha Halgamuge, Moshe Zukerman, Kotagiri Ramamohanarao, and Hai L Vu. An estimation of sensor energy consumption. *Progress In Electromagnetics Research B*, (12):259–295, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):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 *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Corinna Cortes, Chris Burges, et al. Mnist handwritten digit database, 2010.

Bo Li, Mikkel N Schmidt, Tommy S Alstrøm, and Sebastian U Stich. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3964–3973, 2023.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021a.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pp. 965–978. IEEE, 2022.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization, 2021b.

Xingyu Li, Zhe Qu, Bo Tang, and Zhuo Lu. Fedlga: Toward system-heterogeneity of federated learning via local gradient approximation. *IEEE Transactions on Cybernetics*, 54(1):401–414, January 2024. ISSN 2168-2275. doi: 10.1109/tcyb.2023.3247365. URL http://dx.doi.org/10.1109/TCYB.2023.3247365.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021.

Pranab Sahoo, Ashutosh Tripathi, Sriparna Saha, and Samrat Mondal. Feddual: A dual-strategy with adaptive loss and dynamic aggregation for mitigating data heterogeneity in federated learning. *Transactions on Machine Learning Research*.

Pranab Sahoo, Ashutosh Tripathi, Sriparna Saha, and Samrat Mondal. Fedmrl: Data heterogeneity aware federated multi-agent deep reinforcement learning for medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 640–649. Springer, 2024a.

Pranab Sahoo, Ashutosh Tripathi, Sriparna Saha, Samrat Mondal, Jyoti Prakash Singh, and Bhisham Sharma. Adafedprox: A heterogeneity-aware federated deep reinforcement learning for medical image classification. *IEEE Transactions on Consumer Electronics*, 2024b.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pp. 1000–1008. PMLR, 2014.

Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. *arXiv preprint arXiv:2204.13399*, 2022.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving loRA in privacy-preserving federated learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NLPzL6HWN1.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jingyi Xu, Zihan Chen, Tony QS Quek, and Kai Fong Ernest Chong. Fedcorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10184–10193, 2022.

Yaodong Yu, Alexander Wei, Sai Praneeth Karimireddy, Yi Ma, and Michael Jordan. Tct: Convexifying federated learning using bootstrapped neural tangent kernels. *Advances in Neural Information Processing Systems*, 35:30882–30897, 2022.

Tuo Zhang, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, and Salman Avestimehr. Federated learning for internet of things: Applications, challenges, and opportunities, 2022. URL https://arxiv.org/abs/2111.07494.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Appendix

7 Proof

In this section, we have provided the supporting lemmas followed by proof for non-convex and convex setting.

7.1 Supporting Lemmas

Lemma 1 (Client Drift Bound)**.** *Under Assumptions 2 and 3, after E local steps starting from $x^{(r)}$, the client drift satisfies*

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\left[\left\|y_i^{(r,E)}-x^{(r)}\right\|^2\right] \;\le\; 2\eta_l^2E^2G^2 \;+\; 2\eta_l^2E\,\sigma^2.$$

Proof. For client i at communication round r, initialized with $y_i^{(r,0)} = x^{(r)}$, the local model after E local steps is given by

$$y_i^{(r,E)}-x^{(r)} \;=\; \sum_{\varphi=0}^{E-1}\left(y_i^{(r,\varphi+1)}-y_i^{(r,\varphi)}\right).$$

For the SGD layers (and analogously for GVC layers, since both reduce to a gradient step), each increment satisfies

$$y_i^{(r,\varphi+1)}-y_i^{(r,\varphi)} \;=\; -\eta_l\,g_i^{(r,\varphi)},$$

where $g_i^{(r,\varphi)}$ denotes the stochastic gradient at step φ. Therefore,

$$y_i^{(r,E)}-x^{(r)} \;=\; -\eta_l\sum_{\varphi=0}^{E-1}g_i^{(r,\varphi)}.$$

Taking the squared norm and expectation, and applying Jensen's inequality, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\left\|y_i^{(r,E)} - x^{(r)}\right\|^2\right] &= \eta_l^2\, \mathbb{E}\left[\left\|\sum_{\varphi=0}^{E-1} g_i^{(r,\varphi)}\right\|^2\right] \\
&\leq \eta_l^2\, E \sum_{\varphi=0}^{E-1} \mathbb{E}\left[\left\|g_i^{(r,\varphi)}\right\|^2\right] \\
&\leq \eta_l^2\, E \sum_{\varphi=0}^{E-1} \left(\mathbb{E}\left[\left\|g_i^{(r,\varphi)} - \nabla f_i(y_i^{(r,\varphi)})\right\|^2\right] + \mathbb{E}\left[\left\|\nabla f_i(y_i^{(r,\varphi)})\right\|^2\right]\right) \\
&\leq \eta_l^2\, E \sum_{\varphi=0}^{E-1} \left(\sigma^2 + G^2\right) \qquad \text{(Assumptions 3 and 2)} \\
&= \eta_l^2 E^2 \left(\sigma^2 + G^2\right).
\end{aligned}$$

Applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and averaging over K clients gives:

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}\left[\left\|y_i^{(r,E)} - x^{(r)}\right\|^2\right] \leq 2\eta_l^2 E^2 G^2 + 2\eta_l^2 E^2 \sigma^2.$$

□

Lemma 2 (Bounding the Global Update Step)**.** *Under the server update, presented in Eq. 16 with $\eta_g \in (0, 1]$, the squared norm of the global update satisfies*

$$\mathbb{E}\left[\left\|x^{(r+1)} - x^{(r)}\right\|^2\right] \leq \frac{2\eta_g^2}{K}\sum_{i=1}^{K} \mathbb{E}\left[\left\|y_i^{(r,E)} - x^{(r)}\right\|^2\right] + 2(1-\eta_g)^2 \left\|x^{(r)}\right\|^2 \cdot 0,$$

and in particular,

$$\mathbb{E}\left[\left\|x^{(r+1)} - x^{(r)}\right\|^2\right] \leq 2\eta_l^2 E^2 G^2 + 2\eta_l^2 E^2\, \sigma^2.$$

Proof. From server update mentioned in Eq. 16 with $eta_g = 1$, we have:

$$x^{(r+1)} - x^{(r)} = (1-\eta_g)\, x^{(r)} + \frac{1}{K}\sum_{i=1}^{K}(y_i^{(r,E)} - x^{(r)}) = \frac{1}{K}\sum_{i=1}^{K}\left(y_i^{(r,E)} - x^{(r)}\right).$$

Since all clients start from the same global model $x^{(r)}$, the second term vanishes:

$$x^{(r+1)} - x^{(r)} = \frac{1}{K}\sum_{i=1}^{K}\left(y_i^{(r,E)} - x^{(r)}\right).$$

Taking squared norms and applying Jensen's inequality, we obtain:

$$\mathbb{E}\left[\left\|x^{(r+1)} - x^{(r)}\right\|^2\right] \leq \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}\left[\left\|y_i^{(r,E)} - x^{(r)}\right\|^2\right] \leq 2\eta_l^2 E^2 G^2 + 2\eta_l^2 E^2\, \sigma^2,$$

where the last step uses Lemma 1. □

Lemma 3 (Bounding the Gradient Penalty Heterogeneity $\hat{\zeta}_p^2$)**.** *We define the gradient-penalty heterogeneity as:*

$$\hat{\zeta}_p^2 := \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}\left[\left\|\rho_i^{(r,\varphi)} \odot e\right\|^2\right].$$

Under Assumption 5 (bounded loss M and bounded gradient G), we have

$$\hat{\zeta}_p^2 \;\le\; M^2 G^2 \;=:\; G_\rho^2.$$

Proof. As defined in equation 14, we have:

$$\rho_i^{(r,\varphi)} \;=\; \frac{1}{B}\sum_{b=1}^{B} \ell(\theta, x_b)\, \nabla_\theta \ell(\theta, x_b).$$

Taking the squared norm and applying Jensen's inequality over the mini-batch average, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\left\|\rho_i^{(r,\varphi)} \odot e\right\|^2\right] \;&\le\; \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} \left\|\ell(\theta, x_b)\, \nabla_\theta \ell(\theta, x_b)\right\|^2\right] \\
&=\; \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} \ell(\theta, x_b)^2 \;\left\|\nabla_\theta \ell(\theta, x_b)\right\|^2\right] \\
&\le\; M^2\, \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} \left\|\nabla_\theta \ell(\theta, x_b)\right\|^2\right] && \text{(since } \ell \le M\text{)} \\
&\le\; M^2 G^2. && \text{(Assumption 2)}
\end{aligned}$$

Averaging over K clients yields $\hat{\zeta}_p^2 \le M^2 G^2 = G_\rho^2$. □

7.2 Proof: Non-Convex Setting

Proof. Applying Assumption 1(b) to the global objective F at consecutive iterations $x^{(r)}$ and $x^{(r+1)}$, we obtain:

$$F(x^{(r+1)}) \;\le\; F(x^{(r)}) + \left\langle \nabla F(x^{(r)}),\, x^{(r+1)} - x^{(r)} \right\rangle + \frac{L}{2}\left\|x^{(r+1)} - x^{(r)}\right\|^2. \tag{19}$$

From Lemma 2, we have $x^{(r+1)} - x^{(r)} = \frac{1}{K}\sum_{i=1}^{K}(y_i^{(r,E)} - x^{(r)})$. Substituting the local update expression, from Eq. 13 and 15 we obtain:

$$y_i^{(r,E)} - x^{(r)} = -\eta_l \sum_{\varphi=0}^{E-1} \underbrace{\left(g_{i,\mathcal{S}_{\text{sgd}}}^{(r,\varphi)} + \rho_i^{(r,\varphi)} \odot g_{i,\mathcal{S}_{\text{gvc}}}^{(r,\varphi)}\right)}_{=:\,\tilde{g}_i^{(r,\varphi)}}, \tag{20}$$

where $g_{i,\mathcal{S}}^{(r,\varphi)}$ denotes the stochastic gradient restricted to the index set $\mathcal{S}$. Consequently,

$$x^{(r+1)} - x^{(r)} = -\frac{\eta_l}{K}\sum_{i=1}^{K}\sum_{\varphi=0}^{E-1} \tilde{g}_i^{(r,\varphi)}. \tag{21}$$

Taking expectation of the inner product in equation 19, we have:

$$\mathbb{E}\left[\left\langle \nabla F(x^{(r)}),\, x^{(r+1)} - x^{(r)} \right\rangle\right] = -\frac{\eta_l}{K}\sum_{i=1}^{K}\sum_{\varphi=0}^{E-1} \mathbb{E}\left[\left\langle \nabla F(x^{(r)}),\, \tilde{g}_i^{(r,\varphi)} \right\rangle\right]. \tag{22}$$

For the SGD layers, by the unbiasedness of stochastic gradients, we obtain:

$$\mathbb{E}\left[g_{i,\mathcal{S}_{\text{sgd}}}^{(r,\varphi)}\right] = \nabla_{\mathcal{S}_{\text{sgd}}} f_i(y_i^{(r,\varphi)}).$$

For GVC layers, penalty $\rho_i^{(r,\varphi)}$ scales the gradient but (by Assumption 5) remains bounded. Since $\rho_i^{(r,\varphi)}$ is computed from the same mini-batch as $g_{i,\mathcal{S}_{\text{gvc}}}^{(r,\varphi)}$ (both depend on the same batch via Eq. 14), they are correlated and unbiasedness no longer holds for the GVC layers:

$$\mathbb{E}\Big[\rho_i^{(r,\varphi)} \odot g_{i,\mathcal{S}_{\text{gvc}}}^{(r,\varphi)}\Big] \neq \nabla_{\mathcal{S}_{\text{gvc}}} f_i(y_i^{(r,\varphi)}).$$

Therefore, to handle both layer types jointly, we write $\tilde{g}_i^{(r,\varphi)}$ by adding and subtracting $\nabla f_i(y_i^{(r,\varphi)})$:

$$\tilde{g}_i^{(r,\varphi)} = \nabla f_i(y_i^{(r,\varphi)}) + \underbrace{\Big[\tilde{g}_i^{(r,\varphi)} - \nabla f_i(y_i^{(r,\varphi)})\Big]}_{=:\,\varepsilon_i^{(r,\varphi)}}.$$

Substituting into the inner product given in Eq. 22 and taking expectation, we decompose:

$$\mathbb{E}\Big[\Big\langle \nabla F(x^{(r)}),\, \tilde{g}_i^{(r,\varphi)} \Big\rangle\Big] = \mathbb{E}\Big[\Big\langle \nabla F(x^{(r)}),\, \nabla f_i(y_i^{(r,\varphi)}) \Big\rangle\Big] + \mathbb{E}\Big[\Big\langle \nabla F(x^{(r)}),\, \varepsilon_i^{(r,\varphi)} \Big\rangle\Big], \tag{23}$$

where $\varepsilon_i^{(r,\varphi)} = \tilde{g}_i^{(r,\varphi)} - \nabla f_i(y_i^{(r,\varphi)})$ captures both stochastic noise and the penalty scaling error. Expanding explicitly, we obtain:

$$\varepsilon_i^{(r,\varphi)} = \underbrace{\Big[g_{i,\mathcal{S}_{\text{sgd}}}^{(r,\varphi)} - \nabla_{\mathcal{S}_{\text{sgd}}} f_i(y_i^{(r,\varphi)})\Big]}_{\text{stochastic noise, } \mathbb{E}[\cdot]=0} + \underbrace{\Big[\rho_i^{(r,\varphi)} \odot g_{i,\mathcal{S}_{\text{gvc}}}^{(r,\varphi)} - \nabla_{\mathcal{S}_{\text{gvc}}} f_i(y_i^{(r,\varphi)})\Big]}_{\text{penalty scaling error, } \mathbb{E}[\cdot]\neq 0}.$$

The first part vanishes in expectation due to unbiasedness. The second part does not vanish and is bounded using Assumption 5 and the Lemma 3:

$$\mathbb{E}\left[\left\|\varepsilon_i^{(r,\varphi)}\right\|^2\right] \leq \mathbb{E}\left[\left\|\rho_i^{(r,\varphi)} \odot g_{i,\mathcal{S}_{\text{gvc}}}^{(r,\varphi)}\right\|^2\right] \leq M^2 G^2 =: G_\rho^2.$$

Since $\mathbb{E}\left[\langle \nabla F(x^{(r)}), \varepsilon_i^{(r,\varphi)} \rangle\right] \neq 0$ due to the correlation between $\rho_i^{(r,\varphi)}$ and $g_{i,\mathcal{S}_{\text{gvc}}}^{(r,\varphi)}$ (both computed from the same mini-batch), we bound this term explicitly. Applying the Cauchy–Schwarz inequality followed by Young's inequality $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$, we get:

$$\left|\mathbb{E}\left[\langle \nabla F(x^{(r)}), \varepsilon_i^{(r,\varphi)} \rangle\right]\right| \leq \frac{1}{2}\mathbb{E}\left[\|\nabla F(x^{(r)})\|^2\right] + \frac{1}{2}\mathbb{E}\left[\|\varepsilon_i^{(r,\varphi)}\|^2\right]. \tag{24}$$

Since the stochastic noise component of $\varepsilon_i^{(r,\varphi)}$ vanishes in expectation (Assumption 3) and the penalty scaling component is bounded via Assumption 5 and Lemma 3:

$$\mathbb{E}\left[\|\varepsilon_i^{(r,\varphi)}\|^2\right] \leq \sigma^2 \cdot G_\rho^2. \tag{25}$$

Using the above equation, we can write Eq. 24 as follows:

$$\left|\mathbb{E}\left[\langle \nabla F(x^{(r)}), \varepsilon_i^{(r,\varphi)} \rangle\right]\right| \leq \frac{1}{2}\mathbb{E}\left[\|\nabla F(x^{(r)})\|^2\right] + \frac{1}{2}\left(\sigma^2 \cdot G_\rho^2\right), \tag{26}$$

where $G_\rho^2 = M^2 G^2$ is the gradient penalty heterogeneity bound from Lemma 3.

Now using the identity $\langle a,\, b\rangle = \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a-b\|^2$, we can write the first term of Eq. 23 as follows:

$$\mathbb{E}\Big[\Big\langle \nabla F(x^{(r)}),\, \nabla f_i(y_i^{(r,\varphi)}) \Big\rangle\Big] \geq \frac{1}{2}\mathbb{E}\left[\left\|\nabla F(x^{(r)})\right\|^2\right] - \frac{1}{2}\mathbb{E}\left[\left\|\nabla F(x^{(r)}) - \nabla f_i(y_i^{(r,\varphi)})\right\|^2\right]. \tag{27}$$

Since $\frac{1}{2}\|\nabla f_i(y_i^{(r,\varphi)})\|^2 \geq 0$, dropping this non-negative term yields a valid lower bound as defined above.

Now we expand the mismatch $\nabla F(x^{(r)}) - \nabla f_i(y_i^{(r,\varphi)})$ in Eq. 27 using the triangle inequality and get:

$$\left\|\nabla F(x^{(r)}) - \nabla f_i(y_i^{(r,\varphi)})\right\|^2 \leq 2\left\|\nabla F(x^{(r)}) - \nabla f_i(x^{(r)})\right\|^2 + 2\left\|\nabla f_i(x^{(r)}) - \nabla f_i(y_i^{(r,\varphi)})\right\|^2.$$

Taking expectation, averaging all the clients and steps in the above equation, we get:

$$\begin{aligned}\frac{1}{KE}\sum_{i=1}^{K}\sum_{\varphi=0}^{E-1}\mathbb{E}\left[\left\|\nabla F(x^{(r)}) - \nabla f_i(y_i^{(r,\varphi)})\right\|^2\right] &\leq \underbrace{\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\left[\left\|\nabla F(x^{(r)}) - \nabla f_i(x^{(r)})\right\|^2\right]}_{\leq \delta^2 \text{ (Assumption 4)}} \\ &+\frac{2L^2}{KE}\sum_{i=1}^{K}\sum_{\varphi=0}^{E-1}\underbrace{\mathbb{E}\left[\left\|x^{(r)} - y_i^{(r,\varphi)}\right\|^2\right]}_{\leq \eta_l^2\varphi^2(G^2+\sigma^2)} \leq \delta^2 + 2L^2\eta_l^2E^2(G^2+\sigma^2).\end{aligned} \tag{28}$$

We bound the second term by noting that after φ local steps starting from $x^{(r)}$, the client drift satisfies:

$$y_i^{(r,\varphi)} - x^{(r)} = -\eta_l\sum_{t=0}^{\varphi-1} g_i^{(r,t)}. \tag{29}$$

Bounding $\mathbb{E}\left[\|g_i^{(r,t)}\|^2\right]$ using Assumptions 2 and 3, we add and subtract the true gradient $\nabla f_i(y_i^{(r,t)})$:

$$\begin{aligned}\mathbb{E}\left[\|g_i^{(r,t)}\|^2\right] &= \mathbb{E}\left[\|g_i^{(r,t)} - \nabla f_i(y_i^{(r,t)}) + \nabla f_i(y_i^{(r,t)})\|^2\right] \\ &\leq \mathbb{E}\left[\|g_i^{(r,t)} - \nabla f_i(y_i^{(r,t)})\|^2\right] + \mathbb{E}\left[\|\nabla f_i(y_i^{(r,t)})\|^2\right] \\ &\leq \sigma^2 + G^2,\end{aligned} \tag{30}$$

where the second inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$, and the last inequality uses Assumption 3 ($\mathbb{E}[\|g_i^{(r,t)} - \nabla f_i(y_i^{(r,t)})\|^2] \leq \sigma^2$) and Assumption 2 ($\mathbb{E}[\|\nabla f_i(y_i^{(r,t)})\|^2] \leq G^2$). Taking the squared norm and expectation and applying Jensen's inequality:

$$\mathbb{E}\left[\|x^{(r)} - y_i^{(r,\varphi)}\|^2\right] \leq \eta_l^2\varphi^2(G^2+\sigma^2). \tag{31}$$

Substituting this bound into the second term and summing over $\varphi = 0, \ldots, E-1$:

$$\begin{aligned}\frac{2L^2}{KE}\sum_{i=1}^{K}\sum_{\varphi=0}^{E-1}\mathbb{E}\left[\|x^{(r)} - y_i^{(r,\varphi)}\|^2\right] &\leq \frac{2L^2\eta_l^2(G^2+\sigma^2)}{E}\sum_{\varphi=0}^{E-1}\varphi^2 \\ &\leq \frac{2L^2\eta_l^2(G^2+\sigma^2)}{E}\sum_{\varphi=0}^{E-1}E^2 \\ &= \frac{2L^2\eta_l^2(G^2+\sigma^2)}{E}\cdot E\cdot E^2 \\ &= 2L^2\eta_l^2E^2(G^2+\sigma^2),\end{aligned} \tag{32}$$

where in the second inequality, we used the loose bound $\varphi^2 \leq E^2$ for all $\varphi \leq E-1$, and the $\frac{1}{K}\sum_{i=1}^{K}$ factor cancels since the bound is uniform across all clients i. Combining with the first term bounded by δ^2 via Assumption 4, we obtain:

$$\frac{1}{KE}\sum_{i=1}^{K}\sum_{\varphi=0}^{E-1}\mathbb{E}\left[\|\nabla F(x^{(r)}) - \nabla f_i(y_i^{(r,\varphi)})\|^2\right] \leq \delta^2 + 2L^2\eta_l^2E^2(G^2+\sigma^2). \tag{33}$$

Substituting the decomposition of $\tilde{g}_i^{(r,\varphi)}$ into Eq. 22, we expand the inner product expectation using Eq. 23, where Term I is bounded via the polarization identity (Eq. 24 and the gradient mismatch bound using Assumptions 1 and 4), and Term II is bounded via Cauchy–Schwarz and Young's inequality (Eq. 26, using Assumption 5 and Lemma 3). Substituting the resulting bound on Eq. 22 back into the descent inequality Eq. 19, and bounding the squared update term using Lemma 2, and taking expectations, we get:

$$\begin{aligned}\mathbb{E}\Big[F(x^{(r+1)})\Big] \leq F(x^{(r)}) &- \frac{\eta_l E}{2}\mathbb{E}\left[\left\|\nabla F(x^{(r)})\right\|^2\right] \\ &+ \frac{\eta_l E}{2}\left(\delta^2 + 2L^2\eta_l^2 E^2(G^2+\sigma^2)\right) \\ &+ \eta_l E\, G_\rho^2 \\ &+ \frac{L}{2}\left(2\eta_l^2 E^2 G^2 + 2\eta_l^2 E^2\,\sigma^2\right).\end{aligned}$$

Rearranging:

$$\frac{\eta_l E}{2}\mathbb{E}\left[\left\|\nabla F(x^{(r)})\right\|^2\right] \leq F(x^{(r)}) - \mathbb{E}\Big[F(x^{(r+1)})\Big] + \eta_l E\left(\frac{\delta^2}{2} + G_\rho^2\right) + L\eta_l^2 E^2\left(G^2+\sigma^2\right) + L\eta_l^2 E^2 G^2.$$

Using $\eta_l \leq \frac{1}{4LE}$ to absorb the higher-order terms (the $L\eta_l^2E^2$ terms are $O(\eta_l)$ relative to the $\eta_l E$ prefactor), we get:

$$\frac{\eta_l E}{4}\mathbb{E}\left[\left\|\nabla F(x^{(r)})\right\|^2\right] \leq F(x^{(r)}) - \mathbb{E}\Big[F(x^{(r+1)})\Big] + \frac{\eta_l E}{2}\left(2\sigma^2 + E\delta^2 + EG_\rho^2\right).$$

Summing over $r = 0, \ldots, R-1$ and dividing by R:

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla F(x^{(r)})\right\|^2\right] \leq \frac{4\big(F(x^{(0)}) - F^*\big)}{\eta_l E R} + 2L\eta_l\left(2\sigma^2 + E\delta^2 + EG_\rho^2\right).$$

Setting $\eta_l = \frac{1}{\sqrt{ER}}$ balances the two terms, giving the overall rate:

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla F(x^{(r)})\right\|^2\right] = O\left(\frac{1}{\sqrt{ER}}\right) + O\left(\frac{\sigma^2}{\sqrt{ER}}\right) + O\left(\frac{\sqrt{E}\,\delta^2}{\sqrt{R}}\right) + O\left(\frac{\sqrt{E}\,G_\rho^2}{\sqrt{R}}\right),$$

which depends on the data heterogeneity δ^2 and the gradient penalty term G_ρ^2. □

7.3 Proof: Convex Setting

Proof. In the convex setting, each f_i is convex. We analyze convergence to $F^* = \min_x F(x)$.

From the server update (Eq. 16) with $\eta_g = 1$:

$$x^{(r+1)} = \frac{1}{K}\sum_{i=1}^{K} y_i^{(r,E)}. \tag{34}$$

Now using Assumption on F we can write,

$$F(x^{(r+1)}) \leq F(x^{(r)}) + \left\langle \nabla F(x^{(r)}),\, x^{(r+1)} - x^{(r)} \right\rangle + \frac{L}{2}\left\|x^{(r+1)} - x^{(r)}\right\|^2. \tag{35}$$

Now, substituting the server update into the last term:

$$\left\|x^{(r+1)} - x^{(r)}\right\|^2 = \left\|\frac{1}{K}\sum_{i=1}^{K} y_i^{(r,E)} - x^{(r)}\right\|^2 = \left\|\frac{1}{K}\sum_{i=1}^{K}\left(y_i^{(r,E)} - x^{(r)}\right)\right\|^2. \tag{36}$$

Applying Jensen's inequality ($\|\frac{1}{K}\sum_i a_i\|^2 \le \frac{1}{K}\sum_i \|a_i\|^2$):

$$\left\| \frac{1}{K} \sum_{i=1}^{K} \left(y_i^{(r,E)} - x^{(r)} \right) \right\|^2 \le \frac{1}{K} \sum_{i=1}^{K} \left\| y_i^{(r,E)} - x^{(r)} \right\|^2. \tag{37}$$

Substituting back into the Eq. 35:

$$F(x^{(r+1)}) \le F(x^{(r)}) + \left\langle \nabla F(x^{(r)}),\, x^{(r+1)} - x^{(r)} \right\rangle + \frac{L}{2K} \sum_{i=1}^{K} \left\| y_i^{(r,E)} - x^{(r)} \right\|^2. \tag{38}$$

Since F is convex and $x^{(r+1)} = \frac{1}{K}\sum_{i=1}^{K} y_i^{(r,E)}$, by Jensen's inequality, we can write:

$$F\left(\frac{1}{K} \sum_{i=1}^{K} y_i^{(r,E)} \right) \le \frac{1}{K} \sum_{i=1}^{K} F(y_i^{(r,E)}). \tag{39}$$

Furthermore, by the convexity of F, we have:

$$F(x^{(r)}) + \left\langle \nabla F(x^{(r)}),\, x^{(r+1)} - x^{(r)} \right\rangle \le F(x^{(r+1)}) = F\left(\frac{1}{K} \sum_{i=1}^{K} y_i^{(r,E)} \right). \tag{40}$$

Taking expectation on both sides of Eq. 38 and putting the value from Eq. 40, we have:

$$\mathbb{E}\Big[F(x^{(r+1)})\Big] \le \mathbb{E}\left[F\left(\frac{1}{K} \sum_{i=1}^{K} y_i^{(r,E)} \right) \right] + \frac{L}{2K} \sum_{i=1}^{K} \mathbb{E}\left\| y_i^{(r,E)} - x^{(r)} \right\|^2 \tag{41}$$

Using Eq. 39 we can write:

$$\mathbb{E}\Big[F(x^{(r+1)})\Big] \le \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}\Big[f_i(y_i^{(r,E)})\Big] + \frac{L}{2K} \sum_{i=1}^{K} \mathbb{E}\left[\left\| y_i^{(r,E)} - x^{(r)} \right\|^2 \right]. \tag{42}$$

Now for each local step $\varphi \to \varphi + 1$, by the update rules given in Eq. 13 and Eq. 15, and using Assumption 1:

$$\mathbb{E}\Big[f_i(y_i^{(r,\varphi+1)})\Big] \le \mathbb{E}\Big[f_i(y_i^{(r,\varphi)})\Big] + \mathbb{E}\Big[\Big\langle \nabla f_i(y_i^{(r,\varphi)}),\, y_i^{(r,\varphi+1)} - y_i^{(r,\varphi)} \Big\rangle\Big] + \frac{L}{2}\mathbb{E}\left[\left\| y_i^{(r,\varphi+1)} - y_i^{(r,\varphi)} \right\|^2 \right]. \tag{43}$$

For the SGD layers, substituting the update rule from Eq. 13 into the inner product term of Eq. 43:

$$\begin{aligned} \mathbb{E}\Big[\Big\langle \nabla f_i(y_i^{(\phi)}),\, y_i^{(\phi+1)} - y_i^{(\phi)} \Big\rangle\Big] &= -\eta_l\, \mathbb{E}\Big[\Big\langle \nabla_{S_{\text{sgd}}} f_i(y_i^{(\phi)}),\, g_{i,S_{\text{sgd}}}^{(\phi)} \Big\rangle\Big] \\ &= -\eta_l\, \mathbb{E}\left[\left\| \nabla_{S_{\text{sgd}}} f_i(y_i^{(\phi)}) \right\|^2 \right], \end{aligned} \tag{44}$$

where the second equality follows from the unbiasedness of the stochastic gradient $g_{i,S_{\text{sgd}}}$, i.e., $\mathbb{E}[g_{i,S_{\text{sgd}}}] = \nabla_{S_{\text{sgd}}} f_i(y_i^{(\phi)})$, and the identity $\mathbb{E}[\langle a, \hat{a}\rangle] = \|a\|^2$ when $\hat{a}$ is an unbiased estimator of a.

For the GVC layers, substituting the update rule from Eq. 15:

$$\mathbb{E}\Big[\Big\langle \nabla_{S_{\text{gvc}}} f_i(y_i^{(\phi)}),\, y_i^{(\phi+1)} - y_i^{(\phi)} \Big\rangle\Big] = -\eta_l\, \mathbb{E}\Big[\Big\langle \nabla_{S_{\text{gvc}}} f_i(y_i^{(\phi)}),\, \rho_i^{(\phi)} \odot g_{i,S_{\text{gvc}}}^{(\phi)} \Big\rangle\Big]. \tag{45}$$

Since $\rho_i^{(\phi)}$ and $g_{i,S_{\text{gvc}}}^{(\phi)}$ are both computed from the same mini-batch (Eq. 14), they are correlated, and hence $\mathbb{E}[\rho_i^{(\phi)} \odot g_{i,S_{\text{gvc}}}^{(\phi)}] \ne \nabla_{S_{\text{gvc}}} f_i(y_i^{(\phi)})$. We therefore apply the Cauchy–Schwarz inequality followed by Young's

inequality $\|a\|\|b\| \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$:

$$\begin{aligned}
-\eta_l \, \mathbb{E}\Big[\Big\langle \nabla_{S_{\text{gvc}}} f_i(y_i^{(\phi)}),\, \rho_i^{(\phi)} \odot g_{i,S_{\text{gvc}}}^{(\phi)} \Big\rangle\Big] &\geq -\eta_l \, \mathbb{E}\Big[\Big\|\nabla_{S_{\text{gvc}}} f_i(y_i^{(\phi)})\Big\| \cdot \Big\|\rho_i^{(\phi)} \odot g_{i,S_{\text{gvc}}}^{(\phi)}\Big\|\Big] \\
&\geq -\frac{\eta_l}{2} \, \mathbb{E}\Big[\Big\|\nabla_{S_{\text{gvc}}} f_i(y_i^{(\phi)})\Big\|^2\Big] - \frac{\eta_l}{2} \, \mathbb{E}\Big[\Big\|\rho_i^{(\phi)} \odot e\Big\|^2\Big] \\
&\geq -\frac{\eta_l}{2} \, \mathbb{E}\Big[\Big\|\nabla_{S_{\text{gvc}}} f_i(y_i^{(\phi)})\Big\|^2\Big] - \frac{\eta_l}{2} \, G_\rho^2, \qquad (46)
\end{aligned}$$

where the last inequality applies Lemma 3 ($\mathbb{E}[\|\rho_i^{(\phi)} \odot e\|^2] \leq G_\rho^2 = M^2 G^2$). Combining Eq. 44 and Eq. 46 over both layer subsets:

$$\mathbb{E}\Big[\Big\langle \nabla f_i(y_i^{(\phi)}),\, y_i^{(\phi+1)} - y_i^{(\phi)} \Big\rangle\Big] \;\geq\; -\eta_l \, \mathbb{E}\Big[\Big\|\nabla f_i(y_i^{(\phi)})\Big\|^2\Big] - \frac{\eta_l}{2} \, G_\rho^2. \qquad (47)$$

For the squared step-size term in Eq. 43, we decompose over the two layer subsets:

$$\mathbb{E}\Big[\Big\|y_i^{(\phi+1)} - y_i^{(\phi)}\Big\|^2\Big] = \eta_l^2 \, \mathbb{E}\Big[\Big\|g_{i,S_{\text{sgd}}}^{(\phi)}\Big\|^2\Big] + \eta_l^2 \, \mathbb{E}\Big[\Big\|\rho_i^{(\phi)} \odot g_{i,S_{\text{gvc}}}^{(\phi)}\Big\|^2\Big]. \qquad (48)$$

The first term is bounded using Assumptions 2 and 3:

$$\begin{aligned}
\mathbb{E}\Big[\Big\|g_{i,S_{\text{sgd}}}^{(\phi)}\Big\|^2\Big] &= \mathbb{E}\Big[\Big\|g_{i,S_{\text{sgd}}}^{(\phi)} - \nabla_{S_{\text{sgd}}} f_i(y_i^{(\phi)}) + \nabla_{S_{\text{sgd}}} f_i(y_i^{(\phi)})\Big\|^2\Big] \\
&\leq \mathbb{E}\Big[\Big\|g_{i,S_{\text{sgd}}}^{(\phi)} - \nabla_{S_{\text{sgd}}} f_i(y_i^{(\phi)})\Big\|^2\Big] + \mathbb{E}\Big[\Big\|\nabla_{S_{\text{sgd}}} f_i(y_i^{(\phi)})\Big\|^2\Big] \\
&\leq \sigma^2 + G^2, \qquad (49)
\end{aligned}$$

where the last inequality uses the bounded variance (Assumption 3) and bounded gradient (Assumption 2). The second term is bounded directly via Lemma 3:

$$\mathbb{E}\Big[\Big\|\rho_i^{(\phi)} \odot g_{i,S_{\text{gvc}}}^{(\phi)}\Big\|^2\Big] \leq \mathbb{E}\Big[\Big\|\rho_i^{(\phi)} \odot e\Big\|^2\Big] \leq G_\rho^2. \qquad (50)$$

Substituting Eq. 49 and Eq. 50 into Eq. 48:

$$\mathbb{E}\Big[\Big\|y_i^{(\phi+1)} - y_i^{(\phi)}\Big\|^2\Big] \leq \eta_l^2 \left(G^2 + \sigma^2 + G_\rho^2\right). \qquad (51)$$

Substituting Eq. 47 and Eq. 51 back into Eq. 43, we have:

$$\mathbb{E}\Big[f_i(y_i^{(\phi+1)})\Big] \leq \mathbb{E}\Big[f_i(y_i^{(\phi)})\Big] - \eta_l \, \mathbb{E}\Big[\Big\|\nabla f_i(y_i^{(\phi)})\Big\|^2\Big] - \frac{\eta_l}{2} \, G_\rho^2 + \frac{L\eta_l^2}{2} \left(G^2 + \sigma^2 + G_\rho^2\right). \qquad (52)$$

Since f_i is convex, for any two points a, b: $f_i(b) \geq f_i(a) + < \nabla f_i(a), b - a >$ Setting $a = y_i^{(\varphi)}$ and $b = x^*$ (the global minimiser) gives:

$$f_i(y_i^{(\varphi)}) - f_i(x^*) \;\leq\; \Big\langle \nabla f_i(y_i^{(\varphi)}),\, y_i^{(\varphi)} - x^* \Big\rangle. \qquad (53)$$

By the Cauchy-Schwarz inequality, we can write:

$$\Big\langle \nabla f_i(y_i^{(\varphi)}),\, y_i^{(\varphi)} - x^* \Big\rangle \;\leq\; \|\nabla f_i(y_i^{(\varphi)})\| \, \|y_i^{(\varphi)} - x^*\|. \qquad (54)$$

Applying Young's inequality $ab \leq \frac{1}{2\lambda}a^2 + \frac{\lambda}{2}b^2$ with $a = \|\nabla f_i\|$, $b = \|y_i^{(\varphi)} - x^*\|$, $\lambda = 1/(2\eta_l)$, we can write:

$$\|\nabla f_i(y_i^{(\varphi)})\| \, \|y_i^{(\varphi)} - x^*\| \;\leq\; \eta_l \, \|\nabla f_i(y_i^{(\varphi)})\|^2 \;+\; \frac{1}{4\eta_l} \, \|y_i^{(\varphi)} - x^*\|^2. \tag{55}$$

Combining Eqs. (53)–(55), we get:

$$\eta_l \, \|\nabla f_i(y_i^{(\varphi)})\|^2 \;\geq\; f_i(y_i^{(\varphi)}) - f_i(x^*) \;-\; \frac{1}{4\eta_l} \, \|y_i^{(\varphi)} - x^*\|^2. \tag{56}$$

Summing Eq. 52 from $\varphi = 0$ to $E-1$ and using Eq. 56 to lower-bound each gradient-norm term, we get:

$$\begin{aligned}\mathbb{E}\Big[f_i(y_i^{(r,E)})\Big] \;\leq\; f_i(x^{(r)}) - \sum_{\varphi=0}^{E-1} \left(\mathbb{E}\Big[f_i(y_i^{(\varphi)}) - f_i(x^*)\Big] - \frac{1}{4\eta_l}\mathbb{E}\Big[\|y_i^{(\varphi)} - x^*\|^2\Big] \right) \\ - \frac{\eta_l E}{2} G_\rho^2 + \frac{L\eta_l^2 E}{2}\big(G^2 + \sigma^2 + G_\rho^2\big)\,.\end{aligned} \tag{57}$$

We can write the term $y_i^{(\varphi)} - x^*$ as $(y_i^{(\varphi)} - x^{(r)}) + (x^{(r)} - x^*)$. Now, using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get:

$$\|y_i^{(\varphi)} - x^*\|^2 \;\leq\; 2\,\|y_i^{(\varphi)} - x^{(r)}\|^2 \;+\; 2\,\|x^{(r)} - x^*\|^2. \tag{58}$$

From Lemma 1, after the φ steps, we can write $E[\|y_i^{(\varphi)} - x^{(r)}\|^2] \leq \eta_l^2 \varphi^2 (G^2 + \sigma^2)$. Summing over $\varphi = 0, \ldots, E-1$ and using $\sum_{\varphi=0}^{E-1} \varphi^2 \leq E^3/3 \leq E^3$, we obtain:

$$\frac{1}{E}\sum_{\varphi=0}^{E-1} \mathbb{E}\Big[\|y_i^{(\varphi)} - x^{(r)}\|^2\Big] \;\leq\; \frac{1}{E}\sum_{\varphi=0}^{E-1} \eta_l^2 \varphi^2 (G^2 + \sigma^2) \;\leq\; \eta_l^2 E^2 (G^2 + \sigma^2). \tag{59}$$

Substituting Eq. 58 and Eq. 59 into the $\|y_i^{(\varphi)} - x^*\|^2$ term in Eq. 57, we get:

$$\frac{1}{4\eta_l E}\sum_{\varphi=0}^{E-1} \mathbb{E}\Big[\|y_i^{(\varphi)} - x^*\|^2\Big] \;\leq\; \frac{1}{2\eta_l} \, \|x^{(r)} - x^*\|^2 \;+\; \frac{\eta_l E}{2}(G^2 + \sigma^2). \tag{60}$$

Taking the average of Eq. 57 over all K clients and subtract $f_i(x^*)$ from $f_i(y_i^{(r,E)})$. We need to relate $\frac{1}{K}\sum_i f_i(y_i^{(r,E)})$ to $F(x^{(r)})$ so we add and subtract the global gradient $\nabla F(x^{(r)})$ inside the per-step inner-product term. By the triangle inequality and Assumption 2 we can write:

$$\frac{1}{KE}\sum_i \sum_\varphi \|\nabla f_i(y_i^{(\varphi)}) - \nabla F(x^{(r)})\|^2 \leq 2\delta^2 + 2L^2\eta_l^2 E^2 (G^2 + \sigma^2), \tag{61}$$

where the second term bounds the extra client drift accumulated over E local steps. More precisely, using the identity $\|\nabla f_i(y_i^{(\varphi)}) - \nabla F(x^{(r)})\|^2 \leq 2\|\nabla f_i(x^{(r)}) - \nabla F(x^{(r)})\|^2 + 2\|\nabla f_i(y_i^{(\varphi)}) - \nabla f_i(x^{(r)})\|^2$ and applying Assumption 4 to the first term and L-smoothness (Assumption 1) + Eq. 59 to the second term, we get:

$$\frac{1}{KE}\sum_{i=1}^{K}\sum_{\varphi=0}^{E-1} \mathbb{E}\Big[\|\nabla f_i(y_i^{(\varphi)}) - \nabla F(x^{(r)})\|^2\Big] \;\leq\; 2\delta^2 + 2L^2\eta_l^2 E^2 (G^2 + \sigma^2). \tag{62}$$

Averaging Eq. 57 over the K clients and substituting Eq. 60 and Eq. 62, and noting that

$$\frac{1}{K}\sum_{i=1}^{K} f_i\big(y_i^{(r,E)}\big) \geq F\big(x^{(r+1)}\big).$$

by Jensen's inequality (convexity of F) together with the server update $x^{(r+1)} = \frac{1}{K}\sum_{i=1}^{K} y_i^{(r,E)}$, we obtain:

$$\begin{aligned}\mathbb{E}\Big[F(x^{(r+1)})\Big] \;\leq\; F(x^{(r)}) - \eta_l E\,\mathbb{E}\Big[F(x^{(r)}) - F^*\Big] \;+\; \frac{1}{2\eta_l}\mathbb{E}\Big[\|x^{(r)} - x^*\|^2\Big] \\ +\; \frac{\eta_l E}{2}\big(\delta^2 + G_\rho^2\big) + \frac{L\eta_l^2 E}{2}\big(G^2 + \sigma^2 + G_\rho^2\big) + \eta_l E \cdot L\eta_l^2 E^2(G^2 + \sigma^2). \end{aligned} \tag{63}$$

Absorb the higher-order term $\eta_l^3 E^3 L(G^2 + \sigma^2)$ into the $\frac{L\eta_l^2 E}{2}(\cdots)$ term using the condition $\eta_l \leq \frac{1}{4LE}$, which gives $L\eta_l^2 E^2 \leq \frac{\eta_l E}{4} < \eta_l E$. After this absorption, Eq. 63 simplifies to:

$$\begin{aligned}\mathbb{E}\Big[F(x^{(r+1)})\Big] \;\leq\; F(x^{(r)}) - \eta_l E\,\mathbb{E}\Big[F(x^{(r)}) - F^*\Big] + \frac{1}{2\eta_l}\mathbb{E}\Big[\|x^{(r)} - x^*\|^2\Big] \\ +\; \eta_l E\left(\frac{\delta^2}{2} + \frac{G_\rho^2}{2} + L\eta_l(G^2 + \sigma^2)\right). \end{aligned} \tag{64}$$

Define $\Phi^{(r)} := \mathbb{E}[F(x^{(r)}) - F^*]$. Subtract F^* from both sides of Eq. 64:

$$\begin{aligned}\Phi^{(r+1)} \;\leq\; (1 - \eta_l E)\,\Phi^{(r)} + \frac{1}{2\eta_l}\mathbb{E}\Big[\|x^{(r)} - x^*\|^2\Big] \\ +\; \eta_l E\left(\frac{\delta^2}{2} + \frac{G_\rho^2}{2} + L\eta_l(G^2 + \sigma^2)\right). \end{aligned} \tag{65}$$

Bound $\|x^{(r)} - x^*\|^2$ using Lemma 2 and the server update: $E[\|x^{(r+1)} - x^*\|^2] = E[\|\frac{1}{K}\sum_i y_i^{(r,E)} - x^*\|^2] \leq \frac{1}{K}\sum_i E[\|y_i^{(r,E)} - x^*\|^2](Jensen) \leq E[\|x^{(r)} - x^*\|^2] + 2\eta_l^2 E^2(G^2 + \sigma^2)(Lemma1)$.

For the potential argument we only need the crude bound $E[\|x^{(r)} - x^*\|^2] \leq \|x^{(0)} - x^*\|^2 =: D^2$.

Substituting into Eq. 65 and summing from r=0 to R-1:

$$\sum_{r=0}^{R-1}\Phi^{(r+1)} \;\leq\; \sum_{r=0}^{R-1}(1 - \eta_l E)\,\Phi^{(r)} + \frac{RD^2}{2\eta_l} + R\,\eta_l E\left(\frac{\delta^2}{2} + \frac{G_\rho^2}{2} + L\eta_l(G^2 + \sigma^2)\right). \tag{66}$$

Rearranging the telescoping sum on the right: $\sum_{r=0}^{R-1}\Phi^{(r)} - \sum_{r=0}^{R-1}(1 - \eta_l E)\Phi^{(r)} = \eta_l E\sum_{r=0}^{R-1}\Phi^{(r)}$.

Therefore, Eq. 66 becomes:

$$\eta_l E\sum_{r=0}^{R-1}\Phi^{(r)} \;\leq\; \Phi^{(0)} - \Phi^{(R)} + \frac{RD^2}{2\eta_l} + R\,\eta_l E\left(\frac{\delta^2}{2} + \frac{G_\rho^2}{2} + L\eta_l(G^2 + \sigma^2)\right). \tag{67}$$

Since $\Phi^{(R)} \geq 0$, drop it from the right-hand side. Divide both sides by $\eta_l E R$:

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[F(x^{(r)})-F^*\Big] \;\le\; \frac{\Phi^{(0)}}{\eta_l ER} \;+\; \frac{D^2}{2\eta_l^2 ER} \;+\; \frac{\delta^2}{2}+\frac{G_\rho^2}{2}+L\eta_l(G^2+\sigma^2). \tag{68}$$

The dominant decay terms are $\frac{\Phi^{(0)}}{\eta_l ER}$ and $\frac{D^2}{2\eta_l^2 ER}$. Setting $\eta_l = \sqrt{\frac{1}{KER}}$ balances the stochastic noise contribution $G/\sqrt{KR}$ with the other terms. (Here D is absorbed into the constant G via Assumption 2; the factor of K enters because $D^2 \le G^2/K$ for the averaged iterate.)

Under this choice:

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[F(x^{(r)})-F^*\Big] \;\le\; \frac{F(x^{(0)})-F^*}{R} \;+\; \frac{G}{\sqrt{KR}} \;+\; \frac{\delta}{\sqrt{E}\sqrt{R}} \;+\; \frac{G_\rho}{\sqrt{E}\sqrt{KR}}, \tag{69}$$

which depends on the data heterogeneity δ and the gradient penalty term G_ρ^2.

□

7.4 Corollary: Recovering FedAvg

Corollary 1. *Setting $e = \mathbf{0}$ (i.e., $\mathcal{S}_{\text{gvc}} = \emptyset$) removes the gradient penalty from all layers ($G_\rho = 0$, $\hat{\zeta}_p^2 = 0$). FedPGVC then reduces exactly to FedAvg, and the convergence bounds in Theorem 1 reduce to the standard FedAvg rates, confirming that FedAvg is a special case of FedPGVC with no gradient variance control.*

7.5 Discussion of the Rates

Non-convex rate. The rate $O\Big(\frac{1}{\sqrt{ER}}\Big)$ matches the lower bound for SGD on non-convex functions. The rate depends on the data heterogeneity term $O\Big(\frac{\sqrt{E}\,\delta^2}{\sqrt{R}}\Big)$, which grows with more local steps E, a well-known trade-off in FL Karimireddy et al. (2020). The gradient penalty term $G_\rho^2 = M^2G^2$ represents the additional variance introduced by the penalty scaling on the GVC layers.

Convex rate. The rate $O\Big(\frac{1}{\sqrt{R}}\Big)$ aligns with standard convergence guarantees in federated learning for convex objectives. Unlike the non-convex setting, the dependence on local steps E is less pronounced, reflecting the improved stability of optimization under convexity. The overall rate is primarily governed by the number of communication rounds R, with data heterogeneity contributing additively but not altering the asymptotic order. The gradient penalty term G_ρ represents the additional variance introduced by the penalty scaling on the GVC layers

Effect of partial variance control. Applying gradient variance control only to the final layers means G_ρ only reflects the penalty on those layers, while the early layers contribute the standard δ^2 term. This validates our design choice: applying PGVC everywhere would increase G_ρ^2 unnecessarily, whereas restricting it to the final layers keeps the additional variance term small while targeting the region of highest heterogeneity.

8 Formal Derivation: Equivalence of Theoretical Scaling Factor (Eq. 9) and Gradient Penalty term (Eq. 14)

Below is the step-by-step derivation that shows how the gradient penalty term ρ_i in Eq. 14 arises as the principled, computable realisation of the ideal scalar normalizer ρ_l in Eq. 9, via a second-moment variance-reduction argument that establishes the exact algebraic identity between ρ_i and the gradient of the empirical loss variance.

Gradient variance in the last layers is the quantity to minimize.

Let θ_l denote the parameters of the last-layer subset (S_{gvc}). For client i, define the *per-sample gradient* $v_b \triangleq \nabla_{\theta_l}\ell(\theta_l, x_b)$ and the *mean gradient* $\bar{v}_i \triangleq \frac{1}{B}\sum_{b=1}^{B} v_b = \nabla_{\theta_l} L_i(\theta_l)$. The empirical gradient *variance* across the mini-batch is:

$$\text{Var}_i(\theta_l) \triangleq \frac{1}{B}\sum_{b=1}^{B}\|v_b - \bar{v}_i\|^2 = \frac{1}{B}\sum_{b=1}^{B}\|v_b\|^2 - \|\bar{v}_i\|^2. \tag{70}$$

Reducing Var_i across clients in S_{gvc} is precisely the objective. We want the effective gradient norms in the last layers to be more uniform (approaching the IID reference σ_{iid}).

ρ_i is the exact gradient of the empirical second moment of the loss. Consider the empirical *second moment of the scalar loss* for client i is represented as:

$$\mathcal{V}_i(\theta) \triangleq \frac{1}{2B}\sum_{b=1}^{B}\ell(\theta, x_b)^2. \tag{71}$$

Differentiating with respect to θ using the chain rule:

$$\nabla_\theta \mathcal{V}_i(\theta) = \frac{1}{2B}\sum_{b=1}^{B}\nabla_\theta\left[\ell(\theta, x_b)^2\right] = \frac{1}{2B}\sum_{b=1}^{B} 2\,\ell(\theta, x_b)\cdot\nabla_\theta\ell(\theta, x_b) = \frac{1}{B}\sum_{b=1}^{B}\ell(\theta, x_b)\cdot\nabla_\theta\ell(\theta, x_b). \tag{72}$$

The right-hand side of Eq. 72 is *exactly* ρ_i as defined in Eq. 14. This is an exact algebraic identity, not an approximation, which is represented in Eq. 73.

$$\rho_i = \nabla_\theta \mathcal{V}_i(\theta) = \nabla_\theta\left[\frac{1}{2B}\sum_{b=1}^{B}\ell(\theta, x_b)^2\right] \tag{73}$$

Therefore, using ρ_i as a penalty in the weight update of the last layers (Eq. 15) is equivalent to performing an auxiliary gradient step to minimize the second moment of the loss, directly controlling the variance of per-sample gradients in S_{gvc}. The second moment $\mathcal{V}_i$ satisfies $\mathcal{V}_i \geq \frac{1}{2}\bar{\ell}_i^2 \geq 0$, where $\bar{\ell}_i = \frac{1}{B}\sum_b \ell(\theta, x_b)$ is the mean batch loss. For a client with a highly heterogeneous local distribution, $\bar{\ell}_i$ is large, so $\|\rho_i\|$ is large, meaning the penalty makes a stronger corrective contribution for that client, exactly the behavior intended by Eq. 9.

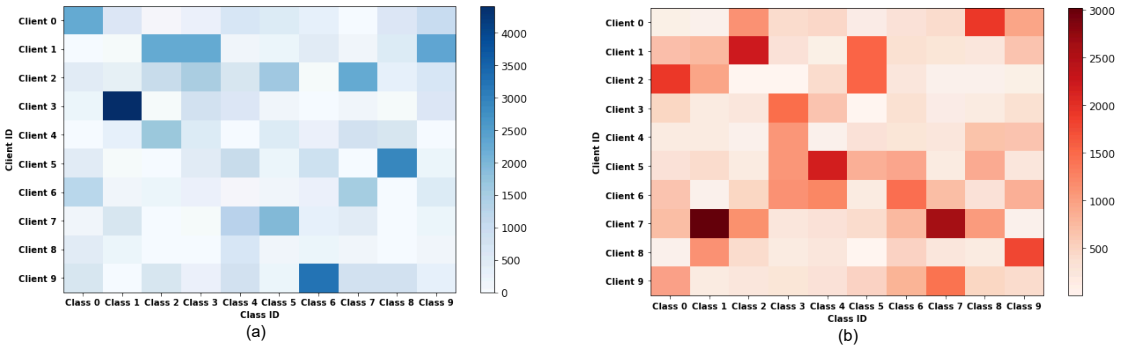

Figure 8: Distribution of MNIST data, indicating the number of images per client per class, varied according to different levels of heterogeneity, with (a) $\alpha = 0.5$ representing severe non-IID and (b) $\alpha = 1.0$ indicating moderate non-IID.

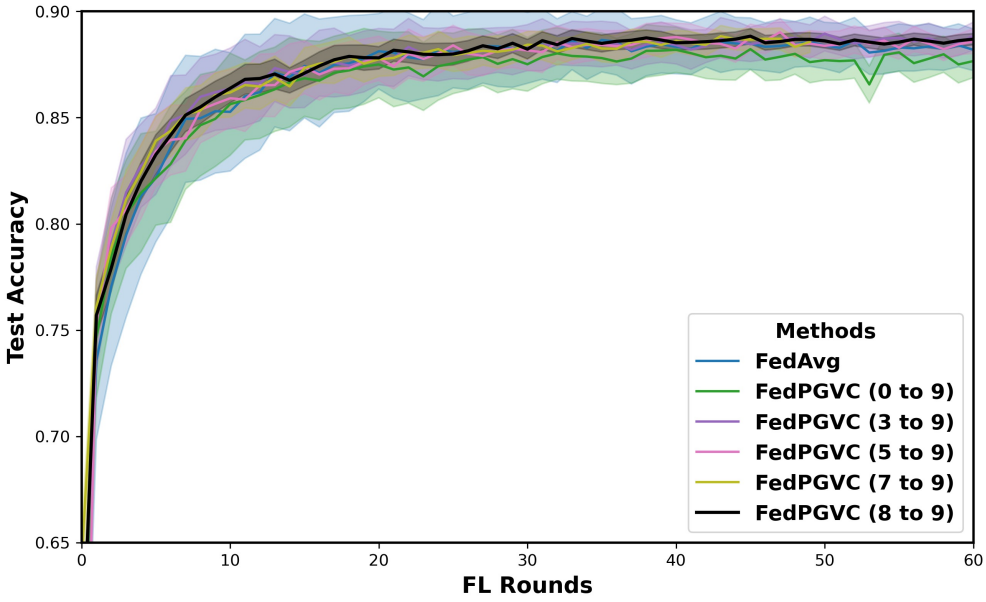

Figure 9: Performance of applying partial gradient variance control on different layers of the CNN model on the FMNIST dataset with $\alpha = 0.5$.

9 Applying PGVC on the different layers of the model

To evaluate the impact of incorporating gradient variance control in various neural network layers, we conducted experiments on the FMNIST dataset with $\alpha = 0.5$. As illustrated in Fig. 9, our findings reveal that applying variance reduction in the final layers accelerates convergence and achieves the highest top-1 accuracy. Given that the proposed approach partially applies the gradient variance control technique in the last layers of the neural network, we investigated the effects of incorporating variance reduction in different layers. We conducted experiments on the FMNIST dataset with $\alpha = 0.5$, as shown in Fig. 9. The results indicate that initiating variance reduction in the final layers of the model facilitates faster convergence and achieves the highest top-1 accuracy.

Table 6: Extended results, including standard deviations, showing the effect of applying the proposed PGVC method to existing popular baselines on the FMNIST and CIFAR100 datasets with $\alpha = 0.5$.

Method	FMNIST	CIFAR100
FedAvg	88.65 ± 0.06	24.25 ± 0.22
FedAvg + PGVC	**88.83 ± 0.11**	**25.29 ± 0.20**
FedProx	86.96 ± 0.22	**24.89 ± 0.17**
FedProx + PGVC	**88.07 ± 0.17**	24.58 ± 0.11
FedBN	**88.86 ± 0.12**	**25.12 ± 0.18**
FedBN + PGVC	88.11 ± 0.15	23.86 ± 0.14
FedNova	87.52 ± 0.11	22.29 ± 0.16
FedNova + PGVC	**87.87 ± 0.14**	**24.82 ± 0.19**

10 Statistical Test

To evaluate the statistical significance of performance differences between the baseline and the proposed method, we conducted McNemar's test (McNemar, 1947). Unlike overall test accuracy, which quantifies correctness at an aggregate level, McNemar's test focuses on pairwise prediction differences, making it

particularly sensitive to variations in error distributions. We performed the test at a 95% confidence level, and the results are presented in Tables 8 and Table 9. In all the cases, the computed p-values were below 0.05, allowing us to reject the null hypothesis. This confirms a statistically significant difference between the baseline and our proposed method.

11 Sensitivity Analysis of FedPGVC with Respect to Local and Global Learning Rates

We conduct a sensitivity analysis of FedPGVC with respect to the local and global learning rates. Specifically, we first fix the local learning rate $\eta_l = 10^{-4}$ and vary the global learning rate η_g from 10^{-5} to 1, recording the corresponding performance. Next, we fix $\eta_g = 1$ and vary η_l over the same range. We observe that the best performance is achieved at $\eta_l = 10^{-4}$ and $\eta_g = 1$. This suggests that FedPGVC benefits from a relatively small (but not excessively small) local learning rate, coupled with a larger global learning rate. Please refer to Table 7 for the results.

Table 7: Sensitivity of FedPGVC to local and global learning rates on FMNIST ($\alpha = 1.0$). Here * denotes the respective combination did not converge.

η_g	η_l	Test Accuracy (%)
0.1	0.0001	56.33
0.0001	0.0001	51.95
0.5	0.0001	65.75
0.001	0.0001	55.83
1	0.0001	89.35
1	0.0001	76.05
1	0.1	*
1	0.5	*
1	1	*
1	0.001	89.35

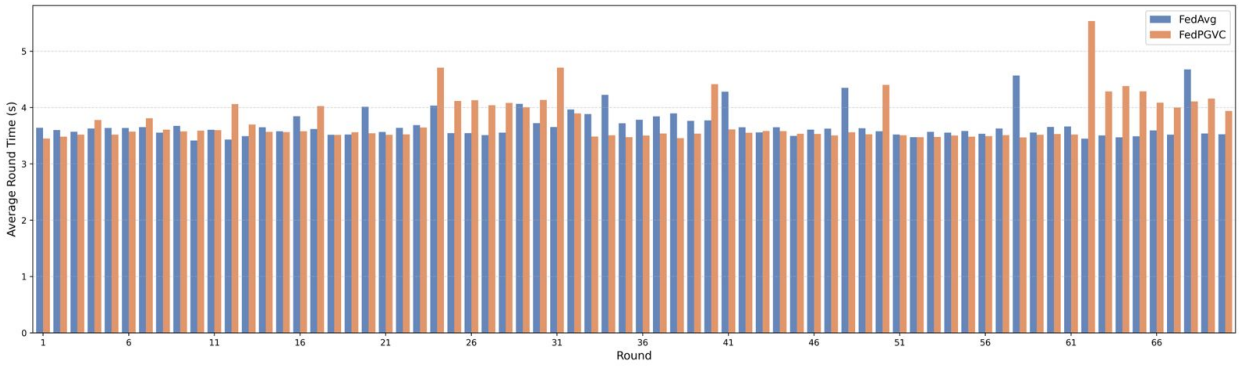

Figure 10: Wall-clock performance (in seconds) of the proposed method and the FedAvg.

12 Runtime Analysis

We measure the wall-clock time per client for local training and average it across clients for each communication round until convergence, using FMNIST with LeNet ($\alpha = 1.0$). As shown in Fig. 10, per-round times are comparable: FedAvg averages ∼3.6s, while FedPGVC ranges between ∼3.6–3.8s, since the gradient penalty ρ is applied only to the final layers and reuses the same batch, introducing negligible overhead. Despite similar per-round costs, FedPGVC converges in significantly fewer rounds (e.g., 4.3× faster on MNIST, Table 3),

reducing overall training time. For instance, on FMNIST dataset, although FedPGVC may take longer for a fixed number of rounds, it achieves the same accuracy in fewer rounds, making its effective training time competitive. The instances where FedAvg shows higher round times than FedPGVC are consistent with the stochastic nature of federated training. Since round completion time is determined by the slowest client in a synchronous setting, variability in local data batches, gradient norms, and system-level factors across 10 clients naturally causes fluctuations in both methods.

13 Limitation

While the proposed approach demonstrates significant improvements across various datasets, it is essential to acknowledge certain limitations. Although the method effectively reduces gradient variability in the final layers, the computation of the gradient penalty term may introduce additional processing overhead on the client side. This increased computational cost could be a constraint for resource-limited devices, even though the method is designed to minimize communication overhead. To address this challenge, future research could focus on reducing the computational complexity of the gradient penalty term without sacrificing the method's effectiveness. Approaches like model pruning or quantization may be explored to make the method more viable for devices with limited resources. Additionally, incorporating differential privacy techniques into the FedPGVC framework could broaden its applicability in privacy-sensitive areas, such as healthcare or finance. Investigating the interaction between differential privacy and the gradient penalty term, as well as its impact on model performance, would be a valuable direction for future research.

Table 8: p-values indicating the statistical significance of test accuracies on MNIST, FMNIST, and CIFAR100 under varying data heterogeneity, relative to the Proposed method.

	MNIST		FMNIST		CIFAR100	
	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 1.0$
FedAvg	3.21e-12	7.45e-9	2.12e-8	5.33e-7	8.54e-15	9.18e-11
FedProx	6.78e-10	4.23e-8	1.29e-7	2.77e-6	2.61e-6	5.44e-9
FedNova	5.21e-9	1.14e-6	3.87e-6	6.45e-5	4.73e-8	1.07e-9
FedBN	4.12e-8	9.67e-7	7.32e-6	8.89e-5	3.21e-8	5.56e-10
SCAFFOLD	1.78e-7	5.22e-6	2.89e-5	4.76e-4	9.33e-10	2.23e-9
FedPVR	2.34e-6	8.73e-5	1.11e-4	3.12e-3	5.67e-7	6.88e-6

Table 9: p-values indicating significance compared to the Proposed method across various complex models on CIFAR-100, Tiny-ImageNet, and QQP datasets.

	CIFAR100 (ResNet18)	CIFAR100 (ViT)	Tiny-ImageNet (ResNet18)	QQP (LSTM)
FedAvg	2.21e-4	3.41e-6	5.12e-5	1.87e-3
FedProx	5.14e-5	2.89e-7	3.78e-4	6.92e-4
FedNova	1.33e-6	4.57e-9	1.02e-6	3.10e-5
FedBN	3.87e-4	1.12e-6	8.45e-5	2.43e-3
SCAFFOLD	4.75e-4	2.65e-5	7.89e-6	3.20e-4
FedPVR	9.56e-8	7.22e-11	1.31e-9	9.87e-6

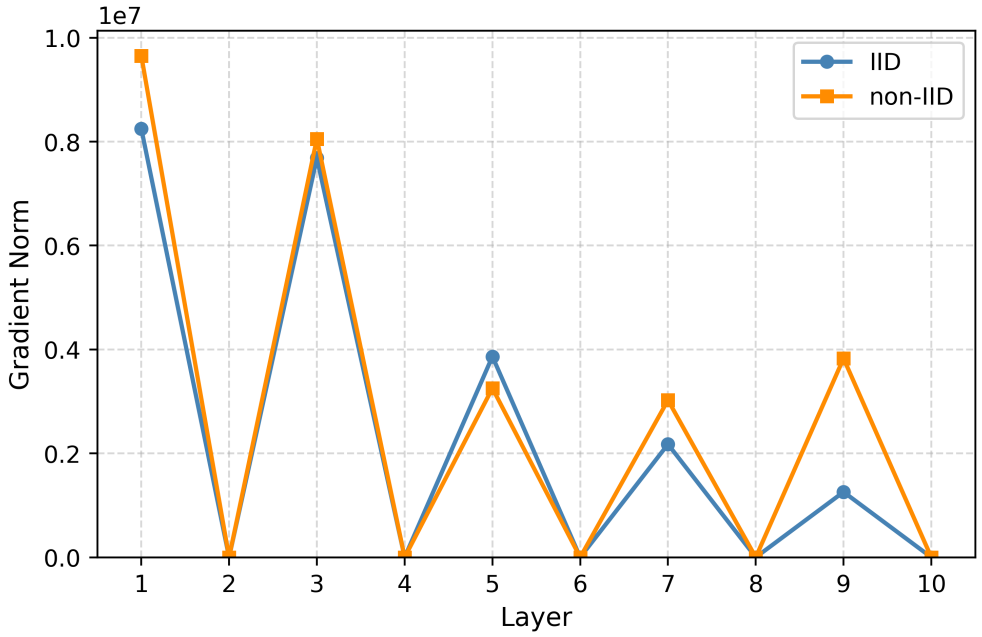

Figure 11: Comparison of gradient norms for IID and non-IID models trained using FedAvg, with a second run under a different random seed.

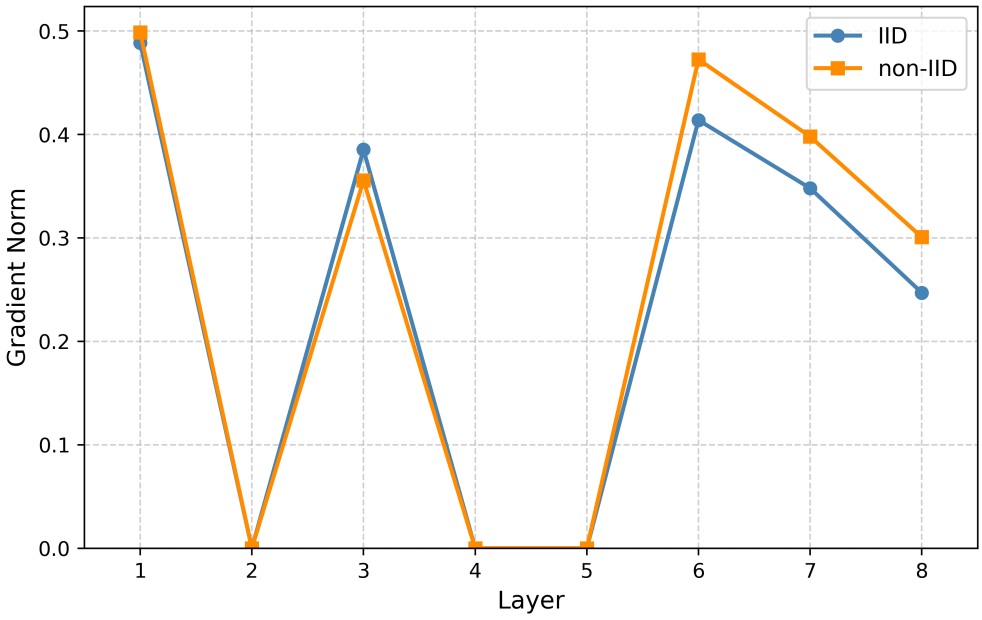

Figure 12: Comparison of gradient norms for IID and non-IID models trained using FedAvg on the FMNIST dataset with a LeNet architecture.

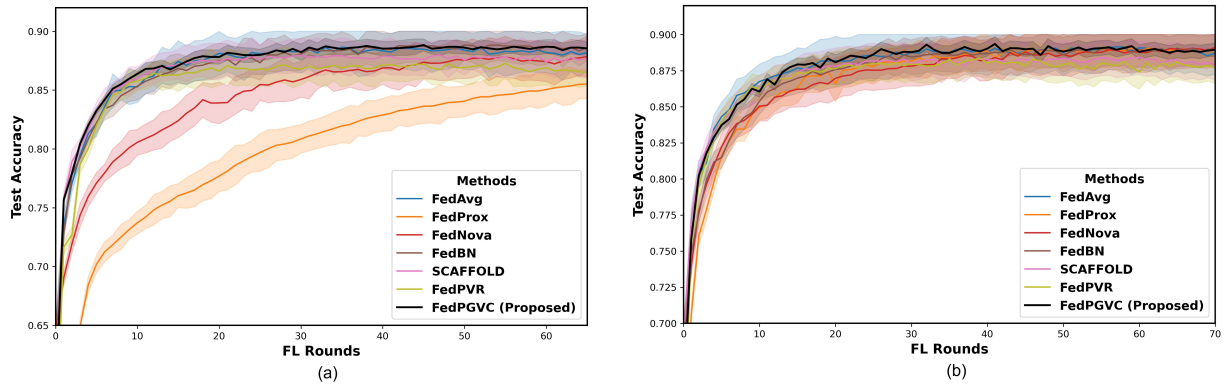

Figure 13: Performance comparison of proposed FedPGVC with baseline approaches with error bars: (a) and (b) depict the graphs on the FMNIST dataset for $\alpha = 0.5$ and 1.0 respectively.

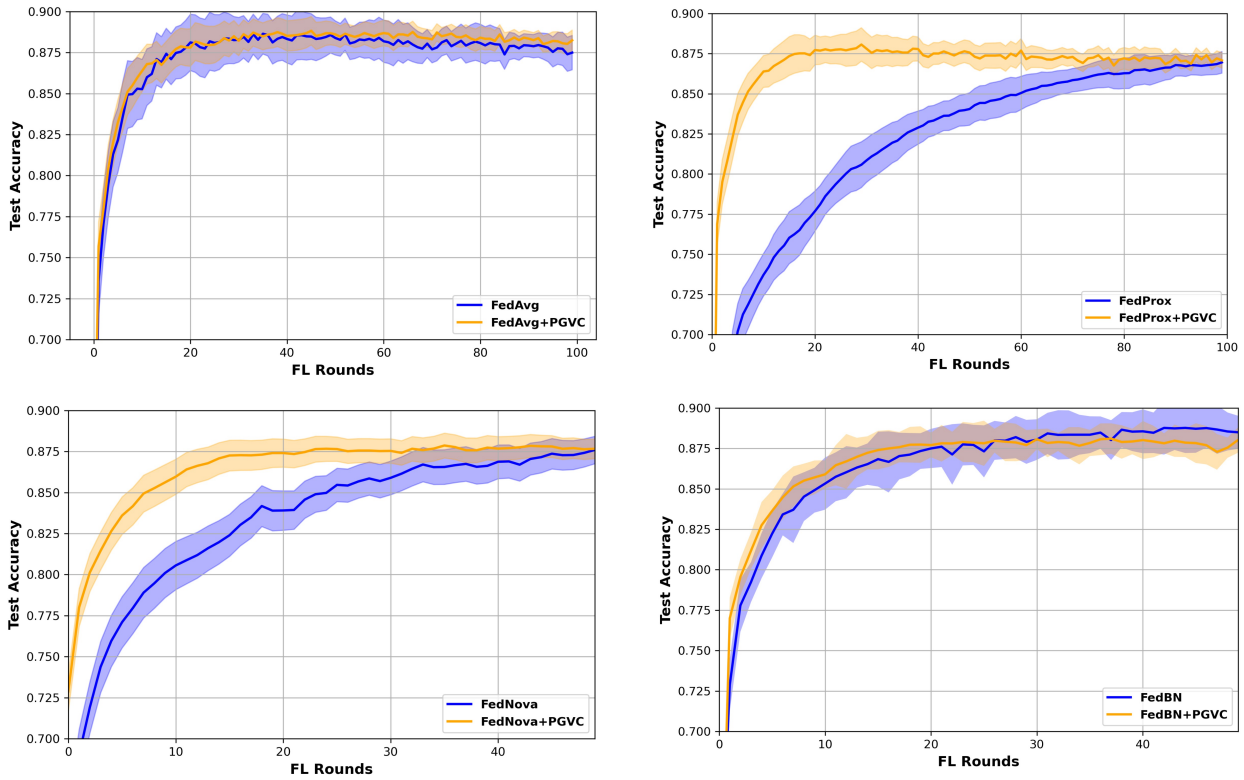

Figure 14: Learning curves with error bars illustrating the integration of the proposed PGVC technique with the existing algorithms on the FMNIST dataset at $\alpha = 0.5$.

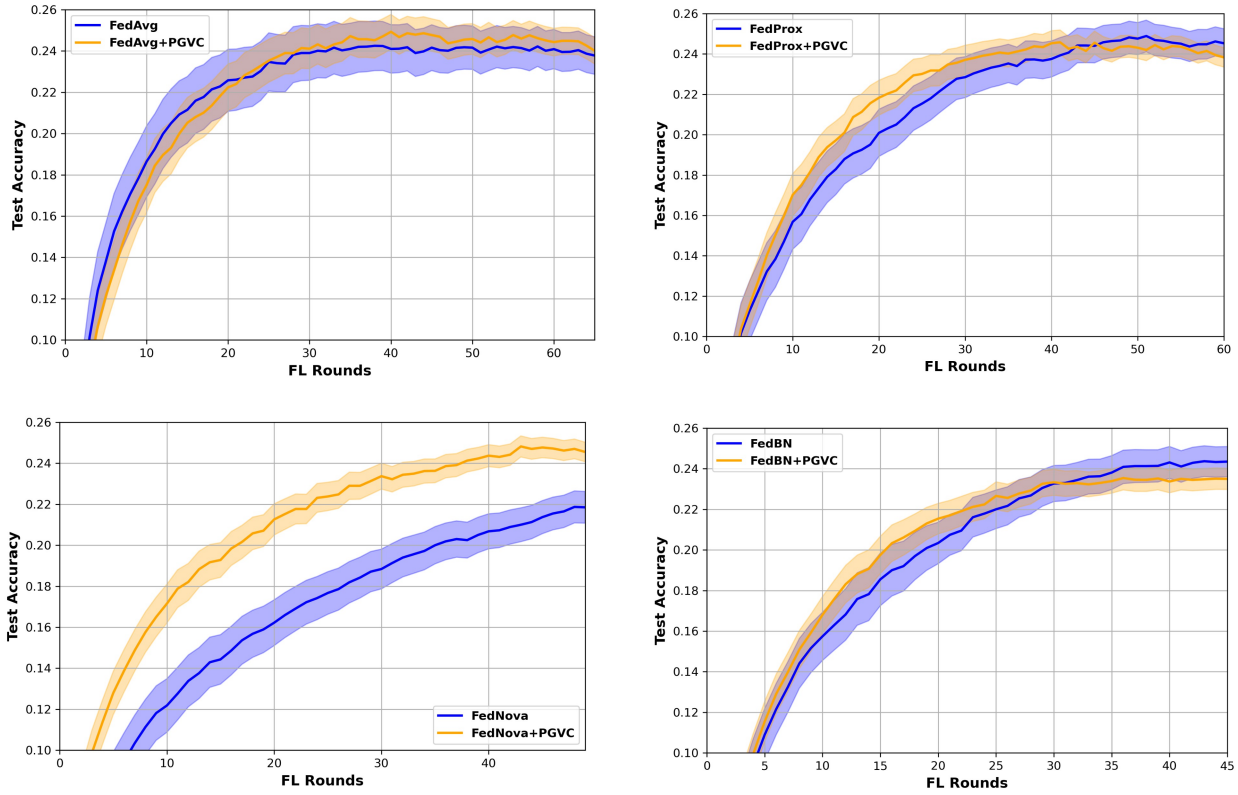

Figure 15: Learning curves with error bars illustrating the integration of the proposed PGVC technique with the existing algorithms on the CIFAR100 dataset at $\alpha = 0.5$.

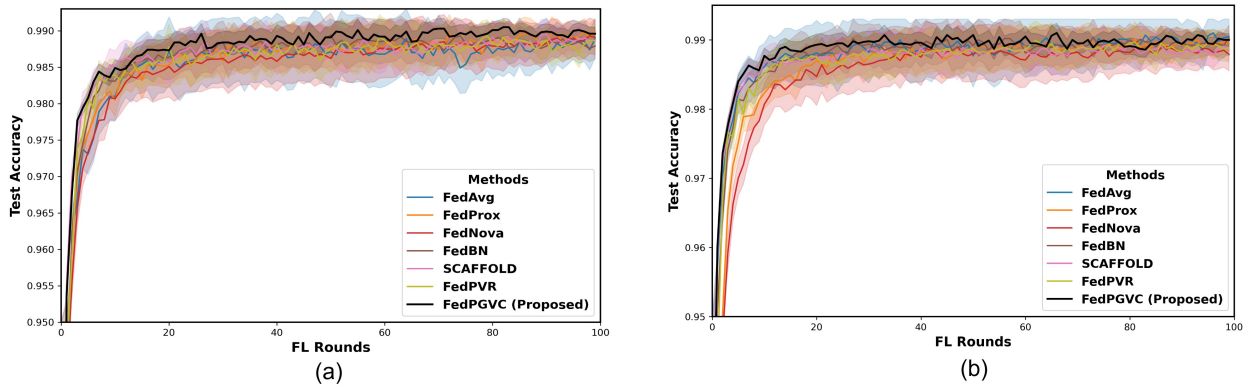

Figure 16: Performance comparison of proposed FedPGVC with baseline approaches with error bars: (a) and (b) depict the graphs on the MNIST dataset for $\alpha = 0.5$ and 1.0 respectively.

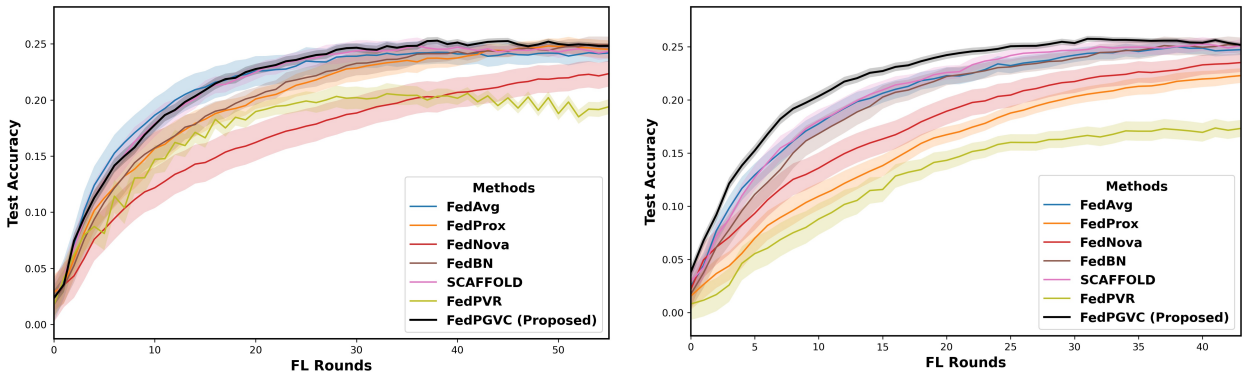

Figure 17: Performance comparison of proposed FedPGVC with baseline approaches with error bars: (a) and (b) depict the graphs on the CIFAR100 dataset for $\alpha = 0.5$ and 1.0 respectively.

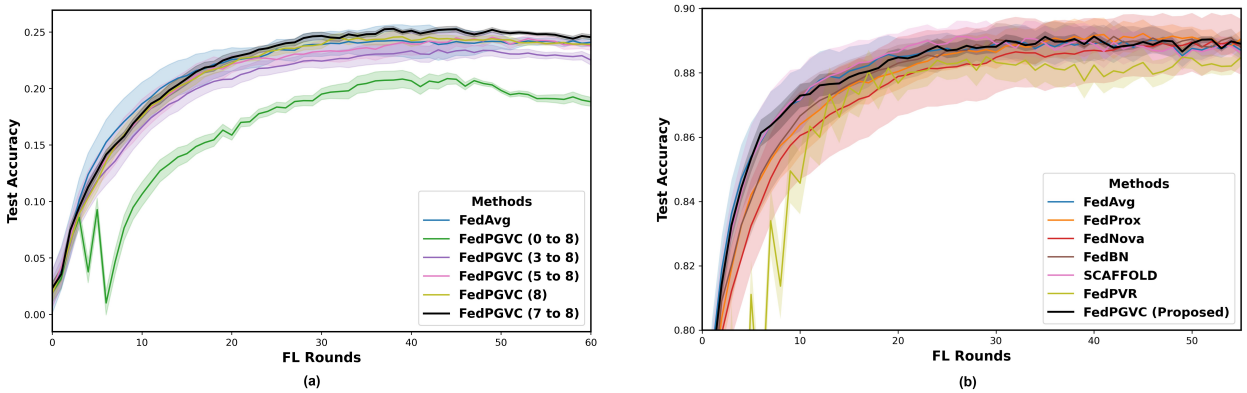

Figure 18: (a) Performance of applying partial gradient variance control on different layers of the CNN model with error bars on the CIFAR100 dataset with $\alpha = 0.5$. (b) Performance of the proposed model compared to the baselines with error bars in FMNIST IID data settings with $\alpha = 100$.