DECENTRALIZED PERSONALIZED FEDERATED LEARNING WITH COMPRESSED MESSAGES VIA FINITE-TIME QUANTIZED COORDINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized Federated Learning (DFL) enables privacy-preserving collaborative training without a central server. However, it faces two major challenges: statistical data heterogeneity, which degrades global model performance on local tasks, and limited communication bandwidth, which hinders collaboration and degrades model accuracy. To address these issues, we propose Q-pFedMe, a decentralized personalized FL algorithm with quantized communication. Q-pFedMe decouples local personalized model optimization from global reference learning using Moreau envelopes. Furthermore, it embeds fixed-point quantization into finite-time averaging, making all nodes reach exact agreement on a common quantized aggregate. We theoretically prove that Q-pFedMe converges linearly to a quantization-dependent neighborhood of the optimal solution for strongly convex objectives, and has an $\mathcal{O}(1/R)$ optimization term with an explicit error floor for smooth non-convex objectives. We also derive a finite-time upper bound on the number of communication steps required by the quantized average consensus subroutine at each global round. Experiments on Synthetic, MNIST, and CIFAR-10 show that Q-pFedMe achieves accuracy comparable to centralized unquantized pFedMe. Additional CIFAR-100 experiments on two directed topologies demonstrate the highest personalized accuracy and lowest per-transmission payload among the evaluated decentralized baselines. Our code is available at `https://anonymous.4open.science/r/N8HgpH8UoWHorEt/`.

1 INTRODUCTION

Federated learning (FL) enables multiple nodes to collaboratively train machine learning models without centrally collecting their raw data training McMahan et al. (2017). To avoid the communication bottlenecks and single-point failures of server-based FL Kairouz & McMahan (2021), decentralized FL (DFL) uses peer-to-peer communication, making it ideal for edge networks Lian et al. (2017). However, DFL faces two primary challenges: non-i.i.d. data degrades global model performance on local tasks Li et al. (2020); Karimireddy et al. (2020), and limited bandwidth bottlenecks high-dimensional parameter exchanges, hindering collaboration Reisizadeh et al. (2020). This has motivated the development of communication-efficient personalized DFL algorithms.

Under statistical heterogeneity, a single global model performs poorly on local tasks Li et al. (2020); Fallah et al. (2020), and training purely local models suffers from poor generalization. To resolve this, personalized federated learning (PFL) learns customized models for each node by leveraging collective network knowledge Tan et al. (2022). PFL is typically formulated as a bi-level optimization problem, decoupling the learning of local personalized models from the learning of a global reference model T Dinh et al. (2020).

A rich body of literature has explored PFL. Prominent centralized methods employ meta-learning frameworks Fallah et al. (2020); Shamsian et al. (2021) or regularization-based approaches Li et al. (2020); Deng et al. (2020); Li et al. (2021); T Dinh et al. (2020). However, these methods rely on a central server, making them vulnerable to communication bottlenecks and single points of failure.

In resource-constrained environments, exchanging full-precision model parameters is often infeasible. Recent communication-efficient decentralized PFL methods rely on sparsification to transmit

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

parameter subsets Dai et al. (2022); Zhang et al. (2025); Long et al. (2025); Chang et al. (2025), knowledge distillation to share representations Jeong & Kountouris (2023); Xiong et al. (2024); Park et al. (2025), or network-aware structural adaptations Liu et al. (2024); Shi et al. (2023); Li et al. (2025); Lin et al. (2024); Ma et al. (2022); Fan et al. (2025). Despite their benefits, the aforementioned approaches possess fundamental limitations. Sparsification drops critical parameters, preventing exact node agreement on shared global variables. Distillation relies on heuristic proxies or architectural alterations rather than compressing the full model. Furthermore, most existing methods assume restrictive undirected or doubly stochastic topologies (which may not be feasible in practical peer-to-peer deployments) and even when communication is reduced they still exchange full-precision real values, making them impractical for real-world, bandwidth-limited peer-to-peer deployments. To compress the full model without discarding parameters, quantization is a natural solution. However, while rigorously studied in standard decentralized learning, strict mathematical quantization remains vastly underexplored in decentralized PFL. The few recent attempts to address communication efficiency in this setting typically rely on heuristic communication budgets, lacking a rigorous bit-level theoretical analysis Kharrat et al. (2025).

To the best of the authors' knowledge, no existing method simultaneously achieves personalization, full decentralization, and exact finite-time agreement over strictly quantized communication with rigorous theoretical guarantees. Thus, designing a decentralized personalized learning framework that integrates bi-level optimization with strict bit-level quantization remains a critical open problem.

Main Contributions. Motivated by this gap, we propose Q-pFedMe, a decentralized personalized FL framework combining Moreau-envelope personalization with finite-time quantized coordination. Quantization is embedded in DQuFTA (Algorithm 2), rather than applied as an independent compression heuristic. Nodes exchange fixed-point integer messages with a lower per-transmission payload than dense FP32 communication. Q-pFedMe is therefore designed for communication settings with finite-precision or per-transmission payload constraints. Our contributions are as follows.
• **Algorithmic Design.** We propose Q-pFedMe, a decentralized personalized FL algorithm which combines Moreau-envelope-based bi-level personalization with a quantized finite-time averaging protocol (see Algorithm 1 (Q-pFedMe)). By reformulating decentralized personalized FL as a bi-level optimization problem, Q-pFedMe decouples local personalized model training from global reference learning. Additionally, nodes communicate only by transmitting quantized values and achieve exact agreement on the global reference model in finite time.
• **Convergence Guarantees.** Under standard assumptions, we establish rigorous convergence bounds for Q-pFedMe in both strongly convex and smooth non-convex settings. For strongly convex objectives, Q-pFedMe converges linearly to a neighborhood of the optimal solution (see Theorem 1). For smooth non-convex objectives, it achieves a sublinear rate of $\mathcal{O}(1/R)$ (see Theorem 2), where R denotes the total number of global communication rounds. Both bounds explicitly quantify the influence of the quantization level Δ, the number of local updates B, the model dimension n, the data heterogeneity parameter σ_f, and the divergence of stochastic gradients γ_f. A key insight is that the perfect agreement property, allows the bounds to match those of the centralized and unquantized pFedMe in T Dinh et al. (2020) up to an additive error that grows with the quantization level Δ.
• **Finite-Time Coordination Analysis.** We characterize the number of communication steps required by DQuFTA (see Algorithm 2) to achieve agreement on a common quantized aggregate within each global round (see Theorem 3). The resulting bound depends on the network size, network diameter, quantization level, and the optimization states entering the coordination procedure.
• **Empirical Validation.** We conduct experiments on Synthetic, MNIST, and CIFAR-10 datasets under different data heterogeneity settings. Q-pFedMe achieves test accuracy strictly comparable to the centralized, unquantized pFedMe, with a average performance gap of 0.43%. The empirical results validate our theoretical predictions: accuracy degrades smoothly and predictably only when the quantization level exceeds a dataset-specific critical threshold, and the convergence behavior exhibits the trade-off governed by the global update parameter β as characterized by our bounds. Additionally, Q-pFedMe achieves the highest accuracy and per-transmission payload reduction on CIFAR-100 under two topologies, compared with the evaluated decentralized methods.

2 Notations and Preliminaries

Let $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Q}_{>0}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the sets of real, rational, positive rational, integer, and natural numbers, respectively. We use $\lfloor a \rfloor$ and $\lceil a \rceil$ for the floor and ceiling of $a \in \mathbb{R}$. Finally, $\|\cdot\|$, $\mathbb{E}[\cdot]$, and

∇f denote the Euclidean norm, expectation, and gradient.

Network Notations. The communication network is a digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of N nodes. A directed edge $(v_l, v_i) \in \mathcal{E}$ indicates that node v_i can transmit to v_l. The sets of in- and out-neighbors of v_i are $\mathcal{N}_i^- = v_j \mid (v_i, v_j) \in \mathcal{E}$ and $\mathcal{N}_i^+ = v_l \mid (v_l, v_i) \in \mathcal{E}$, with respective degrees $D_i^- = |\mathcal{N}_i^-|$ and $D_i^+ = |\mathcal{N}_i^+|$. The graph diameter D is the longest shortest path between any two nodes, and $\mathcal{G}$ is strongly connected if a directed path exists between every pair of nodes.

Asymmetric Uniform Quantizer. Quantization mitigates communication bottlenecks by reducing transmitted bits. While our results extend to other quantizers (e.g., logarithmic), we focus on the asymmetric uniform scalar quantizer. Considering $\xi \in \mathbb{R}$ as the value to be quantized and $\Delta \in \mathbb{Q}_{>0}$ as the quantization level, we have that the quantized version of ξ is $q_\Delta(\xi) = \Delta \left\lfloor \frac{\xi}{\Delta} \right\rfloor \in \mathbb{Q}$.

3 Problem Formulation of Personalized Decentralized Learning

Let us consider a network represented as a digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ nodes. Each node v_i is endowed with a personal model $x_i \in \mathbb{R}^n$ and a local cost function $f_i : \mathbb{R}^n \to \mathbb{R}$. Let $\mathbf{x} = [x_1^\top, x_2^\top, \ldots, x_N^\top]^\top \in \mathbb{R}^{nN}$ represent the concatenation of local variables of the N nodes. Each node stores only its local estimation parameter x_i, can access its cost function f_i, and can exchange information with neighbors. Additionally, communication is subject to limited channel capacity, and nodes exchange only quantized messages. In traditional decentralized learning, the nodes cooperatively aim to solve the following problem:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^{nN}} \quad & f(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} f_i(x_i), \\ \text{s.t.} \quad & x_i = x_j, \ \forall v_i, v_j \in \mathcal{V}, \\ & \text{nodes communicate with quantized values.} \end{aligned} \tag{1}$$

The function f_i denotes the expected loss over the data distribution of node v_i and is defined as $f_i(x_i) = \mathbb{E}_{d_i}\big[\tilde{f}_i(x_i; d_i)\big]$, where d_i is a random data sample drawn according to the distribution of node v_i and $\tilde{f}_i(x_i; d_i)$ is a loss function corresponding to this sample and x_i. Since nodes' data possibly come from different environments, contexts, and applications, nodes can have non-i.i.d. data distributions (i.e., the distributions of d_i and d_j for distinct nodes $v_i, v_j \in \mathcal{V}$ are distinct). Instead of solving the traditional decentralized learning problem equation 1, we adopt the personalized formulation of T Dinh et al. (2020) based on regularized local losses. The personalized decentralized learning problem is defined as the following bi-level problem:

$$\min_{\mathbf{x} \in \mathbb{R}^{nN}} \quad F(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} F_i(x_i) \tag{2a}$$

$$\text{s.t.} \quad x_i = x_j, \quad \forall v_i, v_j \in \mathcal{V}, \tag{2b}$$

$$F_i(x_i) = \min_{\theta_i \in \mathbb{R}^n} \left\{ f_i(\theta_i) + \frac{\lambda}{2} \|\theta_i - x_i\|^2 \right\}, \tag{2c}$$

$$\text{nodes communicate with quantized values.} \tag{2d}$$

In the inner problem equation 2c, each node v_i optimizes a personalized model θ_i that fits its local data while remaining close to a reference model x_i. At the outer level, the reference models x_i are driven toward consensus via the decentralized algorithm, while simultaneously minimizing the average of the Moreau envelopes $F_i(x_i)$. The hyperparameter $\lambda > 0$ controls the trade-off between personalization and global agreement (larger λ forces personalized models to stay closer to the consensus reference). The function $F_i(x_i)$ is the *Moreau envelope* of f_i, which is smooth and convexity-preserving. The optimal personalized model for a given reference x_i is known as the *proximal operator*, and given by

$$\hat{\theta}_i(x_i) := \operatorname{prox}_{f_i/\lambda}(x_i) = \arg\min_{\theta_i \in \mathbb{R}^n} \left\{ f_i(\theta_i) + \frac{\lambda}{2} \|\theta_i - x_i\|^2 \right\}. \tag{3}$$

Operational Assumptions. For the development of our results we make the following assumptions, which are standard in the federated and decentralized learning literature T Dinh et al. (2020); Li et al. (2020); Karimireddy et al. (2020); Lian et al. (2017); Rikos et al. (2022); Koloskova et al. (2020).

Assumption 1 (Lian et al. (2017); Koloskova et al. (2020)). *The network is modeled as a strongly connected digraph $\mathcal{G}$.*

Assumption 2 (Rikos et al. (2022); Bastianello et al. (2023)). *Every node $v_i \in \mathcal{V}$ has knowledge of the diameter D of the underlying digraph (or an upper bound $D' \geq D$) and a common quantization level Δ.*

Assumption 3 (T Dinh et al. (2020)). *The local cost functions f_i are either (a) μ-strongly convex or (b) nonconvex and L-smooth (i.e., have L-Lipschitz continuous gradients). Specifically, for any $x, x' \in \mathbb{R}^n$, either:*

(a) $f_i(x) \geq f_i(x') + \langle \nabla f_i(x'), x - x' \rangle + \frac{\mu}{2}\|x - x'\|^2$, *or*

(b) $\|\nabla f_i(x) - \nabla f_i(x')\| \leq L\|x - x'\|$.

Assumption 4 (T Dinh et al. (2020); Karimireddy et al. (2020)). *The variance of the stochastic gradients at each node is uniformly bounded. Specifically, there exists a constant $\gamma_f > 0$ such that for any $x \in \mathbb{R}^n$ and $i \in \{1, \ldots, N\}$ we have:*

$$\mathbb{E}_{d_i}\left[\|\nabla \tilde{f}_i(x; d_i) - \nabla f_i(x)\|^2\right] \leq \gamma_f^2.$$

Assumption 5 (T Dinh et al. (2020); Li et al. (2020)). *The divergence of the local gradients from the global full-batch gradient is uniformly bounded. Specifically, there exists a constant $\sigma_f > 0$ such that for any $x \in \mathbb{R}^n$,*

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma_f^2,$$

where $\nabla f(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x)$.

Assumption 1 ensures the network-wide information propagation necessary to solve equation 2. Assumption 2 provides to nodes the diameter D for finite-time detection of agreement, and a common quantization level Δ for communication via quantized messages. Before the algorithm begins, nodes can compute the network diameter D and agree upon a common quantization level Δ by executing standard finite-time distributed protocols such as min-consensus, and Oliva et al. (2016), respectively, during the initialization procedure. Assumption 3 introduces standard smoothness and convexity conditions. Assumptions 4 and 5 provide standard decentralized learning bounds, where γ_f^2 limits stochastic sampling noise and σ_f^2 limits gradient divergence from data heterogeneity.

4 Personalized Decentralized FL with Finite Time Quantized Coordination

We propose Algorithm 1 (Q-pFedMe) to solve the personalized decentralized FL problem in (2).

Intuition behind Algorithm 1 (Q-pFedMe). At each global round r, every node first refines its local reference model by performing B steps of stochastic gradient descent on the Moreau envelope F_i. According to Proposition 1, the gradient $\nabla F_i(x)$ equals $\lambda(x - \hat{\theta}_i(x))$, where $\hat{\theta}_i(x)$ is the proximal operator that produces a personalized model tailored to the node's local data. Thus, each local update moves the reference model towards the personalized model, in order to balance local adaptation with the global consensus objective. After the local updates, the nodes engage in a quantized averaging phase (executing Algorithm 2 (DQFTA)) to align their reference models. Because communication is restricted to quantized messages, Algorithm 2 (DQuFTA) guarantees finite-time convergence to a common quantized average, up to a controlled error determined by the quantization step Δ. Finally, after executing Algorithm 2 (DQuFTA), each node performs a consensus-based reference-model update, with $\beta \geq 1$ controlling the degree of over-relaxation. The update reduces to the consensus state when $\beta = 1$ and is over-relaxed when $\beta > 1$.

Intuition of Algorithm 2 (DQuFTA). Algorithm 2 (DQuFTA) is a fully decentralized quantized averaging algorithm that exhibits finite time convergence. It combines the quantized averaging procedure from Rikos et al. (2022) with a distributed stopping mechanism based on min- and max-consensus evaluated every D steps. Given input states $x_{i,B}^r$, $\forall v_i \in \mathcal{V}$, DQuFTA outputs the exact quantized average $q_i^r = \bar{q}^r$ for all nodes after T communication steps (that we bound in Theorem 3).

Algorithm 1 Q-pFedMe: Quantized Personalized Federated Learning with Moreau Envelopes

Input: A strongly connected digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ nodes and $m = |\mathcal{E}|$ edges. Each node $v_i \in \mathcal{V}$ has initial state $x_i^0 \in \mathbb{R}^n$ for which it holds $x_i^0 = x_j^0, \forall v_i, v_j \in \mathcal{V}$, loss function f_i, diameter D, and quantization level $\Delta \in \mathbb{Q}$, local learning rate η, regularization parameter λ, step size β, and local steps B.

Output: Each node $v_i \in \mathcal{V}$ calculates $x_i^* = x_i^R$ which solves problem equation 2.

for $r = 0$ to $R - 1$ **do**
  for all nodes $v_i \in \mathcal{V}$ in parallel **do**
    $x_{i,0}^r = x_i^r$;
    for $b = 0$ to $B - 1$ **do**
      Sample a mini-batch $\mathcal{D}_i$ with size $|\mathcal{D}|$, and minimize $\tilde{h}_i(\theta_i; x_{i,b}^r, \mathcal{D}_i)$ (defined in equation 4) up to an accuracy level to find a δ-approximate $\tilde{\theta}_i(x_{i,b}^r)$;
      $x_{i,b+1}^r = x_{i,b}^r - \lambda\eta(x_{i,b}^r - \tilde{\theta}_i(x_{i,b}^r))$;
    end for
  end for
  $x_i^{r+1} = (1 - \beta)x_i^r + \beta$(Algorithm 2$(x_{i,B}^r, D, \Delta)$). (see Appendix B)
end for

Approximating the Proximal Operator. In practice, computing the exact proximal operator $\hat{\theta}_i(x)$ is infeasible, as it requires minimizing the true local objective f_i. Instead, each node computes a δ-approximate solution $\tilde{\theta}_i(x_{i,b}^r)$ using the first-order method on the stochastic objective

$$\tilde{h}_i(\theta_i; x_{i,b}^r, \mathcal{D}_i) := \tilde{f}_i(\theta_i; \mathcal{D}_i) + \frac{\lambda}{2}\|\theta_i - x_{i,b}^r\|^2, \tag{4}$$

where $\mathcal{D}_i$ is a sampled mini-batch. The inner solver (e.g., gradient descent) is run until the we have

$$\|\nabla \tilde{h}_i(\tilde{\theta}_i; x_{i,b}^r, \mathcal{D}_i)\|^2 \leq \nu, \tag{5}$$

which guarantees that the expected squared error satisfies $\mathbb{E}\big[\|\tilde{\theta}_i(x_{i,b}^r) - \hat{\theta}_i(x_{i,b}^r)\|^2\big] \leq \delta^2$ (see Lemma 1). The number of inner iterations required is $K = \mathcal{O}\big(\kappa \log(d/\nu)\big)$, where κ is the condition number of $\tilde{h}_i$. The gradient of the Moreau envelope is then utilized as equation 6 following (T Dinh et al., 2020, Proposition 1).

Comparison with the Literature. Q-pFedMe differs from existing centralized and decentralized methods in both its personalization mechanism and communication protocol. Unlike FedAvg McMahan et al. (2017) and Per-FedAvg Fallah et al. (2020), Q-pFedMe operates without a central server, while unlike decentralized FedAvg and D-PSGD Lian et al. (2017), it optimizes a Moreau-envelope personalized objective rather than learning a single consensus model. DisPFL Dai et al. (2022) achieves personalization and communication reduction through client-specific sparse masks, transmitting only active model coordinates. DFedPGP Liu et al. (2024) instead personalizes the classifier and applies gradient push to learn a shared representation over directed, potentially time-varying networks. In contrast, Q-pFedMe communicates all model coordinates using finite-precision quantized messages and establishes exact agreement on a common quantized average over a strongly connected digraph. This preserves a shared reference model without sparsifying its coordinates while providing full-model personalization through the Moreau-envelope formulation. The resulting guarantee comes at the cost of multiple inner consensus transmissions per outer round; consequently, Q-pFedMe targets per-transmission payload reduction and exact quantized coordination rather than minimum cumulative network traffic.

5 Convergence Analysis

Inspired by T Dinh et al. (2020), we present the following proposition and lemma for our analysis.

Proposition 1 (T Dinh et al. (2020))**.** *If f_i is convex, or if it is nonconvex and L-smooth, then the Moreau envelope F_i is L_F-smooth with $L_F = \lambda$ (provided that $\lambda > 2L$ in the nonconvex case), and its gradient is*

$$\nabla F_i(x) = \lambda\big(x - \hat{\theta}_i(x)\big). \tag{6}$$

Furthermore, if f_i is μ-strongly convex, then F_i is μ_F-strongly convex with $\mu_F = \frac{\lambda\mu}{\lambda+\mu}$.

Lemma 1 (T Dinh et al. (2020)). *Let $\tilde{\theta}_i(x_{i,b}^r)$ is a solution to (5), we have*

$$\mathbb{E}\big[\|\tilde{\theta}_i(x_{i,b}^r) - \hat{\theta}_i(x_{i,b}^r)\|^2\big] \leq \delta^2 := \begin{cases} \frac{2}{(\lambda+\mu)^2}\big(\frac{\gamma_f^2}{|\mathcal{D}|} + \nu\big), & \textit{if Assumption 3(a) holds,} \\ \frac{2}{(\lambda-L)^2}\big(\frac{\gamma_f^2}{|\mathcal{D}|} + \nu\big), & \textit{if Assumption 3(b) holds, and } \lambda > L. \end{cases} \tag{7}$$

Before presenting our main theorems, let us define $\sigma_{F,1}^2$, and $\sigma_{F,2}^2$, as $\sigma_{F,1}^2 := \frac{2}{N}\sum_{i=1}^N \|\nabla F_i(x^*)\|^2$, and $\sigma_{F,2}^2 := 2\frac{\lambda^2}{\lambda^2 - 8L^2}\sigma_f^2$, for the strongly convex, and the non-convex case, respectively.

Theorem 1 (Convergence under Strong Convexity). *Let Assumptions 1, 2, 3(a), and 4 hold. Suppose the effective step size satisfies $\frac{2}{2+\mu_F} \leq \tilde{\eta} \leq \hat{\eta}_1$ with $\tilde{\eta} = \eta\beta B$ and $\hat{\eta}_1 = \frac{1}{6L_F(1+128\kappa_F/\beta)}$. Then we have the following convergence guarantees.*

(a) **(Global model convergence)** *The global model satisfies:*

$$\begin{aligned} \mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \leq & \mu_F e^{-\tilde{\eta}\mu_F R/4}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \underbrace{\Big(1 + \frac{\mu_F}{4}\Big)\frac{16\delta^2\lambda^2}{\mu_F}}_{=:C_1} \\ & + \frac{\tilde{\eta}^2}{\beta^2}\underbrace{\Big(1 + \frac{\mu_F}{4}\Big)\frac{256\big(B\sigma_{F,1}^2 + \delta^2\lambda^2\big)\kappa_F}{B}}_{=:C_2} + \frac{\beta^2}{\tilde{\eta}}\underbrace{\Big(1 + \frac{4}{\mu_F}\Big)4n\Delta^2}_{=:C_3} \end{aligned} \tag{8}$$

Furthermore, with the optimal step size $\tilde{\eta}_{\text{opt}} = \big(\frac{\beta^4 C_3}{2C_2}\big)^{1/3}$, the asymptotic error is:

$$\begin{aligned} & \limsup_{R\to\infty} \mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \\ & \leq \min\Bigg\{\mathcal{O}\Big(\beta^{2/3}(n\Delta^2)^{2/3}\Big(\frac{(B\sigma_{F,1}^2 + \delta^2\lambda^2)\kappa_F}{B}\Big)^{1/3} + \frac{\delta^2\lambda^2}{\mu_F}\Big), \\ & \qquad \mathcal{O}\Big(\frac{\beta^2 n\Delta^2}{\hat{\eta}_1} + \frac{(B\sigma_{F,1}^2 + \delta^2\lambda^2)\kappa_F}{\beta^2}\hat{\eta}_1^2 + \frac{\delta^2\lambda^2}{\mu_F}\Big)\Bigg\}. \end{aligned} \tag{9}$$

(b) **(Personalized model convergence)** *The average personalized model error satisfies:*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\|\tilde{\theta}_i^R(x_i^R) - x^*\|^2\big] \leq \frac{1}{\mu_F}\mathcal{O}\big(\mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)]\big) + \mathcal{O}\Big(\frac{\sigma_{F,1}^2}{\lambda^2} + \delta^2\Big). \tag{10}$$

Here $\kappa_F := \frac{L_F}{\mu_F}$, $\bar{\mathbf{x}}_R := \sum_{r=0}^{R-1}\frac{\alpha_r}{A_R}\mathbf{x}^r$, $\alpha_r := (1 - \tilde{\eta}\mu_F/4)^{-(r+1)}$, and $A_R := \sum_{r=0}^{R-1}\alpha_r$.

Theorem 2 (Convergence with nonconvex and smooth). *Let Assumptions 1, 2, 3 (b), 4, and 5 hold. If $\eta \leq \frac{\hat{\eta}_2}{\beta B}$, where $\hat{\eta}_2 := \frac{1}{24(L_F + 16\lambda^2)}$ with $\lambda \geq \sqrt{8L^2 + 1}$ and $\beta \geq 1$, then we have*

(a) **(Global model convergence)** *The global model satisfies:*

$$\mathbb{E}[\|\nabla F(\mathbf{x}^{r^*})\|^2] \leq \mathcal{O}\Bigg(\frac{\Delta_F}{\hat{\eta}_2 R} + \beta\Big(\frac{\lambda^2\delta^2}{B} + \lambda^2\delta^2 + \sigma_{F,2}^2\Big) + \frac{\beta^2}{\hat{\eta}_2^2}\cdot n\Delta^2 + \frac{L_F\beta^2}{\hat{\eta}_2}\cdot n\Delta^2\Bigg). \tag{11}$$

(b) **(Personalized model convergence)** *The average personalized model error satisfies:*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|\tilde{\theta}_i(x_i^{r^*}) - x_i^{r^*}\|^2] \leq \mathcal{O}\Bigg(\mathbb{E}[\|\nabla F(\mathbf{x}^{r^*})\|^2]\Bigg) + \mathcal{O}\Bigg(\delta^2 + \frac{\sigma_{F,2}^2}{\lambda^2}\Bigg), \tag{12}$$

where $\Delta_F := F(\mathbf{x}^0) - F(\mathbf{x}^*)$, and $r^* \in \{0, \ldots, R-1\}$ is sampled uniformly.

Theorem 3 (Upper bound on consensus time steps of Algorithm 2(DQuFTA))**.** *Under Assumption 1, 2, 3, 4, and 5, in the global iteration r, after T time steps of consensus communication, all nodes reach exact agreement on the quantized average:*

$$q_i^r = \bar{q}^r = \left\lfloor \frac{1}{N} \sum_{i=1}^{N} \left\lfloor \frac{x_{i,B}^r}{\Delta} \right\rfloor \right\rfloor \Delta, \forall v_i \in \mathcal{V}$$

The number of consensus time steps T required per global iteration r to achieve exact agreement for all nodes is upper-bounded by a problem-dependent quantity. Specifically:

(a) **(Strongly convex case)** *when Assumptions 1, 2, 3 (a), 4, and 5 hold, we have:*

$$T \leq \left(\frac{8N\eta B}{\Delta} \sqrt{6L_F(F(\mathbf{x}^r) - F(\mathbf{x}^*)) + \frac{\lambda^2\delta^2}{B} + 3\sigma_{F,1}^2} + 3N \right) \tau D.$$

(b) **(Nonconvex smooth case)** *when Assumptions 1, 2, 3 (b), 4, and 5 hold, we have:*

$$T \leq \left(\frac{8N\eta B}{\Delta} \sqrt{\frac{3\lambda^2}{2(\lambda^2 - 8L^2)} \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + \frac{\lambda^2\delta^2}{2B} + \sigma_{F,2}^2} + 3N \right) \tau D,$$

where τ depends on the parameters of the digraph and the probability of convergence from (Rikos et al., 2022, Theorem 1), and D is the diameter of the digraph.

Novelty of our Analysis. While our algorithm builds upon T Dinh et al. (2020), extending this bi-level optimization to a fully decentralized and quantized communication setting introduces unique challenges not present in standard centralized PFL. First, quantized finite-time consensus introduces an additive quantization error that scales (at most) quadratically with Δ (see Theorems 1 and 2). We handle this by establishing exact agreement on the quantized average (see Lemma 3 and the perfect-consensus property in equation 30). This approach eliminates the usual consensus-dispersion terms found in decentralized proofs, leaving only the strictly bounded quantization mismatch. Second, we provide the first problem-dependent finite-time bound on the number of consensus steps needed per global round (Theorem 3), relating it directly to the current optimization error and network parameters. This represents a significant departure from prior decentralized PFL works that overwhelmingly assume infinite-precision or uncompressed gossip tracking. Together, these analytical techniques prove that Q-pFedMe mathematically matches the baseline convergence rates of unquantized pFedMe, up to a controlled additive quantization error.

Comparison of Convergence Guarantees. The theoretical guarantees are also different from those of the decentralized baselines. For smooth non-convex objectives, D-PSGD Lian et al. (2017) establishes $\mathcal{O}(1/K + 1/\sqrt{NK}))$ convergence in the averaged squared gradient norm and notes an $\mathcal{O}(1/(NK))$ extension under strong convexity. DFedPGP Liu et al. (2024) establishes an $\mathcal{O}(1/\sqrt{T})$ rate for a stationarity measure involving its shared and personalized parameter blocks, but does not provide a separate strongly convex result. DisPFL Dai et al. (2022) provides a generalization bound rather than an optimization convergence rate. In comparison, the convergence rates in Theorems 1 and 2 match the optimal centralized, unquantized PFL bounds T Dinh et al. (2020), up to an additive quantization error that is polynomial in Δ. Specifically, the linear convergence to a neighborhood (strongly convex) and the $\mathcal{O}(1/R)$ rate (non-convex) are optimal for first-order stochastic methods under standard assumptions Arjevani et al. (2020). Moreover, the dominant quantization error scales as $\mathcal{O}(\Delta^{4/3})$ for strongly convex case (with optimal step size) and as $\mathcal{O}(\Delta^2)$ for non-convex case. Both are expected, as any finitely quantized consensus protocol must incur an error floor proportional to Δ Aysal et al. (2008); Frasca et al. (2008).

Remark 1. *The bound in Theorem 3 depends on the current state of optimization: a larger optimality gap $F(\mathbf{x}^r) - F(\mathbf{x}^*)$ (in the convex case) or a larger expected gradient norm $\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2]$ (in the non-convex case) may require more time steps. The required number of time steps T increases with the number of nodes N and decreases with a larger quantization step Δ. This bound guarantees exact quantized agreement in finite time, enabling the decentralized operation of Q-pFedMe.*

Trade-off between model convergence and communication cost. In Theorem 1, the global error contains the quantization-dependent terms $\mathcal{O}(\beta^{2/3}(n\Delta^2)^{2/3})$ and $\mathcal{O}(\beta^2 n\Delta^2)$, while in Theorem 2

Table 1: Comparison of test accuracy across Q-pFedMe and centralized baselines on all tasks. Q-pFedMe(PM) achieves competitive or superior performance compared to pFedMe(PM), with an average gap of 0.43%. Both methods significantly outperform FedAvg and Per-FedAvg.

Algorithm	Synthetic MLR (%)	Synthetic DNN (%)	MNIST (%)	CIFAR-10 Shards (%)	CIFAR-10 $\alpha = 0.1$ (%)	CIFAR-10 $\alpha = 0.3$ (%)	CIFAR-10 $\alpha = 0.5$ (%)
FedAvg	76.60	80.18	95.35	55.83	54.44	57.50	58.83
Per-FedAvg	75.45	74.92	92.41	48.48	67.59	67.59	67.59
pFedMe(GM)	76.08	78.40	94.94	50.72	50.66	52.95	52.90
pFedMe(PM)	81.01	**82.42**	**95.59**	**78.03**	**82.99**	**70.83**	**65.02**
Q-pFedMe(GM)	79.00	81.92	94.20	50.00	48.27	51.44	52.16
Q-pFedMe(PM)	**81.22**	82.40	94.93	78.00	82.19	70.08	64.06
Gap (%)	+ 0.21	- 0.02	- 0.66	- 0.03	- 0.80	- 0.75	- 0.96

it includes $\frac{\beta^2}{\hat{\eta}_2^2} \cdot n\Delta^2 + \frac{L_F \beta^2}{\hat{\eta}_2} \cdot n\Delta^2$. Although the personalized-model bounds equation 10 and equation 12 are expressed in terms of the global error, the dependence on Δ enters through those global-error terms, so the overall convergence error still grows with Δ. At the same time, the total communication cost scales as $\mathcal{O}(\log_2(1/\Delta)/\Delta)$ (see Appendix B.3). This means a larger Δ reduces the cost. Therefore, Δ controls the balance between model accuracy and communication cost.

6 Experimental Results

In this section, we present comprehensive empirical evaluations of Algorithm 1 (Q-pFedMe) against the centralized baselines pFedMe T Dinh et al. (2020), FedAvg McMahan et al. (2017), and Per-FedAvg Fallah et al. (2020), across three datasets (Synthetic Li et al. (2020), MNIST LeCun et al. (2002), CIFAR-10 Krizhevsky et al. (2009)), heterogeneity settings (shards and Dirichlet Distribution Tan et al. (2022)), and hyperparameter configurations (β, Δ). Our experiments against centralized baselines are conducted on four tasks: Synthetic on MLR, Synthetic on DNN, MNIST on DNN, and CIFAR-10 on CNN. The first task is convex, and the remaining three are non-convex. We report both the personalized model (PM) and the global model (GM) of Q-pFedMe and pFedMe for comparisons. FedAvg only has a global model, and Per-FedAvg only has a personalized model; thus, they are not distinguished by PM/GM. The evaluation metrics is defined in Appendix C.2.

Additionally, we further compare Q-pFedMe with decentralized baselines DPSGD Lian et al. (2017), DisPFL Dai et al. (2022), DFedPGP Liu et al. (2024) on CIFAR-100 with 100 nodes, a highly unbalanced dirichlet partition with $\alpha = 0.1$, and a GroupNorm ResNet-18 over both dense directed exponential topology and sparse directed ring topology. We only report the personalized model of Q-pFedMe for comparisons with other decentralized algorithms. Computational resources, dataset, models, dataset partitions, and communication graph are provided in Appendix C.1.

6.1 Prediction Accuracy Comparison

Comparison with centralized baselines. Table 1 illustrates the test accuracy comparison between Q-pFedMe and centralized baselines (pFedMe, FedAvg, and Per-FedAvg) on four tasks including shards-based partitioning (Synthetic, MNIST, CIFAR-10 Shards) and Dirichlet partitioning (CIFAR-10 with $\alpha = 0.1, 0.3, 0.5$). Q-pFedMe achieves accuracy statistically indistinguishable from centralized pFedMe across all tasks. The average test accuracy gap across all tasks is 0.43%, demonstrating that decentralization and appropriate quantization do not harm final model quality. This is a critical result: Q-pFedMe achieves performance comparable to unquantized (real-valued), central server-based pFedMe while operating in a decentralized fashion and addressing the challenges of quantized consensus. The test accuracy and training loss curves are in Appendix C.3 and Appendix C.4.

Comparison with decentralized baselines. Table 2 supports two main observations. First, Q-pFedMe achieves a BestAcc of 63.31% under both topologies, outperforming the strongest external baseline by 7.11% on the directed exponential graph and 10.86% on the directed ring. Its B10 accuracy reaches 43.90%, exceeding the strongest corresponding baseline by 7.49% and 7.77%, respectively. Second, Q-pFedMe reaches 50% mean personalized accuracy in 90 global rounds, compared with at least 360 rounds on the exponential graph and 450 rounds on the ring for the

Table 2: Comparison of Q-pFedMe with decentralized baselines on CIFAR-100 using 100 nodes, a Dirichlet partition with $\alpha = 0.1$, and a GroupNorm ResNet-18. BestAcc and B10 are reported as mean $\pm$ standard deviation over three seeds. BestAcc denotes the best mean personalized accuracy, and B10 denotes the mean accuracy of the worst-performing 10% of nodes. R@50, T@50, and C@50 denote the number of global rounds, wall-clock time, and cumulative network-wide communication required to first reach 50% mean personalized accuracy, respectively.

Topology	Method	BestAcc(%) ↑	B10(%) ↑	R@50↓	T@50(s) ↓	C@50(TB) ↓
Dir-Exp	D-PSGD	56.20 ± 0.44	34.41 ± 0.49	360	55291.09	**10.29**
Dir-Exp	DisPFL	52.82 ± 0.21	36.41 ± 0.50	450	93262.03	10.54
Dir-Exp	DFedPGP	51.50 ± 0.41	31.74 ± 0.57	450	61862.22	12.79
Dir-Exp	**Q-pFedMe**	$\mathbf{63.31 \pm 0.48}$	$\mathbf{43.90 \pm 0.65}$	**90**	**27262.59**	39.86
Dir-Ring	D-PSGD	50.13 ± 0.58	36.13 ± 0.63	480	61478.35	1.96
Dir-Ring	DisPFL	52.45 ± 0.13	36.09 ± 0.38	450	82628.09	1.51
Dir-Ring	DFedPGP	51.30 ± 0.15	34.03 ± 0.70	460	63639.32	1.86
Dir-Ring	**Q-pFedMe**	$\mathbf{63.31 \pm 0.48}$	$\mathbf{43.90 \pm 0.65}$	**90**	**27262.59**	**1.24**

external baselines. Although we report C@50 only for transparency, Q-pFedMe achieves a 50% accuracy on the directed ring topology with the least communication cost.

6.2 Analysis of Δ, β, and Communication Cost

Effects of Δ. We now investigate how the quantization level Δ affects Q-pFedMe's performance. As detailed in Appendix C.5, it reveals an important finding: Q-pFedMe maintains accuracy comparable to pFedMe only when Δ is below a task-specific threshold Δ_{crit}. When Δ exceeds Δ_{crit} by an order of magnitude, the accuracy degrades smoothly and predictably, which directly reflects the quadratic growth of the quantization error terms established in Theorem 1 and Theorem 2.

Effects of β. We additionally study the sensitivity to the global update parameter β, which controls the aggressiveness of the global reference-model update. As shown in Appendix C.6, moderate over-relaxation can accelerate convergence, whereas excessively large values may cause instability.

Communication Cost Analysis. We evaluate Q-pFedMe's per-transmission payload against centralized baselines. As shown in Table 4, Q-pFedMe reduces per-transmission communication by $43.75\% - 53.13\%$ compared to centralized baselines. Table 5 shows Q-pFedMe achieves the largest reduction among communication-efficient decentralized baselines. Table 6 reveals a trade-off: larger Δ can reduce the per-transmission payload but leads to lower accuracy. Detailed derivations and parameter configurations are provided in Appendix C.7, with numerical values summarized in Table 3.

7 Conclusion

Conclusion. In this paper, we introduced Q-pFedMe, a decentralized personalized FL framework that simultaneously achieves personalization, full decentralization, and exact finite-time agreement among nodes over strictly quantized communication. We proved linear convergence to a neighborhood of the optimum for strongly convex objectives and sublinear convergence for smooth non-convex objectives. Moreover, we derived a finite-time upper bound on the number of consensus steps required per global round. Extensive experiments on Synthetic, MNIST, and CIFAR-10 show that Q-pFedMe matches the accuracy of unquantized centralized baselines while communicating only quantized messages, and experiments on CIFAR-100 show that Q-pFedMe achieves the highest accuracy and payload reduction compared with decentralized baselines on two topologies.

Limitations. Our analysis assumes a static strongly connected digraph; extending Q-pFedMe to time-varying or asynchronous networks requires further analysis. Although the bounds clarify how λ, β, B, and Δ affect the personalization–generalization and accuracy–communication trade-offs, which could guide their practical selection to minimize the obtained convergence bounds, automatic tuning under unknown heterogeneity remains open. Moreover, Q-pFedMe reduces per-transmission payload rather than cumulative network traffic: exact termination of DQuFTA may require multiple exchanges and stopping checks. More efficient termination detection, event-triggered communication, and controlled inexact consensus are promising ways to reduce this overhead.

AI USE STATEMENT

We used ChatGPT-5.5 solely to improve the grammar, wording, and readability of author-written text. The tool was not used to generate research ideas, hypotheses, methods, experimental designs, mathematical claims or proofs, data, result interpretations, or citations. All AI-assisted edits were reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve the collection of new personal data or direct interaction with human participants. The datasets and pretrained models used in our experiments are publicly available and are used in accordance with their stated licenses and terms of use. We report only aggregate experimental results and do not attempt to identify individuals in the datasets.

Like other machine learning methods trained on real-world data, the proposed method may inherit biases or limitations present in its training data. Therefore, its performance should be carefully evaluated before deployment in safety-critical or high-stakes applications. The authors have considered these limitations and are not aware of any additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results throughout the paper and code URL `https://anonymous.4open.science/r/N8HgpH8UoWHorEt/`. All experimental details, including setup and implementation specifics, are provided in Section 6 and Appendix B, enabling reproduction of main results.

REFERENCES

Yossi Arjevani, Ohad Shamir, and Nathan Srebro. A tight convergence analysis for stochastic gradient descent with delayed updates. In *Algorithmic Learning Theory*, pp. 111–132. PMLR, 2020.

Tuncer C Aysal, Mark J Coates, and Michael G Rabbat. Distributed average consensus with dithered quantization. *IEEE Transactions on Signal Processing*, 56(10):4905–4918, 2008.

Nicola Bastianello, Apostolos I Rikos, and Karl H Johansson. Online distributed learning with quantized finite-time coordination. In *IEEE Conference on Decision and Control*, pp. 5026–5032. IEEE, 2023.

Shan Chang, Xianbo Wang, Hao Yu, Denghui Li, Guanghao Liang, Hongzi Zhu, and Bo Li. PRISAM: Efficient personalization via BN masks in heterogeneous decentralized federated learning. In *International Conference on Distributed Computing Systems (ICDCS)*, pp. 802–812. IEEE, 2025.

Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. Dispfl: Towards communication-efficient personalized federated learning via decentralized sparse training. *arXiv preprint arXiv:2206.00187*, 2022.

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

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Mengchen Fan, Keren Li, Tianyun Zhang, Qing Tian, and Baocheng Geng. Pfeddst: Personalized federated learning with decentralized selection training. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Paolo Frasca, Ruggero Carli, Fabio Fagnani, and Sandro Zampieri. Average consensus by gossip algorithms with quantized communication. In *IEEE Conference on Decision and Control*, pp. 4831–4836, 2008.

Eunjeong Jeong and Marios Kountouris. Personalized decentralized federated learning with knowledge distillation. In *IEEE International Conference on Communications*, pp. 1982–1987. IEEE, 2023.

Peter Kairouz and H Brendan McMahan. 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.

Salma Kharrat, Marco Canini, and Samuel Horváth. DPFL: Decentralized personalized federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5086–5094. PMLR, 2025.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pp. 5381–5393. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

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

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

Weishi Li, Yong Peng, Mengyao Du, Fuhui Sun, Xiaoyan Wang, and Li Shen. Hypernetwork aggregation for decentralized personalized federated learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1440–1448, 2025.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

I Lin, Osman Yagan, Carlee Joe-Wong, et al. Fedspd: A soft-clustering approach for personalized decentralized federated learning. *arXiv preprint arXiv:2410.18862*, 2024.

Yingqi Liu, Yifan Shi, Qinglun Li, Baoyuan Wu, Xueqian Wang, and Li Shen. Decentralized directed collaboration for personalized federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23168–23178, 2024.

Qianyu Long, Qiyuan Wang, Christos Anagnostopoulos, and Daning Bi. Decentralized personalized federated learning based on a conditional "sparse-to-sparser" scheme. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Zhenguo Ma, Yang Xu, Hongli Xu, Jianchun Liu, and Yinxing Xue. Like attracts like: Personalized federated learning in decentralized edge computing. *IEEE Transactions on Mobile Computing*, 23(2):1080–1096, 2022.

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.

Gabriele Oliva, Roberto Setola, and Christoforos N Hadjicostis. Distributed finite-time calculation of node eccentricities, graph radius and graph diameter. *Systems & Control Letters*, 92:20–27, 2016.

Junyoung Park, Sunmin Kim, Joohyung Lee, and Dusit Niyato. Communication-efficient decentralized federated learning for generalization and personalization over wireless networks. *IEEE Wireless Communications Letters*, 2025.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International conference on artificial intelligence and statistics*, pp. 2021–2031. PMLR, 2020.

Apostolos I Rikos, Christoforos N Hadjicostis, and Karl H Johansson. Non-oscillating quantized average consensus over dynamic directed topologies. *Automatica*, 146:110621, 2022.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International conference on machine learning*, pp. 9489–9502. PMLR, 2021.

Yifan Shi, Yingqi Liu, Yan Sun, Zihao Lin, Li Shen, Xueqian Wang, and Dacheng Tao. Towards more suitable personalization in federated learning via decentralized partial model training. *arXiv preprint arXiv:2305.15157*, 2023.

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

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

Guojun Xiong, Gang Yan, Shiqiang Wang, and Jian Li. Deprl: Achieving linear convergence speedup in personalized decentralized learning with shared representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16103–16111, 2024.

Anqi Zhang, Ping Zhao, Wenke Lu, and Guanglin Zhang. Decentralized federated learning towards communication efficiency, robustness, and personalization. *ACM Transactions on Sensor Networks*, 21(3):1–20, 2025.

A Appendix

B Details of Algorithm 2 (DQuFTA)

B.1 Algorithm 2 (DQuFTA)

Algorithm 2 DQuFTA: Decentralized Quantized Finite Time Averaging

Input: The states to be averaged $\{x_{i,B}^r\}_{i\in\mathcal{V}}$, quantization level Δ, and upper bound to diameter D.

//Initialization
Each node $v_i \in \mathcal{V}$ sets $y_i^{[0]} = (q_\Delta(x_{i,B}^r))/\Delta$, and $z_i^{[0]} = 1$;
for $\ell = 1, 2, \ldots$ **do**
  // local update and transmission
  while $z_i^{[\ell]} > 0$ **do**
    compute $t = \lfloor y_i^{[\ell]}/z_i^{[\ell]} \rfloor$, and update $y_i^{[\ell]} \leftarrow y_i^{[\ell]} - t$ and $z_i^{[\ell]} \leftarrow z_i^{[\ell]} - 1$;
    select an out-neighbor $v_j \in \mathcal{N}_i^+ \cup \{v_i\}$ randomly with uniform probability;
    transmit t to out-neighbor v_j, and v_j updates $y_j^{[\ell]} \leftarrow y_j^{[\ell]} + t$ and $z_j^{[\ell]} \leftarrow z_j^{[\ell]} + 1$;
  end while
  // max, min voting for termination
  if $\ell \bmod D = 1$ **then**
    if $z_i^{[\ell]} > 0$ **then**
      compute $M_i = \lceil y_i^{[\ell]}/z_i^{[\ell]} \rceil$ and $m_i = \lfloor y_i^{[\ell]}/z_i^{[\ell]} \rfloor$;
    else
      set $M_i =$ NULL and $m_i =$ NULL;
    end if
  end if
  broadcast M_i, m_i to every out-neighbor $v_j \in \mathcal{N}_i^+$;
  receive M_j, m_j from every in-neighbor $v_j \in \mathcal{N}_i^-$;
  set $M_i \leftarrow \max_{v_j \in \mathcal{N}_j^- \cup \{v_i\}} M_j$, and $m_i \leftarrow \min_{v_j \in \mathcal{N}_j^- \cup \{v_i\}} m_j$;
  // condition for termination
  if $\ell \bmod D = 0$ **then**
    if $M_i - m_i \leq 1$ **then**
      set $q_i^r \leftarrow \Delta m_i$ and terminate operation;
    end if
  end if
end for

Output: Each node $v_i \in \mathcal{V}$ calculates $q_i^r = \lfloor \frac{1}{N} \sum_{i=1}^N y_i^{[0]} \rfloor \Delta$.

Intuition of Algorithm 2 (DQuFTA). Algorithm 2 (DQuFTA) combines three components: (i) quantization, (ii) quantized averaging, and (iii) a decentralized stopping rule. The quantized averaging follows (Rikos et al., 2022, Algorithm 1). Each node v_i first applies an asymmetric uniform quantizer to its state $x_{i,B}^r$ with a common step size Δ, obtaining $y_i^{[0]} = \lfloor x_{i,B}^r/\Delta \rfloor$ and setting $z_i^{[0]} = 1$. During the averaging process, at every time step ℓ, each node splits its current integer mass $y_i^{[\ell]}$ into $z_i^{[\ell]}$ equal (or nearly equal) integer tokens, which then move randomly to a neighbor or stay at the same node. As tokens meet at the same node, their masses are merged, and the y and z values are updated accordingly. This mass-splitting mechanism guarantees that after a finite number of time steps every node holds the exact quantized average of the initial integer values, i.e., $y_i^{[\ell]} = \left\lfloor \frac{1}{N} \sum_{j=1}^N y_j^{[0]} \right\rfloor$ for all $V_i \in \mathcal{V}$. To detect termination in finite time, every D steps (where D is the graph diameter) nodes perform a max–min consensus on the ratio $y_i^{[\ell]}/z_i^{[\ell]}$. Once the difference between the global maximum and minimum of this ratio is at most 1, all nodes have reached the same integer average, and they rescale by Δ to obtain the quantized average value $q_i^r = \Delta\, m_i$ and terminate. As a result, every node calculates the quantized average $q_i^r = \bar{q}^r = \left\lfloor \frac{1}{N} \sum_{i=1}^N \left\lfloor \frac{x_{i,B}^r}{\Delta} \right\rfloor \right\rfloor \Delta, \forall v_i \in \mathcal{V}$. Because the mass-splitting process drives this difference

to at most 1 in a number of steps proportional to D, and the condition is checked every D steps, Algorithm 2 terminates after $\mathcal{O}(D)$ steps. Importantly, the quantizer is used only at initialization; during the iterations nodes store and communicate only the already-quantized integer values, so no further quantization is needed, which enhances communication efficiency.

B.2 Communication Cost per Global Round

In each consensus step l, each node v_i needs to transmit $y_i^{[l]}$ and $z_i^{[l]}$ to one randomly selected out-neighbor $z_i^{[l]}$ times. we note $\sum_{i=1}^{N} z_i^l = \sum_{i=1}^{N} z_i^{[0]} = N$, the number of $y_i^{[l]}$'s parameters is n, the number of $z_i^{[l]}$'s parameter is 1. Thus, during the quantized averaging process, Algorithm 2 costs $T \times N \times (n+1) \times Q$ bits. Q is the fixed-point bits required for quantized parameter transmission. In addition, each node v_i broadcasts the maximum M_i and minimum m_i to every out-neighbor every D steps. We note that the number of parameters of M_i and m_i are both n, and the number of out-neighbors is D_i^+. Thus, during the max-min consensus process, Algorithm 2 costs $\frac{T}{D} \times \sum_{i=0}^{N} D_i^+ \times 2n \times Q$ bits. In per global round, Algorithm 2(DQuFTA) costs $NQ(n+1)T + \sum_{i=0}^{N} D_i^+ \frac{T}{D} 2nQ = TQ(N(1+n) + \frac{2n}{D}\sum_{i=0}^{N} D_i^+)$ bits, where N is the number of nodes, n is the number of model parameters, T is the required consensus steps to reach agreement, and Q is jointly determined by the range $[-R, R]$ of model parameters' values and the quantized level Δ as $Q = \lceil \log_2 M \rceil$, where $M = \big(2 \times \lfloor \frac{R}{\Delta} \rfloor + 1\big)$.

B.3 Total Communication Cost

From Appdenix B.2, we have the communication cost per global round is $TQ(N(1+n) + \frac{2n}{D}\sum_{i=0}^{N} D_i^+)$ bits, so the total communication cost is $R\bar{T}Q(N(1+n) + \frac{2n}{D}\sum_{i=0}^{N} D_i^+)$ bits, where $\bar{T}$ is the average consensus steps, and equals $\mathcal{O}(\bar{T}Q)$. From Theorem 3, we get $T = \mathcal{O}(\frac{1}{\Delta})$, and $\bar{T}$ also should be $\mathcal{O}(\frac{1}{\Delta})$. And from the computation of Q in Appendix B.2, we get $Q = \mathcal{O}(\log_2 \frac{1}{\Delta})$. Thus, we have the total communication cost equals $\mathcal{O}(\log_2(1/\Delta)/\Delta)$

C Experiments

C.1 Experimental Setups

Computational Resources. The experiments were conducted on a computing facility equipped with 48GB NVIDIA® A40™ GPUs. All implementations are based on PyTorch.

Datasets and Models Our experiments are conducted on three widespread dataset, including Synthetic Li et al. (2020), MNIST LeCun et al. (2002), which is licensed under CC BY-SA 3.0, CIFAR-10 Krizhevsky et al. (2009), licensed under the MIT License, and CIFAR-100 Krizhevsky et al. (2009), which is publicly available. For μ-strongly convex setting, we consider a L_2-regularized multinomial logistic regression model (MLR) with the softmax activation and cross-entropy loss functions for Synthetic. For nonconvex case, we adopt a two-layers deep neural network (DNN) with hidden layer of size 100 for MNIST and 20 for Synthetic using ReLU activation and a softmax layer at the end, and adopt a two-layers CNN using ReLU activation and a softmax layer at the end for CIFAR-10. We compare the algorithms using the same and fine-tuned learning rates, batch sizes, and number of local and global iterations. We evaluate on strongly convex settings (Synthetic on MLR) and non-convex settings (Synthetic on DNN, MNIST on DNN, and CIFAR-10 on CNN).

For experiments on CIFAR-100, we adopt a GroupNorm ResNet-18. Momentum is disabled for all methods, every node performs the same number of local minibatch updates per global round, and the batch size is fixed to 128. All remaining method-specific hyperparameters follow the corresponding papers or official implementations.

Data Partition. For Synthetic, we follow the partition in T Dinh et al. (2020). Two parameters $\bar{\alpha} = 0.5$ and $\bar{\beta} = 0.5$ control the heterogeneity of local models and local datasets, respectively. The dataset contains 10 classes with 60-dimensional real-valued features and is distributed to $N = 100$ nodes following a power law distribution, with each node receiving between 250 and 25,810 samples.

For MNIST and CIFAR-10, we adopt shards-based partitioning T Dinh et al. (2020), where each client receives only a subset of classes. Specifically, both datasets are distributed to $N = 20$ nodes. For MNIST, each node receives a different local data size ranging from 1,165 to 3,834 samples and is restricted to only 2 of the 10 labels. For CIFAR-10, each node receives between 1,314 and 5,642 samples and is restricted to 3 of the 10 labels.

To evaluate Q-pFedMe and centralized baselines under more realistic heterogeneity, we conduct additional experiments on CIFAR-10 using Dirichlet partitioning Tan et al. (2022). The original training set is divided into 20 subsets, where the class distribution of each subset follows a Dirichlet distribution $\mathrm{Dir}(\alpha)$ with $\alpha = [0.1, 0.3, 0.5]$, corresponding to high, moderate, and mild heterogeneity, respectively.

To evaluate Q-pFedMe and decentralized baselines under highly unbalanced heterogeneity, we conduct experiments on CIFAR-100 using Dirichlet partitioning. The original training set is divided into 100 subsets, where the class distribution of each subset follows a Dirichlet distribution $\mathrm{Dir}(\alpha)$ with $\alpha = 0.1$.

Communication Graph. For the experiments against centralized baselines, we model the communication topology as a random directed graph according to the Erdős–Rényi model, where each directed edge exists independently with probability p. To ensure the graph is strongly connected with high probability, we set $p = 0.3$. This choice satisfies $p \gg \ln N/N$ for the considered network sizes (i.e., $N = [20, 100]$), guaranteeing that the resulting directed graph is almost surely strongly connected. Hence, the generated topology provides a reliable communication backbone where any node can reach any other via a directed path.

For the experiments against decentralized baselines, we adopt a directed exponential topology and a sparse directed ring. The directed exponential graph contains 700 edges, has an in-degree and out-degree of 7, and has directed diameter 5; the directed ring contains 100 edges, has unit in-degree and out-degree, and has directed diameter 99. All methods use the same fixed topology throughout training. For D-PSGD, uniform self/in-neighbor weights yield doubly stochastic but nonsymmetric mixing matrices. DisPFL applies its mask-aware aggregation to the same in-neighbor messages, whereas DFedPGP natively supports directed communication.

C.2 Evaluation Metrics

Test Accuracy of Global Model (GM). The accuracy of global model $\{x_i\}_{i\in\mathcal{V}}$ is defined as:

$$\mathrm{Acc}(\{x_i\}_{i\in\mathcal{V}}) \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\mathrm{Acc}(x_i;\xi_i)$$

where ξ_i is randomly sampled from $\mathcal{D}_i$.

Test Accuracy of Personalized Model (PM). The accuracy of personalized model $\{\tilde{\theta}_i\}_{i\in\mathcal{V}}$ is defined as:

$$\mathrm{Acc}(\{\tilde{\theta}_i\}_{i\in\mathcal{V}}) \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\mathrm{Acc}(\tilde{\theta}_i;\xi_i)$$

where ξ_i is randomly sampled from $\mathcal{D}_i$.

The goal of personalized federated learning is to produce local-specific models. The global model (GM) is only an intermediate anchor for collaboration, while the personalized model (PM) is the one deployed in practice. Hence, we use PM as the primary metric.

C.3 Comparison of test accuracy and training loss curves

Fig. 1 illustrates the test accuracy comparison between Q-pFedMe and centralized baselines. Q-pFedMe achieves accuracy statistically indistinguishable from centralized pFedMe across all tasks. In Fig. 2, the training loss of Q-pFedMe and pFedMe are nearly identical across all four tasks, with both personalized models (PM) and global models (GM) overlapping closely throughout training.

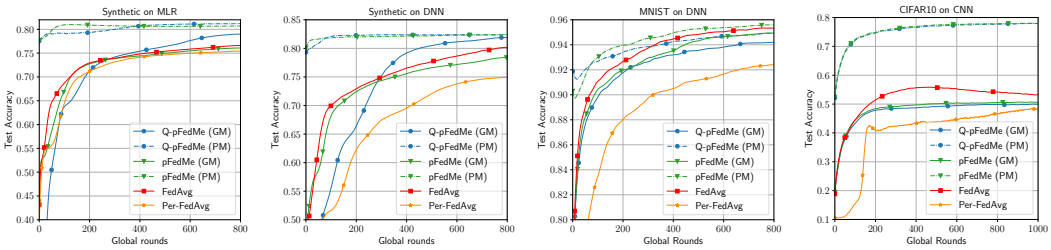


Figure 1: Test accuracy comparison of Q-pFedMe, pFedMe, FedAvg, and Per-FedAvg on four tasks. ($\beta = 1, B = 20$, and $N = 100, \Delta = 1e-4, \eta = 0.005, |D| = 20$ for Synthetic, $N = 20, \Delta = 1e-5, \eta = 0.005, |D| = 20$ for MNIST, $N = 20, \Delta = 1e-5, \eta = 0.05, |D| = 32$ for CIFAR-10)

This empirical observation directly aligns with our theoretical analysis: compared to centralized pFedMe, Q-pFedMe introduces only an additional $\mathcal{O}(\Delta^2)$ quantization error term in its convergence bounds (Theorems 1 and 2). With the task-specific Δ used in our experiments ($\Delta = 1e-4$ for Synthetic, $\Delta = 1e-5$ for MNIST and CIFAR-10), this extra term remains negligible, resulting in virtually indistinguishable optimization behavior.

In contrast, FedAvg achieves lower training loss on Synthetic and MNIST but suffers from overfitting (lower test accuracy, see Fig. 1), as it optimizes a single global model that does not adapt to individual client distributions. Per-FedAvg typically shows higher loss, consistent with its slower convergence and lower accuracy. On CIFAR-10, where heterogeneity is most pronounced, the personalized models (PM) of both Q-pFedMe and pFedMe achieve the lowest loss, confirming the benefit of Moreau-envelope personalization. These results demonstrate that Q-pFedMe faithfully preserves the optimization dynamics of pFedMe, while enabling decentralized, quantized communication.

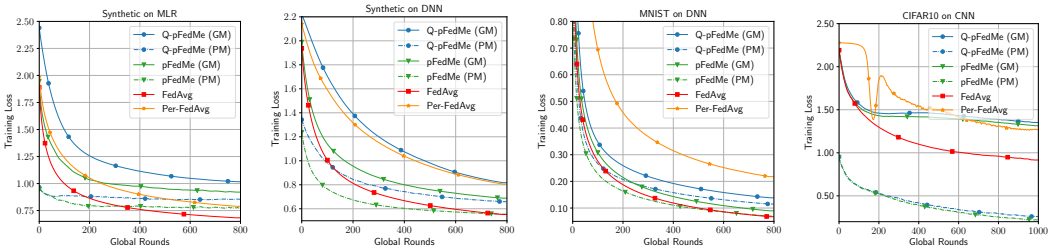


Figure 2: Training loss comparison of Q-pFedMe, pFedMe, FedAvg, and Per-FedAvg on four tasks ($\beta = 1, B = 20$, and $N = 100, \Delta = 1e-4, \eta = 0.005, |D| = 20$ for Synthetic, $N = 20, \Delta = 1e-5, \eta = 0.005, |D| = 20$ for MNIST, $N = 20, \Delta = 1e-5, \eta = 0.05, |D| = 32$ for CIFAR-10)

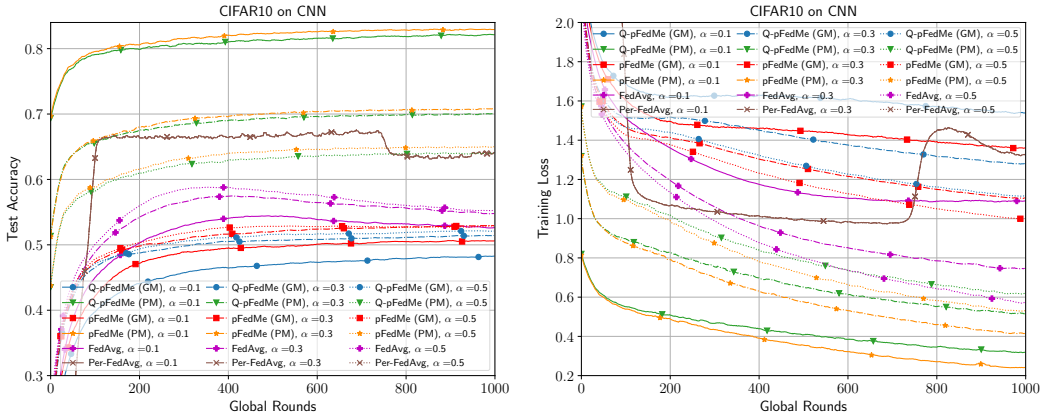
C.4 Evaluation under Dirichlet Distribution

We evaluate Q-pFedMe and centralized baslines on CIFAR-10 with Dirichlet partitioning at three heterogeneity levels: $\alpha = 0.1$ (highly heterogeneous), $\alpha = 0.3$ (moderate), and $\alpha = 0.5$ (mild). The test accuracy and training loss are shown in Fig. 3a and 3b, respectively.

Test accuracy (Fig. 3a). Across all α, Q-pFedMe achieves test accuracy nearly identical to centralized pFedMe. Both methods significantly outperform FedAvg and Per-FedAvg, especially under high heterogeneity ($\alpha = 0.1$), where the accuracy gap is largest. As αincreases (data becomes more IID), the performance of all methods converges, as expected.

Training loss (Fig. 3b). Q-pFedMe matches pFedMe's training loss across all α settings, with curves overlapping closely. This confirms that the $\mathcal{O}(\Delta^2)$ quantization error term remains negligible under different heterogeneity levels. FedAvg and Per-FedAvg show higher final losses, consistent with their lower test accuracy.

Q-pFedMe maintains its accuracy and loss parity with pFedMe across a wide range of heterogeneity levels ($\alpha = 0.1$ to $\alpha = 0.5$), demonstrating that decentralization and quantization do not compromise robustness to statistical heterogeneity. The advantage of personalized methods (pFedMe and Q-pFedMe) over FedAvg is most pronounced when heterogeneity is high ($\alpha = 0.1$).



(a) Test accuracy comparison

(b) Training loss comparison

Figure 3: Performance comparison of Q-pFedMe, pFedMe, FedAvg, and Per-FedAvg on CNN using CIFAR-10 with Dirichlet distribution $\alpha = [0.1, 0.3, 0.5]$. ($\eta = 0.005, N = 20, \Delta = 1e{-}5, B = 20$, $\beta = 1$ and $|D| = 32$)

C.5 Effects of Δ

Fig. 4 and fig. 5 illustrate the test accuracy and training loss comparison across quantized levels $\Delta = [1e-3, 1e-4, 1e-5]$ on Synthetic tasks and $\Delta = [1e-4, 1e-5, 1e-6]$ on MNIST and CIFAR-10 tasks in Q-pFedMe, respectively. On Synthetic (both MLR and DNN), $\Delta = 1e-4$ and $\Delta = 1e-5$ achieves the highest accuracy and lowest loss, while larger $\Delta = 1e-3$ leads to lower accuracy, higher loss and slower convergence. On MNIST and CIFAR-10, $\Delta = 1e-5$ and $\Delta = 1e-6$ yields the best convergence; $\Delta = 1e-4$ results in noticeably lower accuracy and higher loss, and $\Delta = 1e-6$ offers no additional improvement. Therefore, these results indicate smaller Δ preserves accuracy and loss, while larger Δ degrades performance, with a task-specific threshold Δ_{crit}. For Synthetic on MLR and DNN we have $\Delta_{\text{crit}} = 1e-4$, for MNIST on DNN and CIFAR-10 on CNN we have $\Delta_{\text{crit}} = 1e-5$.

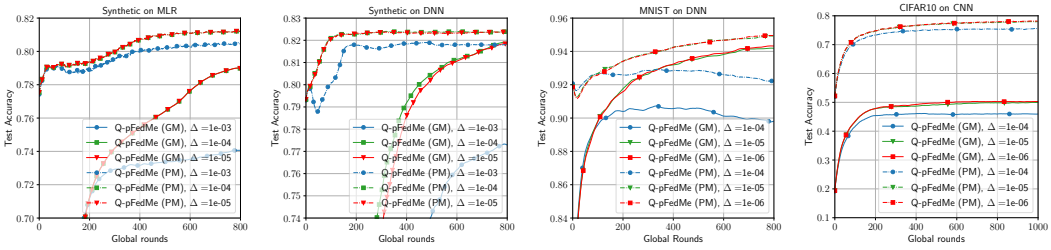


Figure 4: Effect of Δ on the accuracy of Q-pfedMe on four tasks ($\beta = 1, B = 20$, and $\eta = 0.005, |D| = 20, n = 100$ for Synthetic, $\eta = 0.005, |D| = 20, n = 20$ for MNIST, $\eta = 0.05, |D| = 32, n = 20$ for CIFAR-10)

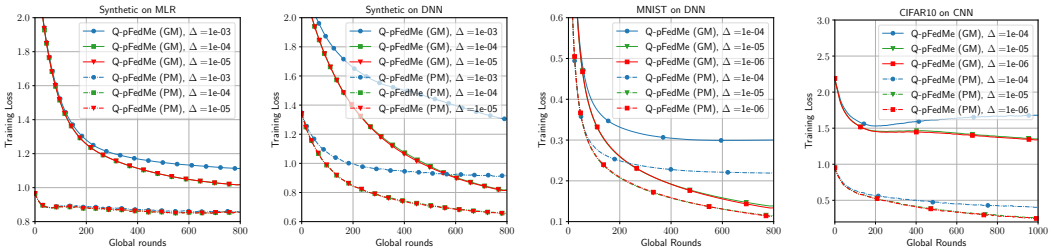


Figure 5: Effect of Δ on the loss of Q-pfedMe on four tasks. ($\beta = 1, B = 20$, and $\eta = 0.005, |D| = 20, N = 100$ for Synthetic, $\eta = 0.005, |D| = 20, N = 20$ for MNIST, $\eta = 0.05, |D| = 32, N = 20$ for CIFAR-10)

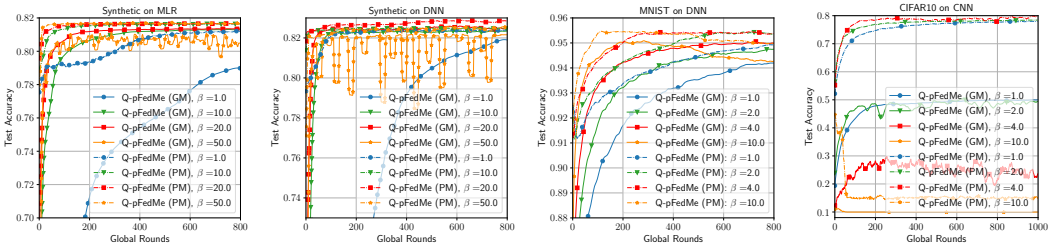


Figure 6: Effect of β on the test accuracy of Q-pfedMe on four tasks ($B = 20$, and $\eta = 0.005, |D| = 20, n = 100, \Delta = 1e-4$, for Synthetic, $\eta = 0.005, |D| = 20, n = 20, \Delta = 1e-5$ for MNIST, $\eta = 0.05, |D| = 32, n = 20, \Delta = 1e-5$ for CIFAR-10)

C.6 Effects of β

The parameter β controls the aggressiveness of the global update. A larger β accelerates convergence by either increasing the effective step size (see Theorem 1) or tightening the learning rate bound (see Theorem 2). As shown in Fig. 6 and Fig. 7, moderate β benefits all tasks. However, excessive β causes instability. On Synthetic, accuracy peaks at $\beta = 20$ and drops at $\beta = 50$ due to step size instability. On CIFAR-10 the optimal value is $\beta = 4$, whereas $\beta = 10$ causes the model to diverge due to the sharp minima of the loss landscape. Similarly, MNIST has an optimal value of $\beta = 4$. However, due to its more moderate curvature, it exhibits only mild performance degradation at $\beta = 10$ rather than outright divergence. The quantization error term $\mathcal{O}(\beta^2\Delta^2)$ is negligible given the extremely small Δ we utilized. Otherwise, all curves would degrade uniformly at the same β. Thus, β tolerance is determined by loss landscape curvature, not quantization error.

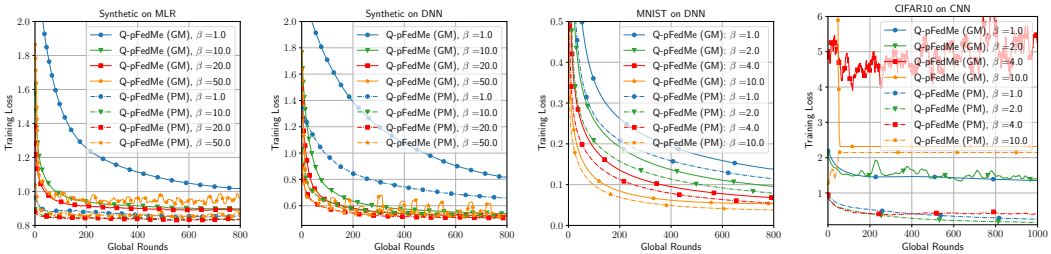


Figure 7: Effect of β on the training loss of Q-pfedMe on four tasks. ($B = 20$, and $\eta = 0.005, |D| = 20, N = 100, \Delta = 1e-4$ for Synthetic, $\eta = 0.005, |D| = 20, N = 20, \Delta = 1e-5$ for MNIST, $\eta = 0.05, |D| = 32, N = 20, \Delta = 1e-5$ for CIFAR-10)

Table 3: Numerical values of parameters for Q-pFedMe experiments. Parameters include quantization level Δ, bits per parameter Q, summation of out-degree $\sum D_i^+$, average consensus steps $\bar{T}$, and the number of model parameters n.

Tasks	Δ	Q (bits)	$\sum D_i^+$	$\bar{T}$ (steps)	n (params)
Synthetic on MLR	$1e-4$	15	2970	124	610
Synthetic on DNN	$1e-4$	15	2970	126	1430
MNIST on DNN	$1e-5$	18	114	79	79510
CIFAR-10 on CNN	$1e-5$	18	114	61	64102
CIFAR-100 on ResNet-18 (Dir-Exp)	$1e-6$	21	700	45	11220132
CIFAR-100 on ResNet-18 (Dir-Ring)	$1e-6$	21	100	10	11220132

Table 4: Per-transmission payload of Q-pFedMe and the centralized baselines. Reduction is measured relative to the corresponding full-precision baseline.

Task	Baseline (KB)	Q-pFedMe (KB)	Reduction ↑
Synthetic on MLR	2.383	1.117	53.126%
Synthetic on DNN	5.586	2.618	53.133%
MNIST on DNN	31.059	17.470	43.752%
CIFAR-10 on CNN	250.398	140.849	43.750%

C.7 Communication Cost

The computation of Q. Q is jointly determined by the dynamic range $[-R, R]$ of model parameters' values and the quantized level Δ as $Q = \lceil \log_2 M \rceil$, where $M = \left(2 \times \lfloor \frac{R}{\Delta} \rfloor + 1\right)$. After the model parameters undergo Batch Normalization or gradient clipping, their values are usually distributed within the range of $[-1.0, 1.0]$, so $R = 1$ is usually the case.

The computation of $\sum_{i=1}^N D_i^+$. Our experiment against centralized baselines is set up on a randomly generated digraph, and the probability of each directed edge existing is $p = 0.3$, so $\sum_{i=1}^N D_i^+ = pN(N-1)$. The experiment against decentralized baselines is set up on a directed exponential topology and a directed ring topology (see Appendix C.1). Therefore, for the directed exponential topology, $\sum_{i=1}^N D_i^+ = 700$, and for the directed ring topology, $\sum_{i=1}^N D_i^+ = 100$.

The computation of D. For a strongly connected random directed graph, the diameter D (maximum shortest directed path length over all ordered node pairs) can be approximated by the undirected formula $D = \lfloor \frac{\ln N}{\ln Np} \rfloor + c$, where c is a small constant (often 0 or 1). This yields $D = 2$ for our experiments.

Numerical values of parameters. Table 3 reports the numerical values of parameters for computing communication cost. Based on the quantization sensitivity analysis, we select $\Delta = \Delta_{\text{crit}}$ (from section 6.2) for each dataset as the reference configuration that preserves accuracy. For the methods with a centralized server, the communication costs $R \times N \times 32 \times n \times 2$. The fixed-point bits required for transmitting real values are 32 bits.

Payload of per transmission. Table 4 and Table 5 respectively show the payload for each peer-to-peer transmission of Q-pFedMe compared to centralized algorithms and decentralized algorithms. Q-pFedMe significantly reduces per-transmission communication cost by $43.750\% - 53.133\%$ compared with centralized baselines across all tasks and achieves the largest reduction among communication-efficient decentralized baselines.

Communication cost to reach target accuracy. Table 6 shows the best accuracy and communication cost to reach target accuracy on CIFAR-10 with CNN. For $\Delta = 1e-4$, the average consensus steps $\bar{T}$ is 45, and $Q = 15$ bits. For $\Delta = 1e-5$, the average consensus steps $\bar{T}$ is 61, and $Q = 18$ bits. For $\Delta = 1e-6$, the average consensus steps $\bar{T}$ is 68, and $Q = 21$ bits. Q-pFedMe with $\Delta = 1e-6$ obtains the highest accuracy, but costs the highest Bytes. Q-pFedMe with $\Delta = 1e-4$ costs the lowest Bytes per transmission, and compared to the highest level of accuracy, it's accuracy has not decreased significantly; it has only dropped by 2.66%.

Table 5: Per-transmission payload and reduction relative to transmitting the full dense model in FP32 of decentralized algorithms. Payload/Tx denotes the payload of one algorithmic message transmitted from one node to one out-neighbor during one peer-to-peer communication step. Under the fixed-width encoding used in our experiments, the payload is constant across transmissions.

Algorithm	Payload/Tx (MB) ↓	Reduction ↑
D-PSGD	42.80	0%
DisPFL	35.10	17.99%
DFedPGP	42.60	0.48%
Q-pFedMe	**28.08**	**34.93%**

Table 6: The table compares best accuracy, rounds to reach 50% accuracy, per-transmission payload, and communication cost to reach 50% accuracy of Q-pFedMe with different quantization levels and the centralized method on CIFAR-10 with CNN.

Methods	BestAcc (%)↑	R@50↓	Payload/Tx (KB)↓	C@50 (MB)↓
FedAvg	56.32	177	250.40	1731.24
Per-FedAvg	51.17	885	250.40	8656.19
pFedMe	78.63	**8**	250.40	**78.25**
Q-pFedMe ($\Delta = 1e-4$)	76.22	**8**	**117.37**	5529.44
Q-pFedMe ($\Delta = 1e-5$)	78.59	**8**	140.85	8994.56
Q-pFedMe ($\Delta = 1e-6$)	**78.88**	**8**	164.40	11697.84

D THEORY

D.1 REVIEW OF USEFUL EXISTING RESULTS

Proposition 2 ([52, Theorems 2.1.5 and 2.1.10]). *If a function $F_i(\cdot)$ is L_F-smooth and μ_F-strongly convex, $\forall x, x'$, we have the following inequalities:*

(a) $\|\nabla F_i(x) - \nabla F_i(x')\|^2 \leq 2L_F\big(F_i(x) - F_i(x') - \langle \nabla F_i(x'), x - x'\rangle\big)$

(b) $\mu_F\|x - x'\| \leq \|\nabla F_i(x) - \nabla F_i(x')\|$.

Proposition 3. *For any vectors $a_i \in \mathbb{R}^d$, $i = 1, \ldots, M$, by Cauchy-Schwarz inequality, we have*

$$\left\|\sum_{i=1}^{M} a_i\right\|^2 \leq M \sum_{i=1}^{M} \|a_i\|^2.$$

Lemma 2 (T Dinh et al. (2020)). *Recall the definition of F_i in equation 2c. We have the following cases.*

(a) Let Assumption 3(a) hold, then we have

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(x_i)\|^2 \leq 4L_F(F(\mathbf{x}) - F(\mathbf{x}^*)) + 2\underbrace{\frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(x^*)\|^2}_{:=\sigma_{F,1}^2}.$$

(b) Let Assumption 3(b) hold. Furthermore, if $\lambda > 2\sqrt{2}L$, then we have

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(x_i) - \nabla F(\mathbf{x})\|^2 \leq \frac{8L^2}{\lambda^2 - 8L^2}\|\nabla F(\mathbf{x})\|^2 + 2\underbrace{\frac{\lambda^2}{\lambda^2 - 8L^2}\sigma_f^2}_{:=\sigma_{F,2}^2}.$$

D.2 PROOF OF LEMMA 1

For completeness, we reproduce below the proof from T Dinh et al. (2020).

Proof. We first prove case (a). Let $h_i(\theta_i; x_{i,b}^r) := f_i(\theta_i) + \frac{\lambda}{2}\|\theta_i - x_{i,b}^r\|^2$. Then $h_i(\theta_i; x_{i,b}^r)$ is $(\lambda+\mu)$-strongly convex with its unique solution $\hat{\theta}_i(x_{i,b}^r)$. Then, by Proposition 2(b), we have

$$\|\tilde{\theta}_i(x_{i,b}^r) - \hat{\theta}_i(x_{i,b}^r)\|^2 \leq \frac{1}{(\lambda+\mu)^2}\|\nabla h_i(\tilde{\theta}_i; x_{i,b}^r)\|^2 \tag{13}$$

$$\leq \frac{2}{(\lambda+\mu)^2}\left(\|\nabla h_i(\tilde{\theta}_i; x_{i,b}^r) - \nabla \tilde{h}_i(\tilde{\theta}_i; x_{i,b}^r, \mathcal{D}_i)\|^2 + \|\nabla \tilde{h}_i(\tilde{\theta}_i; x_{i,b}^r, \mathcal{D}_i)\|^2\right) \tag{14}$$

$$\overset{equation\ 5}{\leq} \frac{2}{(\lambda+\mu)^2}\left(\|\nabla \tilde{f}_i(\tilde{\theta}_i; \mathcal{D}_i) - \nabla f_i(\tilde{\theta}_i)\|^2 + \nu\right) \tag{15}$$

$$= \frac{2}{(\lambda+\mu)^2}\left(\frac{1}{|\mathcal{D}|^2}\left\|\sum_{\xi_i \in \mathcal{D}_i} \nabla \tilde{f}_i(\tilde{\theta}_i; \xi_i) - \nabla f_i(\tilde{\theta}_i)\right\|^2 + \nu\right), \tag{16}$$

where the second inequality is given by Proposition 3. Taking expectation to both sides, we have

$$\mathbb{E}\left[\|\tilde{\theta}_i(x_{i,b}^r) - \hat{\theta}_i(x_{i,b}^r)\|^2\right] = \frac{2}{(\lambda+\mu)^2}\left(\frac{1}{|\mathcal{D}|^2}\sum_{\xi_i \in \mathcal{D}_i} \mathbb{E}_{\xi_i}\left[\|\nabla \tilde{f}_i(\tilde{\theta}_i; \xi_i) - \nabla f_i(\tilde{\theta}_i)\|^2\right] + \nu\right) \tag{17}$$

$$\leq \frac{2}{(\lambda+\mu)^2}\left(\frac{\gamma_f^2}{|\mathcal{D}|} + \nu\right), \tag{18}$$

where the equality uses independence of samples and unbiasedness $\mathbb{E}[\nabla \tilde{f}_i(\tilde{\theta}_i; \xi_i)] = \nabla f_i(\tilde{\theta}_i)$, and the inequality follows from Assumption 4. The case (b) is proved in a similar fashion, considering that $h_i(\theta_i; x_{i,b}^r)$ is $(\lambda - L)$-strongly convex. □

D.3 Proof of Lemma 2

For completeness, we reproduce below the proof from T Dinh et al. (2020).

Proof. We first prove case (a). From Propositions 2 and 3, $\mathbf{x}^*$ is the solution to problem $\min_{\mathbf{x}\in\mathbb{R}^{nN}} F(\mathbf{x})$, i.e., $\nabla F(\mathbf{x}^*) = 0$, and $x_i = x_j$ for all nodes. we have

$$\frac{1}{N}\sum_{i=1}^N \|\nabla F_i(x_i)\|^2 \leq \frac{1}{N}\sum_{i=1}^N 2\big(\|\nabla F_i(x_i) - \nabla F_i(x^*)\|^2 + \|\nabla F_i(x^*)\|^2\big)$$

$$\leq 4L_F(F(\mathbf{x}) - F(\mathbf{x}^*)) + \frac{2}{N}\sum_{i=1}^N \|\nabla F_i(x^*)\|^2.$$

We now prove case (b). Note that all nodes maintain the same reference model $x_i = \bar{x}$ for all $v_i \in \mathcal{V}$ at the start of every round. Therefore, considering Proposition 1, we have:

$$\|\nabla F_i(x_i) - \nabla F(\mathbf{x})\|^2 = \|\nabla F_i(\bar{x}) - \nabla F(\bar{x})\|^2$$

$$= \|\lambda(\bar{x} - \hat{\theta}_i(\bar{x})) - \frac{1}{N}\sum_{j=1}^N \lambda(\bar{x} - \hat{\theta}_j(\bar{x}))\|^2$$

$$= \|\nabla f_i(\hat{\theta}_i(\bar{x})) - \frac{1}{N}\sum_{j=1}^N \nabla f_j(\hat{\theta}_j(\bar{x}))\|^2,$$

where the last equality uses the first-order optimality condition $\nabla f_i(\hat{\theta}_i(\bar{x})) = \lambda(\bar{x} - \hat{\theta}_i(\bar{x}))$. Taking the average over N nodes and applying (T Dinh et al., 2020, Lemma 2(b)) directly (since all

gradients are evaluated at the single common point $\bar{x}$), we obtain:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(x_i)-\nabla F(\mathbf{x})\|^2 &= \frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(\bar{x})-\nabla F(\bar{x})\|^2\\ &\le \frac{8L^2}{\lambda^2-8L^2}\|\nabla F(\bar{x})\|^2+\frac{2\lambda^2}{\lambda^2-8L^2}\sigma_f^2\\ &= \frac{8L^2}{\lambda^2-8L^2}\|\nabla F(\mathbf{x})\|^2+\frac{2\lambda^2}{\lambda^2-8L^2}\sigma_f^2,\end{aligned}$$

where the last equality holds because $\bar{x}=x_i$ for all i, so $\nabla F(\bar{x})=\nabla F(\mathbf{x})$ exactly. This matches (T Dinh et al., 2020, Lemma 2(b)) with the consensus error term $\frac{1}{N}\sum_{i=1}^{N}\|x_i-\bar{x}\|^2=0$ vanishing exactly.

□

D.4 Proof of Theorem 1

D.4.1 Additional notations

We now analyze the convergence of Algorithm 1 (Q-pFedMe). In Step 9 of Algorithm 1 (Q-pFedMe) we have

$$q_i^r = \text{Algorithm 2}(x_{i,B}^r, D, \Delta), \tag{19}$$

$$x_i^{r+1} = (1-\beta)x_i^r + \beta q_i^r. \tag{20}$$

Focusing on equation 19, from the operation of Algorithm 2 (DQuFTA), for every $r \ge 1$, we have $q_i^r = q_j^r, \forall v_i, v_j \in \mathcal{V}$. Let us now denote

$$\bar{x}_B^r := \frac{1}{N}\sum_{i=1}^{N} x_{i,B}^r, \text{ and } \bar{q}^r := \frac{1}{N}\sum_{i=1}^{N} q_i^r, \text{ and } \bar{x}^r = \frac{1}{N}\sum_{i=1}^{N} x_i^r, \tag{21}$$

for every $r \ge 0$. Considering equation 21 and also (Bastianello et al., 2023, Lemma 1), for $r \ge 1$ we have

$$||\bar{q}^r - \bar{x}_B^r|| \le 2\sqrt{n}\Delta, \tag{22}$$

$$||q_i^r - \bar{q}^r|| = 0, \tag{23}$$

$$||q_i^r - \bar{x}_B^r|| \le 2\sqrt{n}\Delta, \tag{24}$$

for every $v_i \in \mathcal{V}$.

In Step 6 of Algorithm 1 (Q-pFedMe) we have the local update

$$x_{i,b+1}^r = x_{i,b}^r - \eta \underbrace{\lambda(x_{i,b}^r - \tilde{\theta}_i(x_{i,b}^r))}_{=:g_{i,b}^r}, \tag{25}$$

which implies

$$\eta\sum_{b=0}^{B-1} g_{i,b}^r = \sum_{b=0}^{B-1}(x_{i,b}^r - x_{i,b+1}^r) = x_{i,0}^r - x_{i,B}^r = x_i^r - x_{i,B}^r. \tag{26}$$

From equation 20, we rewrite the average global update (Step 9 of Algorithm 1 (Q-pFedMe)) as

$$\begin{aligned}\bar{x}^{r+1} &= (1-\beta)\bar{x}^r + \beta\bar{q}^r\\ &= (1-\beta)\bar{x}^r + \beta(\bar{x}_B^r + \bar{q}^r - \bar{x}_B^r)\\ &= \bar{x}^r - \frac{\beta}{N}\sum_{i=1}^{N}(x_i^r - x_{i,B}^r) + \beta(\bar{q}^r - \bar{x}_B^r)\\ &= \bar{x}^r - \underbrace{\eta\beta B}_{\tilde{\eta}}\underbrace{\frac{1}{NB}\sum_{i=1}^{N}\sum_{b=0}^{B-1} g_{i,b}^r}_{g^r} + \underbrace{\beta(\bar{q}^r - \bar{x}_B^r)}_{e^r},\end{aligned} \tag{27}$$

where $\|e^r\| \le 2\beta\sqrt{n}\Delta$ follows from equation 22 and we used equation 26.

Because Algorithm 2 (DQuFTA) outputs the same quantized value to all nodes, we have $q_i^r = q_j^r$ for all $v_i, v_j \in \mathcal{V}$. Together with the identical initialization $x_i^0 = x^0$, the update $x_i^{r+1} = (1-\beta)x_i^r + \beta q_i^r$ preserves equality. Hence, for every $r \ge 0$ we have

$$x_i^r = \bar{x}^r, \tag{28}$$

for every $v_i \in \mathcal{V}$ which directly gives a conformity error of zero,

$$\|x_i^{r+1} - \bar{x}^{r+1}\| = 0, \tag{29}$$

and consequently the consensus error is

$$\sigma_{x^r}^2 := \frac{1}{N}\sum_{i=1}^{N}\|x_i^r - \bar{x}^r\|^2 = 0. \tag{30}$$

D.4.2 Supporting lemma

Lemma 3 (One-step average global update)**.** *Let Assumption 3(a) hold. Then, for $\tilde{\eta} \ge \frac{2}{2+\mu_F}$ we have*

$$\begin{aligned}\mathbb{E}\big[\|\bar{x}^{r+1} - x^*\|^2\big] \le \Big(1 - \frac{\tilde{\eta}\mu_F}{4}\Big)\mathbb{E}\big[\|\bar{x}^r - x^*\|^2\big] - \Big(1 + \frac{\mu_F}{4}\Big)\tilde{\eta}\big(2 - 6L_F\tilde{\eta}\big)\mathbb{E}\big[F(\mathbf{x}^r) - F(\mathbf{x}^*)\big] \\ + \Big(1 + \frac{\mu_F}{4}\Big)\frac{\tilde{\eta}(3\tilde{\eta} + 2/\mu_F)}{NB}\sum_{i=1}^{N}\sum_{b=0}^{B-1}\mathbb{E}\big[\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2\big] \\ + \Big(\frac{4}{\mu_F} + 1\Big)\cdot 4\beta^2 n\Delta^2.\end{aligned}$$

Proof. From the average update equation 27 we have $\bar{x}^{r+1} = \bar{x}^r - \tilde{\eta}g^r + e^r$ with $\|e^r\| \le 2\beta\sqrt{n}\,\Delta$. Hence

$$\mathbb{E}[\|\bar{x}^{r+1} - x^*\|^2] = \mathbb{E}[\|\bar{x}^r - \tilde{\eta}g^r - x^*\|^2] + 2\,\mathbb{E}[\langle \bar{x}^r - \tilde{\eta}g^r - x^*,\, e^r\rangle] + \mathbb{E}[\|e^r\|^2].$$

Setting $a^r = \bar{x}^r - \tilde{\eta}g^r - x^*$, by Cauchy–Schwarz and the AM–GM inequality we have

$$2\langle a^r, e^r\rangle \le \frac{\mu_F}{4}\|a^r\|^2 + \frac{4}{\mu_F}\|e^r\|^2, \qquad \|e^r\|^2 \le (2\beta\sqrt{n}\,\Delta)^2 = 4\beta^2 n\Delta^2.$$

Therefore we have

$$\begin{aligned}\mathbb{E}[\|\bar{x}^{r+1} - x^*\|^2] &\le \Big(1 + \frac{\mu_F}{4}\Big)\mathbb{E}[\|\bar{x}^r - \tilde{\eta}g^r - x^*\|^2] + \Big(\frac{4}{\mu_F} + 1\Big)4\beta^2 n\Delta^2 \\ &= \Big(1 + \frac{\mu_F}{4}\Big)\Big(\mathbb{E}[\|\bar{x}^r - x^*\|^2] - 2\tilde{\eta}\,\mathbb{E}[\langle g^r, \bar{x}^r - x^*\rangle] + \tilde{\eta}^2\,\mathbb{E}[\|g^r\|^2]\Big) + \Big(\frac{4}{\mu_F} + 1\Big)4\beta^2 n\Delta^2.\end{aligned}$$

We now bound $-2\tilde{\eta}\,\mathbb{E}[\langle g^r, \bar{x}^r - x^*\rangle]$. Using the definition $g^r = \frac{1}{NB}\sum_{i,b} g_{i,b}^r$ and the μ_F-strong convexity of each F_i, we have

$$\begin{aligned}&-2\tilde{\eta}\,\mathbb{E}[\langle g^r, \bar{x}^r - x^*\rangle] \\ &= -\frac{2\tilde{\eta}}{N}\sum_{i=1}^{N}\mathbb{E}\big[\langle g^r,\, x_i^r - x^*\rangle\big] \\ &= -\frac{2\tilde{\eta}}{NB}\sum_{i,b}\mathbb{E}\Big[\langle g_{i,b}^r - \nabla F_i(x_i^r),\, x_i^r - x^*\rangle + \langle \nabla F_i(x_i^r),\, x_i^r - x^*\rangle\Big] \\ &\le \frac{\tilde{\eta}}{NB}\sum_{i,b}\mathbb{E}\Big[\frac{2}{\mu_F}\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2 + \frac{\mu_F}{2}\|x_i^r - x^*\|^2 - 2\big(F_i(x_i^r) - F_i(x^*)\big) - \mu_F\|x_i^r - x^*\|^2\Big] \\ &= \frac{\tilde{\eta}}{NB}\sum_{i,b}\mathbb{E}\Big[\frac{2}{\mu_F}\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2 - \frac{\mu_F}{2}\|x_i^r - x^*\|^2\Big] + 2\tilde{\eta}\mathbb{E}\big[F(\mathbf{x}^*) - F(\mathbf{x}^r)\big] \\ &= \frac{2\tilde{\eta}}{\mu_F}\frac{1}{NB}\sum_{i,b}\mathbb{E}\big[\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2\big] - \frac{\tilde{\eta}\mu_F}{2}\mathbb{E}\big[\|\bar{x}^r - x^*\|^2\big] + 2\tilde{\eta}\mathbb{E}\big[F(\mathbf{x}^*) - F(\mathbf{x}^r)\big].\end{aligned}$$

The first inequality above uses the Peter–Paul inequality with parameter $\mu_F/2$ for the cross term and the strong convexity of F_i to bound $\langle \nabla F_i(x_i^r), x_i^r - x^* \rangle$ from below.

Next, we bound $\tilde{\eta}^2\,\mathbb{E}[\|g^r\|^2]$. We have

$$\begin{aligned}\tilde{\eta}^2\,\mathbb{E}[\|g^r\|^2] &= \tilde{\eta}^2\,\mathbb{E}\Big[\Big\|\frac{1}{NB}\sum_{i,b} g_{i,b}^r\Big\|^2\Big] \\ &\le 3\tilde{\eta}^2\,\mathbb{E}\Big[\Big\|\frac{1}{NB}\sum_{i,b}(g_{i,b}^r - \nabla F_i(x_i^r))\Big\|^2 \\ &\qquad + \Big\|\frac{1}{N}\sum_{i=1}^N \nabla F_i(x_i^r) - \nabla F(\mathbf{x}^r)\Big\|^2 + \|\nabla F(\mathbf{x}^r)\|^2\Big]. \end{aligned} \tag{31}$$

For the first term inside the expectation, we use the inequality $\big\|\frac{1}{M}\sum_{k=1}^M v_k\big\|^2 \le \frac{1}{M}\sum_{k=1}^M \|v_k\|^2$ (from Jensen's inequality). For the second term, because we established in equation 28 that $x_i^r = \bar{x}^r$ for all i, we have $\frac{1}{N}\sum_{i=1}^N \nabla F_i(x_i^r) = \nabla F(\mathbf{x}^r)$, making this variance term exactly zero. For the third term, the smoothness of F gives us $\|\nabla F(\mathbf{x}^r)\|^2 \le 2L_F\big(F(\mathbf{x}^r) - F(\mathbf{x}^*)\big)$.

Thus we obtain

$$\tilde{\eta}^2\,\mathbb{E}[\|g^r\|^2] \le \frac{3\tilde{\eta}^2}{NB}\sum_{i,b}\mathbb{E}\big[\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2\big] + 6\tilde{\eta}^2 L_F\,\mathbb{E}\big[F(\mathbf{x}^r) - F(\mathbf{x}^*)\big].$$

Plugging both bounds into the expression for $\mathbb{E}[\|\bar{x}^{r+1} - x^*\|^2]$, and noting that the variance term $\mathbb{E}\big[\|\frac{1}{N}\sum_{i=1}^N \nabla F_i(x_i^r) - \nabla F(\mathbf{x}^r)\|^2\big] = 0$ due to the perfect consensus property $x_i^r = \bar{x}^r$ established in equation 28, yields

$$\begin{aligned}\mathbb{E}[\|\bar{x}^{r+1} - x^*\|^2] \le{}& \Big(1 + \frac{\mu_F}{4}\Big)\Big(1 - \frac{\tilde{\eta}\mu_F}{2}\Big)\mathbb{E}[\|\bar{x}^r - x^*\|^2] \\ &- \Big(1 + \frac{\mu_F}{4}\Big)\tilde{\eta}\big(2 - 6L_F\tilde{\eta}\big)\mathbb{E}[F(\mathbf{x}^r) - F(\mathbf{x}^*)] \\ &+ \Big(1 + \frac{\mu_F}{4}\Big)\frac{\tilde{\eta}(3\tilde{\eta} + 2/\mu_F)}{NB}\sum_{i,b}^{N,B}\mathbb{E}[\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2] + \Big(\frac{4}{\mu_F} + 1\Big)4\beta^2 n\Delta^2.\end{aligned}$$

Now, for $\tilde{\eta} \ge \frac{2}{2+\mu_F}$ (a mild condition that holds for typical step sizes), we have

$$\Big(1 + \frac{\mu_F}{4}\Big)\Big(1 - \frac{\tilde{\eta}\mu_F}{2}\Big) \le 1 - \frac{\tilde{\eta}\mu_F}{4}.$$

Substituting this inequality gives the final statement of the lemma.

□

D.4.3 Completing the proof of Theorem 1

Proof. We have

$$\begin{aligned}&\frac{1}{NB}\sum_{i=1}^N\sum_{b=0}^{B-1}\mathbb{E}\big[\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2\big] \\ &\le 2\lambda^2\delta^2 + \frac{16L_F^2\tilde{\eta}^2}{\beta^2}\Big(\frac{1}{N}\sum_{i=1}^N\mathbb{E}\big[\|\nabla F_i(x_i^r)\|^2\big] + \frac{2\lambda^2\delta^2}{B}\Big) &(32)\\ &\le 2\lambda^2\delta^2 + \frac{16L_F^2\tilde{\eta}^2}{\beta^2}\Big(4L_F\,\mathbb{E}\big[F(\mathbf{x}^r) - F(\mathbf{x}^*)\big] + 2\sigma_{F,1}^2 + \frac{2\lambda^2\delta^2}{B}\Big) &(33)\end{aligned}$$

The third term in Lemma 3 is bounded as follows:

$$\Big(1+\frac{\mu_F}{4}\Big)\frac{\tilde\eta(3\tilde\eta+2/\mu_F)}{NB}\sum_{i=1}^{N}\sum_{b=0}^{B-1}\mathbb{E}\big[\|g_{i,b}^r-\nabla F_i(x_i^r)\|^2\big]$$
$$\le\Big(1+\frac{\mu_F}{4}\Big)\tilde\eta\Big(\frac{8}{\mu_F}\Big)\Big(2\lambda^2\delta^2+\frac{16L_F^2\tilde\eta^2}{\beta^2}\Big(4L_F\,\mathbb{E}[F(\mathbf{x}^r)-F(\mathbf{x}^*)]+2\sigma_{F,1}^2+\frac{2\lambda^2\delta^2}{B}\Big)\Big) \tag{34}$$
$$\le\Big(1+\frac{\mu_F}{4}\Big)\Big(\tilde\eta\,\frac{16\delta^2\lambda^2}{\mu_F}+\frac{\tilde\eta^2}{\beta}\,256\kappa_F L_F\,\mathbb{E}[F(\mathbf{x}^r)-F(\mathbf{x}^*)]+\frac{\tilde\eta^3}{\beta^2}\,\frac{256\big(B\sigma_{F,1}^2+\delta^2\lambda^2\big)\kappa_F}{B}\Big) \tag{35}$$

Inequality equation 32 from (T Dinh et al., 2020, lemma 5), inequality equation 33 due to Lemma 2, and plugging inequality equation 33 into inequality equation 34 together with the bound $3\tilde\eta + 2/\mu_F \le \frac{8}{\mu_F}$, which holds for $\tilde\eta \le 2/\mu_F$. Inequality equation 35 uses $\tilde\eta \le \frac{\beta}{2L_F}$ and the definition $\kappa_F := L_F/\mu_F$.

By substituting equation 35 into Lemma 3 we obtain

$$\begin{aligned}\mathbb{E}[\|\bar x^{r+1}-x^*\|^2] \le{}& \Big(1-\frac{\tilde\eta\mu_F}{4}\Big)\mathbb{E}[\|\bar x^r-x^*\|^2]\\ &-\Big(1+\frac{\mu_F}{4}\Big)\tilde\eta\Big(2-6L_F\tilde\eta-\frac{\tilde\eta}{\beta}\,256\,\kappa_F L_F\Big)\mathbb{E}[F(\mathbf{x}^r)-F(\mathbf{x}^*)]\\ &+\tilde\eta\Big(1+\frac{\mu_F}{4}\Big)\frac{16\delta^2\lambda^2}{\mu_F}+\frac{\tilde\eta^3}{\beta^2}\Big(1+\frac{\mu_F}{4}\Big)\frac{256\big(B\sigma_{F,1}^2+\delta^2\lambda^2\big)\kappa_F}{B}\\ &+\Big(\frac{4}{\mu_F}+1\Big)4\beta^2 n\Delta^2.\end{aligned}$$

Let us now define the constants

$$C_1=\Big(1+\frac{\mu_F}{4}\Big)\frac{16\delta^2\lambda^2}{\mu_F},\qquad C_2=\Big(1+\frac{\mu_F}{4}\Big)\frac{256\big(B\sigma_{F,1}^2+\delta^2\lambda^2\big)\kappa_F}{B},\qquad C_3=\Big(\frac{4}{\mu_F}+1\Big)4n\Delta^2.$$

Then the inequality becomes

$$\begin{aligned}\mathbb{E}[\|\bar x^{r+1}-x^*\|^2] \le{}& \Big(1-\frac{\tilde\eta\mu_F}{4}\Big)\mathbb{E}[\|\bar x^r-x^*\|^2]\\ &-\Big(1+\frac{\mu_F}{4}\Big)\tilde\eta\underbrace{\Big(2-6L_F\tilde\eta-\frac{\tilde\eta}{\beta}\,256\,\kappa_F L_F\Big)}_{\ge 1\text{ when }\tilde\eta\text{ satisfies}(36)}\mathbb{E}[F(\mathbf{x}^r)-F(\mathbf{x}^*)]\\ &+\tilde\eta C_1+\frac{\tilde\eta^3}{\beta^2}C_2+\beta^2C_3.\end{aligned}$$

If the step size satisfies

$$2-6L_F\tilde\eta-\frac{\tilde\eta}{\beta}\,256\,\kappa_F L_F\ge 1,$$

or equivalently,

$$\tilde\eta\le\frac{1}{6L_F\big(1+\frac{128}{3}\frac{\kappa_F}{\beta}\big)}\;\le\;\frac{1}{6L_F(1+128\kappa_F/\beta)}\equiv\hat\eta_1, \tag{36}$$

then the coefficient of $\mathbb{E}[F(\mathbf{x}^r)-F(\mathbf{x}^*)]$ is bounded above by $-\tilde\eta$. Thus, we have

$$\mathbb{E}[\|\bar x^{r+1}-x^*\|^2]\le\Big(1-\frac{\tilde\eta\mu_F}{4}\Big)\mathbb{E}[\|\bar x^r-x^*\|^2]-\tilde\eta\,\mathbb{E}[F(\mathbf{x}^r)-F(\mathbf{x}^*)]+\tilde\eta C_1+\frac{\tilde\eta^3}{\beta^2}C_2+\beta^2C_3. \tag{37}$$

We also note that $\tilde\eta \le \min\{\frac{\beta}{2L_F},\frac{2}{\mu_F}\}$ is required for the previous bounds; this is satisfied when $\tilde\eta\le\hat\eta_1$ with $\beta\ge 1$ and $L_F\ge\mu_F$.

Let $\Delta_r = \|\bar{x}^r - x^*\|^2$. From equation 37 we have

$$\mathbb{E}[\Delta_{r+1}] \leq (1 - \tfrac{\tilde{\eta}\mu_F}{4})\,\mathbb{E}[\Delta_r] - \tilde{\eta}\,\mathbb{E}[F(\mathbf{x}^r) - F(\mathbf{x}^*)] + \tilde{\eta}C_1 + \tfrac{\tilde{\eta}^3}{\beta^2}C_2 + \beta^2 C_3.$$

Rearranging we have

$$\tilde{\eta}\,\mathbb{E}[F(\mathbf{x}^r) - F(\mathbf{x}^*)] \leq (1 - \tfrac{\tilde{\eta}\mu_F}{4})\,\mathbb{E}[\Delta_r] - \mathbb{E}[\Delta_{r+1}] + \tilde{\eta}C_1 + \tfrac{\tilde{\eta}^3}{\beta^2}C_2 + \beta^2 C_3. \qquad (*)$$

We now choose weights α_r satisfying $(1 - \frac{\tilde{\eta}\mu_F}{4})\alpha_r = \alpha_{r-1}$ (e.g. $\alpha_r = (1 - \frac{\tilde{\eta}\mu_F}{4})^{-(r+1)}$, $r = 0, 1, \ldots$), we set $A_R = \sum_{r=0}^{R-1} \alpha_r$ and we multiply by $\frac{\alpha_r}{\tilde{\eta}A_R}$ and sum from $r = 0$ to $R-1$ to have

$$\begin{aligned}\sum_{r=0}^{R-1} \frac{\alpha_r}{A_R}\mathbb{E}[F(\mathbf{x}^r)] - F(\mathbf{x}^*) &\leq \frac{1}{\tilde{\eta}A_R}\sum_{r=0}^{R-1}\Big(\underbrace{(1 - \frac{\tilde{\eta}\mu_F}{4})\alpha_r}_{=\alpha_{r-1}}\,\mathbb{E}[\Delta_r] - \alpha_r\mathbb{E}[\Delta_{r+1}]\Big) \\ &\qquad + C_1 + \frac{\tilde{\eta}^2}{\beta^2}C_2 + \frac{\beta^2}{\tilde{\eta}}C_3 \\ &= \frac{1}{\tilde{\eta}A_R}\big(\Delta_0 - \alpha_{R-1}\mathbb{E}[\Delta_R]\big) + C_1 + \frac{\tilde{\eta}^2}{\beta^2}C_2 + \frac{\beta^2}{\tilde{\eta}}C_3.\end{aligned}$$

Because $\alpha_{R-1}\mathbb{E}[\Delta_R] \geq 0$, the term $-\frac{\alpha_{R-1}}{\tilde{\eta}A_R}\mathbb{E}[\Delta_R]$ is non-positive and can be removed from our upper bound. Thus, we have

$$\sum_{r=0}^{R-1} \frac{\alpha_r}{A_R}\mathbb{E}[F(\mathbf{x}^r)] - F(\mathbf{x}^*) \leq \frac{\Delta_0}{\tilde{\eta}A_R} + C_1 + \frac{\tilde{\eta}^2}{\beta^2}C_2 + \frac{\beta^2}{\tilde{\eta}}C_3.$$

Now we estimate $\frac{1}{\tilde{\eta}A_R}$. Let $q = 1 - \frac{\tilde{\eta}\mu_F}{4}$ and assume $\tilde{\eta}R \geq \frac{4}{\mu_F}$, which ensures $q^R \leq e^{-1} < \frac{1}{2}$. Then $A_R = \frac{q^{-R}-1}{1-q} \geq \frac{q^{-R}}{2(1-q)}$. Substituting $1 - q = \frac{\tilde{\eta}\mu_F}{4}$ gives

$$\frac{1}{\tilde{\eta}A_R} \leq \frac{2(1-q)}{\tilde{\eta}}q^R = \frac{\mu_F}{2}\,q^R \leq \frac{\mu_F}{2}\,e^{-\tilde{\eta}\mu_F R/4}.$$

Thus our bound presented above becomes

$$\sum_{r=0}^{R-1} \frac{\alpha_r}{A_R}\mathbb{E}[F(\mathbf{x}^r)] - F(\mathbf{x}^*) \leq \frac{\mu_F}{2}\Delta_0 e^{-\tilde{\eta}\mu_F R/4} + C_1 + \frac{\tilde{\eta}^2}{\beta^2}C_2 + \frac{\beta^2}{\tilde{\eta}}C_3.$$

By convexity of F, the left-hand side dominates $\mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)]$ with $\bar{\mathbf{x}}_R = \sum_{r=0}^{R-1} \frac{\alpha_r}{A_R}\mathbf{x}^r$. Absorbing the factor $\frac{1}{2}$ into the $\mathcal{O}$ notation (or simply using the slightly larger bound $\mu_F\Delta_0 e^{-\tilde{\eta}\mu_F R/4}$) gives the statement of Theorem 1(a).

Hence, we obtain

$$\begin{aligned}A_R &= \sum_{r=0}^{R-1}\left(1 - \frac{\tilde{\eta}\mu_F}{4}\right)^{-(r+1)} = \left(1 - \frac{\tilde{\eta}\mu_F}{4}\right)^{-R}\sum_{r=0}^{R-1}\left(1 - \frac{\tilde{\eta}\mu_F}{4}\right)^{r} \\ &= \left(1 - \frac{\tilde{\eta}\mu_F}{4}\right)^{-R}\frac{1 - \left(1 - \frac{\tilde{\eta}\mu_F}{4}\right)^{R}}{\tilde{\eta}\mu_F/4} \\ &\geq \left(1 - \frac{\tilde{\eta}\mu_F}{4}\right)^{-R}\frac{1}{\tilde{\eta}\mu_F} = \frac{\alpha_{R-1}}{\tilde{\eta}\mu_F}\end{aligned}$$

which implies $\frac{\alpha_{R-1}}{\tilde{\eta}A_R} \leq \mu_F$ and thus $\frac{1}{\tilde{\eta}A_R} \leq \mu_F\big(1 - \frac{\tilde{\eta}\mu_F}{4}\big)^R \leq \mu_F e^{-\tilde{\eta}\mu_F R/4}$.

By convexity of F, we have

$$\mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \leq \sum_{r=0}^{R-1} \frac{\alpha_r}{A_R} \mathbb{E}[F(\mathbf{x}^r)] - F(\mathbf{x}^*).$$

Substituting the bound on the right-hand side from the telescoping sum and using $\frac{1}{\tilde{\eta} A_R} \leq \mu_F e^{-\tilde{\eta}\mu_F R/4}$ from the previous step yields

$$\mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \leq \mu_F \Delta_0 e^{-\tilde{\eta}\mu_F R/4} + C_1 + \frac{\tilde{\eta}^2}{\beta^2} C_2 + \frac{\beta^2}{\tilde{\eta}} C_3 \tag{38}$$

Next, following the approach of Stich (2019); Arjevani et al. (2020), we focus on two cases. If R is finite and the initial error $\mu_F \Delta_0$ is non-negligible, we set $\tilde{\eta} = \hat{\eta}_1$ (the largest allowed step size) to accelerate convergence, and we have

$$\mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \leq \mu_F \Delta_0 e^{-\hat{\eta}_1\mu_F R/4} + C_1 + \frac{\hat{\eta}_1^2}{\beta^2} C_2 + \frac{\beta^2}{\hat{\eta}_1} C_3.$$

When $R \to \infty$, the exponential term vanishes and we can tune the step size to minimize the remaining constant error. The optimal step size is $\tilde{\eta}_{\text{opt}} = \left(\frac{\beta^4 C_3}{2C_2}\right)^{1/3}$.

To ensure convergence we need $\tilde{\eta} \leq \hat{\eta}_1$, where $\hat{\eta}_1 = \frac{1}{6L_F(1+128\kappa_F/\beta)}$ (and also recall that $\tilde{\eta} \geq \frac{2}{2+\mu_F}$ from Lemma 3). We consider the two following cases.

1. If $\hat{\eta}_1 \geq \tilde{\eta}_{\text{opt}}$ and $R \to \infty$, we can choose $\tilde{\eta} = \tilde{\eta}_{\text{opt}}$, and we have

$$\limsup_{R\to\infty} \mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \leq \mathcal{O}\Big(\beta^{2/3}(n\Delta^2)^{2/3}\Big(\frac{(B\sigma_{F,1}^2 + \delta^2\lambda^2)\kappa_F}{B}\Big)^{1/3} + \frac{\delta^2\lambda^2}{\mu_F}\Big).$$

2. If $\hat{\eta}_1 < \tilde{\eta}_{\text{opt}}$ and $R \to \infty$, we are choose $\tilde{\eta} = \hat{\eta}_1$, which gives us

$$\limsup_{R\to\infty} \mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \leq \mathcal{O}\Big(\frac{\beta^2 n\Delta^2}{\hat{\eta}_1} + \frac{(B\sigma_{F,1}^2 + \delta^2\lambda^2)\kappa_F}{\beta^2}\hat{\eta}_1^2 + \frac{\delta^2\lambda^2}{\mu_F}\Big).$$

Combining the two cases, we obtain

$$\begin{aligned}
&\limsup_{R\to\infty} \mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)] \\
&\leq \min\left\{\mathcal{O}\left(\beta^{2/3}(n\Delta^2)^{2/3}\left(\frac{(B\sigma_{F,1}^2 + \delta^2\lambda^2)\kappa_F}{B}\right)^{1/3} + \frac{\delta^2\lambda^2}{\mu_F}\right),\ \mathcal{O}\left(\frac{\beta^2 n\Delta^2}{\hat{\eta}_1} + \frac{(B\sigma_{F,1}^2 + \delta^2\lambda^2)\kappa_F}{\beta^2}\hat{\eta}_1^2 + \frac{\delta^2\lambda^2}{\mu_F}\right)\right\}
\end{aligned}$$

which finishes the proof of part (a).

We next prove part (b) as follows

$$\begin{aligned}
&\mathbb{E}\big[\|\tilde{\theta}_i^R(x_i^R) - x^*\|^2\big] \\
&\leq 3\mathbb{E}[\|\tilde{\theta}_i^R(x_i^R) - \hat{\theta}_i^R(x_i^R)\|^2 + \|\hat{\theta}_i^R(x_i^R) - x_i^R\|^2 + \|x_i^R - x^*\|^2] \\
&\leq 3\big(\delta^2 + \frac{1}{\lambda^2}\mathbb{E}[\|\nabla F_i(x_i^R)\|^2] + \mathbb{E}[\|x_i^R - x^*\|^2]\big) \\
&\leq 3\big(\delta^2 + \frac{2}{\lambda^2}\mathbb{E}\big[\|\nabla F_i(x_i^R) - \nabla F_i(x^*)\|^2 + \|\nabla F_i(x^*)\|^2\big] + \mathbb{E}[\|x_i^R - x^*\|^2]\big) \\
&\leq 3\big(\delta^2 + 3\mathbb{E}\big[\|x_i^R - x^*\|^2\big] + \frac{2}{\lambda^2}\|\nabla F_i(x^*)\|^2\big),
\end{aligned}$$

where the last inequality is due to smoothness of F_i with $L_F = \lambda$ according to Proposition 1.

Taking the average over N nodes, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\|\tilde{\theta}_i^R(x_i^R) - x^*\|^2\big] &\leq 9\mathbb{E}\big[\|\bar{x}^R - x^*\|^2\big] + \frac{6\sigma_{F,1}^2}{\lambda^2} + 3\delta^2 \\
&\leq \frac{1}{\mu_F}\mathcal{O}\big(\mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)]\big) + \mathcal{O}\left(\frac{\sigma_{F,1}^2}{\lambda^2} + \delta^2\right),
\end{aligned}$$

where the last inequality uses the strong convexity of F to bound $\mathbb{E}\|\bar{x}^R - x^*\|^2 \leq \frac{2}{\mu_F}\mathbb{E}[F(\bar{\mathbf{x}}_R) - F(\mathbf{x}^*)]$, together with equation 38. This completes our proof.

□

D.5 Proof of Theorem 2

Proof. We first prove part (a). Due to the L_F-smoothness of $F(\cdot)$, we have

$$
\begin{aligned}
&\mathbb{E}\big[F(\mathbf{x}^{r+1}) - F(\mathbf{x}^r)\big] \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\bigg(\mathbb{E}[\langle \nabla F_i(x_i^r), x_i^{r+1} - x_i^r\rangle] + \frac{L_F}{2}\mathbb{E}[\|x_i^{r+1} - x_i^r\|^2]\bigg) \\
&= \frac{1}{N}\sum_{i=1}^{N}\bigg(-\beta\mathbb{E}[\langle \nabla F_i(x_i^r), x_i^r - \bar{x}_B^r\rangle] + \beta\mathbb{E}[\langle \nabla F_i(x_i^r), q_i^r - \bar{x}_B^r\rangle] + \frac{L_F}{2}\mathbb{E}[\|-\beta(x_i^r - \bar{x}_B^r) + \beta(q_i^r - \bar{x}_B^r)\|^2]\bigg) \qquad (39) \\
&= \frac{1}{N}\sum_{i=1}^{N}\bigg(-\tilde{\eta}\mathbb{E}[\langle \nabla F_i(x_i^r), g^r\rangle] + \beta\mathbb{E}[\langle \nabla F_i(x_i^r), q_i^r - \bar{x}_B^r\rangle] + \frac{L_F}{2}\mathbb{E}[\|-\tilde{\eta}g^r + \beta(q_i^r - \bar{x}_B^r)\|^2]\bigg) \qquad (40) \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\bigg(-\frac{\tilde{\eta}}{4}\mathbb{E}[\|\nabla F_i(x_i^r)\|^2] + \frac{\tilde{\eta}}{2}\mathbb{E}[\|g^r - \nabla F_i(x_i^r)\|^2] + (\frac{\beta^2}{\tilde{\eta}} + L_F\beta^2)\mathbb{E}[\|q_i^r - \bar{x}_B^r\|^2] + L_F\tilde{\eta}^2\mathbb{E}[\|g^r\|^2]\bigg) \qquad (41) \\
&\leq (3L_F\tilde{\eta}^2 - \frac{\tilde{\eta}}{4})\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + (3L_F\tilde{\eta}^2 + \frac{\tilde{\eta}}{2})\frac{1}{NB}\sum_{i,b}^{N,B}\mathbb{E}[\|g_{i,b}^r - \nabla F_i(x_i^r)\|^2] \\
&\quad + (\frac{\beta^2}{\tilde{\eta}} + L_F\beta^2)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|q_i^r - \bar{x}_B^r\|^2]\bigg) \qquad (42) \\
&\leq (3L_F\tilde{\eta}^2 - \frac{\tilde{\eta}}{4})\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + (\frac{\beta^2}{\tilde{\eta}} + L_F\beta^2)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|q_i^r - \bar{x}_B^r\|^2]\bigg) \\
&\quad + (3L_F\tilde{\eta}^2 + \frac{\tilde{\eta}}{2})\bigg(2\lambda^2\delta^2 + \frac{16L_F^2\tilde{\eta}^2}{\beta^2}\bigg(\frac{3}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla F_i(x_i^r) - \nabla F(\mathbf{x}^r)\|^2] + 3\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + \frac{2\lambda^2\delta^2}{B}\bigg)\bigg) \qquad (43) \\
&\leq -\tilde{\eta}\underbrace{\bigg(\frac{1}{4} - 3\tilde{\eta}L_F\bigg(1 + \frac{16\tilde{\eta}}{\beta^2}(\frac{1}{2} + 3L_F\tilde{\eta})\frac{\lambda^2}{\lambda^2 - 8L^2}\bigg)\bigg)}_{\geq 1/8 \text{ when } \tilde{\eta} \text{ satisfied (46)}}\mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] \\
&\quad + \tilde{\eta}(\frac{1}{2} + 3L_F\tilde{\eta})\bigg(\frac{16L_F^2\tilde{\eta}^2}{\beta^2}(\frac{2\lambda^2\delta^2}{B} + 6\sigma_{F,2}^2) + 2\lambda^2\delta^2\bigg) + (\frac{\beta^2}{\tilde{\eta}} + L_F\beta^2)\cdot 2n\Delta^2 \qquad (44) \\
&\leq -\frac{\tilde{\eta}}{8}\mathbb{E}\big[\|\nabla F(\mathbf{x}^r)\|^2\big] + \tilde{\eta}\underbrace{4\beta\bigg(\frac{2\lambda^2\delta^2}{B} + \lambda^2\delta^2 + 12\sigma_{F,2}^2\bigg)}_{C_4} + \frac{1}{\tilde{\eta}}\underbrace{\beta^2\cdot 2n\Delta^2}_{C_5} + \underbrace{L_F\beta^2\cdot 2n\Delta^2}_{C_6} \qquad (45)
\end{aligned}
$$

where (39) is by re-arranging (20), i.e. $x_i^{r+1} - x_i^r = -\beta(x_i^r - \bar{x}_B^r) + \beta(q_i^r - \bar{x}_B^r)$, (40) is by the second term of (27), (41) is due to Cauchy-Swartz and AM-GM inequalities, (42) is by decomposing $\tilde{\eta}^2\|g^r\|^2$ into three terms by (31), and (43) is by using (T Dinh et al., 2020, Lemma 5), and the fact that $\mathbb{E}[\|X\|^2] = \mathbb{E}[\|X - \mathbb{E}[X]\|]^2 + \mathbb{E}[\|X\|]^2$ for any vector of random variable X. We have (44) by Lemma 2 (b) and (24), re-arranging the terms, and having $\frac{1}{2} + 3L_F\tilde{\eta} \leq \frac{1}{2} + \frac{3}{2}\beta \leq 2\beta$ when $\tilde{\eta} \leq \frac{\beta}{2L_F}$ according to (T Dinh et al., 2020, lemma 5) and $\beta \geq 1$. Finally, we have (45) by using the condition $\lambda^2 - 8L^2 \geq 1$ for the following

$$
L_F\bigg(1 + \frac{16\tilde{\eta}}{\beta^2}(\frac{1}{2} + 3L_F\tilde{\eta})\frac{\lambda^2}{\lambda^2 - 8L^2}\bigg) \leq L_F\bigg(1 + \frac{32\tilde{\eta}\lambda^2}{\beta}\bigg) \leq L_F + 16\lambda^2
$$

to get

$$\frac{1}{4} - 3\tilde{\eta}L_F\left(1 + \frac{16\tilde{\eta}}{\beta^2}(\frac{1}{2} + 3L_F\tilde{\eta})\frac{\lambda^2}{\lambda^2 - 8L^2}\right) \geq \frac{1}{4} - 3\tilde{\eta}\left(L_F + 16\lambda^2\right) \geq \frac{1}{8}$$

with the condition

$$\tilde{\eta} \leq \frac{1}{24(L_F + 16\lambda^2)} =: \hat{\eta}_2. \tag{46}$$

We note that $\hat{\eta}_2 \leq \frac{\beta}{2L_F}$ with $\beta \geq 1$. By re-arranging the terms of (45) and telescoping, we have

$$\frac{1}{8R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] \leq \frac{\Delta_F}{\tilde{\eta}R} + C_4 + \frac{1}{\tilde{\eta}^2}C_5 + \frac{1}{\tilde{\eta}}C_6.$$

Defining $\Delta_F := F(\mathbf{x}^0) - F(\mathbf{x}^*)$, and following the techniques used by Stich (2019); Arjevani et al. (2020), we have $\tilde{\eta}_{opt} = \hat{\eta}_2$, then,

$$\frac{1}{8R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] \leq \frac{\Delta_F}{\hat{\eta}_2 R} + C_4 + \frac{1}{\hat{\eta}_2^2}C_5 + \frac{1}{\hat{\eta}_2}C_6.$$

With r^* uniformly sampled from $0, \ldots, R-1$ we have

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] = \mathbb{E}[\|\nabla F(\mathbf{x}^{r^*})\|^2] \leq \mathcal{O}\big(\mathbb{E}[\|\nabla F(\mathbf{x}^{r^*})\|^2]\big) :=$$
$$\mathcal{O}\left(\frac{\Delta_F}{\hat{\eta}_2 R} + \beta\big(\frac{\lambda^2\delta^2}{B} + \lambda^2\delta^2 + \sigma_{F,2}^2\big) + \frac{\beta^2}{\hat{\eta}_2^2}\cdot n\Delta^2 + \frac{L_F\beta^2}{\hat{\eta}_2}\cdot n\Delta^2\right)$$

We next prove part (b) as follows:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\tilde{\theta}_i(x_i^r) - x_i^r\|^2] &\leq \frac{1}{N}\sum_{i=1}^{N}2\mathbb{E}[\|\tilde{\theta}_i(x_i^r) - \hat{\theta}_i^r\|^2 + \|\hat{\theta}_i^r - x_i^r\|^2] \\ &\leq 2\delta^2 + \frac{2}{N}\sum_{i=1}^{N}\frac{\mathbb{E}[\|\nabla F_i(x_i^r)\|^2]}{\lambda^2} \\ &\leq 2\delta^2 + \frac{2}{\lambda^2 - 8L^2}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + \frac{2\sigma_{F,2}^2}{\lambda^2}, \end{aligned} \tag{47}$$

where the first inequality is due to Proposition (3) and the third inequality is by using the fact that $\mathbb{E}[\|X\|^2] = \mathbb{E}[\|X - \mathbb{E}[X]\|^2] + \mathbb{E}[\|X\|]^2$ for any vector of random variable X, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla F_i(x_i^r)\|^2] &= \sum_{i=1}^{N}\frac{1}{N}\left(\mathbb{E}[\|\nabla F_i(x_i^r) - \nabla F(\mathbf{x}^r)\|^2] + \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2]\right) \\ &\leq 2\sigma_{F,2}^2 + \frac{\lambda^2}{\lambda^2 - 8L^2}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2].\end{aligned}$$

Summing (47) from $r = 0$ to R, we get

$$\begin{aligned}&\frac{1}{RN}\sum_{i=0}^{R-1}\sum_{i=1}^{N}\mathbb{E}[\|\tilde{\theta}_i^r - x_i^r\|^2] \\ &\leq \frac{2}{\lambda^2 - 8L^2}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + 2\delta^2 + \frac{2\sigma_{F,2}^2}{\lambda^2},\end{aligned}$$

and with r^* uniformly sampled from $0, \ldots, R-1$, we finish the proof. □

D.6 PROOF OF THEOREM 3

Proof. In the global iteration r, the input of Algorithm 2 is $y_i^{[0]} = (q_\Delta(x_{i,B}^r))/\Delta = \lfloor \frac{x_{i,B}^r}{\Delta} \rfloor$. Let $\bar{y} := \frac{1}{N}\sum_{i=1}^N y_i^{[0]} = \frac{1}{N}\sum_{i=1}^N \lfloor \frac{x_{i,B}^r}{\Delta} \rfloor$, and $\bar{x}_B^r := \frac{1}{N}\sum_{i=1}^N x_{i,B}^r$. In Algorithm 2, the quantized averaging procedure relies on (Rikos et al., 2022, Algorithm 1), where the number of required time steps T for reaching agreement $\bar{y}$ depends on the the number of nodes N, network diameter D, τ (which depends on parameter of network and probability of convergence), and the nodes' initial states y^{init} (Rikos et al., 2022, Theorem 1). From (Rikos et al., 2022, Remark 2), we know (Rikos et al., 2022, Algorithm 1) is also for the special case where the network is a strongly connected and directed graph (Assumption 1). In addition, Algorithm 2 implements a min- and max-consensus operation among nodes every D time steps, and terminates when $M_i - m_i \leq 1$, which means $m_i = \lfloor \bar{y} \rfloor$ when Algorithm 2 terminates. Thus, in Algorithm 2, each node $v_i \in \mathcal{V}$ reaches exact agreement $q_i^r = \Delta \lfloor \bar{y} \rfloor = \Delta \lfloor \frac{1}{N}\sum_{i=1}^N \lfloor \frac{x_{i,B}^r}{\Delta} \rfloor \rfloor$. Unlike (Rikos et al., 2022, Algorithm 1), in Algorithm 2, $y_i^{[0]}$ is an integer vector, while in (Rikos et al., 2022, Algorithm 1), $y_i^{[0]}$ is an integer. Therefore, the number of time steps for reaching agreement for our algorithm 2 should be equal to $(\mathbb{E}[\|y^{init}\|_\infty] + N)\tau D$ (we only focus on the maximum value of y^{init}, so the infinity norm is taken), where N is the number of nodes, D is the diameter of digraph, τ depends on the parameters of digraph and probability of convergence from (Rikos et al., 2022, Theorem 1), and

$$\mathbb{E}[\|y^{init}\|_\infty] = \mathbb{E}[\| \sum_{v_i \in \mathcal{V}: y_i^{[0]} > \lceil \bar{y} \rceil} (y_i^{[0]} - \lceil \bar{y} \rceil) + \sum_{v_i \in \mathcal{V}: y_i^{[0]} < \lfloor \bar{y} \rfloor} (\lfloor \bar{y} \rfloor - y_i^{[0]}) \|_\infty] \tag{48}$$

$$\leq \mathbb{E}[\| \sum_{i=1}^N |y_i^{[0]} - \lfloor \bar{y} \rfloor| \|_\infty] \tag{49}$$

$$\leq \mathbb{E}[\| \sum_{i=1}^N |\frac{x_{i,B}^r - \bar{x}_B^r}{\Delta}| + 2N \|_\infty] \tag{50}$$

$$\leq \frac{1}{\Delta}\mathbb{E}[\| \sum_{i=1}^N x_{i,B}^r - \bar{x}_B^r \|] + 2N \tag{51}$$

$$\leq \frac{1}{\Delta}\mathbb{E}[\sqrt{N}\sqrt{\sum_{i=1}^N \|x_{i,B}^r - \bar{x}_B^r\|^2}] + 2N \tag{52}$$

$$= \frac{N}{\Delta}\sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|x_{i,B}^r - \bar{x}_B^r\|^2]} + 2N \tag{53}$$

In equation 48, the computation of y^{init} is from (Rikos et al., 2022, Theorem 1). For the first term of equation 48, when $y_i^{[0]} > \lceil \bar{y} \rceil$, $y_i^{[0]} - \lceil \bar{y} \rceil \leq y_i^{[0]} - \lfloor \bar{y} \rfloor$ holds. And for the second term of equation 48, when $y_i^{[0]} < \lfloor \bar{y} \rfloor$, $\lfloor \bar{y} \rfloor - y_i^{[0]} = |y_i^{[0]} - \lfloor \bar{y} \rfloor|$ hold. Thus, $0 \leq y^{init} \leq \sum_{i=1}^N |y_i^{[0]} - \lfloor \bar{y} \rfloor|$ must hold.

Next, we explain equation 50. We have $y_i^{[0]} = \frac{x_{i,B}^r}{\Delta} - \{\frac{x_{i,B}^r}{\Delta}\}$, $\lfloor \bar{y} \rfloor = \bar{y} - \{\bar{y}\} = \frac{1}{N}\sum_{i=1}^N \frac{x_{i,B}^r}{\Delta} - \frac{1}{N}\sum_{i=1}^N \{\frac{x_{i,B}^r}{\Delta}\} - \{\bar{y}\} = \frac{\bar{x}_B^r}{\Delta} - \frac{1}{N}\sum_{i=1}^N \{\frac{x_{i,B}^r}{\Delta}\} - \{\bar{y}\}$, where $\{\cdot\}$ indicates the operation of extracting the decimal part. Then, we get

$$|y_i^{[0]} - \lfloor \bar{y} \rfloor| = |\frac{x_{i,B}^r - \bar{x}_B^r}{\Delta}| + |\underbrace{\frac{1}{N}\sum_{i=1}^N \{\frac{x_{i,B}^r}{\Delta}\} - \{\frac{x_{i,B}^r}{\Delta}\} + \{\bar{y}\}}_{\sigma_i}|$$

$$\leq |\frac{x_{i,B}^r - \bar{x}_B^r}{\Delta}| + 2,$$

where $-1 < \sigma_i < 2, |\sigma_i| < 2$.

equation 51 is due to the L_2 norm is greater than or equal to the absolute value of any component, and equation 52 is due to Cauchy-Schwarz.

Next, we bound $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|x_{i,B}^r - \bar{x}_B^r\|^2]$.

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|x_{i,B}^r - \bar{x}_B^r\|^2] \leq \frac{1}{N}\sum_{i=1}^{N}\left(2\mathbb{E}[\|x_{i,B}^r - x_i^r\|^2] + 2\mathbb{E}[\|x_i^r - \bar{x}_B^r\|^2]\right) \tag{54}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}\left(2\mathbb{E}[\|x_{i,B}^r - x_i^r\|^2] + \frac{1}{N}\sum_{i=1}^{N}2\mathbb{E}[\|x_i^r - x_{i,B}^r\|^2]\right) \tag{55}$$

$$= 4\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|x_i^r - x_{i,B}^r\|^2]$$

$$\leq 4\frac{1}{N}\sum_{i=1}^{N}\left(\frac{8\tilde{\eta}^2}{\beta^2}(3\mathbb{E}[\|\nabla F_i(x_i^r)\|^2 + \frac{2\lambda^2\delta^2}{B}])\right) \tag{56}$$

$$= \frac{32\tilde{\eta}^2}{\beta^2}(3\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla F_i(x_i^r)\|^2] + \frac{2\lambda^2\delta^2}{B}) \tag{57}$$

equation 54 is due to the triangle inequality, i.e., $\mathbb{E}[\|x_{i,B}^r - x_i^r + x_i^r - \bar{x}_B^r\|^2] \leq 2\mathbb{E}[\|x_{i,B}^r - x_i^r\|^2] + 2\mathbb{E}[\|x_i^r - \bar{x}_B^r\|^2]$. equation 55 is due to the inequality $\left\|\frac{1}{M}\sum_{k=1}^{M}v_k\right\|^2 \leq \frac{1}{M}\sum_{k=1}^{M}\|v_k\|^2$ (from Jensen's inequality). equation 56 is from (T Dinh et al., 2020, (22)). Substituting equation 57 into equation 53, we have

$$\mathbb{E}[\|y^{init}\|_\infty] \leq \frac{N}{\Delta}\sqrt{\frac{32\tilde{\eta}^2}{\beta^2}(3\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla F_i(x_i^r)\|^2 + \frac{2\lambda^2\delta^2}{B})} + 2N$$

For the case of strong convexity, from Lemma 2 (a), we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla F_i(x_i^r)\|^2] \leq 4L_F(F(\mathbf{x}^r) - F(\mathbf{x}^*)) + 2\sigma_{F,1}^2$$

For the case of nonconvex and smooth, we have

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla F_i(x_i^r)\|^2] = \sum_{i=1}^{N}\frac{1}{N}\left(\mathbb{E}[\|\nabla F_i(x_i^r) - \nabla F(\mathbf{x}^r)\|^2] + \mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2]\right)$$

$$\leq \frac{\lambda^2}{\lambda^2 - 8L^2}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + 2\sigma_{F,2}^2, \tag{58}$$

where equation 58 is due to Lemma 2 (b). Thus, we have

$$\mathbb{E}[\|y^{init}\|_\infty]$$

$$\leq \begin{cases} \frac{8N\tilde{\eta}}{\Delta\beta}\sqrt{6L_F(F(\mathbf{x}^r) - F(\mathbf{x}^*)) + \frac{\lambda^2\delta^2}{B} + 3\sigma_{F,1}^2} + 2N, \text{when Assumptions 1, 2, 3 (a), 4, and 5 hold} \\ \frac{8N\tilde{\eta}}{\Delta\beta}\sqrt{\frac{3\lambda^2}{2(\lambda^2 - 8L^2)}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + \frac{\lambda^2\delta^2}{2B} + \sigma_{F,2}^2)} + 2N, \text{when Assumptions 1, 2, 3 (b), 4, and 5 hold} \end{cases}$$

where $\tilde{\eta} = \eta\beta B$.

Thus, we get

$$T \leq \begin{cases} \left(\frac{8N\eta B}{\Delta}\sqrt{6L_F(F(\mathbf{x}^r) - F(\mathbf{x}^*)) + \frac{\lambda^2\delta^2}{B} + 3\sigma_{F,1}^2} + 3N\right)\tau D, \text{when Assumptions 1, 2, 3 (a), 4, and 5 hold} \\ \left(\frac{8N\eta B}{\Delta}\sqrt{\frac{3\lambda^2}{2(\lambda^2 - 8L^2)}\mathbb{E}[\|\nabla F(\mathbf{x}^r)\|^2] + \frac{\lambda^2\delta^2}{2B} + \sigma_{F,2}^2} + 3N\right)\tau D, \text{when Assumptions 1, 2, 3 (b), 4, and 5 hold} \end{cases}$$

□