

DPF-SGD: Byzantine-Robust Differentially Private Federated SGD with Exact Optimality Guarantees

Likang Lu, Jiaxin Wu

Abstract

Federated optimization enables collaborative learning over decentralized data but is accompanied by significant privacy risks due to the exchange of model updates. While differential privacy (DP) provides formal protection, optimization accuracy is often degraded, and robustness becomes particularly limited in the presence of Byzantine agents. A Byzantine-robust, differentially private federated SGD algorithm, DPF-SGD, is proposed, in which private step sizes are integrated with a DP noise-decay schedule to regulate gradient updates. To further enhance robustness, a privacy-preserving comparative elimination (PPCE) filter is developed to suppress Byzantine contributions during aggregation while preserving privacy guarantees. Almost sure convergence to the global optimum under ϵ -differential privacy is established, despite persistent noise and adversarial behavior. Experiments on synthetic and MNIST datasets demonstrate one- to two-order-of-magnitude reductions in final optimization error relative to DP baselines.

Keywords: Federated optimization, Differentially privacy, Byzantine robustness, SGD algorithm

1. Introduction

Federated optimization and learning enable the training of global models from geographically distributed datasets without requiring the exchange of raw data [1, 2]. This paradigm has become increasingly important in applications

Email address: llk2813@163.com, lulikang.1@jd.com (Likang Lu)

such as sensor network localization [3], multi-robot coordination [4], power system monitoring [5], and large-scale machine learning [6]. In these settings, federated stochastic gradient descent (SGD) allows multiple agents to collaboratively optimize a shared objective while keeping local datasets private.

In a typical federated SGD architecture [7, 8], each agent maintains a local model and updates it using gradients computed from mini-batches of its local data, together with a global model provided by a central coordinator. To handle the scale and heterogeneity of modern datasets, federated optimization commonly relies on stochastic gradient estimates. In particular, each agent randomly samples a mini-batch of local data and computes the gradient of the empirical loss over this subset, yielding an unbiased estimator of the full local gradient for the SGD update [9]. These estimates substantially reduce computational overhead and communication costs, making them well suited for large-scale and resource-constrained environments [10]. Moreover, stochastic methods naturally tolerate measurement noise commonly encountered in IoT and edge networks [11]. As IoT and edge devices proliferate, such noise and system-scale challenges further motivate the use of stochastic optimization methods in federated learning systems.

Although SGD dates back to the 1950s [12], with decentralized variants emerging in the 1980s [13], it remains foundational for modern large-scale learning. By exchanging model updates rather than raw data, federated SGD architectures offer improved data privacy [6]. Nevertheless, recent studies reveal that federated SGD remains vulnerable to both privacy and robustness threats. Model updates may leak sensitive information about local datasets, enabling attacks such as membership inference, model inversion, and even partial data reconstruction [14]. In addition, the distributed nature of federated learning makes it vulnerable to Byzantine attacks that can disrupt the convergence of training [15, 16]. These challenges motivate the development of federated SGD methods that ensure both convergence and privacy.

DP [17] is widely adopted to enhance privacy in federated SGD by injecting carefully calibrated noise into model updates, ensuring that the output is

insensitive to any single data point. This approach mitigates potential information leakage while enabling rigorous analysis, making DP a de facto standard for privacy-preserving federated learning. However, achieving a practical privacy–utility trade-off is challenging: excessive noise reduces accuracy, whereas insufficient noise compromises privacy. Regulatory frameworks such as the European Union’s GDPR also emphasize “privacy by design” and “data minimization” [18, 9] informing practical privacy requirements. Motivated by these considerations, our DPF-SGD algorithm manages the privacy–utility trade-off under practically relevant privacy budgets (e.g., $\epsilon \in [1, 10]$) [19] by tailoring each agent’s gradient updates with a private step size, facilitating practical federated learning (FL) deployment under regulatory constraints.

Federated SGD is also vulnerable to Byzantine threats [16, 7, 8, 15]. Attackers can submit carefully crafted false updates, disrupting convergence. Recent Byzantine-robust approaches detect and filter anomalous updates; however, some fail to jointly guarantee differential privacy and Byzantine robustness [8], while others focus on parameter tuning yet yield only limited gains [16, 7]. Naively adding DP noise into Byzantine-robust algorithms degrades optimization accuracy, introducing an inherent privacy–accuracy trade-off. Consequently, existing approaches [15] attain only d -accuracy (see Definition 2 in [2]) for a small $d > 0$ relative to the optimal solution, slowing convergence and reducing resilience to attacks that combine malicious manipulation with privacy inference.

Our contribution. We propose DPF-SGD, a novel federated SGD algorithm that jointly achieves differential privacy and Byzantine robustness. The method integrates private step sizes into a DP noise–decay schedule to regulate local gradient updates, while mitigating Byzantine interference via a PPCE filter. We further establish convergence of the global model to the exact global optimum under ϵ -differential privacy, despite persistent DP noise and Byzantine behavior. The main contributions are summarized as follows:

- **Differential Privacy:** Private step sizes and calibrated DP noise preserve

gradient privacy, providing resilience against both stochastic perturbations and adversarial inference.

- **Byzantine Robustness:** The PPCE filter ensures robust aggregation and stable convergence in the presence of Byzantine agents.
- **Communication Efficiency:** Each agent communicates only a single variable per iteration, significantly reducing communication overhead and improving scalability.
- **Convergence and Optimality:** We prove almost sure convergence to the global optimum under ϵ -differential privacy, and demonstrate one- to two-order-of-magnitude reductions in final optimization error on synthetic and MNIST datasets relative to DP baselines.

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 introduces the notation and problem formulation. Section 4 presents the proposed DPF-SGD algorithm. Section 5 establishes its convergence guarantees for $\mathcal{T} = 1$ and $\mathcal{T} = 2$. Section 6 provides the differential privacy analysis. Section 7 presents experimental results on synthetic and MNIST datasets. Finally, Section 8 concludes.

2. Related Work

2.1. Accuracy-Privacy Trade-offs

DP protects sensitive information by adding noise to gradients or model updates [20, 21], but introduces an inherent trade-off: stronger privacy reduces model accuracy [22].

Secure multi-party computation [23] and homomorphic encryption [24] provide strong privacy but are impractical for large-scale FL. Structural and interaction-based methods exploit projections, uncertain step sizes, or structured noise [1, 25, 26, 27, 28] to maintain accuracy, but privacy guarantees are limited, tuning is delicate, and convergence may be hindered.

We propose a federated SGD algorithm that ensures agent privacy (Theorem 4), exact optimality, simplified step-size configuration (Theorems 3, 2), and accelerated convergence (Fig. 2(a)).

2.2. Byzantine Robustness with Privacy

Byzantine agents arbitrarily distort updates, threatening FL. Robust aggregation—distance-based [29], statistical [30], and proxy-based [31]—identifies malicious updates but assumes uniform step sizes, which can enable gradient inference to recover sensitive information (see Section 4.1).

DP-augmented methods [15, 16] attempt to combine privacy and Byzantine robustness, e.g., DP-BREM [15] adds noise to aggregated momentum but only achieves a d -accurate solution with slow convergence.

We introduce the PPCE filter, which guarantees Byzantine robustness and agent privacy while achieving a 0-accurate solution. By using private, variable step sizes, PPCE prevents gradient inference and preserves aggregation integrity.

3. Preliminaries and Problem Formulation

This section presents the preliminaries, system model, threat model, and design objectives for the proposed FL algorithm.

3.1. Preliminaries

3.1.1. Laplace Noise Mechanism

We incorporate the Laplace noise mechanism into the FL framework to ensure DP. Let $b > 0$ and denote by $\text{Lap}(0, b)$ the zero-mean Laplace distribution with variance $2b^2$ and probability density function $p(x) = \frac{1}{2b}e^{-|x|/b}$.

Let f be a function defined on a database $\mathcal{D}$. Its sensitivity Δf is defined as the maximum L_1 -norm difference between $f(\mathcal{D})$ and $f(\mathcal{D}')$ over all adjacent databases $(\mathcal{D}, \mathcal{D}')$: $\Delta f = \sup_{(\mathcal{D}, \mathcal{D}')} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$.

Theorem 1 (Laplace Mechanism [32]). The Laplace mechanism M_L is defined as $M_L(\mathcal{D}, f) = f(\mathcal{D}) + (z[1], \dots, z[d])^\top$, where $z[j] \stackrel{\text{i.i.d.}}{\sim} \text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right)$ for each $j \in [d] = \{1, \dots, d\}$. This mechanism guarantees $(\epsilon, 0)$ -differential privacy.

3.1.2. Exact Fault-Tolerance and $2f$ -Redundancy

The proposed algorithm builds on the principles of exact fault-tolerance and $2f$ -redundancy. Let $\mathcal{H} \subseteq [N]$ denote the set of non-faulty agents, while up to f agents may be Byzantine and behave arbitrarily, with unknown identities.

Definition 1 (Exact fault-tolerance [33]). Let $\mathcal{H} \subseteq [N]$ be the set of non-faulty agents with $|\mathcal{H}| \geq N - f$. A distributed optimization algorithm is said to achieve exact fault-tolerance if all non-faulty agents can compute $x_{\mathcal{H}}^* = \arg \min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} q^i(x)$.

As shown in [34], exact fault-tolerance can be achieved when the non-faulty agents satisfy the $2f$ -redundancy condition, which guarantees that the global minimizer remains unchanged despite the presence of up to f Byzantine agents.

Definition 2 ($2f$ -redundancy [34]). A set of non-faulty agents $\mathcal{H}$, with $|\mathcal{H}| \geq N - f$, satisfies the $2f$ -redundancy condition if for any subset $\mathcal{S} \subseteq \mathcal{H}$ with $|\mathcal{S}| \geq N - 2f$,

$$\arg \min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{S}} q^i(x) = \arg \min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} q^i(x).$$

3.2. System Model

We study Private Federated Optimization Problems (PFOPs) [2] in a federated learning system consisting of a central coordinator (CC) and N agents, where $[N] = \{1, \dots, N\}$. Each agent $i \in [N]$ holds a local dataset D^i of size $|D^i|$. At iteration k , a random mini-batch $D_k^i \subseteq D^i$ is sampled for local training. The PFOP framework is illustrated in Fig. 1.

Let $\mathcal{H} \subseteq [N]$ denote the set of non-faulty agents, while up to f agents may be Byzantine and behave arbitrarily with unknown identities. Let $\mathbf{w} \in \mathbb{R}^d$ denote the model parameters. A PFOP is characterized by $(\mathcal{W}, \mathcal{F}, Q, \mathcal{G}_{\mathcal{A}})$, where $\mathcal{W} \subseteq \mathbb{R}^d$ is the parameter domain, $\mathcal{F}$ is a class of strongly convex and differentiable functions, and $Q(\mathbf{w}) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} q^i(\mathbf{w})$ denotes the global objective. The communication topology is $\mathcal{G}_{\mathcal{A}} = (V, E_{\mathcal{A}})$ with node set $V = \{0\} \cup [N]$, where node 0 represents the CC. The adjacency matrices $\mathcal{A} = \{A_k\}$ may vary over time, and each agent communicates bidirectionally with the CC.

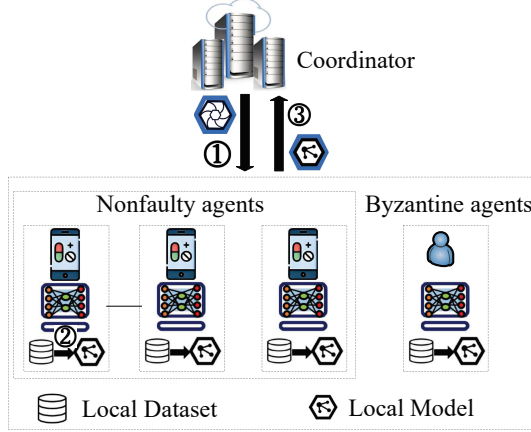


Figure 1: System model of a PFOP.

The system aims to solve

$$\mathbf{w}^* \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} q^i(\mathbf{w}), \quad (1)$$

where $\mathbf{w}$ is the global model shared by all agents and $q^i : \mathbb{R}^d \rightarrow \mathbb{R}$ is the local objective function private to agent $i \in \mathcal{H}$.

The $2f$ -redundancy property, introduced in Section 4, is crucial for our algorithm: any minimizer of the objectives of $N - 2f$ non-faulty agents is also a minimizer for all agents in $\mathcal{H}$ [34]. We assume that a finite optimal solution $\mathbf{w}^*$ exists.

Following the frameworks in [2, 35], we adopt the following standard assumptions.

Assumption 1 (Lipschitz smoothness). The local objective functions of nonfaulty agents have Lipschitz continuous gradients, i.e., there exists a positive constant $L < \infty$ such that $\forall i \in [N]$

$$\|\nabla q^i(\mathbf{w}) - \nabla q^i(\mathbf{w}')\| \leq L \|\mathbf{w} - \mathbf{w}'\|, \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d. \quad (2)$$

Assumption 2 (Strong convexity). The local objective function of each nonfaulty agent is strongly convex, i.e., there exists some $\mu > 0$ such that

$$(\mathbf{w} - \mathbf{w}')^\top (\nabla q^i(\mathbf{w}) - \nabla q^i(\mathbf{w}')) \geq \mu \|\mathbf{w} - \mathbf{w}'\|^2, \forall \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d, \quad (3)$$

where $(\cdot)^\top$ denotes the transpose, and $\mu \leq L$.

Assumption 3 (Gradient boundedness). The gradients of all the individual cost functions are bounded on $\mathbb{R}^d$. We denote $L' > 0$ as the bound such that for any $\mathbf{w} \in \mathbb{R}^d$ and any $i \in [N]$, $\|\nabla q^i(\mathbf{w})\| \leq L'$.

3.3. Threat Model

Following [35, 27], we assume that all agents can observe the messages sent to the coordinator. During a PFOP, the system may face three types of adversaries [35]: honest-but-curious entities, which follow the protocol but attempt to infer private information from observed updates; eavesdroppers, which intercept network communications; and Byzantine agents, which may behave arbitrarily or maliciously to disrupt the training process. Compared with honest-but-curious participants, eavesdroppers and Byzantine agents may obtain additional information or act unpredictably, posing greater threats to both privacy and model integrity.

3.4. Design Objectives

The proposed algorithm ensures privacy by providing ϵ -differential privacy for local gradients via randomized step-sizes and noise injection, even when all communications are observable. It simultaneously guarantees accuracy, achieving a 0-accurate solution by provably converging to the exact global optimum of (1) despite the stochastic perturbations required for privacy and the presence of Byzantine agents. In contrast, [35] provides no ϵ -differential privacy guarantee, while DP-BREM [15] only achieves a d -accurate solution for a small $d > 0$.

4. Proposed DPF-SGD

This section first examines gradient leakage in existing federated SGD schemes and then introduces the proposed algorithm for enhanced gradient privacy. For convenience, Table 1 summarizes the notation used throughout the paper.

Table 1: Notation and description

Notation	Description
N	the number of all agents in the system
$\mathcal{H} \subseteq [1, N]$	the set of all non-faulty agents
f	the maximum number of Byzantine agents
$i \in \mathcal{H}$	i -th participating agents in $\mathcal{H}$
$\mathcal{D}_i \subseteq \mathcal{D}$	the local dataset at agent i , $\mathcal{D} = \bigcup_{i=1}^N \mathcal{D}_i$
$\mathcal{B}_i \subseteq \mathcal{D}_i$	i -th batch data samples of dataset $\mathcal{D}_i$
k	k -th iteration in the system
γ_k	the weakening factor of noises at the k -th iteration
λ_k^i	the private step size of agent i at the k -th iteration
$\mathcal{T}$	the number of epoches at each iteration
$\mathbf{w}_{k,t}^i$	the k -th local model of agent i at the t -th epoch
$q^i(\cdot)$	the local loss function of agent i
$\nabla q^i(\mathbf{w}_{k,t}^i, \mathcal{B}_i)$	the gradient of the local model $\mathbf{w}_{k,t}^i$ on $\mathcal{B}_i$
$Lap(\cdot)$	Laplace noise mechanism
$\mathbf{n}_k^i$	the noise for $\mathbf{w}_{k,\mathcal{T}}^i$ at the k -th iteration
$\tilde{\mathbf{w}}_{k,\mathcal{T}}^i$	the perturbed parameter of local model $\mathbf{w}_{k,\mathcal{T}}^i$
$\bar{\mathbf{w}}_{k+1}$	the global model parameter at the k -th iteration

4.1. The Gradient Leakage of Federated SGD Schemes

The standard federated SGD update at iteration k is

$$\mathbf{w}_{k,\mathcal{T}}^i = \bar{\mathbf{w}}_k - \lambda_k \nabla q^i(\mathbf{w}_{k,0}^i), \quad (4)$$

which corresponds to $\mathcal{T} = 1$ [35], where $\mathbf{w}_{k,0}^i = \bar{\mathbf{w}}_k$ is broadcast by the coordinator, $\lambda_k > 0$ is the learning rate, and $\nabla q^i(\mathbf{w}_{k,0}^i)$ denotes the local gradient of agent i . Since all agents—including honest-but-curious and Byzantine ones—share the same learning rate and can observe the updates exchanged with the coordinator, this update rule inherently leaks gradient information. In particular, an adversarial agent with access to $\bar{\mathbf{w}}_k$ and λ_k can observe $\mathbf{w}_{k,1}^i$ and exactly

recover the local gradient as $\nabla q^i(\bar{\mathbf{w}}_k) = \frac{\bar{\mathbf{w}}_k - \mathbf{w}_{k,1}^i}{\lambda_k}$.

This reveals a fundamental privacy flaw of federated SGD: local gradients are deterministically reconstructible from communicated updates, a risk that is amplified in the presence of honest-but-curious or Byzantine agents. Motivated by this dual challenge, we propose DPF-SGD, which simultaneously provides rigorous gradient privacy protection and robustness against Byzantine agents, while retaining provable convergence guarantees.

4.2. Overview of the Proposed DPF-SGD

Motivated by the gradient leakage identified in Section 4.1, we propose DPF-SGD, a differentially private federated SGD algorithm that tolerates up to f Byzantine agents while achieving exact convergence. DPF-SGD operates in three stages: initialization, private local training, and robust aggregation. To ensure differential privacy, independent Laplace noise is injected at each iteration, and its long-term impact is controlled via a carefully designed interaction weakening factor γ_k , guaranteeing asymptotically vanishing noise effects. During local training, each non-faulty agent $i \in \mathcal{H}$ employs a private, uncoordinated step size λ_k^i to prevent gradient inference from transmitted updates. Finally, the coordinator applies a privacy-preserving comparative elimination (PPCE) rule to discard the f most deviant updates and aggregate the remaining ones. Together, these components provide rigorous gradient privacy and Byzantine robustness while preserving 0-accuracy of the global model (see Definition 2 in [2]).

4.3. Construction of DPF-SGD

DPF-SGD proceeds in three stages: system initialization, private local training, and robust selection and aggregation, as summarized in Algorithm 1.

4.3.1. System Initialization

The coordinator initializes the global model $\bar{\mathbf{w}}_0$ and specifies the number of local update rounds $\mathcal{T}$. Each agent independently selects a private step-size

Algorithm 1: DPF-SGD: Differentially Private Federated SGD

1: Input: $K, \mathcal{T}, \epsilon, \{\mathcal{A}_k\}, \{\gamma_k > 0\}, \bar{\mathbf{w}}_0, \{\lambda_k^i\}, b_k, f$

2: Output: $\bar{\mathbf{w}}_K$

3: for $k = 0, \dots, K - 1$ do

4: Agent-side local update

5: for $i \in \mathcal{H}$ do

6: Receive $\bar{\mathbf{w}}_k$ and γ_k from the CC.

7: Initialize $\mathbf{w}_{k,0}^i = \bar{\mathbf{w}}_k$ and perform local SGD:

$$\mathbf{w}_{k,t+1}^i = \mathbf{w}_{k,t}^i - \lambda_k^i \nabla q^i(\mathbf{w}_{k,t}^i), \quad t = 0, \dots, \mathcal{T} - 1. \quad (5)$$

8: Add Laplace noise:

$$\tilde{\mathbf{w}}_{k,\mathcal{T}}^i = \mathbf{w}_{k,\mathcal{T}}^i + \gamma_k \mathbf{n}_k^i. \quad (6)$$

9: Send $\tilde{\mathbf{w}}_{k,\mathcal{T}}^i$ to the CC.

10: end for

11: CC-side aggregation

12: Collect $\{\tilde{\mathbf{w}}_{k,\mathcal{T}}^i\}_{i=1}^N$ according to $\mathcal{A}_k$.

13: Sort the distances in ascending order:

$$\|\tilde{\mathbf{w}}_{k,\mathcal{T}}^{i_1} - \bar{\mathbf{w}}_k\| \leq \dots \leq \|\tilde{\mathbf{w}}_{k,\mathcal{T}}^{i_N} - \bar{\mathbf{w}}_k\|.$$

14: Discard the f largest distances and define $\mathcal{F}_k = \{i_1, \dots, i_{N-f}\}$.

15: Update the global model:

$$\bar{\mathbf{w}}_{k+1} = \frac{1}{N-f} \sum_{i \in \mathcal{F}_k} \tilde{\mathbf{w}}_{k,\mathcal{T}}^i. \quad (7)$$

16: Broadcast $\bar{\mathbf{w}}_{k+1}$ to all agents.

17: end for

18: Return: $\bar{\mathbf{w}}_K$

sequence $\{\lambda_k^i\}$ for local training (See Theorem 2 and 3 for the choice of λ_k^i). The step sizes of non-faulty agents satisfy the consensus conditions in Theorem 2 or Theorem 3, while the noise weakening sequence $\{\gamma_k\}$ is chosen such that $\lim_{k \rightarrow \infty} \gamma_k = 0$. Under these conditions, the global aggregation sequence $\{\bar{w}_k\}$ converges to the optimal solution of (1).

4.3.2. Private Local Model Training

At communication round k , each agent receives the global model $\bar{w}_k$ and performs $\mathcal{T}$ steps of local SGD using its private step size λ_k^i over mini-batches $\mathcal{B}_i$, updating according to $\mathbf{w}_{k,t+1}^i = \mathbf{w}_{k,t}^i - \lambda_k^i \nabla q^i(\mathbf{w}_{k,t}^i, \mathcal{B}_i)$, $t = 0, \dots, \mathcal{T} - 1$, $\mathbf{w}_{k,0}^i = \bar{w}_k$. Since λ_k^i is private, the local gradient cannot be inferred from consecutive iterates, providing intrinsic protection against gradient reconstruction.

To further ensure differential privacy, each agent injects Laplace noise $\mathbf{n}_k^i \sim \text{Lap}(0, b_k)^d$ (See Theorem 4 for the choice of b_k), yielding the transmitted update $\tilde{\mathbf{w}}_{k,\mathcal{T}}^i = \mathbf{w}_{k,\mathcal{T}}^i + \gamma_k \mathbf{n}_k^i$, so that the overall mechanism satisfies ϵ -differential privacy.

4.3.3. Selection and Aggregation of Effective Updates

Upon receiving the perturbed updates, CC computes their deviations from the current global model and applies the PPCE rule. The f most deviant updates are discarded, and the remaining $N - f$ updates are averaged to form the next global iterate. This robust aggregation mechanism mitigates the impact of Byzantine agents while preserving gradient privacy, ensuring accurate and stable global convergence.

5. Convergence Analysis

We analyze the convergence of Algorithm DPF-SGD for solving (1). Our analysis focuses on the evolution of the global iterate $\bar{w}_k$ in the presence of Byzantine agents and persistent DP noise.

Noise model.. Throughout, we assume the following standard condition on the injected noise [36].

Assumption 4. For all $i \in [1, N]$ and $k \geq 0$, conditional on $\tilde{\mathbf{w}}_k$, the noise satisfies $\mathbb{E}[\mathbf{n}_k^i] = 0$ and $\mathbb{E}[\|\mathbf{n}_k^i\|^2] = \sigma^2$. Moreover,

$$\sum_{k=0}^{\infty} (\gamma_k)^2 \max_{i \in [1, N]} \sigma^2 < \infty, \quad (8)$$

where $\{\gamma_k\}$ is the weakening factor sequence. The initial iterates satisfy $\mathbb{E}[\|\tilde{\mathbf{w}}_{0, \mathcal{T}}^i\|^2] < \infty$.

We will use the following elementary fact.

Lemma 1. If $\bar{\mathbf{w}}$ is a zero-mean random vector with $\mathbb{E}[\|\bar{\mathbf{w}}\|^2] = \sigma^2$, then $\mathbb{E}[\|\bar{\mathbf{w}}\|] \leq \sigma$.

Proof. The result follows from Jensen's inequality because $\phi(z) = z^2$ is convex. $\square$

The convergence analysis of Algorithm 1 is based on the following lemma.

Lemma 2. Let $\{x_k\}$ be a nonnegative sequence, and let $\{\alpha_k\}$ and $\{\beta_k\}$ be positive sequences such that $0 < \alpha_k < 1$, $\sum_{k=1}^{\infty} \alpha_k = \infty$, and $\lim_{k \rightarrow \infty} \beta_k = 0$. If there exists $\hat{K} > 0$ such that $x_{k+1} \leq (1 - \alpha_k)x_k + \beta_k$ for all $k \geq \hat{K}$, then $\lim_{k \rightarrow \infty} x_k = 0$.

Proof. Without loss of generality, assume $\hat{K} = 1$. Let $\bar{x} := \limsup_{k \rightarrow \infty} x_k \geq 0$.

We show that $\bar{x} = 0$.

Since $\beta_k \rightarrow 0$, for any $\varepsilon > 0$ there exists K_ε such that

$$x_{k+1} \leq (1 - \alpha_k)x_k + \varepsilon, \quad \forall k \geq K_\varepsilon. \quad (9)$$

Fix $m \geq K_\varepsilon$ and iterate (9) to obtain, for any $t \geq 1$,

$$\begin{aligned} x_{m+t} &\leq x_m \prod_{j=0}^{t-1} (1 - \alpha_{m+j}) + \varepsilon \sum_{s=0}^{t-1} \prod_{j=s+1}^{t-1} (1 - \alpha_{m+j}) \\ &\leq x_m P(t) + \varepsilon t, \end{aligned} \quad (10)$$

where $P(t) := \prod_{j=0}^{t-1} (1 - \alpha_{m+j})$.

Since $\sum_{k=1}^{\infty} \alpha_k = \infty$, we have $P(t) \rightarrow 0$ as $t \rightarrow \infty$. Note that the inequality $\ln(1 - x) \leq -x$ (for $0 < x < 1$). Indeed, $0 < 1 - \alpha_{m+j} < 1$ and $\ln(1 - \alpha_{m+j}) \leq -\alpha_{m+j}$. Then, $\lim_{t \rightarrow \infty} \ln P(t) = \lim_{t \rightarrow \infty} \sum_{j=0}^{t-1} \ln(1 - \alpha_{m+j}) \leq -\lim_{t \rightarrow \infty} \sum_{j=0}^{t-1} \alpha_{m+j} = -\sum_{j=0}^{\infty} \alpha_{m+j} \rightarrow -\infty$. Thus, $\lim_{t \rightarrow \infty} P(t) = \lim_{t \rightarrow \infty} e^{\ln P(t)} = e^{-\infty} = 0$.

Suppose by contradiction that $\bar{x} > 0$. Choose a subsequence $\{k_i\}$ such that $x_{k_i} \rightarrow \bar{x}$ and $k_i \geq K_\varepsilon$. Fix N large enough so that $P_i(N) \leq 1/2$ for all sufficiently large i . Next, choose ε so small that $\varepsilon N < \bar{x}/4$. Then, from (10), we have that $x_{k_i+N} \leq P_i(N)x_{k_i} + \varepsilon N \leq \frac{1}{2}\bar{x} + \frac{\bar{x}}{4} = \frac{3}{4}\bar{x}$. Hence, $x_{k_i+N} \leq \frac{3}{4}\bar{x}$ for all large i , which contradicts the fact that $\bar{x}$ is the limit superior (because $k_i + N \rightarrow \infty$ as $i \rightarrow \infty$). Therefore $\bar{x} = 0$. Since $x_k \geq 0$, we conclude $\lim_{k \rightarrow \infty} x_k = 0$. $\square$

5.0.1. Single local step ($\mathcal{T} = 1$)

For $\mathcal{T} = 1$, combining (5) and (6) yields

$$\tilde{\mathbf{w}}_{k,1}^i = \bar{\mathbf{w}}_k - \lambda_k^i \nabla q^i(\bar{\mathbf{w}}_k) + \gamma_k \mathbf{n}_k^i. \quad (11)$$

Let $\mathcal{B}$ and $\mathcal{H}$ denote the Byzantine and non-faulty agents, respectively, with $|\mathcal{B}| \leq f < N/4$. At each iteration, the PPCE filter discards the f updates farthest from the aggregate, ensuring that any retained Byzantine update is dominated by a non-faulty one.

Theorem 2. Let $\{\bar{\mathbf{w}}_k\}$ be generated by DPF-SGD with $\mathcal{T} = 1$ and $\gamma_k \in (0, 1)$. Assume

$$\frac{f}{N-f} < \min \left\{ \frac{1}{3}, \frac{\mu}{6L} \right\}. \quad (12)$$

If the stepsizes satisfy

$$\lambda_k^i \in [\underline{\lambda}_k, \hat{\lambda}_k] \subseteq \left[\frac{h\mu}{10L^2}, \frac{\mu}{10L^2} \right], \quad (13)$$

with $\underline{\lambda}_k \geq h\hat{\lambda}_k$ for some $h \in (2/3, 1)$, then

$$\begin{aligned} \mathbb{E}[\|\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^*\|^2] &\leq \left(1 - \frac{\mu}{6}\hat{\lambda}_k\right) \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] \\ &\quad + (8\bar{C}_0 + 5\gamma_k\sigma)\sigma\gamma_k. \end{aligned} \quad (14)$$

The complete proof of Theorem 2 is provided in Appendix A.

Corollary 1. Under the assumptions of Theorem 2, if $\sum_{k=1}^{\infty} \hat{\lambda}_k = \infty$ and $\lim_{k \rightarrow \infty} \gamma_k = 0$, then

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] = 0, \quad (15)$$

and in particular $\bar{\mathbf{w}}_k \rightarrow \mathbf{w}_{\mathcal{H}}^*$ in mean square.

Proof. Define $\alpha_k := \frac{\mu}{6} \hat{\lambda}_k$, $\beta_k := (8\bar{C}_0 + 5\gamma_k\sigma)\sigma\gamma_k$, and $x_k := \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2]$. By (14), the sequence $\{x_k\}$ satisfies $x_{k+1} \leq (1 - \alpha_k)x_k + \beta_k$.

Since $\gamma_k \rightarrow 0$, we have $\beta_k \rightarrow 0$. Moreover, $\sum_{k=1}^{\infty} \hat{\lambda}_k = \infty$ implies $\sum_{k=1}^{\infty} \alpha_k = \infty$. Applying Lemma 2 together with (14) in Theorem 2, we conclude that $\lim_{k \rightarrow \infty} x_k = \lim_{k \rightarrow \infty} \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] = 0$. Therefore, $\bar{\mathbf{w}}_k \rightarrow \mathbf{w}_{\mathcal{H}}^*$ in mean square. $\square$

5.0.2. Multiple local steps ($\mathcal{T} > 1$)

We now extend the result to $\mathcal{T} > 1$, where each agent performs $\mathcal{T}$ local updates before aggregation. For $i \in \mathcal{H}$,

$$\bar{\mathbf{w}}_{k,\mathcal{T}}^i = \bar{\mathbf{w}}_k - \lambda_k^i \sum_{t=0}^{\mathcal{T}-1} \nabla q^i(\bar{\mathbf{w}}_{k,t}^i) + \gamma_k \mathbf{n}_k^i. \quad (16)$$

Theorem 3. Let $\{\bar{\mathbf{w}}_k\}$ be generated by DPF-SGD with $\mathcal{T} > 1$ and $\gamma_k \in (0, 1)$.

Assume

$$\frac{f}{N-f} < \min \left\{ \frac{1}{3}, \frac{\mu}{12L} \right\}. \quad (17)$$

If the stepsizes satisfy

$$\lambda_k^i \in [\underline{\lambda}_k, \hat{\lambda}_k], \quad \hat{\lambda}_k = \min \left\{ \frac{\mu}{(40 + 16L)\mathcal{T}^2}, \frac{\mu}{16L^2\mathcal{T}} \right\}, \quad (18)$$

with $\underline{\lambda}_k \geq h\hat{\lambda}_k$ for some $h \in (2/3, 1)$. Then

$$\begin{aligned} \mathbb{E}[\|\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^*\|^2] &\leq \left(1 - \frac{\mu\mathcal{T}}{6} \hat{\lambda}_k\right) \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] \\ &\quad + O(\gamma_k\sigma + \gamma_k^2\sigma^2). \end{aligned} \quad (19)$$

The complete proof of Theorem 3 is provided in Appendix B.

Corollary 2. Under the assumptions of Theorem 3, if $\sum_{k=1}^{\infty} \hat{\lambda}_k = \infty$ and $\lim_{k \rightarrow \infty} \gamma_k = 0$, then

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] = 0. \quad (20)$$

In particular, the sequence $\{\bar{\mathbf{w}}_k\}$ converges to $\mathbf{w}_{\mathcal{H}}^*$ in mean square.

Proof. Let $\alpha_k := \frac{\mu\mathcal{T}}{6}\hat{\lambda}_k$, $\beta_k := O(\gamma_k\sigma + \gamma_k^2\sigma^2)$, and $x_k := \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2]$. By (19), the sequence $\{x_k\}$ satisfies $x_{k+1} \leq (1 - \alpha_k)x_k + \beta_k$.

Since $\gamma_k \rightarrow 0$, we have $\beta_k \rightarrow 0$. Moreover, $\sum_{k=1}^{\infty} \hat{\lambda}_k = \infty$ implies $\sum_{k=1}^{\infty} \alpha_k = \infty$. Applying Lemma 2 together with (19) in Theorem 3, we yield that $\lim_{k \rightarrow \infty} x_k = \lim_{k \rightarrow \infty} \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] = 0$. $\square$

6. Privacy Analysis

6.1. ϵ -Differential Privacy

We define two PFOPs $\mathcal{P} = (\mathcal{W}, \mathcal{F}, Q, \mathcal{G}_{\mathcal{A}})$ and $\mathcal{P}' = (\mathcal{W}', \mathcal{F}', Q', \mathcal{G}'_{\mathcal{A}})$ to be adjacent if (1) $\mathcal{W} = \mathcal{W}'$, $\mathcal{F} = \mathcal{F}'$, and $\mathcal{G}_{\mathcal{A}} = \mathcal{G}'_{\mathcal{A}}$; and (2) there exists a client $i \in [N]$ such that $q^i \neq q'^i$, while $q^j = q'^j$ for all $j \neq i$.

We analyze the differential privacy guarantees of Algorithm 1. Define $\tilde{\mathbf{W}}_{k,\mathcal{T}} \triangleq [\tilde{\mathbf{w}}_{k,\mathcal{T}}^1, \dots, \tilde{\mathbf{w}}_{k,\mathcal{T}}^N]$, $\mathbf{W}_{k,t+1} \triangleq [\mathbf{w}_{k,t+1}^1, \dots, \mathbf{w}_{k,t+1}^N]$, $t = 0, \dots, \mathcal{T}-1$. An execution of the algorithm is a (possibly infinite) sequence

$$\alpha = \{\bar{\mathbf{w}}_0, \langle \{\mathbf{W}_{0,t+1}\}_{t=0}^{\mathcal{T}-1}, \tilde{\mathbf{W}}_{0,\mathcal{T}}, \bar{\mathbf{w}}_1 \rangle, \langle \{\mathbf{W}_{1,t+1}\}_{t=0}^{\mathcal{T}-1}, \tilde{\mathbf{W}}_{1,\mathcal{T}}, \bar{\mathbf{w}}_2 \rangle, \dots\}.$$

The observable output of the execution is the sequence of exchanged messages, defined by the observation mapping

$$\mathcal{R}(\alpha) \triangleq \{\bar{\mathbf{w}}_0, \langle \tilde{\mathbf{w}}_{0,\mathcal{T}}, \bar{\mathbf{w}}_1 \rangle, \langle \tilde{\mathbf{w}}_{1,\mathcal{T}}, \bar{\mathbf{w}}_2 \rangle, \dots\}.$$

Here, $\mathbf{I}_k = (\mathbf{W}_{k,1}, \dots, \mathbf{W}_{k,\mathcal{T}})$ denotes the internal agent models (i.e., the noiseless agent-side updates), $\tilde{\mathbf{W}}_{k,\mathcal{T}}$ the perturbed local updates and $\bar{\mathbf{w}}_k$ the updated global models, generated according to (5), (6), and (7).

Let $\mathbf{I} = \{\mathbf{I}_k\}_{k \geq 0}$ denote the full internal state sequence, and let $\mathcal{R}^{-1}(\mathcal{P}, Y, \bar{\mathbf{w}}_0)$ be the set of internal trajectories consistent with problem instance $\mathcal{P}$, observation Y , and initial state $\bar{\mathbf{w}}_0$. Any $\mathbf{I}(\rho) \in \mathcal{R}^{-1}(\mathcal{P}, Y, \bar{\mathbf{w}}_0)$ therefore corresponds to a reconstructed internal trajectory consistent with Y .

Let $\mathbb{O} = \{\rho \mid \rho = \{\bar{\mathbf{w}}_0, \langle \tilde{W}_0, \bar{\mathbf{w}}_1 \rangle, \langle \tilde{W}_1, \bar{\mathbf{w}}_2 \rangle, \dots\}\}$. Following [2], an algorithm for solving $\mathcal{P}$ is said to be ϵ -differentially private if, for any adjacent $\mathcal{P}$ and $\mathcal{P}'$, any measurable observation set $Y \subseteq \mathbb{O}$, and any initial state $\bar{\mathbf{w}}_0$, the following inequality holds:

$$\mathbb{P}[\mathcal{R}^{-1}(\mathcal{P}, Y, \bar{\mathbf{w}}_0)] \leq e^\epsilon \mathbb{P}[\mathcal{R}^{-1}(\mathcal{P}', Y, \bar{\mathbf{w}}_0)], \quad (21)$$

where the probability is taken over the internal randomness of the algorithm.

6.2. Privacy Analysis

The sensitivity at iteration k captures the discrepancy between internal states reconstructed from the same observation sequence under two adjacent problem instances. Following [2], we formalize this notion as follows.

Definition 3. Consider Algorithm 1 executed on two adjacent PFOPs $\mathcal{P}$ and $\mathcal{P}'$ with a fixed initial state $\bar{\mathbf{w}}_0$. For any observation sequence $Y \in \mathbb{O}$, let $\mathbf{I}(\rho) = \{\mathbf{I}_k(Y)\}_{k \geq 0} \in \mathcal{R}^{-1}(\mathcal{P}, Y, \bar{\mathbf{w}}_0)$ and $\mathbf{I}'(\rho) = \{\mathbf{I}'_k(Y)\}_{k \geq 0} \in \mathcal{R}^{-1}(\mathcal{P}', Y, \bar{\mathbf{w}}_0)$ be the corresponding reconstructed internal trajectories. The ℓ_1 -sensitivity at iteration k is defined as

$$\Delta_k = \sup_{\rho \in \mathbb{O}} \sup_{\substack{\mathbf{I}(\rho) \in \mathcal{R}^{-1}(\mathcal{P}, Y, \bar{\mathbf{w}}_0) \\ \mathbf{I}'(\rho) \in \mathcal{R}^{-1}(\mathcal{P}', Y, \bar{\mathbf{w}}_0)}} \|\mathbf{I}_k(\rho) - \mathbf{I}'_k(\rho)\|_1,$$

where the norm used is L_1 -norm.

Based on the above definition, we obtain the following result.

Lemma 3. At each round $k \in \{0, 1, \dots, K-1\}$, if each agent in (6) adds a noise vector $\mathbf{n}_k^i \sim \text{Lap}(0|b_k)^d$ such that $\sum_{k=1}^{\infty} \frac{\Delta^k}{b_k} \leq \epsilon$, then the Algorithm 1 is ϵ -differentially private.

Lemma 4. Under Assumption 3, the sensitivity of the DPF-SGD algorithm at iteration k satisfies $\Delta_k \leq 2\sqrt{d}L'\lambda_{\max}$, where $\lambda_{\max} = \max_{k=0}^{K-1}\{\hat{\lambda}_k\}$ and K denotes the total number of iterations.

The proofs of Lemmas 3 and 4 follow those of Lemmas 2 and 4 in [2], respectively, and are therefore omitted.

Let ϵ_k denote the privacy budget at iteration $k \in \{0, \dots, K-1\}$. To stabilize training, we fix the effective noise scale across iterations by enforcing a uniform allocation $\epsilon_k = \epsilon/K$. From Lemmas 3 and 4, it follows that Algorithm 1 guarantees ϵ -differential privacy.

Theorem 4. Under Assumption 3, Algorithm 1 satisfies ϵ -differential privacy for any $K \in \mathbb{N}$ rounds if the injected noise $\mathbf{n}_k^i \sim \text{Lap}(0, b_k)^d$ satisfies

$$b_k \geq \frac{2\sqrt{d}L'\mathcal{T}\hat{\lambda}_k K}{\gamma_k \epsilon}, \quad k = 0, 1, \dots, K-1.$$

Proof. We analyze the privacy loss per round and apply sequential composition.

Sensitivity. At round k , the released update is $\tilde{\mathbf{w}}_{k,\mathcal{T}}^i = \bar{\mathbf{w}}_k - \lambda_k^i \sum_{t=0}^{\mathcal{T}-1} \nabla q^i(\bar{\mathbf{w}}_{k,t}^i) + \gamma_k \mathbf{n}_k^i$. Since γ_k scales the injected noise, the effective sensitivity of the noisy update equals $\Delta_k^{\text{noisy}} = \frac{\Delta_k}{\gamma_k}$.

By Lemma 4, the unscaled sensitivity satisfies $\Delta_k \leq 2\sqrt{d}L'\mathcal{T}\hat{\lambda}_k$, yielding

$$\Delta_k^{\text{noisy}} \leq \frac{2\sqrt{d}L'\mathcal{T}\hat{\lambda}_k}{\gamma_k}.$$

Noise calibration. To ensure ϵ_k -DP at round k , the Laplace mechanism requires $b_k \geq \Delta_k^{\text{noisy}}/\epsilon_k$. We allocate the privacy budget uniformly, setting $\epsilon_k = \epsilon/K$, which gives

$$b_k = \frac{\Delta_k^{\text{noisy}}}{\epsilon_k} = \frac{\Delta_k K}{\gamma_k \epsilon}.$$

Composition. With this choice, $\sum_{k=1}^K \frac{\Delta_k^{\text{noisy}}}{b_k} = \sum_{k=1}^K \epsilon_k = \epsilon$. Thus, by Lemma 3 and the sequential composition theorem, Algorithm 1 satisfies ϵ -differential privacy. $\square$

7. Numerical Results

We evaluate DPF-SGD on synthetic and MNIST datasets and demonstrate that it outperforms shared step-size FL baselines while mitigating the attacks in Section 4.1. All DPF-SGD experiments adopt the noise decay sequence $\gamma_k = \frac{1}{1+0.01 \cdot k}$ satisfying $\lim_{k \rightarrow \infty} \gamma_k = 0$

Datasets and model. We use (i) a synthetic dataset [? 35] with $N = 64$ clients, each holding 100 samples, loss $q^i(\mathbf{w}) = \|\mathbf{X}^i - M^i \mathbf{w}\|^2 + \varsigma \|\mathbf{w}\|^2$ ($\mathbf{X}^i = M^i \mathbf{w}^* + \mathbf{z}^i$, $\mathbf{z}^i \sim \mathcal{N}(0, I_d)$, $M^i \in \mathbb{R}^{3 \times 2}$, and $\varsigma = 0.01$); and (ii) MNIST. For MNIST, DPF-SGD is implemented with one CC and $N = 64$ clients using the CNN architecture in [28].

Setup. The default parameters are $(K, \mu, L, h, L') = (100, 0.6, 1, 0.7, 1)$, with $\mathcal{T} \in \{1, 2\}$ and $f \in \{0, 1, 3\}$. Each agent performs DPF-SGD with $\lambda_k^i = b_u / (1.0 + 0.1 \cdot (i+1) \cdot k)$ and its $\hat{\lambda}_k = b_u / (1.0 + 0.1 \cdot k)$, where $b_u = \frac{\mu}{10 \cdot L^2}$ for $\mathcal{T} = 1$ or $b_u = \min(\frac{\mu}{(40+16 \cdot L)T^2}, \frac{\mu}{16L^2T})$ for $\mathcal{T} = 2$. The parameters satisfy $\frac{f}{N-f} < \frac{\mu}{12L}$, and $\sum_{k=K}^{\infty} \hat{\lambda}_k = \infty$ for both $\mathcal{T} = 1$ and $\mathcal{T} = 2$, since $\sum_{k=K}^{\infty} \frac{1}{1+0.1 \cdot k} = \infty$ by comparison to the divergent p -series $\sum_{k=K}^{\infty} \frac{1}{k^p}$ for $p \leq 1$. Gradients are clipped at $C' = 1$ and perturbed by Laplace noise [37].

Baselines include Central DP [38] and Local DP [39], all with fixed learning rate $\lambda = 0.001$ and Laplace noise $\sigma = 0.01$.

7.1. Convergence on Synthetic Data

Fig. 2(a) shows the convergence behavior on the synthetic dataset. The vertical axis reports the squared parameter error $\|\bar{w}_k - w^*\|^2$ (log scale), and the horizontal axis denotes the number of training steps. Results are averaged over 5 runs and sampled every $5k$ rounds.

All methods exhibit a rapid decrease in error during the early stage of training, dropping by several orders of magnitude within the first 10^4 iterations. After this phase, their convergence trajectories diverge. Central DP converges the slowest and stabilizes around 10^{-2} due to the persistent variance introduced

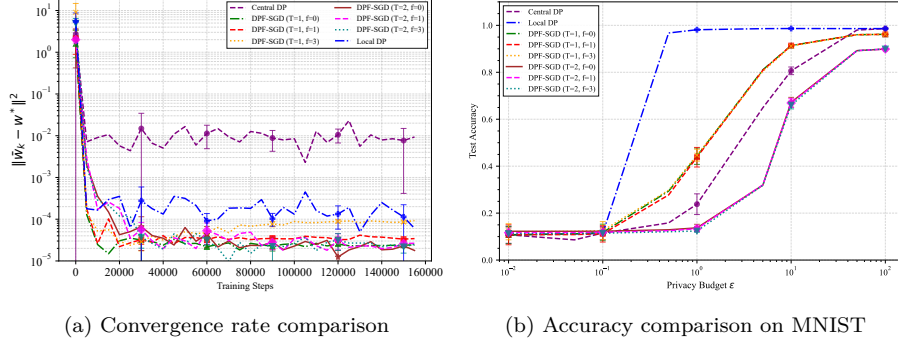


Figure 2: Experimental results across algorithms.

by DP noise. Local DP achieves a lower steady-state error (around 10^{-4}) but shows noticeable fluctuations.

In contrast, DPF-SGD variants consistently achieve the lowest optimization error across the entire training process. In particular, the methods with $T = 1$ quickly converge to an error level around 10^{-5} and remain stable throughout the training horizon. Compared with the DP baselines, this represents roughly one to two orders of magnitude reduction in the final optimization error relative to the optimal solution. Increasing the number of local updates ($T = 2$) yields similar performance with slightly higher variance, while different filtering parameters f lead to nearly identical trajectories.

Overall, these results demonstrate that DPF-SGD achieves faster convergence and significantly lower steady-state optimization error compared with existing DP baselines.

7.2. Accuracy on MNIST

Fig. 2(b) shows test accuracy on MNIST versus the privacy budget ϵ (log scale). As expected, all methods exhibit a clear privacy–utility trade-off: for small ϵ ($\leq 10^{-1}$), strong noise severely perturbs gradients, yielding near-random performance (≈ 0.1 – 0.15), while accuracy improves steadily as ϵ increases and noise diminishes.

Among baselines, Local DP improves rapidly and reaches near-optimal ac-

curacy around $\epsilon \approx 1$, whereas Central DP recovers more slowly, remaining low at $\epsilon = 1$ (≈ 0.25) and requiring $\epsilon \geq 10$ to achieve competitive performance. This gap reflects the heavier noise typically imposed by centralized DP at small privacy budgets.

DPF-SGD consistently outperforms Central DP in the moderate-privacy regime. In particular, at $\epsilon \approx 1$, DPF-SGD ($T = 1$) attains 0.42–0.46 accuracy, substantially narrowing the gap to Local DP. As ϵ increases, it rapidly approaches the best-performing baselines, reaching ≈ 0.9 at $\epsilon = 10$.

Increasing local updates ($T = 2$) slightly degrades performance under strong privacy, likely due to noise accumulation, though this effect diminishes as ϵ grows. The influence of the filtering parameter f is negligible, indicating robustness to its choice.

Overall, these results indicate that DPF-SGD achieves a favorable balance between privacy and utility, particularly in the practically relevant moderate-privacy regime ($\epsilon \in [1, 10]$)

8. Conclusion

While DP has become fundamental for safeguarding sensitive data, its application in distributed optimization frequently leads to trade-offs in accuracy and convergence speed due to the necessity of iterative noise injection. This paper presents a novel DPF-SGD variant that utilizes a private step size, achieving ϵ -DP while preserving high levels of accuracy. Remarkably, even in the presence of persistent privacy noise and Byzantine attacks, the proposed algorithm ensures that all fault-free agents converge rapidly to the true optimum. Theoretical insights and experimental evaluations substantiate that this approach surpasses existing DP methodologies in terms of convergence speed while maintaining high accuracy for $\epsilon > 1$. Looking ahead, future research will aim to enhance accuracy in high-privacy contexts ($\epsilon \leq 1$) and adapt the framework for more complex models.

Acknowledgment

The foundation of this work was completed during the authors' Master's studies at Johns Hopkins University. We thank our advisor, Associate Professor Yinzhi Cao, for his invaluable mentorship.

References

- [1] F. Yan, et al., Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties, *IEEE Trans. Knowl. Data Eng.* 25 (11) (2013) 2483–2493.
- [2] Z. Huang, S. Mitra, N. Vaidya, Differentially private distributed optimization, in: *ICDCN*, 2015, pp. 1–10.
- [3] M. G. Rabbat, R. D. Nowak, Distributed optimization in sensor networks, in: *IPSN 2004*, pp. 20–27.
- [4] A. Zia, M. Haleem, Bridging research gaps in industry 5.0: Synergizing federated learning, collaborative robotics, and autonomous systems for enhanced operational efficiency and sustainability, *IEEE Access* 13 (2025) 40456–40479.
- [5] S. Li, M. Wang, Y. Cai, Y. Sun, Dynamic power allocation in noma-based federated learning system, in: *WCSP 2023*, IEEE, 2023, pp. 183–188.
- [6] M. Ribero, H. Vikalo, G. de Veciana, Federated learning at scale: Addressing client intermittency and resource constraints, *IEEE J. Sel. Top. Signal Process.* 19 (1) (2025) 60–73.
- [7] R. Guerraoui, N. Gupta, R. Pinot, S. Rouault, J. Stephan, Combining differential privacy and byzantine resilience in distributed SGD, *arXiv preprint arXiv:2110.03991* (2021) (2021).
- [8] R. Guerraoui, N. Gupta, R. Pinot, S. Rouault, J. Stephan, Differential privacy and byzantine resilience in SGD: do they add up?, in: *PODC '21*, ACM, 2021, pp. 391–401.

- [9] Y. I. Alzoubi, A. Mishra, Differential privacy and artificial intelligence: potentials, challenges, and future avenues, *Journal of Artificial Intelligence and Privacy* (2025). doi:10.1186/s13635-025-00203-9.
URL <http://link.springer.com/article/10.1186/s13635-025-00203-9>
- [10] R. Xin, S. Kar, U. A. Khan, Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence, *IEEE Signal Process. Mag.* 37 (3) (2020) 102–113.
- [11] L. Bottou, F. E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.* 60 (2) (2018) 223–311.
- [12] H. Robbins, S. Monro, A stochastic approximation method, *SIAM Rev.* 22 (3) (1951) 400–407.
- [13] J. N. Tsitsiklis, D. P. Bertsekas, M. Athans, Distributed asynchronous deterministic and stochastic gradient optimization algorithms, *IEEE Trans. Autom. Control* 31 (9) (1984) 803–812.
- [14] H. Yin, A. Mallya, A. Vahdat, J. M. Álvarez, J. Kautz, P. Molchanov, See through gradients: Image batch recovery via gradinversion, in: *CVPR 2021*, IEEE, 2021, pp. 16337–16346.
- [15] X. Gu, M. Li, L. Xiong, DP-BREM: differentially-private and byzantine-robust federated learning with client momentum, in: *34th USENIX Security Symposium*, USENIX Association, 2025, pp. 3065–3082.
- [16] H. Zhang, J. Wu, Q. Pan, Byzantine-resilient differentially private federated learning: A dual-phase group-wise aggregation approach, in: *WASA 2025*, Vol. 15688 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 194–204.
- [17] C. Dwork, et al., Calibrating noise to sensitivity in private data analysis, in: *TCC2006*, Springer, 2006, pp. 265–284.

- [18] Censinet, Top frameworks for gdpr data de-identification, Censinet Perspectives (2025).
URL <https://www.censinet.com/perspectives/top-frameworks-gdpr-data-de-identification>
- [19] S. Overflow, Allocating privacy budgets and choosing epsilon in production, online article (2025).
URL <https://www.systemoverflow.com/learn/ml-privacy-fairness/differential-privacy/allocating-privacy-budgets-and-choosing-epsilon-in-production>
- [20] T. Fukami, et al., Dp-norm: Differential privacy primal-dual algorithm for decentralized federated learning, *IEEE Trans. Inf. Forensics Secur.* 19 (2024) 5783–5797.
- [21] Z. Zhang, et al., Communication and energy efficient wireless federated learning with intrinsic privacy, *IEEE Trans. Dependable Secur. Comput.* 21 (4) (2024) 4035–4047.
- [22] L. Zhu, Z. Liu, S. Han, Deep leakage from gradients, in: *NeurIPS 2019*, 2019, pp. 14747–14756.
- [23] B. Liu, E. B. Blancaflor, Research on federal learning privacy protection based on secure multi-party computing, in: *Proc. CSAIDE'24*, ACM, 2024, pp. 142–147.
- [24] Q. X. et al., Efficiency optimization techniques in privacy-preserving federated learning with homomorphic encryption: A brief survey, *IEEE Internet Things J.* 11 (14) (2024) 24569–24580.
- [25] S. Gade, N. H. Vaidya, Private optimization on networks, in: *Proc. Annu. Amer. Control Conf. (ACC)*, 2018, pp. 1402–1409.
- [26] Y. Lou, L. Yu, S. Wang, P. Yi, Privacy preservation in distributed subgradient optimization algorithms, *IEEE Trans. Cybern.* 48 (7) (2018) 2154–2165.

- [27] Y. Wang, H. V. Poor, Decentralized stochastic optimization with inherent privacy protection, *IEEE Trans. Autom. Control.* 68 (4) (2023) 2293–2308.
- [28] Y. Wang, A. Nedic, Decentralized gradient methods with time-varying uncoordinated stepsizes: Convergence analysis and privacy design, *IEEE Trans. Autom. Control.* 69 (8) (2024) 5352–5367.
- [29] P. Blanchard, et al., Machine learning with adversaries: Byzantine tolerant gradient descent, in: *NeurIPS*, 2017, pp. 119–129.
- [30] K. Pillutla, S. M. Kakade, Z. Harchaoui, Robust aggregation for federated learning, *IEEE Trans. Signal Process.* 70 (2022) 1142–1154.
- [31] X. Cao, et al., Fltrust: Byzantine-robust federated learning via trust bootstrapping, in: *NDSS2021*, 2021.
- [32] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, *Found. Trends Theor. Comput. Sci.* (2014) 211–407.
- [33] L. Su, N. H. Vaidya, Fault-tolerant multi-agent optimization: Optimal iterative distributed algorithms, in: G. Giakkoupis (Ed.), *Proc. 35th ACM Symp. Princ. Distrib. Comput.*, ACM, 2016, pp. 425–434.
- [34] N. Gupta, N. H. Vaidya, Fault-tolerance in distributed optimization: The case of redundancy, in: Y. Emek, C. Cachin (Eds.), *Proc. 39th ACM Symp. Princ. Distrib. Comput.*, ACM, 2020, pp. 365–374.
- [35] N. Gupta, T. T. Doan, N. H. Vaidya, Byzantine fault-tolerance in federated local SGD under $2f$ -redundancy, *IEEE Trans. Control. Netw. Syst.* 10 (4) (2023) 1669–1681.
- [36] Y. Wang, A. Nedic, Tailoring gradient methods for differentially private distributed optimization, *IEEE Trans. Autom. Control.* 69 (2) (2024) 872–887.

- [37] K. Wei, et al., Federated learning with differential privacy: Algorithms and performance analysis, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3454–3469.
- [38] R. C. Geyer, T. Klein, M. Nabi, Differentially private federated learning: A client level perspective, arXiv:1712.07557 (2017).
- [39] V. Pihur, et al., Differentially-private “draw and discard” machine learning, arXiv:1807.04369 (2018).

Appendix

8.1. Proof of Theorem 2

Proof. Define $\nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k) \triangleq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \lambda_k^i \nabla q^i(\bar{\mathbf{w}}_k)$ and $\mathbf{n}^{\mathcal{H}} \triangleq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbf{n}_k^i$. Using (7) and (11), the global update can be written as (22).

$$\begin{aligned}
\bar{\mathbf{w}}_{k+1} &\stackrel{(7)}{=} \frac{1}{|\mathcal{H}|} \left[\sum_{i \in \mathcal{H}} \tilde{\mathbf{w}}_{k,1}^i + \sum_{i \in \mathcal{B}_k} \tilde{\mathbf{w}}_{k,1}^i - \sum_{i \in \mathcal{H} \setminus \mathcal{B}_k} \tilde{\mathbf{w}}_{k,1}^i \right] \\
&\stackrel{(11)}{=} \bar{\mathbf{w}}_k - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \lambda_k^i \nabla q^i(\bar{\mathbf{w}}_k) + \gamma_k \cdot \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbf{n}_k^i \\
&\quad + \frac{1}{|\mathcal{H}|} \left[\sum_{i \in \mathcal{B}_k} \tilde{\mathbf{w}}_{k,1}^i - \sum_{i \in \mathcal{H} \setminus \mathcal{B}_k} \tilde{\mathbf{w}}_{k,1}^i \right] \\
&= \bar{\mathbf{w}}_k - \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k) + \gamma_k \cdot \mathbf{n}^{\mathcal{H}} \\
&\quad + \frac{1}{|\mathcal{H}|} \left[\sum_{j \in \mathcal{B}_k} (\tilde{\mathbf{w}}_{k,1}^j - \bar{\mathbf{w}}_k) - \sum_{i \in \mathcal{H} \setminus \mathcal{B}_k} (\tilde{\mathbf{w}}_{k,1}^i - \bar{\mathbf{w}}_k) \right], \tag{22}
\end{aligned}$$

where the first and last equalities follow from the PPCE property $|\mathcal{B}_k| = |\mathcal{H} \setminus \mathcal{B}_k|$. Here, $\mathcal{B}_k$ denotes the set of Byzantine agents accepted by the filter, $\mathcal{H}$ the set of non-faulty agents, and $\mathcal{H} \setminus \mathcal{B}_k$ the set of honest agents rejected by the filter.

Step 1: Error decomposition. Let $A_1 \triangleq \bar{\mathbf{w}}_k - x_{\mathcal{H}}^* - \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k) + \gamma_k \mathbf{n}^{\mathcal{H}}$, $B_1 \triangleq \frac{1}{|\mathcal{H}|} \sum_{j \in \mathcal{B}_k} (\tilde{\mathbf{w}}_{k,1}^j - \bar{\mathbf{w}}_k)$, and $C_1 \triangleq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H} \setminus \mathcal{B}_k} (\tilde{\mathbf{w}}_{k,1}^i - \bar{\mathbf{w}}_k)$. Then

$$\begin{aligned}
\mathbb{E} \|\bar{\mathbf{w}}_{k+1} - x_{\mathcal{H}}^*\|^2 &= \mathbb{E} \|A_1\|^2 + \mathbb{E} \|B_1 - C_1\|^2 \\
&\quad + 2\mathbb{E}[A_1^\top B_1] - 2\mathbb{E}[A_1^\top C_1]. \tag{23}
\end{aligned}$$

Step 2: Honest descent and noise. Under $2f$ -redundancy, $\nabla q^i(x_{\mathcal{H}}^*) = 0$ for all $i \in \mathcal{H}$. By Lipschitz smoothness (Assumption 1),

$$\|\nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k)\| \leq L\hat{\lambda}_k \|\bar{\mathbf{w}}_k - x_{\mathcal{H}}^*\|. \quad (24)$$

Moreover, $\mathbb{E}[\mathbf{n}^{\mathcal{H}}] = 0$ and $\mathbb{E}[\|\mathbf{n}^{\mathcal{H}}\|^2] = \sigma_k^2$.

Step 3: Bounding $\mathbb{E}[\|A_1\|^2]$. By strong convexity (Assumption 2) and the fact that $\lambda_k^i \geq \underline{\lambda}_k \geq h \cdot \hat{\lambda}_k$ in (18),

$$\begin{aligned} (\nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k))^{\top} (\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*) &\geq \frac{\mu}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \lambda_k^i \|\bar{\mathbf{w}}_k - \mathbf{w}^*\|^2 \\ &\geq \mu h \hat{\lambda}_k \|\bar{\mathbf{w}}_k - \mathbf{w}^*\|^2. \end{aligned} \quad (25)$$

Combining (24)–(25) yields

$$\begin{aligned} \mathbb{E}[\|A_1\|^2] &\leq (1 - 2h\mu\hat{\lambda}_k + 2L^2\hat{\lambda}_k^2) \mathbb{E}[\|\bar{\mathbf{w}}_k - x_{\mathcal{H}}^*\|^2] \\ &\quad + 4\bar{C}_0\gamma_k\sigma + 2\gamma_k^2\sigma^2. \end{aligned} \quad (26)$$

Step 4: Bounding Byzantine residuals. From (11) and Lipschitz continuity,

$$\|\tilde{\mathbf{w}}_{k,1}^i - \bar{\mathbf{w}}_k\| \leq L\lambda_k^i \|\bar{\mathbf{w}}_k - x_{\mathcal{H}}^*\| + \gamma_k \|n_k^i\|. \quad (27)$$

Using the PPCE filter and $|\mathcal{B}_k| = |\mathcal{H} \setminus \mathcal{B}_k|$,

$$\mathbb{E}\|B_1 - C_1\|^2 \leq \frac{8L^2|\mathcal{B}_k|^2}{|\mathcal{H}|^2} \hat{\lambda}_k^2 \mathbb{E}\|\bar{\mathbf{w}}_k - x_{\mathcal{H}}^*\|^2 + \frac{8|\mathcal{B}_k|^2}{|\mathcal{H}|^2} \gamma_k^2 \sigma^2. \quad (28)$$

Step 5: Cross terms. Using symmetry of the PPCE filter and bounded iterates,

$$\begin{aligned} 2\mathbb{E}[A_1^{\top} B_1] &\leq \frac{2|\mathcal{B}_k|}{|\mathcal{H}|} (L\hat{\lambda}_k + 2L^2\hat{\lambda}_k^2) \mathbb{E}\|\bar{\mathbf{w}}_k - x_{\mathcal{H}}^*\|^2 \\ &\quad + \frac{4|\mathcal{B}_k|}{|\mathcal{H}|} \bar{C}_0\gamma_k\sigma + \frac{4|\mathcal{B}_k|}{|\mathcal{H}|} \gamma_k^2\sigma^2, \end{aligned} \quad (29)$$

$$\begin{aligned} -2\mathbb{E}[A_1^{\top} C_1] &\leq \frac{2|\mathcal{B}_k|}{|\mathcal{H}|} (L\hat{\lambda}_k + L^2\hat{\lambda}_k^2) \mathbb{E}\|\bar{\mathbf{w}}_k - x_{\mathcal{H}}^*\|^2 \\ &\quad + \frac{|\mathcal{B}_k|}{|\mathcal{H}|} \gamma_k^2\sigma^2. \end{aligned} \quad (30)$$

Step 6: Collecting terms. Substituting into (23) gives

$$\begin{aligned} \mathbb{E}[\|\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^*\|^2] &\leq \left(1 - 2\left(h\mu - \frac{4Lf}{|\mathcal{H}|}\right)\hat{\lambda}_k + 5L^2\hat{\lambda}_k^2\right) \\ &\quad \cdot E[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] + 8\bar{C}_0\gamma_k\sigma + 5\gamma_k^2\sigma^2. \end{aligned} \quad (31)$$

Here, since $2/3 < h < 1$ and $|\mathcal{H}| \geq N - f$, the condition $\frac{f}{N-f} \leq \frac{\mu}{6L}$ ensures that $h\mu - \frac{4Lf}{|\mathcal{H}|} \geq \frac{\mu}{3}$.

Step 7: Final contraction. : If $\frac{f}{N-f} \leq \frac{\mu}{6L}$ and $\hat{\lambda}_k \leq \frac{\mu}{10L^2}$, then

$$\begin{aligned} \mathbb{E}\|\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^*\|^2 &\leq \left(1 - \frac{\mu}{6}\hat{\lambda}_k\right)\mathbb{E}\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2 \\ &\quad + (8\bar{C}_0 + 5\gamma_k\sigma)\gamma_k\sigma. \end{aligned} \quad (32)$$

Requiring $\frac{2\mu}{3}\hat{\lambda}_k - 5L^2\hat{\lambda}_k^2 \geq \frac{\mu}{6}\hat{\lambda}_k$ yields $\hat{\lambda}_k \leq \frac{\mu}{10L^2}$. This completes the proof. $\square$

8.2. Proof of Theorem 3

The following bounds control the drift of local iterates.

Lemma 5. If $\lambda_k^i \leq \mu/(16L^2\mathcal{T})$, then for all $t \in [0, \mathcal{T} - 1]$,

$$\|\mathbf{w}_{k,t+1}^i - \mathbf{w}_{\mathcal{H}}^*\| \leq 2\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|, \quad (33)$$

$$\|\mathbf{w}_{k,t+1}^i - \bar{\mathbf{w}}_k\| \leq 2\mathcal{T}L\lambda_k^i\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|. \quad (34)$$

Proof. By $\lambda_k^i \leq \hat{\lambda}_k \leq \mu/(16\mathcal{T}L^2)$ and Assumptions 1-2 (which imply $\mu \leq L$), we have $L\mathcal{T}\lambda_k^i \leq \frac{\mu}{16L} \leq \frac{1}{16} \leq \ln 2$.

From the local update rule (cf. Algorithm (5)), it follows that $\mathbf{w}_{k,t+1}^i = \mathbf{w}_{k,t}^i - \lambda_k^i \nabla q^i(\mathbf{w}_{k,t}^i)$.

Since $\nabla q^i(\mathbf{w}_{\mathcal{H}}^*) = 0$, for any $t \in [0, \mathcal{T} - 1]$,

$$\begin{aligned} \|\mathbf{w}_{k,t+1}^i - \mathbf{w}_{k,t}^i\| &= \lambda_k^i \|\nabla q^i(\mathbf{w}_{k,t}^i) - \nabla q^i(\mathbf{w}_{\mathcal{H}}^*)\| \\ &\leq L\lambda_k^i \|\mathbf{w}_{k,t}^i - \mathbf{w}_{\mathcal{H}}^*\|. \end{aligned} \quad (35)$$

Using the triangle inequality, $\|\mathbf{w}_{k,t+1}^i - \mathbf{w}_{\mathcal{H}}^*\| \leq (1 + L\lambda_k^i)\|\mathbf{w}_{k,t}^i - \mathbf{w}_{\mathcal{H}}^*\| \leq e^{L\lambda_k^i}\|\mathbf{w}_{k,t}^i - \mathbf{w}_{\mathcal{H}}^*\|$.

Iterating yields $\|\mathbf{w}_{k,t+1}^i - \mathbf{w}_{\mathcal{H}}^*\| \leq e^{Lt\lambda_k^i} \|\mathbf{w}_{k,0}^i - \mathbf{w}_{\mathcal{H}}^*\| \leq e^{L\mathcal{T}\lambda_k^i} \|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\| \leq 2\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|$, where we used $\mathbf{w}_{k,0}^i = \bar{\mathbf{w}}_k$ and $L\mathcal{T}\lambda_k^i \leq \ln 2$.

Combining this bound with (35), we obtain $\|\mathbf{w}_{k,t+1}^i - \mathbf{w}_{k,t}^i\| \leq 2L\lambda_k^i \|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|$.

Summing over $l = 0, \dots, t$ and using $\mathbf{w}_{k,0}^i = \bar{\mathbf{w}}_k$ gives $\|\mathbf{w}_{k,t+1}^i - \bar{\mathbf{w}}_k\| \leq \sum_{l=0}^t \|\mathbf{w}_{k,l+1}^i - \mathbf{w}_{k,l}^i\| \leq 2(t+1)L\lambda_k^i \|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|$. Setting $t = \mathcal{T} - 1$ completes the proof. $\square$

Proof. Let $\hat{\lambda}_k = \max_{i \in \mathcal{H}} \{\lambda_k^i\}$ and assume $\hat{\lambda}_k \leq \mu/(16L^2\mathcal{T})$. By Lemma 5, inequalities (33)–(34) hold for all $t \in [0, \mathcal{T} - 1]$ and $i \in \mathcal{H}$.

Step 1: Decomposition of the averaged iterate. Using the update rule and the filtering mechanism, the averaged iterate admits the decomposition $\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^* = A_2 - D_2 + B_2$, where $A_2 = \bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^* - \mathcal{T}\lambda_k \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k) + \gamma_k \cdot n^{\mathcal{H}}$ is the ideal descent-plus-noise term, $B_2 = \frac{1}{|\mathcal{H}|} (\sum_{j \in \mathcal{B}_k} (x_{k,\mathcal{T}}^j - \bar{\mathbf{w}}_k) - \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} (x_{k,\mathcal{T}}^i - \bar{\mathbf{w}}_k))$ represents the Byzantine perturbation, and $D_2 = \frac{\lambda_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \sum_{t=0}^{\mathcal{T}-1} (\nabla q^i(\mathbf{w}_{k,t}^i) - \nabla q^i(\bar{\mathbf{w}}_k))$ captures the multi-step gradient drift.

Step 2: Expansion of the squared norm. Expanding $\|\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^*\|^2$ yields

$$\|A_2\|^2 + \|B_2\|^2 + \|D_2\|^2 + 2A_2^T B_2 - 2A_2^T D_2 - 2D_2^T B_2.$$

Step 3: Bounding the main descent term. Recall that A_2 corresponds to the averaged ideal descent direction with additive noise. Conditioning on $\bar{\mathbf{w}}_k$ and using $\nabla q^{\mathcal{H}}(\mathbf{w}_{\mathcal{H}}^*) = 0$, we write $A_2 = \bar{\mathbf{w}}_k - \mathcal{T}h\hat{\lambda}_k \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k) + \boldsymbol{\xi}_k$, where $\boldsymbol{\xi}_k$ denotes the aggregated noise term satisfying $\mathbb{E}[\boldsymbol{\xi}_k \mid \bar{\mathbf{w}}_k] = 0$ and $\mathbb{E}[\|\boldsymbol{\xi}_k\|^2] = O(\gamma_k \sigma + \gamma_k^2 \sigma^2)$.

Expanding the squared norm and taking expectations yield

$$\begin{aligned} \mathbb{E}[\|A_2\|^2] &= \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^* - \mathcal{T}h\hat{\lambda}_k \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k)\|^2] + \mathbb{E}[\|\boldsymbol{\xi}_k\|^2] \\ &= \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] - 2\mathcal{T}h\hat{\lambda}_k \mathbb{E}[\langle \bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*, \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k) \rangle] \\ &\quad + \mathcal{T}^2 h^2 \hat{\lambda}_k^2 \mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k)\|^2] + O(\gamma_k \sigma + \gamma_k^2 \sigma^2). \end{aligned} \tag{36}$$

By μ -strong convexity of $q^{\mathcal{H}}$, $\langle \bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*, \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k) \rangle \geq \mu \|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2$, and by L -Lipschitz continuity, $\|\nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k)\| \leq L \|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|$.

Substituting these bounds into (36), we obtain $\mathbb{E}[\|A_2\|^2] \leq \left(1 - 2\mathcal{T}h\mu\hat{\lambda}_k + \mathcal{T}^2h^2L^2\hat{\lambda}_k^2\right)\mathbb{E}\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2 + O(\gamma_k\sigma + \gamma_k^2\sigma^2)$.

Finally, $h \in (\frac{2}{3}, 1)$, the quadratic term is less than $\mathcal{T}^2L^2\hat{\lambda}_k^2$, which yields

$$\begin{aligned} \mathbb{E}[\|A_2\|^2] &\leq (1 - 2\mathcal{T}h\mu\hat{\lambda}_k + \mathcal{T}^2L^2\hat{\lambda}_k^2) \\ &\quad \cdot \mathbb{E}\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2 + O(\gamma_k\sigma + \gamma_k^2\sigma^2). \end{aligned} \quad (37)$$

Step 4: Bounding $\mathbb{E}[\|D_2\|^2]$ and $\mathbb{E}[\|B_2\|^2]$.

Drift term. Using the inequalities (33)–(34) in Lemma 5, we have for each honest worker $i \in \mathcal{H}$: $\|\nabla q_i(\bar{\mathbf{w}}_k^t) - \nabla q_i(\bar{\mathbf{w}}_k)\| = O(\mathcal{T}L^2\hat{\lambda}_k)\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|$, where the factor $\mathcal{T}L^2\hat{\lambda}_k$ arises from multiple local steps.

Squaring and summing over $t = 0, \dots, \mathcal{T}-1$ gives $\|D_2\|^2 = \left\| \sum_{t=0}^{\mathcal{T}-1} (\nabla q_i(\bar{\mathbf{w}}_k^t) - \nabla q_i(\bar{\mathbf{w}}_k)) \right\|^2 = O(\mathcal{T}^4L^4\hat{\lambda}_k^3) \cdot \|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2$. Taking expectations yields

$$\mathbb{E}[\|D_2\|^2] = O(\mathcal{T}^4L^4\hat{\lambda}_k^3)\mathbb{E}\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2. \quad (38)$$

Byzantine term. Let $|\mathcal{B}_k| < f$ denote the number of Byzantine workers. By standard aggregation bounds (e.g., coordinate-wise median or trimmed mean) and Lipschitz continuity, we have $\|B_2\|^2 = O\left(\frac{f^2}{|\mathcal{H}|^2} \sum_{i \in \mathcal{H}} \|\nabla q_i(\bar{\mathbf{w}}_k) - \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k)\|^2\right) + O(\gamma_k^2\sigma^2)$.

Using Lipschitz continuity of ∇q_i and $\nabla q^{\mathcal{H}}$, this gives $\|\nabla q_i(\bar{\mathbf{w}}_k) - \nabla q^{\mathcal{H}}(\bar{\mathbf{w}}_k)\|^2 = O(\mathcal{T}^2L^2\hat{\lambda}_k^2)\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2$.

Combining these estimates yields

$$\mathbb{E}[\|B_2\|^2] = O\left(\frac{f^2}{|\mathcal{H}|^2} \mathcal{T}^2L^2\hat{\lambda}_k^2\right)\mathbb{E}\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2 + O(\gamma_k^2\sigma^2). \quad (39)$$

Step 5: Bounding cross terms. Note that $|\mathcal{B}_k| < f$ and $|\mathcal{F}_k| = |\mathcal{H}| = N - f$.

Using the bounds

$$\begin{aligned} \|A_2\| &\leq (1 + 2\mathcal{T}^2L\hat{\lambda}_k)\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\| + \gamma_k\|n^{\mathcal{H}}\|, \\ \|B_2\| &\leq \frac{4\mathcal{T}L|\mathcal{B}_k|}{|\mathcal{H}|}\hat{\lambda}_k\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\| + \gamma_k\|n_k^i\|, \\ \|D_2\| &\leq 4\mathcal{T}^4L^4\hat{\lambda}_k^3\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|, \end{aligned}$$

together with $\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\| \leq 2\bar{C}_0$, the cross terms satisfy

$$\begin{aligned} \mathbb{E}[2A_2^T B_2] &\leq \left(\frac{8\mathcal{T}Lf}{|\mathcal{H}|} \hat{\lambda}_k + 2\mathcal{T}L \left(\mathcal{T} + \frac{(2 + 8\mathcal{T}^2 L)f}{|\mathcal{H}|} \right) \hat{\lambda}_k^2 \right) \\ &\quad \cdot \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] + O(\gamma_k \sigma + \gamma_k^2 \sigma^2), \end{aligned} \quad (40)$$

$$\begin{aligned} \mathbb{E}[-2A_2^T D_2] &\leq (4\mathcal{T}^2 L^2 \hat{\lambda}_k^2 + 8\mathcal{T}^4 L^3 \hat{\lambda}_k^3 + 2\mathcal{T}^4 L^4 \hat{\lambda}_k^4) \\ &\quad \cdot \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] + 2\gamma_k^2 \sigma^2, \end{aligned} \quad (41)$$

$$\begin{aligned} \mathbb{E}[-2D_2^T B_2] &\leq \left(\frac{16\mathcal{T}^3 L^3 f}{|\mathcal{H}|} \hat{\lambda}_k^3 + 2\mathcal{T}^4 L^4 \hat{\lambda}_k^4 \right) \\ &\quad \cdot \mathbb{E}[\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2] + 2\gamma_k^2 \sigma^2. \end{aligned} \quad (42)$$

Step 6: Aggregation. Note that $\hat{\lambda}_k L^2 \mathcal{T} \leq 1$ and $\hat{\lambda}_k L \mathcal{T} \leq 1$ since $\hat{\lambda}_k \leq \mu/(16L^2 \mathcal{T})$ and $0 < \mu \leq L$.

Collecting the preceding bounds yields the recursion $\mathbb{E}[\|\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^*\|^2] \leq (1 - (2\mathcal{T}h\mu - \frac{8\mathcal{T}Lf}{|\mathcal{H}|})\hat{\lambda}_k + (20 + 8L)\mathcal{T}^2 L \hat{\lambda}_k^2) \mathbb{E}\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2 + O(\gamma_k \sigma + \gamma_k^2 \sigma^2)$. Since $h \in (\frac{2}{3}, 1)$ and $\frac{f}{|\mathcal{H}|} = \frac{f}{N-f} \leq \min\{\frac{1}{3}, \frac{\mu}{12L}\}$, we have that $2\mathcal{T}h\mu - \frac{8\mathcal{T}Lf}{|\mathcal{H}|} > \frac{4\mathcal{T}\mu}{3} - 8\mathcal{T}L \cdot \frac{\mu}{12L} = \frac{2\mathcal{T}\mu}{3}$. Therefore,

$$\begin{aligned} \mathbb{E}[\|\bar{\mathbf{w}}_{k+1} - \mathbf{w}_{\mathcal{H}}^*\|^2] &\leq \left(1 - \frac{2\mu\mathcal{T}}{3} \hat{\lambda}_k + (20 + 8L)\mathcal{T}^2 L \hat{\lambda}_k^2 \right) \\ &\quad \cdot \mathbb{E}\|\bar{\mathbf{w}}_k - \mathbf{w}_{\mathcal{H}}^*\|^2 + O(\gamma_k \sigma + \gamma_k^2 \sigma^2). \end{aligned} \quad (43)$$

Imposing $\frac{2\mu\mathcal{T}}{3} \hat{\lambda}_k - (20 + 8L)\mathcal{T}^2 L \hat{\lambda}_k^2 \geq \frac{\mu\mathcal{T}}{6} \hat{\lambda}_k$ is satisfied whenever $\hat{\lambda}_k \leq \frac{\mu}{(40 + 16L)\mathcal{T}^2}$.

This completes the proof. $\square$