# Optimal Byzantine-resilient Federated Learning with User-level Differential Privacy

**Ming Xiang**[1] **Stratis Ioannidis**[1] **Edmund Yeh**[1] **Carlee Joe-Wong**[2] **Lili Su**[1]
[1]Northeastern University, Boston, MA, USA; [2]Carnegie Mellon University, Pittsburgh, PA, USA.
{xiang.mi,e.ioannidis,e.yeh,l.su}@northeastern.edu, cjoewong@andrew.cmu.edu

## Abstract

Real-world deployment of federated learning (FL) requires resilience against adversarial clients and data privacy. Byzantine fault captures the worst-case behavior of adversarial clients, while differential privacy (DP) provides statistical guarantees against privacy loss. In particular, user-level DP guards a client's entire contribution and provides more realistic protection against information leakage. Despite intensive efforts, the minimax-optimal convergence bound of Byzantine-resilient FL with user-level DP remains open. We establish an information-theoretic lower bound for any Byzantine-resilient empirical risk minimization problem under user-level DP and heterogeneous clients. We propose ByzULDP, a central DP algorithm featuring a two-level momentum across the client and server. Client-side momentum mitigates stochastic gradient variance during robust aggregation. Unlike the fault-free case, naively combining non-linear Byzantine-resilient aggregation with central DP noise can be insufficient to guarantee convergence to a bounded neighborhood of a stationary point. Our key contribution is a server-side momentum that leverages a property specific to user-level DP, where privacy sensitivity depends on how client updates are constructed, to empower convergence. ByzULDP achieves the order-optimal convergence bound for strongly-convex objectives, with vanishing error terms decaying at the optimal $\mathcal{O}(1/T)$ and $\mathcal{O}(1/\sqrt{T})$ rates for strongly-convex and non-convex objectives, respectively. We corroborate our analysis with numerical experiments on real-world datasets across diverse privacy budgets.

## 1 Introduction

Federated learning (FL) enables large-scale AI/ML training without transmitting raw local data through collaboration between a parameter server (PS) and clients. Its success has been evidenced by widespread commercial deployments in diverse fields such as autonomous vehicles [18, 56, 74], the Internet of things [30, 55, 75], and natural language processing [47, 49, 53].

While FL keeps raw data locally, shared model parameters can still leak sensitive information about clients' training data [12, 38]. Differential privacy (DP) provides statistical guarantees against such leakage [26], with the client as the natural unit of protection. In regulated domains such as healthcare and finance, the aggregating server operates as a trusted intermediary under legal accountability. The privacy threat comes from external auditors or curious clients, who attempt to infer sensitive information about participating clients. Meanwhile, FL operates over a complex and heterogeneous client population [40], where faulty or abnormal clients arise due to software or hardware bugs [58], poisoned data [59], or adversarial manipulations [72]. Such unstructured failures are well captured by the Byzantine failure model [50]. Byzantine-resilience [17, 19, 70] and user-level DP [37, 45, 66] have each received extensive attention in isolation. However, real-world deployments, such as healthcare, mobile personalization, and financial collaborations, require addressing both of them simultaneously. Their integration is fundamentally challenging: Byzantine-resilience aims to preserve

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the utility by discriminating adversarial updates, whereas user-level DP aims to obscure clients' contributions, potentially drowning useful signals. We therefore ask:

***Question:*** *Under a trusted server, what is the minimax-optimal convergence bound of user-level DP empirical risk minimization (ERM) under Byzantine adversaries and heterogeneous clients?*

Combining DP and Byzantine-resilience remains largely underexplored. One line of work provides per-step privacy guarantees [34, 51, 67, 76], yielding a loose cumulative privacy cost that grows with the number of rounds $T$. Only a handful of works study the synthesis of DP and Byzantine-resilience with multi-round guarantees. Allouah et al. [7] and Gu et al. [33] both consider item-level DP with Byzantine resilience, with the analysis of the latter requiring i.i.d. clients. Yet, FL is by definition heterogeneous, and item-level DP often provides insufficient protection against user-level leakage [45]. For example, in a specialized hospital, client records often exhibit correlated structures. Per-item perturbations cannot fully mask cross-record correlations, leaving the cumulative client signature recoverable. Closer to our setting, Allouah et al. [8] achieve user-level DP without a trusted server via locally correlated noise injection at each client, seeking to attain the utility of central DP algorithms, where the server is trusted. In the absence of Byzantine adversaries, central DP is known to achieve a favorable privacy-utility trade-off [25]. However, the minimax-optimal convergence bound for Byzantine-resilient ERM with user-level DP remains open.

**Contributions.** In this work, we establish order-matching lower and upper bounds for Byzantine-resilient ERM with user-level DP. The lower bound is privacy-regime agnostic, while the upper bound is attained by a central DP algorithm, showing that central DP suffices to achieve the optimum. Our contributions are fourfold:

- In Section 4, we establish an information-theoretic lower bound for any Byzantine-resilient ERM problem under user-level DP constraints and heterogeneous clients.
- In Section 5, we propose the `ByzULDP`, which features two-level momentum across the client and the server. Client-side momentum mitigates stochastic gradient variance during robust aggregation, which ensures Byzantine resilience. Server-side momentum reshapes each client's contribution so that the privacy sensitivity achieves a reduction specific to user-level DP, enabling convergence.
- In Section 6, we provide two convergence results for `ByzULDP`. For strongly convex objectives, the non-vanishing errors match our lower bound, and the vanishing error terms decay at the optimal $\mathcal{O}(1/T)$ rate for stochastic first-order methods. For non-convex objectives, we obtain the optimal $\mathcal{O}(1/\sqrt{T})$ rate on the vanishing error terms.
- In Section 7, we corroborate our analysis with numerical experiments on real-world datasets across a range of privacy budgets.

## 2 Related Work

**User-level Private Learning.** Without Byzantine resilience, user-level private learning has received growing attention. For example, Badih et al. [11] study user-level DP for PAC learning and characterize the number of clients required to learn a concept class. The seminal work [45] by Levy et al. studies user-level DP for mean estimation, ERM, and stochastic convex optimization (SCO) under homogeneous client distributions or mildly heterogeneous data. Bassily et al. [16] and Liu et al. [48] both achieve the optimal rates for user-level DP-SCO, with the latter adopting more relaxed assumptions on how the number of clients scales with the problem dimension. Another line of work explores intermediate threat models bridging local and central DP by using cryptographic techniques or anonymization via shuffling [22, 39]. Yet, Byzantine clients fundamentally alter the problem. Most prior work relies on a simple mean aggregation, which is highly vulnerable to adversarial updates. Naively combining robust aggregators with DP noise is insufficient to guarantee convergence to a bounded neighborhood of a stationary point, as these nonlinear aggregators often exhibit high privacy sensitivity; see Remark 2 for details.

**Byzantine-resilient Learning.** Tolerating Byzantine attacks in distributed learning has received intensive attention [4, 17, 19, 27, 28, 31, 41, 42, 61, 63, 70, 73]. A robust aggregator is the most common way to mitigate the negative impacts of adversarial updates from Byzantine clients. Numerous robust aggregators have been developed over the recent years, including Geometric Median [19], Krum [17], Coordinate-wise Trimmed Mean [73], Coordinate-wise Median [73], Bulyan [35], and centered clipping [42]. Yet, their original analysis is under a simplified i.i.d. setting. Karimireddy et

al. [41] and Allouah et al. [4] have shown that any distributed algorithm can only reach a neighborhood of a stationary point in the presence of coupled non-i.i.d. data and Byzantine faults. Karimireddy et al. [41] propose a simple bucketing mechanism to reduce the likelihood that any subset of clients is consistently compromised. Allouah et al. [4] propose nearest-neighbor mixing, which remixes the clients' updates with their closest neighbors to obtain a better representation of the global gradient.

**Robust learning and estimation with DP.** Recent advances focus on the user-level private-mean estimation problem with [1, 21, 43] or without Byzantine adversaries [2, 32, 54], where the clients send the perturbed local data to the server only once. However, FL prohibits direct data transmission and requires multi-round interactions. Another line of work leverages cryptographic techniques, e.g., homomorphic encryption and secure multi-party computation to reconcile robustness and privacy [23, 24], but these approaches incur prohibitive computation overhead when combined with robust aggregations. Recently, Allouah et al. [7] establish a privacy-utility-robustness trilemma for item-level DP in the presence of Byzantine adversaries. Gu et al. [33] also study item-level DP and propose algorithms in both local and central DP modes, yet their analysis is only applicable to i.i.d. clients. Allouah et al. [8] consider user-level DP with an untrusted server and propose an algorithm that uses correlated local noise to achieve utility comparable to central DP algorithms. We establish an improved convergence bound in the central DP setting; see Remark 1 for details.

## 3 Problem Formulation

**Federated Learning System.** We consider the classic master-slave architecture with a collection of $n$ clients and a PS. The clients hold local data sets $D_1, \ldots, D_n$. The local loss function at client $i$ is defined as $F_i(\boldsymbol{\theta}) \triangleq \mathbb{E}_{(x,y)\sim D_i}[\ell(\boldsymbol{\theta}, (x, y))]$, where, with a little abuse of notation, $\mathbb{E}_{(x,y)\sim D_i}$ can be understood as taking expectation over the empirical data distribution. The global objective is the average of clients' local loss functions:

$$F(\boldsymbol{\theta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{\theta}).$$

**Threat model.** We consider an FL environment with two simultaneous threats: *Byzantine faults* and *client privacy leakage*. The PS aims to robustly learn a model despite malicious updates from clients under Byzantine faults. At the same time, it also faces privacy adversaries, including external auditors who analyze the deployed model and participating clients who observe per-round updates.

*Byzantine fault.* We follow the classic notion of Byzantine fault in distributed systems, where up to $f$ clients can be compromised to send malicious messages. These clients, collectively denoted by $\mathcal{B}$, can collude and adaptively send malicious messages based on previously observed information exchange, but cannot impersonate peers or observe or interfere with the communications of honest clients [50]. Let $\mathcal{G} \triangleq [n] \setminus \mathcal{B}$ denote the set of honest clients. Due to the existence of Byzantine clients, it is impossible to minimize the fault-free objective $F$ [62]. Therefore, we follow the mainstream in the literature of Byzantine-resilient distributed learning [4, 7, 8, 41], and henceforth consider (1).

$$F_{\mathcal{G}}(\boldsymbol{\theta}) \triangleq \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} F_i(\boldsymbol{\theta}) \tag{1}$$

*User-level differential privacy.* We consider central DP [16, 45], where the PS is assumed to be trustworthy. To protect honest clients from having their participation inferred by privacy adversaries, we adopt Byzantine user-level DP in Definition 1, with adjacency defined exclusively over the set of honest clients to account for Byzantine behavior.

**Definition 1** ($(\epsilon, \delta)$-Byzantine user-level DP)**.** Let $\mathcal{X}$ be the space of possible client datasets, $[n]$ be the set of client indices, which includes a set of honest clients $\mathcal{G}$ and a set of Byzantine clients $\mathcal{B}$ such that $\mathcal{G} \cap \mathcal{B} = \emptyset$.

- **Byzantine strategy.** A Byzantine strategy $\mathfrak{B}$ is an arbitrary, possibly randomized and colluding rule whose round-$t$ actions depend on the honest datasets $\mathcal{D}_{\mathcal{G}}$ only through what the server publishes in rounds $s < t$.
- **Output.** For an algorithm $\mathcal{A}$, let $\mathcal{A}_{\mathfrak{B}}(\mathcal{D}_{\mathcal{G}})$ denote its output, i.e., everything the server publishes over the course of its execution, when the honest clients hold the dataset collection $\mathcal{D}_{\mathcal{G}} = \{D_i\}_{i\in\mathcal{G}} \in \mathcal{X}^{|\mathcal{G}|}$, and the Byzantine clients follow strategy $\mathfrak{B}$.

- **Neighboring.** Two honest dataset collections $\mathcal{D}_{\mathcal{G}}$ and $\mathcal{D}'_{\mathcal{G}}$ are said to be *neighboring*, denoted by $\mathcal{D}_{\mathcal{G}} \sim \mathcal{D}'_{\mathcal{G}}$, if they differ in the dataset of exactly one honest client.

We say that $\mathcal{A}$ satisfies $(\epsilon, \delta)$-Byzantine user-level DP if, for every Byzantine strategy $\mathfrak{B}$, every pair of neighboring collections $\mathcal{D}_{\mathcal{G}} \sim \mathcal{D}'_{\mathcal{G}}$, it holds for every measurable set $Q$,

$$\mathbb{P}\left\{\mathcal{A}_{\mathfrak{B}}(\mathcal{D}_{\mathcal{G}}) \in Q\right\} \leq e^{\epsilon}\mathbb{P}\left\{\mathcal{A}_{\mathfrak{B}}(\mathcal{D}'_{\mathcal{G}}) \in Q\right\} + \delta.$$

## 4 Lower Bound

In this section, we introduce an information-theoretic lower bound in Theorem 1 for any Byzantine-resilient ERM that ensures $(\epsilon, \delta)$ user-level DP under heterogeneous clients. The full proofs are deferred to Appendix E.5.

**Assumption 1.** There exist $B \geq 0$ and $G \geq 0$ such that $\forall \boldsymbol{\theta} \in \mathbb{R}^d$

$$\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \|\nabla F_i(\boldsymbol{\theta}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta})\|_2^2 \leq B^2 \|\nabla F_{\mathcal{G}}(\boldsymbol{\theta})\|_2^2 + G^2. \tag{2}$$

Assumption 1 captures the heterogeneity across honest clients [6, 41, 42], which is what makes both Byzantine robustness and user-level DP fundamentally hard. When $G = 0$, all honest clients share common stationary points and become almost statistically indistinguishable. Specifically, clients' full-batch gradients become identical when $B = G = 0$, making the DP guarantees trivial. Even when $B > 0$ and $G = 0$, the same degeneracy may arise, e.g., if each $F_i$ is strongly-convex or admits a unique stationary point. In any of these regimes, Byzantine clients cannot deviate without being identified as outliers and are readily excluded. Henceforth, we focus on the regime where $G > 0$.

**Theorem 1.** *For any given $L, B \geq 0$ and $G > 0$, there exists a collection of $n$ functions $\{F_i\}_{i\in[n]}$, each defined over a dataset of $m$ samples, that are all i) $L$-smooth, ii) $\mu$-strongly convex with $\mu = \Theta(L)$, and iii) satisfy Assumption 1. Let $m, n, d \geq 1$ and $1 \leq f < n/2$. Suppose that $\epsilon, \delta \in (0, 1)$ and $1/2^{(n-f)^{1-\gamma}} \leq \delta \leq 1/(n-f)^{1+\gamma}$, where $\gamma \in (0, 1)$. If any distributed algorithm $\mathcal{A}$ is $(\epsilon, \delta)$-Byzantine user-level DP, then its output $\widehat{\boldsymbol{\theta}}$ satisfies*

$$\mathbb{E}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 = \Omega\big(\frac{f}{n}\frac{G^2}{\mu^2} + \big(1 + \frac{f}{n}\big)\frac{dG^2\log(1/\delta)}{\mu^2\epsilon^2(n-f)^2}\big),$$
$$\mathbb{E}\|\nabla F_{\mathcal{G}}(\widehat{\boldsymbol{\theta}})\|_2^2 = \Omega\big(\frac{f}{n}G^2 + \big(1 + \frac{f}{n}\big)\frac{dG^2\log(1/\delta)}{\epsilon^2(n-f)^2}\big),$$

*where $\widehat{\boldsymbol{\theta}}$ is the output of $\mathcal{A}$, the expectation is taken over the randomness in $\mathcal{A}$ and the data generation, and $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} F_i(\boldsymbol{\theta})$.*

The number of samples per client $m$ does not appear in the bound because we focus on the ERM problem with *arbitrary* data. In the worst case, each client $i \in [n]$ may hold $m$ identical copies of a single point, with the point differing across clients. User-level DP reduces to item-level DP over $n$ data points, and no improvement in $m$ is possible. Note that the construction of Theorem 1 is based on $m$ data samples. The absence of $m$ is inherent to the worst-case ERM setting, rather than an artifact of our proof. When no Byzantine clients are present ($f = 0$), the bound reduces to $\Omega(dG^2\log(1/\delta)/(\epsilon^2 n^2))$, matching the centralized item-level ERM lower bound in [15] up to the dependence on $G$, which is normalized to $G = \mathcal{O}(1)$ in their construction. The constraints on $\epsilon$ and $\delta$ are necessary only for the lower bound [60], which can be relaxed at the cost of a looser bound via techniques from [15]. The proof of Theorem 1 includes three subcases, each captured by a separate lemma. The theorem follows by taking a maximum. Lemma 1 recovers the DP-only cost in the absence of Byzantine clients.

**Lemma 1** (Non-Byzantine)**.** *Suppose that the conditions in Theorem 1 hold with $f = 0$, the output $\widehat{\boldsymbol{\theta}}$ of a $(\epsilon, \delta)$ user-level DP learning algorithm $\mathcal{A}$ has an error at least*

$$\mathbb{E}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 = \Omega\big(\frac{dG^2\log(1/\delta)}{\mu^2\epsilon^2 n^2}\big),\ \mathbb{E}\|\nabla F(\widehat{\boldsymbol{\theta}})\|_2^2 = \Omega\big(\frac{dG^2\log(1/\delta)}{\epsilon^2 n^2}\big). \tag{3}$$

**Algorithm 1:** ByzULDP

**Input:** $\boldsymbol{\theta}^0$, $\boldsymbol{m}_i^0 = \mathbf{0}, \forall i \in \mathcal{G}$, $\widetilde{R}^{-1} = \mathbf{0}$, $R$, $\sigma_{dp,t}^2$, $\eta_t$, $\beta_t$, $\alpha_t$, and $T$;
**Output:** $\boldsymbol{\theta}^T$.
**for** $t = 0, \ldots, T-1$ **do**

/* On all clients */

**for** $i = 1, \cdots, n$ **do**
Download $\boldsymbol{\theta}^t$ from the PS ;
$\widehat{\boldsymbol{g}}_i^t \leftarrow \texttt{clip}(\nabla \ell_i(\boldsymbol{\theta}^t; \xi_i^t), C)$, where $\texttt{clip}(\boldsymbol{v}, C) = \boldsymbol{v} \cdot \min\left\{1, C/\|\boldsymbol{v}\|_2\right\}$;
$\boldsymbol{m}_i^t \leftarrow \alpha_t \boldsymbol{m}_i^{t-1} + (1-\alpha_t)\widehat{\boldsymbol{g}}_i^t$;
$\widehat{\boldsymbol{m}}_i^t = \begin{cases} \boldsymbol{m}_i^t, & \text{if } i \in \mathcal{G}; \\ *, & \text{otherwise.} \end{cases}$;
Send $\widehat{\boldsymbol{m}}_i^t$ to the PS.
**end**

/* On the PS */

**for** $i = 1, \cdots, n$ **do**
$\boldsymbol{M}_i^t \leftarrow \beta_t \widetilde{R}^{t-1} + (1-\beta_t)\widehat{\boldsymbol{m}}_i^t$;
**end**
$R^t \leftarrow R(\boldsymbol{M}_1^t, \ldots, \boldsymbol{M}_n^t)$
$\widetilde{R}^t \leftarrow R^t + \boldsymbol{X}_{dp}^t$, where $\boldsymbol{X}_{dp}^t \sim \mathcal{N}(0, \sigma_{dp,t}^2 \mathbf{I})$
$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta_t \widetilde{R}^t$;

**end**

To show Lemma 1, we reduce the distributed learning problem to a centralized DP mean estimation of one-way marginals. The key technique is to adapt [45, Proposition 2] to our setting: user-level DP mean estimation is equivalent to the item-level mean estimation problem, where each item is the mean of a client's local dataset, thereby reducing a multi-point estimation problem to a single-point one. This reduction clarifies the earlier intuition: the $m$ data points held by each client contain no more private information than their mean, which explains why $m$ does not appear in the bound. Lemma 2 characterizes the sole cost of Byzantine resilience in the absence of DP constraints. Its proof follows from the standard roadmap in the Byzantine-resilient learning literature [4, 7, 42].

**Lemma 2** (Non-private)**.** *Suppose that the conditions in Theorem 1 hold with $1 \leq f < n/2$, the output $\widehat{\boldsymbol{\theta}}$ of a $(\epsilon, \delta)$ user-level DP learning algorithm $\mathcal{A}$ has an error at least*

$$\mathbb{E}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 = \Omega\Big(\frac{f}{n-f}\frac{G^2}{\mu^2}\Big), \ \mathbb{E}\|\nabla F_{\mathcal{G}}(\widehat{\boldsymbol{\theta}})\|_2^2 = \Omega\Big(\frac{f}{n-f}G^2\Big). \tag{4}$$

**Lemma 3** (Coupled)**.** *Suppose that the conditions in Theorem 1 hold with $1 \leq f < n/2$, the output $\widehat{\boldsymbol{\theta}}$ of a $(\epsilon, \delta)$ user-level DP learning algorithm $\mathcal{A}$ has an error at least*

$$\mathbb{E}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 = \Omega\Big(\frac{f}{n}\frac{dG^2\log(1/\delta)}{\mu^2\epsilon^2(n-f)^2}\Big), \ \mathbb{E}\|\nabla F_{\mathcal{G}}(\widehat{\boldsymbol{\theta}})\|_2^2 = \Omega\Big(\frac{f}{n}\frac{dG^2\log(1/\delta)}{\epsilon^2(n-f)^2}\Big). \tag{5}$$

Compared with (3), (5) inflates the DP error by a factor of Byzantine fraction $f/n$. The key step is to construct two adversarial configurations that are indistinguishable to any user-level DP algorithm, yet have well-separated optima. We partition the $n$ clients into three groups whose local dataset means point in opposing directions: $+G_0\mathbb{1}$, $-G_0\mathbb{1}$, and $\mathbf{0}$, where $G_0 = \Theta(G)$. We construct hard instances by combining these groups in different ways. By comparing each instance to a common reference, we show that no distributed algorithm can distinguish them, which reduces the distributed learning problem to a centralized, item-level DP mean-estimation task on $n - f$ items.

## 5 Algorithm: ByzULDP

In this section, we propose **Byz**antine-resilient **U**ser-**l**evel **D**ifferentially **P**rivate Learning (ByzULDP), formally presented in Algorithm 1. ByzULDP departs from FedAvg by tackling the dual challenges of Byzantine attacks and user-level DP together. At a high level, ByzULDP achieves Byzantine resilience through a nonlinear robust aggregator combined with a client-side momentum scheme that mitigates stochastic gradient noise. ByzULDP introduces a server-side momentum to regulate the global update. While the fault-free recipe for user-level central DP relies on injecting central noise directly into a linear simple mean [16, 45], this approach is insufficient for convergence with nonlinear robust aggregators; see Remark 2 for details.

A key innovation in ByzULDP is the use of server-side momentum to recalibrate each client's message, thereby controlling its global impact and reducing privacy sensitivity without sacrificing robustness. This recalibration is enabled *exclusively* by user-level DP; the gradient is the only viable data representation for item-level DP, lacking such architectural flexibility.

**Client-side.** Each honest client $i \in \mathcal{G}$ downloads $\boldsymbol{\theta}^t$ from the PS, computes a clipped stochastic gradient $\widehat{\boldsymbol{g}}_i^t = \texttt{clip}(\nabla \ell_i(\boldsymbol{\theta}^t; \xi_i^t), C)$, where $\texttt{clip}(\boldsymbol{v}, C) = \boldsymbol{v} \cdot \min\{1, C/\|\boldsymbol{v}\|_2\}$ (line 6), and updates its local momentum $\boldsymbol{m}_i^t = \alpha_t \boldsymbol{m}_i^{t-1} + (1-\alpha_t)\widehat{\boldsymbol{g}}_i^t$ (line 7). Clipping bounds each client's influence on the global model, which is important for our privacy analysis. Byzantine clients may report arbitrary values $*$ in place of $\boldsymbol{m}_i^t$. One-step gradient descent is commonly adopted in Byzantine-resilient FL [4, 5, 6, 7, 41, 42]; extending to multi-step local updates is left to future work.

**Server-side.** The PS computes $\boldsymbol{M}_i^t = \beta_t \widetilde{R}^{t-1} + (1-\beta_t)\widehat{\boldsymbol{m}}_i^t$ for each client $i \in [n]$ (lines 11-13), where $\widetilde{R}^{t-1}$ is the previous-round privatized aggregate and $\beta_t \in (0,1)$. Similar server-side momentum has been adopted in [20, 71] yet to mitigate the client-drift issue in non-Byzantine and non-private settings. Such a momentum mechanism enables contraction $\boldsymbol{M}_i^t - \boldsymbol{M}_j^t = (1-\beta_t)(\widehat{\boldsymbol{m}}_i^t - \widehat{\boldsymbol{m}}_j^t)$, which shrinks the distance between honest clients by a factor $1-\beta_t$ and reduces the privacy sensitivity of the global aggregation, formally shown in Proposition 1. Since $\widetilde{R}^{t-1}$ is already privatized, the DP post-processing property ensures that all subsequent computations remain private. In line 14, the PS aggregates the server-side momentum via a nonlinear robust aggregator $R(\cdot)$, defined in Definition 2.

**Definition 2** ($(f,\kappa)$-robust [4])**.** Let $f < n/2$ and $\kappa \geq 0$. An aggregation rule $R$ is said to be $(f,\kappa)$-robust if for any vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, and any set $S \subseteq [n]$ of size $n-f$, it holds that

$$\|R(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \overline{\boldsymbol{x}}_S\|_2^2 \leq \frac{\kappa}{|S|} \sum_{i \in S} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_S\|_2^2 ,$$

where $\overline{\boldsymbol{x}}_S = \frac{1}{|S|}\sum_{i \in S} \boldsymbol{x}_i$.

We refer to $\kappa$ as a robust coefficient. Robust aggregators aim to filter out the adversarial updates and have been widely adopted in Byzantine-resilient algorithm design [4, 17, 19, 41, 42, 61, 70, 73]. Our analysis can be easily extended to any such $R$. However, to obtain the optimal upper bound, we adopt coordinate-wise trimmed mean (CWTM) [73], defined in Definition 3, as the robust aggregator.

**Definition 3** (CWTM)**.** Let $\boldsymbol{x} \in \mathbb{R}^d$, we define $[\boldsymbol{x}]_r$ the $r$-th coordinate of $\boldsymbol{x}$. Given the input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, we let $\pi_r$ denote a permutation on $[n]$ that sorts the $r$-th coordinate of the input vector in non-decreasing order, i.e., $[\boldsymbol{x}_{\pi_r(1)}]_r \leq [\boldsymbol{x}_{\pi_r(2)}]_r \leq \ldots \leq [\boldsymbol{x}_{\pi_r(n)}]_r$. Then, the coordinate-wise trimmed mean $\texttt{CWTM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^d$ is defined coordinate-wise by

$$\left[\texttt{CWTM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\right]_r \triangleq \frac{1}{n-2f} \sum_{i=f+1}^{n-f} \left[\boldsymbol{x}_{\pi_r(i)}\right]_r .$$

Finally, the PS forms the privatized aggregate $\widetilde{R}^t = R^t + \boldsymbol{X}_{dp}^t$ with $\boldsymbol{X}_{dp}^t \sim \mathcal{N}(0, \sigma_{dp,t}\mathbf{I})$ and updates $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_t \widetilde{R}^t$. The value of $\sigma_{dp,t}$ will be specified in Section 6.2.

## 6 Convergence Analysis of ByzULDP (Upper Bound)

In this section, we analyze the convergence of ByzULDP. Missing proofs, intermediate results, and additional privacy analysis are deferred to the Appendix.

### 6.1 Additional Assumptions

Let $\mathcal{F}^t$ define the natural filtration generated by the randomness up to round $t$.

**Assumption 2.** Each local objective function $\nabla F_i(x)$ is $L$-Lipschitz, i.e.,

$$\|\nabla F_i(\boldsymbol{\theta}_1) - \nabla F_i(\boldsymbol{\theta}_2)\|_2 \leq L \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 , \ \forall \boldsymbol{\theta}_1, \ \boldsymbol{\theta}_2 \in \mathbb{R}^d, \text{ and } \forall\, i \in [n].$$

**Assumption 3.** Stochastic gradients $\nabla \ell_i(\boldsymbol{\theta}^t; \xi_i^t)$ are unbiased with bounded variance, i.e.,

$$\mathbb{E}\left[\nabla \ell_i(\boldsymbol{\theta}^t; \xi_i^t) \mid \mathcal{F}^t\right] = \nabla F_i(\boldsymbol{\theta}^t) \text{ and } \mathbb{E}\left[\left\|\nabla \ell_i(\boldsymbol{\theta}^t; \xi_i^t) - \nabla F_i(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \leq \sigma^2, \ \forall\, i \in [n].$$

**Assumption 4.** Stochastic gradients $\nabla \ell_i(\boldsymbol{\theta}, \xi)$ are uniformly bounded in $l_2$ norm $\|\nabla \ell_i(\boldsymbol{\theta}, \xi)\|_2 \leq C, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d$ and $\xi$.

Assumptions 2 and 3 are common in FL analysis [42, 46, 64]. Assumption 4 is widely adopted by DP algorithm analysis [7, 15, 16, 29, 45, 65]. We conjecture that Assumption 4 is imposed primarily for technical convenience and leave its relaxation to future work.

### 6.2 Privacy Results of ByzULDP

We start from the privacy analysis of ByzULDP. Since clients influence the global model through the robust aggregator, the key step is characterizing its user-level sensitivity. Definition 4 adapts the standard $\ell_2$ sensitivity [26] to the Byzantine setting. As Byzantine clients may behave arbitrarily, we measure sensitivity with respect to replacing a single honest client's update.

**Definition 4** (Byzantine-robust user-level sensitivity)**.** Let $[n] = \mathcal{G} \cup \mathcal{B}$ be a union of honest clients $\mathcal{G}$ and Byzantine clients $\mathcal{B}$ such that $|\mathcal{B}| \leq f$. The Byzantine-robust user-level sensitivity of a robust aggregator $R : \mathbb{R}^{n \times d} \to \mathbb{R}^d$, operating on $n$ inputs of which at most $f < n/2$ are Byzantine, is

$$\Delta(R) \triangleq \max_{\mathcal{B} \subset [n], |\mathcal{B}| \leq f} \sup_{\boldsymbol{X}_{\mathcal{B}}} \sup_{\boldsymbol{X}_{\mathcal{G}} \simeq \boldsymbol{X}_{\mathcal{G}'}} \|R(\boldsymbol{X}_{\mathcal{G}} \cup \boldsymbol{X}_{\mathcal{B}}) - R(\boldsymbol{X}_{\mathcal{G}'} \cup \boldsymbol{X}_{\mathcal{B}})\|_2,$$

where $\boldsymbol{X}_{\mathcal{G}}, \boldsymbol{X}_{\mathcal{B}}$ are input collections from the respective sets, and $\boldsymbol{X}_{\mathcal{G}} \simeq \mathbf{X}_{\mathcal{G}'}$ denotes honest client collections differing by exactly one input.

Our analysis extends to any robust aggregator in Definition 2, but CWTM enables us to attain the optimal convergence rate in Section 6.3.

**Lemma 4** (Per-dimension CWTM)**.** *Let $n \geq (2+\nu)f$, where $\nu > 0$. Fix a coordinate $r$, for any $a_1 \leq a_2 \leq \cdots \leq a_{n-1}$ in $\mathbb{R}^d$, let $\mathtt{TM}_r(h)$ define the coordinate-$r$ output of CWTM when inserting value $h$. For any $h, h' \in \mathbb{R}$, it holds that*

$$|\mathtt{TM}_r(h) - \mathtt{TM}_r(h')| \leq \frac{|h - h'|}{n} \sqrt{1 + \frac{4(2+\nu)^2}{\nu^2} \frac{f}{n}}.$$

The proof of Lemma 4 characterizes how inserting an honest client update affects the CWTM output by distinguishing whether the update lies inside or outside the trimmed set. Since it holds coordinate-wise, and honest updates are bounded by $C$ in $\ell_2$ norm, aggregating across coordinates yields a sensitivity without explicit dependence on the dimension $d$. This intuition is formalized in Proposition 1.

**Proposition 1** (CWTM sensitivity)**.** *Let $n \in \mathbb{N}$. If the input to the robust aggregator is $\boldsymbol{x}_i = \beta_t \widetilde{R}^{t-1} + (1-\beta_t)\boldsymbol{m}_i^t$, and $n \geq (2+\nu)f$, where $\|\boldsymbol{m}_i^t\|_2 \leq C$ for $i \in \mathcal{G}$ and $\nu > 0$, the Byzantine-robust user-level sensitivity of CWTM in round $t$ is*

$$\Delta_t(\mathtt{CWTM}) = \frac{2(1-\beta_t)C}{n} \sqrt{1 + \frac{4(2+\nu)^2}{\nu^2} \frac{f}{n}}.$$

**Theorem 2** (User-level DP noise)**.** *Let $\delta \in (0,1)$ and $\epsilon > 0$ such that $\epsilon \leq 8\log(1/\delta)$. Algorithm 1 is a $(\epsilon, \delta)$ user-level DP algorithm with CWTM as the robust aggregator when $\sigma_{dp,t}^2 = \frac{32(1-\beta_t)^2 C^2 T \log(1/\delta)}{n^2 \epsilon^2} \left(1 + \frac{4(2+\nu)^2}{\nu^2} \frac{f}{n}\right)$.*

Theorem 2 combines the per-round dimension-independent sensitivity in Proposition 1 with the cross-round Rényi DP composition [52], yielding the $\sigma_{dp,t}^2 \propto T$ scaling without explicit dependence on $d$. The noise variance additionally inherits a $(1-\beta_t)^2$ factor from server-momentum aggregation, since $\|\boldsymbol{M}_i^t - \boldsymbol{M}_j^t\|^2 = (1-\beta_t)^2 \|\boldsymbol{m}_i^t - \boldsymbol{m}_j^t\|^2$ for $i, j \in \mathcal{G}$. Such a dependence effectively reduces privacy sensitivity when $(1-\beta_t)$ is small and is crucial for convergence; see Remark 2 for details.

### 6.3 Convergence Results of ByzULDP

In this subsection, we provide convergence results for ByzULDP. The non-convex case uses constant $\alpha_t \equiv \alpha$ and $\eta_t \equiv \eta$, while the strongly-convex case keeps time-varying schedules. Let $\overline{\boldsymbol{m}}^t \triangleq \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \boldsymbol{m}_i^t$ and $F_{\mathcal{G}}^* = \min_{\boldsymbol{\theta}} F_{\mathcal{G}}(\boldsymbol{\theta})$. In Lemma 5, we show that client-side momentum mitigates the stochastic gradient noise. Take non-convex objectives as an example.

**Lemma 5** (Non-convex local momentum drift)**.** *Under Assumptions 1, 2, and 3. When $F_{\mathcal{G}}$ is non-convex, it holds for any $\alpha_t = \alpha \in (0,1)$ that*

$$\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 \le \left(3+\frac{3}{|\mathcal{G}|}\right)(1-\alpha^2)\sigma^2 + 3G^2 + 3B^2(1-\alpha)\sum_{r=0}^{t}\alpha^{t-r}\mathbb{E}\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\|_2^2. \quad (6)$$

(6) shows that the stochastic variance term scales as $(1-\alpha^2)$. Setting $1-\alpha = \Theta(\eta L)$ makes this factor small, effectively dampening its effect. The proof of Lemma 5 follows the standard Polyak momentum analysis by unrolling across all rounds [4, 7, 42]. An analogous strongly-convex result uses single-round unrolling; see Appendix G.2 for details. We establish our convergence bounds in Theorems 3 and 4 by constructing separate Lyapunov functions for strongly-convex and non-convex objectives to account for their different learning rate and momentum schedules.

**Theorem 3** (Strongly-convex, Informal)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. When $\eta_t = \frac{10}{\mu(t+a\frac{L}{\mu})}$ such that $a \ge 49000$, $1-\alpha_t = \Theta(\eta_t L)$, $1-\beta_t = \Theta(\eta_t L)$, $B^2 \le 1/(600\kappa)$, and $n \ge (2+\nu)f$, where $\nu > 0$, it holds for Algorithm 1 with* CWTM *as the robust aggregator that*

$$\mathbb{E}\left\|\boldsymbol{\theta}^T - \boldsymbol{\theta}^*\right\|_2^2 \lesssim \frac{\kappa G^2}{\mu^2} + \left(\frac{L}{\mu}\right)\frac{\log(1/\delta)C^2 d}{\mu^2\epsilon^2 n^2}\left(1+\frac{f}{n}\right) + L(1+\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^3 T} \quad (7)$$
$$+ \frac{L^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\mu^3 T^2} + \frac{L\sigma^2}{\mu^3 T^2 |\mathcal{G}|} + \frac{L\kappa}{\mu^3 T^2}\left[\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 + G^2\right].$$

**Theorem 4** (Non-convex, Informal)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. Let $1-\alpha = \Theta(\eta L)$, $1-\beta = \Theta(\eta L)$, such that $\eta < 1/(5100L)$ and $B^2 \le 1/(23625\kappa)$. When $n \ge (2+\nu)f$, where $\nu > 0$ is an absolute constant, it holds for Algorithm 1 with* CWTM *as the robust aggregator that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \lesssim \frac{\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*}{\eta T} + \frac{\eta L\sigma^2}{|\mathcal{G}|} + \eta L\kappa(1+\frac{1}{|\mathcal{G}|})\sigma^2 + \kappa G^2 + \frac{d\sigma_{dp}^2}{\eta L}$$
$$\overset{\sigma_{dp}=\Theta(1-\beta)}{\lesssim} \frac{\mathbb{E}\left[F(\boldsymbol{\theta}^0)\right] - F^*}{\eta T} + \eta L\sigma^2\Big(\frac{1}{|\mathcal{G}|} + \kappa(1+\frac{1}{|\mathcal{G}|})\Big) + \kappa G^2 + \frac{(1-\beta)^2 T}{\eta L}\frac{dC^2\log(\frac{1}{\delta})(1+\frac{f}{n})}{n^2\epsilon^2} \quad (8)$$
$$\overset{(1-\beta)=\Theta(\eta L)}{\lesssim} \frac{\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*}{\eta T} + \eta L\sigma^2(\frac{1}{|\mathcal{G}|} + \kappa(1+\frac{1}{|\mathcal{G}|})) + \kappa G^2 + \frac{\eta L T dC^2\log(1/\delta)(1+\frac{f}{n})}{n^2\epsilon^2}. \quad (9)$$

**Corollary 1** (Informal)**.** *Suppose that Assumptions 1, 2, 3, and 4 and the conditions in Theorem 4 hold. The choice of learning rate is formally defined in Appendix G.3.3. When $n \ge (2+\nu)f$, where $\nu > 0$ is an absolute constant, it holds for Algorithm 1 with* CWTM *as the robust aggregator that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \lesssim \kappa G^2 + \sqrt{L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)\frac{\log(1/\delta)C^2 d}{\epsilon^2 n^2}\left(1+\frac{f}{n}\right)}$$
$$+ \frac{\sigma}{\sqrt{T}}\sqrt{L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)\left(\frac{1}{|\mathcal{G}|} + \kappa\left(1+\frac{1}{|\mathcal{G}|}\right)\right)} + \frac{L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{T}.$$

**Remark 1** (**Optimal bound for strongly-convex objectives**)**.** We remark on Theorem 3, which establishes the convergence guarantees for a $\mu$ strongly-convex objective $F_{\mathcal{G}}$.

- **Order-optimal bound.** Theorem 3 establishes the full convergence bound for ByzULDP under a $\mu$-strongly convex objective. The colored terms in (7) are the non-vanishing errors.

– *DP terms.* The hard instances in Lemma 1 and in Lemma 3 satisfy $G = \Theta(C)$ when $f = \Theta(n)$, yielding a matching lower bound $\Omega\left(\frac{(1+f/n)dC^2\log(1/\delta)}{\mu^2\epsilon^2(n-f)^2}\right)$ up to constants and a factor $L/\mu$; see Appendix G.5 for details.
– *Byzantine-only term.* CWTM aggregator achieves a robustness coefficient of $\kappa = \Theta(f/n)$ [4], yielding a matching bound $\mathcal{O}(\kappa G^2/\mu^2) = \mathcal{O}(f/n \cdot G^2/\mu^2)$; see Corollary 4 for details.

Together, our upper bound is order-optimal with respect to the lower bound in Section 4. In the absence of the non-vanishing error terms, our $\mathcal{O}(1/T)$ rate for strongly-convex objectives also matches the optimal rate for stochastic first-order methods [57].

- **Privacy-utility tradeoff.** For the non-vanishing errors, we have both DP error and Byzantine residual terms, where the former dominates when $d \gtrsim n$. Fortunately, the large scale of FL systems can effectively mitigate the impact of DP noise on the final model utility [40]. On the other hand, the Byzantine residual error becomes negligible when the number of Byzantine clients $f$ is small or when the clients' data is not highly heterogeneous.
- **Comparisons with [8].** Recall that [8] guarantees user-level DP without a trusted server in the presence of Byzantine faults. We establish an improved convergence bound in the central DP setting. Ignoring the Byzantine-only error and $\mu$, we have the following comparisons:

$$\textbf{Ours}: \ \mathcal{O}\left(\frac{dC^2\log(1/\delta)}{n^2\epsilon^2} + \frac{fdC^2\log(1/\delta)}{n^3\epsilon^2}\right);$$
$$\textbf{[8]}: \ \mathcal{O}\left(\frac{dC^2\log(1/\delta)}{n^2\epsilon^2} + \frac{fdC^2\log(1/\delta)}{n^2\epsilon^2} + \frac{fC^2\log(1/\delta)}{n\epsilon^2}\right).$$

The two bounds coincide in the Byzantine-free case $f = 0$. Yet, our bound strictly improves whenever $f \geq 1$. The dominant Byzantine-DP term improves from $fd/n^2$ to $fd/n^3$ when $d \gtrsim n$, and from $f/n$ to $f/n^2$ when $d \lesssim n$. Moreover, their robust aggregator costs $\mathcal{O}(f(nd^2 + d^3))$ to compute, while CWTM runs efficiently in $\mathcal{O}(nd\log(n))$.

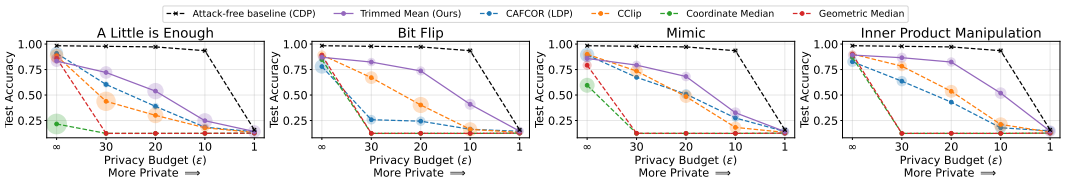
**Remark 2 (Results for non-convex objectives).** We remark on Theorem 4 and Corollary 1, which establish the convergence guarantees for a non-convex objective $F_{\mathcal{G}}$.

- **Necessity of server-side momentum.** Theorem 2 establishes that $\sigma_{dp} = \Theta(1-\beta)$. Substituting into Theorem 4 and setting $(1-\beta) = \Theta(\eta L)$ reduces the DP error to a $\mathcal{O}(\eta LT)$ scaling. Without server momentum ($\beta = 0$), this term scales as $\mathcal{O}(T/(\eta L))$ and diverges, e.g., as $\mathcal{O}(T^{3/2})$ when $\eta = \Theta(1/\sqrt{T})$. Thus, server momentum is a structural requirement for a non-trivial DP guarantee.
- **Non-convex bound.** Without the non-vanishing error terms, our $\mathcal{O}(1/\sqrt{T})$ convergence rate for non-convex objectives matches the optimal bound for stochastic first-order methods [9]. Similar to Remark 1, CWTM achieves an order-optimal Byzantine residual error of $\mathcal{O}(\kappa G^2) = \mathcal{O}(f/n \cdot G^2)$. However, a gap remains regarding the DP residual error. Ignoring the Byzantine residual, (9) scales as $\mathcal{O}(\frac{1}{\eta T} + \eta L(1+T))$. Applying the AM-GM inequality yields an optimal learning rate of $\eta = \Theta(\frac{1}{\sqrt{LT(1+T)}})$, subject to the constraint $\eta < 1/(5100L)$. However, this optimal $\eta$ inevitably couples the DP noise term with the initialization error under a square root in Corollary 1, preventing a strict match with our lower bound in Section 4. The appearance of the DP noise term inside a square root is widely observed in the DP non-convex learning literature; see, e.g., [7, 8, 10, 14, 45].

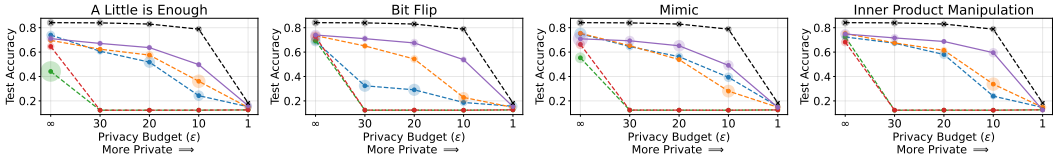
## 7 Numerical Experiments

**Overview.** In this section, we evaluate the performance of ByzULDP with CWTM against other state-of-the-art (SOTA) robust aggregators. The hyperparameter setups and additional experiment details are deferred to Appendix H. We consider an FL system of 1000 clients with 250 clients corrupted by a variety of Byzantine adversaries. The image classification tasks use CNNs and are based on the MNIST [44] and the Fashion-MNIST [68] datasets. While these tasks are relatively simple in non-adversarial settings, they become significantly more challenging in the presence of Byzantine clients [36, 42] and DP [7, 8]. All of them include 10 classes of images of different categories. We partition the dataset to clients such that each client holds two classes of images.

**Attacks and Benchmark algorithms.** We evaluate four SOTA attacks, including a little is enough [13], bit-flip attack [3], mimic attack [41], and inner-product manipulation attack [69]. The detailed attack setups are deferred to Appendix H. Our ByzULDP is a user-level DP framework, compatible




(a) Test accuracy results on the MNIST dataset.




(b) Test accuracy results on the Fashion-MNIST dataset.

Figure 1: Test accuracy results and comparisons on real-world datasets. The results are averaged over the last 50 rounds across 4 different random seeds. All the experiments run for a total of 800 rounds. The radius of the shaded circle denotes the standard deviation at a given privacy budget $\epsilon$.

with any robust aggregator satisfying Definition 2, and it attains the optimal rate when instantiated with `CWTM`. Other than `CWTM`, the SOTA robust aggregators include geometric median [19], coordinate-wise median [73], and `CClip` [42]. We additionally compare against two baselines: a simple mean as an aggregator in the attack-free setting ($f = 0$), and `CAFCOR` [8]. `CAFCOR` is a local DP algorithm that guarantees user-level DP without a trusted server.

**Discussions.** Fig. 1 displays the final performance under different privacy budgets. Observations are mostly consistent between the two datasets. Compared with the attack-free baseline, a performance gap emerges due to the presence of Byzantine clients. When user-level DP protection is not imposed, `CWTM` is among the top-performing robust aggregators. The other strong performers are `CClip` and `CAFCOR`, where the former relies on a clipping threshold that requires careful tuning, and the latter incurs high computational overhead (see Remark 1 for details). As a byproduct, we find that server-side momentum also improves the empirical performance of robust aggregators. Intuitively, the shared anchor $\beta_t \widetilde{R}^{t-1}$ pulls all clients' updates toward a common direction, thereby strengthening their concentration and improving the robust aggregator's effectiveness against Byzantine clients. We defer the details to Table 6 in Appendix H.

Under user-level DP protections, `CWTM` consistently outperforms the other baselines in the presence of Byzantine attacks. We attribute this to its dimension-free sensitivity, which yields the optimal convergence rate. Stronger privacy requires higher DP noise, degrading performance. Each nonlinear robust aggregator induces a different Byzantine-robust user-level sensitivity (Definition 4) due to the differences in structure; see Table 4 for details. Since such sensitivity determines the variance of the DP noise injected at the server, it dictates the signal-to-noise ratio and directly shapes the privacy-utility tradeoff. Among the robust aggregators within our `ByzULDP` framework, only `CClip` and `CWTM` achieve meaningful test accuracy, and they are also the only two whose sensitivity denominators grow with the number of clients. However, the sensitivity of `CClip` is proportional to a heuristically chosen clipping threshold $\tau$, typically set as $\tau = 10/(1-\alpha)$ [41, 42], which can be significantly larger than the corresponding $C$ in `CWTM`. `CAFCOR` adopts an interesting correlated-noise mechanism to inject DP noise at clients and lies outside our `ByzULDP` framework. Nevertheless, it still falls behind, which we attribute to its suboptimal convergence rate; see Remark 1 for details. We also show that server-side momentum is essential to guarantee performance. The results are deferred to Fig. 3 in Appendix H.

## 8 Conclusion

In this paper, we study Byzantine-resilient federated learning with user-level DP. We have established an information-theoretic lower bound for any Byzantine-resilient empirical risk minimization problem under user-level DP and heterogeneous clients. We have proposed `ByzULDP`, a central DP algorithm featuring two-level momentum across client and server. We have shown that `ByzULDP` achieves the order-optimal convergence bound for strongly-convex objectives. Numerical experiments have corroborated our analysis.

## Acknowledgments and Disclosure of Funding

We gratefully acknowledge the support from the National Science Foundation under grants 2106891, 2107062, and the National Science Foundation CAREER award under grant 2340482. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## References

[1] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Robust testing and estimation under manipulation attacks. In *International Conference on Machine Learning*, pages 43–53. PMLR, 2021.

[2] Sushant Agarwal, Gautam Kamath, Mahbod Majid, Argyris Mouzakis, Rose Silver, and Jonathan Ullman. Private mean estimation with person-level differential privacy. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2819–2880. SIAM, 2025.

[3] Zeyuan Allen-Zhu, Faeze Ebrahimianghazani, Jerry Li, and Dan Alistarh. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

[4] Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 1232–1300. PMLR, 2023.

[5] Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Ahmed Jellouli, Geovani Rizk, and John Stephan. Adaptive gradient clipping for robust federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and Geovani Rizk. Robust distributed learning: Tight error bounds and breakdown point under data heterogeneity. *Advances in Neural Information Processing Systems*, 36:45744–45776, 2023.

[7] Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. On the privacy-robustness-utility trilemma in distributed learning. In *International Conference on Machine Learning*, pages 569–626. PMLR, 2023.

[8] Youssef Allouah, Rachid Guerraoui, and John Stephan. Towards trustworthy federated learning with untrusted participants. In *Forty-second International Conference on Machine Learning*, 2025.

[9] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

[10] Raman Arora, Raef Bassily, Tomás González, Cristóbal A Guzmán, Michael Menart, and Enayat Ullah. Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*, pages 1060–1092. PMLR, 2023.

[11] Ghazi Badih, Kumar Ravi, and Manurangsi Pasin. User-level private learning via correlated sampling. *Advances in Neural Information Processing Systems*, 1, 2021.

[12] Li Bai, Haibo Hu, Qingqing Ye, Haoyang Li, Leixia Wang, and Jianliang Xu. Membership inference attacks and defenses in federated learning: A survey. *ACM Computing Surveys*, 57(4):1–35, 2024.

[13] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021.

[15] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

[16] Raef Bassily and Ziteng Sun. User-level private stochastic convex optimization with optimal rates. In *International Conference on Machine Learning*, pages 1838–1851. PMLR, 2023.

[17] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

[18] Jin-Hua Chen, Min-Rong Chen, Guo-Qiang Zeng, and Jia-Si Weng. Bdfl: a byzantine-fault-tolerance decentralized federated learning method for autonomous vehicle. *IEEE Transactions on Vehicular Technology*, 70(9):8639–8652, 2021.

[19] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

[20] Ziheng Cheng, Xinmeng Huang, Pengfei Wu, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. In *The Twelfth International Conference on Learning Representations*, 2024.

[21] Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 883–900. IEEE, 2021.

[22] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 375–403. Springer, 2019.

[23] Antoine Choffrut, Rachid Guerraoui, Rafael Pinot, Renaud Sirdey, John Stephan, and Martin Zuber. Towards practical homomorphic aggregation in byzantine-resilient distributed learning. In *Proceedings of the 25th International Middleware Conference*, pages 431–444, 2024.

[24] Henry Corrigan-Gibbs and Dan Boneh. Prio: Private, robust, and scalable computation of aggregate statistics. In *14th USENIX symposium on networked systems design and implementation (NSDI 17)*, pages 259–282, 2017.

[25] John Duchi, Martin J Wainwright, and Michael I Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. *Advances in Neural Information Processing Systems*, 26, 2013.

[26] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

[27] Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pages 6246–6283. PMLR, 2022.

[28] Jiashi Feng, Huan Xu, and Shie Mannor. Distributed robust learning. *arXiv preprint arXiv:1409.5937*, 2014.

[29] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. *Advances in Neural Information Processing Systems*, 34:20172–20184, 2021.

[30] Bimal Ghimire and Danda B Rawat. Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things. *IEEE Internet of Things Journal*, 9(11):8229–8249, 2022.

[31] Avishek Ghosh, Raj Kumar Maity, Swanand Kadhe, Arya Mazumdar, and Kannan Ramchandran. Communication efficient and byzantine tolerant distributed learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2545–2550. IEEE, 2020.

[32] Antonious M Girgis, Deepesh Data, and Suhas Diggavi. Distributed user-level private mean estimation. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2196–2201. IEEE, 2022.

[33] Xiaolan Gu, Ming Li, and Li Xiong. {DP-BREM}:{Differentially-Private} and {Byzantine-Robust} federated learning with client momentum. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3065–3082, 2025.

[34] Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, Sébastien Rouault, and John Stephan. Differential privacy and byzantine resilience in sgd: Do they add up? In *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, pages 391–401, 2021.

[35] Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *International conference on machine learning*, pages 3521–3530. PMLR, 2018.

[36] Rachid Guerraoui, Sébastien Rouault, et al. Distributed momentum for byzantine-resilient stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

[37] Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539, 2020.

[38] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, 34:7232–7241, 2021.

[39] Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. Distributed learning without distress: Privacy-preserving empirical risk minimization. *Advances in neural information processing systems*, 31, 2018.

[40] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[41] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2021.

[42] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International conference on machine learning*, pages 5311–5319. PMLR, 2021.

[43] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[44] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[45] Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

[46] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks, April 2020. arXiv:1812.06127 [cs, stat].

[47] Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. Fednlp: Benchmarking federated learning methods for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 157–175, 2022.

[48] Daogao Liu and Hilal Asi. User-level differentially private stochastic convex optimization: Efficient algorithms with optimal rates. In *International Conference on Artificial Intelligence and Statistics*, pages 4240–4248. PMLR, 2024.

[49] Yi Liu, Xiaohan Bi, Lei Li, Sishuo Chen, Wenkai Yang, and Xu Sun. Communication efficient federated learning for multilingual neural machine translation with adapter. In *Findings of the association for computational linguistics: ACL 2023*, pages 5315–5328, 2023.

[50] Nancy A Lynch. *Distributed algorithms*. Elsevier, 1996.

[51] Xu Ma, Xiaoqian Sun, Yuduo Wu, Zheli Liu, Xiaofeng Chen, and Changyu Dong. Differentially private byzantine-robust federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(12):3690–3701, 2022.

[52] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[53] Viktor Moskvoretskii, Nazarii Tupitsa, Chris Biemann, Samuel Horváth, Eduard Gorbunov, and Irina Nikishina. Low-resource machine translation through the lens of personalized federated learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8806–8825, 2024.

[54] Shyam Narayanan, Vahab Mirrokni, and Hossein Esfandiari. Tight and robust private mean estimation with few users. In *International Conference on Machine Learning*, pages 16383–16412. PMLR, 2022.

[55] Thien Duc Nguyen, Samuel Marchal, Markus Miettinen, Hossein Fereidooni, N Asokan, and Ahmad-Reza Sadeghi. Dïot: A federated self-learning anomaly detection system for iot. In *2019 IEEE 39th International conference on distributed computing systems (ICDCS)*, pages 756–767. IEEE, 2019.

[56] Muzi Peng, Jiangwei Wang, Dongjin Song, Fei Miao, and Lili Su. Privacy-preserving and uncertainty-aware federated trajectory prediction for connected autonomous vehicles. In *The 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)*. IEEE/RSJ, 2023.

[57] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress.

[58] Ethan Rathbun, Christopher Amato, and Alina Oprea. Sleepernets: Universal backdoor poisoning attacks against reinforcement learning agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 111994–112024. Curran Associates, Inc., 2024.

[59] Junshuai Song, Zhao Li, Zehong Hu, Yucheng Wu, Zhenpeng Li, Jian Li, and Jun Gao. Poisonrec: an adaptive data poisoning framework for attacking black-box recommender systems. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 157–168. IEEE, 2020.

[60] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2016.

[61] Lili Su and Nitin H Vaidya. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM symposium on principles of distributed computing*, pages 425–434, 2016.

[62] Lili Su and Nitin H Vaidya. Byzantine-resilient multiagent optimization. *IEEE Transactions on Automatic Control*, 66(5):2227–2233, 2020.

[63] Shreyas Sundaram and Bahman Gharesifard. Consensus-based distributed optimization with malicious nodes. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 244–249. IEEE, 2015.

[64] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization, July 2020. arXiv:2007.07481.

[65] Puyu Wang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private sgd with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56:306–336, 2022.

[66] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020.

[67] Ming Xiang and Lili Su. Distributed non-convex optimization with one-bit compressors on heterogeneous data: Efficient and resilient algorithms. *arXiv preprint arXiv:2210.00665*, 2022.

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

[69] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in artificial intelligence*, pages 261–270. PMLR, 2020.

[70] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International conference on machine learning*, pages 6893–6901. PMLR, 2019.

[71] Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

[72] Han Yang, Dongbing Gu, and Jianhua He. A robust and efficient federated learning algorithm against adaptive model poisoning attacks. *IEEE Internet of Things Journal*, 11(9):16289–16302, 2024.

[73] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pages 5650–5659. Pmlr, 2018.

[74] Tengchan Zeng, Omid Semiari, Mingzhe Chen, Walid Saad, and Mehdi Bennis. Federated learning on the road autonomous controller design for connected and autonomous vehicles. *IEEE Transactions on Wireless Communications*, 21(12):10407–10423, 2022.

[75] Tuo Zhang, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, and A Salman Avestimehr. Federated learning for the internet of things: Applications, challenges, and opportunities. *IEEE Internet of Things Magazine*, 5(1):24–29, 2022.

[76] Heng Zhu and Qing Ling. Bridging differential privacy and byzantine-robustness via model aggregation. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2427–2433. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

## Appendices

Here, we provide an overview of the Appendix. In particular, the proofs of the main results are presented and backed by supporting intermediate results.


**A Limitations** **17**

**B Broader Impacts** **17**

**C Nomenclatures** **17**

**D Useful Inequalities** **18**

**E Lower Bound** **20**
- E.1 Proof of Lemma 1 (Non-Byzantine) . . . 20
- E.2 Proof of Lemma 2 (Non-private) . . . 21
- E.3 Proof of Lemma 3 (Coupled) . . . 23
- E.4 Proof of Theorem 1 . . . 27
- E.5 Supporting Lemmas of Appendix E . . . 27

**F Privacy Analysis** **29**
- F.1 Proof of Lemma 4 . . . 29
- F.2 Proof of Proposition 1 . . . 30
- F.3 Supporting Rényi DP Analysis . . . 31
- F.4 Proof of Theorem 2 . . . 32
- F.5 Byzantine-robust User-level Sensitivity Analysis of Additional Robust Aggregators 32

**G Convergence Analysis** **35**
- G.1 Shared Intermediate Results . . . 35
  - G.1.1 Proof of Descent Lemma . . . 35
  - G.1.2 Proof of Global Momentum Estimation Error . . . 36
  - G.1.3 Proof of Client-side Momentum Estimation Error . . . 40
- G.2 Strongly-convex objectives . . . 41
  - G.2.1 Proof of Local Momentum Drift . . . 41
  - G.2.2 Proof of Theorem 3 . . . 43
- G.3 Non-convex objectives . . . 51
  - G.3.1 Proof of Lemma 5 . . . 51
  - G.3.2 Proof of Theorem 4 . . . 52
  - G.3.3 Proof of Corollary 1 . . . 57
- G.4 Supporting Results for the Optimality of `CWTM` . . . 59
- G.5 Conditions of matching optimality . . . 59
- G.6 Optimality of `CWTM` . . . 60

**H Experiment Details** **63**
- H.1 Experimental setups . . . 63
- H.2 Byzantine Attack Details . . . 63
- H.3 Additional results . . . 64

**I NeurIPS Paper Checklist** **65**

# A Limitations

The limitations of our work are twofold:

1. We assume a trustworthy parameter server. While this is the standard setting in regulated domains and achieves the minimax-optimal convergence rate, it only applies to central DP. Recent work by Allouah et al. [8] makes important progress in user-level DP without a trusted server in the presence of Byzantine adversaries. It is interesting to explore the minimax-optimal convergence rate when the server cannot be trusted.
2. We assume that stochastic gradients are uniformly bounded for technical convenience. Although this assumption is widely adopted in the DP literature [7, 15, 16, 29, 45, 65], real-world data distributions may exhibit heavy tails. Relaxing this assumption but applying gradient clipping can introduce bias that complicates the convergence analysis. It is interesting to explore the trade-off between the clipping threshold, privacy, and utility.

# B Broader Impacts

Federated learning has reshaped distributed ML/AI in recent years and has driven commercial applications such as autonomous vehicles, the Internet of Things, and natural language processing. Our paper focuses on the fault-tolerance aspect of real-world federated learning systems in the presence of both malicious clients and privacy leakage. Our work has substantially advanced the theory and algorithms for federated learning by bringing together insights from statistics, optimization, distributed computing and engineering practices. In addition, our research is important for federated learning systems to expand their outreach to more adversarial deployment environments. We are unaware of any potential negative social impacts of our work.

# C Nomenclatures

In this section, we provide the notations and nomenclatures used throughout our proofs for a comprehensive presentation. It is worth noting that all notation has been properly introduced before its first use.

Table 1: Notation table

| | |
|---|---|
| $\|\boldsymbol{v}\|_2$ | The $l_2$ norm of a given vector $\boldsymbol{v}$. |
| $\mathcal{F}^t$ | The sigma algebra generated by randomness up to round $t$. |
| $\mathbb{R}^d$ | A $d$-dimensional vector space, where $d$ denotes the dimension. |
| $[n]$ | A set $\{k \mid k \in \mathbb{N}, k \in [1, n]\}$. |
| $\mathcal{G}$ | A set of honest clients. |
| $\mathcal{B}$ | A set of clients suffering from Byzantine attacks $\mathcal{B} \triangleq [n] \setminus \mathcal{G}$. |
| $\mathbb{1}$ | An all one vector $\mathbb{1} \in \mathbb{R}^d$. |
| $\lesssim$ | $f(n) \lesssim g(n)$, if there exists a constant $c_{\mathcal{O}} > 0$ and an integer $n_0 \in \mathbb{N}$, $f(n) \leq c_{\mathcal{O}} g(n)$ for all $n \geq n_0$. |
| $\gtrsim$ | $f(n) \gtrsim g(n)$, if there exists a constant $c_{\Omega} > 0$ and an integer $n_0 \in \mathbb{N}$, $f(n) \geq c_{\Omega} g(n)$ for all $n \geq n_0$. |
| $\mathcal{O}$ | $f(n) = \mathcal{O}(g(n))$, if there exists a constant $c_{\mathcal{O}} > 0$ and an integer $n_0 \in \mathbb{N}$, $f(n) \leq c_{\mathcal{O}} g(n)$ for all $n \geq n_0$. |
| $\Theta$ | $f(n) = \Theta(g(n))$, if there exists a constant $c_{\Theta} > 0$ and an integer $n_0 \in \mathbb{N}$, $f(n) = c_{\Theta} g(n)$ for all $n \geq n_0$. |
| $\Omega$ | $f(n) = \Omega(g(n))$, if there exists a constant $c_{\Omega} > 0$ and an integer $n_0 \in \mathbb{N}$, $f(n) \geq c_{\Omega} g(n)$ for all $n \geq n_0$. |

Table 2: Algorithmic nomenclature table

| | |
|---|---|
| $\boldsymbol{\theta}^t$ | The global model at the **beginning** of round $t$ in Algorithm 1. |
| $F_i(\boldsymbol{\theta})$ | The local objective function at client $i$. |
| $F_{\mathcal{G}}(\boldsymbol{\theta})$ | The global objective function defined in (1): $F_{\mathcal{G}}(\boldsymbol{\theta}) \triangleq \sum_{i\in\mathcal{G}} F_i(\boldsymbol{\theta})/\left\lvert\mathcal{G}\right\rvert$. |
| $\nabla\ell_i(\boldsymbol{\theta})$ | The local stochastic gradient function at client $i$ taken with respect to $\boldsymbol{\theta}$. |
| $\widehat{\boldsymbol{g}}_i^t$ | The clipped local stochastic gradient function at client $i$ taken with respect to $\boldsymbol{\theta}$. $\widehat{\boldsymbol{g}}_i^t \leftarrow \texttt{clip}(\nabla\ell_i(\boldsymbol{\theta}^t;\xi_i^t), C)$, where $\texttt{clip}(\boldsymbol{v}, C) = \boldsymbol{v}\cdot\min\left\{1, C/\left\lVert\boldsymbol{v}\right\rVert_2\right\}$. |
| $\boldsymbol{m}_i^t$ | The client-side momentum at client $i$ in round $t$. |
| $\overline{\boldsymbol{m}}^t$ | An average over the client-side momentum for clients in $\mathcal{G}$ in round $t$. |
| $\widehat{\boldsymbol{m}}_i^t$ | The local update from client $i$ to the PS in round $t$. It duplicates $\boldsymbol{m}_i^t$ for honest clients in $\mathcal{G}$ but can be arbitrary for Byzantine clients. |
| $\boldsymbol{M}_i^t$ | The server-side momentum in round $t$ for client $i$. |
| $\overline{\boldsymbol{M}}^t$ | An average over the server-side momentum for clients in $\mathcal{G}$ in round $t$, i.e., $\overline{\boldsymbol{M}}^t \triangleq \sum_{i\in\mathcal{G}} \boldsymbol{M}_i^t/\left\lvert\mathcal{G}\right\rvert$. |
| $\nabla F_i(\boldsymbol{\theta})$ | The local full-batch gradient function at client $i$ taken with respect to $\boldsymbol{\theta}$. |
| $\nabla F_{\mathcal{G}}(\boldsymbol{\theta})$ | The global full-batch gradient function taken with respect to $\boldsymbol{\theta}$. |
| $R^t$ | The robust aggregation at the PS in round $t$ by a robust aggregator $R$. |
| $\boldsymbol{X}_{dp}^t$ | The DP noise at the PS in round $t$. |
| $\widetilde{R}^t$ | The noisy robust aggregation at the PS in round $t$ is a sum of a robust aggregation $R$ and a DP noise $\boldsymbol{X}_{dp}^t$. |
| $D_i$ | Client $i$'s local empirical data distribution. |
| $\xi_i$ | An **independent** stochastic sample drawn from client $i$'s local distribution $D_i$. |

Table 3: Variable table

| | |
|---|---|
| $L$ | Lipschitz constant in Assumption 2. |
| $\sigma^2$ | The upper bound of the stochastic gradient variance. |
| $\sigma_{dp,t}^2$ | The variance of DP noise in round $t$. |
| $(B,\ G)$ | Parameters that capture the averaged local-global gradient dissimilarity. |
| $C$ | A uniform upper bound on the stochastic gradient norm. |
| $n$ | The number of clients in the FL system. |
| $m$ | The number of local data points at each client in the FL system. |

# D Useful Inequalities

For completeness and for ease of exposition, we present some common inequalities that will be frequently used in our proofs. The followings hold for any $\boldsymbol{a}_i \in \mathbb{R}^d$ and any $i \in [n]$.

1. Jensen's inequality.

$$\left\lVert\frac{1}{n}\sum_{i=1}^n \boldsymbol{a}_i\right\rVert_2^2 \leq \frac{1}{n}\sum_{i=1}^n \left\lVert\boldsymbol{a}_i\right\rVert_2^2 \quad \text{and} \quad \left\lVert\sum_{i=1}^n \boldsymbol{a}_i\right\rVert_2^2 \leq n\sum_{i=1}^n \left\lVert\boldsymbol{a}_i\right\rVert_2^2. \tag{10}$$

2. Young's inequality (a.k.a. Peter-Paul inequality).

$$\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle \leq \frac{\|\boldsymbol{a}_1\|_2^2}{2\epsilon} + \frac{\epsilon \|\boldsymbol{a}_2\|_2^2}{2}, \quad \text{for any } \epsilon > 0. \tag{11}$$

Equivalently, we have

$$\begin{aligned} \|\boldsymbol{a}_1 + \boldsymbol{a}_2\|_2^2 &= \|\boldsymbol{a}_1\|_2^2 + \|\boldsymbol{a}_2\|_2^2 + 2\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle \\ &\leq \left(1 + \frac{1}{\epsilon}\right) \|\boldsymbol{a}_1\|_2^2 + (1+\epsilon) \|\boldsymbol{a}_2\|_2^2, \quad \text{for any } \epsilon > 0. \end{aligned} \tag{12}$$

3. Cauchy-Schwartz inequality.

$$\langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle \leq \|\boldsymbol{a}_1\|_2 \|\boldsymbol{a}_2\|_2 . \tag{13}$$

4. Smoothness corollary. Given Assumption 2, it holds that

$$\begin{aligned} F(\boldsymbol{a}_1) - F(\boldsymbol{a}_2) &= \left\langle \boldsymbol{a}_1 - \boldsymbol{a}_2, \int_0^1 \nabla F(\boldsymbol{a}_2 + \tau(\boldsymbol{a}_1 - \boldsymbol{a}_2)) \mathrm{d}\tau \right\rangle \\ &= \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + \int_0^1 \langle \boldsymbol{a}_1 - \boldsymbol{a}_2, \nabla F(\boldsymbol{a}_2 + \tau(\boldsymbol{a}_1 - \boldsymbol{a}_2)) - \nabla F(\boldsymbol{a}_2) \rangle \, \mathrm{d}\tau \\ &\overset{(a)}{\leq} \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + L \int_0^1 \tau \|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2 \|(\boldsymbol{a}_1 - \boldsymbol{a}_2)\|_2 \, \mathrm{d}\tau \\ &\leq \langle \nabla F(\boldsymbol{a}_2), \boldsymbol{a}_1 - \boldsymbol{a}_2 \rangle + \frac{L}{2} \|\boldsymbol{a}_1 - \boldsymbol{a}_2\|_2^2, \end{aligned} \tag{14}$$

where $(a)$ follows from Cauchy-Schwartz inequality and Assumption 2.

# E Lower Bound

## E.1 Proof of Lemma 1 (Non-Byzantine)

**Lemma E.1.** *Suppose that $\epsilon, \delta \in (0,1)$ and $1/2^{n^{1-\gamma}} \leq \delta \leq 1/n^{1+\gamma}$, where $\gamma \in (0,1)$. For any given $L, B \geq 0$ and $G > 0$, there exists a collection of $n$ functions $\{F_i\}_{i\in[n]}$, each defined over a dataset of $m$ samples, that are all i) $L$-smooth, ii) $\mu$-strongly convex with $\mu = \Theta(L)$, and iii) satisfy Assumption 1 with respect to $F$. If $\mathcal{A}$ is $(\epsilon, \delta)$-DP, then the output $\widehat{\boldsymbol{\theta}}$ of a $(\epsilon, \delta)$ user-level DP learning algorithm $\mathcal{A}$ has an error at least*

$$\mathbb{E}\Big[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2\Big] = \Omega\Big(\frac{dG^2\log(1/\delta)}{\mu^2\epsilon^2 n^2}\Big),\ \mathbb{E}\Big[\|\nabla F(\widehat{\boldsymbol{\theta}})\|_2^2\Big] = \Omega\Big(\frac{dG^2\log(1/\delta)}{\epsilon^2 n^2}\Big). \tag{15}$$

**Proof of Lemma 1**

In this proof, we show the lower bound when we do not have any Byzantine adversaries, i.e., $f = 0$. It is worth noting that our bound remains valid when we have $f > 0$. We focus on the case of $G > 0$ only. We first construct $n$ local datasets, each consisting of $m$ elements: Each client $i$ holds a local dataset $\mathcal{D}_i$ of $m$ ordered elements

$$\mathcal{D}_i = \left\{ \frac{\boldsymbol{x}_i^1 G_0}{L\sqrt{d}}, \ldots, \frac{\boldsymbol{x}_i^m G_0}{L\sqrt{d}} \right\},$$

where $\boldsymbol{x}_i^j \in \{\pm 1\}^d$ for $j = 1, \cdots, m$ and $i \in [n]$, and $G_0 = \frac{G}{2}$, such that

$$\sum_{i=1}^n \boldsymbol{x}_i^1 = \cdots = \sum_{i=1}^n \boldsymbol{x}_i^j = \cdots = \sum_{i=1}^n \boldsymbol{x}_i^m.$$

We define local functions based on these datasets as follows: for any given $\mu \leq L$,

$$F_i(\boldsymbol{\theta}) = \frac{\mu}{2}\left\|\boldsymbol{\theta} - \frac{1}{m}\frac{G_0}{L\sqrt{d}}\sum_{j=1}^m \boldsymbol{x}_i^j\right\|_2^2.$$

The resulting global objective is

$$F(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n F_i(\boldsymbol{\theta}).$$

The optimality gaps $\|\nabla F(\widehat{\boldsymbol{\theta}})\|_2^2$ and $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ can be characterized as follows:

$$\left\|\nabla F(\widehat{\boldsymbol{\theta}})\right\|_2^2 = \mu^2\left\|\boldsymbol{\theta} - \frac{G_0}{L\sqrt{d}}\frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m \boldsymbol{x}_i^j\right\|_2^2, \text{ and } \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 = \left\|\boldsymbol{\theta} - \frac{G_0}{L\sqrt{d}}\frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m \boldsymbol{x}_i^j\right\|_2^2.$$

Next, we verify that the constructed functions collectively satisfy the $(B, G)$-conditions. It can be readily seen that

$$\nabla F_i(\boldsymbol{\theta}) = \mu\boldsymbol{\theta} - \frac{\mu G_0}{L\sqrt{d}}\frac{1}{m}\sum_{j=1}^m \boldsymbol{x}_i^j, \text{ and } \nabla F(\boldsymbol{\theta}) = \mu\boldsymbol{\theta} - \frac{\mu G_0}{L\sqrt{d}}\frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m \boldsymbol{x}_i^j.$$

It follows that

$$\frac{1}{n}\sum_{i=1}^n \|\nabla F_i(\boldsymbol{\theta}) - \nabla F(\boldsymbol{\theta})\|_2^2 = \frac{\mu^2 G_0^2}{L^2 d}\sum_{i=1}^n\left\|\frac{1}{m}\sum_{j=1}^m \boldsymbol{x}_i^j - \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \boldsymbol{x}_i^j\right\|_2^2 \leq 4G_0^2 = G^2.$$

Thus, our construction satisfies the $(B, G)$-condition for arbitrary given $B \geq 0$ and $G > 0$.

Different from the item-level differential privacy guarantees in Definition 2.3 in [7], Definition 1 requires the indistinguishability of clients' local datasets. At a high level, Definition 1 does not require each client to be differentially private with respect to its own datasets.

Recall that $\sum_{i=1}^n \boldsymbol{x}_i^1 = \cdots = \sum_{i=1}^n \boldsymbol{x}_i^j = \cdots = \sum_{i=1}^n \boldsymbol{x}_i^m$, which results in

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{x}_i^j = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^j = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^j, \quad \forall\, j \in [m].$$

Thus, when $\mu = \Theta(L)$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F(\widehat{\boldsymbol{\theta}})\right\|_2^2\right] &= \mu^2\left\|\widehat{\boldsymbol{\theta}} - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\boldsymbol{x}_i^j G_0}{L\sqrt{d}}\right\|_2^2 \\
&= \mu^2\frac{1}{m}\left\|\mathbb{1}_m \otimes \widehat{\boldsymbol{\theta}} - \frac{G_0}{L\sqrt{d}}\begin{bmatrix}\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^1 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^m\end{bmatrix}\right\|_2^2 \\
&\overset{(a)}{=} \Omega\left(\mu^2\frac{1}{m}\frac{G_0^2}{L^2}\cdot\frac{md\log(1/\delta)}{\epsilon^2 n^2}\right) \overset{(b)}{=} \Omega\left(\frac{d\log(1/\delta)}{\epsilon^2 n^2}G^2\right),
\end{aligned}$$

where equality $(a)$ follows by invoking Corollary 2, and equality $(b)$ follows from $\mu = \Theta(L)$ and $G^2 = 4G_0^2$. Similarly,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2\right] &= \left\|\widehat{\boldsymbol{\theta}} - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\boldsymbol{x}_i^j G_0}{L\sqrt{d}}\right\|_2^2 = \Omega\left(\frac{d\log(1/\delta)}{\epsilon^2 n^2}\frac{G_0^2}{L^2}\right) \\
&= \Omega\left(\frac{d\log(1/\delta)}{\epsilon^2 n^2}\frac{G^2}{\mu^2}\right).
\end{aligned}$$

### E.2 Proof of Lemma 2 (Non-private)

**Lemma E.2.** *For any given $L, B \geq 0$ and $G > 0$, there exists a collection of $n$ functions $\{F_i\}_{i\in[n]}$, each defined over a dataset of $m$ samples, that are all i) $L$-smooth, ii) $\mu$-strongly convex with $\mu = \Theta(L)$, and iii) satisfy Assumption 1 with respect to $F_{\mathcal{G}}$. For this collection, the output $\widehat{\boldsymbol{\theta}}$ of any distributed algorithm $\mathcal{A}$ satisfies*

$$\mathbb{E}\Big[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2\Big] = \Omega\Big(\frac{f}{n-f}\frac{G^2}{\mu^2}\Big),\ \mathbb{E}\Big[\|\nabla F_{\mathcal{G}}(\widehat{\boldsymbol{\theta}})\|_2^2\Big] = \Omega\Big(\frac{f}{n-f}G^2\Big). \tag{16}$$

**Proof of Lemma 2 [42, 4]**

We construct local function families such that, regardless of the algorithm employed, the resulting observations are indistinguishable under different Byzantine client configurations.

The construction is similar to Theorem III in [41] and Proposition 3.2 in [7], but with adaptations to different optimality gaps. Let $\tau = f/(n-f)$. Consider the following three groups of local datasets:

$$\mathcal{D}_+ = \left\{\underbrace{+\frac{G_0\cdot\mathbb{1}}{L\sqrt{\tau d}},\ldots,+\frac{G_0\cdot\mathbb{1}}{L\sqrt{\tau d}}}_{n \text{ elements}}\right\},\ \mathcal{D}_- = \left\{\underbrace{-\frac{G_0\cdot\mathbb{1}}{L\sqrt{\tau d}},\ldots,-\frac{G_0\cdot\mathbb{1}}{L\sqrt{\tau d}}}_{n \text{ elements}}\right\},\ \text{and } \mathcal{D}_0 = \left\{\underbrace{0\cdot\mathbb{1},\ldots,0\cdot\mathbb{1}}_{n \text{ elements}}\right\},$$

where $\mathbb{1} \in \mathbb{R}^d$ and $G_0 > 0$ is a constant that can be chosen according to the given $G$. For each of the $n$ clients, its local dataset $\mathcal{D}_i$ is constructed as:

$$\mathcal{D}_i = \begin{cases}\mathcal{D}_+, & \text{if } i \in [1, f];\\ \mathcal{D}_-, & \text{if } i \in [f+1, 2f];\\ \mathcal{D}_0, & \text{if } i \in [2f+1, n].\end{cases}$$

That is, each dataset is of $m$ rows chosen from its corresponding $d$-dimensional data universe.

Let $\mu \leq L$ and $\mu = \Theta(L)$. Define each client's local objective as

$$F_i(\boldsymbol{\theta}) = \frac{\mu}{2n}\sum_{\boldsymbol{x}\in\mathcal{D}_i}\|\boldsymbol{\theta} - \boldsymbol{x}\|_2^2.$$

It can be readily computed that each $F_i$ is $L$ smooth, $\mu$-strongly convex, and

$$F_i(\boldsymbol{\theta}) = \begin{cases} \frac{\mu}{2}\left\|\boldsymbol{\theta} - \frac{G_0 \cdot \mathbb{1}}{L\sqrt{\tau d}}\right\|_2^2, & \text{if } i \in [1, f], \\ \frac{\mu}{2}\left\|\boldsymbol{\theta} + \frac{G_0 \cdot \mathbb{1}}{L\sqrt{\tau d}}\right\|_2^2, & \text{if } i \in [f+1, 2f], \\ \frac{\mu}{2}\left\|\boldsymbol{\theta}\right\|_2^2, & \text{otherwise.} \end{cases}$$

We consider the following two Byzantine client configurations induced by the adversary.

- **Configuration #1:** $\mathcal{G}_1 = \{1, \ldots, n\}$, i.e., $\mathcal{B}^1 = \emptyset$
- **Configuration #2:** $\mathcal{G}_2 = \{1, \ldots, f\} \cup \{2f+1, \ldots, n\}$, i.e., $\mathcal{B}^2 = [f+1, 2f]$.

Clearly, $|\mathcal{B}^1| = 0 \leq f$, and $|\mathcal{B}^2| = |[f+1, 2f]| = f$. For ease of exposition, let $F_{\mathcal{G}}^1$ and $F_{\mathcal{G}}^2$ denote the global objective functions of the two configurations, respectively, i.e.,

$$F_{\mathcal{G}}^1(\boldsymbol{\theta}) = \frac{1}{|\mathcal{G}_1|}\sum_{i \in \mathcal{G}_1} F_i(\boldsymbol{\theta}), \text{ and } F_{\mathcal{G}}^2(\boldsymbol{\theta}) = \frac{1}{|\mathcal{G}_2|}\sum_{i \in \mathcal{G}_2} F_i(\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}^{*1} = \arg\min_{\boldsymbol{\theta}} F_{\mathcal{G}}^1(\boldsymbol{\theta})$, and $\boldsymbol{\theta}^{*2} = \arg\min_{\boldsymbol{\theta}} F_{\mathcal{G}}^2(\boldsymbol{\theta})$.

Now, we verify the $(B, G)$ condition of the constructed functions.

- **Configuration #1.** It is easy to see that

$$\nabla F_{\mathcal{G}}^1(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{f} \mu\left(\boldsymbol{\theta} - \frac{G_0 \cdot \mathbb{1}}{L\sqrt{\tau d}}\right) + \frac{1}{n}\sum_{i=f+1}^{2f} \mu\left(\boldsymbol{\theta} + \frac{G_0 \cdot \mathbb{1}}{L\sqrt{\tau d}}\right) + \frac{1}{n}\sum_{i=2f+1}^{n} \mu\boldsymbol{\theta} = \mu\boldsymbol{\theta}.$$

  It follows that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \left\|\nabla F_i(\boldsymbol{\theta}) - \nabla F_{\mathcal{G}}^1(\boldsymbol{\theta})\right\|_2^2 \\ &= \frac{f}{n}\left\|\mu\boldsymbol{\theta} - \frac{\mu G_0 \mathbb{1}}{L\sqrt{\tau d}} - \mu\boldsymbol{\theta}\right\|_2^2 + \frac{n-2f}{n}\left\|\mu\boldsymbol{\theta} - \mu\boldsymbol{\theta}\right\|_2^2 + \frac{f}{n}\left\|\mu\boldsymbol{\theta} + \frac{\mu G_0 \mathbb{1}}{L\sqrt{\tau d}} - \mu\boldsymbol{\theta}\right\|_2^2 \\ &\leq \frac{2f}{n}\frac{G_0^2}{\tau} \leq \frac{2f}{n-f}\frac{G_0^2}{\tau} = 2G_0^2. \end{aligned}$$

- **Configuration #2.** Similarly, we have

$$\nabla F_{\mathcal{G}}^2(\boldsymbol{\theta}) = \frac{f}{n-f}\left(\mu\boldsymbol{\theta} - \frac{\mu G_0 \mathbb{1}}{L\sqrt{\tau d}}\right) + \frac{n-2f}{n-f}\mu\boldsymbol{\theta} = \mu\boldsymbol{\theta} - \frac{f}{n-f}\frac{\mu G_0 \mathbb{1}}{L\sqrt{\tau d}}.$$

  It follows that

$$\begin{aligned} &\frac{1}{n-f}\sum_{i \in \mathcal{G}_2} \left\|\nabla F_i(\boldsymbol{\theta}) - \nabla F_{\mathcal{G}}^2(\boldsymbol{\theta})\right\|_2^2 \\ &= \frac{f}{n-f}\left\|\frac{\mu}{L}\left(\frac{f}{n-f}\frac{G_0\mathbb{1}}{\sqrt{\tau d}} - \frac{G_0\mathbb{1}}{\sqrt{\tau d}}\right)\right\|_2^2 + \frac{n-2f}{n-f}\left\|\frac{\mu}{L}\left(\frac{f}{n-f}\frac{G_0\mathbb{1}}{\sqrt{\tau d}}\right)\right\|_2^2 \\ &\leq \frac{G_0^2}{\tau}\frac{f(n-2f)(n-f)}{(n-f)^3} = \frac{f}{n-f}\frac{G_0^2}{\tau} = G_0^2. \end{aligned}$$

Therefore, for any given $G > 0$, we can choose $G_0 = \frac{1}{\sqrt{2}}G$. For any algorithm $\mathcal{A}(F_1, \cdots, F_m)$, it cannot distinguish Configuration #1 from Configuration #2, since the Byzantine clients in Configuration #2 can disguise themselves by mimicking the behaviors of the non-Byzantine clients $f+1, \cdots, 2f$ in Configuration #1. Hence, $\mathcal{A}$ should generate the same output $\widehat{\boldsymbol{\theta}}$ under these two configurations.

We know that

$$\left\|\nabla F_{\mathcal{G}}^1(\widehat{\boldsymbol{\theta}})\right\|_2^2 = \left\|\mu\widehat{\boldsymbol{\theta}}\right\|_2^2, \text{ and } \left\|\nabla F_{\mathcal{G}}^2(\widehat{\boldsymbol{\theta}})\right\|_2^2 = \left\|\mu\widehat{\boldsymbol{\theta}} - \sqrt{\frac{f}{n-f}}\frac{\mu G_0 \mathbb{1}}{L\sqrt{d}}\right\|_2^2.$$

In addition,

$$\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*1}\right\|_2^2=\left\|\widehat{\boldsymbol{\theta}}\right\|_2^2,\ \text{ and }\ \left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*2}\right\|_2^2=\left\|\widehat{\boldsymbol{\theta}}-\sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{L\sqrt{d}}\right\|_2^2.$$

So, we have

$$\begin{aligned}\max\left\{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*1}\right\|_2^2\right],\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*2}\right\|_2^2\right]\right\}&\geq\frac{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*1}\right\|_2^2\right]+\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*2}\right\|_2^2\right]}{2}\\&=\frac{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}\right\|_2^2+\left\|\widehat{\boldsymbol{\theta}}-\sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{L\sqrt{d}}\right\|_2^2\right]}{2}\\&\geq\frac{f}{n-f}\frac{G_0^2}{4L^2}\\&=\frac{f}{n-f}\frac{G^2}{8L^2}.\end{aligned}$$

In addition,

$$\begin{aligned}\max\left\{\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}^1(\widehat{\boldsymbol{\theta}})\right\|_2^2\right],\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}^2(\widehat{\boldsymbol{\theta}})\right\|_2^2\right]\right\}&\geq\frac{\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}^1(\widehat{\boldsymbol{\theta}})\right\|_2^2\right]+\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}^2(\widehat{\boldsymbol{\theta}})\right\|_2^2\right]}{2}\\&\geq\mu^2\frac{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{1*}\right\|_2^2\right]+\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{2*}\right\|_2^2\right]}{2}\\&\geq\frac{f}{n-f}\frac{\mu^2G^2}{L^28}.\end{aligned}$$

We can get the desired results by invoking $L=\Theta(\mu)$.

### E.3 Proof of Lemma 3 (Coupled)

**Lemma E.3.** *Let $m,n,d\geq 1$ and $1\leq f<m/2$. Suppose that $\epsilon,\delta\in(0,1)$ and $1/2^{(n-f)^{1-\gamma}}\leq\delta\leq 1/(n-f)^{1+\gamma}$, where $\gamma\in(0,1)$. For any given $L,B\geq 0$ and $G>0$, there exists a collection of $n$ functions $\{F_i\}_{i\in[n]}$, each defined over a dataset of $m$ samples, that are all i) $L$-smooth, ii) $\mu$-strongly convex with $\mu=\Theta(L)$, and iii) satisfy Assumption 1 with respect to $F_{\mathcal{G}}$. The output $\widehat{\boldsymbol{\theta}}$ of a $(\epsilon,\delta)$ user-level DP learning algorithm $\mathcal{A}$ has an error at least*

$$\mathbb{E}\Big[\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2\Big]=\Omega\Big(\frac{f}{n}\frac{dG^2\log(1/\delta)}{\mu^2\epsilon^2(n-f)^2}\Big),\ \mathbb{E}\Big[\|\nabla F_{\mathcal{G}}(\widehat{\boldsymbol{\theta}})\|_2^2\Big]=\Omega\Big(\frac{f}{n}\frac{dG^2\log(1/\delta)}{\epsilon^2(n-f)^2}\Big).\tag{17}$$

**Proof of Lemma 3**

We construct local function families such that, regardless of the algorithm adopted, the resulting observations are indistinguishable under different Byzantine client configurations. Consider the following three groups of local datasets:

$$\mathcal{D}_+=\left\{\underbrace{+\frac{G_0\mathbb{1}}{L\sqrt{\tau d}},\ldots,+\frac{G_0\mathbb{1}}{L\sqrt{\tau d}}}_{m\text{ elements}}\right\},\ \mathcal{D}_-=\left\{\underbrace{-\frac{G_0\mathbb{1}}{L\sqrt{\tau d}},\ldots,-\frac{G_0\mathbb{1}}{L\sqrt{\tau d}}}_{m\text{ elements}}\right\},\text{and }\mathcal{D}_0=\left\{\underbrace{0\cdot\mathbb{1},\ldots,0\cdot\mathbb{1}}_{m\text{ elements}}\right\},$$

where $\mathbb{1}\in\mathbb{R}^d$, $\tau=f/(n-f)$, and $G_0>0$ is a constant that can be chosen according to a given $G$. For each of the $n$ clients, its local dataset $\mathcal{D}_i$ is constructed as:

$$\mathcal{D}_i=\begin{cases}\mathcal{D}_+, & \text{if } i\in[1,f];\\ \mathcal{D}_-, & \text{if } i\in[f+1,2f];\\ \mathcal{D}_0, & \text{if } i\in[2f+1,n].\end{cases}$$

That is, each dataset is of $m$ rows chosen from its corresponding $d$-dimensional data universe.

Let $\mu \leq L$ and $\mu = \Theta(L)$. Define each client's local objective as

$$F_i(\boldsymbol{\theta}) = \frac{\mu}{2n} \sum_{\boldsymbol{x} \in \mathcal{D}_i} \|\boldsymbol{\theta} - \boldsymbol{x}\|_2^2 .$$

It can be readily computed that each $F_i$ is $L$-smooth, $\mu$-strongly convex, and

$$F_i(\boldsymbol{\theta}) = \begin{cases} \frac{\mu}{2} \left\| \boldsymbol{\theta} - \frac{G_0 \mathbb{1}}{L\sqrt{\tau d}} \right\|_2^2, & \text{if } i \in [1, f]; \\ \frac{\mu}{2} \left\| \boldsymbol{\theta} + \frac{G_0 \mathbb{1}}{L\sqrt{\tau d}} \right\|_2^2, & \text{if } i \in [f+1, 2f]; \\ \frac{\mu}{2} \|\boldsymbol{\theta}\|_2^2, & \text{otherwise.} \end{cases}$$

We consider the following three Byzantine client configurations induced by the adversary.

- **Configuration #1:** $\mathcal{G}_1 = \{1, \ldots, f\} \cup \{2f+1, n\}$, i.e., $\mathcal{B}^1 = \{f+1, \ldots, 2f\}$;
- **Configuration #2:** $\mathcal{G}_2 = \{f+1, \ldots, 2f\} \cup \{2f+1, n\}$, i.e., $\mathcal{B}^2 = \{1, \ldots, f\}$;
- **Configuration #3:** $\mathcal{G}_3 = [n]$, i.e., $\mathcal{B}^3 = \emptyset$.

Clearly, $|\mathcal{B}^1| = |[f+1, 2f]| = f$, $|\mathcal{B}^2| = |[1, f]| = f$, and $|\mathcal{B}^3| = 0 \leq f$.

For any algorithm $\mathcal{A}(F_1, \cdots, F_n)$, it cannot distinguish Configuration #1 from Configuration #3, since the Byzantine clients in Configuration #1 can disguise themselves by mimicking the behaviors of the non-Byzantine clients $f+1, \cdots, 2f$ in Configuration #3. Similarly, any algorithm $\mathcal{A}(F_1, \cdots, F_n)$ cannot distinguish Configuration #2 from Configuration #3, since the Byzantine clients in Configuration #2 can disguise themselves by mimicking the behaviors of the non-Byzantine clients $1, \cdots, f$ in Configuration #3. Therefore, $\mathcal{A}$ should generate the same output $\widehat{\boldsymbol{\theta}}$ under Configuration #1 and Configuration #2.

For ease of exposition, let $F_{\mathcal{G}}^1$ and $F_{\mathcal{G}}^2$ denote the global objective functions of the configurations #1 and #2, respectively, i.e.,

$$F_{\mathcal{G}}^1(\boldsymbol{\theta}) = \frac{1}{|\mathcal{G}_1|} \sum_{i \in \mathcal{G}_1} F_i(\boldsymbol{\theta}) \text{ and } F_{\mathcal{G}}^2(\boldsymbol{\theta}) = \frac{1}{|\mathcal{G}_2|} \sum_{i \in \mathcal{G}_2} F_i(\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}^{*1} = \arg\min_{\boldsymbol{\theta}} F_{\mathcal{G}}^1(\boldsymbol{\theta})$, and $\boldsymbol{\theta}^{*2} = \arg\min_{\boldsymbol{\theta}} F_{\mathcal{G}}^2(\boldsymbol{\theta})$.

Now, we verify the $(B, G)$ condition of the constructed functions. Recall that $n - f > f \Leftrightarrow f < \frac{n}{2} \Leftrightarrow n - f > \frac{n}{2}$.

- **Configuration #1.** It is easy to see that

$$\nabla F_{\mathcal{G}}^1(\boldsymbol{\theta}) = \frac{f}{n-f} \sum_{i=1}^{f} \mu \left( \boldsymbol{\theta} - \frac{G_0 \mathbb{1}}{L\sqrt{\tau d}} \right) + \frac{n-2f}{n-f} \sum_{i=2f+1}^{n} \mu \boldsymbol{\theta} = \mu \boldsymbol{\theta} - \sqrt{\frac{f}{n-f}} \frac{\mu G_0 \mathbb{1}}{L\sqrt{d}}.$$

  It follows that

$$\frac{1}{n-f} \sum_{i \in \mathcal{G}} \left\| \nabla F_i(\boldsymbol{\theta}) - \nabla F_{\mathcal{G}}^1(\boldsymbol{\theta}) \right\|_2^2 = \underbrace{\frac{(n-2f)\mu^2}{(n-f)L^2} \left\| \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{\sqrt{d}} \right\|_2^2}_{(1.a)} + \underbrace{\frac{f\mu^2}{(n-f)L^2} \left\| \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{\sqrt{d}} - \frac{G_0 \mathbb{1}}{\sqrt{\tau d}} \right\|_2^2}_{(1.b)}.$$

  **Bounding** $(1.a)$**.** It holds that

$$\frac{(n-2f)\mu^2}{(n-f)L^2} \left\| \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{\sqrt{d}} \right\|_2^2 = \frac{(n-2f)\mu^2}{(n-f)L^2} \sum_{i=1}^{d} \left\| \sqrt{\frac{f}{n-f}} \frac{G_0}{\sqrt{d}} \right\|_2^2 = \frac{\mu^2 G_0^2}{L^2 \tau} \frac{f}{n-f} \frac{(n-2f)f}{(n-f)^2} \leq G_0^2.$$

  **Bounding** $(1.b)$**.**

$$\frac{f\mu^2}{(n-f)L^2} \left\| \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{\sqrt{d}} - \frac{G_0 \mathbb{1}}{\sqrt{\tau d}} \right\|_2^2 = \frac{f\mu^2}{(n-f)L^2} \sum_{i=1}^{d} \left\| \sqrt{\frac{f}{n-f}} \frac{G_0}{\sqrt{d}} - \frac{G_0}{\sqrt{\tau d}} \right\|_2^2 \leq \frac{\mu^2 G_0^2}{L^2 \tau} \frac{f}{m-f} \frac{(n-2f)^2}{(n-f)^2} \leq G_0^2.$$

  Combining them together, it holds that

$$\frac{1}{n-f} \sum_{i \in \mathcal{G}} \left\| \nabla F_i(\boldsymbol{\theta}) - \nabla F_{\mathcal{G}}^1(\boldsymbol{\theta}) \right\|_2^2 \leq 2G_0^2.$$

- **Configuration #2:**

$$\nabla F_{\mathcal{G}}^2(\boldsymbol{\theta}) = \frac{f}{n-f}\sum_{i=f+1}^{2f}\mu\left(\boldsymbol{\theta} - \frac{G_0\mathbb{1}}{L\sqrt{\tau d}}\right) + \frac{n-2f}{n-f}\sum_{i=2f+1}^{n}\mu\boldsymbol{\theta} = \mu\boldsymbol{\theta} + \sqrt{\frac{f}{n-f}}\frac{\mu G_0\mathbb{1}}{L\sqrt{d}}.$$

It follows that

$$\frac{1}{n-f}\sum_{i\in\mathcal{G}}\left\|\nabla F_i(\boldsymbol{\theta}) - \nabla F_{\mathcal{G}}^2(\boldsymbol{\theta})\right\|_2^2 = \underbrace{\frac{(n-2f)\mu^2}{(n-f)L^2}\left\|-\sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{d}}\right\|_2^2}_{(2.a)} + \underbrace{\frac{f\mu^2}{(n-f)L^2}\left\|-\sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{d}} + \frac{G_0\mathbb{1}}{\sqrt{\tau d}}\right\|_2^2}_{(2.b)}.$$

**Bounding** $(2.a)$**.**

$$\frac{(n-2f)\mu^2}{(n-f)L^2}\left\|-\sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{d}}\right\|_2^2 = \frac{(n-2f)\mu^2}{(n-f)L^2}\sum_{i=1}^{d}\left\|G_0\sqrt{\frac{f}{(n-f)d}}\right\|_2^2 = \frac{\mu^2G_0^2}{L^2\tau}\frac{f}{n-f}\frac{(n-2f)f}{(n-f)^2} \leq G_0^2.$$

**Bounding** $(2.b)$**.**

$$\frac{f\mu^2}{(n-f)L^2}\left\|-\sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{d}} + \frac{G_0\mathbb{1}}{\sqrt{\tau d}}\right\|_2^2 \leq \frac{\mu^2G_0^2}{L^2\tau}\frac{f}{n-f}\frac{(n-2f)^2}{(n-f)^2} \leq G_0^2.$$

Combining them together, it holds that

$$\frac{1}{n-f}\sum_{i\in\mathcal{G}}\left\|\nabla F_i(\boldsymbol{\theta}) - \nabla F_{\mathcal{G}}^2(\boldsymbol{\theta})\right\|_2^2 \leq 2G_0^2.$$

Therefore, for any given $G > 0$, we can choose $G_0 = G/\sqrt{2}$.

Let $\mathcal{D}'$ be an auxiliary dataset.

$$\mathcal{D}' = \left\{\underbrace{\mathbb{1}_m \otimes \frac{G_0\mathbb{1}_d}{L\sqrt{\tau nd}}}_{f \text{ elements}}, \ldots, \underbrace{\mathbb{1}_m \otimes \frac{-G_0\mathbb{1}_d}{L\sqrt{\tau nd}}}_{f \text{ elements}}, \ldots, \underbrace{\mathbb{1}_m \otimes 0\mathbb{1}_d}_{(n-2f) \text{ elements}}\right\}.$$

Next, we show that the errors induced by $\mathcal{D}$ and $\mathcal{D}'$ are the same in a special case. Suppose that an auxiliary global model parameter $\widehat{\boldsymbol{\theta}}' \in \mathbb{R}^{md}$ It holds, when $\widehat{\boldsymbol{\theta}}' = \mathbb{1}_m \otimes \widehat{\boldsymbol{\theta}}/\sqrt{m}$, that

- **Configuration #1.**

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}}' - \boldsymbol{\theta}'^{1*}\right\|_2^2 &= \left\|\widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{md}}\right\|_2^2 \\ &= \left\|\mathbb{1}_m \otimes \frac{\widehat{\boldsymbol{\theta}}}{\sqrt{m}} - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{md}}\right\|_2^2 \\ &= \frac{1}{m}\sum_{j=1}^{m}\left\|\widehat{\boldsymbol{\theta}} - \sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{d}}\right\|_2^2 = \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{1*}\right\|_2^2.\end{aligned}$$

- **Configuration #2.**

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}}' - \boldsymbol{\theta}'^{2*}\right\|_2^2 &= \left\|\widehat{\boldsymbol{\theta}}' + \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{md}}\right\|_2^2 \\ &= \left\|\mathbb{1}_m \otimes \frac{\widehat{\boldsymbol{\theta}}}{\sqrt{m}} + \mathbb{1}_m \otimes \sqrt{\frac{f}{m-f}}\frac{G_0\mathbb{1}}{\sqrt{md}}\right\|_2^2 \\ &= \frac{1}{m}\sum_{j=1}^{m}\left\|\widehat{\boldsymbol{\theta}} + \sqrt{\frac{f}{n-f}}\frac{G_0\mathbb{1}}{\sqrt{d}}\right\|_2^2 = \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{2*}\right\|_2^2.\end{aligned}$$

We construct another auxiliary dataset $\mathcal{D}'_{\pm}$ of $n-f$ data points.

$$\mathcal{D}'_{\pm} = \left\{ \underbrace{\mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{\boldsymbol{x}_1^{\pm} G_0 \mathbb{1}}{L\sqrt{nd}}, \ldots, \mathbb{1}_n \otimes \sqrt{\frac{f}{n-f}} \frac{\boldsymbol{x}_{n-f}^{\pm} G_0 \mathbb{1}}{L\sqrt{nd}}}_{(n-f) \text{ elements}} \right\},$$

where $\mathbb{1} \in \mathbb{R}^d$ and $\boldsymbol{x}_i^{\pm} \in \{-1, +1\}$ for all $i \in [1, n-f]$. Let $\widehat{\boldsymbol{\theta}}_{(i)}$ denote the $i$-th coordinate of $\widehat{\boldsymbol{\theta}}$, $\overline{\boldsymbol{x}}^{\pm} \triangleq \sum_{i=1}^{n-f} \boldsymbol{x}_i^{\pm}/(n-f)$, and $\overline{\boldsymbol{x}}_{(i)}^{\pm}$ denote the $i$-th coordinate of $\overline{\boldsymbol{x}}^{\pm}$. Since we know that $\overline{\boldsymbol{x}}_{(i)}^{\pm} \in [-1, 1]$, it holds that $\overline{\boldsymbol{x}}_{(i)}^{\pm} = -1 \cdot \lambda_{(i)} + 1 \cdot (1 - \lambda_{(i)})$ for some $\lambda_{(i)} \geq 0$. Further, it holds that $\lambda_i = \lambda$ because $\overline{\boldsymbol{x}}_{(i)}^{\pm}$ is by definition identical for all $i \in \{1, \ldots, d\}$. It follows that

$$\widehat{\boldsymbol{\theta}}'_{(i)} - \sqrt{\frac{f}{n-f}} \frac{\overline{\boldsymbol{x}}_{(i)}^{\pm} G_0}{L\sqrt{md}} = \lambda \left( \widehat{\boldsymbol{\theta}}'_{(i)} + \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right) + (1-\lambda) \left( \widehat{\boldsymbol{\theta}}'_{(i)} - \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right)$$

$$\Rightarrow \left\| \widehat{\boldsymbol{\theta}}'_{(i)} - \sqrt{\frac{f}{n-f}} \frac{\overline{\boldsymbol{x}}_{(i)}^{\pm} G_0}{L\sqrt{md}} \right\|_2^2 \leq \lambda \left\| \widehat{\boldsymbol{\theta}}'_{(i)} + \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2 + (1-\lambda) \left\| \widehat{\boldsymbol{\theta}}'_{(i)} - \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2.$$

Since this holds for each coordinate $i \in \{1, \ldots, d\}$, it holds that

$$\sum_{i=1}^{md} \left\| \widehat{\boldsymbol{\theta}}'_{(i)} - \sqrt{\frac{f}{n-f}} \frac{\overline{\boldsymbol{x}}_{(i)}^{\pm} G_0}{L\sqrt{md}} \right\|_2^2 \leq \lambda \sum_{i=1}^{md} \left\| \widehat{\boldsymbol{\theta}}'_{(i)} + \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2 + (1-\lambda) \sum_{i=1}^{md} \left\| \widehat{\boldsymbol{\theta}}'_{(i)} - \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2$$

$$\Leftrightarrow \quad \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{\overline{\boldsymbol{x}}^{\pm} G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \leq \lambda \left\| \widehat{\boldsymbol{\theta}}' + \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2 + (1-\lambda) \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2.$$

Taking expectation with respect to the randomness in the algorithm, we have

$$\begin{aligned}
&\mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{\overline{\boldsymbol{x}}^{\pm} G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \right]^2 \\
&\leq \lambda \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' + \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \right] + (1-\lambda) \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \right] \\
&\leq \max\left\{ \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' + \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \right], \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \right] \right\}. \quad (18)
\end{aligned}$$

Recall that we have shown the indistinguishability of Configurations #1 and #2. Since the algorithm should generate the same output for all possible collections of Byzantine clients, we have for any $\widehat{\boldsymbol{\theta}}'$ that

$$\begin{aligned}
&\max\left\{ \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \boldsymbol{\theta}'^{1*} \right\|_2^2 \right], \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \boldsymbol{\theta}'^{2*} \right\|_2^2 \right] \right\} \\
&\qquad = \max\left\{ \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2 \right], \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' + \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{G_0}{L\sqrt{md}} \right\|_2^2 \right] \right\} \\
&\qquad \overset{(18)}{\geq} \mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{\overline{\boldsymbol{x}}^{\pm} G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \right]. \quad (19)
\end{aligned}$$

In our data universe below, $\Delta = \sqrt{\frac{f}{m(n-f)}} \frac{G_0}{L}$ with dimension $md$, and the one-way marginal (average) is computed over $n-f$ data points in the auxiliary dataset $\mathcal{D}'_{\pm}$. It holds that

$$\mathbb{E}\left[ \left\| \widehat{\boldsymbol{\theta}}' - \mathbb{1}_m \otimes \sqrt{\frac{f}{n-f}} \frac{\overline{\boldsymbol{x}}^{\pm} G_0 \mathbb{1}}{L\sqrt{md}} \right\|_2^2 \right] \overset{(a)}{=} \Omega\left( \frac{f}{n-f} \frac{nd \log(1/\delta)}{m\epsilon^2 (n-f)^2} \frac{G_0^2}{L^2} \right) \overset{(b)}{=} \Omega\left( \frac{f}{n-f} \frac{d \log(1/\delta)}{\epsilon^2 (n-f)^2} \frac{G^2}{L^2} \right),$$

where equality $(a)$ holds because of Corollary 2, and equality $(b)$ holds because $G^2 = 2G_0^2$. It follows that

$$\max\left\{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{1*}\right\|_2^2\right], \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{2*}\right\|_2^2\right]\right\} = \max\left\{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}' - \boldsymbol{\theta}'^{1*}\right\|_2^2\right], \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}' - \boldsymbol{\theta}'^{2*}\right\|_2^2\right]\right\}$$
$$= \Omega\left(\frac{f}{n-f}\frac{d\log(1/\delta)}{\epsilon^2(n-f)^2}\frac{G^2}{L^2}\right).$$

In addition,

$$\max\left\{\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}^1(\widehat{\boldsymbol{\theta}})\right\|_2^2\right], \mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}^2(\widehat{\boldsymbol{\theta}})\right\|_2^2\right]\right\} = \mu^2 \max\left\{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{1*}\right\|_2^2\right], \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{2*}\right\|_2^2\right]\right\}$$
$$\geq \Omega\left(\frac{f}{n-f}\frac{d\log(1/\delta)G^2}{\epsilon^2(n-f)^2}\right).$$

We can get the desired results by invoking $L = \Theta(\mu)$.

### E.4 Proof of Theorem 1

Theorem 1 can be obtained by summing up Lemmas 1, 2, and 3.

### E.5 Supporting Lemmas of Appendix E

**Lemma E.4** ([60]). *Let $n, d \geq 1$, $\mathcal{X} = \left\{\pm 1/\sqrt{d}\right\}^d$, and $\mathcal{Y} = [\pm 1/\sqrt{d}]^d$. Let $\mathcal{D} \in \mathcal{X}^n$ denote a dataset of size $n$, and $\bar{\mathcal{D}} \triangleq \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{D}} \boldsymbol{x}$ denote the average of the samples in $\mathcal{D}$. Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an arbitrary mapping used to estimate $\bar{\mathcal{D}}$. Suppose that $\epsilon = O(1)$ and $2^{-o(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$. If $\mathcal{A}$ is $(\epsilon, \delta)$-DP and $\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \bar{\mathcal{D}}\right\|_1\right] \leq \sqrt{d}\varrho$, then*

$$n = \Omega\left(\frac{\sqrt{d\log(1/\delta)}}{\varrho\epsilon}\right).$$

**Lemma E.5.** *Let $n, d \geq 1$, $\mathcal{X} = \left\{\pm 1/\sqrt{d}\right\}^d$, and $\mathcal{Y} = [\pm 1/\sqrt{d}]^d$. Let $\mathcal{D} \in \mathcal{X}^n$ denote a dataset of size $n$, and $\overline{\mathcal{D}} \triangleq \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{D}} \boldsymbol{x}$ denote the average of the samples in $\mathcal{D}$. Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an arbitrary mapping used to estimate $\overline{\mathcal{D}}$. Suppose that $\epsilon = O(1)$ and $2^{-o(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$. If $\mathcal{A}$ is $(\epsilon, \delta)$-DP, then*

$$\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_2^2\right] = \Omega\left(\frac{d\log(1/\delta)}{\epsilon^2 n^2}\right), \tag{20}$$

*where the expectation is taken over the randomness of $\mathcal{A}$.*

**Proof of Lemma E.5.** We prove this lemma by contradiction. Suppose that (20) is not true. Then

$$\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_2^2\right] = o\left(\frac{d\log(1/\delta)}{\epsilon^2 n^2}\right).$$

On the other hand, we know that

$$\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_1\right] \leq \sqrt{d}\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_2\right] \leq \sqrt{d}\sqrt{\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_2^2\right]} = o\left(\frac{d\sqrt{\log(1/\delta)}}{\epsilon n}\right).$$

It further implies that there exists $0 < \varrho = o(\frac{\sqrt{d\log(1/\delta)}}{\epsilon n})$ (i.e., $n = o(\frac{\sqrt{d\log(1/\delta)}}{\epsilon\varrho})$) such that

$$\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_1\right] \leq \sqrt{d}\varrho.$$

Invoking Lemma E.4, we obtain $n = \Omega(\sqrt{d\log(1/\delta)}/\varrho\epsilon)$, contradicting with $n = o(\sqrt{d\log(1/\delta)}/\varrho\epsilon)$. Therefore, (20) must be true.

□

**Corollary 2.** *Fix $\Delta > 0$, and let $n, d \geq 1$. Define $\mathcal{X} = \left\{\pm\Delta/\sqrt{d}\right\}^d$ and $\mathcal{Y} = [\pm\Delta/\sqrt{d}]^d$. Let $\mathcal{A} : k\mathcal{X}^n \to k\mathcal{Y}$ be an arbitrary scale-invariant mapping for any $k > 0$, i.e., $\mathcal{A}(k \cdot \mathcal{D}) = k \cdot \mathcal{A}(\mathcal{D})$ for any dataset $\mathcal{D} \in \mathcal{X}^n$ and $k > 0$. For any $\epsilon = O(1)$, and $2^{-o(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$, if $\mathcal{A}$ is $(\epsilon, \delta)$-DP, then*

$$\mathbb{E}\left[\left\|\mathcal{A}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_2^2\right] = \Omega(\frac{\Delta^2 d \log(1/\delta)}{\epsilon^2 n^2}),$$

*where $\overline{\mathcal{D}}$ is defined in Lemma E.5.*

Corollary 2 follows immediately from Lemma E.5 and the fact that $\mathcal{A}$ is scale invariant.

## F Privacy Analysis

**Claim 1.** When $n \geq (2+\nu)f$, where $\nu > 0$, it holds that

$$\left(1 - \frac{2f}{n}\right)^{-2} \asymp 1 + \frac{f}{n}.$$

**Proof of Claim 1.** It is easy to see that

$$\frac{1}{(n-2f)^2} = \frac{1}{n^2}\frac{1}{(1-\frac{2f}{n})^2} = \frac{1}{n^2}\left(1 - \frac{2f}{n}\right)^{-2}.$$

By Bernoulli inequality, it holds that $\left(1 - \frac{2f}{n}\right)^{-2} \geq 1 + \frac{4f}{n}$. We know that

$$\frac{1}{(1-x)^2} = \frac{(1-x)^2 + 1 - (1-x)^2}{(1-x)^2} = 1 + \frac{x(2-x)}{(1-x)^2} \leq 1 + \frac{2x}{(1-x)^2}.$$

Therefore,

$$1 + \frac{4f}{n} \leq \left(1 - \frac{2f}{n}\right)^{-2} \leq 1 + \frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}.$$

It follows that

$$\left(1 - \frac{2f}{n}\right)^{-2} \asymp 1 + \frac{f}{n}.$$

□

### F.1 Proof of Lemma 4

We consider $\texttt{TM}$ as a continuous function of $h$. For ease of notation, let $L(h) \triangleq |\{j : a_j < h\}|$ be the rank of $h$ among the unchanged values. After insertion, its position becomes $L(h)+1$. The value remains untrimmed iff

$$f + 1 \leq L(h) + 1 \leq n - f \quad \Leftrightarrow f \leq L(h) \leq n - f - 1.$$

**Case #1:** $L(h) < f$**.** $h$ is trimmed from below.

$$\texttt{TM}(h) = \frac{1}{n-2f}\sum_{j=f}^{n-f-1} a_j = \texttt{TM}_{\text{bottom}}.$$

**Case #2:** $f \leq L(h) \leq n - f$**.** $h$ is kept.

$$\texttt{TM}(h) = \frac{1}{n-2f}\left(h + \sum_{j=f+1}^{n-f-1} a_j\right) = \texttt{TM}_{\text{middle}}(h).$$

**Case #3:** $L(h) > n - f$ $h$ is trimmed from above.

$$\texttt{TM}(h) = \frac{1}{n-2f}\sum_{j=f+1}^{n-f} a_j = \texttt{TM}_{\text{top}}.$$

It is easy to see that

$$\texttt{TM}_{\text{middle}}(h) = \frac{1}{n-2f}\left(h + \sum_{j=f+1}^{n-f-1} a_j\right) \geq \frac{1}{n-2f}\left(a_f + \sum_{j=f+1}^{n-f-1} a_j\right) = \texttt{TM}_{\text{bottom}},$$

and

$$\texttt{TM}_{\text{middle}}(h) = \frac{1}{n-2f}\left(h + \sum_{j=f+1}^{n-f-1} a_j\right) \leq \frac{1}{n-2f}\left(a_{n-f} + \sum_{j=f+1}^{n-f-1} a_j\right) = \texttt{TM}_{\text{top}},$$

where both inequalities hold because $f \leq L(h) \leq n-f$. Therefore, we have

$$\texttt{TM}(h) = \begin{cases} \texttt{TM}_{\text{bottom}}, & \text{if } h < a_f; \\ \frac{1}{n-2f}\left(h + \sum_{j=f+1}^{n-f-1} a_j\right), & \text{if } a_f \leq h \leq a_{n-f}; \\ \texttt{TM}_{\text{top}}, & \text{if } h > a_{n-f}. \end{cases}$$

With a bit of abuse of notations, it follows that

$$\frac{\mathrm{d}\texttt{TM}(h)}{\mathrm{d}h} = \begin{cases} \frac{1}{n-2f}, & \text{if } a_f \leq h \leq a_{n-f}; \\ 0, & \text{otherwise.} \end{cases}$$

Fig. 2 confirms this pictorially. Without loss of generality, we assume $h < h'$. Therefore,

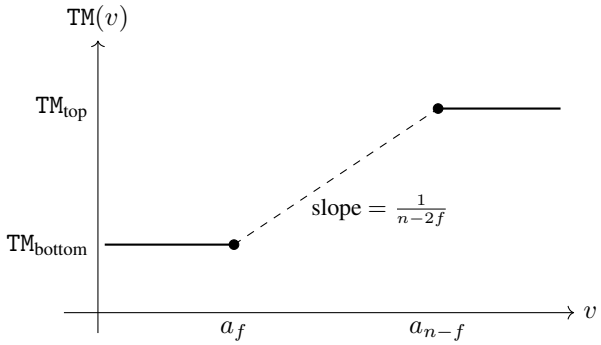



Figure 2: Demonstration of CWTM with respect to a coordinate-wise insertion value.

$$|\texttt{TM}(h) - \texttt{TM}(h')| = \left|\int_h^{h'} \frac{\mathrm{d}\texttt{TM}(h)}{\mathrm{d}h}\mathrm{d}h\right| \leq \frac{1}{n-2f}\int_h^{h'} \mathrm{d}h = \frac{|h-h'|}{n-2f} \leq \frac{|h-h'|}{n}\sqrt{1 + \frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}},$$

where the last inequality follows from Claim 1.

### F.2 Proof of Proposition 1

Let $i$ denote the replaced client in $\mathcal{G}_1$, $j$ denote the replaced client in its adjacent healthy client set $\mathcal{G}_2$. Further, define $\texttt{Diff}(\texttt{CWTM})$ as the difference between the CWTM outputs of $\mathcal{G}_1$ and $\mathcal{G}_2$.

**Step 1: One-dimensional analysis.** From Lemma 4, we know that

$$[\|\texttt{Diff}(\texttt{CWTM})\|_2]_r \leq \frac{|[\boldsymbol{x}_i]_r - [\boldsymbol{x}_j]_r|}{n-2f} = \frac{(1-\beta_t)\,|[\boldsymbol{m}_i]_r - [\boldsymbol{m}_j]_r|}{n-2f}.$$

**Step 2: High-dimensional analysis.** By the definition of Byzantine user-level sensitivity, it holds that

$$\begin{aligned} \|\texttt{Diff}(\texttt{CWTM})\|_2^2 &\leq \frac{(1-\beta_t)^2 \sum_{r=1}^d |[\boldsymbol{m}_i]_r - [\boldsymbol{m}_j]_r|^2}{(n-2f)^2} \\ &\leq \frac{2(1-\beta_t)^2 \sum_{r=1}^d \left(|[\boldsymbol{m}_i]_r|^2 + |[\boldsymbol{m}_j]_r|^2\right)}{(n-2f)^2} \\ &= \frac{2(1-\beta_t)^2 \left(\|\boldsymbol{m}_i\|_2^2 + \|\boldsymbol{m}_j\|_2^2\right)}{(n-2f)^2} \\ &\leq \frac{4(1-\beta_t)^2 C^2}{(n-2f)^2}. \end{aligned}$$

**Step 4: Combining together.**

$$\|\texttt{Diff}(\texttt{CWTM})\|_2^2 \leq \frac{4(1-\beta_t)^2C^2}{(n-2f)^2} \leq \frac{4(1-\beta_t)^2C^2}{n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right).$$

Therefore, we have the Byzantine user-level sensitivity of CWTM as

$$\left(\Delta(\texttt{CWTM})\right)^2 = \max\|\texttt{Diff}(\texttt{CWTM})\|_2^2 = \frac{4(1-\beta_t)^2C^2}{n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right).$$

### F.3 Supporting Rényi DP Analysis

Our privacy analysis is based on Rényi differential privacy (RDP) for its tight composition guarantees. Its guarantee can be translated back to DP. First, we introduce the notions of RDP.

**Definition 5** ($(\alpha,\epsilon)$-RDP, Definition 4 in [52])**.** A randomized mechanism $\mathcal{M}:\mathcal{D}\to\mathcal{R}$ is said to have $\epsilon$-Rényi differential privacy of order $\alpha$, or $(\alpha,\epsilon)$-RDP for short, if for any adjacent $\mathcal{D},\mathcal{D}'\in\boldsymbol{X}$, it holds that

$$D_\alpha\left(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')\right)\leq\epsilon,$$

where $D_\alpha(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')) \triangleq \frac{1}{\alpha-1}\log\mathbb{E}_{x\sim\mathcal{M}(\mathcal{D}')}(\mathcal{M}(\mathcal{D})(x)/\mathcal{M}(\mathcal{D}')(x))^\alpha$ defines the Rényi divergence of order $\alpha>1$ for two probability distributions $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$.

Next, we introduce the connections between Rényi differential privacy and differential privacy [26].

**Lemma F.6** (RDP to DP, Proposition 3 in [52])**.** *If $\mathcal{M}$ is an $(\alpha,\epsilon)$-RDP mechanism, it also satisfies $(\epsilon+\log(1/\delta)/(\alpha-1),\delta)$-differential privacy for any $\delta\in(0,1)$.*

It is well known that the Gaussian mechanism can also work under Rényi differential privacy.

**Lemma F.7** (Gaussian mechanism, Corollary 3 in [52])**.** *Define Gaussian mechanism $\boldsymbol{G}_{\sigma_{dp}}f = f+\mathcal{N}(0,\sigma_{dp}^2\mathbf{I})$, where $f$ is a function with $l_2$ sensitivity $\Delta$. The gaussian mechanism $\boldsymbol{G}_{\sigma_{dp}}f$ satisfies $(\alpha,\alpha\Delta^2/2\sigma_{dp}^2)$-RDP.*

Lemma F.8 characterizes the adaptive composition in RDP.

**Lemma F.8** (Adaptive composition, Proposition 1 in [52])**.** *Let $\mathcal{M}_1:\mathcal{D}\to\mathcal{R}_1$ be $(\alpha,\epsilon_1)$-RDP and $g:\mathcal{R}_1\times\mathcal{D}\to\mathcal{R}_2$ be $(\alpha,\epsilon_2)$-RDP, then the mechanism defined as $(X,Y)$, where $X\sim\mathcal{M}_1(\mathcal{D})$ and $Y\sim\mathcal{M}_2(X,\mathcal{D})$, satisfies $(\alpha,\epsilon_1+\epsilon_2)$-RDP.*

**Lemma F.9.** *Let $\delta\in(0,1)$, Algorithm 1 is a $(\epsilon^*,\delta)$ differentially private distributed learning algorithm with*

$$\epsilon^* = \inf_{\alpha>1}\left(T\epsilon(\alpha)+\frac{\log(1/\delta)}{\alpha-1}\right),$$

*where $\epsilon(\alpha)=\frac{\alpha\Delta_r^2}{2\sigma_{dp,r}^2}=\frac{\alpha\Delta^2}{2\sigma_{dp}^2}$ is the per-round RDP budget.*

*Proof of Lemma F.9.* We characterize the privacy guarantees of Algorithm 1 based on RDP. This includes two steps: i) derive guarantees in a single global round for all clients, and ii) compose the single-round guarantees to account for the final privacy loss.

**Step 1: single-round guarantee.** Definition 1 defines the notion of differential privacy at a system level. It shows that the output model parameter should be nearly indistinguishable when two adjacent client sets differ by only one client.

We track the privacy loss through the robust gradient descent step (line 7) in Algorithm 1. In Proposition 1, we have derived the Byzantine user-level sensitivity.

**Step 2: $T$ rounds composition.** By Lemma F.8, we know that Algorithm 1 is $(\alpha,T\epsilon(\alpha))$-RDP after $T$ rounds of execution, where $\epsilon(\alpha)=\alpha\Delta^2/2\sigma_{dp}^2$ is the per-round RDP budget.

Table 4: Byzantine-robust user-level sensitivity when the honest inputs to the corresponding robust aggregator are bounded by $C$ or $\tau$ in the case of CClip.

| Robust Aggregator | Trimmed Mean | Coordinate Median | Geometric Median | Krum | CClip | Average (fault-free) |
|---|---|---|---|---|---|---|
| Sensitivity | $\frac{2C}{n-2f}$ | $2C\sqrt{d}$ | $2C$ | $2C$ | $\frac{2\tau}{n}$ | $\frac{2C}{n}$ |

Now, we convert RDP guarantees back to differential privacy guarantees. By Lemma F.6, it holds that Algorithm 1 is $(\epsilon_{dp}, \delta)$-differentially private, where $\delta \in (0,1)$ and

$$\epsilon_{dp} = \sum_{r=0}^{T-1} \frac{\alpha \Delta_r^2}{2\sigma_{dp,r}^2} + \frac{\log(1/\delta)}{\alpha - 1} = \frac{\alpha T \Delta^2}{2\sigma_{dp}^2} + \frac{\log(1/\delta)}{\alpha - 1}, \tag{21}$$

where the last equality follows from $\frac{\Delta_r^2}{\sigma_{dp,r}^2} = \frac{\Delta^2}{\sigma_{dp}^2}$ for $r \in [T]$. □

### F.4 Proof of Theorem 2

By Lemma F.9, since (21) holds for any $\alpha > 1$, we can optimize it to obtain a tighter budget.

$$\frac{\partial \epsilon_{dp}}{\partial \alpha} = \frac{T\Delta_{\text{Byz},r}^2}{2\sigma_{dp,r}^2} - \frac{\log(1/\delta)}{(\alpha-1)^2} = 0 \Leftrightarrow \alpha = 1 + \frac{\sigma_{dp,r}}{\Delta_{\text{Byz},r}} \sqrt{\frac{2\log(1/\delta)}{T}} > 1,$$

which implies that

$$\epsilon_{dp}^* = \frac{T\Delta_{\text{Byz},r}^2}{2\sigma_{dp,r}^2} \left(1 + \frac{\sigma_{dp,r}}{\Delta_{\text{Byz},r}} \sqrt{\frac{2\log(1/\delta)}{T}}\right) + \frac{\log(1/\delta)}{\frac{\sigma_{dp,r}}{\Delta_{\text{Byz},r}} \sqrt{\frac{2\log(1/\delta)}{T}}} = \frac{T\Delta_{\text{Byz},r}^2}{2\sigma_{dp,r}^2} + \frac{\Delta_{\text{Byz},r}\sqrt{2T\log(1/\delta)}}{\sigma_{dp,r}}.$$

When $\sigma_{dp,r} \geq 2\Delta_{\text{Byz},r}\sqrt{2T\log(1/\delta)}/\epsilon$ and $\epsilon \leq 8\log(1/\delta)$, it holds that

$$\epsilon_{dp}^* = \frac{T\Delta_{\text{Byz},r}^2}{2\sigma_{dp,r}^2} + \frac{\Delta_{\text{Byz},r}\sqrt{2T\log(1/\delta)}}{\sigma_{dp,r}} \leq \frac{\epsilon^2}{16\log(1/\delta)} + \frac{\epsilon}{2} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

### F.5 Byzantine-robust User-level Sensitivity Analysis of Additional Robust Aggregators

In this section, we characterize the sensitivity of robust aggregators when a single honest client is replaced. When updates from honest clients are bounded, updates from Byzantine clients must also be bounded in the worst case. Otherwise, it is easy for the aggregator to filter out their presence. In the remainder of this section, we consider $\{\boldsymbol{x}_i\}_{i\in[n]}$ as the inputs to the robust aggregator. Note that we have $\boldsymbol{x}_i^t = \boldsymbol{M}_i^t = \beta_t \widetilde{R}^{t-1} + (1-\beta_t)\widehat{\boldsymbol{m}}_i^t$ in Algorithm 1.

**Krum.** Let $\mathcal{N}_j$ denote the set of indices of the $n-f$ nearest neighbors of $\boldsymbol{x}_j$ in $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, with ties arbitrarily broken. `Krum` outputs the vector $\boldsymbol{x}_{k^*}$ such that

$$k^* \in \arg\min_{j\in[n]} \sum_{i\in\mathcal{N}_j} \|\boldsymbol{x}_j - \boldsymbol{x}_i\|_2^2,$$

with ties arbitrarily broken if the set of minimizers above includes more than one element.

**Example 1.** Suppose that we have a collection of five clients' updates $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5\}$ such that $\|\boldsymbol{x}_i\|_2 = C$ for $i \in [5]$. Further, we assume $f = 1$ Byzantine client is $\boldsymbol{x}_5$. There exist client placements where the Byzantine user-level sensitivity of `Krum` is $2C$.

**Proof of Example 1.** We consider two adjacent configurations, where clients are divided into two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$. Let the clients in each cluster reside at the same point, i.e., $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = 0$ for $i, j \in \mathcal{C}_1$ and for $i, j \in \mathcal{C}_2$, respectively. Moreover, let $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = 2C$ for all $i \in \mathcal{C}_1$ and all $j \in \mathcal{C}_2$.

**Configuration 1.** Let $\mathcal{C}_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, and $\mathcal{C}_2 = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}$. It can be readily checked that

$$\texttt{Krum}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5) = \boldsymbol{x}_{\mathcal{C}_1},$$

which denotes any points in $\mathcal{C}_1$.

**Configuration 2.** Without loss of generality, we replace client $\boldsymbol{x}_1$ in $\mathcal{C}_1$ with an auxiliary client $\boldsymbol{x}' \in \mathcal{C}_2$. Therefore, it holds that $\mathcal{C}_1 = \{\boldsymbol{x}_2, \boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}'\}$. It can be readily checked that

$$\texttt{Krum}(\boldsymbol{x}', \ldots, \boldsymbol{x}_5) = \boldsymbol{x}_{\mathcal{C}_2},$$

which denotes any points in $\mathcal{C}_2$.

Therefore, the Byzantine user-level sensitivity of `Krum` in this instance is $2C$. □

**Proposition 2.** *Let $m \in \mathbb{N}$ and $f < n/2$. The Byzantine-robust user-level sensitivity of* `Krum` *is $2C$ if $\|\boldsymbol{x}_i\|_2 \leq C$ for $i \in [n]$.*

**Proof of Proposition 2.** We prove this by contradiction based on Example 1. Suppose that we have the Byzantine user-level sensitivity of `Krum` is smaller than $2C$. Contradiction with Example 1. Therefore, it must be true that $\Delta(\texttt{Krum}) = 2C$.

It is a rather trivial bound. By the triangle inequality, we have

$$\Delta(\texttt{Krum}) = \|\boldsymbol{x}_{k_1} - \boldsymbol{x}_{k_2}\|_2 \leq \|\boldsymbol{x}_{k_1}\|_2 + \|\boldsymbol{x}_{k_2}\|_2 \leq 2C,$$

where $k_1$ and $k_2$ denote the indices of the minimizers of two adjacent honest client sets, respectively. □

**Median.** For input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, their coordinate-wise median, denoted by $\texttt{CWMed}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, is defined to be a vector whose $k$-th coordinate, for all $k \in [d]$, is defined to be

$$[\texttt{CWMed}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)]_k \triangleq \texttt{Median}([\boldsymbol{x}_1]_k \ldots, [\boldsymbol{x}_n]_k).$$

**Example 2.** Suppose that we have a collection of five clients' updates $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5\}$ such that $\|\boldsymbol{x}_i\|_2 = C$ for $i \in [5]$. Further, we assume $f = 1$ Byzantine client is $\boldsymbol{x}_5$. There exist client placements where the Byzantine user-level sensitivity of `CWMedian` is $2C$.

**Proof of Example 2.** We consider two adjacent configurations, where clients are divided into two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$. Let the clients in each cluster reside at the same point, i.e., $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = 0$ for $i, j \in \mathcal{C}_1$ and for $i, j \in \mathcal{C}_2$, respectively. Moreover, let $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = 2C$ for all $i \in \mathcal{C}_1$ and all $j \in \mathcal{C}_2$. More specifically, let

$$\boldsymbol{x}_i = \begin{cases} (\underbrace{C, 0, \ldots, 0}_{d \text{ dimensions}}), & \text{if } i \in \mathcal{C}_1; \\ (\underbrace{-C, 0, \ldots, 0}_{d \text{ dimensions}}), & \text{if } i \in \mathcal{C}_2. \end{cases}$$

**Configuration 1.** Let $\mathcal{C}_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, and $\mathcal{C}_2 = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}$. It can be readily checked that

$$\texttt{CWMed}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5) = (C, 0, \ldots, 0).$$

**Configuration 2.** Without loss of generality, we replace client $\boldsymbol{x}_1$ in $\mathcal{C}_1$ with an auxiliary client $\boldsymbol{x}' \in \mathcal{C}_2$. Therefore, it holds that $\mathcal{C}_1 = \{\boldsymbol{x}_2, \boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}'\}$. It can be readily checked that

$$\texttt{CWMed}(\boldsymbol{x}', \ldots, \boldsymbol{x}_5) = (-C, 0, \ldots, 0).$$

Therefore, the Byzantine user-level sensitivity of `CWMed` in this instance is $2C$. □

**Proposition 3.** *Let $m \in \mathbb{N}$ and $f < n/2$. The Byzantine-robust user-level sensitivity of* `CWMed` *is at least $2C$ if $\|\boldsymbol{x}_i\|_2 \leq C$ for $i \in [n]$.*

**Proof of Proposition 3.** We prove this by contradiction based on Example 2. Suppose that we have the Byzantine user-level sensitivity of `CWMed` is smaller than $2C$. Contradiction with Example 2. Therefore, it must be true that $\Delta(\texttt{CWMed})$ is at least $2C$. □

**Geometric median.** The geometric median of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, denoted by $\texttt{GM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, is defined to be a vector that minimizes the sum of the $l_2$ distances to these vectors:

$$\texttt{GM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \arg\min_{\boldsymbol{y} \in \mathbb{R}^d} \sum_{i=1}^{n} \|\boldsymbol{y} - \boldsymbol{x}_i\|_2 .$$

**Example 3.** Suppose that we have a collection of five clients' updates $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5\}$ such that $\|\boldsymbol{x}_i\|_2 = C$ for $i \in [5]$. Further, we assume $f = 1$ Byzantine client is $\boldsymbol{x}_5$. There exist client placements where the Byzantine user-level sensitivity of $\texttt{GM}$ is $2C$.

**Proof of Example 3.** We reuse the instances from Example 2 and consider two adjacent configurations, where clients are divided into two clusters $\mathcal{C}_1$ and $\mathcal{C}_2$. Let the clients in each cluster reside at the same point, i.e., $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = 0$ for $i, j \in \mathcal{C}_1$ and for $i, j \in \mathcal{C}_2$, respectively. Moreover, let $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = 2C$ for all $i \in \mathcal{C}_1$ and all $j \in \mathcal{C}_2$. More specifically, let

$$\boldsymbol{x}_i = \begin{cases} (\underbrace{C, 0, \ldots, 0}_{d \text{ dimensions}}), & \text{if } i \in \mathcal{C}_1; \\ (\underbrace{-C, 0, \ldots, 0}_{d \text{ dimensions}}), & \text{if } i \in \mathcal{C}_2. \end{cases}$$

**Configuration 1.** Let $\mathcal{C}_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, and $\mathcal{C}_2 = \{\boldsymbol{x}_4, \boldsymbol{x}_5\}$. It can be readily checked that

$$\texttt{GM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5) = (C, 0, \ldots, 0).$$

**Configuration 2.** Without loss of generality, we replace client $\boldsymbol{x}_1$ in $\mathcal{C}_1$ with an auxiliary client $\boldsymbol{x}' \in \mathcal{C}_2$. Therefore, it holds that $\mathcal{C}_1 = \{\boldsymbol{x}_2, \boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}'\}$. It can be readily checked that

$$\texttt{GM}(\boldsymbol{x}', \ldots, \boldsymbol{x}_5) = (-C, 0, \ldots, 0).$$

Therefore, the Byzantine user-level sensitivity of $\texttt{CWMed}$ in this instance is $2C$. □

**Proposition 4.** *Let $n \in \mathbb{N}$ and $f < n/2$. The Byzantine-robust user-level sensitivity of $\texttt{GM}$ is $2C$ if $\|\boldsymbol{x}_i\|_2 \leq C$ for $i \in [n]$.*

**Proof of Proposition 4.** We prove this by contradiction based on Example 3. Suppose that we have the Byzantine user-level sensitivity of $\texttt{GM}$ is smaller than $2C$. Contradiction with Example 3. Therefore, it must be true that $\Delta(\texttt{GM})$ is at least $2C$.

Next, we show that $2C$ is achievable. Since $\texttt{GM}$ is a convex operation over a set of inputs within a convex set ($\|\boldsymbol{x}\|_2 \leq C$). It holds that $\|\texttt{GM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\|_2 \leq C$. Without loss of generality, we assume $\boldsymbol{x}_1 \in \mathcal{G}$ is replaceped with $\boldsymbol{x}'$. It follows that

$$\Delta(\texttt{GM}) = \|\texttt{GM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \texttt{GM}(\boldsymbol{x}', \ldots, \boldsymbol{x}_n)\|_2 \leq \|\texttt{GM}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\|_2 + \|\texttt{GM}(\boldsymbol{x}', \ldots, \boldsymbol{x}_n)\|_2 \leq 2C.$$

□

# G Convergence Analysis

## G.1 Shared Intermediate Results

For ease of notation, let $\overline{\boldsymbol{M}}^t \triangleq \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \boldsymbol{M}_i^t$, $\overline{\boldsymbol{m}}^t \triangleq \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \boldsymbol{m}_i^t$, and $\overline{\boldsymbol{g}}^t \triangleq \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \boldsymbol{g}_i^t$.

### G.1.1 Proof of Descent Lemma

**Lemma G.10.** *Under Assumption 2, it holds that*

$$\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1}) - F_{\mathcal{G}}(\boldsymbol{\theta}^t) \mid \mathcal{F}^t\right] \leq -\eta_t\left(\frac{3}{4} - 2\eta_t L\right)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + \frac{\eta_t^2 L d}{2}\sigma_{dp,t}^2$$
$$+ 2\eta_t(1+\eta_t L)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{M}}^t\right\|_2^2 \Big| \mathcal{F}^t\right] + \frac{2\kappa\eta_t(1+\eta_t L)}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{M}_i^t - \overline{\boldsymbol{M}}^t\right\|_2^2 \Big| \mathcal{F}^t\right].$$

**Proof of Lemma G.10.** By Assumption 2, it holds that

$$F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1}) - F_{\mathcal{G}}(\boldsymbol{\theta}^t) \leq \underbrace{\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right\rangle}_{(A)} + \underbrace{\frac{L}{2}\left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right\|_2^2}_{(B)}.$$

**Bounding $(A)$.**

$$\begin{aligned}
&\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right\rangle \\
&= \eta_t\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \overline{\boldsymbol{M}}^t - R^t\right\rangle - \eta_t\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\rangle - \eta_t\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 - \eta_t\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \boldsymbol{X}_{dp,t}^t\right\rangle \\
&\overset{(a)}{\leq} \frac{\eta_t}{8}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + 2\eta_t\left\|R^t - \overline{\boldsymbol{M}}^t\right\|_2^2 + \frac{\eta_t}{8}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + 2\eta_t\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 - \eta_t\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \\
&\quad - \eta_t\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \boldsymbol{X}_{dp,t}^t\right\rangle \\
&= -\frac{3\eta_t}{4}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + 2\eta_t\left\|R^t - \overline{\boldsymbol{M}}^t\right\|_2^2 + 2\eta_t\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 - \eta_t\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \boldsymbol{X}_{dp,t}^t\right\rangle,
\end{aligned}$$

where inequality $(a)$ follows from Young's inequality. Taking expectation conditional on $\mathcal{F}^t$, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right\rangle \mid \mathcal{F}^t\right] \\
&\leq -\frac{3\eta_t}{4}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + 2\eta_t\mathbb{E}\left[\left\|R^t - \overline{\boldsymbol{M}}^t\right\|_2^2 \mid \mathcal{F}^t\right] + 2\eta_t\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] - \eta_t\left\langle \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \mathbb{E}\left[\boldsymbol{X}_{dp,t}^t \mid \mathcal{F}^t\right]\right\rangle \\
&= -\frac{3\eta_t}{4}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + 2\eta_t\mathbb{E}\left[\left\|R^t - \overline{\boldsymbol{M}}^t\right\|_2^2 \mid \mathcal{F}^t\right] + 2\eta_t\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right].
\end{aligned}$$

**Bounding $(B)$.**

$$\begin{aligned}
\frac{L}{2}\left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right\|_2^2 &= \frac{\eta_t^2 L}{2}\left\|R^t - \overline{\boldsymbol{M}}^t + \overline{\boldsymbol{M}}^t + \boldsymbol{X}_{dp,t}^t\right\|_2^2 \\
&= \frac{\eta_t^2 L}{2}\left(\left\|R^t - \overline{\boldsymbol{M}}^t + \overline{\boldsymbol{M}}^t\right\|_2^2 + \left\|\boldsymbol{X}_{dp,t}^t\right\|_2^2 + 2\left\langle R^t - \overline{\boldsymbol{M}}^t + \overline{\boldsymbol{M}}^t, \boldsymbol{X}_{dp,t}^t\right\rangle\right).
\end{aligned}$$

Taking expectation conditional on $\mathcal{F}^t$, it holds that

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}\left[\left\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^t\right\|_2^2 \mid \mathcal{F}^t\right] = \frac{\eta_t^2 L}{2}\Big(\mathbb{E}\left[\left\|R^t-\overline{\boldsymbol{M}}^t+\overline{\boldsymbol{M}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)+\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right]+\mathbb{E}\left[\left\|\boldsymbol{X}_{dp,t}^t\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\quad + 2\mathbb{E}\left[\left\langle R^t-\overline{\boldsymbol{M}}^t+\overline{\boldsymbol{M}}^t, \boldsymbol{X}_{dp,t}^t\right\rangle \mid \mathcal{F}^t\right]\Big) \\
&\overset{(b)}{=} \frac{\eta_t^2 L}{2}\Big(\mathbb{E}\left[\left\|R^t-\overline{\boldsymbol{M}}^t+\overline{\boldsymbol{M}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)+\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right]+\mathbb{E}\left[\left\|\boldsymbol{X}_{dp,t}^t\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\quad + 2\left\langle \mathbb{E}\left[R^t-\overline{\boldsymbol{M}}^t+\overline{\boldsymbol{M}}^t \mid \mathcal{F}^t\right], \mathbb{E}\left[\boldsymbol{X}_{dp,t}^t \mid \mathcal{F}^t\right]\right\rangle\Big) \\
&= \frac{\eta_t^2 L}{2}\left(\left\|R^t-\overline{\boldsymbol{M}}^t+\overline{\boldsymbol{M}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)+\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + d\sigma_{dp,t-1}^2\right) \\
&\overset{(c)}{\leq} \eta_t^2 L\left(2\left\|R^t-\overline{\boldsymbol{M}}^t\right\|_2^2+2\left\|\overline{\boldsymbol{M}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2+2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2+\frac{d}{2}\sigma_{dp,t-1}^2\right) \\
&\overset{(d)}{\leq} \eta_t^2 L\left(\frac{2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{M}_i^t-\overline{\boldsymbol{M}}^t\right\|_2^2+2\left\|\overline{\boldsymbol{M}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2+2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2+\frac{d}{2}\sigma_{dp,t-1}^2\right),
\end{aligned}
$$

where equality $(b)$ holds because $\boldsymbol{X}_{dp}^t$ and $R^t$ are independent, inequality $(c)$ follows from Jensen's inequality, and inequality $(d)$ follows from the definition of robust aggregator.

**Putting them together.**

$$
\begin{aligned}
\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1})-F_{\mathcal{G}}(\boldsymbol{\theta}^t) \mid \mathcal{F}^t\right] \leq -\eta_t\left(\frac{3}{4}-2\eta_t L\right)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + \frac{2\kappa\eta_t(1+\eta_t L)}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{M}_i^t-\overline{\boldsymbol{M}}^t\right\|_2^2 \mid \mathcal{F}^t\right] \\
+ 2\eta_t(1+\eta_t L)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2 \mid \mathcal{F}^t\right] + \frac{\eta_t^2 L d}{2}\sigma_{dp,t-1}^2.
\end{aligned}
$$

□

### G.1.2 Proof of Global Momentum Estimation Error

**Lemma G.11.** *Let $\mathcal{F}^{t-1}$ be the sigma algebra by the randomness up to round $t-1$. Suppose that Assumptions 2, and 4 hold, we have for some absolute constant $k_2 > 0$*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \leq (1-\beta_t)^2(1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|} + C_{\boldsymbol{M},1}\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
+ C_{\boldsymbol{M},2}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + C_{\boldsymbol{M},3}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
+ C_{\boldsymbol{M},4}\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + C_{\boldsymbol{M},5}d\sigma_{dp,t-1}^2.
\end{aligned}
$$

*where*

$$
\begin{aligned}
C_{\boldsymbol{M},1}(\alpha_t,\beta_t,\eta_{t-1}L) &\triangleq \beta_t\left(\beta_t+\frac{(1-\beta_t)\alpha_t+(1-\beta_{t-1})\alpha_t}{2k_2}+\eta_{t-1}L(\beta_t+(1-\beta_t)\alpha_t)\right) \\
&\quad + 3\eta_{t-1}L\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(1+\eta_{t-1}L\right)+\beta_t(1-\beta_t)\right), \\
C_{\boldsymbol{M},2}(\alpha_t,\beta_t,\eta_{t-1}L) &\triangleq (1-\beta_t)\alpha_t\left[\left((1-\beta_t)\alpha_t+2k_2\beta_t\right)\left(1+\eta_{t-1}L\right)+(1-\beta_t)\beta_t\right], \\
C_{\boldsymbol{M},3}(\alpha_t,\beta_t,\eta_{t-1}L) &\triangleq 3\eta_{t-1}L\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(1+\eta_{t-1}L\right)+\beta_t(1-\beta_t)\right), \\
C_{\boldsymbol{M},4}(\alpha_t,\beta_t,\eta_{t-1}L) &\triangleq \beta_t(1-\beta_{t-1})\left(2k_2\beta_t+(1-\beta_t)\alpha_t+\frac{\eta_{t-1}L(1-\beta_t)}{1-\beta_{t-1}}(\beta_t+(1-\beta_t)\alpha_t)\right) \\
&\quad + 3(1-\beta_{t-1})^2\eta_{t-1}L\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(1+\eta_{t-1}L\right)+\beta_t(1-\beta_t)\right) \\
C_{\boldsymbol{M},5}(\alpha_t,\beta_t,\eta_{t-1}L) &\triangleq \left(\beta_t^2+\eta_{t-1}L\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(\left(\beta_t+(1-\beta_t)\alpha_t\right)\left(1+\eta_{t-1}L\right)+\beta_t(1-\beta_t)\right)\right).
\end{aligned}
$$

**Proof of Lemma G.11.** We know that

$$
\begin{aligned}
&\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) = \beta_t \widetilde{R}^{t-1} + (1-\beta_t)\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \\
&= \beta_t \boldsymbol{X}_{dp}^{t-1} + \beta_t \left(R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right) + \beta_t \left(\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) + (1-\beta_t)\left(\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right) - \beta_t \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}\right. \\
&= \beta_t \boldsymbol{X}_{dp}^{t-1} + \beta_t \left(R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right) + \beta_t \left(\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) + (1-\beta_t)\left(\alpha_t \overline{\boldsymbol{m}}^{t-1} + (1-\alpha_t)\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right) \\
&\quad - \beta \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) \\
&= \beta_t \boldsymbol{X}_{dp}^{t-1} + \beta_t \left(R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right) + \beta_t \left(\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) + (1-\beta_t)\alpha_t \left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) \\
&\quad + (1-\beta_t)(1-\alpha_t)\left(\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right) - (\beta_t + (1-\beta_t)\alpha_t)\left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \\
&= \left\|\beta_t \boldsymbol{X}_{dp}^{t-1}\right\|_2^2 + \left\|\beta_t \left(R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right)\right\|_2^2 + \left\|\beta_t \left(\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\|_2^2 + \left\|(1-\beta_t)\alpha_t \left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\|_2^2 \\
&\quad + \left\|(1-\beta_t)(1-\alpha_t)\left(\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2 + \left\|(\beta_t + (1-\beta_t)\alpha_t)\left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2 \\
&\quad + 2\beta_t(1-\beta_t)\alpha_t \left\langle \overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}), \overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\rangle \\
&\quad + 2\left(\beta_t + (1-\beta_t)\alpha_t\right)(1-\beta_t)\alpha_t \left\langle \overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}), \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\rangle \\
&\quad + 2\beta_t(1-\beta_t)\alpha_t \left\langle R^{t-1} - \overline{\boldsymbol{M}}^{t-1}, \overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\rangle \\
&\quad + 2\beta_t(\beta_t + (1-\beta_t)\alpha_t)\left\langle R^{t-1} - \overline{\boldsymbol{M}}^{t-1}, \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\rangle \\
&\quad + 2\beta_t^2 \left\langle R^{t-1} - \overline{\boldsymbol{M}}^{t-1}, \overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\rangle \\
&\quad + 2\beta_t\left(\beta_t + (1-\beta_t)\alpha_t\right)\left\langle \overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}), \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\rangle \\
&\quad + 2(1-\beta_t)(1-\alpha_t)\left\langle \overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \beta_t R^{t-1} - \beta_t \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) + (1-\beta_t)\alpha_t \left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\rangle \\
&\quad - 2(1-\beta_t)(1-\alpha_t)\left(\beta_t + (1-\beta_t)\alpha_t\right)\left\langle \overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\rangle \\
&\quad + 2\left\langle \beta_t \boldsymbol{X}_{dp}^{t-1}, \beta_t R^{t-1} + (1-\beta_t)\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\rangle \\
&\overset{(a)}{\leq} \left\|\beta_t \boldsymbol{X}_{dp}^{t-1}\right\|_2^2 + \left\|\beta_t \left(R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right)\right\|_2^2 + \left\|\beta_t \left(\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\|_2^2 + \left\|(1-\beta_t)\alpha_t \left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\|_2^2 \\
&\quad + \left\|(1-\beta_t)(1-\alpha_t)\left(\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2 + \left\|(\beta_t + (1-\beta_t)\alpha_t)\left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2 \\
&\quad + 2\beta_t(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \\
&\quad + 2\left(\beta_t + (1-\beta_t)\alpha_t\right)(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2 \\
&\quad + 2\beta_t(1-\beta_t)\alpha_t \left\|R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right\|_2 \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \\
&\quad + 2\beta_t(\beta_t + (1-\beta_t)\alpha_t)\left\|R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right\|_2 \left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2 \\
&\quad + 2\beta_t^2 \left\|R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right\|_2 \left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \\
&\quad + 2\beta_t\left(\beta_t + (1-\beta_t)\alpha_t\right)\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2 \\
&\quad + 2(1-\beta_t)(1-\alpha_t)\left\langle \overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \beta_t R^{t-1} - \beta_t \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) + (1-\beta_t)\alpha_t \left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\rangle \\
&\quad - 2(1-\beta_t)(1-\alpha_t)\left(\beta_t + (1-\beta_t)\alpha_t\right)\left\langle \overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\rangle \\
&\quad + 2\left\langle \beta_t \boldsymbol{X}_{dp}^{t-1}, \beta_t R^{t-1} + (1-\beta_t)\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\rangle,
\end{aligned}
$$

where inequality $(a)$ follows from the Cauchy-Schwarz inequality. By the definition of a robust aggregator, it holds that

$$\left\|R^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right\|_2=\sqrt{\left\|R^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right\|_2^2}\leq\sqrt{\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{M}_i^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right\|_2^2}$$
$$\overset{(25)}{\leq}(1-\beta_{t-1})\sqrt{\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2}. \tag{22}$$

Let $(A)$ define the terms that include $\|R^{t-1}-\overline{\boldsymbol{M}}^{t-1}\|_2$.

$$\begin{aligned}
(A)\triangleq&\left\|\beta_t\left(R^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right)\right\|_2^2+2\beta_t(1-\beta_t)\alpha_t\left\|R^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right\|_2\left\|\overline{\boldsymbol{m}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2\\
&+2\beta_t(\beta_t+(1-\beta_t)\alpha_t)\left\|R^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right\|_2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2+2\beta_t^2\left\|R^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right\|_2\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2\\
\overset{(b)}{\leq}&2\beta_t(1-\beta_t)(1-\beta_{t-1})\alpha_t\sqrt{\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2}\left\|\overline{\boldsymbol{m}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2\\
&+2\eta_{t-1}L\beta_t(1-\beta_t)(\beta_t+(1-\beta_t)\alpha_t)\sqrt{\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2}\left\|R^{t-1}+X_{dp}^{t-1}\right\|_2\\
&+2\beta_t^2(1-\beta_{t-1})\sqrt{\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2}\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2+\beta_t^2(1-\beta_{t-1})^2\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2\\
\overset{(c)}{\leq}&\beta_t^2(1-\beta_{t-1})^2\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2+\beta_t(1-\beta_t)(1-\beta_{t-1})\alpha_t\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2\\
&+\beta_t(1-\beta_t)(1-\beta_{t-1})\alpha_t\left\|\overline{\boldsymbol{m}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\\
&+\eta_{t-1}L\beta_t(1-\beta_t)(\beta_t+(1-\beta_t)\alpha_t)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2+\eta_{t-1}L\beta_t(1-\beta_t)(\beta_t+(1-\beta_t)\alpha_t)\left\|R^{t-1}+X_{dp}^{t-1}\right\|_2^2\\
&+2k_2\beta_t^2(1-\beta_{t-1})\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2+\frac{\beta_t^2}{2k_2}(1-\beta_{t-1})\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\\
=&\beta_t(1-\beta_{t-1})\left(2k_2\beta_t+(1-\beta_t)\alpha_t+\frac{\eta_{t-1}L(1-\beta_t)}{1-\beta_{t-1}}(\beta_t+(1-\beta_t)\alpha_t)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2\\
&+\beta_t(1-\beta_t)(1-\beta_{t-1})\alpha_t\left\|\overline{\boldsymbol{m}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2+\frac{\beta_t^2}{2k_2}(1-\beta_{t-1})\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\\
&+\eta_{t-1}L\beta_t(1-\beta_t)(\beta_t+(1-\beta_t)\alpha_t)\left\|R^{t-1}+X_{dp}^{t-1}\right\|_2^2,
\end{aligned}$$

where inequality $(b)$ follows from (22), and inequality $(c)$ follows from Young's inequality. Let $(B)$ define the terms that include $\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\|_2$.

$$
\begin{aligned}
(B) &\triangleq 2\beta_t(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \\
&\quad + 2\left(\beta_t + (1-\beta_t)\alpha_t\right)(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t})\right\|_2 \\
&\quad + 2\beta_t\left(\beta_t + (1-\beta_t)\alpha_t\right) \left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t})\right\|_2 \\
&\overset{(d)}{\leq} 2\beta_t(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \\
&\quad + 2\eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|R^{t-1} + \boldsymbol{X}_{dp}^{t-1}\right\|_2 \\
&\quad + 2\eta_{t-1}L\beta_t\left(\beta_t + (1-\beta_t)\alpha_t\right) \left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2 \left\|R^{t-1} + \boldsymbol{X}_{dp}^{t-1}\right\|_2 \\
&\overset{(e)}{\leq} \beta_t(1-\beta_t)\frac{\alpha_t}{2k_2}\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 2k_2\beta_t(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\quad + \eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)(1-\beta_t)\alpha_t \left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + \eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)(1-\beta_t)\alpha_t \left\|R^{t-1} + \boldsymbol{X}_{dp}^{t-1}\right\|_2^2 \\
&\quad + \eta_{t-1}L\beta_t\left(\beta_t + (1-\beta_t)\alpha_t\right) \left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + \eta_{t-1}L\beta_t\left(\beta_t + (1-\beta_t)\alpha_t\right) \left\|R^{t-1} + \boldsymbol{X}_{dp}^{t-1}\right\|_2^2 \\
&= \beta_t\left((1-\beta_t)\frac{\alpha_t}{2k_2} + \eta_{t-1}L(\beta_t + (1-\beta_t)\alpha_t)\right)\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\quad + (1-\beta_t)\alpha_t\left(2k_2\beta_t + \eta_{t-1}L(\beta_t + (1-\beta_t)\alpha_t)\right)\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + \eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)^2 \left\|R^{t-1} + \boldsymbol{X}_{dp}^{t-1}\right\|_2^2,
\end{aligned}
$$

where inequality $(d)$ follows from Assumption 2, and inequality $(e)$ follows from Young's inequality and $k_2$ is an absolute constant. Putting them together, it follows, by taking expectation conditional on $\mathcal{F}^t$, that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t})\right\|_2^2 \mid \mathcal{F}^t\right] &\leq \beta_t\left(\beta_t + \frac{(1-\beta_t)\alpha_t + (1-\beta_{t-1})\alpha_t}{2k_2} + \eta_{t-1}L(\beta_t + (1-\beta_t)\alpha_t)\right)\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\quad + (1-\beta_t)\alpha_t\left[\left((1-\beta_t)\alpha_t + 2k_2\beta_t\right)(1+\eta_{t-1}L) + (1-\beta_t)\beta_t\right]\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\quad + (1-\beta_t)^2(1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|} + \beta_t^2\mathbb{E}\left[\left\|\boldsymbol{X}_{dp,t}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&\quad + \eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(\left(\beta_t + (1-\beta_t)\alpha_t\right)(1+\eta_{t-1}L) + \beta_t(1-\beta_t)\right)\mathbb{E}\left[\left\|R^{t-1} + \boldsymbol{X}_{dp,t}^{t-1}\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\quad + \beta_t(1-\beta_{t-1})\left(2k_2\beta_t + (1-\beta_t)\alpha_t + \frac{\eta_{t-1}L(1-\beta_t)}{1-\beta_{t-1}}(\beta_t + (1-\beta_t)\alpha_t)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \\
&\quad + 2\left\langle \beta_t\boldsymbol{X}_{dp,t}^{t-1}, \beta_t R^{t-1} + (1-\beta_t)\mathbb{E}\left[\overline{\boldsymbol{m}}^{t} \mid \mathcal{F}^t\right] - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t})\right\rangle.
\end{aligned}
$$

Taking expectation conditional on $\mathcal{F}^{t-1}$, it holds that

$$
\begin{aligned}
\mathbb{E}\left[\left\|R^{t-1} + \boldsymbol{X}_{dp}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] &\overset{(f)}{=} \mathbb{E}\left[\left\|R^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + \mathbb{E}\left[\left\|\boldsymbol{X}_{dp,t}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + 2\left\langle \mathbb{E}\left[R^{t-1} \mid \mathcal{F}^{t-1}\right], \mathbb{E}\left[\boldsymbol{X}_{dp,t}^{t-1} \mid \mathcal{F}^{t-1}\right]\right\rangle \\
&\overset{(g)}{=} \mathbb{E}\left[\left\|R^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + d\sigma_{dp,t-1}^2, \qquad (23)
\end{aligned}
$$

where equality $(f)$ follows from independent DP noise, and equality $(g)$ follows from zero-mean DP noise. It remains to bound $\|R^{t-1}\|^2$

$$\begin{aligned}
\left\|R^{t-1}\right\|_2^2 &= \left\|R^{t-1} - \overline{\boldsymbol{M}}^{t-1} + \overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1}) + \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\overset{(g)}{\leq} 3\left\|R^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right\|_2^2 + 3\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 3\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\overset{(h)}{\leq} \frac{3\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{M}_i^{t-1} - \overline{\boldsymbol{M}}^{t-1}\right\|_2^2 + 3\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 3\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\overset{(i)}{\leq} \frac{3(1-\beta_{t-1})^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right\|_2^2 + 3\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 3\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2,
\end{aligned} \tag{24}$$

where inequality $(g)$ follows from Jensen's inequality, inequality $(h)$ follows from Definition 2, and inequality $(i)$ follows from (22). Therefore, it further holds, by putting the above back, that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \leq (1-\beta_t)^2(1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|} \\
&+ \beta_t\left(\beta_t + \frac{(1-\beta_t)\alpha_t + (1-\beta_{t-1})\alpha_t}{2k_2} + \eta_{t-1}L(\beta_t + (1-\beta_t)\alpha_t)\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&+ 3\eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(1 + \eta_{t-1}L\right) + \beta_t(1-\beta_t)\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&+ (1-\beta_t)\alpha_t\left[\left((1-\beta_t)\alpha_t + 2k_2\beta_t\right)\left(1 + \eta_{t-1}L\right) + (1-\beta_t)\beta_t\right]\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&+ 3\eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(1 + \eta_{t-1}L\right) + \beta_t(1-\beta_t)\right)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&+ \beta_t(1-\beta_{t-1})\left(2k_2\beta_t + (1-\beta_t)\alpha_t + \frac{\eta_{t-1}L(1-\beta_t)}{1-\beta_{t-1}}(\beta_t + (1-\beta_t)\alpha_t)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&+ 3(1-\beta_{t-1})^2\eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(1 + \eta_{t-1}L\right) + \beta_t(1-\beta_t)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&+ \left(\beta_t^2 + \eta_{t-1}L\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(\left(\beta_t + (1-\beta_t)\alpha_t\right)\left(1 + \eta_{t-1}L\right) + \beta_t(1-\beta_t)\right)\right)d\sigma_{dp,t-1}^2.
\end{aligned}$$

□

### G.1.3 Proof of Client-side Momentum Estimation Error

**Lemma G.12.** *Under Assumptions 2, and 3, it holds that*

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^{t-1}\right] &\leq \alpha_t^2(1+\eta_{t-1}L)\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + (1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|} \\
&+ 3\alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 3\alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&+ 3\alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)\frac{(1-\beta_{t-1})^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + \alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)d\sigma_{dp,t-1}^2.
\end{aligned}$$

**Proof of Lemma G.12.**

$$\begin{aligned}
\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) &= \alpha_t\overline{\boldsymbol{m}}^{t-1} + (1-\alpha_t)\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \\
&= \alpha_t\left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) + (1-\alpha_t)\left(\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right) - \alpha_t\left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 &= (1-\alpha_t)^2\left\|\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + \alpha_t^2\left\|\left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) - \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\|_2^2 \\
&\quad + 2(1-\alpha_t)\alpha_t\left\langle\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t), \left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) - \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\rangle.
\end{aligned}$$

Taking expectation conditional on $\mathcal{F}^t$, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] = (1-\alpha_t)^2\mathbb{E}\left[\left\|\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\quad + \alpha_t^2\left\|\left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) - \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\|_2^2 \\
&\quad + 2(1-\alpha_t)\alpha_t\left\langle\mathbb{E}\left[\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \mid \mathcal{F}^t\right], \left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) - \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\rangle \\
&= \alpha_t^2\left\|\left(\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right) - \left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right)\right\|_2^2 + (1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|} \\
&\overset{(a)}{\leq} \alpha_t^2(1+\eta_{t-1}L)\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + \alpha_t^2(1+\frac{1}{\eta_{t-1}L})\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + (1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|} \\
&\overset{(b)}{\leq} \alpha_t^2(1+\eta_{t-1}L)\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + \alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)\left\|R^{t-1} + \boldsymbol{X}_{dp}^{t-1}\right\|_2^2 + (1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|},
\end{aligned}$$

where inequality $(a)$ follows from Young's inequality, and inequality $(b)$ follows from Assumption 2. Plugging (23) and (24), it holds, by taking expectation conditional on $\mathcal{F}^{t-1}$, that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \overset{(c)}{\leq} \alpha_t^2(1+\eta_{t-1}L)\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + \alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)d\sigma_{dp,t-1}^2 \\
&\quad + 3\alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 3\alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^{t-1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&\quad + 3\alpha_t^2\eta_{t-1}L(1+\eta_{t-1}L)\frac{(1-\beta_{t-1})^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + (1-\alpha_t)^2\frac{\sigma^2}{|\mathcal{G}|},
\end{aligned}$$

where inequality $(c)$ follows from zero-mean and independent DP noise $\boldsymbol{X}_{dp}^{t-1}$. □

### G.2 Strongly-convex objectives

#### G.2.1 Proof of Local Momentum Drift

**Lemma G.13.** *Under Assumptions 2, 3 and 1. When $F_{\mathcal{G}}$ is strongly convex, for any $\alpha_t, \beta_t \in (0,1)$, it holds that*

$$\begin{aligned}
&\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 \middle| \mathcal{F}^{t-1}\right] \leq 2(1-\alpha_t)B^2\eta_{t-1}^2L^2d\sigma_{dp,t-1}^2 + 2(1-\alpha_t)^2\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 \\
&\quad + \left(\alpha_t + 6\kappa(1-\alpha_t)(1-\beta_{t-1})^2B^2\eta_{t-1}^2L^2\right)\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \middle| \mathcal{F}^{t-1}\right] + (1-\alpha_t)G^2 \\
&\quad + 2(1-\alpha_t)B^2\left(1+3\eta_{t-1}^2L^2\right)\left\|\nabla F(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 6(1-\alpha_t)B^2\eta_{t-1}^2L^2\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1} - \nabla F(\boldsymbol{\theta}^{t-1})\right\|_2^2 \middle| \mathcal{F}^{t-1}\right].
\end{aligned}$$

**Proof of Lemma G.13.** We know that

$$\boldsymbol{m}_i^t = \alpha_t\boldsymbol{m}_i^{t-1} + (1-\alpha_t)\boldsymbol{g}_i^t.$$

It follows that

$$\begin{aligned}
&\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\alpha_t\left(\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right) + (1-\alpha_t)\left(\left(\boldsymbol{g}_i^t - \nabla F_i(\boldsymbol{\theta}^t)\right) - \left(\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right)\right. \\
&\quad \left. + (1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2 \\
&= \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\alpha_t\left(\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right) + (1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2 \\
&\quad + \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|(1-\alpha_t)\left(\left(\boldsymbol{g}_i^t - \nabla F_i(\boldsymbol{\theta}^t)\right) - \left(\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right)\right\|_2^2 \\
&\quad + \frac{1-\alpha_t}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\langle\alpha_t\left(\boldsymbol{m}_i^{t-1} - \overline{\boldsymbol{m}}^{t-1}\right) + (1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right), \left(\boldsymbol{g}_i^t - \nabla F_i(\boldsymbol{\theta}^t)\right) - \left(\overline{\boldsymbol{g}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\rangle.
\end{aligned}$$

Taking expectation conditional on $\mathcal{F}^t$, it holds that

$$
\begin{aligned}
&\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\mid\mathcal{F}^t\right]=\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}[\Big\|\alpha_t\left(\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right)+(1-\alpha_t)\left(\left(\boldsymbol{g}_i^t-\nabla F_i(\boldsymbol{\theta}^t)\right)-\left(\overline{\boldsymbol{g}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right)\\
&\quad+(1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\Big\|_2^2\mid\mathcal{F}^t]\\
&=\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\alpha_t\left(\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right)+(1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2\mid\mathcal{F}^t\right]\\
&\quad+\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|(1-\alpha_t)\left(\left(\boldsymbol{g}_i^t-\nabla F_i(\boldsymbol{\theta}^t)\right)-\left(\overline{\boldsymbol{g}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right)\right\|_2^2\mid\mathcal{F}^t\right]\\
&\quad+\frac{1-\alpha_t}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\langle\alpha_t\left(\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right)+(1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right),\left(\boldsymbol{g}_i^t-\nabla F_i(\boldsymbol{\theta}^t)\right)-\left(\overline{\boldsymbol{g}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\rangle\mid\mathcal{F}^t\right]\\
&\overset{(a)}{=}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\alpha_t\left(\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right)+(1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right\|_2^2\mid\mathcal{F}^t\right]\\
&\quad+\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|(1-\alpha_t)\left(\left(\boldsymbol{g}_i^t-\nabla F_i(\boldsymbol{\theta}^t)\right)-\left(\overline{\boldsymbol{g}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\right)\right\|_2^2\mid\mathcal{F}^t\right]\\
&\quad+\frac{1-\alpha_t}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\langle\alpha_t\left(\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right)+(1-\alpha_t)\left(\nabla F_i(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right),\mathbb{E}\left[\left(\boldsymbol{g}_i^t-\nabla F_i(\boldsymbol{\theta}^t)\right)-\left(\overline{\boldsymbol{g}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right)\mid\mathcal{F}^t\right]\right\rangle\\
&\overset{(b)}{\leq}\frac{\alpha_t}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\left(\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right)\right\|_2^2+\frac{1-\alpha_t}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\nabla F_i(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2+2(1-\alpha_t)^2\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2\\
&\overset{(c)}{\leq}\frac{\alpha_t}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\left(\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right)\right\|_2^2+(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2+(1-\alpha_t)G^2+2(1-\alpha_t)^2\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2,
\end{aligned}
$$

where equality $(a)$ follows from independence, inequality $(b)$ follows from Jensen's inequality and Assumption 3, and inequality $(c)$ follows from Assumption 1.

It follows that

$$
\begin{aligned}
(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2&=(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})+\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\\
&\leq 2(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2+2(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\\
&\leq 2(1-\alpha_t)B^2\eta_{t-1}^2L^2\left\|R^{t-1}+\boldsymbol{X}_{dp}^{t-1}\right\|_2^2+2(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2.
\end{aligned}
$$

We know that

$$
\begin{aligned}
\mathbb{E}\left[\left\|R^{t-1}+\boldsymbol{X}_{dp}^{t-1}\right\|_2^2\mid\mathcal{F}^{t-1}\right]&=\mathbb{E}\left[\left\|R^{t-1}\right\|_2^2\mid\mathcal{F}^{t-1}\right]+\mathbb{E}\left[\left\|\boldsymbol{X}_{dp}^{t-1}\right\|_2^2\mid\mathcal{F}^{t-1}\right]+2\left\langle\mathbb{E}\left[R^{t-1}\mid\mathcal{F}^{t-1}\right],\mathbb{E}\left[\boldsymbol{X}_{dp}^{t-1}\mid\mathcal{F}^{t-1}\right]\right\rangle\\
&=\mathbb{E}\left[\left\|R^{t-1}\right\|_2^2\mid\mathcal{F}^{t-1}\right]+d\sigma_{dp}^2,
\end{aligned}
$$

and

$$
\begin{aligned}
\left\|R^{t-1}\right\|_2^2&=\left\|R^{t-1}-\overline{\boldsymbol{M}}^{t-1}+\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})+\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\\
&\overset{(22)}{\leq}\frac{3(1-\beta_{t-1})^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2+3\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2+3\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|R^{t-1}+\boldsymbol{X}_{dp}^{t-1}\right\|_2^2\mid\mathcal{F}^{t-1}\right]\\
&\leq\frac{3(1-\beta_{t-1})^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2\mid\mathcal{F}^{t-1}\right]+3\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\mid\mathcal{F}^{t-1}\right]+3\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2\\
&\quad+d\sigma_{dp,t-1}^2.
\end{aligned}
$$

Putting it back in, it holds that

$$
\begin{aligned}
&(1-\alpha_t)B^2\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \leq 2(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\quad + \frac{6\kappa(1-\alpha_t)B^2\eta_t^2L^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{M}_i^{t-1}-\overline{\boldsymbol{M}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + 2(1-\alpha_t)B^2\eta_t^2L^2d\sigma_{dp,t-1}^2 \\
&\quad + 6(1-\alpha_t)B^2\eta_t^2L^2\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + 6(1-\alpha_t)B^2\eta_t^2L^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\leq 2(1-\alpha_t)B^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + \frac{6\kappa(1-\alpha_t)(1-\beta_{t-1})^2B^2\eta_t^2L^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&\quad + 6(1-\alpha_t)B^2\eta_t^2L^2\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right] + 6(1-\alpha_t)B^2\eta_t^2L^2\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \\
&\quad + 2(1-\alpha_t)B^2\eta_t^2L^2d\sigma_{dp,t-1}^2.
\end{aligned}
$$

Combining everything together, it holds that

$$
\begin{aligned}
\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2 \mid \mathcal{F}^{t-1}\right] &\leq 2(1-\alpha_t)B^2\eta_{t-1}^2L^2d\sigma_{dp,t-1}^2 + (1-\alpha_t)G^2 + 2(1-\alpha_t)^2\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 \\
&\quad + \left(\alpha_t + 6\kappa(1-\alpha_t)(1-\beta_{t-1})^2B^2\eta_{t-1}^2L^2\right)\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^{t-1}-\overline{\boldsymbol{m}}^{t-1}\right\|_2^2 \mid \mathcal{F}^{t-1}\right] \\
&\quad + 2(1-\alpha_t)B^2\left(1+3\eta_{t-1}^2L^2\right)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 + 6(1-\alpha_t)B^2\eta_{t-1}^2L^2\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t-1}-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t-1})\right\|_2^2 \mid \mathcal{F}^{t-1}\right].
\end{aligned}
$$

□

### G.2.2 Proof of Theorem 3

**Theorem G.5** (Strongly-convex, formal)**.** *Suppose that Assumptions 2, 3, 1, and 4 hold. When* $\alpha_t = 1 - 4900\eta_t L$, $\beta_t = 1 - 55\eta_t L$, $\eta_t = \frac{10}{\mu(t+a\frac{L}{\mu})}$ *such that* $a \geq 49000$, *and* $B^2 \leq \frac{1}{600\kappa}$, *it holds that*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\theta}^T-\boldsymbol{\theta}^*\right\|_2^2\right] \lesssim \frac{L^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{\mu^3T^2} + \frac{L\sigma^2}{\mu^3T^2|\mathcal{G}|} + \frac{L\kappa}{\mu^3T^2}\left(\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2+G^2\right) \\
+ L(1+\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^3T} + \left(\frac{L}{\mu}\right)\frac{\log(1/\delta)C^2d}{\mu^2\epsilon^2m^2}\left(1+\frac{f}{m}\right) + \frac{\kappa G^2}{\mu^2}.
\end{aligned}
$$

*Proof.* Define the Lyapunov function as

$$
\begin{aligned}
W^t &\triangleq \mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^t) - F_{\mathcal{G}}^* + \frac{1}{20L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2 + \frac{1}{15L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{m}}^t\right\|_2^2 + \frac{2\kappa}{L}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]; \\
V^t &\triangleq \left(t+a\frac{L}{\mu}\right)^2 W^t \\
&= \left(t+a\frac{L}{\mu}\right)^2\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^t) - F_{\mathcal{G}}^* + \frac{1}{20L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2 + \frac{1}{15L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{m}}^t\right\|_2^2 + \frac{2\kappa}{L}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right].
\end{aligned}
$$

For ease of notation, let $\widehat{t} = t + \frac{aL}{\mu}$. It follows that

$$
\begin{aligned}
V^{t+1} - V^t &= (\widehat{t}+1)^2W^{t+1} - \widehat{t}^2W^t \\
&= (\widehat{t}+1)^2(W^{t+1}-W^t) + \left((\widehat{t}+1)^2-\widehat{t}^2\right)W^t \\
&= (\widehat{t}+1)^2(W^{t+1}-W^t) + \left(2\widehat{t}+1\right)W^t.
\end{aligned}
$$

Next, we have

$$\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1}) - F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right] \leq -\eta_t\left(\frac{3}{4} - 2\eta_t L\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] + \frac{2\kappa\eta_t(1+\eta_t L)(1-\beta_t)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right]$$
$$+ 2\eta_t(1+\eta_t L)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{M}}^t\right\|_2^2\right] + \frac{\eta_t^2 L d}{2}\sigma_{dp,t}^2,$$

It follows that

$$\begin{aligned}
&\frac{z_1}{L}\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^{t+1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1})\right\|_2^2 - \left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \leq \frac{z_1}{L}(1-\beta_{t+1})^2(1-\alpha_{t+1})^2\frac{\sigma^2}{|\mathcal{G}|} \\
&+ \frac{z_1}{L}\left(\beta_{t+1}\left(\beta_{t+1} + \frac{(1-\beta_{t+1})\alpha_{t+1} + (1-\beta_t)\alpha_{t+1}}{2k_2} + \eta_t L(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1})\right) - 1\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \\
&+ 3z_1\eta_t\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1})\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \\
&+ \frac{z_1}{L}(1-\beta_{t+1})\alpha_{t+1}\left[\left((1-\beta_{t+1})\alpha_{t+1} + 2k_2\beta_{t+1}\right)(1+\eta_t L) + (1-\beta_{t+1})\beta_{t+1}\right]\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \\
&+ 3z_1\eta_t\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1})\right)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \\
&+ \frac{z_1}{L}(1-\beta_t)\beta_{t+1}\left(2k_2\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1} + \frac{\eta_t L(1-\beta_{t+1})}{1-\beta_t}(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1})\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 \mid \mathcal{F}^t\right] \\
&+ 3z_1(1-\beta_t)^2\eta_t\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1})\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 \mid \mathcal{F}^t\right] \\
&+ \frac{z_1}{L}\left(\beta_{t+1}^2 + \eta_t L\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1})\right)\right)d\sigma_{dp,t}^2,
\end{aligned}$$

where the inequality follows from Lemma G.11. Also,

$$\begin{aligned}
&\frac{z_2}{L}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^{t+1} - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1})\right\|_2^2 - \left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \leq \frac{z_2}{L}\left(\alpha_{t+1}^2(1+\eta_t L) - 1\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\quad + 3z_2\alpha_{t+1}^2\eta_t(1+\eta_t L)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + 3z_2\alpha_{t+1}^2\eta_t(1+\eta_t L)\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\quad + 3z_2\alpha_{t+1}^2\eta_t(1+\eta_t L)\frac{(1-\beta_t)^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 \mid \mathcal{F}^t\right] \\
&\quad + \frac{z_2}{L}(1-\alpha_{t+1})^2\frac{\sigma^2}{|\mathcal{G}|} + z_2\alpha_{t+1}^2\eta_t(1+\eta_t L)d\sigma_{dp,t}^2,
\end{aligned}$$

where the inequality follows from Lemma G.12. Furthermore,

$$\begin{aligned}
&\frac{\kappa z_3}{L}\mathbb{E}\Big[\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^{t+1} - \overline{\boldsymbol{m}}^{t+1}\right\|_2^2 - \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 \mid \mathcal{F}^t\Big] \\
&\leq \frac{\kappa z_3}{L}\left(\alpha_{t+1} + 6\kappa(1-\alpha_{t+1})(1-\beta_t)^2B^2\eta_t^2L^2 - 1\right)\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 \\
&\quad + \frac{\kappa z_3}{L}(1-\alpha_{t+1})G^2 + 2(1-\alpha_{t+1})^2\frac{\kappa z_3}{L}\left(1 + \frac{1}{|\mathcal{G}|}\right)\sigma^2 \\
&\quad + 2(1-\alpha_{t+1})B^2(1+3\eta_t^2L^2)\frac{\kappa z_3}{L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 + 6(1-\alpha_{t+1})B^2\eta_t^2L\kappa z_3\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2 \\
&\quad + 2(1-\alpha_{t+1})B^2\eta_t^2L\kappa z_3 d\sigma_{dp,t}^2,
\end{aligned}$$

where the inequality follows from Lemma 5. Putting them together and taking expectation over the remaining randomness, it holds that

$$
\begin{aligned}
W^{t+1} - W^t \leq & -\eta_t \left( \frac{3}{4} - 2\eta_t L - 3z_1 \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right) \left( \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1}) \right) \right) \mathbb{E}\left[ \left\| \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \right\|_2^2 \right] \\
& - \eta_t \left( -3z_2\alpha_{t+1}^2(1+\eta_t L) - 2(1-\alpha_{t+1})B^2(1+3\eta_t^2 L^2)\frac{\kappa z_3}{\eta_t L} \right) \mathbb{E}\left[ \left\| \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \right\|_2^2 \right] \\
& + \left( 2\eta_t(1+\eta_t L) + 3z_2\alpha_{t+1}^2\eta_t(1+\eta_t L) + 6(1-\alpha_{t+1})B^2\eta_t^2 L\kappa z_3 \right) \mathbb{E}\left[ \left\| \overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \right\|_2^2 \right] \\
& + \frac{z_1}{L} \left( \beta_{t+1} \left( \beta_{t+1} + \frac{(1-\beta_{t+1})\alpha_{t+1} + (1-\beta_t)\alpha_{t+1}}{2k_2} + \eta_t L(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}) \right) - 1 \right) \mathbb{E}\left[ \left\| \overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \right\|_2^2 \right] \\
& + 3z_1\eta_t \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right) \left( \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1}) \right) \mathbb{E}\left[ \left\| \overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \right\|_2^2 \right] \\
& + \frac{z_1}{L}(1-\beta_{t+1})\alpha_{t+1} \left[ \left( (1-\beta_{t+1})\alpha_{t+1} + 2k_2\beta_{t+1} \right)(1+\eta_t L) + (1-\beta_{t+1})\beta_{t+1} \right] \mathbb{E}\left[ \left\| \overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \right\|_2^2 \right] \\
& + \frac{z_2}{L} \left( \alpha_{t+1}^2(1+\eta_t L) - 1 \right) \mathbb{E}\left[ \left\| \overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) \right\|_2^2 \right] + \frac{2\kappa\eta_t(1+\eta_t L)(1-\beta_t)^2}{|\mathcal{G}|} \sum_{i\in\mathcal{G}} \mathbb{E}\left[ \left\| \boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t \right\|_2^2 \right] \\
& + \frac{z_1}{L}(1-\beta_t)\beta_{t+1} \left( 2k_2\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1} + \frac{\eta_t L(1-\beta_{t+1})}{1-\beta_t}(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}) \right) \frac{\kappa}{|\mathcal{G}|} \sum_{i\in\mathcal{G}} \mathbb{E}\left[ \left\| \boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t \right\|_2^2 \right] \\
& + 3z_1(1-\beta_t)^2\eta_t \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right) \left( \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1}) \right) \frac{\kappa}{|\mathcal{G}|} \sum_{i\in\mathcal{G}} \mathbb{E}\left[ \left\| \boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t \right\|_2^2 \right] \\
& + 3z_2\alpha_{t+1}^2\eta_t(1+\eta_t L)\frac{(1-\beta_t)^2\kappa}{|\mathcal{G}|} \sum_{i\in\mathcal{G}} \mathbb{E}\left[ \left\| \boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t \right\|_2^2 \mid \mathcal{F}^t \right] \\
& + \frac{\kappa z_3}{L} \left( \alpha_{t+1} + 6\kappa(1-\alpha_{t+1})(1-\beta_t)^2 B^2\eta_t^2 L^2 - 1 \right) \frac{1}{|\mathcal{G}|} \sum_{i\in\mathcal{G}} \mathbb{E}\left[ \left\| \boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t \right\|_2^2 \right] \\
& + \frac{(1-\alpha_{t+1})^2}{L} \left( z_1(1-\beta_{t+1})^2 + z_2 + 2\kappa z_3 \right) \left( 1 + \frac{1}{|\mathcal{G}|} \right) \sigma^2 + \frac{\kappa z_3}{L}(1-\alpha_{t+1})G^2 \\
& + \left( \frac{\eta_t^2 L}{2} + \frac{z_1}{L} \left( \beta_{t+1}^2 + \eta_t L \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right) \left( \left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1}) \right) \right) \right) d\sigma_{dp,t}^2 \\
& + \left( z_2\alpha_{t+1}^2\eta_t(1+\eta_t L) + 2(1-\alpha_{t+1})B^2\eta_t^2 L\kappa z_3 \right) d\sigma_{dp,t}^2.
\end{aligned}
$$

Let $\alpha_t = 1 - k_1\eta_t L$ and $\beta_t = 1 - k_2\eta_t L$, where $k_2 \leq k_1$. More specifically, $k_1 = 4900$, and $k_2 = 55$. It follows that $\alpha_t \leq \beta_t$. We know that

$$
\frac{\eta_{t+1}}{\eta_t} = \frac{\frac{10}{\mu\left(t+1+a\frac{L}{\mu}\right)}}{\frac{10}{\mu\left(t+a\frac{L}{\mu}\right)}} = \frac{t + a\frac{L}{\mu}}{t+1+a\frac{L}{\mu}} = 1 - \frac{1}{t+1+a\frac{L}{\mu}} \in [\frac{a}{1+a}, 1],
$$

where the last inequality follows from $t \geq 0$ and $L \geq \mu$. Recall that $\eta_t \leq \frac{1}{49000L}$, and $B^2 \leq \frac{1}{40k_1\kappa}$. Next, we bound the coefficients by applying the learning rate condition for $\eta_t$.

**Bounding $C_1$ for $\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\|_2^2\right]$.**

$$
\begin{aligned}
C_1 \triangleq\ & \frac{3}{4} - 2\eta_t L - 3z_1\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1})\right) \\
& - 3z_2\alpha_{t+1}^2(1+\eta_t L) - 2(1-\alpha_{t+1})B^2(1+3\eta_t^2 L^2)\frac{\kappa z_3}{\eta_t L} \\
\geq\ & \frac{3}{4} - 2\eta_t L - 3z_1\left(1+k_2\eta_t L\right)\left(\left(1+k_2\eta_t L\right)(1+\eta_t L) + k_2\eta_t L\right) - 3z_2(1+\eta_t L) - 2k_1 B^2(1+3\eta_t^2 L^2)\kappa z_3 \\
\geq\ & \frac{3}{4} - 2\times\frac{1}{150} - \frac{3}{20}\times\left(1+\frac{8}{150}\right)\times\left(\left(1+\frac{8}{150}\right)\times\left(1+\frac{1}{150}\right) + \frac{8}{150}\right) - \frac{1}{5}\times\left(1+\frac{1}{150}\right) - \frac{4}{40}\left(1+3\times(\frac{1}{150})^2\right) \\
\geq\ & \frac{3}{4} - \frac{1}{75} - 0.19921 - 0.252 - 0.105 \geq \frac{1}{10}.
\end{aligned}
$$

**Bounding $C_2$ for $\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]$.**

$$
\begin{aligned}
C_2 \triangleq\ & \left(2\eta_t(1+\eta_t L) + 3z_2\alpha_{t+1}^2\eta_t(1+\eta_t L) + 6(1-\alpha_{t+1})B^2\eta_t^2 L\kappa z_3\right) \\
& + \frac{z_1}{L}\left(\beta_{t+1}\left(\beta_{t+1} + \frac{(1-\beta_{t+1})\alpha_{t+1} + (1-\beta_{t+1})\alpha_t}{2k_2} + \eta_t L(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1})\right) - 1\right) \\
& + 3z_1\eta_t\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1} + (1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L) + \beta_{t+1}(1-\beta_{t+1})\right) \\
\leq\ & \left(2\eta_t(1+\eta_t L) + 3z_2\eta_t(1+\eta_t L) + 6k_1\eta_t L B^2\eta_t^2 L\kappa z_3\right) \\
& + \frac{\eta_t z_1}{L}\left((L + L(1+k_2\eta_t L)) - k_2 L\frac{\eta_{t+1}}{\eta_t}\right) + 3z_1\left(1+k_2\eta_t L\right)\left(\left(1+k_2\eta_t L\right)(1+\eta_t L) + k_2\eta_t L\right) \\
\leq\ & \eta_t\left(2(1+\frac{1}{150}) + \frac{1}{5}(1+\frac{1}{150}) + \frac{12}{40}\times\frac{1}{150}\right) + \frac{\eta_t}{20}\left(\left(1+1+\frac{1}{10}\right) - k_2\frac{a}{1+a}\right) \\
& + \frac{3}{20}\left(1+\frac{1}{10}\right)\left(\left(1+\frac{1}{10}\right)\left(1+\frac{1}{150}\right) + \frac{1}{10}\right) \\
\leq\ & \frac{\eta_t}{20}\left(50.5 - \frac{ak_2}{1+a}\right) \leq -\eta_t,
\end{aligned}
$$

where the last inequality holds because $k_2 \geq 51.5(1+\frac{1}{a})$.

**Bounding $C_3$ for $\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]$.**

$$
\begin{aligned}
C_3 \triangleq\ & \frac{z_1}{L}(1-\beta_{t+1})\alpha_{t+1}\left[\left((1-\beta_{t+1})\alpha_{t+1} + 2k_2\beta_{t+1}\right)(1+\eta_t L) + (1-\beta_{t+1})\beta_{t+1}\right] + \frac{z_2}{L}\left(\alpha_{t+1}^2(1+\eta_t L) - 1\right) \\
& \leq \eta_t\left(z_1 k_2\left[\left(k_2\eta_t L + 2k_2\right)(1+\eta_t L) + k_2\eta_t L\right] + z_2\left(1 - \frac{\eta_{t+1}k_1}{\eta_t}\right)\right) \\
& \leq \frac{\eta_t}{20}\left(k_2^2\left[\left(\frac{1}{150}+2\right)\left(1+\frac{1}{150}\right) + \frac{1}{10}\right] + \frac{4}{3}\left(1 - \frac{ak_1}{1+a}\right)\right) \\
& \leq \frac{\eta_t}{15}\left(1.6k_2^2 + 1 - \frac{ak_1}{1+a}\right) \leq -\eta_t,
\end{aligned}
$$

where the last inequality holds because $k_1 \geq \left(1.6k_2^2 + 16\right)(1+\frac{1}{a})$.

**Bounding $C_4$ for $\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]/|\mathcal{G}|$.**

$$
\begin{aligned}
C_4 &\triangleq 2\eta_t(1+\eta_t L)(1-\beta_t)^2+\frac{z_1}{L}(1-\beta_{t+1})\beta_{t+1}\left(2k_2\beta_{t+1}+(1-\beta_{t+1})\alpha_{t+1}+\frac{\eta_t L(1-\beta_{t+1})}{1-\beta_t}(\beta_{t+1}+(1-\beta_{t+1})\alpha_{t+1})\right)\\
&\quad+3z_1(1-\beta_t)^2\eta_t\left(\beta_{t+1}+(1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1}+(1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L)+\beta_{t+1}(1-\beta_{t+1})\right)\\
&\quad+3z_2\alpha_{t+1}^2\eta_t(1+\eta_t L)(1-\beta_t)^2+\frac{z_3}{L}\left(\alpha_{t+1}+6\kappa(1-\alpha_{t+1})(1-\beta_t)^2B^2\eta_t^2L^2-1\right)\\
&\le 2\eta_t(1+\eta_t L)(k_2\eta_t L)^2+z_1k_2\eta_t\left(2k_2+k_2\eta_t L+\eta_t L(1+k_2\eta_t L)\right)\\
&\quad+3z_1k_2^2\eta_t^2L^2\eta_t\left(1+k_2\eta_t L\right)\left(\left(1+k_2\eta_t L\right)(1+\eta_t L)+k_2\eta_t L\right)\\
&\quad+3z_2\eta_t(1+\eta_t L)k_2^2\eta_t^2L^2+z_3\left(6\kappa k_1\eta_t k_2^2\eta_t^2L^2B^2\eta_t^2L^2-k_1\eta_{t+1}\right)\\
&\le \eta_t\Bigg(2(1+\frac{1}{150})(\frac{1}{10})^2+\frac{k_2}{20}\left(2k_2+\frac{1}{10}+\frac{1}{150}(1+\frac{1}{10})\right)\\
&\quad+\frac{3}{20}(\frac{1}{10})^2\left(1+\frac{1}{10}\right)\left(\left(1+\frac{1}{10}\right)\left(1+\frac{1}{150}\right)+\frac{1}{10}\right)+\frac{1}{5}(1+\frac{1}{10})(\frac{1}{10})^2+2\left(\frac{1}{40}(\frac{1}{10})^2(\frac{1}{150})^2-\frac{ak_1}{1+a}\right)\Bigg)\\
&\le \eta_t\left(0.021+0.1k_2^2+5.4\times10^{-3}k_2+2\times10^{-3}+2.2\times10^{-3}+2.3\times10^{-8}-\frac{2ak_1}{1+a}\right)\\
&\le \eta_t\left(0.0253+0.1k_2^2+5.4\times10^{-3}k_2-\frac{2ak_1}{1+a}\right)\le-\eta_t,
\end{aligned}
$$

where the last inequality holds because $k_1\ge\left(0.6265+0.05k_2^2+2.7\times10^{-3}k_2\right)\left(1+\frac{1}{a}\right)$. Finally, we bound the remaining noise coefficients.

**Coefficient $C_5$ of $\sigma^2$.**

$$
C_5\triangleq\frac{(1-\alpha_{t+1})^2}{L}\left(z_1(1-\beta_{t+1})^2+z_2+2\kappa z_3\right)\le k_1^2\eta_t^2L\left(\frac{1}{20}\left(\frac{1}{10}\right)^2+\frac{1}{15}+4\kappa\right)\le k_1^2\eta_t^2L\left(1+4\kappa\right),
$$

**Coefficient $C_6$ of $d\sigma_{dp,t}^2$.**

$$
\begin{aligned}
C_6 &\triangleq\left(\frac{\eta_t^2L}{2}+\frac{z_1}{L}\left(\beta_{t+1}^2+\eta_t L\left(\beta_{t+1}+(1-\beta_{t+1})\alpha_{t+1}\right)\left(\left(\beta_{t+1}+(1-\beta_{t+1})\alpha_{t+1}\right)(1+\eta_t L)+\beta_{t+1}(1-\beta_{t+1})\right)\right)\right.\\
&\quad\left.+\left(z_2\alpha_{t+1}^2\eta_{t+1}(1+\eta_t L)+2(1-\alpha_{t+1})B^2\eta_t^2L\kappa z_3\right)\right.\\
&\le\frac{1}{L}\left(\frac{1}{2}\left(\frac{1}{49000}\right)^2+\frac{1}{20}\left(1+\frac{1}{49000}\left(1+\frac{1}{10}\right)\left(\left(1+\frac{1}{10}\right)\left(1+\frac{1}{49000}\right)+\frac{1}{10}\right)\right)\right)\\
&\quad+\frac{1}{L}\left(\frac{1}{15}\frac{1}{49000}(1+\frac{1}{49000})+\frac{2\times2}{40}\left(\frac{1}{49000}\right)^2\right)\le\frac{1}{L}\left(0.051+1.4\times10^{-6}\right)\le\frac{0.06}{L}.
\end{aligned}
$$

Putting them together.

$$
\begin{aligned}
W^{t+1}-W^t &\leq -\frac{\eta_t}{10}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]-\eta_t\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2\right]-\eta_t\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]\\
&\quad-\frac{\kappa\eta_t}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]+k_1^2\eta_t^2L(1+4\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2+2\kappa k_1\eta_tG^2+\frac{0.06}{L}d\sigma_{dp,t}^2\\
&\leq -\frac{\mu\eta_t}{5}\left(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^t)-F_{\mathcal{G}}^*\right]+\frac{z_1}{\mu}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2\right]+\frac{z_2}{\mu}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{\kappa z_3}{\mu}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]\right)\\
&\quad+k_1^2\eta_t^2L(1+4\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2+2\kappa k_1\eta_tG^2+\frac{0.06}{L}d\sigma_{dp,t}^2\\
&\overset{(a)}{\leq} -\frac{\mu\eta_t}{5}\left(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^t)-F_{\mathcal{G}}^*\right]+\frac{z_1}{L}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2\right]+\frac{z_2}{L}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{\kappa z_3}{L}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]\right)\\
&\quad+k_1^2\eta_t^2L(1+4\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2+2\kappa k_1\eta_tG^2+\frac{0.06}{L}d\sigma_{dp,t}^2\\
&=-\frac{\mu\eta_t}{5}W^t+k_1^2\eta_t^2L(1+4\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2+2\kappa k_1\eta_tG^2+\frac{0.06}{L}d\sigma_{dp,t}^2,
\end{aligned}
$$

where inequality $(a)$ holds because $\mu\leq L$. By plugging $\eta_t$ in, we have

$$
\begin{aligned}
V^{t+1}-V^t \leq &-\left(\frac{2(\widehat{t}+1)^2}{\widehat{t}}-(2\widehat{t}+1)\right)W^t+100k_1^2\frac{(\widehat{t}+1)^2}{\mu^2(\widehat{t})^2}L(1+4\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2\\
&+20\kappa k_1\frac{(\widehat{t}+1)^2}{\mu\widehat{t}}G^2+\frac{0.06(\widehat{t}+1)^2}{L}d\sigma_{dp,t}^2.
\end{aligned}
$$

We know that

$$
\frac{2(\widehat{t}+1)^2}{\widehat{t}}-(2\widehat{t}+1)\geq\frac{2(\widehat{t}+1)^2}{\widehat{t}+1}-(2\widehat{t}+1)\geq 1>0.
$$

It can be readily checked that,

$$
\left(\frac{\widehat{t}+1}{\widehat{t}}\right)^2=\left(1+\frac{1}{\widehat{t}}\right)^2\leq\left(1+\frac{1}{a}\right)^2\leq 2.
$$

We also know that

$$
d\sigma_{dp,r}^2\leq\frac{24(1-\beta_r)^2T\log(1/\delta)C^2d^2}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)=T\eta_r^2L^2\frac{24k_2^2\log(1/\delta)C^2d^2}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right).
$$

For ease of presentation, let

$$
C_{dp}\triangleq\frac{24k_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right).
$$

It follows that

$$
V^{t+1}-V^t\leq 200k_1^2L(1+4\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^2}+20\kappa k_1\widehat{t}\frac{G^2}{\mu}+\frac{0.12TLC_{dp}}{\mu^2}.
$$

By telescoping sum, we have

$$
\begin{aligned}
V^{t+1}-V^0&=\sum_{k=0}^{t}\left(V^{k+1}-V^k\right)\\
&\leq 200(t+1)k_1^2L(1+4\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^2}+\frac{0.12TL(t+1)C_{dp}}{\mu^2}+\frac{20}{\mu}\left(\sum_{r=0}^{t}\left(r+a\frac{L}{\mu}\right)\right)\kappa k_1G^2.
\end{aligned}
$$

We know that

$$\sum_{r=0}^{t}\left(r+a\frac{L}{\mu}\right)=\frac{t(t+1)}{2}+a(t+1)\frac{L}{\mu}.$$

Therefore,

$$\begin{aligned}V^{t+1}-V^0&=\sum_{k=0}^{t}\left(V^{k+1}-V^k\right)\\&\leq 200(t+1)\left(k_1^2L(1+2\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^2}+\frac{TLC_{dp}}{\mu^2}+\frac{1}{\mu}\left(\frac{t}{2}+a\frac{L}{\mu}\right)\kappa k_1G^2\right).\end{aligned}$$

By definition, it holds that

$$V^{t+1}\geq(t+1+a\frac{L}{\mu})^2\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1})-F_{\mathcal{G}}^*\right].$$

Rearrange the terms, we have

$$\begin{aligned}\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1})-F_{\mathcal{G}}^*\right]&\leq\frac{V_0}{(t+1+a\frac{L}{\mu})^2}\\&\quad+\frac{200(t+1)}{(t+1+a\frac{L}{\mu})^2}\left(k_1^2L(1+2\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^2}+\frac{TLC_{dp}}{\mu^2}+\frac{1}{\mu}\left(\frac{t}{2}+a\frac{L}{\mu}\right)\kappa k_1G^2\right)\\&\leq\frac{V_0}{(t+1+a\frac{L}{\mu})^2}+\frac{200}{t+1+a\frac{L}{\mu}}\left(k_1^2L(1+2\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^2}+\frac{TLC_{dp}}{\mu^2}\right)+\frac{100}{\mu}\kappa k_1G^2.\end{aligned}$$

It remains to bound $V_0$.

$$V_0=\left(a\frac{L}{\mu}\right)^2\left(F_{\mathcal{G}}(\boldsymbol{\theta}^0)-F_{\mathcal{G}}^*+\frac{z_1}{L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)-\overline{\boldsymbol{M}}^0\right\|_2^2+\frac{z_2}{L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)-\overline{\boldsymbol{m}}^0\right\|_2^2+\frac{\kappa z_3}{L}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right).$$

By definition, we have

$$\overline{\boldsymbol{m}}^0=\beta_0R^{-1}+(1-\beta_0)\overline{\boldsymbol{m}}^0=\beta_0R^{-1}+(1-\beta_0)\alpha_0\overline{\boldsymbol{m}}^{-1}+(1-\beta_0)(1-\alpha_0)\overline{\boldsymbol{g}}^0=(1-\beta_0)(1-\alpha_0)\overline{\boldsymbol{g}}^0,$$

where the last equality follows from the convention that $R^{-1}=\overline{\boldsymbol{m}}^{-1}=\mathbf{0}$. It follows that

$$\begin{aligned}\left\|\overline{\boldsymbol{M}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2&=\left\|(1-\beta_0)(1-\alpha_0)\left(\overline{\boldsymbol{g}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right)-(1-(1-\beta_0)(1-\alpha_0))\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\\&\leq(1-\beta_0)(1-\alpha_0)\left\|\overline{\boldsymbol{g}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2+(1-(1-\beta_0)(1-\alpha_0))\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2,\end{aligned}$$

where the last inequality holds because of Jensen's inequality. Take expectation, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]&\leq(1-\beta_0)(1-\alpha_0)\mathbb{E}\left[\left\|\overline{\boldsymbol{g}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]+(1-(1-\beta_0)(1-\alpha_0))\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]\\&\leq\frac{\sigma^2}{|\mathcal{G}|}+\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]&=\mathbb{E}\left[\left\|\alpha_0\overline{\boldsymbol{m}}^{-1}+(1-\alpha_0)\left(\overline{\boldsymbol{g}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right)-\alpha_0\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]\\&\overset{(a)}{=}\mathbb{E}\left[\left\|(1-\alpha_0)\left(\overline{\boldsymbol{g}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right)-\alpha_0\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]\\&\overset{(b)}{=}\mathbb{E}\left[\left\|(1-\alpha_0)\left(\overline{\boldsymbol{g}}^0-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right)\right\|_2^2\right]+\mathbb{E}\left[\left\|\alpha_0\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]\\&\leq\frac{\sigma^2}{|\mathcal{G}|}+\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right],\end{aligned}$$

where equality $(a)$ follows from the convention that $\overline{\boldsymbol{m}}^{-1} = \mathbf{0}$, and equality $(b)$ follows from the independence in stochastic gradients. Next, we have

$$\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0) - \overline{\boldsymbol{M}}^0\right\|_2^2 = \left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0) - \overline{\boldsymbol{M}}^0\right\|_2^2 = \left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2 \leq 2L(F_{\mathcal{G}}(\boldsymbol{\theta}^0) - F_{\mathcal{G}}^*),$$

and

$$\begin{aligned}
&\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^0 - \overline{\boldsymbol{m}}^0\right\|_2^2\right] \\
&= \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|(1-\alpha_0)\left((\boldsymbol{g}_i^0 - \nabla F_i(\boldsymbol{\theta}^0)) - (\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right) + (1-\alpha_0)(\nabla F_i(\boldsymbol{\theta}^0) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0))\right\|_2^2\right] \\
&= \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|(1-\alpha_0)\left((\boldsymbol{g}_i^0 - \nabla F_i(\boldsymbol{\theta}^0)) - (\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right)\right\|_2^2 + \left\|(1-\alpha_0)(\nabla F_i(\boldsymbol{\theta}^0) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0))\right\|_2^2\right] \\
&\leq 2\left(1 + \frac{1}{|\mathcal{G}|}\right)\sigma^2 + B^2\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] + G^2.
\end{aligned}$$

Combining everything together, it holds that

$$\begin{aligned}
V_0 &= \left(a\frac{L}{\mu}\right)^2\left(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^* + \frac{z_1}{L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0) - \overline{\boldsymbol{M}}^0\right\|_2^2 + \frac{z_2}{L}\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0) - \overline{\boldsymbol{m}}^0\right\|_2^2 + \frac{\kappa z_3}{L}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right) \\
&\leq \left(a\frac{L}{\mu}\right)^2\left(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^* + \frac{z_1+z_2}{L}\left(\frac{\sigma^2}{|\mathcal{G}|} + \mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]\right)\right) \\
&\qquad + \left(a\frac{L}{\mu}\right)^2\frac{\kappa z_3}{L}\left(2\left(1 + \frac{1}{|\mathcal{G}|}\right)\sigma^2 + B^2\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] + G^2\right) \\
&\leq \left(a\frac{L}{\mu}\right)^2\left(6\left(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*\right) + \frac{2}{L}\frac{\sigma^2}{|\mathcal{G}|} + \frac{2\kappa}{L}\left(2\left(1 + \frac{1}{|\mathcal{G}|}\right)\sigma^2 + G^2\right)\right).
\end{aligned}$$

Putting everything together, we have

$$\begin{aligned}
\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1}) - F_{\mathcal{G}}^*\right] &\leq \frac{6a^2L^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\mu^2(t+1+a\frac{L}{\mu})^2} + \frac{2a^2L\sigma^2}{\mu^2(t+1+a\frac{L}{\mu})^2|\mathcal{G}|} + \frac{2a^2L\kappa}{\mu^2(t+1+a\frac{L}{\mu})^2}\left(2\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 + G^2\right) \\
&\quad + \frac{200}{t+1+a\frac{L}{\mu}}\left(k_1^2L(1+2\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^2} + \frac{TLC_{dp}}{\mu^2}\right) + \frac{100}{\mu}\kappa k_1G^2 \\
&\leq \frac{6a^2L^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\mu^2(t+1)^2} + \frac{2a^2L\sigma^2}{\mu^2(t+1)^2|\mathcal{G}|} + \frac{2a^2L\kappa}{\mu^2(t+1)^2}\left(2\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 + G^2\right) \\
&\quad + \frac{200}{t+1}\left(k_1^2L(1+2\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^2} + \frac{TLC_{dp}}{\mu^2}\right) + \frac{100}{\mu}\kappa k_1G^2.
\end{aligned}$$

For $\mu$-strongly convex objectives, we also know that

$$F_{\mathcal{G}}(\boldsymbol{\theta}^t) - F_{\mathcal{G}}^* \geq \frac{\mu}{2}\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\right\|_2^2.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\right\|_2^2\right] &\lesssim \frac{L^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\mu^3t^2} + \frac{L\sigma^2}{\mu^3t^2|\mathcal{G}|} + \frac{L\kappa}{\mu^3t^2}\left(\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 + G^2\right) \\
&\quad + \frac{1}{t}\left(L(1+\kappa)\left(1+\frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^3} + \frac{TLC_{dp}}{\mu^3}\right) + \frac{\kappa G^2}{\mu^2}.
\end{aligned}$$

In particular, when $t = T$, it holds that

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\theta}^T - \boldsymbol{\theta}^*\right\|_2^2\right] &\lesssim \frac{L^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\mu^3 T^2} + \frac{L\sigma^2}{\mu^3 T^2 |\mathcal{G}|} + \frac{L\kappa}{\mu^3 T^2}\left(\left(1 + \frac{1}{|\mathcal{G}|}\right)\sigma^2 + G^2\right) \\ &\quad + L(1+\kappa)\left(1 + \frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^3 T} + \frac{LC_{dp}}{\mu^3} + \frac{\kappa G^2}{\mu^2} \\ &\lesssim \frac{L^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\mu^3 T^2} + \frac{L\sigma^2}{\mu^3 T^2 |\mathcal{G}|} + \frac{L\kappa}{\mu^3 T^2}\left(\left(1 + \frac{1}{|\mathcal{G}|}\right)\sigma^2 + G^2\right) \\ &\quad + L(1+\kappa)\left(1 + \frac{1}{|\mathcal{G}|}\right)\frac{\sigma^2}{\mu^3 T} + \left(\frac{L}{\mu}\right)\frac{\log(1/\delta)C^2 d}{\mu^2 \epsilon^2 n^2}\left(1 + \frac{f}{n}\right) + \frac{\kappa G^2}{\mu^2}.\end{aligned}$$

□

### G.3 Non-convex objectives

In the analysis of non-convex objectives, we adopt a constant learning rate, i.e., $\eta_t = \eta$, $\beta_t = \beta$, and $\alpha_t = \alpha$.

#### G.3.1 Proof of Lemma 5

It holds, for $i \in \mathcal{G}$, that

$$\boldsymbol{M}_i^t - \overline{\boldsymbol{M}}^t = \beta\widetilde{R}^{t-1} + (1-\beta)\widehat{\boldsymbol{m}}_i^t - \left(\beta\widetilde{R}^{t-1} + (1-\beta)\overline{\boldsymbol{m}}^t\right) = (1-\beta)\left(\widehat{\boldsymbol{m}}_i^t - \overline{\boldsymbol{m}}^t\right).$$

It follows that

$$\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{M}_i^t - \overline{\boldsymbol{M}}^t\right\|_2^2 = \frac{(1-\beta)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\widehat{\boldsymbol{m}}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 = \frac{(1-\beta)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2, \tag{25}$$

where the last equality holds from the fact that $\widehat{\boldsymbol{m}}_i^t = \boldsymbol{m}_i^t$ for $i \in \mathcal{G}$. Furthermore, it holds for $i \in \mathcal{G}$ that

$$\boldsymbol{m}_i^t = \alpha\boldsymbol{m}_i^{t-1} + (1-\alpha)\,\boldsymbol{g}_i^t = (1-\alpha)\sum_{r=0}^{t}\alpha^{t-r}\boldsymbol{g}_i^r.$$

It follows that

$$\begin{aligned}&\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2 = \frac{(1-\alpha)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\sum_{r=0}^{t}\alpha^{t-r}\left(\boldsymbol{g}_i^r - \overline{\boldsymbol{g}}^t\right)\right\|_2^2 \\ &\leq \underbrace{\frac{3\,(1-\alpha)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\sum_{r=0}^{t}\alpha^{t-r}\left(\boldsymbol{g}_i^r - \nabla F_i(\boldsymbol{\theta}^r)\right)\right\|_2^2}_{(A)} \\ &\quad + \underbrace{\frac{3\,(1-\alpha)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\sum_{r=0}^{t}\alpha^{t-r}\left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r) - \overline{\boldsymbol{g}}^r\right)\right\|_2^2}_{(B)} + \underbrace{\frac{3\,(1-\alpha)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\sum_{r=0}^{t}\alpha^{t-r}\left(\nabla F_i(\boldsymbol{\theta}^r) - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\right)\right\|_2^2}_{(C)},\end{aligned}$$

where the inequality follows from Jensen's inequality.

**Bounding $(A)$.**

$$\frac{3\,(1-\alpha)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\sum_{r=0}^{t}\alpha^{t-r}\left(\boldsymbol{g}_i^r - \nabla F_i(\boldsymbol{\theta}^r)\right)\right\|_2^2\right] = \frac{3\,(1-\alpha)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\sum_{r=0}^{t}\alpha^{2(t-r)}\mathbb{E}\left[\left\|\boldsymbol{g}_i^r - \nabla F_i(\boldsymbol{\theta}^r)\right\|_2^2\right] \leq \frac{3\,(1-\alpha)^2}{1-\alpha^2}\sigma^2,$$

where the equality and the inequality both follow from Assumption 3.

**Bounding $(B)$.**

$$\frac{3\left(1-\alpha\right)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\sum_{r=0}^{t}\alpha^{t-r}\left(\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)-\overline{\boldsymbol{g}}^r\right)\right\|_2^2\right]=\frac{3\left(1-\alpha\right)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\sum_{r=0}^{t}\alpha^{2(t-r)}\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)-\overline{\boldsymbol{g}}^r\|_2^2\right]\leq\frac{3\left(1-\alpha\right)^2}{1-\alpha^2}\frac{\sigma^2}{|\mathcal{G}|},$$

where the equality and the inequality both follow from Assumption 3. Furthermore, it holds that

$$\frac{(1-\alpha)^2}{1-\alpha^2}=\frac{1-\alpha}{1+\alpha}=\frac{(1-\alpha)(1+\alpha)}{(1+\alpha)^2}\leq(1-\alpha)(1+\alpha)=1-\alpha^2.$$

**Bounding $(C)$.**

$$\begin{aligned}\frac{3\left(1-\alpha\right)^2}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\sum_{r=0}^{t}\alpha^{t-r}\left(\nabla F_i(\boldsymbol{\theta}^r)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\right)\right\|_2^2\right]&\leq\frac{3\left(1-\alpha\right)^2}{|\mathcal{G}|}\left(\sum_{r=0}^{t}\alpha^{t-r}\right)\sum_{i\in\mathcal{G}}\sum_{r=0}^{t}\alpha^{t-r}\mathbb{E}\left[\|\nabla F_i(\boldsymbol{\theta}^r)-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\|_2^2\right]\\&\leq 3\left(1-\alpha\right)\sum_{r=0}^{t}\alpha^{t-r}\left(B^2\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\|_2^2\right]+G^2\right)\leq 3G^2+3B^2(1-\alpha)\sum_{r=0}^{t}\alpha^{t-r}\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\|_2^2\right].\end{aligned}$$

Putting them together, it holds that

$$\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]\leq 3\left(1+\frac{1}{|\mathcal{G}|}\right)\left(1-\alpha^2\right)\sigma^2+3G^2+3B^2(1-\alpha)\sum_{r=0}^{t}\alpha^{t-r}\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\|_2^2\right].$$

### G.3.2 Proof of Theorem 4

**Theorem G.6.** *Suppose that Assumptions 2, 3, and 4 hold. Let $\alpha=1-5100\eta L$, $\beta=1-50\eta L$. $\eta<1/(5100L)$, and $B^2<1/(23625\kappa)$, it holds for Algorithm 1 that*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]\leq\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{\eta T}+\frac{25d\sigma_{dp}^2}{\eta L}+\frac{2\eta L}{T}k_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|}+4k_1^2\eta L\frac{\sigma^2}{|\mathcal{G}|}\\+47250k_1\eta L\kappa\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2+23625\kappa G^2.\end{aligned}$$

*Proof.* By Lemma G.10, we know that

$$\begin{aligned}\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1})-F_{\mathcal{G}}(\boldsymbol{\theta}^t)\mid\mathcal{F}^t\right]\leq-\eta\left(\frac{3}{4}-2\eta L\right)\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2+\frac{2\kappa\eta(1+\eta L)}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{M}_i^t-\overline{\boldsymbol{M}}^t\right\|_2^2\mid\mathcal{F}^t\right]\\+2\eta(1+\eta L)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{m}}^t\right\|_2^2\mid\mathcal{F}^t\right]+\frac{\eta^2Ld}{2}\sigma_{dp}^2.\end{aligned}$$

Define Lyapunov function $V^t \triangleq \mathbb{E}[F_{\mathcal{G}}(\boldsymbol{\theta}^t)] + \mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{M}}^t\|_2^2]/16L + \mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{m}}^t\|_2^2]/16L$, we have

$$
\begin{aligned}
V^{t+1} - V^t &= \mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1})\right] - \mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right] + \frac{1}{16L}\left(\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1}) - \overline{\boldsymbol{M}}^{t+1}\right\|_2^2\right] - \mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{M}}^t\right\|_2^2\right]\right) \\
&\quad + \frac{1}{16L}\left(\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^{t+1}) - \overline{\boldsymbol{m}}^{t+1}\|_2^2\right] - \mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{m}}^t\|_2^2\right]\right) \\
&\leq -\eta\left(\frac{3}{4} - 2\eta L\right)\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\|_2^2\right] + \frac{2\kappa\eta(1+\eta L)}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{M}_i^t - \overline{\boldsymbol{M}}^t\right\|_2^2\right] \\
&\quad + 2\eta(1+\eta L)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{M}}^t\right\|_2^2\right] + \frac{\eta^2 L d}{2}\sigma_{dp}^2 + \frac{\beta^2 + \eta L\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)}{16L} d\sigma_{dp}^2 \\
&\quad - \frac{1}{16L}\left(\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{M}}^t\right\|_2^2\right] + \mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{m}}^t\|_2^2\right]\right) + \frac{(1-\beta)^2(1-\alpha)^2}{16L}\frac{\sigma^2}{|\mathcal{G}|} \\
&\quad + \frac{\beta}{16L}\left(\beta + \frac{(1-\beta)\alpha}{k_2} + \eta L(\beta + (1-\beta)\alpha)\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \\
&\quad + \frac{3\eta}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \\
&\quad + \frac{(1-\beta)\alpha}{16L}\left[\left((1-\beta)\alpha + 2k_2\beta\right)(1+\eta L) + (1-\beta)\beta\right]\mathbb{E}\left[\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\|_2^2\right] \\
&\quad + \frac{3\eta}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\|_2^2\right] \\
&\quad + \frac{(1-\beta)\beta}{16L}\left(2k_2\beta + (1-\beta)\alpha + \eta L(\beta + (1-\beta)\alpha)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] \\
&\quad + \frac{3(1-\beta)^2\eta}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] \\
&\quad + \frac{\alpha^2(1+\eta L)}{16L}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] + \frac{\alpha^2\eta(1+\eta L)}{16}d\sigma_{dp}^2 + \frac{(1-\alpha)^2}{16L}\frac{\sigma^2}{|\mathcal{G}|} \\
&\quad + \frac{3\alpha^2\eta(1+\eta L)}{16}\mathbb{E}\left[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\|_2^2\right] + \frac{3\alpha^2\eta(1+\eta L)}{16}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] + \frac{3\alpha^2\eta(1+\eta L)}{16}\frac{(1-\beta)^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right].
\end{aligned}
$$

Rearrange the terms, we have

$$
\begin{aligned}
&V^{t+1}-V^t\\
&\leq -\eta\left(\frac{3}{4}-\frac{3}{16}\left(\beta+(1-\beta)\alpha\right)\left(\left(\beta+(1-\beta)\alpha\right)(1+\eta L)+\beta(1-\beta)\right)-\frac{3\alpha^2(1+\eta L)}{16}-2\eta L\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]\\
&\quad+\frac{(3\alpha^2+32)\eta(1+\eta L)}{16}\frac{(1-\beta)^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]\\
&\quad+\frac{(1-\beta)\beta}{16L}\left(2k_2\beta+(1-\beta)\alpha+\eta L(\beta+(1-\beta)\alpha)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]\\
&\quad+\frac{3(1-\beta)^2\eta}{16}\left(\beta+(1-\beta)\alpha\right)\left(\left(\beta+(1-\beta)\alpha\right)(1+\eta L)+\beta(1-\beta)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]\\
&\quad+\left(2\eta(1+\eta L)-\frac{1}{16L}+\frac{\beta}{16L}\left(\beta+(1-\beta)\frac{\alpha}{k_2}+\eta L(\beta+(1-\beta)\alpha)\right)\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2\right]\\
&\quad+\frac{3\eta}{16}\left(\left(\beta+(1-\beta)\alpha\right)\left(\left(\beta+(1-\beta)\alpha\right)(1+\eta L)+\beta(1-\beta)\right)+\alpha^2(1+\eta L)\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\right\|_2^2\right]\\
&\quad+\left(-\frac{1}{16L}+\frac{\alpha^2(1+\eta L)}{16L}+\frac{(1-\beta)\alpha}{16L}\left[\left((1-\beta)\alpha+2k_2\beta\right)(1+\eta L)+(1-\beta)\beta\right]\right)\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t-\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]\\
&\quad+\left(\frac{\eta^2 L}{2}+\frac{\beta^2+\eta L\left(\beta+(1-\beta)\alpha\right)\left(\left(\beta+(1-\beta)\alpha\right)(1+\eta L)+\beta(1-\beta)\right)}{16L}+\frac{\alpha^2\eta(1+\eta L)}{16}\right)d\sigma_{dp}^2\\
&\quad+\frac{(1-\alpha)^2(1+(1-\beta)^2)}{16L}\frac{\sigma^2}{|\mathcal{G}|}.
\end{aligned}
$$

Next, we bound the coefficients of the terms $\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\|_2^2]$ and $\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{m}}^t\|_2^2]$. Recall the definition of $\alpha=1-k_1\eta L=1-5100\eta L$ and $\beta=1-k_2\eta L=1-50\eta L$, it holds that $\alpha\leq\beta$.

**Bounding $C_1$ for $\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{M}}^t\|_2^2]$.**

$$
\begin{aligned}
C_1 &\triangleq 2\eta(1+\eta L)-\frac{1}{16L}+\frac{\beta}{16L}\left(\beta+(1-\beta)\frac{\alpha}{k_2}+\eta L(\beta+(1-\beta)\alpha)\right)\\
&\qquad+\frac{3\eta}{16}\left(\left(\beta+(1-\beta)\alpha\right)\left(\left(\beta+(1-\beta)\alpha\right)(1+\eta L)+\beta(1-\beta)\right)+\alpha^2(1+\eta L)\right)\\
&\leq 2\eta(1+\eta L)-\frac{k_2\eta}{16}+\frac{\eta}{16}\left(1+(1+k_2\eta L)\right)+\frac{3\eta}{16}\left(\left(1+k_2\eta L\right)\left(\left(1+k_2\eta L\right)(1+\eta L)+k_2\eta L\right)+(1+\eta L)\right)\\
&\leq -\frac{\eta}{16}\left(k_2-32(1+\eta L)-(2+k_2\eta L)-3\left(1+k_2\eta L\right)\left(\left(1+k_2\eta L\right)(1+\eta L)+k_2\eta L+(1+\eta L)\right)\right)\\
&\leq -\frac{\eta}{16}\left(50-33-2.5-12\right)=-\frac{5\eta}{32}<0,
\end{aligned}
$$

where the third to last inequality holds because $\eta\leq\frac{1}{132L}$ and $k_2=50$.

**Bounding $C_2$ for $\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)-\overline{\boldsymbol{m}}^t\|_2^2]$.**

$$
\begin{aligned}
C_2 &\triangleq -\frac{1}{16L}+\frac{\alpha^2(1+\eta L)}{16L}+\frac{(1-\beta)\alpha}{16L}\left[\left((1-\beta)\alpha+2k_2\beta\right)(1+\eta L)+(1-\beta)\beta\right]\\
&\leq -\frac{1}{16L}\left(1-\alpha-\alpha^2\eta L-(1-\beta)\alpha\left(\left((1-\beta)\alpha+2k_2\beta\right)(1+\eta L)+(1-\beta)\beta\right)\right)\\
&\leq -\frac{\eta}{16}\left(k_1-1-k_2^2\left((\eta L+2)(1+\eta L)+\eta L\right)\right)\\
&\leq -\frac{\eta}{16}\left(5100-\frac{3}{132}-4-33^2\left(\left(\frac{1}{132}+2\right)\times\left(\frac{1}{132}+1\right)+\frac{1}{132}\right)\right)<-\frac{20\eta}{16}<0,
\end{aligned}
$$

where the second to last inequality holds because $\eta \leq \frac{1}{132L}$ and $k_1 = 5100$. Therefore, we have

$$
\begin{aligned}
&V^{t+1} - V^t \\
&\leq -\eta\left(\frac{3}{4} - \frac{3}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right) - \frac{3\alpha^2(1+\eta L)}{16} - 2\eta L\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \\
&\quad + \frac{(3\alpha^2+32)\eta(1+\eta L)}{16}\frac{(1-\beta)^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] \\
&\quad + \frac{(1-\beta)\beta}{16L}\left(2k_2\beta + (1-\beta)(\beta+\alpha) + \eta L(\beta + (1-\beta)\alpha)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] \\
&\quad + \frac{3(1-\beta)^2\eta}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] \\
&\quad + C_1\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t) - \overline{\boldsymbol{M}}^t\right\|_2^2\right] + C_2\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^t - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] + \frac{(1-\alpha)^2(1+(1-\beta)^2)}{16L}\frac{\sigma^2}{|\mathcal{G}|} \\
&\quad + \left(\frac{\eta^2 L}{2} + \frac{\beta^2 + \eta L\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)}{16L} + \frac{\alpha^2\eta(1+\eta L)}{16}\right)d\sigma_{dp}^2 \\
&\leq -\eta\left(\frac{3}{4} - \frac{3}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right) - \frac{3\alpha^2(1+\eta L)}{16} - 2\eta L\right)\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \\
&\quad + \frac{(3\alpha^2+32)\eta(1+\eta L)}{16}\frac{(1-\beta)^2\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] + \frac{(1-\alpha)^2(1+(1-\beta)^2)}{16L}\frac{\sigma^2}{|\mathcal{G}|} \\
&\quad + \frac{(1-\beta)\beta}{16L}\left(2k_2\beta + (1-\beta)\alpha + \eta L(\beta + (1-\beta)\alpha)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] \\
&\quad + \frac{3(1-\beta)^2\eta}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)\frac{\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right] \\
&\quad + \left(\frac{\eta^2 L}{2} + \frac{\beta^2 + \eta L\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right)}{16L} + \frac{\alpha^2\eta(1+\eta L)}{16}\right)d\sigma_{dp}^2.
\end{aligned}
$$

Furthermore, it holds for the coefficients of $\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\|_2^2]$ that

$$
\begin{aligned}
&\frac{3}{4} - \frac{3}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right) - \frac{3\alpha^2(1+\eta L)}{16} - 2\eta L \\
&\geq \frac{3}{4} - \frac{3}{16}\left(1 + k_2\eta L\right)\left(\left(1 + k_2\eta L\right)(1+\eta L) + k_2\eta L\right) - \frac{3(1+\eta L)}{16} - 2\eta L \geq \frac{2}{25},
\end{aligned}
$$

where the last inequality follows from $\eta \leq 1/132L$ and $k_2 = 50$.

**Bounding $C_3$ for $\mathbb{E}\left[\left\|\boldsymbol{m}_i^t - \overline{\boldsymbol{m}}^t\right\|_2^2\right]$.**

$$
\begin{aligned}
C_3 &\triangleq \frac{(3\alpha^2+32)\eta(1+\eta L)(1-\beta)^2}{16} + \frac{(1-\beta)\beta}{16L}\left(2k_2\beta + (1-\beta)(\beta+\alpha) + \eta L(\beta + (1-\beta)\alpha)\right) \\
&\quad + \frac{3(1-\beta)^2\eta}{16}\left(\beta + (1-\beta)\alpha\right)\left(\left(\beta + (1-\beta)\alpha\right)(1+\eta L) + \beta(1-\beta)\right) \\
&\leq \frac{35k_2^2\eta^3L^2(1+\eta L)}{16} + \frac{k_2\eta}{16}\left(2k_2 + 2k_2\eta L + \eta L(1 + k_2\eta L)\right) + \frac{3k_2^2\eta^3L^2}{16}\left(1 + k_2\eta L\right)\left(\left(1 + k_2\eta L\right)(1+\eta L) + k_2\eta L\right) \\
&\leq \frac{35\eta\left(\frac{50}{132}\right)^2(1+\frac{1}{132})}{16} + \frac{50\eta}{16}\left(2\times 50 + \frac{2\times 50}{132} + \frac{1}{132}\times(1 + \frac{50}{132})\right) \\
&\quad + \frac{3\left(\frac{50}{132}\right)^2\eta}{16}(1 + \frac{50}{132})\times\left((1+\frac{50}{132})\times(1+\frac{1}{132}) + \frac{50}{132}\right) \\
&\leq \eta\left(0.316 + 314 + 0.06558\right) \leq 315\eta.
\end{aligned}
$$

Moreover, we have

$$\begin{aligned}
&\frac{\eta^2 L}{2}+\frac{\beta^2+\eta L\left(\beta+(1-\beta)\alpha\right)\left(\left(\beta+(1-\beta)\alpha\right)\left(1+\eta L\right)+\beta(1-\beta)\right)}{16L}+\frac{\alpha^2\eta(1+\eta L)}{16}\\
&\leq\frac{1}{L}\left(\frac{\left(\frac{1}{132}\right)^2}{2}+\frac{1+\frac{1}{132}\left(1+\frac{50}{132}\right)\times\left(\left(1+\frac{50}{132}\right)\times\left(1+\frac{1}{132}\right)+\frac{50}{132}\right)}{16}+\frac{\frac{1}{132}\times\left(1+\frac{1}{132}\right)}{16}\right)\\
&\leq\frac{3\times10^{-5}+0.065+5\times10^{-4}}{L}\leq\frac{1}{L}.
\end{aligned}$$

Hence, the inequality simplifies to

$$\begin{aligned}
&V^{t+1}-V^t\leq-\frac{2\eta}{25}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{315\eta\kappa}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\boldsymbol{m}_i^t-\overline{\boldsymbol{m}}^t\right\|_2^2\right]+\frac{d\sigma_{dp}^2}{L}+\frac{k_1^2\eta^2 L}{8}\frac{\sigma^2}{|\mathcal{G}|}\\
&\leq-\frac{2\eta}{25}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{d\sigma_{dp}^2}{L}+\frac{k_1^2\eta^2 L}{8}\frac{\sigma^2}{|\mathcal{G}|}\\
&\quad+945\kappa\eta\left(1+\frac{1}{|\mathcal{G}|}\right)\left(1-\alpha^2\right)\sigma^2+945\kappa\eta G^2+945\kappa\eta B^2(1-\beta)^2(1-\alpha)\sum_{r=0}^{t}\alpha^{t-r}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\right\|_2^2\right].
\end{aligned}$$

By telescoping sum, it holds that

$$\begin{aligned}
&\frac{V^T-V^0}{T}\leq-\frac{2\eta}{25}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{d\sigma_{dp}^2}{L}+\frac{k_1^2\eta^2 L}{8}\frac{\sigma^2}{|\mathcal{G}|}\\
&\quad+945\kappa\eta\left(1+\frac{1}{|\mathcal{G}|}\right)\left(1-\alpha^2\right)\sigma^2+945\kappa\eta G^2+945\kappa\eta B^2(1-\alpha)\frac{1}{T}\sum_{t=0}^{T-1}\sum_{r=0}^{t}\alpha^{t-r}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^r)\right\|_2^2\right]\\
&\leq-\frac{2\eta}{25}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{d\sigma_{dp}^2}{L}+\frac{k_1^2\eta^2 L}{8}\frac{\sigma^2}{|\mathcal{G}|}\\
&\quad+945\kappa\eta\left(1+\frac{1}{|\mathcal{G}|}\right)\left(1-\alpha^2\right)\sigma^2+945\kappa\eta G^2+945\kappa\eta B^2(1-\alpha)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]\left(\sum_{r=0}^{T-1}\alpha^r\right)\\
&\leq-\eta\left(\frac{2}{25}-945\kappa B^2\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{d\sigma_{dp}^2}{L}+\frac{k_1^2\eta^2 L}{8}\frac{\sigma^2}{|\mathcal{G}|}+945\kappa\eta\left(1+\frac{1}{|\mathcal{G}|}\right)\left(1-\alpha^2\right)\sigma^2+1638\kappa\eta G^2.
\end{aligned}$$

For some $B^2<1/(23625\kappa)$, it holds that

$$\frac{V^T-V^0}{T}\leq-\frac{\eta}{25}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]+\frac{d\sigma_{dp}^2}{L}+\frac{k_1^2\eta^2 L}{8}\frac{\sigma^2}{|\mathcal{G}|}+945\kappa\eta\left(1+\frac{1}{|\mathcal{G}|}\right)\left(1-\alpha^2\right)\sigma^2+945\kappa\eta G^2.$$

Rearrange the terms, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right]\leq\frac{25(V^0-V^T)}{\eta T}+\frac{25d\sigma_{dp}^2}{\eta L}+4k_1^2\eta L\frac{\sigma^2}{|\mathcal{G}|}+47250k_1\eta L\kappa\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2+23625\kappa G^2,$$

where the inequality holds because

$$(1-\alpha^2)=(1-\alpha)(1+\alpha)\leq2(1-\alpha).$$

For the term $-V^T$, it holds that

$$-V^T\leq-F_{\mathcal{G}}^*.$$

For the term $V^0$, it holds that

$$V^0=\mathbb{E}[F_{\mathcal{G}}(\boldsymbol{\theta}^0)]+\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)-\overline{\boldsymbol{M}}^0\|_2^2]/16L+\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)-\overline{\boldsymbol{m}}^0\|_2^2]/16L.$$

By definition, we have

$$\overline{\boldsymbol{M}}^0 = \beta\widetilde{R}^{-1} + (1-\beta)\overline{\boldsymbol{m}}^0 = \beta\widetilde{R}^{-1} + (1-\beta)\alpha\overline{\boldsymbol{m}}^{-1} + (1-\beta)(1-\alpha)\overline{\boldsymbol{g}}^0 = (1-\beta)(1-\alpha)\overline{\boldsymbol{g}}^0,$$

where the last equality follows from the convention that $\widetilde{R}^{-1} = \overline{\boldsymbol{m}}^{-1} = \mathbf{0}$. It follows that

$$\begin{aligned}\left\|\overline{\boldsymbol{M}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2 &= \left\|(1-\beta)(1-\alpha)\left(\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right) - (1-(1-\beta)(1-\alpha))\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2 \\ &\leq (1-\beta)(1-\alpha)\left\|\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2 + (1-(1-\beta)(1-\alpha))\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2,\end{aligned}$$

where the last inequality holds because of Jensen's inequality. Take expectation, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\boldsymbol{M}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] &\leq (1-\beta)(1-\alpha)\mathbb{E}\left[\left\|\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] + (1-(1-\beta)(1-\alpha))\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] \\ &\leq k_1k_2\eta^2L^2\frac{\sigma^2}{|\mathcal{G}|} + \mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right]\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\boldsymbol{m}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] &= \mathbb{E}\left[\left\|\alpha\overline{\boldsymbol{m}}^{-1} + (1-\alpha)\left(\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right) - \alpha\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\left\|(1-\alpha)\left(\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right) - \alpha\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] \\ &\overset{(b)}{=} \mathbb{E}\left[\left\|(1-\alpha)\left(\overline{\boldsymbol{g}}^0 - \nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right)\right\|_2^2\right] + \mathbb{E}\left[\left\|\alpha\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right] \\ &\leq (1-\alpha)^2\frac{\sigma^2}{|\mathcal{G}|} + \mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right\|_2^2\right],\end{aligned}$$

where equality $(a)$ follows from the convention that $\overline{\boldsymbol{m}}^{-1} = \mathbf{0}$, and equality $(b)$ follows from the independence in stochastic gradients. Putting them together, we have

$$V^0 \leq \mathbb{E}[F_{\mathcal{G}}(\boldsymbol{\theta}^0)] + \eta^2L\frac{k_1(k_1+k_2)}{16}\frac{\sigma^2}{|\mathcal{G}|} + \frac{1}{8L}\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\|_2^2].$$

We know, for a smooth objective, that

$$\mathbb{E}[\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^0)\|_2^2] \leq 2L(F_{\mathcal{G}}(\boldsymbol{\theta}^0) - F_{\mathcal{G}}^*).$$

It follows that

$$V^0 \leq \mathbb{E}[F_{\mathcal{G}}(\boldsymbol{\theta}^0)] + \eta^2L\frac{k_1(k_1+k_2)}{16}\frac{\sigma^2}{|\mathcal{G}|} + \frac{1}{4}\mathbb{E}[F_{\mathcal{G}}(\boldsymbol{\theta}^0) - F_{\mathcal{G}}^*].$$

Combining everything together, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \leq \frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\eta T} + \frac{25d\sigma_{dp}^2}{\eta L} + \frac{2\eta L}{T}k_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|} + 4k_1^2\eta L\frac{\sigma^2}{|\mathcal{G}|} \\ + 47250k_1\eta L\kappa\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 + 23625\kappa G^2.\end{aligned}$$

□

### G.3.3 Proof of Corollary 1

**Corollary 3.** *Suppose that Assumptions 2, 3, 1, and 4 hold. When $m \geq (2+\nu)f$, where $\nu > 0$ is an absolute constant, we have for Algorithm 1 with coordinate-wise trimmed mean as the robust aggregator that*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \lesssim \frac{\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*}{\eta T} + \eta LT\left(1+\frac{f}{n}\right)\frac{\log(1/\delta)C^2d}{\epsilon^2n^2} + \frac{\eta L}{T}\frac{\sigma^2}{|\mathcal{G}|} + \eta L\frac{\sigma^2}{|\mathcal{G}|} \\ + \eta L\kappa\left(1+\frac{1}{|\mathcal{G}|}\right)\sigma^2 + \kappa G^2. \qquad (26)\end{aligned}$$

*When*

$$\eta = \min\left\{\frac{1}{132L}, \frac{1}{\sqrt{T}}\sqrt{\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T + 2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|} + 4k_1^2L\frac{\sigma^2}{|\mathcal{G}|} + 94500k_1L\kappa\sigma^2}}\right\},$$

*it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \lesssim \frac{\sigma}{\sqrt{T}}\sqrt{L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)\left(\frac{1}{|\mathcal{G}|} + \kappa\left(1+\frac{1}{|\mathcal{G}|}\right)\right)} + \frac{L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{T}$$
$$+ \sqrt{L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)\frac{\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{f}{n}\right)} + \kappa G^2.$$

We put $\sigma_{dp}^2$ for coordinate-wise trimmed mean into the formula

$$\sigma_{dp}^2 = \frac{8T\log(1/\delta)\Delta_{Byz}^2}{\epsilon^2} = \frac{32(1-\beta)^2T\log(1/\delta)C^2}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)$$
$$\leq \frac{32k_2^2\eta^2L^2T\log(1/\delta)C^2}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right).$$

Plugging this in, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \leq \frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\eta T} + \eta LT\frac{800k_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right) + 2\eta Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|}$$
$$+ 4k_1^2\eta L\frac{\sigma^2}{|\mathcal{G}|} + 94500k_1\eta L\kappa\sigma^2 + 23625\kappa G^2.$$

When we have the optimal learning rate, i.e.,

$$\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\eta T}$$
$$= \eta LT\frac{800k_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right) + 2\eta Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|} + 4k_1^2\eta L\frac{\sigma^2}{|\mathcal{G}|} + 94500k_1\eta L\kappa\sigma^2,$$

we have

$$\eta = \frac{1}{\sqrt{T}}\sqrt{\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T + 2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|} + 4k_1^2L\frac{\sigma^2}{|\mathcal{G}|} + 94500k_1L\kappa\sigma^2}}.$$

Together with the previous condition, we use

$$\eta = \min\left\{\frac{1}{132L}, \frac{1}{\sqrt{T}}\sqrt{\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T + 2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|} + 4k_1^2L\frac{\sigma^2}{|\mathcal{G}|} + 94500k_1L\kappa\sigma^2}}\right\}.$$

It follows that

$$\frac{1}{\eta} = \max\left\{132L, \sqrt{T}\sqrt{\frac{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T + 2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|} + 4k_1^2L\frac{\sigma^2}{|\mathcal{G}|} + 94500k_1L\kappa\sigma^2}{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}}\right\}$$
$$\leq 132L + \sqrt{T}\sqrt{\frac{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T + 2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|} + 4k_1^2L\frac{\sigma^2}{|\mathcal{G}|} + 94500k_1L\kappa\sigma^2}{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right] - F_{\mathcal{G}}^*)}}.$$

Therefore, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \leq \frac{3960L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}$$
$$+2\sqrt{\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}}\sqrt{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T+2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|}+4k_1^2L\frac{\sigma^2}{|\mathcal{G}|}+94500k_1L\kappa\sigma^2}$$
$$+23625\kappa G^2.$$

By $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_{\mathcal{G}}(\boldsymbol{\theta}^t)\right\|_2^2\right] \leq \frac{3960L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}$$
$$+2\sqrt{\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}}\sqrt{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2m^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T+2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|}+4k_1^2L\frac{\sigma^2}{|\mathcal{G}|}+94500k_1L\kappa\sigma^2}$$
$$+23625\kappa G^2$$
$$\leq \frac{3960L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}$$
$$+2\sqrt{\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}}\sqrt{\frac{800Lk_2^2\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)T}$$
$$+2\sqrt{\frac{30(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}}\sqrt{2Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|}+4k_1^2L\frac{\sigma^2}{|\mathcal{G}|}+94500k_1L\kappa\sigma^2}+23625\kappa G^2$$
$$=\frac{2\sigma}{\sqrt{T}}\sqrt{60(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)\left(Lk_1(k_1+k_2)\frac{\sigma^2}{|\mathcal{G}|}+2k_1^2L\frac{\sigma^2}{|\mathcal{G}|}+47250k_1L\kappa\right)}+\frac{3960L(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)}{T}$$
$$+2\sqrt{24000Lk_2^2(\mathbb{E}\left[F_{\mathcal{G}}(\boldsymbol{\theta}^0)\right]-F_{\mathcal{G}}^*)\frac{\log(1/\delta)C^2d}{\epsilon^2n^2}\left(1+\frac{4(2+\nu)^2}{\nu^2}\frac{f}{n}\right)}+23625\kappa G^2.$$

### G.4 Supporting Results for the Optimality of CWTM

In this section, we present the remaining supporting results that establish the optimality of our results.

### G.5 Conditions of matching optimality

**Lemma 1.** From its proof, we have

$$\nabla F_i(\theta)=\mu\theta-\frac{\mu G_0}{L\sqrt{d}}\sum_{j=1}^m\frac{x_i^j}{m},$$

where $\sum_{j=1}^m x_i^j/m\in\{\pm1\}^d$ for $j\in[m]$ and $i\in[n]$. It follows that

$$C=\|\nabla F_i(\mathbf{0})\|_2=\frac{\mu G_0}{L\sqrt{d}}\|\sum_{j=1}^m\frac{x_i^j}{m}\|_2=\frac{\mu G_0}{L\sqrt{d}}\sqrt{d}=\frac{\mu G_0}{L}=\Theta(G),$$

where the last equality holds because $\mu=\Theta(L)$ and $G_0=G/2$.

**Lemmas 2 and 3.** Let $\tau=f/(n-f)$ and $\mathbb{1}\in\mathbb{R}^d$ be an all-one vector. From the proofs, we have

$$\nabla F_i(\theta)=\begin{cases}\mu(\theta-\frac{G_0\mathbb{1}}{L\sqrt{\tau d}}), & \text{if } i\in[1,f];\\ \mu(\theta+\frac{G_0\mathbb{1}}{L\sqrt{\tau d}}), & \text{if } i\in[f+1,2f];\\ \mu\theta, & \text{otherwise.}\end{cases}$$

It follows that

$$\nabla F_i(\mathbf{0})=\begin{cases}-\frac{\mu G_0\mathbb{1}}{L\sqrt{\tau d}}, & \text{if } i\in[1,f];\\ +\frac{\mu G_0\mathbb{1}}{L\sqrt{\tau d}}, & \text{if } i\in[f+1,2f];\\ \mathbf{0}, & \text{otherwise.}\end{cases}$$

We have

$$C = \max_{i\in[n]}\{\|\nabla F_i(\mathbf{0})\|_2\} = \|\frac{\mu G_0 \mathbb{1}}{L\sqrt{\tau d}}\|_2 = \frac{\mu G_0}{L\sqrt{\tau}} = \Theta(\frac{G}{\sqrt{\tau}}),$$

where the last equality holds because $\mu = \Theta(L)$ and $G_0 = G/\sqrt{2}$. In the regime of $f = \Theta(n)$, it holds that $C = \Theta(G)$.

### G.6 Optimality of CWTM

The proofs and intermediate results below show that coordinate-wise trimmed mean (CWTM) is order-optimal in terms of Byzantine-robustness for completeness. The proofs follow the standard roadmap from [4].

**Lemma G.14.** *Assume that $R : \mathbb{R}^{d\times n} \to \mathbb{R}^d$ is a coordinate-wise aggregation function, i.e., there exist $d$ real-valued functions $R_1, \ldots, R_d$ such that $\forall \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, $[R(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)]_r = R_r([\boldsymbol{x}_1]_r, \ldots, [\boldsymbol{x}_n]_r)$. If, for each $r \in [d]$, $R_r$ is $(f, \kappa)$-robust then $R$ is $(f, \kappa)$-robust.*

**Proof of Lemma G.14.** Consider $n$ arbitrary vectors in $\mathbb{R}^d$ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and an arbitrary set $S \subseteq [n]$ such that $|S| = n - f$. Assume that, for each $r \in [d]$, $R_r$ is $(f, \kappa)$-robust. Since $R$ is assumed to be a coordinate-wise aggregation function, we have

$$\begin{aligned}\|R(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \bar{\boldsymbol{x}}_S\|_2^2 &= \sum_{r=1}^{d} |R_r([\boldsymbol{x}_1]_r, \ldots, [\boldsymbol{x}_n]_r) - [\bar{\boldsymbol{x}}_S]_r|^2 \\ &\overset{(a)}{\leq} \frac{\kappa}{|S|}\sum_{r=1}^{d}\sum_{i\in S} |[\boldsymbol{x}_i]_r - [\bar{\boldsymbol{x}}_S]_r|^2 = \frac{\kappa}{|S|}\sum_{i\in S}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_S\|_2^2,\end{aligned}$$

where inequality $(a)$ follows from Definition 2. □

**Lemma G.15.** *Consider $n$ real values $x_1, \ldots, x_n$ such that $x_1 \leq \ldots \leq x_n$. Let $S \subseteq [n]$ of size $|S| = n - f$, and $I \triangleq \{f + 1, \ldots, n - f\}$. It holds that*

$$\sum_{i\in I\setminus S} |x_i - \bar{x}_S|^2 \leq \sum_{i\in S\setminus I} |x_i - \bar{x}_S|^2.$$

**Proof of Lemma G.15.** Observer that $|I| = n - 2f$, $|S| = n - f$, $|I \setminus S| \leq f$, and $|S \setminus I| \geq f$. Therefore, $|I \setminus S| \leq |S \setminus I|$. It suffices to show that there exists an injection $\phi : I \setminus S \to S \setminus I$ such that

$$|x_i - \bar{x}_S| \leq \left|x_{\phi(i)} - \bar{x}_S\right|, \quad \forall i \in I \setminus S.$$

Since $|I \setminus S| \leq |S \setminus I|$, we have $\sum_{i\in I\setminus S} |x_{\phi(i)} - \bar{x}_S|^2 \leq \sum_{i\in S\setminus I} |x_{\phi(i)} - \bar{x}_S|^2$.

For east of notation, we denote $S^{\complement} \triangleq [n] \setminus S$. It follows that $|S^{\complement}| = f$ and $I \setminus S = I \cap S^{\complement}$. Let $I^+ \triangleq \{n - f + 1, \ldots, n\}$ and $I^- \triangleq \{1, \ldots, f\}$ define the indices of values that are greater (or equal to) and smaller (or equal to) the values in $I$, respectively. Further let $|I \cap S^{\complement}| = q$. Since $S^{\complement} = [n] \setminus S$, it holds that

$$|I^+ \cap S| = f - |I^+ \cap S^{\complement}| \geq f - |S^{\complement} \setminus (I \cap S^{\complement})| = f - (f - q) = q. \tag{27}$$

Similarly, we have

$$|I^- \cap S| \geq q. \tag{28}$$

Recall that $q = |I \cap S^{\complement}|$ where $I \setminus S = I \cap S^{\complement}$. Therefore, (27) and (28) show that there exists an injection from $I \setminus S$ to $(I^- \cap S) \times (I^+ \cap S)$. Let $\psi$ denote this injection. For each $i \in I \setminus S$, $\psi(i)$ is a pair, denoted by $(\psi^-(i), \psi^+(i))$, in $(I^- \cap S) \times (I^+ \cap S)$. Consider an arbitrary $i \in I \setminus S$. By the definitions of $I^-$ and $I^+$, we have

$$x_{\psi^-(i)} \leq x_i \leq x_{\psi^+(i)}.$$

It follows, for any real value $y$, that

$$|x_i - y| \leq \max\left\{\left|x_{\psi^+(i)} - y\right|, \left|x_{\psi^-(i)} - y\right|\right\}.$$

The injection is thus defined as $\phi(i) = \arg\max_{j\in\{\psi^+(i),\psi^-(i)\}} |x_j - \bar{x}_S|$ for all $i \in I \setminus S$. □

**Proposition 5.** *Suppose that $n \in \mathbb{N}^+$ and $f < n/2$. $\texttt{CWTM}$ is $(f, \kappa)$-robust with*

$$\kappa = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right).$$

**Proof of Proposition 5.** Due to Lemma G.14, it suffices to show $\texttt{CWTM}$ is $(f, \kappa)$-robust in the scalar domain.

Without loss of generality, let $x_1, \dots, x_n \in \mathbb{R}$ such that $x_1 \leq \dots x_n$. We denote by $I \triangleq \{f+1, \dots, n-f\}$, and let $S$ be an arbitrary subset of $[n]$ of size $n-f$. Recall Definition 3, it holds that $\texttt{TM} = \frac{1}{n-2f}\sum_{i\in I} x_i$. It follows that

$$\begin{aligned}
\left|\texttt{CWTM}(x_1, \dots, x_n) - \bar{x}_S\right|^2 &= \left|\frac{1}{n-2f}\sum_{i\in I} x_i - \bar{x}_S\right|^2 = \left|\frac{1}{n-2f}\sum_{i\in I}(x_i - \bar{x}_S)\right|^2 \\
&\overset{(a)}{=} \left|\frac{1}{n-2f}\sum_{i\in I}(x_i - \bar{x}_S) - \frac{1}{n-2f}\sum_{i\in S}(x_i - \bar{x}_S)\right|^2 \\
&= \left|\frac{1}{n-2f}\sum_{i\in I\setminus S}(x_i - \bar{x}_S) - \frac{1}{n-2f}\sum_{i\in S\setminus I}(x_i - \bar{x}_S)\right|^2,
\end{aligned}$$

where equality $(a)$ follows from the fact that $\sum_{i\in S}(x_i - \bar{x}_S) = 0$. It follows that

$$\left|\texttt{CWTM}(x_1, \dots, x_n) - \bar{x}_S\right|^2 \leq \frac{|I\setminus S| + |S\setminus I|}{(n-2f)^2}\left(\sum_{i\in I\setminus S}\left|x_i - \bar{x}_S\right|^2 + \sum_{i\in S\setminus I}\left|x_i - \bar{x}_S\right|^2\right).$$

Note that $|I\setminus S| = |I\cup S| - |S| \leq n - (n-f) = f$. Similarly, we have $|S\setminus I| = |I\cup S| - |I| \leq n - (n-2f) = 2f$, Therefore, it follows that $|I\setminus S| + |S\setminus I| \leq 3f$, which yields

$$\begin{aligned}
\left|\texttt{CWTM}(x_1, \dots, x_n) - \bar{x}_S\right|^2 &\leq \frac{3f}{(n-2f)^2}\left(\sum_{i\in I\setminus S}\left|x_i - \bar{x}_S\right|^2 + \sum_{i\in S\setminus I}\left|x_i - \bar{x}_S\right|^2\right) \\
&\overset{(b)}{\leq} \frac{6f}{(n-2f)^2}\sum_{i\in S\setminus I}\left|x_i - \bar{x}_S\right|^2 \leq \frac{6f}{(n-2f)^2}\sum_{i\in S}\left|x_i - \bar{x}_S\right|^2,
\end{aligned}$$

where inequality $(b)$ follows from Lemma G.15 Since

$$\begin{aligned}
\frac{f}{(n-2f)^2} = \frac{f(n-f)}{(n-2f)^2}\frac{1}{n-f} &= \left(\frac{f}{n-2f} + \left(\frac{f}{n-2f}\right)^2\right)\frac{1}{n-f} \\
&\leq \left(\frac{f}{n-2f} + \left(\frac{f}{n-2f}\right)^2\right) = \frac{f}{n-2f}\left(1 + \frac{f}{n-2f}\right).
\end{aligned}$$

□

**Claim 2.** When $n \geq (2+\nu)f$, where $\nu > 0$, it holds that

$$\frac{f}{n} \leq \frac{f}{n-2f}\left(1 + \frac{f}{n-2f}\right) \leq \frac{(\nu+2)(\nu+1)}{\nu^2}\frac{f}{n}.$$

Equivalently, we have

$$\frac{f}{n-2f}\left(1 + \frac{f}{n-2f}\right) = \Theta(\frac{f}{n}).$$

**Proof of Claim 2.** $n \geq (2+\nu)f$, where $\nu > 0$, implies that $n - 2f \geq \nu f$, which leads to

$$\frac{f}{n-2f} \leq \frac{1}{\nu}.$$

It also implies that

$$n - 2f \geq n - \frac{2n}{\nu+2} = \frac{\nu n}{\nu+2}.$$

**Upper bound.**

$$\frac{f}{n-2f}\left(1+\frac{f}{n-2f}\right) \leq \frac{f}{n-2f}\frac{\nu+1}{\nu} \leq \frac{\nu+2}{\nu}\frac{\nu+1}{\nu}\frac{f}{n} = \frac{(\nu+2)(\nu+1)}{\nu^2}\cdot\frac{f}{n}$$

**Lower bound.**

$$\frac{f}{n-2f}\left(1+\frac{f}{n-2f}\right) \geq \frac{f}{n-2f} \geq \frac{f}{n}.$$

**Putting them together.**

$$\frac{f}{n} \leq \frac{f}{n-2f}\left(1+\frac{f}{n-2f}\right) \leq \frac{(\nu+2)(\nu+1)}{\nu^2}\frac{f}{n}.$$

□

**Corollary 4.** *Suppose that $n \in \mathbb{N}^+$ and $n \geq (2+\nu)f$, where $\nu > 0$,* `CWTM` *is $(f,\kappa)$-robust with*

$$\kappa = \Theta\left(\frac{f}{n}\right).$$

Corollary 4 can be shown by combining Proposition 5 and Claim 2.

Table 5: Neural network architecture, loss function, training steps, and batch size specifications

| **Dataset** | **MNIST / Fashion-MNIST** |
|---|---|
| Neural network | CNN |
| Model architecture* | **C**(1,16,8,2,3) – **T** – **M**(2,1) – **C**(16,32,4,2) – **T** – **M**(2,1) – **L**(512,32) – **T** – **L**(10) |
| Loss function | Cross-entropy loss |
| Number of global rounds $T$ | 800 |
| Batch size | 20 |

* **C** (# in-channel, # out-channel, kernel, stride, padding): a 2D convolution layer ; **T**: tanh activation function; **M** (kernel, stride): a 2D max-pool layer; **L** (# outputs) or **L** (# inputs, # outputs): a fully-connected linear layer; **D**: a dropout layer (probability 0.2).

Table 6: Test accuracy comparisons of different robust aggregators with and without server-side momentum on the MNIST and Fashion-MNIST datasets. Experiments run for a total of 800 rounds. Each client holds 2 classes of images out of a total of 10 classes. $\beta = 0$ indicates that server-side momentum is turned *off*, and $\beta > 0$ indicates that server-side momentum is turned *on*. All the results are obtained based on an average over the last 50 rounds across 4 different random seeds.

| **Datasets** | **Attack** | **Settings** | `CWTM` | `CClip` | `GM` | `CWMED` |
|---|---|---|---|---|---|---|
| **MNIST** | ALIE | $\beta = 0$ | $80.7 \pm 2.6\%$ | $89.8 \pm 1.0\%$ | $67.1 \pm 1.0\%$ | $38.9 \pm 8.0\%$ |
| | | $\beta > 0$ | $83.3 \pm 5.4\%$ ($\uparrow 2.6\%$) | $88.4 \pm 3.1\%$ ($\downarrow -1.4\%$) | $86.0 \pm 4.8\%$ ($\uparrow 18.9\%$) | $21.6 \pm 9.4\%$ ($\downarrow -17.3\%$) |
| | BitFlip | $\beta = 0$ | $67.5 \pm 2.3\%$ | $80.7 \pm 2.0\%$ | $77.4 \pm 1.7\%$ | $68.9 \pm 3.9\%$ |
| | | $\beta > 0$ | $87.1 \pm 3.7\%$ ($\uparrow 19.6\%$) | $89.4 \pm 4.1\%$ ($\uparrow 8.7\%$) | $89.0 \pm 1.7\%$ ($\uparrow 11.6\%$) | $84.8 \pm 1.9\%$ ($\uparrow 15.9\%$) |
| | Mimic | $\beta = 0$ | $70.1 \pm 3.2\%$ | $87.5 \pm 1.1\%$ | $40.5 \pm 6.3\%$ | $19.3 \pm 1.0\%$ |
| | | $\beta > 0$ | $85.9 \pm 3.6\%$ ($\uparrow 15.8$) | $90.0 \pm 4.0\%$ ($\uparrow 2.5$) | $79.2 \pm 4.2\%$ ($\uparrow 38.7$) | $59.6 \pm 6.8\%$ ($\uparrow 40.3$) |
| | IPM | $\beta = 0$ | $73.6 \pm 4.3\%$ | $88.4 \pm 1.0\%$ | $86.2 \pm 2.0\%$ | $68.9 \pm 3.7\%$ |
| | | $\beta > 0$ | $89.2 \pm 4.1\%$ ($\uparrow 15.6\%$) | $90.2 \pm 3.3\%$ ($\uparrow 1.8\%$) | $90.0 \pm 1.2\%$ ($\uparrow 3.8\%$) | $87.1 \pm 1.2\%$ ($\uparrow 18.2\%$) |
| **Fashion MNIST** | ALIE | $\beta = 0$ | $68.6 \pm 2.0\%$ | $71.7 \pm 1.0\%$ | $62.8 \pm 2.5\%$ | $48.0 \pm 3.6\%$ |
| | | $\beta > 0$ | $71.1 \pm 3.7\%$ ($\uparrow 2.5\%$) | $69.6 \pm 2.6\%$ ($\downarrow -2.1\%$) | $64.6 \pm 3.4\%$ ($\uparrow 1.8\%$) | $44.1 \pm 9.4\%$ ($\downarrow -3.9\%$) |
| | BitFlip | $\beta = 0$ | $63.5 \pm 4.6\%$ | $65.8 \pm 3.4\%$ | $64.0 \pm 4.1\%$ | $59.2 \pm 5.2\%$ |
| | | $\beta > 0$ | $73.9 \pm 4.8\%$ ($\uparrow 10.4\%$) | $73.5 \pm 3.0\%$ ($\uparrow 7.7\%$) | $72.7 \pm 4.8\%$ ($\uparrow 8.7\%$) | $69.4 \pm 4.0\%$ ($\uparrow 10.2\%$) |
| | Mimic | $\beta = 0$ | $60.5 \pm 4.0\%$ | $70.3 \pm 2.2\%$ | $37.9 \pm 2.6\%$ | $37.5 \pm 1.0\%$ |
| | | $\beta > 0$ | $70.8 \pm 6.9\%$ ($\uparrow 10.3\%$) | $75.4 \pm 2.2\%$ ($\uparrow 5.1\%$) | $66.3 \pm 4.3\%$ ($\uparrow 28.4\%$) | $55.3 \pm 4.9\%$ ($\uparrow 17.8\%$) |
| | IPM | $\beta = 0$ | $60.9 \pm 5.0\%$ | $66.8 \pm 2.0\%$ | $65.1 \pm 2.0\%$ | $57.2 \pm 4.6\%$ |
| | | $\beta > 0$ | $74.8 \pm 4.5\%$ ($\uparrow 13.9\%$) | $75.2 \pm 3.1\%$ ($\uparrow 8.4\%$) | $68.2 \pm 3.9\%$ ($\uparrow 3.1\%$) | $72.5 \pm 2.6\%$ ($\uparrow 15.3\%$) |

# H Experiment Details

## H.1 Experimental setups

**Hardware and Software Setups.**

- **Hardware.** The simulations are performed on a private cluster of multiple nodes, each node is equipped with 8 CPUs, 64 GB of RAM, and 1 NVIDIA A5000 GPU.
- **Software.** We code our experiments based on PyTorch 2.8.0 and Python 3.9.1.

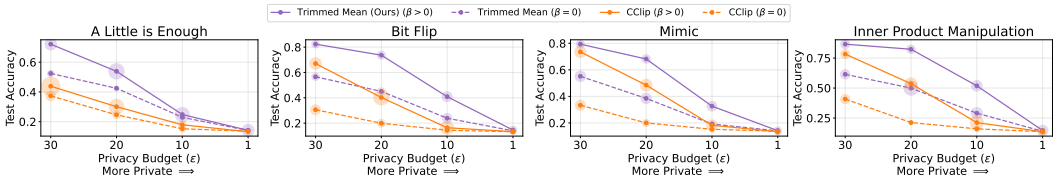
**Neural Network and Hyper-parameter Specifications.**

Table 5 specifies details of the structures of the convolutional neural network and training. The local learning rate $\eta$ is searched based on the best performance after 100 global rounds, over two grids $\{0.75, 0.5, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003\}$. The client-side momentum coefficient $\alpha = 0.1$, and the server-side momentum coefficient $\beta = 0.85$. The DP $\delta$ is set to be $10^{-4}$ throughout.

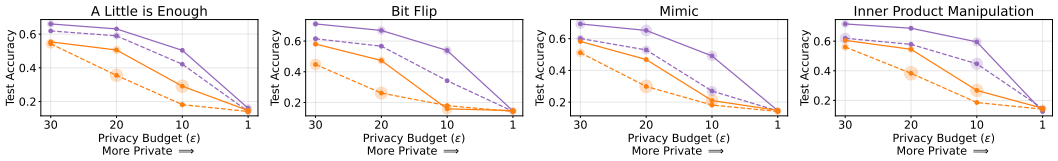
**Datasets and Data Heterogeneity.** The MNIST dataset [44] contains 10 classes of images. In total, there are 50,000 train images and 10,000 test images. The Fashion-MNIST dataset [68] contains 10 classes of images. In total, there are 60,000 train images and 10,000 test images. Each client holds two classes of images.

## H.2 Byzantine Attack Details

We consider four state-of-the-art gradient attacks, including A Little is Enough (ALIE) [13], Bit-flipping (BitFlip) [3], Inner product manipulation (IPM) [69], and Mimic [42].




(a) Test accuracy results on the MNIST dataset.




(b) Test accuracy results on the Fashion-MNIST dataset.

Figure 3: Test accuracy results and comparisons on real-world datasets when the server-side momentum is turned *on* ($\beta > 0$) and when it is turned *off* ($\beta = 0$). The results are averaged over the last 50 rounds across 4 different random seeds. All the experiments run for a total of 800 rounds. The radius of the shaded circle denotes the standard deviation at a given privacy budget $\epsilon$.

Let $\boldsymbol{a}_t$ be the attack vector in round $t$ and $\gamma \geq 0$ a fixed constant. The Byzantine clients send the PS a Byzantine vector $B_t = \bar{\boldsymbol{u}}_t + \gamma \boldsymbol{a}_t$, where $\bar{\boldsymbol{u}}_t = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \boldsymbol{M}_i^t$. We have the following attacks:

- **ALIE.** $\boldsymbol{a}_t = -\boldsymbol{\sigma}_t$, where $\boldsymbol{\sigma}_t \triangleq [\sigma_t^{(1)}, \ldots, \sigma_t^{(d)}]$ and $\sigma_t^{(r)}$ is the coordinate-wise standard deviation of the honest updates at coordinate $r$. All Byzantine clients send $\boldsymbol{B}_t = \bar{\boldsymbol{u}}_t - \gamma \boldsymbol{\sigma}_t$, with $\gamma = z_{\max}$ computed in closed form from $(n, m)$ via the normal CDF [13].
- **BitFlip.** $\boldsymbol{a}_t = -\bar{\boldsymbol{u}}_t$ and $\gamma = 2$. All Byzantine clients send $\boldsymbol{B}_t = -\bar{\boldsymbol{u}}_t$ in round $t$.
- **Inner Product Manipulation.** $\boldsymbol{a}_t = -\bar{\boldsymbol{u}}_t$. All Byzantine clients send $\boldsymbol{B}_t = (1 - \gamma)\bar{\boldsymbol{u}}_t$ in round $t$.

The optimal $\gamma$ for **IPM** is determined by greedily searching over a defined range of values [4]; in each round we choose $\gamma$ to maximize the $\ell_2$ distance between the robust aggregator's output and the average of the honest clients' updates. This strategy ensures the maximum degradation caused by the Byzantine clients in terms of $\ell_2$ distance. A **mimic** attack is defined as all Byzantine clients mimicking a specific honest client by simply sending their updates to the PS. We follow the heuristic strategy in [41] to determine the set of mocking honest clients.

### H.3 Additional results

In the absence of user-level DP protections, Table 6 shows the test accuracies for different robust aggregators, including `CWTM`, `CClip`, `GM`, and `CWMED`, on the MNIST and Fashion-MNIST datasets with server-side momentum enabled and disabled. Detailed definitions of the robust aggregators can be found in Appendix F.5. We can see that the server-side momentum boosts the performance in most cases. Notably, the improvements are significant. For example, `GM` improves over $25\%$ under Mimic attack on the Fashion-MNIST dataset. The improvement stems from the shared anchor $\beta_t \widetilde{R}^{t-1}$, which pulls all clients' updates toward a common direction, thereby strengthening their concentration and improving the robust aggregator's effectiveness against Byzantine clients. Occasionally, we can see a drop in performance. We attribute this to the construction of `CWMED`, which yields a suboptimal robust coefficient [4]. `CClip` is also observed to have a mild accuracy drop. It is interesting to explore how to improve the performance of `CWMED` and `CClip` in a provable way in the future.

Fig. 3 shows performance comparisons with server-side momentum enabled and disabled across different privacy budgets for `CWTM` and `CClip`. Without server-side momentum, both robust aggregators experience a significant performance drop, with extreme cases greater than $30\%$, e.g., for `CClip` under bit-flip and mimic attacks on the MNIST dataset. Turning off momentum amplifies DP noise and degrades the signal-to-noise ratio, drowning out the effective signals. Furthermore, our analysis in Section 6.3 shows that disabling the server-side momentum will incur an unbounded error, which is proportional to the number of rounds $T$, in the convergence upper bound. Consequently, the server-side momentum acts as a critical stabilizer, preventing client-level DP noise from causing catastrophic effects over extended training horizons.

# I NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have faithfully stated our contributions in both the abstract and introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Appendix A for details.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The regulatory assumptions are stated in Section 6. Due to space limitations, we are unable to present all the missing proofs and intermediate results in the main text. They are deferred to Appendix. Please refer to Table of Contents for details.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental and the hyperparameter setups in Section 7 and Appendix H.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our evaluations are based on open-accessed datasets, and our code has been provided as supplemental material.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental setting/details are important parts of reproducing our results. We provide the details in Section 7 and Appendix H to the best of our ability.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our results are averaged over multiple random seeds and accompanied by error bars.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Please find the software/hardware specifications in Appendix H.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The NeurIPS code of ethics is strictly enforced throughout our research.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed broader impacts in Appendix B. We are unaware of any negative impacts.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The existing assets used in this paper has been adequately cited or credited to.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have documented the experiment details in Section 7 and Appendix H.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The paper uses LLM for editing and visualizations.