Differential Privacy in Distributed Learning: Beyond Uniformly Bounded Stochastic Gradients

Yue Huang
Sun Yat-sen University

Jiaojiao Zhang
Great Bay University

Qing Ling
Sun Yat-sen University

Abstract

This paper explores locally differentially private distributed algorithms that solve non-convex empirical risk minimization problems. Traditional approaches often assume uniformly bounded stochastic gradients, which may not hold in practice. To address this issue, we propose differentially **Pri**vate **S**tochastic recursive **M**omentum with gr**A**dient clipping (PriSMA) that judiciously integrates clipping and momentum to enhance utility while guaranteeing privacy. Without assuming uniformly bounded stochastic gradients, given privacy requirement (ϵ, δ), PriSMA achieves a learning error of $\tilde{\mathcal{O}}\big((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{2}{5}}\big)$, where M is the number of clients, N is the number of data samples on each client and d is the model dimension. This learning error bound is better than the state-of-the-art $\tilde{\mathcal{O}}\big((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{1}{3}}\big)$ in terms of the dependence on M and N.

1 INTRODUCTION

Distributed learning enables collaboration between a server and multiple clients to train a common model with distributed data and parallel computation, and has become a fundamental tool for large-scale learning (Kairouz et al., 2021a; McMahan et al., 2017a). In this paper, we consider devising distributed learning algorithms to solve a non-convex empirical risk minimization (ERM) problem in the form of

$$\min_{x\in\mathbb{R}^d} F(x) := \frac{1}{M}\sum_{i=1}^{M} F_i(x),\ F_i(x) := \frac{1}{N}\sum_{j=1}^{N} f_{ij}(x). \quad (1)$$

Therein, x is the model to learn, M is the number of clients, N is the number of data samples on each client, f_{ij} is a non-convex loss function associated with data sample $\mathcal{D}_{ij}$ in the local dataset $\mathcal{D}_i = \{\mathcal{D}_{i1}, \ldots, \mathcal{D}_{iN}\}$ on client i. We define $\mathcal{D} = \bigcup_{i=1}^{M} \mathcal{D}_i$ as the entire dataset across all clients.

Compared to centralized approaches, one significant advantage of solving (1) in a distributed manner is the ability to avoid aggregating raw data samples at a single data center and thus have the potential to protect privacy. However, recent studies (Nasr et al., 2019; Zhu et al., 2019; Jegorova et al., 2023; Rigaki and Garcia, 2023) have highlighted the risk of privacy leakage in distributed learning processes, demonstrating that sensitive information can be exposed through shared stochastic gradients or model updates as iterations accumulate.

To mitigate privacy concerns, differential privacy (DP) offers a robust framework for privacy preservation (Dwork, 2006). In the context of distributed learning, DP adds noise to the transmitted information to ensure that the outputs across neighboring datasets are indistinguishable from potential adversaries (Agarwal et al., 2018; Wei et al., 2020; Cao et al., 2020). In distributed learning, two main variants of DP are global DP (GDP) and local DP (LDP) that address different threats. GDP operates under the assumption of a trustworthy server and secure communication channels, aiming to protect against external adversaries (McMahan et al., 2017b). On the other hand, LDP is able to deal with a more stringent threat, safeguarding against an honest-but-curious server or eavesdroppers that may intercept communications (Kairouz et al., 2021a). In this paper we consider LDP.

Recent studies (Huang et al., 2019; Cao et al., 2020; Zhao et al., 2020; Noble et al., 2022; Li et al., 2022; Lowy et al., 2023) explore distributed ERM under LDP. Therein, Huang et al. (2019); Cao et al. (2020) investigate the alternating direction method of multipliers for convex ERM. Zhao et al. (2020) consider stochastic gradient descent (SGD) with partial partici-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

pation under LDP, but do not provide utility analysis. Li et al. (2022) propose a unified framework of distributed learning for non-convex ERM, with LDP as well as communication compression. Given a privacy budget ϵ (see Definition 1), applying this framework to SGD yields a learning error[1] of $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{1}{2}}\right)$. Noble et al. (2022) introduce DP-SCAFFOLD that combines SCAFFOLD (Karimireddy et al., 2020) with LDP, achieving a same learning error of $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{1}{2}}\right)$. The learning error is improved to $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{2}{3}}\right)$ by Lowy et al. (2023), who propose a distributed variant of DP-SPIDER (Arora et al., 2023). However, these results (Noble et al., 2022; Li et al., 2022; Lowy et al., 2023) rely on assuming uniformly bounded stochastic gradients of loss functions. Such an assumption implies bounded sensitivity (see (7)) and facilitates the analysis, but is often violated in practice.

One approach to addressing this issue is to clip the stochastic gradients. A notable work is differentially private SGD with gradient clipping (DPSGD-GC) (Abadi et al., 2016) that injects noise into clipped SGD with the variance of DP noise being $\Omega(C^2)$, where C is the clipping threshold. Thus, a smaller C implies lower DP noise variance. Some works still assume uniformly bounded stochastic gradients and choose large clipping thresholds that exceed the stochastic gradient bound, rendering the clipping operation ineffective (Zhang et al., 2017; Bassily et al., 2014; Xu et al., 2021). To enable small clipping thresholds and remove the uniformly bounded stochastic gradient assumption, several new analyses have been presented in recent works in the single-machine setting with $M = 1$. Among them, Fang et al. (2022) analyze the learning error of DPSGD-GC under the assumption that the stochastic gradient noise is light-tailed, while Das et al. (2023) provide the analysis when the Lipschitz constants of the sample loss functions are heavy-tailed. In addition, Yang et al. (2022) show that DPSGD-GC achieves a learning error of $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{N\epsilon})^{\frac{1}{2}}\right)$ under an almost sure upper bound on the stochastic gradient variance. The work of Xiao et al. (2023) achieves a learning error of $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{N\epsilon})^{\frac{1}{2}}\right)$ with the help of momentum. However, it relies on a stringent assumption regarding the independence of stochastic gradient noise, and demands $\mathcal{O}(Nd)$ memory to store all per-sample momentums. This space complexity is mitigated in (Xiao et al., 2023) using inner and outer momentums but without theoretical guarantees. In the context of stochastic convex optimization, other than non-convex ERM that we are interested in, Lowy and Razaviyayn (2023); Asi et al. (2024); Zhao et al. (2024) analyze the learning error of DPSGD-GC by assuming bounded k-th moment. However, this assumption implies that the full gradient of the loss function is uniformly bounded, a condition that our paper does not impose (see Section 4.3 for details). Note that above works (Fang et al., 2022; Das et al., 2023; Yang et al., 2022; Xiao et al., 2023; Lowy and Razaviyayn, 2023; Asi et al., 2024; Zhao et al., 2024), though removing the uniformly bounded stochastic gradient assumption, require other strong assumptions to show the utility-privacy trade-off.

Fully removing the assumption of uniformly bounded stochastic gradients and enabling a small clipping threshold C are challenging. On one hand, the magnitudes of the stochastic gradients may become very large during distributed learning. On the other hand, a small C that is beneficial for privacy protection may cause the clipped stochastic gradients to deviate significantly from the true stochastic gradients, not mentioning the full gradients. This is verified by the work of (Koloskova et al., 2023) in the single-machine setting with $M = 1$. It indicates that DPSGD-GC without DP noise (called as SGD-GC) incurs a lower bound of $\Omega(\frac{\tilde{\sigma}^2}{C})$ on the learning error, where $\tilde{\sigma}^2$ is the stochastic gradient variance. By the analysis of (Koloskova et al., 2023), careful selection of the clipping threshold in DPSGD-GC yields a learning error of $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{N\epsilon})^{\frac{1}{3}}\right)$. The same conclusion is made in (Li et al., 2024), which extends DPSGD-GC to the distributed setting and achieves a learning error of $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{1}{3}}\right)$[2], worse than that assuming uniformly bounded stochastic gradients.

Contributions. In this paper, we propose differentially **Pri**vate **S**tochastic recursive **M**omentum with gr**A**dient clipping, abbreviated as PriSMA, for solving the non-convex ERM problem in the form of (1). Our main contributions are as follows.

C1) We prove that any algorithm that only accesses clipped stochastic gradients suffers from a lower bound of $\Omega(\min\{\sigma, \frac{\sigma^2}{C}\})$ on the learning error, where σ^2 is the stochastic gradient variation and C is the clipping threshold.

C2) We design an LDP algorithm called as PriSMA, which innovatively integrates the techniques of clipping and momentum. First, without adding the DP noise, we prove that PriSMA achieves the lower bound of $\Omega(\min\{\sigma, \frac{\sigma^2}{C}\})$. Second, given a privacy budget ϵ, we

[1] The learning error in this paper is in terms of the expected gradient norm $\mathbb{E}\|\nabla F(x)\|$ unless otherwise specified.

[2] This error bound is in (Li et al., 2024, Appendix D.5.3). Their results are divided into two cases: one for small $\tilde{\sigma}^2$ and another for large $\tilde{\sigma}^2$. We focus on the case where the stochastic gradient variance $\tilde{\sigma}^2 \sim \Theta(1)$, which falls into large $\tilde{\sigma}^2$.

prove that clipping and momentum can help our algorithm use less DP noise than DPSGD-GC and achieve a learning error of $\tilde{\mathcal{O}}\big((\frac{\sqrt{d}}{\sqrt{MN}\epsilon})^{\frac{2}{5}}\big)$. Our result is better in order than the learning error of $\tilde{\mathcal{O}}\big((\frac{\sqrt{d}}{\sqrt{MN}\epsilon})^{\frac{1}{3}}\big)$ given by DPSGD-GC.

Notations. Given a set $\mathcal{B}$, we use $|\mathcal{B}|$ to denote the cardinality. Given functions $f, g : X \to [0, \infty)$, where X is any set, we say $f = \mathcal{O}(g)$ if there exists a constant $c < \infty$ such that $f(x) \le cg(x)$ for all $x \in X$, while $f = \Omega(g)$ if there exists a constant $c > 0$ such that $f(x) \ge cg(x)$ for all $x \in X$. We use $f = \tilde{\mathcal{O}}(g)$ as shorthand for $f = \mathcal{O}(g \max\{1, \log g\})$. We use $\mathcal{N}(\mu, \Sigma)$ to denote a Gaussian distribution with mean μ and covariance matrix Σ, $\|\cdot\|$ to denote ℓ_2-norm, Pr to denote the probability of a random event, and $\mathbb{E}_j[x]$ to denote the expectation of random variable x w.r.t. j. We use $[M]$ to collect integers from 1 to M.

2 PRELIMINARIES ON LDP

This paper focuses on a distributed learning problem where the server is not trusted. In this context, we begin with introducing the concept of LDP in distributed learning (Kairouz et al., 2021a; Yin et al., 2021; Nguyen et al., 2023; Yi et al., 2024). Recall that $\mathcal{D}_i := \{\mathcal{D}_{i1}, \dots, \mathcal{D}_{iN}\}$ denotes the local dataset on client i. Two local datasets $\mathcal{D}_i$ and $\mathcal{D}'_i$ are neighboring if they differ by only one data sample. Let us consider a randomized distributed algorithm $\mathcal{A}$ that is synchronized and iterates for T times. The input of $\mathcal{A}$ is the overall dataset $\mathcal{D} := \bigcup_{i=1}^{M} \mathcal{D}_i$, while the output contains all the messages transmitted by the clients to the server over T times. At time t, the message transmitted by client i to the server is denoted as Z_i^t, which depends on the local dataset $\mathcal{D}_i$ and previous messages $Z_j^1, \dots, Z_j^{t-1}$ received by the server from other clients $j \neq i$. For notational convenience, let $Z_i := \{Z_i^1, \dots, Z_i^T\}$ be the concatenation of the messages transmitted by client i over T times and $Z_{-i} := \{Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_M\}$ be the concatenation of the messages transmitted by all clients but i over T times.

Likewise, denote $\mathcal{A}_i$ as the local algorithm of $\mathcal{A}$ on client i. The input of $\mathcal{A}_i$ contains $\mathcal{D}_i$ and Z_{-i}, the local dataset and all the messages transmitted by the other clients, while the output contains Z_i, all the messages transmitted by client i.

Following these notations, we define LDP as follows.

Definition 1 ((ϵ, δ)-LDP)**.** *A randomized distributed algorithm $\mathcal{A}$ satisfies (ϵ, δ)-LDP, if for any client i, any neighboring datasets $\mathcal{D}_i$ and $\mathcal{D}'_i$, and any event $\mathcal{E}_i$ in the output of $\mathcal{A}_i$, it holds that*

$$\Pr[\mathcal{A}_i(\mathcal{D}_i, Z_{-i}) \in \mathcal{E}_i] \le e^{\epsilon} \Pr[\mathcal{A}_i(\mathcal{D}'_i, Z_{-i}) \in \mathcal{E}_i] + \delta.$$

Therein, $\epsilon > 0$ is the privacy budget, and $\delta \in (0, 1)$ is the permissible probability of privacy leakage. Smaller values of ϵ and δ indicate stricter privacy protection. The notion of LDP presented in Definition 1 is also referred to as other terms in the literature, such as "inter-silo-record-level DP" (Lowy et al., 2023) or "silo-level LDP" (Zhou and Chowdhury, 2024).

Next, we introduce the concept of Rényi DP (RDP) in the context of distributed learning. RDP is particularly useful for analyzing the composition of randomized mechanisms and the privacy amplification through subsampling (Mironov, 2017).

Definition 2 ((α, ρ)-RDP)**.** *A randomized distributed algorithm $\mathcal{A}$ satisfies (α, ρ)-RDP where $\alpha > 1$ and $\rho > 0$, if for any client i and any neighboring datasets $\mathcal{D}_i$ and $\mathcal{D}'_i$, it holds that*

$$D_\alpha(\mathcal{A}_i(\mathcal{D}_i, Z_{-i}) \| \mathcal{A}_i(\mathcal{D}'_i, Z_{-i})) \le \rho, \tag{2}$$

in which $D_\alpha(X\|Y)$ denotes the α-order Rényi divergence between the distributions of random variables X and Y.

In the analysis, we will use the following lemma to convert the privacy guarantee from (α, ρ)-RDP to (ϵ, δ)-LDP (Mironov, 2017).

Fact 1 (Conversion from RDP to LDP)**.** *If a randomized distributed algorithm $\mathcal{A}$ satisfies (α, ρ)-RDP, then it satisfies $\left(\rho + \frac{\log \frac{1}{\delta}}{\alpha - 1}, \delta\right)$-LDP for any $\delta \in (0, 1)$.*

3 ALGORITHM DEVELOPMENT

Below, we propose PriSMA to solve (1).

The proposed algorithm involves the concept of gradient clipping, which is frequently employed in learning algorithms to control the magnitudes of stochastic gradients. Given a stochastic gradient $\nabla f_{ij}(x)$ as the input, gradient clipping outputs

$$\text{clip}_C(\nabla f_{ij}(x)) := \min\left\{\frac{C}{\|\nabla f_{ij}(x)\|}, 1\right\} \cdot \nabla f_{ij}(x), \tag{3}$$

where $C > 0$ is the threshold.

With this concept, we summarize the compact form of the proposed algorithm for any time $t \ge 1$ as

$$\begin{cases} v_i^t = g_i^t + (1-\gamma)(v_i^{t-1} - \tilde{g}_i^t) + \xi_i^t, \ \forall i, \\ x^{t+1} = x^t - \eta \cdot \text{clip}_{C_2}\left(\dfrac{1}{M}\displaystyle\sum_{i=1}^{M} v_i^t\right). \end{cases} \tag{4}$$

Therein, $\gamma > 0$ is the momentum step size, $\eta > 0$ is the step size, $C_2 > 0$ is the threshold, $\xi_i^t \sim \mathcal{N}(0, \sigma_1^2 \mathbb{I}_d)$ is the Gaussian noise, and g_i^t and $\tilde{g}_i^t$ are the clipped stochastic gradients given by

$$g_i^t := \frac{1}{b} \sum_{j \in \mathcal{B}_i^t} \mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^t)\right), \tag{5a}$$

$$\tilde{g}_i^t := \frac{1}{b} \sum_{j \in \mathcal{B}_i^t} \mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^{t-1})\right), \tag{5b}$$

where $\mathcal{B}_i^t$ is the mini-batch, $b > 0$ is the batch size and $C_1 > 0$ is the threshold.

The per-client implementation of PriSMA is outlined in Algorithm 1. At time $t \ge 1$, each client i independently and uniformly samples a subset of data without replacement, indexed by $\mathcal{B}_i^t$ with a size of $|\mathcal{B}_i^t| = b$. Then, for each data sample $j \in \mathcal{B}_i^t$, client i computes the stochastic gradients at the current model x^t and the previous model x^{t-1}. These stochastic gradients are individually clipped by client i with a threshold C_1, and averaged over $\mathcal{B}_i^t$ to obtain g_i^t and $\tilde{g}_i^t$, respectively corresponding to x^t and x^{t-1}. Using g_i^t and $\tilde{g}_i^t$, each client i constructs a local momentum direction v_i^t with a Gaussian noise $\xi_i^t \sim \mathcal{N}(0, \sigma_1^2 \mathbb{I}_d)$ added for privacy protection and transmits v_i^t to the server. Upon receiving all v_i^t from the clients, the server averages them, clips the result with a threshold C_2, and updates the model x^t. At each time, PriSMA requires each client to compute $2b$ stochastic gradients, send a d-dimensional vector, and use $\mathcal{O}(d)$ storage.

At time $t = 0$, each client i initializes v_i^0 by computing the stochastic gradients for a subset of data $\mathcal{B}_i^0$ with a size of $|\mathcal{B}_i^0| = b$, clipping them using a threshold C_1, averaging the results, and then adding a Gaussian noise $\xi_i^0 \sim \mathcal{N}(0, \sigma_0^2 \mathbb{I}_d)$ to protect the privacy of v_i^0.

Small DP noise to guarantee privacy. PriSMA makes use of four fundamental techniques: stochastic gradients, clipping, momentum, and DP. The innovation lies in our novel algorithm design that leverages the clipping technique, not merely to avoid the bounded stochastic gradient assumption but to ingeniously integrate with momentum. Such an integration significantly lowers the DP noise to be added and improves the utility. We illustrate as follows.

According to the local update on each client i at time t in (4), define a query function $q_i^t(\mathcal{B}_i^t)$ as

$$q_i^t(\mathcal{B}_i^t) := \gamma g_i^t + (1-\gamma)(g_i^t - \tilde{g}_i^t). \tag{6}$$

Thus, (4) is equivalent to $v_i^t = q_i^t(\mathcal{B}_i^t) + \xi_i^t + (1-\gamma)v_i^{t-1}$, where we add a Gaussian noise ξ_i^t to protect the privacy of v_i^t. Since v_i^{t-1} is already differentially private, according to the post-processing property (Dwork and Roth, 2014), for a given pair of (ϵ, δ), the amount of DP noise to be added is related to the so-called sensitivity of the query function $q_i^t(\mathcal{B}_i^t)$, defined as

$$\begin{aligned}\Delta(q_i^t(\mathcal{B}_i^t)) &:= \sup_{\mathcal{B}_i^t, \mathcal{B}_i^{t\prime}} \|q_i^t(\mathcal{B}_i^t) - q_i^t(\mathcal{B}_i^{t\prime})\| \qquad (7)\\ &\le \gamma\|g_i^t - g_i^{t\prime}\| + (1-\gamma)\left\|\left(g_i^t - \tilde{g}_i^t\right) - \left(g_i^{t\prime} - \tilde{g}_i^{t\prime}\right)\right\|,\end{aligned}$$

where $\mathcal{B}_i^t$ and $\mathcal{B}_i^{t\prime}$ are neighboring local datasets, while $g_i^{t\prime}$ and $\tilde{g}_i^{t\prime}$ are respectively the counterparts of g_i^t and $\tilde{g}_i^t$ when we replace $\mathcal{B}_i^t$ with $\mathcal{B}_i^{t\prime}$. Smaller sensitivity means smaller amount of DP noise to be added to ensure (ϵ, δ)-LDP.

Algorithm 1 Proposed PriSMA

```
Input: step sizes $\gamma$ and $\eta$; clipping thresholds $C_1$ and $C_2$; batch size $b$; variances of DP noises $\sigma_0^2$ and $\sigma_1^2$; total time $T$
Initialization: $x^0 = 0$
for $t = 0, 1, \ldots, T-1$ do
    for each client $i = 1, 2, \ldots, M$ in parallel do
        Receive $x^t$ from the server
        Independently uniformly sample a subset of data without replacement, indexed by $\mathcal{B}_i^t$, with a size of $|\mathcal{B}_i^t| = b$
        if $t = 0$ then
            Update $v_i^t = \frac{1}{b}\sum_{j \in \mathcal{B}_i^t} \mathrm{clip}_{C_1}(\nabla f_{ij}(x^t)) + \xi_i^0$, where $\xi_i^0 \sim \mathcal{N}(0, \sigma_0^2 \mathbb{I}_d)$
        else
            Compute $g_i^t$ and $\tilde{g}_i^t$ according to (5)
            Update $v_i^t = g_i^t + (1-\gamma)(v_i^{t-1} - \tilde{g}_i^t) + \xi_i^t$, where $\xi_i^t \sim \mathcal{N}(0, \sigma_1^2 \mathbb{I}_d)$
        end if
        Send $v_i^t$ to the server
    end for
    Server
    Receive $v_i^t$ from all clients
    Update $x^{t+1} = x^t - \eta \cdot \mathrm{clip}_{C_2}\left(\frac{1}{M}\sum_{i=1}^M v_i^t\right)$
    Send $x^{t+1}$ to all clients
end for
```

Observe the right-hand side of (7). The first term can be small if the momentum step size γ and the threshold C_1 are small, while the second term can be well-controlled using the distance between two successive models, x^t and x^{t-1}. Indeed, substituting (4) and (5) into (7), and assuming L-smoothness of the loss functions (see Assumption 1), we obtain

$$\left\|\left(g_i^t - \tilde{g}_i^t\right) - \left(g_i^{t\prime} - \tilde{g}_i^{t\prime}\right)\right\| \le \frac{2L\|x^t - x^{t-1}\|}{b} \le \frac{2L\eta C_2}{b}. \tag{8}$$

As shown in (8), thanks to clipping with a threshold C_2 on the server, a small step size η results in a small sensitivity caused by $g_i^t - \tilde{g}_i^t$. Incorporating (8) into

(7), we have

$$\Delta(q_i^t(\mathcal{B}_i^t)) \leq \gamma \frac{2C_1}{b} + (1-\gamma)\frac{2L\eta C_2}{b}, \qquad (9)$$

which illustrates PriSMA's effective integration of clipping and momentum to lower the sensitivity comparing to DPSGD-GC. To see so, observe that if we set $\gamma = 1$ and remove clipping with the threshold C_2 on the server, Algorithm 1 recovers DPSGD-GC with a clipping threshold of C_1, such that the sensitivity is $\frac{2C_1}{b}$. As long as we choose $C_2 < \frac{C_1}{L\eta}$, the sensitivity in (9) is less than $\frac{2C_1}{b}$.

However, using momentum also comes with a cost. Because v_i^t not only incorporates $g_i^t + \xi_i^t$ but also uses $(1-\gamma)(v_i^{t-1} - \tilde{g}_i^t)$, it leads v_i^t to accumulate the historical stochastic gradients and DP noises up to time t. In contrast, in DPSGD-GC, $v_i^t = g_i^t + \xi_i^t$ only contains the stochastic gradient and DP noise at time t. To overcome this challenge, we leverage the fact of $1-\gamma < 1$ to establish a contractive property on $\frac{1}{M}\sum_{i=1}^M v_i^t$ (see Lemma D.1 in Appendices), construct an appropriate Lyapunov function, and judiciously select the value of C_2, ultimately achieving a better utility-privacy trade-off than the existing works.

4 ANALYSIS

In this section, we establish the utility-privacy trade-off of PriSMA. The following mild assumptions are made on the loss functions.

Assumption 1 (L-smoothness). *Each loss function is L-smooth. For any $x, y \in \mathbb{R}^d$, we have*

$$\|\nabla f_{ij}(x) - \nabla f_{ij}(y)\| \leq L\|x-y\|, \quad \forall i, j.$$

Assumption 2 (Bounded gradient variation). *For any $x \in \mathbb{R}^d$, there exists a constant $\sigma > 0$ such that*

$$\frac{1}{MN}\sum_{i=1}^M \sum_{j=1}^N \|\nabla f_{ij}(x) - \nabla F(x)\|^2 \leq \sigma^2.$$

Assumption 1 is common for non-convex analysis. In the single-machine setting with $M = 1$ or if $\mathcal{D}_i$ for all clients i are independent and identically distributed, the commonly used assumption of bounded stochastic gradient variance in stochastic optimization (Koloskova et al., 2023; Cutkosky and Orabona, 2019; Fang et al., 2018) implies Assumption 2.

In our algorithm, we encounter both stochastic gradient noise and DP noise. The use of the clipping technique complicates the analysis by intertwining its resultant error with these noises. To clearly present our analysis, we decompose the analysis into two stages: (i) We consider the special case where the DP noise is zero, establish a lower bound for the error due to the clipping technique and the stochastic gradient noise, and then prove that our algorithm achieves this lower bound in the absence of the DP noise. (ii) Under this tight convergence analysis framework, we continue to consider the DP noise and analyze the utility-privacy trade-off of the proposed PriSMA.

4.1 Lower Bound of Clipping-based Stochastic Gradient Algorithms

Let us consider the special case where no DP noise appears in PriSMA, termed as Non-PriSMA. Because Non-PriSMA applies clipping to each stochastic gradient, we establish a lower bound for the error induced by clipping and stochastic gradient noise. To this end, we define $\mathcal{P}_\sigma^L$ as the class of problems in the form of (1) and satisfying Assumptions 1 and 2, $\mathcal{A}_{C_1}$ as the class of algorithms that can only access clipped stochastic gradients with a threshold $C_1 > 0$. Further, for any algorithm $\mathcal{A} \in \mathcal{A}_{C_1}$, if problems $\mathcal{P}_0$, $\mathcal{P}_1 \in \mathcal{P}_\sigma^L$ satisfy that for any $i \in [M]$, $j \in [N]$ and $x \in \mathbb{R}^d$,

$$\text{clip}_{C_1}\left(\nabla f_{ij}^{\mathcal{P}_0}(x)\right) = \text{clip}_{C_1}\left(\nabla f_{ij}^{\mathcal{P}_1}(x)\right), \qquad (10)$$

where $f_{ij}^{\mathcal{P}_0}$ and $f_{ij}^{\mathcal{P}_1}$ are the loss functions w.r.t. the data sample $\mathcal{D}_{ij}$ in $\mathcal{P}_0$ and $\mathcal{P}_1$ respectively, then the outputs of $\mathcal{A}$ on $\mathcal{P}_0$ and $\mathcal{P}_1$ are the same. Therefore, Non-PriSMA belongs to $\mathcal{A}_{C_1}$.

The lower bound of $\mathcal{A}_{C_1}$ on $\mathcal{P}_\sigma^L$ is presented as follows.

Theorem 1 (Lower bound of $\mathcal{A}_{C_1}$ on $\mathcal{P}_\sigma^L$). *For any clipping threshold $C_1 > 0$ and any algorithm $\mathcal{A} \in \mathcal{A}_{C_1}$, there exists a problem $\mathcal{P} \in \mathcal{P}_\sigma^L$, for which the solution yielded by $\mathcal{A}$ on $\mathcal{P}$, denoted as $x_{\mathcal{A},\mathcal{P}}$, satisfies*

$$\mathbb{E}\|\nabla F(x_{\mathcal{A},\mathcal{P}})\| \geq \Omega\left(\min\left\{\sigma, \frac{\sigma^2}{C_1}\right\}\right). \qquad (11)$$

The lower bound presented in (11) is consistent with the one given by (Koloskova et al., 2023). However, ours is more general, since Koloskova et al. (2023) only consider the lower bound of SGD-GC on $\mathcal{P}_\sigma^L$. In contrast, our analysis holds for any algorithm within $\mathcal{A}_{C_1}$ on $\mathcal{P}_\sigma^L$. For more details, readers are referred to the discussions in Appendix C.

4.2 Utility of Non-PriSMA (PriSMA without DP Noise)

Because Non-PriSMA belongs to $\mathcal{A}_{C_1}$, the error caused by the clipping technique and stochastic gradient noise is at least $\Omega\left(\min\left\{\sigma, \frac{\sigma^2}{C_1}\right\}\right)$. Next, we conduct a utility analysis to prove that Non-PriSMA is able to reach this lower bound.

Before presenting the theoretical results, let us examine the "fixed point" of Non-PriSMA. Although this fixed point does not exist, it helps us clarify how our algorithm works and motivates our analysis. Let us ignore DP noise and stochastic gradient noise, and assume that the proposed algorithm stops at a fixed point, denoted as $(x^*, v_1^*, \dots, v_M^*)$. Then (4) gives

$$v_i^* = \frac{1}{N}\sum_{j=1}^{N} \text{clip}_{C_1}\left(\nabla f_{ij}(x^*)\right), \quad \forall i. \tag{12}$$

Motivated by (12), we characterize the distances between v_i and $\frac{1}{N}\sum_{j=1}^{N}\text{clip}_{C_1}\left(\nabla f_{ij}(x)\right)$ for all i, where we omit the time index. To this end, let us define

$$\begin{aligned} \nabla F_i^{C_1}(x) &:= \frac{1}{N}\sum_{j=1}^{N} \text{clip}_{C_1}\left(\nabla f_{ij}(x)\right), \\ \nabla F^{C_1}(x) &:= \frac{1}{M}\sum_{i=1}^{M}\nabla F_i^{C_1}(x), \quad \bar{v} := \frac{1}{M}\sum_{i=1}^{M} v_i. \end{aligned} \tag{13}$$

With stochastic gradient noise and without DP noise, the distance between $\bar{v}^t$ and $\nabla F^{C_1}(x^t)$ evolves as in the following lemma.

Lemma 1. *Under Assumptions 1 and 2, for any $\gamma \in (0,1)$, Non-PriSMA satisfies*

$$\begin{aligned} &\mathbb{E}\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2 \\ \le{}& (1-\gamma)^2\mathbb{E}\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + 2\gamma^2\frac{\sigma^2}{Mb} \\ &+ 2(1-\gamma)^2\frac{L^2}{Mb}\mathbb{E}\|x^{t+1}-x^t\|^2. \end{aligned} \tag{14}$$

At the right-side hand of (14), the first term indicates a contraction property of $\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2$. But there exist two error terms, one caused by stochastic gradient noise and another caused by the evolution of the model.

With the aid of Lemma 1, in the following theorem, we establish the utility of Non-PriSMA, which attains the lower bound given in Theorem 1 and is hence tight.

Theorem 2 (Utility of Non-PriSMA)**.** *Under Assumptions 1 and 2, set the clipping thresholds such that $C_1 \ge 2C_2$ and $C_1 \ge 16\sigma$, while set the step sizes such that*

$$\gamma = \min\left\{\frac{MbC_1C_2}{32\sigma^2}, \sqrt{\frac{M}{T}}\right\}, \tag{15a}$$

$$\eta = \min\left\{\frac{\sqrt{Mb\gamma}}{2\sqrt{2}L}, \frac{Mb\gamma}{4}, \frac{C_1\gamma}{LC_2}, \frac{1}{6L}, \sqrt{\frac{M}{T}}\right\}. \tag{15b}$$

Then, the sequence $\{x^t\}_t$ generated by Non-PriSMA satisfies

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\| \\ \le{}&\mathcal{O}\left(\frac{1}{(MT)^{\frac{1}{2}}C_2} + \frac{1}{(MT)^{\frac{1}{4}}} + \frac{\sigma^4}{C_1^2C_2} + \frac{\sigma^2}{C_1}\right). \end{aligned} \tag{16}$$

Under the conditions in Theorem 2, if we further choose $\frac{C_1}{2} \ge C_2 \ge \frac{\sigma^2}{C_1}$ which can be satisfied since $\frac{\sigma^2}{C_1} \le \frac{\sigma}{16} < \frac{C_1}{2}$, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\| \le \mathcal{O}\left(\frac{\sigma^2}{C_1}\right) = \mathcal{O}\left(\min\left\{\sigma, \frac{\sigma^2}{C_1}\right\}\right),$$

where we use $\frac{\sigma^2}{C_1} \le \sigma$. Thus, Non-PriSMA attains the lower bound in Theorem 1 asymptotically. This shows that we have optimally handled the impacts of clipping and stochastic gradient noise in the absence of DP noise under the conditions in Theorem 2.

4.3 Utility-privacy Trade-off of PriSMA

Based on the tight analysis framework above, we continue to establish the utility-privacy trade-off of PriSMA subject to pre-defined (ϵ, δ)-LDP. According to (Wang et al., 2023, Lemma 3.7), we determine the variances of DP noises to meet (α, ρ)-RDP, as shown in the following theorem.

Theorem 3 (RDP guarantee)**.** *Under Assumption 1, for a given (α, ρ)-RDP requirement with $\alpha > 1$ and $\rho > 0$, let the variances of injected Gaussian noises in PriSMA satisfy*

$$\sigma_0^2 = \frac{3.5T\Delta_0^2\alpha}{N^2\rho}, \; \sigma_1^2 = \frac{3.5T\Delta_1^2\alpha}{N^2\rho}, \tag{17}$$

where $\Delta_0 = 2C_1$ and $\Delta_1 = 2\left(\gamma C_1 + (1-\gamma)L\eta C_2\right)$. If it holds that

$$\hat{\sigma}^2 := \frac{3.5b^2T\alpha}{N^2\rho} \ge 0.7, \; \alpha \le \frac{2\hat{\sigma}^2}{3}\log\frac{N}{b\alpha(1+\hat{\sigma}^2)} + 1, \tag{18}$$

then PriSMA satisfies (α, ρ)-RDP.

With Fact 1, we can convert (α, ρ)-RDP to (ϵ, δ)-LDP, as outlined in the following corollary.

Corollary 1 (LDP guarantee)**.** *Under the same conditions as in Theorem 3, suppose the variances of the injected Gaussian noises in PriSMA are set as σ_0^2 and σ_1^2 in (17), and*

$$\alpha = 1 + \frac{2\log\frac{1}{\delta}}{\epsilon}, \; \rho = \frac{\epsilon}{2},$$

where $\epsilon > 0$ and $\delta \in (0, 1)$. Then, the proposed PriSMA satisfies (ϵ, δ)-LDP. Futhermore, by setting $b = \Theta(\frac{N\sqrt{\epsilon}}{\sqrt{T}})$ and $T \geq \mathcal{O}(\frac{\log^4 \frac{1}{\delta}}{\epsilon^3})$, the condition (18) in Theorem 3 is satisfied.

Thanks to the ingenious use of the clipping technique in PriSMA, our privacy analysis no longer requires to assume uniformly bounded stochastic gradients. Additionally, the DP variance in (17) can be small by choosing small step sizes η and γ, as well as small clipping thresholds C_1 and C_2.

With the variances of DP noises in (17) at hand, we can express the variances in the upper bound of learning error using the parameters of LDP requirement (ϵ, δ), thereby finally establishing the utility-privacy trade-off for PriSMA as follows.

Theorem 4 (Utility-privacy trade-off for PriSMA)**.** *Suppose Assumptions 1–2 hold. Given LDP requirement (ϵ, δ), by choosing appropriate parameters, the sequence $\{x^t\}_t$ generated by PriSMA satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\| \leq \tilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}}{\sqrt{M}N\epsilon}\right)^{\frac{2}{5}}\right). \tag{19}$$

The specific values of the parameters are provided in Appendix F. Of particular note, under the assumptions we make, the utility-privacy trade-off of DPSGD-GC is $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{1}{3}}\right)$, worse than our $\tilde{\mathcal{O}}\left((\frac{\sqrt{d}}{\sqrt{M}N\epsilon})^{\frac{2}{5}}\right)$.

Comparing assumptions and analytical tools. Similar to PriSMA, some recent works (Arora et al., 2023; Tran and Cutkosky, 2022; Wang et al., 2023; Lowy et al., 2023) also focus on reducing sensitivity for better utility; yet, they do not use clipping but assume uniformly bounded stochastic gradients. In contrast, PriSMA operates without assuming uniformly bounded stochastic gradients, which necessitates novel analytical tools to quantify the errors caused by clipping; see Lemma D.3.

In the DP stochastic convex optimization literature, some works (Wang et al., 2020; Kamath et al., 2022; Lowy and Razaviyayn, 2023; Asi et al., 2024; Zhao et al., 2024) relax the uniformly bounded stochastic gradient assumption to a bounded k-th moment assumption on the stochastic gradients. Their learning error[3] analyses depend on the bounded estimation errors[4] of their gradient estimators. Specifically, Wang et al. (2020); Kamath et al. (2022) estimate the full gradient based on the robust mean estimator in (Holland, 2019), while Lowy and Razaviyayn (2023); Asi et al. (2024); Zhao et al. (2024) use clipped stochastic gradient as the estimator. Zhao et al. (2024) employ an iterative updating method to further reduce the estimation error of the clipped stochastic gradient. However, the property of bounded estimation error for these estimators requires the full gradient $\nabla F(x)$ to be uniformly bounded, which is implied by the k-th bounded moment assumption but not our Assumption 2. Under Assumption 2, the gradient estimation error in PriSMA can be large, so that we focus on bounding the inner product between the gradient estimator and the true gradient, as shown in Lemma D.3, which is sufficient to establish PriSMA's learning error under bounded variation.

Removing dependence on L. As depicted in (17), the variance σ_1^2 is dependent on the smoothness constant L, whose exact value unknown in practice, although it is possible to roughly estimate an upper bound. To eliminate the dependency on L, we introduce another clipping with a threshold $C_3 > 0$ on the clipped stochastic gradient difference, modifying the update of v_i^t to be

$$v_i^t = (1-\gamma)v_i^{t-1} + \frac{\gamma}{b}\sum_{j\in\mathcal{B}_i^t}\text{clip}_{C_1}\left(\nabla f_{ij}(x^t)\right) + \xi_i^t \tag{20}$$
$$+\frac{1-\gamma}{b}\sum_{j\in\mathcal{B}_i^t}\text{clip}_{C_3}(\text{clip}_{C_1}(\nabla f_{ij}(x^t)) - \text{clip}_{C_1}(\nabla f_{ij}(x^{t-1}))).$$

Following the similar steps in deriving (9), the corresponding sensitivity of $q_i^t(\mathcal{B}_i^t)$ given by (20) satisfies

$$\Delta(q_i^t(\mathcal{B}_i^t)) \leq \frac{2(\gamma C_1 + (1-\gamma)C_3)}{b}, \tag{21}$$

and thus Corollary 1 holds with

$$\sigma_1^2 = \frac{14T(\gamma C_1 + (1-\gamma)C_3)^2\alpha}{N^2\rho},$$

which does not rely on L anymore.

When we set a sufficiently large threshold C_3 such that $C_3 \geq L\eta C_2$, clipping with C_3 becomes inactive, such that the modified algorithm with C_3 is equivalent to the original PriSMA. Therefore, all theoretical results still hold for the modified algorithm (20).

Tuning hyperparameters. As we observe from preliminary numerical experiments, C_2 is relatively easy to tune. C_3 is used to remove the dependence of the DP noise variance on the smoothness constant L. Thus, C_3 is no longer necessary if we can rely on a roughly estimated upper bound of L as in many existing works (Beck and Teboulle, 2009; Zhang and Hong, 2020). Besides, since acceleration is widely used in the literature, there are various rules-of-thumb to tune the momentum step size γ. Last but not least, it is possible to apply auto-clipping techniques (Bu et al., 2023)

[3] Here the learning error is on the expected suboptimality gap $\mathbb{E}[F(x) - \min_x F(x)]$, assuming convexity of F.

[4] The estimation error of gradient estimator $\hat{g}(x)$ is in terms of $\mathbb{E}\|\hat{g}(x) - \nabla F(x)\|^2$.

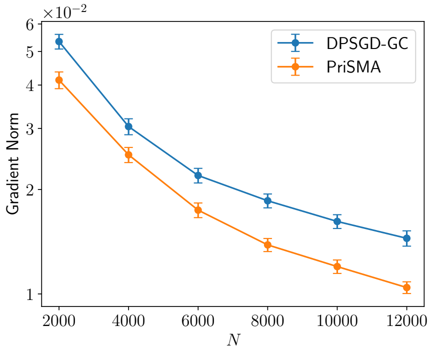

Figure 1: Non-convex regularized least squares.

in PriSMA to avoid hand-tuning some clipping thresholds. We will explore this direction in our future work.

5 NUMERICAL EXPERIMENTS

Non-convex regularized least squares. We first consider a distributed non-convex regularized least squares problem in the form of (1), with

$$f_{ij}(x) = \frac{1}{2}(a_{ij}^T x - b_{ij})^2 + \frac{\lambda}{2}\sum_{k=1}^{d}\frac{(x^{(k)})^2}{1+(x^{(k)})^2}, \qquad (22)$$

for all $i \in [M]$, $j \in [N]$. Therein, $a_{ij} \in \mathbb{R}^d$ and $b_{ij} \in \mathbb{R}$ constitute the data sample $\mathcal{D}_{ij}$, and $x^{(k)}$ represents the k-th dimension of $x \in \mathbb{R}^d$. Note that such a loss function does not have bounded gradient when x is unbounded. We aim to validate the improved utility-privacy trade-off of PriSMA over DPSGD-GC in terms of N, the number of data samples at each client.

We set the number of clients $M = 10$, the model dimension $d = 10$, the regularization parameter $\lambda = 1$, as well as the parameters of LDP requirement $\epsilon = 1$ and $\delta = 10^{-4}$. We assess PriSMA and DPSGD-GC for $N \in \{2000 \times l,\ l = 1, 2, \cdots, 6\}$. When $N = 2000$, each dimension of a_{ij} is independently and uniformly randomly sampled from $[-1, 1]$, while the ground truth $x^* \in \mathbb{R}^d$ is sampled from $\mathcal{N}(0, \mathbb{I}_d)$. With a_{ij} and x^*, b_{ij} is set to $a_{ij}^T x^* + \phi_{ij}$, where $\phi_{ij} \sim \mathcal{N}(0, 2)$. For $N = 2000 \times l$ with $l > 1$, we duplicate the data samples generated at $N = 2000$ for l times, ensuring that $F(x)$ remains the same for different N.

In PriSMA and DPSGD-GC, for each N, we search the step size $\eta \in \{1 \times 10^l, 2 \times 10^l, 5 \times 10^l, l = -1, -2, -3\}$, the clipping threshold $C_1 \in \{5, 8, 10, 12, 15\}$ and the total number of iterations $T \in \{10^3 \times 2^l, l = 1, 2, 3, 4, 5\}$. The batch size is $b = 0.1 \times N$. For PriSMA, we fix the clipping threshold $C_2 = 1$, while search the momentum step size $\gamma \in \{0.001, 0.002, 0.005, 0.008, 0.01, 0.02, 0.05\}$ and the clipping threshold $C_3 \in \{0.005, 0.01, 0.02\}$. After fixing the best parameters, we run each algorithm for 25 trials, reporting the averaged gradient norm and its standard error in Figure 1. By Figure 1, PriSMA consistently outperforms DPSGD-GC in terms of the gradient norm for any given N. As N increases, the gradient norm of PriSMA decreases much faster than that of DPSGD-GC.

Convolutional neural network training. Now we consider training a convolution neural network (CNN) for the image classification task on the CIFAR10 dataset (Krizhevsky et al., 2009). We launch $M = 10$ clients, distributing the CIFAR10 dataset randomly and evenly among them such that each client has $N = 5000$ data samples. The structure of the CNN is the same as the one in (Kairouz et al., 2021b).

The privacy budgets are set to $\epsilon \in \{4, 8, 16, 32\}$ with a fixed $\delta = 10^{-5}$. The total number of iterations is $T = 1000$ and the batch size is $b = 512$. To identify the optimal parameter combinations, we employ a grid search over the following ranges: the step size η for both algorithms in $\{1.0 \times 10^l, 2.0 \times 10^l, 5.0 \times 10^l, l = 0, -1, -2\}$, the clipping threshold C_1 for both algorithms in $\{1, 2, 4, 8, 16\}$. For PriSMA, we set the clipping threshold $C_2 = \min\{4, \frac{C_1}{2}\}$, while search the clipping threshold C_3 in $\{0.05, 0.5, 1.5\}$ and the momentum step size γ in $\{0.1, 0.5, 0.8, 0.9, 0.99\}$. For each setting, we run 12 trials and report the averaged test accuracy and the averaged test loss with their corresponding standard errors. As shown in Table 1, PriSMA outperforms DPSGD-GC in terms of both test accuracy and test loss.

6 CONCLUSIONS

This paper proposes PriSMA, an LDP algorithm, for solving distributed non-convex ERM problems. Given LDP requirement (ϵ, δ), PriSMA achieves a learning error of $\tilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}}{\sqrt{MN}\epsilon}\right)^{\frac{2}{5}}\right)$ without assuming uniformly bounded stochastic gradients. The enabling factor is the innovative combination of clipping and momentum. This learning error bound is better in order than $\tilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}}{\sqrt{MN}\epsilon}\right)^{\frac{1}{3}}\right)$ established for DPSGD-GC.

Our theoretical results demonstrate that the learning error of PriSMA is better than DPSGD-GC when $d < MN^2\epsilon^2$. In our future work, we will explore how to improve the dependence on d for PriSMA to obtain a better utility-privacy trade-off for high-dimensional distributed non-convex ERM problems.

Acknowledgments

Qing Ling (corresponding author) is supported in part by NSF China grant 62373388, Guangdong Basic and Applied Basic Research Foundation grants 2021-

Table 1: Test accuracy and test loss on convolutional neural network training.

privacy budget	PriSMA		DPSGD-GC	
	accuracy	loss	accuracy	loss
$\epsilon = 4$	$\mathbf{55.570} \pm 0.261$	$\mathbf{1.306} \pm 0.007$	54.868 ± 0.261	1.435 ± 0.009
$\epsilon = 8$	$\mathbf{61.203} \pm 0.190$	$\mathbf{1.257} \pm 0.007$	61.083 ± 0.253	1.303 ± 0.013
$\epsilon = 16$	$\mathbf{65.847} \pm 0.180$	$\mathbf{1.147} \pm 0.008$	65.588 ± 0.238	1.260 ± 0.011
$\epsilon = 32$	$\mathbf{68.678} \pm 0.125$	$\mathbf{1.091} \pm 0.007$	68.426 ± 0.201	1.096 ± 0.007

B1515020094 and 2023B1515040025, R&D project of Pazhou Lab (Huangpu) grant 2023K0606, and Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence grant 2023B1212010001.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Conference on Computer and Communications Security*.

Agarwal, N., Suresh, A. T., Xu, F. X., Kumar, S., and McMahan, H. B. (2018). CPSGD: Communication-efficient and differentially-private distributed SGD. In *Advances in Neural Information Processing Systems*.

Arora, R., Bassily, R., González, T., Guzmán, C. A., Menart, M., and Ullah, E. (2023). Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*.

Asi, H., Liu, D., and Tian, K. (2024). Private stochastic convex optimization with heavy tails: Near-optimality from simple reductions. *arXiv preprint arXiv:2406.02789*.

Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Symposium on Foundations of Computer Science*.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Bu, Z., Wang, Y.-X., Zha, S., and Karypis, G. (2023). Automatic clipping: Differentially private deep learning made easier and stronger. In *Advances in Neural Information Processing Systems*.

Cao, X., Zhang, J., Poor, H. V., and Tian, Z. (2020). Differentially private ADMM for regularized consensus optimization. *IEEE Transactions on Automatic Control*, 66:3718–3725.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*.

Das, R., Kale, S., Xu, Z., Zhang, T., and Sanghavi, S. (2023). Beyond uniform lipschitz condition in differentially private optimization. In *International Conference on Machine Learning*.

Dwork, C. (2006). Differential privacy. In *International Colloquium on Automata, Languages, and Programming*.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9:211–407.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*.

Fang, H., Li, X., Fan, C., and Li, P. (2022). Improved convergence of differential private SGD with gradient clipping. In *International Conference on Learning Representations*.

Holland, M. J. (2019). Robust descent using smoothed multiplicative noise. In *International Conference on Artificial Intelligence and Statistics*.

Huang, Z., Hu, R., Guo, Y., Chan-Tin, E., and Gong, Y. (2019). DP-ADMM: ADMM-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:1002–1012.

Jegorova, M., Kaul, C., Mayor, C., O'Neil, A. Q., Weir, A., Murray-Smith, R., and Tsaftaris, S. A. (2023). Survey: Leakage and privacy at inference time. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45:9090–9108.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021a). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14:1–210.

Kairouz, P., McMahan, H. B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. (2021b). Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*.

Kamath, G., Liu, X., and Zhang, H. (2022). Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*.

Koloskova, A., Hendrikx, H., and Stich, S. U. (2023). Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Li, B., Jiang, X., Schmidt, M. N., Alstrøm, T. S., and Stich, S. U. (2024). An improved analysis of per-sample and per-update clipping in federated learning. In *International Conference on Learning Representations*.

Li, Z., Zhao, H., Li, B., and Chi, Y. (2022). SoteriaFL: A unified framework for private federated learning with communication compression. In *Advances in Neural Information Processing Systems*.

Lowy, A., Ghafelebashi, A., and Razaviyayn, M. (2023). Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*.

Lowy, A. and Razaviyayn, M. (2023). Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. In *International Conference on Algorithmic Learning Theory*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017a). Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. (2017b). Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.

Mironov, I. (2017). Rényi differential privacy. In *Computer Security Foundations Symposium*.

Nasr, M., Shokri, R., and Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Symposium on Security and Privacy*.

Nguyen, T., Lai, P., Tran, K., Phan, N., and Thai, M. T. (2023). Active membership inference attack under local differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*.

Noble, M., Bellet, A., and Dieuleveut, A. (2022). Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*.

Rigaki, M. and Garcia, S. (2023). A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56:1–34.

Tran, H. and Cutkosky, A. (2022). Momentum aggregation for private non-convex ERM. In *Advances in Neural Information Processing Systems*.

Wang, D., Xiao, H., Devadas, S., and Xu, J. (2020). On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*.

Wang, L., Jayaraman, B., Evans, D., and Gu, Q. (2023). Efficient privacy-preserving stochastic nonconvex optimization. In *Uncertainty in Artificial Intelligence*.

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. (2019). Subsampled Rényi differential privacy and analytical moments accountant. In *International Conference on Artificial Intelligence and Statistics*.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q. S., and Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469.

Xiao, H., Xiang, Z., Wang, D., and Devadas, S. (2023). A theory to instruct differentially-private learning via clipping bias reduction. In *Symposium on Security and Privacy*.

Xu, J., Zhang, W., and Wang, F. (2021). A(DP)^{2}SGD: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:8036–8047.

Yang, X., Zhang, H., Chen, W., and Liu, T.-Y. (2022). Normalized/Clipped SGD with perturbation for differentially private non-convex optimization. *arXiv preprint arXiv:2206.13033*.

Yi, K., Gazagnadou, N., Richtárik, P., and Lyu, L. (2024). Fedp3: Federated personalized and privacy-friendly network pruning under model heterogeneity.

In *International Conference on Learning Representations.*

Yin, X., Zhu, Y., and Hu, J. (2021). A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys*, 54:1–36.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. (2021). Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298.*

Zhang, J. and Hong, M. (2020). First-order algorithms without lipschitz gradient: A sequential local optimization approach. *arXiv preprint arXiv:2010.03194.*

Zhang, J., Zheng, K., Mou, W., and Wang, L. (2017). Efficient private ERM for smooth objectives. *arXiv preprint arXiv:1703.09947.*

Zhao, P., Wu, J., Liu, Z., Wang, C., Fan, R., and Li, Q. (2024). Differential private stochastic optimization with heavy-tailed data: Towards optimal rates. *arXiv preprint arXiv:2408.09891.*

Zhao, Y., Zhao, J., Yang, M., Wang, T., Wang, N., Lyu, L., Niyato, D., and Lam, K.-Y. (2020). Local differential privacy-based federated learning for internet of things. *IEEE Internet of Things Journal*, 8:8836–8853.

Zhou, X. and Chowdhury, S. R. (2024). On differentially private federated linear contextual bandits. In *International Conference on Learning Representations.*

Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems.*

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendices for "Differential Privacy in Distributed Learning: Beyond Uniformly Bounded Stochastic Gradients"

A Experimental implementation

The experiments are run on an Intel Xeon Silver 4314 processor for non-convex regularized least squares and on an NVIDIA GeForce RTX 3080Ti processor for convolutional neural network training. The former is light-weight in terms of memory and time, while the latter requires 5.4 GB memory and 15 minutes for a single run. Within the code[5], we utilize opacus (Yousefpour et al., 2021) to accelerate the per-sample clipping operation.

B Notations in Appendices

In Appendices, we use $i \sim [M]$ to denote that i is uniformly randomly sampled from the set of $[M]$. We define $\beta^t := \min\left\{\frac{C_2}{\|\bar{v}^t\|}, 1\right\}$, $\lambda_1^t = \frac{C_1}{\|\nabla F(x^t)\|}$, and $\lambda_2^t = \frac{C_2}{\|\nabla F(x^t)\|}$. We define $\mathcal{F}^t := \sigma\left(\bigcup_{k=0}^t \{v_1^k, \ldots, v_M^k\}\right)$, where $\sigma(X)$ is the σ-algebra generated by X. We define

$$\nabla F_{\mathcal{B}^t}^{C_1}(x) := \frac{1}{M}\sum_{i=1}^{M}\frac{1}{b}\sum_{j\in\mathcal{B}_i^t}\text{clip}_{C_1}\left(\nabla f_{ij}(x)\right),\ \forall x, \tag{23}$$

and a Lyapunov function

$$\Phi^t := F(x^t) + \frac{\eta}{\gamma}\left(\frac{1}{\lambda_1^{t-1}} + \frac{LC_2\eta}{C_1\gamma} + 1\right)\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2,\ \forall t \geq 0, \tag{24}$$

where $\bar{v}^t$ and $\nabla F^{C_1}(x^t)$ are defined in (13). In addition, we define $\lambda_1^{-1} := \frac{C_1}{\|\nabla F(x^0)\|}$.

C Proof for the lower bound

Theorem 1 (Lower bound of $\mathcal{A}_{C_1}$ on $\mathcal{P}_\sigma^L$)**.** *For any clipping threshold $C_1 > 0$ and any algorithm $\mathcal{A} \in \mathcal{A}_{C_1}$, there exists a problem $\mathcal{P} \in \mathcal{P}_\sigma^L$, for which the solution yielded by $\mathcal{A}$ on $\mathcal{P}$, denoted as $x_{\mathcal{A},\mathcal{P}}$, satisfies*

$$\mathbb{E}\|\nabla F(x_{\mathcal{A},\mathcal{P}})\| \geq \Omega\left(\min\left\{\sigma, \frac{\sigma^2}{C_1}\right\}\right). \tag{25}$$

Proof. We use $x^{(k)}$ to denote the k-th dimension of vector $x \in \mathbb{R}^d$.

First, we construct two loss functions for later use. For any $a \in \mathbb{R}$, $L > 0$ and $\hat{C} > 0$, define

$$g_a(x) := \frac{L}{2}(x^{(1)} - a)^2, \quad g_a^{\hat{C}}(x) := \begin{cases} \frac{L}{2}(x^{(1)} - a)^2, & |x^{(1)} - a| \leq \frac{\hat{C}}{L}, \\ \hat{C}|x^{(1)} - a| - \frac{\hat{C}^2}{2L}, & |x^{(1)} - a| > \frac{\hat{C}}{L}, \end{cases} \quad \forall x \in \mathbb{R}^d. \tag{26}$$

Then the gradients of $g_a(x)$ and $g_a^{\hat{C}}(x)$ are

$$(\nabla g_a(x))^{(1)} = L(x - a), \quad \left(\nabla g_a^{\hat{C}}(x)\right)^{(1)} = \begin{cases} L(x^{(1)} - a), & |x^{(1)} - a| \leq \frac{\hat{C}}{L}, \\ \hat{C}\cdot\text{sign}(x^{(1)} - a), & |x^{(1)} - a| > \frac{\hat{C}}{L}. \end{cases} \tag{27}$$

$$(\nabla g_a(x))^{(k)} = \left(\nabla g_a^{\hat{C}}(x)\right)^{(k)} = 0, \quad \forall k \in [d] \wedge k \neq 1. \tag{28}$$

[5] https://github.com/Etherial-h/PriSMA

The loss functions $g_a(x)$ and $g_a^{\hat{C}}(x)$ are L-Lipschitz smooth and have the same clipped gradient given any clipping threshold $C_1 \leq \hat{C}$.

Next, we construct two problems, $\mathcal{P}_0$ and $\mathcal{P}_1$, to establish the lower bound. For each client i in $\mathcal{P}_0$, a fraction p of all loss functions are $g_a(x)$, and the remaining ones are $g_0(x)$. For each client i in $\mathcal{P}_1$, a fraction p of all loss functions in $\mathcal{P}_1$ are $g_a^{\hat{C}}(x)$, and the remaining ones are $g_0^{\hat{C}}(x)$, where $a = \frac{2\hat{C}}{L}$, $\hat{C} = \max\{C_1, \sigma\}$ and $p = \frac{\sigma^2}{4\hat{C}^2} \leq \frac{1}{4}$. We use $F_{\mathcal{P}_0}$ and $F_{\mathcal{P}_1}$ to denote the averaged loss functions in $\mathcal{P}_0$ and $\mathcal{P}_1$, respectively. For $F_{\mathcal{P}_0}$ and $F_{\mathcal{P}_1}$, we have

$$\|\nabla F_{\mathcal{P}_0}(x)\| = L|x^{(1)} - pa| = L\left|x^{(1)} - \frac{\sigma^2}{2L\hat{C}}\right|, \tag{29}$$

$$\begin{aligned}\|\nabla F_{\mathcal{P}_1}(x)\| &= \begin{cases} \hat{C}, & x^{(1)} < -\frac{\hat{C}}{L} \ \vee \ x^{(1)} \geq \frac{3\hat{C}}{L}, \\ \left|(1-p)\, Lx^{(1)} - p\hat{C}\right|, & |x^{(1)}| \leq \frac{\hat{C}}{L}, \\ \left|pL\left(x^{(1)} - a\right) + (1-p)\hat{C}\right|, & |x^{(1)} - a| \leq \frac{\hat{C}}{L}, \end{cases} \\ &= \begin{cases} \hat{C}, & x^{(1)} < -\frac{\hat{C}}{L} \ \vee \ x^{(1)} \geq \frac{3\hat{C}}{L}, \\ \left|\left(1 - \frac{\sigma^2}{4\hat{C}^2}\right) Lx^{(1)} - \frac{\sigma^2}{4\hat{C}}\right|, & |x^{(1)}| \leq \frac{\hat{C}}{L}, \\ \frac{\sigma^2}{4\hat{C}^2} Lx^{(1)} + \hat{C} - \frac{3\sigma^2}{4\hat{C}}, & |x^{(1)} - a| \leq \frac{\hat{C}}{L}. \end{cases}\end{aligned} \tag{30}$$

Since the loss functions in $\mathcal{P}_0$ and $\mathcal{P}_1$ have the same clipped gradients, $\mathcal{A}$ can not distinguish $\mathcal{P}_0$ and $\mathcal{P}_1$, meaning that the outputs of $\mathcal{A}$ for solving $\mathcal{P}_0$ and $\mathcal{P}_1$ are the same. For any output $x_{\mathcal{A}} \in \mathbb{R}$, we have

$$\max\{\|\nabla F_{\mathcal{P}_0}(x_{\mathcal{A}})\|, \|\nabla F_{\mathcal{P}_1}(x_{\mathcal{A}})\|\} \geq \frac{\sigma^2}{14\hat{C}}. \tag{31}$$

The above inequality is due to $\|\nabla F_{\mathcal{P}_0}(x_{\mathcal{A}})\| < \frac{\sigma^2}{14\hat{C}}$, which implies that

$$x_{\mathcal{A}} \in \left(\frac{3\sigma^2}{7L\hat{C}}, \frac{4\sigma^2}{7L\hat{C}}\right) := S_0,$$

and $\|\nabla F_{\mathcal{P}_1}(x_{\mathcal{A}})\| < \frac{\sigma^2}{14\hat{C}}$, which implies that

$$x_{\mathcal{A}} \in \left(\frac{5\sigma^2}{28(1-p)L\hat{C}}, \frac{9\sigma^2}{28(1-p)L\hat{C}}\right) \subset \left(\frac{5\sigma^2}{28L\hat{C}}, \frac{3\sigma^2}{7L\hat{C}}\right) := S_1.$$

Observe that $S_0 \cap S_1 = \varnothing$.

Thus, it holds that

$$\begin{aligned} &\max\left\{\mathbb{E}[\|\nabla F_{\mathcal{P}_0}(x_{\mathcal{A}})\|], \mathbb{E}[\|\nabla F_{\mathcal{P}_1}(x_{\mathcal{A}})\|]\right\} \\ \geq\ &\frac{1}{2}\mathbb{E}[\|\nabla F_{\mathcal{P}_0}(x_{\mathcal{A}})\| + \|\nabla F_{\mathcal{P}_1}(x_{\mathcal{A}})\|] \\ \geq\ &\frac{1}{2}\mathbb{E}[\max\{\|\nabla F_{\mathcal{P}_0}(x_{\mathcal{A}})\|, \|\nabla F_{\mathcal{P}_1}(x_{\mathcal{A}})\|\}] \\ \geq\ &\frac{\sigma^2}{28\hat{C}} = \Omega\left(\min\left\{\sigma, \frac{\sigma^2}{C_1}\right\}\right). \end{aligned} \tag{32}$$

This completes the proof of Theorem 1.

Our established lower bound is similar to that in Koloskova et al. (2023), but significantly different in several aspects. First, the work of Koloskova et al. (2023) only investigates DPSGD-GC without DP noise (called as SGD-GC), while our work considers an algorithm class $\mathcal{A}_{C_1}$ that contains SGD-GC. Second, the work of Koloskova et al. (2023) only considers the single-machine and one-dimensional special case with $M = 1$ and $d = 1$, while our work allows $M \geq 1$ and $d \geq 1$. The technical tools are also different. We construct two problems in $\mathcal{P}_\sigma^L$ and show that any algorithm in $\mathcal{A}_{C_1}$ must perform poorly on at least one of them. The work of Koloskova et al. (2023) constructs one problem $\mathcal{P}_0 \in \mathcal{P}_\sigma^L$, on which a fixed point of SGD-GC performs poorly.

D Auxiliary lemmas

Before proving Lemma 1 and Theorems 2–4, we give several auxiliary lemmas as follows.

Lemma D.1. *Under Assumptions 1 and 2, for any $\gamma \in (0,1)$, the sequence $\{v_i^t\}_t$ generated by PriSMA satisfies*

$$\begin{aligned}\mathbb{E}[\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t] \leq\ & (1-\gamma)^2\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + 2\gamma^2\frac{\sigma^2}{Mb} \\ & + 2(1-\gamma)^2\frac{L^2}{Mb}\|x^{t+1} - x^t\|^2 + \frac{d\sigma_1^2}{M},\end{aligned} \tag{33}$$

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\lambda_1^t}\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t\right] \leq\ & \left(\frac{1}{\lambda_1^{t-1}} + \frac{LC_2\eta}{C_1} - \frac{\gamma}{\lambda_1^t}\right)\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 \\ & + \frac{2}{\lambda_1^t}\left(\gamma^2\frac{\sigma^2}{Mb} + (1-\gamma)^2\frac{L^2}{Mb}\|x^{t+1} - x^t\|^2 + \frac{d\sigma_1^2}{2M}\right),\end{aligned} \tag{34}$$

where $\frac{1}{\lambda_1^{t-1}} + \frac{LC_2\eta}{C_1} - \frac{\gamma}{\lambda_1^t} > 0$.

Proof. First, according to the definition of $\nabla F^{C_1}_{\mathcal{B}^{t+1}}$ in (23), we have

$$\begin{aligned}&\mathbb{E}[\|\nabla F^{C_1}(x^t) - \nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^t) - \nabla F^{C_1}(x^{t+1}) + \nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^{t+1})\|^2|\mathcal{F}^t] \\ \overset{(a)}{=}\ & \frac{1}{Mb}\mathbb{E}_{i\sim[M],j\sim[N]}\left[\|\nabla F^{C_1}(x^t) - \nabla F^{C_1}(x^{t+1}) + \mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^{t+1})\right) - \mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^t)\right)\|^2|\mathcal{F}^t\right] \\ \overset{(b)}{\leq}\ & \frac{1}{Mb}\mathbb{E}_{i\sim[M],j\sim[N]}\|\mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^{t+1})\right) - \mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^t)\right)\|^2 \\ \overset{(c)}{\leq}\ & \frac{L^2}{Mb}\|x^{t+1} - x^t\|^2,\end{aligned} \tag{35}$$

where (a) and (b) are due to the fact that $\mathcal{B}_i^{t+1}$ is uniformly and independently sampled from $\mathcal{D}_i$ for all $i \in [N]$, and (c) is due to the L-smoothness of f_{ij}. Thanks to the contraction property of the clipping operator, we have

$$\begin{aligned}&\mathbb{E}[\|\nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^{t+1}) - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t] \\ =\ & \frac{1}{Mb}\mathbb{E}_{i,j}\|\mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^{t+1})\right) - \nabla F^{C_1}(x^{t+1})\|^2 \\ =\ & \frac{1}{Mb}\mathbb{E}_{i,j}\|\mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^{t+1})\right) - \mathrm{clip}_{C_1}\left(\nabla F(x^{t+1})\right) - (\nabla F^{C_1}(x^{t+1}) - \mathrm{clip}_{C_1}\left(\nabla F(x^{t+1})\right))\|^2 \\ \leq\ & \frac{1}{Mb}\mathbb{E}_{i,j}\|\mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^{t+1})\right) - \mathrm{clip}_{C_1}\left(\nabla F(x^{t+1})\right)\|^2 \\ \leq\ & \frac{1}{Mb}\mathbb{E}_{i,j}\|f_{ij}(x^{t+1}) - \nabla F(x^{t+1})\|^2 \leq \frac{\sigma^2}{Mb}.\end{aligned} \tag{36}$$

According to the update of v_i^t in PriSMA, we have

$$\begin{aligned}
&\mathbb{E}[\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t] \\
\overset{(a)}{=}& \mathbb{E}[\|\nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^{t+1}) + (1-\gamma)(\bar{v}^t - \nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^t)) - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t] + \frac{d\sigma_1^2}{M} \\
=& \mathbb{E}[\|(1-\gamma)\left(\bar{v}^t - \nabla F^{C_1}(x^t) + \nabla F^{C_1}(x^t) - \nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^t)\right) \\
&+ (1-\gamma+\gamma)\left(\nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^{t+1}) - \nabla F^{C_1}(x^{t+1})\right)\|^2|\mathcal{F}^t] + \frac{d\sigma_1^2}{M} \\
\overset{(b)}{\le}& (1-\gamma)^2\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + 2\gamma^2\mathbb{E}[\|\nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^{t+1}) - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t] \\
&+ 2(1-\gamma)^2\mathbb{E}[\|\nabla F^{C_1}(x^t) - \nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^t) - \nabla F^{C_1}(x^{t+1}) + \nabla F^{C_1}_{\mathcal{B}^{t+1}}(x^{t+1})\|^2|\mathcal{F}^t] + \frac{d\sigma_1^2}{M} \\
\overset{(c)}{\le}& (1-\gamma)^2\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + 2\gamma^2\frac{\sigma^2}{Mb} + 2(1-\gamma)^2\frac{L^2}{Mb}\|x^{t+1} - x^t\|^2 + \frac{d\sigma_1^2}{M},
\end{aligned} \tag{37}$$

where (a) is due to the fact that ξ_i^t is independently sampled from $\mathcal{N}(0, \sigma_1^2\mathbb{I})$, (b) is due to the fact that $\mathcal{B}^{t+1}$ is uniformly, randomly and independently sampled from $\mathcal{D}$ and Young's inequality, while (c) comes from (35) and (36). This completes the proof of (33).

To prove (34), since $F(\cdot)$ is L-Lipschitz smooth, for any $t \ge 1$, we have

$$\begin{aligned}
\frac{1}{\lambda_1^t} = \frac{1}{C_1}\|\nabla F(x^t)\| &\le \frac{1}{C_1}\|\nabla F(x^t) - \nabla F(x^{t-1})\| + \frac{1}{C_1}\|\nabla F(x^{t-1})\| \\
&\le \frac{L}{C_1}\|x^t - x^{t-1}\| + \frac{1}{\lambda_1^{t-1}} \\
&\le \frac{LC_2\eta}{C_1} + \frac{1}{\lambda_1^{t-1}}.
\end{aligned} \tag{38}$$

For $t = 0$, we have $\frac{1}{\lambda_1^0} = \frac{1}{\lambda_1^{-1}}$. Thus, (38) holds for any $t \ge 0$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\lambda_1^t}\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t\right] \\
\overset{(a)}{\le}& \frac{1-\gamma}{\lambda_1^t}\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{2}{\lambda_1^t}\left(\gamma^2\frac{\sigma^2}{Mb} + (1-\gamma)^2\frac{L^2}{Mb}\|x^{t+1} - x^t\|^2 + \frac{d\sigma_1^2}{2M}\right) \\
\overset{(b)}{\le}& \left(\frac{1}{\lambda_1^{t-1}} + \frac{LC_2\eta}{C_1} - \frac{\gamma}{\lambda_1^t}\right)\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 \\
&+ \frac{2}{\lambda_1^t}\left(\gamma^2\frac{\sigma^2}{Mb} + (1-\gamma)^2\frac{L^2}{Mb}\|x^{t+1} - x^t\|^2 + \frac{d\sigma_1^2}{2M}\right),
\end{aligned} \tag{39}$$

where (a) is from (37) and (b) is from (38).

By (38) and $\gamma < 1$, we know that $\frac{1}{\lambda_1^{t-1}} + \frac{LC_2\eta}{C_1} - \frac{\gamma}{\lambda_1^t} > 0$. This completes the proof of (34). □

Lemma D.2. *Consider a set of vectors $\{u_l \in \mathbb{R}^d,\ l \in [n]\}$ and a constant $C > 0$. Define $\bar{u} = \frac{1}{n}\sum_{l=1}^n u_l$. If $\frac{1}{n}\sum_{l=1}^n \|u_l - \bar{u}\|^2 \le \sigma^2$ and $\|\bar{u}\| \le \frac{C}{2}$, then we have*

$$\left\|\frac{1}{n}\sum_{l=1}^n \text{clip}_C(u_l) - \bar{u}\right\| \le \frac{8\sigma^4}{C^2} + \frac{32\sigma^4\|\bar{u}\|^2}{C^4}. \tag{40}$$

Proof. This proof is based on the one in Koloskova et al. (2023). First, we use $\mathbb{E}_l$ to denote the expectation on the distribution that l is uniformly and randomly sampled from $[n]$ and define $\delta_l := \text{Id}\{\|u_l\| > C\}$. We have

$$\mathbb{E}[\delta_l] = \Pr[\|u_l\| > C] \le \Pr\left[\|u_l - \bar{u}\| > \frac{C}{2}\right] \le \frac{4\sigma^2}{C^2}, \tag{41}$$

where the last inequality is from Markov's inequality and $\mathbb{E}_l[\|u_l - \bar{u}\|^2] \leq \sigma^2$. Then we have

$$\begin{aligned}
\|\mathbb{E}_l[\mathrm{clip}_C(u_l)] - \bar{u}\|^2 &= \left\| \mathbb{E}_l \left[\left(1 - \frac{C}{\|u_l\|} \right) u_l \delta_l \right] \right\|^2 \\
&= (\mathbb{E}\delta_l)^2 \left\| E_l \left[\left(1 - \frac{C}{\|u_l\|} \right) u_l | \delta_l = 1 \right] \right\|^2 \\
&\leq (\mathbb{E}\delta_l)^2 \mathbb{E}[\|u_l\|^2 | \delta_l = 1] \\
&\leq 2(\mathbb{E}\delta_l)^2 \left(\mathbb{E}[\|u_l - \bar{u}\|^2 | \delta_l = 1] + \mathbb{E}[\|\bar{u}\|^2 | \delta_l = 1] \right) \\
&= 2(\mathbb{E}\delta_l)\mathbb{E}[\delta_l \|u_l - \bar{u}\|^2] + 2(\mathbb{E}\delta_l)^2 \|\bar{u}\|^2 \\
&\leq 2(\mathbb{E}\delta_l)\mathbb{E}[\|u_l - \bar{u}\|^2] + 2(\mathbb{E}\delta_l)^2 \|\bar{u}\|^2 \\
&\leq \frac{8\sigma^4}{C^2} + \frac{32\sigma^4 \|\bar{u}\|^2}{C^4},
\end{aligned} \tag{42}$$

which completes the proof of Lemma D.2. □

Lemma D.2 can be used to bound the difference between $\nabla F^{C_1}(x)$ and the full gradient $\nabla F(x)$. According to Lemma D.2, if Assumption 2 holds and $\|\nabla F(x)\| \leq \frac{C_1}{2}$ for some $x \in \mathbb{R}^d$, then it holds that

$$\|\nabla F(x) - \nabla F^{C_1}(x)\| \leq \frac{8\sigma^4}{C_1^2} + \frac{32\sigma^4 \|\nabla F(x)\|^2}{C_1^4}.$$

Lemma D.3. *Suppose Assumptions 1 and 2 hold. Consider $\{x^t, v_i^t, i \in [M]\}_t$ generated by PriSMA with $C_1 \geq 2C_2$ and $C_1 \geq 16\sigma$.*

(i) *If $\|\nabla F(x^t)\| \geq C_1$, it holds that*

$$F(x^{t+1}) \leq F(x^t) - \frac{\eta C_2}{4}\|\nabla F(x^t)\| + \frac{\eta}{\lambda_1}\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(\frac{1}{\lambda_1} - L\eta). \tag{43}$$

(ii) *If $C_1 > \|\nabla F(x^t)\| \geq \frac{C_1}{2}$, it holds that*

$$F(x^{t+1}) \leq F(x^t) - \frac{\eta C_2}{4}\|\nabla F(x^t)\| + \eta\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(1 - L\eta). \tag{44}$$

(iii) *If $\frac{C_1}{2} > \|\nabla F(x^t)\| \geq C_2$, it holds that*

$$F(x^{t+1}) \leq F(x^t) - \frac{\eta C_2}{4}\|\nabla F(x^t)\| + \frac{8\eta\sigma^4}{C_1^2} + \eta\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(1 - L\eta). \tag{45}$$

(iv) *If $C_2 > \|\nabla F(x^t)\|$, it holds that*

$$F(x^{t+1}) \leq F(x^t) - \frac{\eta}{4}\|\nabla F(x^t)\|^2 + \frac{8\eta\sigma^4}{C_1^2} + \eta\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(1 - L\eta). \tag{46}$$

Proof. Define $\beta^t := \min\{\frac{C_2}{\|\bar{v}^t\|}, 1\}$. According to the update of x^t, we have

$$x^{t+1} = x^t - \eta \cdot \mathrm{clip}_{C_2}\left(\bar{v}^t\right) = x^t - \eta\beta^t\bar{v}^t. \tag{47}$$

The L-Lipschitz smoothness of $F(\cdot)$ implies that

$$\begin{aligned}
F(x^{t+1}) &\leq F(x^t) + \langle \nabla F(x^t), x^{t+1} - x^t \rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2 \\
&= F(x^t) - \eta\langle \nabla F(x^t), \beta^t\bar{v}^t \rangle + \frac{L\eta^2(\beta^t)^2}{2}\|\bar{v}^t\|^2.
\end{aligned} \tag{48}$$

Next, we discuss $A^t = \langle \nabla F(x^t),\, \beta^t \bar{v}^t \rangle$ by cases. Before doing so, we observe that

$$\begin{aligned}
&\|\mathrm{clip}_{C_1}\left(\nabla F(x^t)\right) - \nabla F^{C_1}(x^t)\|^2 \\
=\ & \|\mathrm{clip}_{C_1}\left(\nabla F(x^t)\right) - \frac{1}{MN}\sum_{ij} \mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^t)\right)\|^2 \\
\le\ & \mathbb{E}_{i\sim[M],j\sim[N]}\|\mathrm{clip}_{C_1}\left(\nabla F(x^t)\right) - \mathrm{clip}_{C_1}\left(\nabla f_{ij}(x^t)\right)\|^2 \\
\overset{(a)}{\le}\ & \mathbb{E}_{i\sim[M],j\sim[N]}\|\nabla F(x^t) - \nabla f_{ij}(x^t)\|^2 \\
\le\ & \sigma^2,
\end{aligned} \tag{49}$$

where (a) is from the contraction property of the clipping operator.

(B1) If $\|\nabla F(x^t)\| \ge C_1$ and $\|\bar{v}^t\| \ge C_1$: we have $\beta^t = \frac{C_2}{\|\bar{v}^t\|}$ and

$$\begin{aligned}
A^t =\ & \frac{1}{\lambda_2^t}\langle \lambda_2^t \nabla F(x^t),\, \beta^t \bar{v}^t\rangle \\
=\ & \frac{\lambda_2^t}{2}\|\nabla F(x^t)\|^2 - \frac{1}{2\lambda_2^t}\|\lambda_2^t \nabla F(x^t) - \beta^t \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
\overset{(a)}{=}\ & \frac{C_2}{2}\|\nabla F(x^t)\| - \frac{1}{2\lambda_2^t}\|\lambda_2^t \nabla F(x^t) - \beta^t \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
\overset{(b)}{=}\ & \frac{C_2}{2}\|\nabla F(x^t)\| - \frac{C_2^2}{2\lambda_2^t C_1^2}\|\lambda_1^t \nabla F(x^t) - \mathrm{clip}_{C_1}\left(\bar{v}^t\right)\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
\ge\ & \frac{C_2}{2}\|\nabla F(x^t)\| - \frac{C_2^2}{\lambda_2^t C_1^2}\|\lambda_1^t \nabla F(x^t) - \nabla F^{C_1}(x^t)\|^2 \\
& - \frac{C_2^2}{\lambda_2^t C_1^2}\|\nabla F^{C_1}(x^t) - \mathrm{clip}_{C_1}\left(\bar{v}^t\right)\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
\overset{(c)}{\ge}\ & \frac{C_2}{2}\|\nabla F(x^t)\| - \frac{C_2^2}{\lambda_2^t C_1^2}\sigma^2 - \frac{C_2^2}{\lambda_2^t C_1^2}\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
\overset{(d)}{\ge}\ & \frac{C_2}{2}\|\nabla F(x^t)\| - \frac{C_2\sigma^2}{C_1^2}\|\nabla F(x^t)\| - \frac{C_2}{\lambda_1^t C_1}\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
\overset{(e)}{\ge}\ & \frac{C_2}{4}\|\nabla F(x^t)\| - \frac{C_2}{\lambda_1^t C_1}\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2,
\end{aligned} \tag{50}$$

where (a) is from $\lambda_2^t = \frac{C_2}{\|\nabla F(x^t)\|}$, (b) is from $\frac{C_1}{C_2} = \frac{\lambda_1}{\lambda_2}$, (c) is from $\|\nabla F^{C_1}(x^t) - \mathrm{clip}_{C_1}\left(\bar{v}^t\right)\| \le \|\nabla F^{C_1}(x^t) - \bar{v}^t\|$, (d) is from $\lambda_2^t = \frac{C_2}{\|\nabla F(x^t)\|}$, and (e) is from $C_1 \ge 16\sigma > 2\sigma$.

(B2) If $\|\nabla F(x^t)\| \ge C_1$ and $\|\bar{v}^t\| < C_1$: we have $\beta^t C_1 \ge C_2$, $\lambda_1^t \nabla F(x^t) = \mathrm{clip}_{C_1}\left(\nabla F(x^t)\right)$ and

$$\begin{aligned}
A^t =\ & \frac{\beta^t}{\lambda_1^t}\langle \lambda_1^t \nabla F(x^t),\, \bar{v}^t\rangle \\
=\ & \frac{\beta^t \lambda_1}{2}\|\nabla F(x^t)\|^2 - \frac{\beta^t}{2\lambda_1^t}\|\lambda_1^t \nabla F(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2\lambda_1^t}\|\bar{v}^t\|^2 \\
=\ & \frac{\beta^t C_1}{2}\|\nabla F(x^t)\| - \frac{\beta^t}{2\lambda_1^t}\|\lambda_1^t \nabla F(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2\lambda_1^t}\|\bar{v}^t\|^2 \\
\ge\ & \frac{\beta^t C_1}{2}\|\nabla F(x^t)\| - \frac{\beta^t}{\lambda_1^t}\|\lambda_1^t \nabla F(x^t) - \nabla F^{C_1}(x^t)\|^2 \\
& - \frac{\beta^t}{\lambda_1^t}\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2\lambda_1^t}\|\bar{v}^t\|^2 \\
\overset{(a)}{\ge}\ & \frac{\beta^t C_1}{2}\|\nabla F(x^t)\| - \frac{\beta^t \sigma^2}{C_1}\|\nabla F(x^t)\| - \frac{1}{\lambda_1^t}\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{(\beta^t)^2}{2\lambda_1^t}\|\bar{v}^t\|^2 \\
\overset{(b)}{\ge}\ & \frac{C_2}{4}\|\nabla F(x^t)\| - \frac{1}{\lambda_1^t}\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{(\beta^t)^2}{2\lambda_1^t}\|\bar{v}^t\|^2,
\end{aligned} \tag{51}$$

where (a) is from (49), $\lambda_1^t = \frac{C_1}{\|\nabla F(x^t)\|}$ and $\beta^t \leq 1$, and (b) is from $C_1 \geq 16\sigma > 2\sigma$ and $\beta^t C_1 \geq C_2$.

(B3) If $C_1 \geq \|\nabla F(x^t)\| \geq \frac{C_1}{2}$ and $\|\bar{v}^t\| \leq \|\nabla F(x^t)\|$: we have

$$\begin{aligned}
A^t &= \beta^t \langle \nabla F(x^t),\, \bar{v}^t \rangle \\
&= \frac{\beta^t}{2}\|\nabla F(x^t)\|^2 - \frac{\beta^t}{2}\|\nabla F(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2 \\
&\geq \frac{\beta^t}{2}\|\nabla F(x^t)\|^2 - \beta^t\|\nabla F(x^t) - \nabla F^{C_1}(x^t)\|^2 - \beta^t\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2 \\
&\overset{(a)}{\geq} \frac{\beta^t}{2}\|\nabla F(x^t)\|^2 - \beta^t\sigma^2\frac{4\|\nabla F(x^t)\|^2}{C_1^2} - \beta^t\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2 \\
&\overset{(b)}{\geq} \frac{\beta^t}{4}\|\nabla F(x^t)\|^2 - \beta^t\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2 \\
&\overset{(c)}{\geq} \frac{C_2}{4}\|\nabla F(x^t)\| - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2, \qquad (52)
\end{aligned}$$

where (a) is from (49) and $\|\nabla F(x^t)\| \geq \frac{C_1}{2}$, (b) is from $C_1 \geq 16\sigma > 4\sigma$, while (c) is from $\|\nabla F(x^t)\| \geq \frac{C_1}{2} \geq C_2$ and $\beta^t = \min\{1, C_2/\|\bar{v}^t\|\} \geq \min\{1, C_2/\|\nabla F(x^t)\|\} = C_2/\|\nabla F(x^t)\|$.

(B4) If $C_1 \geq \|\nabla F(x^t)\| \geq \frac{C_1}{2}$ and $\|\bar{v}^t\| \geq \|\nabla F(x^t)\|$: we have

$$\begin{aligned}
A^t &\overset{(a)}{=} \lambda_2^t \langle \nabla F(x^t),\, \mathrm{clip}_{\|\nabla F(x^t)\|}\left(\bar{v}^t\right) \rangle \\
&= \frac{\lambda_2^t}{2}\|\nabla F(x^t)\|^2 - \frac{\lambda_2^t}{2}\|\nabla F(x^t) - \mathrm{clip}_{\|\nabla F(x^t)\|}\left(\bar{v}^t\right)\|^2 + \frac{\lambda_2^t}{2}\|\mathrm{clip}_{\|\nabla F(x^t)\|}\left(\bar{v}^t\right)\|^2 \\
&\overset{(b)}{\geq} \frac{\lambda_2^t}{2}\|\nabla F(x^t)\|^2 - \frac{\lambda_2^t}{2}\|\nabla F(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
&\geq \frac{\lambda_2^t}{2}\|\nabla F(x^t)\|^2 - \lambda_2^t\|\nabla F(x^t) - \nabla F^{C_1}(x^t)\|^2 - \lambda_2^t\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \qquad (53) \\
&\overset{(c)}{\geq} \frac{\lambda_2^t}{2}\|\nabla F(x^t)\|^2 - \lambda_2^t\sigma^2\frac{4\|\nabla F(x^t)\|^2}{C_1^2} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2 \\
&\overset{(d)}{\geq} \frac{C_2}{4}\|\nabla F(x^t)\| - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_2^t}\|\bar{v}^t\|^2, \qquad (54)
\end{aligned}$$

where (a) and (b) are from $\beta^t\bar{v}^t = \frac{C_2}{\|\bar{v}^t\|}\bar{v}^t = \frac{C_2}{\|\nabla F(x^t)\|}\mathrm{clip}_{\|\nabla F(x^t)\|}\left(\bar{v}^t\right) = \lambda_2^t \mathrm{clip}_{\|\nabla F(x^t)\|}\left(\bar{v}^t\right)$, (c) is from (49) and $\|\nabla F(x^t)\| \geq \frac{C_1}{2} \geq C_2$, and (d) is from $C_1 \geq 16\sigma > 4\sigma$ and $\lambda_2 = \frac{C_2}{\|\nabla F(x^t)\|}$.

Before discussing the cases with $\frac{C_1}{2} > \|\nabla F(x^t)\|$, we show that when $\frac{C_1}{2} > \|\nabla F(x^t)\|$, according to Lemma D.2, it holds that

$$\|\nabla F(x^t) - \nabla F^{C_1}(x^t)\| \leq \frac{8\sigma^4}{C_1^2} + \frac{32\sigma^4\|\nabla F(x^t)\|^2}{C_1^4}. \qquad (55)$$

(B5) If $\frac{C_1}{2} > \|\nabla F(x^t)\| \geq C_2$ and $\|\bar{v}^t\| \leq \|\nabla F(x^t)\|$: similar to $\mathcal{J}_{10}$, we have

$$\begin{aligned}
A^t &= \beta^t \langle \nabla F(x^t),\, \bar{v}^t \rangle \\
&= \frac{\beta^t}{2}\|\nabla F(x^t)\|^2 - \frac{\beta^t}{2}\|\nabla F(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2 \\
&\geq \frac{\beta^t}{2}\|\nabla F(x^t)\|^2 - \beta^t\|\nabla F(x^t) - \nabla F^{C_1}(x^t)\|^2 - \beta^t\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2 \\
&\overset{(a)}{\geq} \frac{\beta^t}{2}\|\nabla F(x^t)\|^2 - \beta^t\left(\frac{8\sigma^2}{C_1^2} + \frac{32\sigma^4\|\nabla F(x^t)\|^2}{C_1^4}\right) - \beta^t\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{\beta^t}{2}\|\bar{v}^t\|^2 \\
&\overset{(b)}{\geq} \frac{\beta^t}{4}\|\nabla F(x^t)\|^2 - \frac{8\sigma^4}{C_1^2} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2, \\
&\overset{(c)}{\geq} \frac{C_2}{4}\|\nabla F(x^t)\| - \frac{8\sigma^4}{C_1^2} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2, \qquad (56)
\end{aligned}$$

where (a) is from (55), (b) is from $C_1 \geq 16\sigma$ and $\beta^t \leq 1$, and (c) is from $\beta^t = \min\{1, C_2/\|\bar{v}^t\|\} \geq \min\{1, C_2/\|\nabla F(x^t)\|\} = C_2/\|\nabla F(x^t)\|$.

(B6) If $\frac{C_1}{2} > \|\nabla F(x^t)\| \geq C_2$ and $\|\bar{v}^t\| \geq \|\nabla F(x^t)\|$: similar to $\mathcal{J}_{11}$, we have

$$\begin{aligned} A^t &\overset{(a)}{\geq} \frac{\lambda_2^t}{2}\|\nabla F(x^t)\|^2 - \lambda_2^t\|\nabla F(x^t) - \nabla F^{C_1}(x^t)\|^2 - \lambda_2^t\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_2}\|\bar{v}^t\|^2 \\ &\overset{(b)}{\geq} \frac{\lambda_2^t}{2}\|\nabla F(x^t)\|^2 - \lambda_2^t\left(\frac{8\sigma^2}{C_1^2} + \frac{32\sigma^4\|\nabla F(x^t)\|^2}{C_1^4}\right) - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2 \\ &\overset{(c)}{\geq} \frac{C_2}{4}\|\nabla F(x^t)\| - \frac{8\sigma^4}{C_1^2} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2, \end{aligned} \tag{57}$$

where (a) is obtained by following the proof of (53), (b) is from (55) and $\|\nabla F(x^t)\| \geq C_2$, and (c) is from $C_1 \geq 16\sigma$.

(B7) If $C_2 > \|\nabla F(x^t)\|$: similar to above, we have

$$\begin{aligned} A^t &= \langle \nabla F(x^t),\, \beta^t\bar{v}^t\rangle \\ &\geq \frac{1}{2}\|\nabla F(x^t)\|^2 - \frac{1}{2}\|\nabla F(x^t) - \beta^t\bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2 \\ &\overset{(a)}{\geq} \frac{1}{2}\|\nabla F(x^t)\|^2 - \frac{1}{2}\|\nabla F(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2 \\ &\geq \frac{1}{2}\|\nabla F(x^t)\|^2 - \|\nabla F(x^t) - \nabla F^{C_1}(x^t)\|^2 - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2 \\ &\overset{(b)}{\geq} \frac{1}{2}\|\nabla F(x^t)\|^2 - \frac{8\sigma^4}{C_1^2} - \frac{32\sigma^4\|\nabla F(x^t)\|^2}{C_1^4} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2 \\ &\overset{(c)}{\geq} \frac{1}{4}\|\nabla F(x^t)\|^2 - \frac{8\sigma^4}{C_1^2} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2, \end{aligned} \tag{58}$$

where (a) is from the contraction property of the clipping operator, (b) is from (55), and (c) is from $C_1 \geq 16\sigma$.

Next, we sum up the above results.

(i) If $\|\nabla F(x^t)\| \geq C_1$, this case can be separated into two cases (B1) and (B2) above, then from (50) and (51), we have

$$A^t \geq \frac{C_2}{4}\|\nabla F(x^t)\| - \frac{1}{\lambda_1^t}\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2\lambda_1^t}\|\bar{v}^t\|^2. \tag{59}$$

Then with (48), we have

$$F(x^{t+1}) \leq F(x^t) - \frac{\eta C_2}{4}\|\nabla F(x^t)\| + \frac{\eta}{\lambda_1^t}\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(\frac{1}{\lambda_1^t} - L\eta). \tag{60}$$

(ii) If $C_1 > \|\nabla F(x^t)\| \geq \frac{C_1}{2}$, this case can be separated into two cases (B3) and (B4) above, then from (52) and (54), we have

$$A^t \geq \frac{C_2}{4}\|\nabla F(x^t)\| - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2. \tag{61}$$

Then with (48), we have

$$F(x^{t+1}) \leq F(x^t) - \frac{\eta C_2}{4}\|\nabla F(x^t)\| + \eta\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(1 - L\eta). \tag{62}$$

(iii) If $\frac{C_1}{2} > \|\nabla F(x^t)\| \geq C_2$, this case can be separated into two cases (B5) and (B6) above, then from (56) and (57), we have

$$A^t \geq \frac{C_2}{4}\|\nabla F(x^t)\| - \frac{8\sigma^4}{C_1^2} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2. \tag{63}$$

Then with (48), we have

$$F(x^{t+1}) \le F(x^t) - \frac{\eta C_2}{4}\|\nabla F(x^t)\| + \frac{8\eta\sigma^4}{C_1^2} + \eta\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(1-L\eta). \tag{64}$$

(iv) If $C_2 > \|\nabla F(x^t)\|$, this is the case (B7), then from (58), we have

$$A^t \ge \frac{1}{4}\|\nabla F(x^t)\|^2 - \frac{8\sigma^4}{C_1^2} - \|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{(\beta^t)^2}{2}\|\bar{v}^t\|^2. \tag{65}$$

Then with (48), we have

$$F(x^{t+1}) \le F(x^t) - \frac{\eta}{4}\|\nabla F(x^t)\|^2 + \frac{8\eta\sigma^4}{C_1^2} + \eta\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 - \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2(1-L\eta). \tag{66}$$

This completes the proof. □

Lemma D.4. *Suppose Assumptions 1 and 2 hold. Consider* $\{x^t, v_i^t, i \in [M]\}_t$ *generated by PriSMA with* $C_1 \ge 2C_2 > 0$ *and* $C_1 \ge 16\sigma$*. By choosing step sizes such that*

$$\frac{16d\sigma_1^2}{MC_1C_2} \le \gamma \le \frac{MbC_1C_2}{32\sigma^2},\ \frac{L^2\eta^2}{Mb\gamma} \le \frac{1}{8},\ \eta\left(1 + \frac{4}{Mb}\left(\frac{LC_2\eta^2}{C_1\gamma^2} + \frac{\eta}{\gamma}\right)\right) \le \frac{1}{2L}, \tag{67}$$

we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x^t)\|\right] \le \frac{\Gamma}{C_2} + \sqrt{\Gamma}, \tag{68}$$

where Γ *is defined as*

$$\begin{aligned}\Gamma := \frac{8\Delta_F^0}{\eta T} &+ \left(\frac{16\sigma^2}{Mb} + \frac{8d\sigma_1^2}{M\gamma^2}\right)\left(\frac{C_2\gamma}{C_1} + \frac{LC_2\eta}{C_1} + \gamma\right) + \frac{64\sigma^4}{C_1^2} \\ &+ \frac{8}{\gamma MT}\left(\frac{\sigma^2}{b} + d\sigma_0^2\right)\left(\frac{\|\nabla F(x^0)\|}{C_1} + \frac{LC_2\eta}{C_1\gamma} + 1\right).\end{aligned}$$

Proof. Recalling the definition of the Lyapunov function

$$\Phi^t := F(x^t) + \frac{\eta}{\gamma}\left(\frac{1}{\lambda_1^{t-1}} + \frac{LC_2\eta}{C_1\gamma} + 1\right)\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2, \tag{69}$$

For notational convenience, we define

$$A_0 := \frac{\eta}{\gamma}\left(\frac{LC_2\eta}{C_1\gamma} + 1\right). \tag{70}$$

According to Lemma D.1, from (33), we have

$$\begin{aligned}&A_0\mathbb{E}[\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t] \\ &\le A_0\left[(1-\gamma)^2\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + 2\gamma^2\frac{\sigma^2}{Mb} + 2(1-\gamma)^2\frac{L^2(\beta^t)^2\eta^2}{Mb}\|\bar{v}^t\|^2 + \frac{d\sigma_1^2}{M}\right], \\ &\le A_0(1-\gamma)\|\nabla F^{C_1}(x^t) - \bar{v}^t\|^2 + \frac{2A_0\gamma^2\sigma^2}{Mb} + 2A_0\frac{L^2(\beta^t)^2\eta^2}{Mb}\|\bar{v}^t\|^2 + \frac{A_0 d\sigma_1^2}{M}.\end{aligned} \tag{71}$$

And from (34), we have

$$\begin{aligned}&\mathbb{E}\left[\frac{\eta}{\gamma\lambda_1^t}\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t\right] \\ &\le \frac{\eta}{\gamma}\left(\frac{1}{\lambda_1^{t-1}} + \frac{LC_2\eta}{C_1} - \frac{\gamma}{\lambda_1^t}\right)\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{2\eta}{\gamma\lambda_1^t}\left(\frac{\gamma^2\sigma^2}{Mb} + (1-\gamma)^2\frac{L^2}{Mb}\|x^{t+1} - x^t\|^2 + \frac{d\sigma_1^2}{2M}\right), \\ &\le \left(\frac{\eta}{\gamma\lambda_1^{t-1}} + \gamma A_0 - \eta - \frac{\eta}{\lambda_1^t}\right)\|\bar{v}^t - \nabla F^{C_1}(x^t)\|^2 + \frac{2\eta}{\gamma\lambda_1^t}\left(\frac{\gamma^2\sigma^2}{Mb} + \frac{L^2(\beta^t)^2\eta^2}{Mb}\|\bar{v}^t\|^2 + \frac{d\sigma_1^2}{2M}\right).\end{aligned} \tag{72}$$

Then by combining above two inequalities, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\eta}{\gamma}\left(\frac{1}{\lambda_1^t}+\frac{LC_2\eta}{C_1\gamma}+1\right)\|\bar{v}^{t+1}-\nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t\right]\\
=\;&\mathbb{E}\left[\left(\frac{\eta}{\gamma\lambda_1^t}+A_0\right)\|\bar{v}^{t+1}-\nabla F^{C_1}(x^{t+1})\|^2|\mathcal{F}^t\right]\\
\leq\;&\left(\frac{\eta}{\gamma\lambda_1^{t-1}}+\gamma A_0-\eta-\frac{\eta}{\lambda_1^t}\right)\|\bar{v}^t-\nabla F^{C_1}(x^t)\|^2+\frac{2\eta}{\gamma\lambda_1^t}\left(\frac{\gamma^2\sigma^2}{Mb}+\frac{L^2(\beta^t)^2\eta^2}{Mb}\|\bar{v}^t\|^2+\frac{d\sigma_1^2}{2M}\right)\\
&+A_0(1-\gamma)\|\bar{v}^t-\nabla F^{C_1}(x^t)\|^2+\frac{2A_0\gamma^2\sigma^2}{Mb}+2A_0\frac{L^2(\beta^t)^2\eta^2}{Mb}\|\bar{v}^t\|^2+\frac{A_0d\sigma_1^2}{M}\\
=\;&\left(\frac{\eta}{\gamma\lambda_1^{t-1}}+A_0-\eta-\frac{\eta}{\lambda_1^t}\right)\|\bar{v}^t-\nabla F^{C_1}(x^t)\|^2+\frac{2\eta}{\gamma\lambda_1^t}\left(\frac{\gamma^2\sigma^2}{Mb}+\frac{d\sigma_1^2}{2M}\right)\\
&+\left(\frac{\eta}{\gamma\lambda_1^t}+A_0\right)\frac{2L^2(\beta^t)^2\eta^2}{Mb}\|\bar{v}^t\|^2+\frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b}+d\sigma_1^2\right)\\
\overset{(a)}{=}\;&\left(\frac{\eta}{\gamma\lambda_1^{t-1}}+A_0-\eta-\frac{\eta}{\lambda_1^t}\right)\|\bar{v}^t-\nabla F^{C_1}(x^t)\|^2+\frac{2\eta}{\gamma C_1}\|\nabla F(x^t)\|^2\left(\frac{\gamma^2\sigma^2}{Mb}+\frac{d\sigma_1^2}{2M}\right)\\
&+\frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2\cdot\frac{4L^2\eta}{Mb}\left(\frac{\eta}{\gamma\lambda_1^t}+A_0\right)+\frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b}+d\sigma_1^2\right),
\end{aligned}\tag{73}$$

where (a) is from $\lambda_1^t=\frac{C_1}{\|\nabla F(x^t)\|}$.

(i) For $t\in\mathcal{J}_0=\{t,\,\|\nabla F(x^t)\|\geq C_1\}$, it holds that

$$\begin{aligned}
\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t]\leq\;&\Phi^t-\eta\left(\frac{C_2}{4}-\frac{2\gamma\sigma^2}{MbC_1}-\frac{d\sigma_1^2}{MC_1\gamma}\right)\|\nabla F(x^t)\|+\frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b}+d\sigma_1^2\right)\\
&-\frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2\left(\frac{1}{\lambda_1^t}-L\eta-\frac{4L^2\eta}{Mb}\left(\frac{\eta}{\gamma\lambda_1^t}+A_0\right)\right),
\end{aligned}\tag{74}$$

which is from (43) and (73). Since $\lambda_1^t\leq 1$ for $t\in\mathcal{J}_0$, by choosing step sizes satisfying

$$\frac{16d\sigma_1^2}{MC_1C_2}\leq\gamma\leq\frac{MbC_1C_2}{32\sigma^2},\;\frac{L^2\eta^2}{Mb\gamma}\leq\frac{1}{8},\;\eta\left(1+\frac{4A_0}{Mb}\right)\leq\frac{1}{2L},\tag{75}$$

from (74), we have

$$\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t]\leq\Phi^t-\frac{\eta C_2}{8}\|\nabla F(x^t)\|+\frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b}+d\sigma_1^2\right).\tag{76}$$

(ii) For $t\in\mathcal{J}_1=\{t,\,C_1>\|\nabla F(x^t)\|\geq\frac{C_1}{2}\}$, it holds that

$$\begin{aligned}
\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t]\leq\;&\Phi^t-\eta\left(\frac{C_2}{4}-\frac{2\gamma\sigma^2}{MbC_1}-\frac{d\sigma_1^2}{MC_1\gamma}\right)\|\nabla F(x^t)\|+\frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b}+d\sigma_1^2\right)\\
&-\frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2\left(1-L\eta-\frac{4L^2\eta}{Mb}\left(\frac{\eta}{\gamma\lambda_1^t}+A_0\right)\right),
\end{aligned}\tag{77}$$

which is from (44) and (73). Since $\lambda_1^t\geq 1$ for $t\in\mathcal{J}_1$, if the step sizes satisfy (75), from (77), we have

$$\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t]\leq\Phi^t-\frac{\eta C_2}{8}\|\nabla F(x^t)\|+\frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b}+d\sigma_1^2\right).\tag{78}$$

(iii) For $t\in\mathcal{J}_2=\{t,\,\frac{C_1}{2}>\|\nabla F(x^t)\|\geq C_2\}$, similar to **(ii)**, it holds that

$$\begin{aligned}
\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t]\leq\;&\Phi^t-\eta\left(\frac{C_2}{4}-\frac{2\gamma\sigma^2}{MbC_1}-\frac{d\sigma_1^2}{MC_1\gamma}\right)\|\nabla F(x^t)\|+\frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b}+d\sigma_1^2\right)\\
&-\frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2\left(1-L\eta-\frac{4L^2\eta}{Mb}\left(\frac{\eta}{\gamma\lambda_1^t}+A_0\right)\right)+\frac{8\eta\sigma^4}{C_1^2},
\end{aligned}\tag{79}$$

which is from (45) and (73). If the step sizes satisfy (75), we have

$$\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t] \leq \Phi^t - \frac{\eta C_2}{8}\|\nabla F(x^t)\| + \frac{A_0}{M}\left(\frac{2\gamma^2\sigma^2}{b} + d\sigma_1^2\right) + \frac{8\eta\sigma^4}{C_1^2}. \tag{80}$$

(iv) For $t \in \mathcal{J}_3 = \{t,\, C_2 > \|\nabla F(x^t)\|\}$, it holds that

$$\begin{aligned}\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t] \leq{}& \Phi^t - \frac{\eta}{4}\|\nabla F(x^t)\|^2 + \left(\frac{\eta}{\lambda_1^t\gamma} + A_0\right)\left(\frac{2\gamma^2\sigma^2}{Mb} + \frac{d\sigma_1^2}{M}\right) + \frac{8\eta\sigma^4}{C_1^2}\\ &- \frac{\eta(\beta^t)^2}{2}\|\bar{v}^t\|^2\left(1 - L\eta - \frac{4L^2\eta}{Mb}\left(\frac{\eta}{\gamma\lambda_1^t} + A_0\right)\right),\end{aligned} \tag{81}$$

which is from (46) and (73). Since $\lambda_1^t \geq \frac{C_1}{C_2}$ for $t \in \mathcal{J}_3$ and the step sizes satisfy (75), we have

$$\mathbb{E}[\Phi^{t+1}|\mathcal{F}^t] \leq \Phi^t - \frac{\eta}{8}\|\nabla F(x^t)\|^2 + \left(\frac{C_2\eta}{C_1\gamma} + A_0\right)\left(\frac{2\gamma^2\sigma^2}{Mb} + \frac{d\sigma_1^2}{M}\right) + \frac{8\eta\sigma^4}{C_1^2}. \tag{82}$$

Next, we sum up the above results. Denote $\mathcal{I}_0 = \{t,\, \|\nabla F(x^t)\| \geq C_2\}$, $\mathcal{I}_1 = \{t,\, C_2 > \|\nabla F(x^t)\|\}$, and define Id_i^t as $\mathrm{Id}_i^t = 1$ if $t \in \mathcal{I}_i$; otherwise $\mathrm{Id}_i^t = 0$. Then according to the above results, we have

$$\begin{aligned}&C_2\mathrm{Id}_0^t\|\nabla F(x^t)\| + \mathrm{Id}_1^t\|\nabla F(x^t)\|^2\\ \leq{}& \frac{8\mathbb{E}[\Phi^t - \Phi^{t+1}|\mathcal{F}^t]}{\eta} + \left(\frac{16\gamma^2\sigma^2}{Mb} + \frac{8d\sigma_1^2}{M}\right)\left(\frac{C_2}{C_1\gamma} + \frac{A_0}{\eta}\right) + \frac{64\sigma^4}{C_1^2}\\ \overset{(a)}{=}{}& \frac{8\mathbb{E}[\Phi^t - \Phi^{t+1}|\mathcal{F}^t]}{\eta} + \left(\frac{16\sigma^2}{Mb} + \frac{8d\sigma_1^2}{M\gamma^2}\right)\left(\frac{C_2\gamma}{C_1} + \frac{LC_2\eta}{C_1} + \gamma\right) + \frac{64\sigma^4}{C_1^2},\end{aligned} \tag{83}$$

where (a) is from (70).

Taking the expectation over time on both sides of the above equation and summing up over all the indices t, we obtain

$$\begin{aligned}&\frac{1}{T}\mathbb{E}\left[\sum_{t\in\mathcal{I}_0} C_2\|\nabla F(x^t)\| + \sum_{t\in\mathcal{I}_1}\|\nabla F(x^t)\|^2\right]\\ \leq{}& \frac{8\mathbb{E}[\Phi^0 - \Phi^T]}{\eta T} + \left(\frac{16\sigma^2}{Mb} + \frac{8d\sigma_1^2}{M\gamma^2}\right)\left(\frac{C_2\gamma}{C_1} + \frac{LC_2\eta}{C_1} + \gamma\right) + \frac{64\sigma^4}{C_1^2}\\ \leq{}& \frac{8\Delta_F^0}{\eta T} + \frac{8}{MT\gamma}\left(\frac{\sigma^2}{b} + d\sigma_0^2\right)\left(\frac{\|\nabla F(x^0)\|}{C_1} + \frac{LC_2\eta}{C_1\gamma} + 1\right)\\ &+ \left(\frac{16\sigma^2}{Mb} + \frac{8d\sigma_1^2}{M\gamma^2}\right)\left(\frac{C_2\gamma}{C_1} + \frac{LC_2\eta}{C_1} + \gamma\right) + \frac{64\sigma^4}{C_1^2} := \Gamma,\end{aligned} \tag{84}$$

where $\Delta_F^0 := F(x^0) - \min_x F(x)$. From above inequality, we have

$$\frac{1}{T}\mathbb{E}\left[\sum_{t\in\mathcal{I}_0} C_2\|\nabla F(x^t)\|\right] \overset{(a)}{\leq} \Gamma, \tag{85}$$

$$\frac{1}{T}\mathbb{E}\left[\sum_{t\in\mathcal{I}_1}\left(2\sqrt{\Gamma}\|\nabla F(x^t)\| - \Gamma\right)\right] \overset{(b)}{\leq} \frac{1}{T}\mathbb{E}\left[\sum_{t\in\mathcal{I}_1}\|\nabla F(x^t)\|^2\right] \overset{(c)}{\leq} \Gamma, \tag{86}$$

where (a) and (c) are from (84), (b) is from

$$2\sqrt{\Gamma}\|\nabla F(x^t)\| - \Gamma = \|\nabla F(x^t)\|^2 - (\|\nabla F(x^t)\| - \sqrt{\Gamma})^2 \leq \|\nabla F(x^t)\|^2.$$

Then it follows that

$$\frac{1}{T}\mathbb{E}\left[\sum_{t\in\mathcal{I}_0}\|\nabla F(x^t)\|\right] \leq \frac{\Gamma}{C_2}, \quad \frac{1}{T}\mathbb{E}\left[\sum_{t\in\mathcal{I}_1}\|\nabla F(x^t)\|\right] \leq \sqrt{\Gamma}. \tag{87}$$

Eventually, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x^t)\|\right] \le \frac{\Gamma}{C_2} + \sqrt{\Gamma}, \tag{88}$$

which completes the proof. □

E Proof for Non-PriSMA

Lemma 1. *Under Assumptions 1 and 2, for any $\gamma \in (0,1)$, the sequence $\{v_i^t\}_t$ generated by Non-PriSMA satisfies*

$$\begin{aligned}&\mathbb{E}\|\bar{v}^{t+1} - \nabla F^{C_1}(x^{t+1})\|^2\\ \le\ &(1-\gamma)^2\mathbb{E}\|\bar{v}^{t} - \nabla F^{C_1}(x^{t})\|^2 + 2\gamma^2\frac{\sigma^2}{Mb} + 2(1-\gamma)^2\frac{L^2}{b}\mathbb{E}\|x^{t+1}-x^t\|^2.\end{aligned} \tag{89}$$

Proof. Following Lemma D.1 and setting DP noise as zero, we complete the proof of Lemma 1. □

Theorem 2 (Convergence of Non-PriSMA)**.** *Under Assumptions 1 and 2, set the clipping thresholds such that $C_1 \ge 2C_2$ and $C_1 \ge 16\sigma$, while set the step sizes such that*

$$\gamma = \min\left\{\frac{MbC_1C_2}{32\sigma^2}, \sqrt{\frac{M}{T}}\right\},\ \eta = \min\left\{\frac{\sqrt{Mb\gamma}}{2\sqrt{2}L}, \frac{Mb\gamma}{4}, \frac{C_1\gamma}{LC_2}, \frac{1}{6L}, \sqrt{\frac{M}{T}}\right\}. \tag{90}$$

Then, the sequence $\{x^t\}_t$ generated by Non-PriSMA satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\| \le \mathcal{O}\left(\frac{1}{(MT)^{\frac{1}{2}}C_2} + \frac{1}{(MT)^{\frac{1}{4}}} + \frac{\sigma^4}{C_1^2C_2} + \frac{\sigma^2}{C_1}\right). \tag{91}$$

Proof. For Non-PriSMA, we have $\sigma_0 = \sigma_1 = 0$. Observe that the step sizes (90) satisfies the requirements in (67) of Lemma D.4. With such step sizes, we have

$$\Gamma \le \mathcal{O}\left(\frac{1}{\sqrt{MT}} + \frac{\sigma^4}{C_1^2}\right), \tag{92}$$

where Γ is defined in Lemma D.4. Then according to Lemma D.4, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\| \le \mathcal{O}\left(\frac{1}{(MT)^{\frac{1}{2}}C_1} + \frac{1}{(MT)^{\frac{1}{4}}} + \frac{\sigma^4}{C_1^2C_2} + \frac{\sigma^2}{C_1}\right), \tag{93}$$

which completes the proof. □

F Proof for PriSMA

Theorem 3 (RDP guarantee)**.** *Under Assumption 1, given (α,ρ)-RDP with $\alpha > 1$ and $\rho > 0$, let the variances of injected Gaussian noises in PriSMA satisfy*

$$\sigma_0^2 = \frac{3.5T\Delta_0^2\alpha}{N^2\rho},\ \sigma_1^2 = \frac{3.5T\Delta_1^2\alpha}{N^2\rho}, \tag{94}$$

where $\Delta_0 = 2C_1$ and $\Delta_1 = 2\left(\gamma C_1 + (1-\gamma)L\eta C_2\right)$. If it holds that

$$\hat{\sigma}^2 := \frac{3.5b^2T\alpha}{N^2\rho} \ge 0.7,\ \alpha \le \frac{2\hat{\sigma}^2}{3}\log\frac{N}{b\alpha(1+\hat{\sigma}^2)} + 1, \tag{95}$$

then PriSMA satisfies (α,ρ)-RDP for each client i.

Proof. First, we analyze the sensitivity of the queries in PriSMA.
Case $t = 0$: For each client i, we consider the query on $\mathcal{B}_i^0$ with $|\mathcal{B}_i^0| = b$, in the form of

$$q_0(\mathcal{B}_i^0) = \frac{1}{b}\sum_{j\in\mathcal{B}_i^0} \text{clip}_{C_1}\left(f_{ij}(x^0)\right), \tag{96}$$

Then, the ℓ_2-sensitivity of q_i^0 is bounded as

$$\begin{aligned}\tilde{\Delta}_0 &= \max_{\mathcal{B}_i^0 \sim \mathcal{B}_i^{0\prime}} \|q^0(\mathcal{B}_i^0) - q^0(\mathcal{B}_i^{0\prime})\| \\ &= \max_{j,j'} \frac{1}{b}\|\text{clip}_{C_1}\left(f_{ij}(x^0)\right) - \text{clip}_{C_1}\left(f_{ij'}(x^0)\right)\| \\ &\leq \frac{2C_1}{b}. \end{aligned}\tag{97}$$

Case $t > 0$: For each client i, we consider the query on $\mathcal{B}_i^t$ with $|\mathcal{B}_i^t| = b$, in the form of

$$q_1(\mathcal{B}_i^t) = \frac{1}{b}\sum_{j\in\mathcal{B}_i^t} \text{clip}_{C_1}\left(f_{ij}(x^t)\right) + (1-\gamma)(v_i^{t-1} - \frac{1}{b}\sum_{j\in\mathcal{B}_i^t} \text{clip}_{C_1}\left(f_{ij}(x^{t-1})\right)), \tag{98}$$

Then, the ℓ_2-sensitivity of q_i^0 is bounded as

$$\begin{aligned}\tilde{\Delta}_1 &= \max_{\mathcal{B}_i^t, \mathcal{B}_i^{t\prime}} \|q_1(\mathcal{B}_i^t) - q_1(\mathcal{B}_i^{t\prime})\| \\ &= \max_{j,j'} \frac{1}{b}\|\text{clip}_{C_1}\left(f_{ij}(x^t)\right) - (1-\gamma)\text{clip}_{C_1}\left(f_{ij}(x^{t-1})\right) \\ &\quad - \text{clip}_{C_1}\left(f_{ij'}(x^0)\right) + (1-\gamma)\text{clip}_{C_1}\left(f_{ij'}(x^{t-1})\right)\| \\ &\leq \frac{1}{b}\max_{j,j'}\Big\{\gamma\|\text{clip}_{C_1}\left(f_{ij}(x^t)\right) - \text{clip}_{C_1}\left(f_{ij'}(x^0)\right)\| \\ &\quad + (1-\gamma)\|(\text{clip}_{C_1}\left(f_{ij}(x^t)\right) - \text{clip}_{C_1}\left(f_{ij}(x^{t-1})\right)) - (\text{clip}_{C_1}\left(f_{ij'}(x^t)\right) - \text{clip}_{C_1}\left(f_{ij'}(x^{t-1})\right))\|\Big\} \\ &\leq \frac{2}{b}(\gamma C_1 + (1-\gamma)L\|x^t - x^{t-1}\|) \\ &\leq \frac{2}{b}(\gamma C_1 + (1-\gamma)L\eta C_2). \end{aligned}\tag{99}$$

Second, according to the result of privacy amplification by sub-sampling Wang et al. (2023, 2019), we set the variances of injected Gaussian noises in PriSMA such that

$$\sigma_0^2 = \frac{3.5T\Delta_0^2\alpha}{N^2\rho},\ \sigma_1^2 = \frac{3.5T\Delta_1^2\alpha}{N^2\rho}, \tag{100}$$

where

$$\Delta_0 = b\tilde{\Delta}_0,\ \Delta_1 = b\tilde{\Delta}_1.$$

With them, PriSMA satisfies (α, ρ)-RDP for each client i, if it holds that

$$\hat{\sigma}^2 := \frac{3.5b^2T\alpha}{N^2\rho} \geq 0.7,\ \alpha \leq \frac{2\hat{\sigma}^2}{3}\log\frac{N}{b\alpha(1+\hat{\sigma}^2)} + 1. \tag{101}$$

This completes the proof. □

Corollary 1. *Under the same conditions as in Theorem 3, assume that σ_0^2 and σ_1^2 satisfies* (17)*, and*

$$\alpha = 1 + \frac{2\log\frac{1}{\delta}}{\epsilon},\ \rho = \frac{\epsilon}{2}, \tag{102}$$

where $\epsilon > 0$ and $\delta \in (0, 1)$. Then, the proposed PriSMA satisfies (ϵ, δ)-LDP. Futhermore, if $b = \Theta(\frac{N\sqrt{\epsilon}}{\sqrt{T}})$ and $T \geq \mathcal{O}(\frac{\log^4\frac{1}{\delta}}{\epsilon^3})$, the condition (18) *in Theorem 3 is satisfied.*

Proof. From Theorem 3 and Fact 1, by setting α and ρ as in (102), we can verify that PriSMA satisfies (ϵ, δ)-LDP. Since $\alpha > 1$, by setting $b^2 = C\frac{N^2\epsilon}{T}$, where $C \geq \frac{3}{28}$ is a constant, we obtain

$$\hat{\sigma}^2 := \frac{3.5b^2T\alpha}{N^2\rho} = 14C\alpha \geq \frac{3}{2}\alpha \geq 0.7. \tag{103}$$

Consequently, the following inequality holds:

$$\frac{b\alpha(1+\hat{\sigma}^2)}{N} \leq \frac{b}{N}\alpha(1+14C\alpha) \leq \frac{b}{N}\alpha^2(14C+1) = \sqrt{\frac{C\epsilon}{T}}(1+\frac{2\log\frac{1}{\delta}}{\epsilon})^2(14C+1). \tag{104}$$

If $\frac{2\log\frac{1}{\delta}}{\epsilon} \geq 1$, by selecting $T \geq \mathcal{O}(\frac{\log^4\frac{1}{\delta}}{\epsilon^3})$, we have

$$\frac{b\alpha(1+\hat{\sigma}^2)}{N} \leq (14C+1)\frac{16\log^2\frac{1}{\delta}}{\epsilon^2}\sqrt{\frac{C\epsilon}{T}} \leq \frac{1}{3}. \tag{105}$$

If $\frac{2\log\frac{1}{\delta}}{\epsilon} < 1$, by selecting $T \geq \mathcal{O}(\epsilon)$, we have

$$\frac{b\alpha(1+\hat{\sigma}^2)}{N} \leq 4(14C+1)\sqrt{\frac{C\epsilon}{T}} \leq \frac{1}{3}. \tag{106}$$

Therefore, combining (104), (105), and (106), we obtain

$$\frac{2\hat{\sigma}^2}{3}\log\frac{N}{b\alpha(1+\hat{\sigma}^2)} + 1 \geq \frac{2\hat{\sigma}^2}{3} + 1 \geq \alpha + 1 \geq \alpha, \tag{107}$$

which completes the proof. □

Theorem 4 (Utility of PriSMA)**.** *Suppose Assumptions 1–2 hold. Given (ϵ, δ)-LDP with $\epsilon > 0$ and $0 < \delta < 1$, by choosing appropriate parameters, the sequence $\{x^t\}_t$ generated by PriSMA satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\| \leq \mathcal{O}\left(\frac{d\log\frac{1}{\delta}}{MN^2\epsilon^2}\right)^{\frac{1}{5}}. \tag{108}$$

Proof. Given (ϵ, δ)-LDP, according to Corollary 1, each client in PriSMA injects Gaussian noise with variances

$$\sigma_0^2 = \frac{a_0TC_1^2\log\frac{1}{\delta}}{N^2\epsilon^2},\ \sigma_1^2 = \frac{a_1T(\gamma C_1 + (1-\gamma)L\eta C_2)^2\log\frac{1}{\delta}}{N^2\epsilon^2}, \tag{109}$$

where a_0 and a_1 are positive constants. Furthermore, the requirements (18) in Theorem 3 can be satisfied by setting $b = \Theta(\frac{N}{\sqrt{\epsilon T}})$ and $T \geq \mathcal{O}(\frac{\log^4\frac{1}{\delta}}{\epsilon^3})$.

For convenience, we define

$$\xi := \frac{d\log\frac{1}{\delta}}{MN^2\epsilon^2}.$$

We choose clipping thresholds as

$$C_2 = \mathcal{O}(1), \quad C_1 \geq \max\{16\sigma,\, 2C_2\}, \tag{110}$$

step sizes such that

$$\gamma \leq\ \min\left\{\frac{C_2}{64C_1a_1\xi}, \frac{MbC_1C_2}{32\sigma^2}\right\}, \tag{111}$$

$$\eta =\ \min\left\{\frac{1}{64La_1\xi}, \sqrt{\frac{C_1\gamma}{64LC_2a_1\xi}}, \frac{\sqrt{Mb\gamma}}{2\sqrt{2}L}, \frac{Mb\gamma}{4}, \frac{C_1\gamma}{LC_2}, \frac{1}{6L}\right\} = \mathcal{O}(1). \tag{112}$$

Thus, the requirements (67) in Lemma D.4 are satisfied, and from Lemma D.4 we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x^t)\|\right] \leq \frac{\Gamma}{C_2} + \sqrt{\Gamma}, \tag{113}$$

where

$$\begin{aligned}\Gamma &\leq \mathcal{O}\left(\frac{\Delta_F^0}{\eta T} + \frac{\sigma^2}{Mb}\left(\gamma + \frac{1}{\gamma T}\right) + \frac{d\sigma_1^2}{M\gamma} + \frac{d\sigma_0^2}{\gamma MT} + \frac{\sigma^4}{C_1^2}\right)\\ &\leq \mathcal{O}\left(\frac{\Delta_F^0}{\eta T} + \frac{\sigma^2}{Mb}\left(\gamma + \frac{1}{\gamma T}\right) + C_1^2\xi\left(\frac{1}{\gamma} + \gamma T\right) + \frac{\sigma^4}{C_1^2}\right). \end{aligned}\tag{114}$$

Therefore, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(x^t)\|\right] \leq \mathcal{O}\left(\sqrt{\frac{\Delta_F^0}{\eta T}} + \frac{\sigma}{\sqrt{Mb}}\sqrt{\gamma + \frac{1}{\gamma T}} + C_1\sqrt{\xi}\sqrt{\frac{1}{\gamma} + \gamma T} + \frac{\sigma^2}{C_1}\right). \tag{115}$$

The third and fourth terms on the right-hand side of the above inequality are due to DP noise and clipping bias, respectively. Next, we will select C_1 to balance these terms. With large enough N, we further choose the total number of iterations as

$$T = \Theta\left(\max\left\{\xi^{-\frac{2}{5}}, \frac{\log^4\frac{1}{\delta}}{\epsilon^3}\right\}\right),$$

the clipping threshold C_1 as

$$C_1 = \max\left\{16\sigma,\ 2C_2,\ \xi^{-\frac{1}{5}}\right\} = \Theta\left(\xi^{-\frac{1}{5}}\right), \tag{116}$$

the step size γ as

$$\gamma = \min\left\{\frac{C_2}{64C_1a_1\xi}, \frac{MbC_1C_2}{32\sigma^2}, \xi^{\frac{1}{5}}\right\} = \mathcal{O}\left(\xi^{\frac{1}{5}}\right), \tag{117}$$

and the batch size as

$$b = \Theta(\frac{N}{\sqrt{\epsilon T}}) = \Theta(N^{\frac{3}{5}}). \tag{118}$$

Then from (115), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x^t)\| \leq \mathcal{O}\left(\xi^{\frac{1}{5}}\right). \tag{119}$$

This completes the proof. □