On the Privacy-Robustness-Utility Trilemma in Distributed Learning

Youssef Allouah[1] Rachid Guerraoui[1] Nirupam Gupta[1] Rafaël Pinot[1] John Stephan[1]

Abstract

The ubiquity of distributed machine learning (ML) in sensitive public domain applications calls for algorithms that protect data *privacy*, while being *robust* to faults and adversarial behaviors. Although privacy and robustness have been extensively studied independently in distributed ML, their synthesis remains poorly understood. We present the first *tight* analysis of the error incurred by any algorithm ensuring robustness against a fraction of adversarial machines, as well as *differential privacy* (DP) for honest machines' data against any other curious entity. Our analysis exhibits a fundamental trade-off between privacy, robustness, and utility. To prove our lower bound, we consider the case of mean estimation, subject to distributed DP and robustness constraints, and devise reductions to centralized estimation of one-way marginals. We prove our matching upper bound by presenting a new distributed ML algorithm using a high-dimensional robust aggregation rule. The latter amortizes the dependence on the dimension in the error (caused by adversarial workers and DP), while being agnostic to the statistical properties of the data.

1. Introduction

Distributed machine learning (ML) has been playing a pivotal role in a wide range of applications (Dean et al., 2012; Abadi et al., 2016), due to an unprecedented growth in the complexity of ML models and the volume of data being used for training purposes. Distributed ML breaks a complex ML task into sub-tasks that are performed in a collaborative fashion. In the standard *server-based* architecture, n machines (a.k.a., *workers*) collaboratively train a global model on their datasets, with the help of a coordinator (the *server*). This is typically achieved through a distributed implementation of the renowned *stochastic gradient descent* (SGD) algorithm (Bertsekas & Tsitsiklis, 2015). In distributed SGD (or DSGD), the server maintains a model which is updated iteratively by averaging gradients of the loss function associated with the model, computed by the different workers upon sampling random points from their local datasets. DSGD is particularly useful in cases where the data held by the workers is too sensitive to be shared, e.g., medical data collected by several hospitals (Sheller et al., 2020).

Privacy. Although DSGD inherently ensures privacy of the workers' data to an extent, by not sharing it explicitly, information leakage can still be significant. When the ML model maintained at the server is publicly released, it may be exposed to membership inference (Shokri et al., 2016) or model inversion attacks (Fredrikson et al., 2015; Hitaj et al., 2017; Melis et al., 2019) by external entities. Furthermore, upon observing the gradients and transient models during the learning procedure, *curious* machines (be they workers or the server itself) can infer sensitive information about the datasets held locally by the machines, or even reconstruct data points in certain scenarios (Phong et al., 2017; Wang et al., 2019b; Zhu et al., 2019; Zhao et al., 2020).

Robustness. In real-world distributed systems, it is arguably inevitable to encounter faulty workers that may deviate from their prescribed algorithm. This may result from hardware and software bugs, data corruption, network latency, or malicious adversaries controlling a subset of workers. To cover all such possible scenarios, it is common to assume that a fraction of the machines can be *adversarial*[1] and arbitrarily deviate from their algorithms. In the context of DSGD, adversarial workers may send incorrect gradients (Feng et al., 2015; Su & Vaidya, 2016) to the server and critically influence the learning procedure, as shown in (Baruch et al., 2019; Xie et al., 2019).

Integrating privacy and robustness. With the growing concerns and legal obligations regarding the processing of public data in AI-driven technologies (EU, 2016), privacy and robustness issues question the very applicability of ML in critical public domain services, such as healthcare or banking. It is thus natural to seek distributed ML methods that simultaneously ensure privacy and robustness. In fact,

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Correspondence to: Youssef Allouah <youssef.allouah@epfl.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Sometimes called "Byzantine" in the parlance of distributed computing (Lamport et al., 1982).

these aspects have separately received significant attention in the past. On the one hand, the standard statistical privacy requirement of (ε, δ)-*differential privacy* ((ε, δ)-DP) has been studied to a great extent in the context of distributed ML (Choudhury et al., 2019; Hu et al., 2020; Noble et al., 2022). On the other hand, numerous provably robust adaptations of DSGD have been proposed (Blanchard et al., 2017; Xie et al., 2018; Yin et al., 2018; Gupta et al., 2021; Farhadkhani et al., 2022). Yet, the synthesis of privacy and robustness remains highly understudied in distributed ML. The few works on this topic, such as (Guerraoui et al., 2021; Zhu & Ling, 2022; Xiang & Su, 2022; Ma et al., 2022), only focus on *per-step privacy*, and provide loose upper bounds on the learning error. On the other hand, the guarantees presented in (Cheu et al., 2021; Acharya et al., 2021) only apply to discrete distribution estimation subject to *non-interactive local DP* (Kasiviswanathan et al., 2011), a restricted case of distributed ML where each worker holds a single data point and can be queried only once.

An orthogonal line of work studied the case where the server is assumed not to be curious, i.e., data only needs to be protected against the public release of the model (Dwork & Lei, 2009; Liu et al., 2021b; Hopkins et al., 2022a; Liu et al., 2022). In this setting, it was recently shown that privacy and robustness are *mutually* beneficial (Georgiev & Hopkins, 2022; Hopkins et al., 2022b). However, the assumption of a non-curious server may not be viable, especially in applications such as healthcare and finance, where sovereignty of data must be protected at every stage of the learning procedure (Lowy et al., 2023). In this paper, we focus on the setting where the server itself may be curious, and we show that privacy and robustness are actually at odds.

1.1. Contributions

We precisely characterize the privacy-robustness-utility trilemma in distributed learning. Specifically, we present the first *tight* analysis of the error incurred by any distributed ML algorithm that simultaneously ensures (i) robustness against a minority of adversarial workers, and (ii) differential privacy (DP) of each worker's data against curious entities including other workers and the server. In short, we show that, in addition to the usual separate costs of privacy and robustness, the learning accuracy necessarily degrades due to their interplay.

Main results. We consider a system of n workers up to f of which (of unknown identity) may be adversarial, and the remainder are *honest*. The server is assumed *honest-but-curious* (Bonawitz et al., 2016). Each honest worker holds a dataset comprising m points. The goal of the server is to learn a model, parameterized by a d-dimensional vector, incurring minimum loss over the collective dataset of the honest workers. We denote by G the *heterogeneity* (Karimireddy et al., 2020; 2022) between the honest datasets.

We show that a distributed learning algorithm that is robust to f adversarial workers, while ensuring (ε, δ)-DP of each honest worker's data against the server (and other curious workers) incurs a training error in

$$\widetilde{\Omega}\left(\frac{d}{\varepsilon^2 n m^2} + \frac{f}{n} \cdot \frac{1}{\varepsilon^2 m^2} + \frac{f}{n} \cdot G^2\right), \tag{1}$$

where $\widetilde{\Omega}$ ignores the logarithmic terms.

The first and the third terms in (1) are the respective errors due to privacy and robustness separately. Importantly, the second term represents the additional cost of satisfying privacy and robustness simultaneously. We then present a new distributed ML algorithm, SAFE-DSHB[2], which we prove yields a matching upper bound (up to a logarithmic factor) for the class of smooth and strongly convex loss functions, while ensuring both privacy and robustness. We also obtain an upper bound for smooth *non-convex* learning problems.

The key to proving the tightness of this trade-off is the robust high-dimension aggregation rule we introduce, namely *SMEA*[3]. As an important consequence of our result, we observe that the privacy-robustness trade-off (second term) is dominated by the privacy cost alone (first term) when the dimension d is larger than the number of adversarial workers f. This observation however does not mean that the trade-off is not significant, but rather that it can be adequately controlled when using SMEA. This would not have been possible otherwise with the use of existing aggregation rules such as coordinate-wise or geometric median, for which the upper bound has an additional dimension factor in the privacy-robustness trade-off.

Independent contributions. As a byproduct of our analysis, we obtain several results that are of independent interest to both the robust distributed ML and the privacy communities. Indeed, our upper bound is tight for strongly convex losses, even when removing the privacy constraints. This is mainly due to the use of momentum in SAFE-DSHB (see Section 1.2 below) which allows obtaining an excess error that is independent of the variance of local stochastic gradients. This improves over the state-of-the-art analysis on robust distributed learning with strongly convex losses (Data & Diggavi, 2021), which induces a suboptimal excess error. Besides, our analysis features a tighter dependence on heterogeneity in the excess error. Our lower bound on the cost of privacy (without robustness) also improves over the state-of-the-art (Lowy & Razaviyayn, 2023) as we make no assumptions on the *interactivity* of the algorithm and impose weaker conditions on the DP parameter ε (see Section 3).

[2] Safe Distributed Stochastic Heavy Ball method, inspired from the optimization literature (Gadat et al., 2018).

[3] Smallest Maximum Eigenvalue Averaging.

1.2. Overview of Proof Techniques

Lower bound. We prove our lower bound by reducing distributed mean estimation to centralized estimation of one-way marginals (i.e. row-wise averages). We distinguish cases depending on the presence of adversarial workers. In each case, we start with a distributed algorithm $\mathcal{A}$ whose interactions with each worker are (ε, δ)-DP, and then construct a *centralized* algorithm $\mathcal{M}$ using $\mathcal{A}$. Depending on the case, we then use either the advanced composition theorem (Dwork et al., 2014) or an indistinguishability argument on the honest identities to relate the DP and utility guarantees of $\mathcal{M}$ to those of $\mathcal{A}$. We conclude by applying lower bounds on centralized private estimation of one-way marginals (Steinke & Ullman, 2016) to $\mathcal{M}$.

Upper bound. To prove our matching upper bound, we present SAFE-DSHB, a privacy-preserving robust adaptation of DSGD. Our algorithm incorporates *Polyak's momentum* (Polyak, 1964) and a *Gaussian mechanism* (Dwork et al., 2014) at the worker level, as well as *SMEA*, our robust aggregation rule at the server level. We identify a key property that, if satisfied by an aggregation rule, mitigates the curse of dimensionality that could impact the Gaussian mechanism. This property, called (f, κ)-*robust averaging*, requires the squared distance between the aggregate and the average of honest vectors to be bounded by κ times the spectral norm of the empirical covariance matrix of the honest vectors. Our aggregation rule, SMEA, satisfies (f, κ)-robust averaging for $\kappa = \mathcal{O}(f/n)$, while being agnostic to the statistical properties of honest inputs. Another critical element of our analysis is the tuning of the momentum coefficients to control the trade-off between the deviation from the true gradient and the reduction of the *drift* between honest workers' momentums. We achieve this through a novel *Lyapunov function* (a.k.a. potential function in optimization literature (Schmidt et al., 2017)).

1.3. Prior Work

Only a handful of works addressed the interplay between DP and robustness in distributed ML. It was conjectured that ensuring both these requirements is *impractical*, in the sense that it would require the batch size to grow with the model dimension (Guerraoui et al., 2021). However, the underlying analysis relied upon the criterion of $(\alpha,\ f)$-Byzantine resilience (Blanchard et al., 2017), which has been recently shown to be a restrictive sufficient condition (Karimireddy et al., 2021). Subsequent works (Zhu & Ling, 2022; Xiang & Su, 2022; Ma et al., 2022) augmented the RSA learning algorithm (Li et al., 2019) with the sign-flipping or sign-Gaussian privacy mechanisms. However, these works only focus on *per-step* DP, and the presented upper bounds on the error of the proposed algorithms are loose.

Another line of work targeted the specific learning problem of *discrete distribution estimation* subject to *non-interactive local DP* (Duchi et al., 2013) and robustness constraints. The bounds for this problem (Cheu et al., 2021; Acharya et al., 2021) are comparable to ours in the particular scenario where each worker holds a single data point and the algorithm is non-interactive (can query each worker once). Although a recent paper (Chhor & Sentenac, 2023) considered a more general case where workers hold a batch of data points, the algorithm was still assumed non-interactive, and the data distribution identical for all the workers. It was also shown recently (Li et al., 2022) that local DP and robustness are disentangled when the adversarial workers corrupt the data before randomization only, which however need not be the case in general. The aforementioned works being tailored to non-interactive local DP, it is not clear how to extend their results to the general distributed ML setting.

Significant attention was given to robust mean estimation under DP (Dwork & Lei, 2009; Liu et al., 2021b; Hopkins et al., 2022a; Liu et al., 2022). However, as we pointed out, the corresponding results do not readily apply to our setting, as they would require the server to be *non-curious*. Moreover, robust mean estimation (Diakonikolas et al., 2019; Ashtiani & Liaw, 2022; Liu et al., 2022) typically assumes the honest inputs to be identically distributed, which need not be the case in a general distributed setting.

1.4. Paper Outline

Section 2 defines the problem and recalls some useful concepts. Sections 3 and 4 present our lower bound and the analysis of SAFE-DSHB. Section 5 presents SMEA and derives our matching upper bound. Section 6 discusses future work. We defer full proofs to appendices A-D, and experimental evaluation to Appendix E.

2. Problem Statement

We consider the classical server-based architecture comprising n workers $w_1, \dots, w_n$, and a central server. The workers hold local datasets $\mathcal{D}_1, \dots, \mathcal{D}_n$, each composed of m data points from an input space $\mathcal{X}$, i.e., $\mathcal{D}_i \coloneqq \{x_1^{(i)}, \dots, x_m^{(i)}\} \in \mathcal{X}^m$. For a given parameter vector $\theta \in \mathbb{R}^d$, a data point $x \in \mathcal{X}$ has a real-valued loss function $\ell(\theta; x)$. The empirical loss function for each worker w_i is defined by

$$\mathcal{L}(\theta; \mathcal{D}_i) \coloneqq \frac{1}{m} \sum_{x \in \mathcal{D}_i} \ell(\theta; x).$$

The goal of the server is to compute an optimal parameter vector θ^* minimizing the global empirical loss function $\mathcal{L}(\theta; \mathcal{D}_1, \dots, \mathcal{D}_n)$ defined to be

$$\mathcal{L}(\theta; \mathcal{D}_1, \dots, \mathcal{D}_n) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; \mathcal{D}_i).$$

We assume that each loss $\mathcal{L}(\cdot; \mathcal{D}_i)$ is differentiable, and that $\mathcal{L}$ is lower bounded, i.e., $\inf_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta; \mathcal{D}_1, \dots, \mathcal{D}_n)$ is finite.

2.1. Robustness

We consider a setting where at most f out of n workers may be adversarial. Such workers may send arbitrary messages to the server, and need not follow the prescribed protocol. The identity of adversarial workers is a priori unknown to the server. Let $\mathcal{H} \subseteq \{1, \dots, n\}$, with $|\mathcal{H}| = n - f$. We define

$$\mathcal{L}_{\mathcal{H}}(\theta) \coloneqq \mathcal{L}(\theta; \mathcal{D}_i, \, i \in \mathcal{H}) \coloneqq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathcal{L}(\theta; \mathcal{D}_i).$$

If $\mathcal{H}$ represents the indices of honest workers, the function $\mathcal{L}_{\mathcal{H}}$ is referred to as the global *honest loss*. An algorithm is deemed robust to adversarial workers if it enables the server to compute a minimum of the global honest loss (Gupta & Vaidya, 2020). Formally, we define robustness as follows.

Definition 2.1 (**(f, ϱ)-robust).** *A* distributed *algorithm is said to be* (f, ϱ)-robust *if it outputs a parameter $\hat{\theta}$ such that*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\hat{\theta}) - \mathcal{L}_*\right] \leq \varrho,$$

where $\mathcal{L}_* \coloneqq \inf_{\theta \in \mathbb{R}^d} \mathcal{L}_{\mathcal{H}}(\theta)$, and the expectation is taken over the randomness of the algorithm.

In other words, an algorithm $\mathcal{A}$ is said to be (f, ϱ)-robust if, in every execution of $\mathcal{A}$, the server outputs a *ϱ-approximate* minimizer of the honest loss, despite the presence of up to f adversarial workers. Note that (f, ϱ)-robustness is in general impossible for any ϱ when $f \geq \frac{n}{2}$ (Liu et al., 2021a). Thus, throughout the paper, we assume that $f < \frac{n}{2}$.

2.2. Differential Privacy

Each honest worker $w_i, i \in \mathcal{H}$, aims to protect the privacy of their dataset $\mathcal{D}_i$ against all other entities, i.e., the server and the other workers. To define our privacy requirement formally, we recall below the definition of item-level differential privacy (DP) (Dwork et al., 2014), where two datasets are said to be adjacent if they differ by one item.

Definition 2.2 (**(ε, δ)-DP).** *Let $\varepsilon \geq 0$, $\delta \in [0, 1]$. A randomized algorithm $\mathcal{M} : \mathcal{X}^m \to \mathcal{Y}$ satisfies (ε, δ)-DP if for any* adjacent *datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^m$ and subset $S \subseteq \mathcal{Y}$, we have*

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(\mathcal{D}') \in S\right] + \delta. \qquad (2)$$

We consider the server to be *honest-but-curious*, i.e., it follows the prescribed algorithm correctly, but may try to infer sensitive information about the workers' datasets. Thus, the workers must enforce privacy locally at their end. We assume that the server can only query the dataset of a worker w_i through a dedicated communication channel, and that there is no direct communication between the workers. Hence, for privacy in this context, we require the communications between the server and each honest worker to satisfy the criterion of DP in (2). In our context, we formalize this property below, inspired from (Smith et al., 2017).

Definition 2.3 (**(ε, δ)-distributed DP).** *Let $\varepsilon \geq 0$, $\delta \in [0, 1]$. Consider a randomized* distributed *algorithm $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathcal{Y}$. Let Z_i be a function that outputs the transcript of communications between the server and worker w_i during the execution of $\mathcal{A}$. Algorithm $\mathcal{A}$ is said to satisfy (ε, δ)-*distributed DP *if for all $i \in \mathcal{H}$, Z_i satisfies (ε, δ)-DP with respect to the dataset held by worker w_i.*

The above criterion of distributed DP reduces to *local DP* (Kasiviswanathan et al., 2011; Duchi et al., 2013) when each local dataset comprises a single item (i.e., $m = 1$). Moreover, an algorithm satisfying (ε, δ)-distributed DP may be fully *interactive*, i.e., the queries made to the workers by the server may share arbitrary dependence (Kasiviswanathan et al., 2011). Hereafter, a distributed algorithm satisfying (ε, δ)-distributed DP is simply said to be (ε, δ)-DP.

2.3. Assumptions

Our results are derived under standard assumptions. First, we recall that data heterogeneity can be modeled following the assumption below (Karimireddy et al., 2020; 2022).

Assumption 2.1 (Bounded heterogeneity)**.** *There exists $G < \infty$ such that for all $\theta \in \mathbb{R}^d$,*

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\| \nabla \mathcal{L}(\theta; \mathcal{D}_i) - \nabla \mathcal{L}_{\mathcal{H}}(\theta) \right\|^2 \leq G^2.$$

To present the convergence guarantees of SAFE-DSHB, we make the following standard assumption on the variance of stochastic gradients (Bottou et al., 2018).

Assumption 2.2 (Bounded variance)**.** *There exists $\sigma < \infty$ such that for each honest worker $w_i, i \in \mathcal{H}$, and all $\theta \in \mathbb{R}^d$,*

$$\frac{1}{m} \sum_{x \in \mathcal{D}_i} \left\| \nabla_{\theta} \ell(\theta; x) - \nabla \mathcal{L}(\theta; \mathcal{D}_i) \right\|^2 \leq \sigma^2.$$

Additionally, we also assume the point-wise gradients to be bounded, as usually done when analyzing differentially private ML algorithms to circumvent the complications due to clipping (Agarwal et al., 2018; Noble et al., 2022).

Assumption 2.3 (Bounded gradient)**.** *There exists $C < \infty$ such that for all $\theta \in \mathbb{R}^d$, $i \in \mathcal{H}$, and $x \in \mathcal{D}_i$,*

$$\left\| \nabla \ell(\theta; x) \right\| \leq C.$$

3. Lower Bound

We now prove our lower bound on the error incurred by a (f, ϱ)-robust distributed algorithm, when ensuring (ε, δ)-DP.

The main result is given in Theorem 3.1, whose full proof is deferred to Appendix A.5. To give insights about the proof, we detail three separate cases in sections 3.1, 3.2 and 3.3 where we respectively study $f = 0$, $f \geq 1$ but no privacy is enforced, and the adversarial setting $f \geq 1$ with privacy.

Theorem 3.1. *Let $\mathcal{X} = \mathbb{R}^d$, $\ell = \|\cdot\|^2$, $n \geq 3$, $0 \leq f < n/2$, $m \geq 1$, and $\varepsilon, \delta \in (0, 1)$. Consider arbitrary datasets $\mathcal{D}_1, \ldots, \mathcal{D}_n \in \mathcal{X}^m$ such that Assumption 2.1 is satisfied with $G \geq 1$. Let $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathbb{R}^d$ be an (ε, δ)-DP distributed algorithm. Assume that $\varepsilon \leq 1/4\sqrt{2n \ln(m+1)}$, and that $2^{-m^{1-\gamma}} \leq n\delta \leq 1/8m^{1+\gamma}$ for some $\gamma \in (0, 1)$. For any $\varrho \leq \frac{f+1}{100(n-f)}$, if $\mathcal{A}$ is (f, ϱ)-robust, then*

$$\varrho = \widetilde{\Omega}\left(\frac{d}{\varepsilon^2 n m^2} + \frac{f}{n} \cdot \frac{1}{\varepsilon^2 m^2} + \frac{f}{n} \cdot G^2\right).$$

Comparison with prior work. Our lower bound generalizes that of the non-adversarial centralized case. Specifically, specializing our lower bound to the case $n = 1$ yields the bound $\Omega\left(\frac{d}{\varepsilon^2 m^2}\right)$, which corresponds to the lower bound from centralized private ERM (Theorem V.5, Bassily et al. (2014))[4]. Second, we improve over a result from the *non-adversarial* private distributed learning literature (Theorem D.3, Lowy & Razaviyayn (2023)), where a similar lower bound is shown. While we consider distributed algorithm $\mathcal{A}$ as a black-box verifying (ε, δ)-DP (as per Definition 2.3), the mentioned work imposes additional structure on $\mathcal{A}$ by assuming it to be round-based and to satisfy *compositionality*, which essentially abstracts the class of round-based algorithms whose DP guarantees can be computed from advanced composition. Moreover, as the number of data points per worker m is typically greater than the number of workers n, our condition $\varepsilon = \mathcal{O}(1/\sqrt{n \log m})$ is arguably weaker than $\varepsilon = \mathcal{O}(1/m)$ in (Lowy & Razaviyayn, 2023).

Discussion on assumptions. The assumptions on $\varepsilon, \delta, \varrho$ are only needed to use the lower bound from (Steinke & Ullman, 2016), which additionally features the $\log(1/\delta)$ factor. One could use the same proof technique as in (Bassily et al., 2014) and remove these assumptions, at the expense of loosening the bound, e.g. an additional $\log m$ factor in the denominator of the first term appears.

3.1. Case I: Non-adversarial Setting

In this particular case, we assume all the workers to be honest, i.e., $f = 0$. However, the algorithm satisfies (ε, δ)-distributed DP. We show the following result.

Proposition 3.1. *Let $n, m \geq 1$, and $\varepsilon, \delta \in (0, 1)$. Consider $\mathcal{X} = \{\pm \frac{1}{\sqrt{d}}\}^d$ and $\ell = \|\cdot\|^2$. Consider an arbitrary (ε, δ)-DP distributed algorithm $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathbb{R}^d$. Assume that $\varepsilon \leq 1/4\sqrt{2n \ln(m+1)}$ and that $2^{-m^{1-\gamma}} \leq n\delta \leq 1/8m^{1+\gamma}$ for some $\gamma \in (0, 1)$. For any $\varrho \leq 1/100$, if $\mathcal{A}$ is $(0, \varrho)$-robust, then*

$$\varrho = \Omega\left(\frac{d}{\varepsilon^2 n m^2}\right).$$

Sketch of proof. We consider the quadratic loss function. We derive a *centralized* DP algorithm $\mathcal{M}$ from $\mathcal{A}$, and then reduce to private estimation of one-way marginals (Steinke & Ullman, 2016). Algorithm $\mathcal{M}$ runs $\mathcal{A}$ on n copies of the same dataset $\mathcal{D} \in \mathcal{X}^m$. Thus, $\mathcal{M}$ inherits the error guarantee ϱ from $\mathcal{A}$ on estimating the average of $\mathcal{D}$, but with a weaker $(\varepsilon_n, \delta_n)$-DP guarantee, due to the composition of n adaptive (ε, δ)-DP queries (since $\mathcal{A}$ can query each of the n copies of $\mathcal{D}$ up to (ε, δ)-DP budget). Using the centralized DP lower bound from (Steinke & Ullman, 2016), we have $\varrho = \Omega(d \log(1/\delta_n)/\varepsilon_n^2 m^2)$. We bound ε_n and δ_n via *advanced composition* (Dwork et al., 2014) as follows: $\varepsilon_n = \mathcal{O}(\varepsilon\sqrt{n \log(1/\delta')})$ (provided that ε is small enough) and $\delta_n \leq n\delta + \delta'$, where δ' is carefully chosen to ensure that $\log(1/\delta_n)/\log(1/\delta') = \Omega(1)$ (provided δ is small enough). Substituting the above values of ε_n and δ_n in the above lower bound on ϱ proves the proposition. □

3.2. Case II: No Privacy

Finally, we adapt the lower bound from robust distributed ML (Karimireddy et al., 2022) to our robustness definition (Definition 2.1) in Proposition 3.2 below.

Proposition 3.2. *Let Assumption 2.1 hold. Let $n \geq 1$, $1 \leq f < n/2$, and $\nu = \frac{16f(n-2f)}{(n-f)^2}$. Consider $\mathcal{X} = \{\pm \frac{G}{\sqrt{\nu d}}\}^d$ and $\ell = \|\cdot\|^2$. If a distributed algorithm is (f, ϱ)-robust, then*

$$\varrho = \Omega\left(\frac{f}{n} \cdot G^2\right).$$

3.3. Case III: Adversarial Setting

We now state, in Proposition 3.3 below, the part of our bound where privacy and robustness are coupled.

Proposition 3.3. *Let $n \geq 3$, $1 \leq f < n/2$, $m \geq 1$, $\varepsilon, \delta \in (0, 1)$, and $\nu = \frac{16f(n-2f)}{(n-f)^2}$. Consider $\mathcal{X} = \{\pm \frac{1}{\sqrt{d}}\}^d \cup \{\pm \frac{1}{\sqrt{\nu d}}\}^d$ and $\ell = \|\cdot\|^2$. Consider any (ε, δ)-DP distributed algorithm $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathbb{R}^d$. Assume that $2^{-o(m)} \leq \delta \leq 1/m^{1+\Omega(1)}$. For any $\varrho \leq \frac{f+1}{100(n-f)}$, if $\mathcal{A}$ is (f, ϱ)-robust, then*

$$\varrho = \Omega\left(\frac{f+1}{n-f} \cdot \frac{\log(1/\delta)}{\varepsilon^2 m^2}\right).$$

Sketch of proof. We consider the quadratic loss function, and reduce to the case $d = 1$ with a careful choice of datasets. We derive a *centralized* DP algorithm $\mathcal{M}$

[4] Notice that the loss function in (Bassily et al., 2014) is not divided by the number of samples m.

from $\mathcal{A}$, and then reduce to private estimation of one-way marginals (Steinke & Ullman, 2016). Algorithm $\mathcal{M}$ runs $\mathcal{A}$ on input dataset $\mathcal{D} \in \mathcal{X}^m$ together with the remaining $n-1$ datasets crafted as follows: f 'adversarial' datasets are filled with -1, while $n-f-1$ 'honest' datasets are filled with $+1$. This ensures that, in all cases, $\mathcal{M}$ estimates the average of $\mathcal{D}$ better than at least an f-sized minority of datasets. Therefore, as $\mathcal{A}$ guarantees error ϱ on estimating the *average of every group* of $n-f$ datasets' averages (by Definition 2.1), we can bound the error of estimating the average of $\mathcal{D}$ by $\tilde{\varrho} = \Theta(\frac{n-f}{f+1}\varrho)$. We conclude by applying the aforementioned DP lower bound to $\mathcal{M}$, which is (ε, δ)-DP and ensures error $\tilde{\varrho}$ in estimating the average of $\mathcal{D}$. □

4. Our Algorithm: SAFE-DSHB

We prove in this section that our lower bound is tight. Specifically, we present a new distributed algorithm, SAFE-DSHB, which yields a matching upper bound. Upon describing SAFE-DSHB in Section 4.1, we analyze its privacy in Section 4.2 and convergence guarantees in Section 4.3 for smooth *strongly convex* and *non-convex* loss functions.

4.1. Description of SAFE-DSHB

Similar to DSGD, SAFE-DSHB is an iterative algorithm where the server initiates each iteration (or step) $t \geq 0$ by broadcasting its current model parameter vector θ_t to all the workers. The initial parameter vector θ_0 is chosen arbitrarily by the server. Upon receiving θ_t from the server, each honest worker w_i samples a mini-batch $S_t^{(i)}$ of $b \leq m$ data points randomly from its local dataset $\mathcal{D}_i$ *without replacement*. Then, w_i computes the gradients $\nabla \ell(\theta_t; x)$ for all $x \in S_t^{(i)}$, clips each of them using a threshold value C and averages the clipped gradients to obtain a gradient estimate $g_t^{(i)}$. Specifically,

$$g_t^{(i)} = \frac{1}{b} \sum_{x \in S_t^{(i)}} \nabla \ell(\theta_t; x) \cdot \min \left\{1, \frac{C}{\|\nabla \ell(\theta_t; x)\|}\right\}.$$

To protect the privacy of its data, w_i then obfuscates $g_t^{(i)}$ with Gaussian noise to obtain $\tilde{g}_t^{(i)}$, i.e.,

$$\tilde{g}_t^{(i)} = g_t^{(i)} + \xi_t^{(i)} \; ; \quad \xi_t^{(i)} \sim \mathcal{N}\left(0, \sigma_{\mathrm{DP}}^2 I_d\right),$$

where I_d denotes the identity matrix of dimension $d \times d$, and $\mathcal{N}\left(0, \sigma_{\mathrm{DP}}^2 I_d\right)$ denotes a d-dimensional Gaussian distribution with mean 0 and covariance $\sigma_{\mathrm{DP}}^2 I_d$. Finally, w_i uses this noisy gradient to update its local *Polyak's momentum* (Polyak, 1964) denoted by $m_t^{(i)}$, which is then sent to the server. Specifically, for $t \geq 1$,

$$m_t^{(i)} = \beta_{t-1} m_{t-1}^{(i)} + (1 - \beta_{t-1}) \tilde{g}_t^{(i)},$$

where $m_0^{(i)} = 0$ by convention, and $\beta_t \in [0, 1]$ is referred to as the momentum coefficient. Recall that if worker w_i is adversarial, then it may send an arbitrary value for its momentum $m_t^{(i)}$. Upon receiving the local momentums from all the workers, the server aggregates them using F to obtain $R_t = F(m_t^{(1)}, \dots, m_t^{(n)})$. Finally, the server updates the model θ_t to

$$\theta_{t+1} = \theta_t - \gamma_t \, R_t$$

where $\gamma_t \geq 0$ is the learning rate at step t. The above procedure is repeated for a total of T steps, after which the server outputs $\hat{\theta}$ which is sampled uniformly from the set $\{\theta_0, \dots, \theta_{T-1}\}$. The complete learning procedure is summarized in Algorithm 1.

Algorithm 1 SAFE-DSHB

```
Initialization: Initial model $\theta_0$, initial momentum $m_0^{(i)} = 0$ for each honest worker $w_i$, robust aggregation $F$, DP noise $\sigma_{\mathrm{DP}}$, batch size $b$, clipping threshold $C$, learning rates $\{\gamma_t\}$, momentum coefficients $\{\beta_t\}$, and total number of steps $T$.
for $t = 0 \dots T-1$ do
    Server broadcasts $\theta_t$ to all workers.
    for every honest worker $w_i, i \in \mathcal{H}$, in parallel do
        Sample a mini-batch $S_t^{(i)}$ of size $b$ at random from $\mathcal{D}_i$ without replacement.
        Clip and average the mini-batch gradients:
            $g_t^{(i)} = \frac{1}{b} \sum_{x \in S_t^{(i)}} \mathbf{Clip}\left(\nabla \ell(\theta_t; x); C\right),$
        where $\mathbf{Clip}(g; C) := g \cdot \min\{1, C/\|g\|\}$.
        Add noise to the mini-batch average gradient:
            $\tilde{g}_t^{(i)} = g_t^{(i)} + \xi_t^{(i)}; \; \xi_t^{(i)} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d).$
        Send $m_t^{(i)} = \beta_{t-1} m_{t-1}^{(i)} + (1 - \beta_{t-1}) \tilde{g}_t^{(i)}$.
    end for
    Server aggregates: $R_t = F(m_t^{(1)}, \dots, m_t^{(n)})$.
    Server updates the model: $\theta_{t+1} = \theta_t - \gamma_t R_t$.
end for
return $\hat{\theta}$ uniformly sampled from $\{\theta_0, \dots, \theta_{T-1}\}$.
```

4.2. Privacy of SAFE-DSHB

We present below the DP guarantee of SAFE-DSHB. To state closed-form expressions, we will assume that the batch size b is sufficiently small compared to m the number of data points per worker. This assumption is only made for pedagogical reasons, but is not necessary for the privacy analysis to hold. In particular, the expressions that result from removing this assumption are difficult to read and interpret (Wang et al., 2019a). We defer the full DP analysis without this assumption to Appendix C.

Theorem 4.1. *Consider Algorithm 1. Let $\varepsilon > 0, \delta \in (0, 1)$ be such that $\varepsilon \leq \log(1/\delta)$. There exists a constant $k > 0$ such that, for a sufficiently small batch size b, when $\sigma_{\mathrm{DP}} \geq k \cdot \frac{2C}{b} \max\left\{1,\, \frac{b\sqrt{T\log(1/\delta)}}{m\varepsilon}\right\}$, Algorithm 1 is (ε, δ)-DP.*

4.3. Convergence of SAFE-DSHB

To present the convergence of SAFE-DSHB we first introduce below a criterion, namely (f, κ)*-robust averaging*, for an aggregation rule F that proves crucial in our analysis.

Definition 4.1. *Let $n \geq 1$, $0 \leq f < n/2$ and $\kappa \geq 0$. An aggregation rule F is said to be (f, κ)*-robust averaging *if for any vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, and any set $S \subseteq \{1, \ldots, n\}$ of size $n - f$, the output $\hat{x} = F(x_1, \ldots, x_n)$ satisfies*

$$\|\hat{x} - \overline{x}_S\|^2 \leq \kappa \cdot \lambda_{\max}\left(\frac{1}{|S|}\sum_{i \in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right),$$

where $\overline{x}_S := \frac{1}{|S|}\sum_{i\in S} x_i$ and $\lambda_{\max}$ denotes the maximum eigenvalue. We refer to κ as the robustness coefficient of F.

Comparison to prior work. Our robustness criterion is stronger than existing ones: (f, κ)-robustness (Allouah et al., 2023), (f, λ)-resilience (Farhadkhani et al., 2022) and $(c, \delta_{\max})$-ARAgg (Karimireddy et al., 2022). The last two works bound the error with the diameter of honest inputs, i.e., maximum squared pairwise distance. The latter is greater than the empirical variance (bound used in (f, κ)-robustness (Allouah et al., 2023)), which itself is greater than the maximum eigenvalue of the empirical covariance (that we use) in high-dimensional spaces (i.e., $d > 1$). In fact, the tight analysis of aggregation functions (e.g., trimmed mean, Krum) conducted in (Allouah et al., 2023) through the lens of (f, κ)-robustness directly implies our (f, κ')-robust averaging criterion, with $\kappa' \leq d \cdot \kappa$. However, aggregation rules that are optimal w.r.t. (f, κ)-robustness (Allouah et al., 2023) may be suboptimal in our context, as we need to suppress the dimension dependence of κ for our tight bounds.

Tighter heterogeneity metric. We introduce a new metric G_{cov} for quantifying the heterogeneity between the local gradients of honest workers' loss functions, which is arguably tighter than G defined in Section 3.2. Specifically,

$$G_{\mathrm{cov}}^2 := \sup_{\theta \in \mathbb{R}^d} \sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} \langle v,\, \nabla\mathcal{L}(\theta; \mathcal{D}_i) - \nabla\mathcal{L}_{\mathcal{H}}(\theta)\rangle^2 .$$

Note that G_{cov}^2 above represents an upper bound on the spectral norm of the empirical covariance of honest gradients, which is smaller than their empirical variance G^2. Moreover, if the gradients have a well-conditioned empirical covariance, then G_{cov} has weaker dependence on d.

We state our convergence result below in Theorem 4.2. Essentially, we analyze the convergence of SAFE-DSHB with an (f, κ)-robust averaging aggregation F, under assumptions 2.2 and 2.3, for smooth strongly convex and non-convex loss functions. We use the following notation:

$$\begin{aligned}&\mathcal{L}_* = \inf_{\theta \in \mathbb{R}^d} \mathcal{L}_{\mathcal{H}}(\theta),\ \mathcal{L}_0 = \mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*,\ a_1 = 240,\\ &a_2 = 480,\ a_3 = 5760,\ \text{ and }\ a_4 = 270.\end{aligned} \tag{3}$$

Theorem 4.2. *Suppose that assumptions 2.2 and 2.3 hold true, and that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Let F satisfy the condition of (f, κ)-robust averaging. We let*

$$\overline{\sigma}^2 = \frac{\sigma_b^2 + d\sigma_{\mathrm{DP}}^2}{n - f} + 4\kappa\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2\left(1 + \frac{d}{n-f}\right)\right),$$

where $\sigma_b^2 = 2(1 - \frac{b}{m})\frac{\sigma^2}{b}$. Consider Algorithm 1 with $T \geq 1$, the learning rates γ_t and momentum coefficients β_t specified below. We prove that the following holds, where the expectation $\mathbb{E}[\cdot]$ is over the randomness of the algorithm.

1. ***Strongly convex:** Assume that $\mathcal{L}_{\mathcal{H}}$ is μ-strongly convex. If $\gamma_t = \frac{10}{\mu(t + a_1\frac{L}{\mu})}$ and $\beta_t = 1 - 24L\gamma_t$ then*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_T) - \mathcal{L}_*\right] \leq \frac{4a_1\kappa G_{\mathrm{cov}}^2}{\mu} + \frac{2a_1^2 L\overline{\sigma}^2}{\mu^2 T} + \frac{2a_1^2 L^2\mathcal{L}_0}{\mu^2 T^2}.$$

2. ***Non-convex:** If $\gamma = \min\left\{\frac{1}{24L},\, \frac{\sqrt{a_4\mathcal{L}_0}}{2\overline{\sigma}\sqrt{a_3 LT}}\right\}$ and $\beta_t = 1 - 24L\gamma$ then*

$$\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{H}}(\hat{\theta})\|^2\right] \leq a_2\kappa G_{\mathrm{cov}}^2 + \frac{\sqrt{a_3 a_4 L\mathcal{L}_0}\overline{\sigma}}{\sqrt{T}} + \frac{a_4 L\mathcal{L}_0}{T}.$$

Sketch of proof. We show that at each step t, the descent $\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t)$ can be bounded from above. Doing so is however non-trivial, as one needs to consider two conflicting effects: (i) the drift between honest momentums, and (ii) the deviation between the average honest momentum and the true gradient. To control this trade-off, we use increasing momentum coefficients and decreasing learning rates, and introduce an adapted *Lyapunov function* V_t. Ignoring the constants, the function can be written as follows:

$$V_t := (t + K)^2 \cdot \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_t) - \mathcal{L}_* + \frac{1}{L}\delta_t + \frac{\kappa}{L}\Delta_t\right],$$

where $\delta_t := \|\overline{m}_t - \nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\|^2$ represents the deviation of the momentum from the true gradient, $\Delta_t := \lambda_{\max}\left(\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}(m_t^{(i)} - \overline{m}_t)(m_t^{(i)} - \overline{m}_t)^\top\right)$ represents the drift between the honest momentums, and $K := \frac{L}{\mu}$ denotes the condition number of $\mathcal{L}_{\mathcal{H}}$. □

Remark 4.3. Our strongly convex upper bound also holds true for the larger class of smooth μ-*PL* functions (Karimi et al., 2016), which includes some non-convex functions.

Comparison to prior work. Our convergence rate in $\mathcal{O}\left(\frac{1}{T}\right)$ for strongly convex losses is optimal in the non-adversarial and privacy-free setting (Agarwal et al., 2009). We improve over the state-of-the-art strongly convex analysis (Data & Diggavi, 2021), without privacy, which features a suboptimal excess term proportional to the stochastic noise $\overline{\sigma}^2$. Essentially, we remove this dependency on $\overline{\sigma}^2$ thanks to the use of momentum, although our convergence rate is in $\mathcal{O}\left(\frac{1}{T}\right)$ instead of being exponential as in (Data & Diggavi, 2021). In fact, making $\overline{\sigma}^2$ vanish at a rate $\frac{1}{T}$ is crucial in our setting, as the DP noise σ_{DP}^2 scales with T (Theorem 4.1). We also improve over the state-of-the-art non-convex analysis (Farhadkhani et al., 2022). Namely, our analysis features a tighter characterization of the data heterogeneity G_{cov}, instead of the traditional heterogeneity metric G.

5. Tight Upper Bound

We present a new aggregation rule named *SMEA* (Smallest Maximum Eigenvalue Averaging) in Section 5.1, and show that it yields a tight upper bound in Section 5.2.

5.1. Robust Aggregation: SMEA

Consider a set of n vectors $x_1, \dots, x_n$. Let S^* be an arbitrary subset of $[n]$ of size $n - f$ with the smallest empirical *maximum eigenvalue*, i.e.,

$$S^* \in \underset{\substack{S \subseteq [n] \\ |S| = n-f}}{\operatorname{argmin}} \lambda_{\max}\left(\frac{1}{|S|}\sum_{i \in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right).$$

SMEA outputs the average of the inputs in S^*, i.e.,

$$\mathrm{SMEA}(x_1, \dots, x_n) \coloneqq \frac{1}{|S^*|}\sum_{i \in S^*} x_i.$$

Note that SMEA draws inspiration from the *minimum diameter averaging* method (El Mhamdi et al., 2018), which itself is reminiscent of the *minimal volume ellipsoid* method (Rousseeuw, 1985). We show that our aggregation rule satisfies the criterion of (f, κ)-robust averaging.

Proposition 5.1. *Let $f < n/2$. SMEA is (f, κ)-robust averaging with*

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

Proposition 5.1 implies that, when $n \geq (2 + \eta)f$ for some constant $\eta > 0$, SMEA satisfies (f, κ)-robust averaging with $\kappa = \mathcal{O}(f/n)$. Importantly, SMEA satisfies this high-dimensional robustness property while being agnostic to the statistical properties of the valid inputs, knowledge of which is key in designing efficient robust estimators (Diakonikolas et al., 2017; Steinhardt et al., 2018) (see Appendix B.2).

Computational complexity. However, as SMEA involves computing the maximum eigenvalue of d-dimensional symmetric matrices, which is in $\mathcal{O}(d^3)$, the worst-case computational complexity of SMEA is $\mathcal{O}\big(\binom{n}{f} \cdot d^3\big)$, which is exponential in f. This shortcoming of our method should be addressed in the future.

5.2. Upper Bound

Upon combining the results in theorems 4.1, 4.2, Proposition 5.1, and ignoring the vanishing terms in T, we obtain Corollary 5.1 that quantifies the privacy-robustness-utility trade-off of SAFE-DSHB using the SMEA aggregation rule.

Corollary 5.1. *Consider Algorithm 1 with aggregation $F = \mathrm{SMEA}$, under the strongly convex setting of Theorem 4.2. Suppose that assumptions 2.1, 2.2, 2.3 hold, and that $n \geq (2 + \eta)f$, for some absolute constant $\eta > 0$. Let $\varepsilon > 0, \delta \in (0, 1)$ be such that $\varepsilon \leq \log(1/\delta)$. Then, there exists a constant $k > 0$ such that, if $\sigma_{\mathrm{DP}} = k \cdot {}^{2C}\!/_{b} \max\{1, {}^{b\sqrt{T\log(1/\delta)}}\!/_{\varepsilon m}\}$, then Algorithm 1 is (ε, δ)-DP and (f, ϱ)-robust where*

$$\varrho = \mathcal{O}\left(\frac{d\log(1/\delta)}{\varepsilon^2 n m^2} + \frac{f}{n} \cdot \frac{\log(1/\delta)}{\varepsilon^2 m^2} + \frac{f}{n}G^2\right).$$

Tightness. Our upper bound is tight, in the sense that it matches the lower bound, up to the logarithmic factor $\log(1/\delta)$ in the first term. We believe that it is not possible to improve upon our upper bound in general, but rather that it may be possible to improve our lower bound in Proposition 3.1, by including the factor $\log(1/\delta)$. This could be done, for example, by assuming the stronger Rényi DP property (Mironov, 2017), satisfied by the Gaussian mechanism, instead of relying on the advanced composition theorem.

6. Conclusions and Future Work

Applying machine learning in sensitive public domains requires algorithms that protect data privacy, while being robust to faults and adversarial behaviors. We present the first tight analysis of the error incurred by any distributed ML algorithm ensuring robustness to adversarial workers and differential privacy for honest machines' data against any other curious entity. Our algorithm SAFE-DSHB yields a tight upper bound for the class of smooth strongly convex problems, up to a logarithmic factor. Proving a tighter lower bound on the privacy cost, featuring the usual $\log(1/\delta)$ factor, is an appealing goal. Proving similar bounds for the non-strongly convex class is also of interest. Also, in Appendix E, we conduct small-scale experiments showing encouraging results using our aggregation rule SMEA (as well as other aggregation rules). Yet, while SMEA is simple and agnostic to the statistical properties of honest data, it has a high computational complexity. Deploying it on larger scale systems goes through designing variants with lower complexity, and this is also an interesting research direction.

Acknowledgements

This work was supported in part by SNSF grants 200021_200477 and 200021_182542. The authors are thankful to the anonymous reviewers for their constructive comments.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Acharya, J., Sun, Z., and Zhang, H. Robust testing and estimation under manipulation attacks. In *International Conference on Machine Learning*, pp. 43–53. PMLR, 2021.

Agarwal, A., Wainwright, M. J., Bartlett, P., and Ravikumar, P. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

Agarwal, N., Suresh, A. T., Yu, F. X. X., Kumar, S., and McMahan, B. cpsgd: Communication-efficient and differentially-private distributed sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

Allen-Zhu, Z., Ebrahimianghazani, F., Li, J., and Alistarh, D. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*, 2020.

Allouah, Y., Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1232–1300. PMLR, 2023.

Arora, R., Bassily, R., González, T., Guzmán, C., Menart, M., and Ullah, E. Faster rates of convergence to stationary points in differentially private optimization. *arXiv preprint arXiv:2206.00846*, 2022.

Ashtiani, H. and Liaw, C. Private and polynomial time algorithms for learning gaussians and beyond. In *Conference on Learning Theory*, pp. 1075–1076. PMLR, 2022.

Baruch, M., Baruch, G., and Goldberg, Y. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, 8-14 December 2019, Long Beach, CA, USA*, 2019.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Bertsekas, D. and Tsitsiklis, J. *Parallel and distributed computation: numerical methods*. Athena Scientific, 2015.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 119–129. Curran Associates, Inc., 2017.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Bun, M., Ullman, J., and Vadhan, S. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 1–10, 2014.

Cheu, A., Smith, A., and Ullman, J. Manipulation attacks in local differential privacy. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 883–900. IEEE, 2021.

Chhor, J. and Sentenac, F. Robust estimation of discrete distributions under local differential privacy. In *International Conference on Algorithmic Learning Theory*, pp. 411–446. PMLR, 2023.

Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., and Das, A. Differential privacy-enabled federated learning for sensitive health data. *arXiv preprint arXiv:1910.02578*, 2019.

Data, D. and Diggavi, S. Byzantine-resilient high-dimensional sgd with local iterations on heterogeneous data. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2478–2488. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/data21a.html`.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M. a., Senior, A., Tucker, P., Yang, K., Le, Q., and Ng, A. Large scale distributed deep networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pp. 999–1008. PMLR, 2017.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, pp. 371–380, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585062. doi: 10.1145/1536414.1536466. URL https://doi.org/10.1145/1536414.1536466.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

El Mhamdi, E. M., Guerraoui, R., and Rouault, S. The hidden vulnerability of distributed learning in Byzantium. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3521–3530. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/mhamdi18a.html.

EU. Regulation (eu) 2016/679 of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. Technical report, European Parliament and European Council, 2016.

Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. Byzantine machine learning made easy by resilient averaging of momentums. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6246–6283. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/farhadkhani22a.html.

Feng, J., Xu, H., and Mannor, S. Distributed robust learning, 2015.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813677. URL https://doi.org/10.1145/2810103.2813677.

Gadat, S., Panloup, F., and Saadane, S. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461 – 529, 2018. doi: 10.1214/18-EJS1395. URL https://doi.org/10.1214/18-EJS1395.

Georgiev, K. and Hopkins, S. B. Privacy induces robustness: Information-computation gaps and sparse mean estimation. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=g-OkeNXPy-X.

Guerraoui, R., Gupta, N., Pinot, R., Rouault, S., and Stephan, J. Differential privacy and Byzantine resilience in sgd: Do they add up? In *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, PODC'21, pp. 391–401, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385480. doi: 10.1145/3465084.3467919. URL https://doi.org/10.1145/3465084.3467919.

Gupta, N. and Vaidya, N. H. Fault-tolerance in distributed optimization: The case of redundancy. In *Proceedings of the 39th Symposium on Principles of Distributed Computing*, pp. 365–374, 2020.

Gupta, N., Liu, S., and Vaidya, N. Byzantine fault-tolerant distributed machine learning with norm-based comparative gradient elimination. In *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pp. 175–181. IEEE, 2021.

Hitaj, B., Ateniese, G., and Perez-Cruz, F. Deep models under the gan: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pp. 603–618, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349468. doi: 10.1145/3133956.3134012. URL https://doi.org/10.1145/3133956.3134012.

Hopkins, S. B., Kamath, G., and Majid, M. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1406–1417, 2022a.

Hopkins, S. B., Kamath, G., Majid, M., and Narayanan, S. Robustness implies privacy in statistical estimation. *arXiv preprint arXiv:2212.05015*, 2022b.

Hu, R., Guo, Y., Li, H., Pei, Q., and Gong, Y. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539, 2020.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Karimireddy, S. P., He, L., and Jaggi, M. Learning from history for Byzantine robust optimization. *International Conference On Machine Learning, Vol 139*, 139, 2021.

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=jXKKDEi5vJt`.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Lamport, L., Shostak, R., and Pease, M. The Byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401, jul 1982. ISSN 0164-0925. doi: 10.1145/357172.357176. URL `https://doi.org/10.1145/357172.357176`.

Li, L., Xu, W., Chen, T., Giannakis, G. B., and Ling, Q. RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1544–1551, 2019.

Li, M., Berrett, T. B., and Yu, Y. On robustness and local differential privacy. *arXiv preprint arXiv:2201.00751*, 2022.

Liu, S., Gupta, N., and Vaidya, N. H. Approximate Byzantine fault-tolerance in distributed optimization. In *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, PODC'21, pp. 379–389, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450385480. doi: 10.1145/3465084.3467902.

Liu, X., Kong, W., Kakade, S., and Oh, S. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34:3887–3901, 2021b.

Liu, X., Kong, W., and Oh, S. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pp. 1167–1246. PMLR, 2022.

Lowy, A. and Razaviyayn, M. Private federated learning without a trusted server: Optimal algorithms for convex losses. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=TVY6GoURrw`.

Lowy, A., Ghafelebashi, A., and Razaviyayn, M. Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*, pp. 5749–5786. PMLR, 2023.

Ma, X., Sun, X., Wu, Y., Liu, Z., Chen, X., and Dong, C. Differentially private Byzantine-robust federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 2022.

Melis, L., Song, C., Cristofaro, E. D., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pp. 691–706. IEEE, 2019. doi: 10.1109/SP.2019.00029. URL `https://doi.org/10.1109/SP.2019.00029`.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10110–10145. PMLR, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pauwels, E. Lecture notes: Statistics, optimization and algorithms in high dimension, 2020.

Phong, L. T., Aono, Y., Hayashi, T., Wang, L., and Moriai, S. Privacy-preserving deep learning: Revisited and enhanced. In Batten, L., Kim, D. S., Zhang, X., and Li, G. (eds.), *Applications and Techniques in Information Security*, pp. 100–110, Singapore, 2017. Springer Singapore. ISBN 978-981-10-5421-1.

Polyak, B. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5.

Rice, J. A. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

Rigollet, P. and Hütter, J.-C. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.

Rousseeuw, P. J. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8 (37):283–297, 1985.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.

Shokri, R., Stronati, M., and Shmatikov, V. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820, 2016.

Smith, A., Thakurta, A., and Upadhyay, J. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 58–77. IEEE, 2017.

Steinhardt, J. *Robust learning: Information theory and algorithms*. Stanford University, 2018.

Steinhardt, J., Charikar, M., and Valiant, G. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Steinke, T. and Ullman, J. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2016.

Su, L. and Vaidya, N. H. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM symposium on principles of distributed computing*, pp. 425–434, 2016.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019a.

Wang, Z., Mengkai, S., Zhang, Z., Song, Y., Wang, Q., and Qi, H. Beyond inferring class representatives: User-level privacy leakage from federated learning. pp. 2512–2520, 04 2019b. doi: 10.1109/INFOCOM.2019.8737416.

Xiang, M. and Su, L. β-stochastic sign sgd: A Byzantine resilient and differentially private gradient compressor for federated learning. *arXiv preprint arXiv:2210.00665*, 2022.

Xie, C., Koyejo, O., and Gupta, I. Generalized Byzantine-tolerant sgd, 2018.

Xie, C., Koyejo, O., and Gupta, I. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, pp. 83, 2019.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR, 2018.

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. Opacus: User-friendly differential privacy library in pytorch, 2021. URL `https://arxiv.org/abs/2109.12298`.

Zhao, B., Mopuri, K. R., and Bilen, H. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Zhu, B., Jiao, J., and Steinhardt, J. Robust estimation via generalized quasi-gradients. *Information and Inference: A Journal of the IMA*, 11(2):581–636, 2022.

Zhu, H. and Ling, Q. Bridging differential privacy and Byzantine-robustness via model aggregation. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2427–2433. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/337. URL `https://doi.org/10.24963/ijcai.2022/337`. Main Track.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 14774–14784. Curran Associates, Inc., 2019.

Organization of the Appendix

- Appendix A contains the proof of our lower bounds.
  - Appendix A.1 reviews a known lower bound on estimating the average of one-way marginals under DP.
  - Appendix A.2 contains the proof of the lower bound due to privacy alone in Proposition 3.1.
  - Appendix A.3 contains the proof of the lower bound due to robustness alone in Proposition 3.2.
  - Appendix A.4 contains the proof of the coupled lower bound in Proposition 3.3.
  - Appendix A.5 contains the proof of the final lower bound in Theorem 3.1.
- Appendix B contains proofs of claims related to (f, κ)-robust averaging and SMEA.
  - Appendix B.1 contains the analysis of SMEA in Proposition 5.1.
  - Appendix B.2 discusses the filter algorithm introduced in (Diakonikolas et al., 2017), and its robustness property.
- Appendix C contains the privacy analysis of SAFE-DSHB.
  - Appendix C.1 recalls preliminary results on DP and Rényi DP.
  - Appendix C.2 presents the proof of the privacy guarantee in Theorem 4.1.
- Appendix D contains the convergence analysis of SAFE-DSHB and the upper bound.
  - Appendix D.1 presents the proof outline for the convergence result presented in Theorem 4.2
  - Appendix D.2 presents the proof of Theorem 4.2
  - Appendix D.3 presents the proof of the upper bound presented in Corollary 5.1.
  - Appendix D.4 presents an upper bound for the non-convex case.
  - Appendix D.5 presents proofs for supporting lemmas used in the proof of Theorem 4.2
- Appendix E contains the experimental setup and results of our empirical evaluation.
  - Appendix E.1 describes our experimental setup.
  - Appendix E.2 contains our empirical results.

A. Lower Bounds

In Section A.1, we recall lower bounds on centralized private algorithms. We then extend these results to distributed private algorithms. We start by the lower bound due to privacy alone in Section A.2. Next, we show the lower bound due to robustness alone in Section A.3. We then show the lower bound due to the privacy-robustness tradeoff in Section A.4. Finally, we merge the previous results to show the final lower bound in Section A.5.

A.1. Lower Bound in Centralized DP

We recall lower bound (Steinke & Ullman, 2016) on the error incurred by *centralized* differentially private mechanisms for estimating d-dimensional one-way marginals; i.e., the average of rows of a dataset. Recall that Steinke & Ullman prove a sharper bound (by factor $\log{(1/\delta)}$) than Bassily et al., whose work is based on lower bounds using fingerprinting codes (Bun et al., 2014). We recall below the main lower bound from (Steinke & Ullman, 2016).

Lemma A.1 (Theorem 1.1, Steinke & Ullman (2016))**.** *Let $m, d \geq 1, \varepsilon, \delta \in (0,1)$ and $\mathcal{X} = \{\pm 1\}^d, \mathcal{Y} = [\pm 1]^d$. Consider any (ε, δ)-DP centralized algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$. Assume that $\delta \leq 1/m^{1+\Omega(1)}$ and that $\delta \geq 2^{-o(m)}$. Let $\mathcal{D} \in \mathcal{X}^m$ and $\overline{\mathcal{D}}$ denote the average of records of $\mathcal{D}$. For any $\varrho \leq 1/10$ such that for every $\mathcal{D} \in \mathcal{X}^m$, $\mathbb{E}\left[\left\|\mathcal{M}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_1\right] \leq d\varrho$, we have:*

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

Observe in Lemma A.1 that the lower bound assumption $\delta \leq 1/m^{1+\Omega(1)}$ is slightly more restrictive than the folklore assumption $\delta = o(1/m)$ (Dwork et al., 2014). The latter ensures that (ε, δ)-DP precludes some intuitively non-private algorithms, e.g., when $\delta \geq 1/m$, the algorithm that returns $\lfloor m\delta \rfloor$ random elements of the dataset is $(0, \delta)$-DP.

A.2. Case I: Non-adversarial Setting

We prove below our lower bound due to privacy, stated in Proposition 3.1.

Proposition 3.1. *Let $n, m \geq 1$, and $\varepsilon, \delta \in (0,1)$. Consider $\mathcal{X} = \{\pm\frac{1}{\sqrt{d}}\}^d$ and $\ell = \|\cdot\|^2$. Consider an arbitrary (ε, δ)-DP distributed algorithm $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathbb{R}^d$. Assume that $\varepsilon \leq 1/4\sqrt{2n \ln{(m+1)}}$ and that $2^{-m^{1-\gamma}} \leq n\delta \leq 1/8m^{1+\gamma}$ for some $\gamma \in (0,1)$. For any $\varrho \leq 1/100$, if $\mathcal{A}$ is $(0, \varrho)$-robust, then*

$$\varrho = \Omega\left(\frac{d}{\varepsilon^2 n m^2}\right).$$

Proof. Let $n, m, d \geq 1$, $\varepsilon, \delta \in (0,1)$, and $\varrho \leq 1/100$. Consider $\mathcal{X} = \left\{\pm 1/\sqrt{d}\right\}^d$ and $\ell = \|\cdot\|^2$. We consider an arbitrary distributed algorithm $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathbb{R}^d$ that satisfies (ε, δ)-distributed DP (see Definition 2.3), and $(0,\ \varrho)$-robustness (see Definition 2.1). We assume that $\varepsilon \leq 1/4\sqrt{2n \ln{(m+1)}}$ and that $2^{-m^{1-\gamma}} \leq n\delta \leq 1/8m^{1+\gamma}$ for some $\gamma \in (0,1)$.

Proof outline. We consider the centralized algorithm $\mathcal{M}$ which takes as input dataset $\mathcal{D} \in \mathcal{X}^m$ and executes $\mathcal{A}(\mathcal{D}_1, \ldots, \mathcal{D}_n)$ on n copies of $\mathcal{D}$, i.e., $\mathcal{D}_1 = \ldots = \mathcal{D}_n = \mathcal{D}$. Then, we derive the DP guarantee and utility of $\mathcal{M}$ using the facts that $\mathcal{A}$ satisfies (ε, δ)-distributed DP (see Definition 2.3) and $(0,\ \varrho)$-robustness, respectively. Finally, we apply the *centralized* DP lower bound on $\mathcal{M}$ (stated in Lemma A.1) to conclude the proof.

Privacy guarantee of $\mathcal{M}$. We first analyze the DP guarantees of $\mathcal{M}$ inherited from $\mathcal{A}$.

Recall from Definition 2.3 that, since $\mathcal{A}$ is (ε, δ)-DP, it can communicate with *each* database $\mathcal{D}_i$ subject to (ε, δ)-DP. Thus, when running $\mathcal{M}$, in the worst case, algorithm $\mathcal{A}$ may adaptively query the same database $\mathcal{D}$ a total of n times, subject to (ε, δ)-DP budget for each query. Therefore, $\mathcal{M}$ is $(\varepsilon_n, \delta_n)$-DP where $(\varepsilon_n, \delta_n)$ is the privacy guarantee resulting from composing (ε, δ)-DP across n adaptive queries. Thanks to the advanced composition theorem (Dwork et al., 2014), we obtain that, for any $\delta' \in (0,1)$,

$$\varepsilon_n = \varepsilon\sqrt{2n \ln{(1/\delta')}} + n\varepsilon(e^{\varepsilon} - 1), \quad \delta_n = n\delta + \delta'. \tag{4}$$

As $\varepsilon \in (0,1)$, we have $e^{\varepsilon} - 1 \leq 2\varepsilon$ and thus

$$\varepsilon_n \leq \varepsilon\sqrt{2n \ln{(1/\delta')}} + 2n\varepsilon^2. \tag{5}$$

We now set δ' as follows:

$$\delta' = \frac{1}{(m+1)^{1+\gamma}} \in (0,1). \tag{6}$$

We verify below the privacy conditions on $\mathcal{M}$ of Lemma A.1. We first prove that $\ln(1/\delta') \in [n\varepsilon^2, 1/16n\varepsilon^2)$, and then that $\varepsilon_n \leq 4\varepsilon\sqrt{n\ln(1/\delta')} < 1$.

Bound on $\ln(1/\delta')$*:* Since we assume $\varepsilon \leq 1/4\sqrt{2n\ln(m+1)}$ (with $m \geq 1$), we have

$$n\varepsilon^2 \leq 1/16 \leq 1/16n\varepsilon^2.$$

On the other hand, as $m \geq 1$, it follows from the expression (6) of δ' that $1/\delta' \geq 2$ and $\ln(1/\delta') \geq 1/4 \geq n\varepsilon^2$. Also, since $\varepsilon \leq 1/4\sqrt{2n\ln(m+1)}$ we have $\ln(m+1) \leq 1/32n\varepsilon^2$, and thus (because $\gamma \in (0,1)$) we have

$$\ln(1/\delta') = (1+\gamma)\ln(m+1) < 2\ln(m+1) \leq 1/16n\varepsilon^2.$$

This proves that

$$\ln(1/\delta') \in [n\varepsilon^2, 1/16n\varepsilon^2). \tag{7}$$

Bound on ε_n*:* Thanks to (7), we have $\ln(1/\delta') \geq n\varepsilon^2$. Thus, by taking square roots we have $\varepsilon\sqrt{n} \leq \sqrt{\ln(1/\delta')}$. Therefore, $n\varepsilon^2 \leq \varepsilon\sqrt{n\ln(1/\delta')}$. Then, using the bound on ε_n in (5), we obtain

$$\varepsilon_n \leq \varepsilon\sqrt{2n\ln(1/\delta')} + 2n\varepsilon^2 \leq \varepsilon\sqrt{2n\ln(1/\delta')} + 2\varepsilon\sqrt{n\ln(1/\delta')} \leq 4\varepsilon\sqrt{n\ln(1/\delta')}.$$

On the other hand, since we showed in (7) that $\ln(1/\delta') < 1/16n\varepsilon^2$, we have $4\varepsilon\sqrt{n\ln(1/\delta')} < 1$. This proves that

$$\varepsilon_n \leq 4\varepsilon\sqrt{n\ln(1/\delta')} < 1. \tag{8}$$

From (8), we have $\varepsilon_n \in (0,1)$. From (4), we have $\delta_n = n\delta + \delta'$. Thus, by assumption on δ and (6), the parameter δ_n satisfies both $\delta_n \geq n\delta \geq 2^{-m^{1-\gamma}} = 2^{-o(m)}$ and $\delta_n = n\delta + \delta' \leq 1/8m^{1+\gamma} + 1/(m+1)^{1+\gamma} = 1/m^{1+\Omega(1)}$.

Utility guarantees of $\mathcal{M}$. We now analyze the utility guarantees of $\mathcal{M}$, inherited from $\mathcal{A}$.

Let $\mathcal{D} \in \mathcal{X}^m$ be an arbitrary set of m points from the specified space $\mathcal{X} = \left\{\pm 1/\sqrt{d}\right\}^d$. Recall that $\mathcal{A}$ is assumed $(0, \varrho)$-robust. By Definition 2.1, for any $\mathcal{D}_1, \ldots, \mathcal{D}_n \in \mathcal{X}^m$, the output $\hat{\theta} = \mathcal{A}(\mathcal{D}_1, \ldots, \mathcal{D}_n)$ verifies

$$\varrho \geq \mathbb{E}\left[\mathcal{L}(\hat{\theta}; \mathcal{D}_1, \ldots, \mathcal{D}_n) - \inf_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta; \mathcal{D}_1, \ldots, \mathcal{D}_n)\right], \tag{9}$$

In this particular case, since $\mathcal{D}_1, \ldots, \mathcal{D}_n = \mathcal{D}$ and $\ell(\theta; x) := \left\|\theta - x\right\|^2$, we have for all $\theta \in \mathbb{R}^d$,

$$\mathcal{L}(\theta; \mathcal{D}_1, \ldots, \mathcal{D}_n) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{x \in \mathcal{D}_i}\left\|\theta - x\right\|^2 = \frac{1}{m}\sum_{x \in \mathcal{D}}\left\|\theta - x\right\|^2 = \mathcal{L}(\theta; \mathcal{D}). \tag{10}$$

We can rewrite the above upon applying the bias-variance decomposition: for any $x_1, \ldots, x_n$ we have $\frac{1}{n}\sum_{i=1}^{n}\left\|x_i - \overline{x}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|x_i\right\|^2 - \left\|\overline{x}\right\|^2$ where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Thus, denoting $\overline{\mathcal{D}} := \frac{1}{m}\sum_{x \in \mathcal{D}} x$, we can rewrite (10) as

$$\mathcal{L}(\theta; \mathcal{D}_1, \ldots, \mathcal{D}_n) = \mathcal{L}(\theta; \mathcal{D}) = \left\|\theta - \overline{\mathcal{D}}\right\|^2 + \frac{1}{m}\sum_{x \in \mathcal{D}}\left\|\overline{\mathcal{D}} - x\right\|^2. \tag{11}$$

This loss is minimized at $\theta = \overline{\mathcal{D}}$, and the minimum value $\mathcal{L}_* := \frac{1}{m}\sum_{x \in \mathcal{D}}\left\|\overline{\mathcal{D}} - x\right\|^2$. Thus, substituting the expression of $\mathcal{L}$ from (11) in (9), we obtain that

$$\varrho \geq \mathbb{E}\left[\mathcal{L}(\hat{\theta}; \mathcal{D}) - \mathcal{L}_*\right] = \mathbb{E}\left[\left\|\hat{\theta} - \overline{\mathcal{D}}\right\|^2\right].$$

Note that by construction of $\mathcal{M}$, we have $\mathcal{M}(\mathcal{D}) = \mathcal{A}(\mathcal{D}, \dots, \mathcal{D}) = \hat{\theta}$. Thus, from above we obtain that

$$\varrho \geq \mathbb{E}\left[\left\|\mathcal{M}(\mathcal{D}) - \overline{\mathcal{D}}\right\|^2\right].$$

Thus, as $\left\|\cdot\right\|_1 \leq \sqrt{d}\left\|\cdot\right\|$, by taking square roots above, applying Jensen's inequality and multiplying by d, we obtain that

$$d\sqrt{\varrho} \geq d\sqrt{\mathbb{E}\left[\left\|\mathcal{M}(\mathcal{D}) - \overline{\mathcal{D}}\right\|^2\right]} \geq d\;\mathbb{E}\left[\left\|\mathcal{M}(\mathcal{D}) - \overline{\mathcal{D}}\right\|\right] \geq \sqrt{d}\;\mathbb{E}\left[\left\|\mathcal{M}(\mathcal{D}) - \overline{\mathcal{D}}\right\|_1\right] = \mathbb{E}\left[\left\|\sqrt{d}\cdot\mathcal{M}(\mathcal{D}) - \sqrt{d}\cdot\overline{\mathcal{D}}\right\|_1\right]. \tag{12}$$

Recall that $\mathcal{X} = \{\pm 1/\sqrt{d}\}^d$. As in Theorem 5.2 of (Steinke & Ullman, 2016), we define a mechanism $\mathcal{M}' : \{\pm 1\}^{d\times m} \to [\pm 1]^d$ as follows: on input $\mathcal{D}' \subseteq \{\pm 1\}^{d\times m}$ let $\mathcal{D} = \mathcal{D}'/\sqrt{d} \in \mathcal{X}^m$, return $\sqrt{d}\cdot\mathcal{M}(\mathcal{D})$ truncated to $[\pm 1]^d$. Thus, by (12), mechanism $\mathcal{M}'$ verifies for all $\mathcal{D}' \subseteq \{\pm 1\}^{d\times m}$ that

$$d\sqrt{\varrho} \geq \mathbb{E}\left[\left\|\mathcal{M}'(\mathcal{D}) - \overline{\mathcal{D}'}\right\|_1\right]. \tag{13}$$

Invoking Lemma A.1. Note that $\mathcal{M}'$, similar to $\mathcal{M}$, is also $(\varepsilon_n, \delta_n)$-DP by the argument of *post-processing*. Recall that we have shown earlier that ε_n, δ_n satisfy the conditions of Lemma A.1. Since $\varrho \leq 1/100$, we also have $\sqrt{\varrho} \leq 1/10$. Therefore, upon applying Lemma A.1 to $\mathcal{M}'$, in conjunction with (13), we deduce that

$$m = \Omega\left(\frac{\sqrt{d\log\left(1/\delta_n\right)}}{\varepsilon_n\sqrt{\varrho}}\right).$$

By rearranging terms above and taking squares, we obtain that

$$\varrho = \Omega\left(\frac{d\log\left(1/\delta_n\right)}{\varepsilon_n^2 m^2}\right). \tag{14}$$

Recall that we have already shown in (8) and (4), respectively, that $\varepsilon_n \leq 4\varepsilon\sqrt{n\ln\left(1/\delta'\right)}$ and $\delta_n = n\delta + \delta'$, where $\delta' = 1/(m+1)^{1+\gamma}$ (defined in (6)). Therefore, (14) yields

$$\varrho = \Omega\left(\frac{d\log\left(1/(n\delta + \delta')\right)}{\varepsilon^2 n m^2 \log\left(1/\delta'\right)}\right). \tag{15}$$

As $\ln(1+x) \leq x$, substituting δ' from (6), and using the assumption that $\delta \leq 1/8nm^{1+\gamma}$, $\gamma \in (0,1), m \geq 1$, we obtain that

$$\begin{aligned}
\frac{\ln\left(1/(n\delta + \delta')\right)}{\ln\left(1/\delta'\right)} &= \frac{\ln\left(1/\delta'(1 + n\delta/\delta')\right)}{\ln\left(1/\delta'\right)} = 1 + \frac{\ln\left(1/(1 + n\delta/\delta')\right)}{\ln\left(1/\delta'\right)} \\
&= 1 - \frac{\ln\left(1 + n\delta/\delta'\right)}{\ln\left(1/\delta'\right)} \geq 1 - \frac{n\delta}{\delta'\ln\left(1/\delta'\right)} = 1 - \frac{n\delta(m+1)^{\gamma+1}}{(1+\gamma)\ln\left(m+1\right)} \\
&\geq 1 - \frac{(m+1)^{\gamma+1}}{8(1+\gamma)m^{1+\gamma}\ln\left(m+1\right)} \geq 1 - \frac{(2m)^{\gamma+1}}{8(1+\gamma)m^{1+\gamma}\ln\left(m+1\right)} \\
&= 1 - \frac{2^{\gamma+1}}{8(1+\gamma)\ln\left(m+1\right)} \geq 1 - \frac{4}{8\ln\left(m+1\right)} \geq 1 - \frac{1}{2\ln\left(2\right)} = \Omega(1).
\end{aligned}$$

Finally, substituting from above in Equation (15) proves the desired result, i.e.,

$$\varrho = \Omega\left(\frac{d}{\varepsilon^2 n m^2}\right).$$

□

A.3. Case II: No Privacy

We prove below the lower bound due to robustness stated in Proposition 3.2.

Proposition 3.2. *Let Assumption 2.1 hold. Let $n \geq 1$, $1 \leq f < n/2$, and $\nu = \frac{16f(n-2f)}{(n-f)^2}$. Consider $\mathcal{X} = \{\pm \frac{G}{\sqrt{\nu d}}\}^d$ and $\ell = \|\cdot\|^2$. If a distributed algorithm is (f, ϱ)-robust, then*

$$\varrho = \Omega\left(\frac{f}{n} \cdot G^2\right).$$

Proof. The proof is similar to that of Theorem III (Karimireddy et al., 2022). Let $n \geq 1$, $1 \leq f < n/2$, $\nu = \frac{16f(n-2f)}{(n-f)^2}$, and $G > 0$. Consider $\mathcal{X} = \{\pm \frac{G}{\sqrt{\nu d}}\}^d$ and $\ell = \|\cdot\|^2$. Let Assumption 2.1 hold. Assume that algorithm $\mathcal{A}$ is (f, ϱ)-robust.

Denote by $x = \frac{G}{\sqrt{\nu d}} \cdot \mathbf{1} \in \mathbb{R}^d$, where $\mathbf{1} \in \mathbb{R}^d$ is the vector of ones. Consider the following datasets $\mathcal{D}_1 = \ldots = \mathcal{D}_{n-f} = \{x\}^m$ (i.e. all rows are x) and $\mathcal{D}_{n-f+1} = \ldots = \mathcal{D}_n = \{-x\}^m$ (i.e. all rows are $-x$). Consider the two situations of honest identities $\mathcal{H}_1 = \{1, \ldots, n-f\}$ and $\mathcal{H}_2 = \{f+1, \ldots, n\}$.

We first show that the loss functions $\mathcal{L}(\cdot\,; \mathcal{D}_1), \ldots, \mathcal{L}(\cdot\,; \mathcal{D}_n)$ (defined using ℓ in Section 2) satisfy Assumption 2.1 in both situations. This is straightforward in situation $\mathcal{H}_1$ since honest losses are identical. In situation $\mathcal{H}_2$, we have for all $\theta \in \mathbb{R}^d$,

$$\nabla \mathcal{L}_{\mathcal{H}_2}(\theta) = \frac{1}{n-f} \sum_{i \in \mathcal{H}_2} \nabla \mathcal{L}(\theta; \mathcal{D}_i) = \frac{n-2f}{n-f} 2(\theta - x) + \frac{f}{n-f} 2(\theta + x) = 2\left(\theta - \frac{n-3f}{n-f} x\right).$$

Observe that, as $n > 2f$, the intersection $\mathcal{H}_1 \cap \mathcal{H}_2 = \{f+1, \ldots, n-f\}$ is non-empty. Therefore, thanks to the choice of x, we now show that Assumption 2.1 holds, as for all $\theta \in \mathbb{R}^d$ we have

$$\begin{aligned}
\frac{1}{|\mathcal{H}_2|} \sum_{i \in \mathcal{H}_2} \|\nabla \mathcal{L}(\theta; \mathcal{D}_i) - \nabla \mathcal{L}_{\mathcal{H}_2}(\theta)\|^2 &= \frac{|\mathcal{H}_1 \cap \mathcal{H}_2|}{n-f} \|\nabla \mathcal{L}(\theta; \mathcal{D}_{f+1}) - \nabla \mathcal{L}_{\mathcal{H}_2}(\theta)\|^2 \\
&\quad + \frac{|\mathcal{H}_2 \setminus \mathcal{H}_1|}{n-f} \|\nabla \mathcal{L}(\theta; \mathcal{D}_n) - \nabla \mathcal{L}_{\mathcal{H}_2}(\theta)\|^2 \\
&= \frac{n-2f}{n-f} \left\|2(\theta - x) - 2(\theta - \frac{n-3f}{n-f} x)\right\|^2 \\
&\quad + \frac{f}{n-f} \left\|2(\theta + x) - 2(\theta - \frac{n-3f}{n-f} x)\right\|^2 \\
&= \frac{4(n-2f)}{n-f} \left\|\frac{-2f}{n-f} x\right\|^2 + \frac{4f}{n-f} \left\|\frac{2(n-2f)}{n-f} x\right\|^2 = \frac{16f(n-2f)}{(n-f)^2} \|x\|^2 \\
&= \nu \|x\|^2 = G^2.
\end{aligned}$$

Now, denote $\mathcal{L}_{*,\mathcal{H}_1} \coloneqq \inf_{\mathbb{R}^d} \mathcal{L}_{\mathcal{H}_1}$ and $\mathcal{L}_{*,\mathcal{H}_2} \coloneqq \inf_{\mathbb{R}^d} \mathcal{L}_{\mathcal{H}_2}$. Since learning algorithm $\mathcal{A}$ is (f, ϱ)-robust, it outputs $\hat{\theta}$ such that $\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_1}(\hat{\theta}) - \mathcal{L}_{*,\mathcal{H}_1}\right] \leq \varrho$ and $\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_2}(\hat{\theta}) - \mathcal{L}_{*,\mathcal{H}_2}\right] \leq \varrho$. Note that situations $\mathcal{H}_1$ and $\mathcal{H}_2$ are indistinguishable to algorithm $\mathcal{A}$ because it ignores the honest identities, and thus $\hat{\theta}$ is the same in both situations.

Recall that the expression of loss $\mathcal{L}_{\mathcal{H}_1}$ is

$$\mathcal{L}_{\mathcal{H}_1} = \frac{1}{|\mathcal{H}_1|} \sum_{i \in \mathcal{H}_1} \mathcal{L}(\theta; \mathcal{D}_i) = \frac{1}{|\mathcal{H}_1|} \sum_{i \in \mathcal{H}_1} \|\theta - x\|^2 = \|\theta - x\|^2.$$

Therefore, the loss is minimized at $\theta = x$ and we have $\mathcal{L}_{*,\mathcal{H}_1} = \mathcal{L}_{\mathcal{H}_1}(x) = 0$. Thus, we have

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_1}(\hat{\theta}) - \mathcal{L}_{*,\mathcal{H}_1}\right] = \mathbb{E}\left[\left\|\hat{\theta} - x\right\|^2\right].$$

On the other hand, after some algebraic manipulations, the expression of loss $\mathcal{L}_{\mathcal{H}_2}$ is

$$\begin{aligned}\mathcal{L}_{\mathcal{H}_2}(\theta) &= \frac{1}{|\mathcal{H}_2|}\sum_{i\in\mathcal{H}_2}\mathcal{L}(\theta;\mathcal{D}_i) = \frac{|\mathcal{H}_1\cap\mathcal{H}_2|}{n-f}\cdot\|\theta-x\|^2+\frac{|\mathcal{H}_2\setminus\mathcal{H}_1|}{n-f}\cdot\|\theta+x\|^2\\ &= \frac{n-2f}{n-f}\cdot\left(\|\theta\|^2+\|x\|^2-2\left\langle\theta,\,x\right\rangle\right)+\frac{f}{n-f}\cdot\left(\|\theta\|^2+\|x\|^2+2\left\langle\theta,\,x\right\rangle\right)\\ &= \left\|\theta-\frac{n-3f}{n-f}x\right\|^2+\nu\,\|x\|^2\,.\end{aligned}$$

Therefore, the loss is minimized at $\theta = \frac{n-3f}{n-f}x$ and we have $\mathcal{L}_{*,\mathcal{H}_2} = \nu\,\|x\|^2$. Thus, we obtain

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_2}(\hat{\theta})-\mathcal{L}_{*,\mathcal{H}_2}\right] = \mathbb{E}\left[\left\|\hat{\theta}-\frac{n-3f}{n-f}x\right\|^2\right].$$

Recall that $\nu = \frac{16f(n-2f)}{(n-f)^2}$. Therefore, invoking Jensen's inequality, we have

$$\begin{aligned}\varrho &\geq \max\left\{\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_1}(\hat{\theta})-\mathcal{L}_{*,\mathcal{H}_1}\right],\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_2}(\hat{\theta})-\mathcal{L}_{*,\mathcal{H}_2}\right]\right\} \geq \frac{1}{2}\left(\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_1}(\hat{\theta})-\mathcal{L}_{*,\mathcal{H}_1}\right]+\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_2}(\hat{\theta})-\mathcal{L}_{*,\mathcal{H}_2}\right]\right)\\ &= \frac{1}{2}\left(\left\|\hat{\theta}-x\right\|^2+\left\|\hat{\theta}-\frac{n-3f}{n-f}x\right\|^2\right) \geq \frac{1}{4}\left\|\frac{2f}{n-f}x\right\|^2 = \left(\frac{f}{n-f}\right)^2\frac{G^2}{\nu} = \frac{1}{16}\cdot\frac{f}{n-2f}G^2. \qquad (16)\end{aligned}$$

Since $n-2f\leq n$, we obtain $\varrho \geq \frac{1}{16}\cdot\frac{f}{n}\,G^2$, which concludes the proof. □

A.4. Case III: Adversarial Setting

We show below the lower bound from Proposition 3.3 due to the privacy-robustness tradeoff.

Proposition 3.3. *Let $n\geq 3$, $1\leq f<n/2$, $m\geq 1$, $\varepsilon,\delta\in(0,1)$, and $\nu = \frac{16f(n-2f)}{(n-f)^2}$. Consider $\mathcal{X} = \{\pm\frac{1}{\sqrt{d}}\}^d\cup\{\pm\frac{1}{\sqrt{\nu d}}\}^d$ and $\ell = \|\cdot\|^2$. Consider any (ε,δ)-DP distributed algorithm $\mathcal{A}:\mathcal{X}^{m\times n}\to\mathbb{R}^d$. Assume that $2^{-o(m)}\leq\delta\leq 1/m^{1+\Omega(1)}$. For any $\varrho\leq\frac{f+1}{100(n-f)}$, if $\mathcal{A}$ is (f,ϱ)-robust, then*

$$\varrho = \Omega\left(\frac{f+1}{n-f}\cdot\frac{\log\left(1/\delta\right)}{\varepsilon^2m^2}\right).$$

Proof. Let $n\geq 3$, $1\leq f<n/2$, $m\geq 1$, $d\geq 1$, $\varepsilon,\delta\in(0,1)$, $\nu = \frac{16f(n-2f)}{(n-f)^2}$, and $\varrho\leq\frac{f+1}{100(n-f)}$. Consider $\mathcal{X} = \{\pm 1/\sqrt{d}\}^d\cup\{\pm 1/\sqrt{\nu d}\}^d$ and $\ell = \|\cdot\|^2$. We consider a distributed algorithm $\mathcal{A}:\mathcal{X}^{m\times n}\to\mathbb{R}^d$ that satisfies (ε,δ)-distributed DP where $2^{-o(m)}\leq\delta\leq 1/m^{1+\Omega(1)}$ and (f,ϱ)-robustness.

We consider the following datasets. Let $\mathbf{1}$ denote the vector of ones in $\mathbb{R}^d$. For $i\in\{2,\dots,n-f\}$, we set

$$\mathcal{D}_i = \mathcal{D}_+ \coloneqq \{+\frac{1}{\sqrt{d}}\cdot\mathbf{1}\}^m,$$

i.e., all rows are $+\frac{1}{\sqrt{d}}\cdot\mathbf{1}\in\mathbb{R}^d$. For $i\in\{n-f+1,\dots,n\}$ we set

$$\mathcal{D}_i = \mathcal{D}_- \coloneqq \{-\frac{1}{\sqrt{d}}\cdot\mathbf{1}\}^m,$$

i.e., all rows are $-\frac{1}{\sqrt{d}}\cdot\mathbf{1}\in\mathbb{R}^d$. Finally, we fix $\mathcal{D}_1\in\mathcal{X}^m$ to be an arbitrary dataset with every element having identical coordinates. That is, for arbitrary $\alpha_{1,1},\dots,\alpha_{1,m}\in\{\pm 1\}$, we set

$$\mathcal{D}_1 = \left\{\frac{\alpha_{1,1}}{\sqrt{d}}\cdot\mathbf{1},\dots,\frac{\alpha_{1,m}}{\sqrt{d}}\cdot\mathbf{1}\right\}.$$

Proof outline. We consider the centralized algorithm $\mathcal{M} : \mathcal{X}^m \to \mathbb{R}^d$ which takes as input dataset $\mathcal{D}_1 \in \mathcal{X}^m$ and executes $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n)$, where the datasets $\mathcal{D}_2, \dots, \mathcal{D}_n$ are fixed above. We first derive the DP and utility guarantees $\mathcal{M}$ inherits from $\mathcal{A}$, which satisfies (ε, δ)-distributed DP (see Definition 2.3) and (f, ϱ)-robustness, and then conclude the proof by applying the centralized lower bound Lemma A.1 to $\mathcal{M}$.

Privacy guarantees of $\mathcal{M}$. We first state the privacy guarantees of $\mathcal{M}$ inherited from $\mathcal{A}$.

As per Definition 2.3, since $\mathcal{A}$ is (ε, δ)-DP, all communications with worker w_1 (whose dataset is $\mathcal{D}_1$) are (ε, δ)-DP. It follows directly that $\mathcal{M}$ is (ε, δ)-DP by post-processing.

Utility guarantees of $\mathcal{M}$. We now analyze the utility guarantees of $\mathcal{M}$ inherited from $\mathcal{A}$.

Since $\mathcal{A}$ is (f, ϱ)-robust (Definition 2.1), the output $\hat{\theta} = \mathcal{M}(\mathcal{D}_1) = \mathcal{A}(\mathcal{D}_1, \dots, \mathcal{D}_n)$ verifies

$$\varrho \geq \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\hat{\theta}) - \mathcal{L}_*\right], \tag{17}$$

for any set of honest identities $\mathcal{H} \subseteq \{1, \dots, n\}, |\mathcal{H}| = n - f$, where we denote $\mathcal{L}_* \coloneqq \inf_{\mathbb{R}} \mathcal{L}_{\mathcal{H}}$.

<u>*Reduction to one-dimensional space:*</u> We now show that we can simply consider $d = 1$, without loss of generality. For this, we develop the RHS of (17). We have for any $\theta \in \mathbb{R}^d$ and $\mathcal{H} \subseteq \{1, \dots, n\}, |\mathcal{H}| = n - f$:

$$\mathcal{L}_{\mathcal{H}}(\theta) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \frac{1}{m} \sum_{x \in \mathcal{D}_i} \|\theta - x\|^2 . \tag{18}$$

The above function is minimized at $\theta^*_{\mathcal{H}} \coloneqq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \overline{\mathcal{D}}_i$ the average of one-way marginals $\overline{\mathcal{D}}_i \coloneqq \frac{1}{m} \sum_{x \in \mathcal{D}_i} x$. Therefore, the minimum of $\mathcal{L}_{\mathcal{H}}$ is $\mathcal{L}_{*,\mathcal{H}} \coloneqq \mathcal{L}_{\mathcal{H}}(\theta^*_{\mathcal{H}})$.
Recall the following bias-variance decomposition: for any $x_1, \dots, x_n \in \mathbb{R}^d$ we have $\frac{1}{n} \sum_{i=1}^n \|x_i - \overline{x}\|^2 = \frac{1}{n} \sum_{i=1}^n \|x_i\|^2 - \|\overline{x}\|^2$, where we denoted $\overline{x} \coloneqq \frac{1}{n} \sum_{i=1}^n x_i$. Therefore, recalling (18) and $\theta^*_{\mathcal{H}} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \overline{\mathcal{D}}_i$, we have

$$\mathcal{L}_{\mathcal{H}}(\theta) - \mathcal{L}_{*,\mathcal{H}} = \mathcal{L}_{\mathcal{H}}(\theta) - \mathcal{L}_{\mathcal{H}}(\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \overline{\mathcal{D}}_i) = \left\| \theta - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \overline{\mathcal{D}}_i \right\|^2 . \tag{19}$$

Recall our setting of datasets in the beginning of the proof: in particular, for every $i \in \{1, \dots, n\}$, each element of dataset $\mathcal{D}_i$ has identical coordinates. Thus, there is $\alpha_i \in [\pm 1]$ such that $\overline{\mathcal{D}}_i = \frac{\alpha_i}{\sqrt{d}} \cdot \mathbf{1}$. Plugging this in (19) yields:

$$\begin{aligned} \mathcal{L}_{\mathcal{H}}(\theta) - \mathcal{L}_{*,\mathcal{H}} &= \left\| \theta - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \overline{\mathcal{D}}_i \right\|^2 = \left\| \theta - \frac{1}{\sqrt{d} \cdot |\mathcal{H}|} \sum_{i \in \mathcal{H}} \alpha_i \mathbf{1} \right\|^2 = \sum_{k=1}^d \left| \theta_k - \frac{1}{\sqrt{d} \cdot |\mathcal{H}|} \sum_{i \in \mathcal{H}} \alpha_i \right|^2 \\ &= \frac{1}{d} \sum_{k=1}^d \left| \sqrt{d} \cdot \theta_k - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \alpha_i \right|^2 , \end{aligned} \tag{20}$$

where θ_k denotes the k-th coordinate of $\theta \in \mathbb{R}^d$. Upon applying (17) and then Jensen's inequality, we obtain

$$\begin{aligned} \varrho \geq \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\hat{\theta}) - \mathcal{L}_{*,\mathcal{H}}\right] &= \frac{1}{d} \sum_{k=1}^d \mathbb{E}\left[\left| \sqrt{d} \cdot \hat{\theta}_k - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \alpha_i \right|^2 \right] \geq \mathbb{E}\left[\left| \frac{1}{d} \sum_{k=1}^d \sqrt{d} \cdot \hat{\theta}_k - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \alpha_i \right|^2 \right] \\ &= \mathbb{E}\left[\left| \sum_{k=1}^d \frac{\hat{\theta}_k}{\sqrt{d}} - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \alpha_i \right|^2 \right] . \end{aligned} \tag{21}$$

Therefore, everything happens as if $d = 1$. That is, data universe $\mathcal{X} = \{\pm 1\}$, and datasets $\mathcal{D}_+ = \{+1\}^m$, $\mathcal{D}_- = \{-1\}^m$, and $\mathcal{D}_1 = \{\alpha_{1,1}, \dots, \alpha_{1,m}\}$ being arbitrary in $\mathcal{X}^m$. Indeed, denote $\tilde{\theta} \coloneqq \sum_{k=1}^d \frac{\hat{\theta}_k}{\sqrt{d}} \in \mathbb{R}$. Recall that, now that $d = 1$, each $\alpha_i \in [\pm 1]$ is such that $\overline{\mathcal{D}}_i = \alpha_i$. In this one-dimensional setting of datasets, we develop the RHS of (21), by using the aforementioned bias-variance decomposition backwards:

$$\begin{aligned}\varrho \geq \mathbb{E}\left[\left|\tilde{\theta} - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\alpha_i\right|^2\right] &= \mathbb{E}\left[\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\frac{1}{m}\sum_{x\in\mathcal{D}_i}\left|\tilde{\theta} - x\right|^2\right] - \mathbb{E}\left[\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\frac{1}{m}\sum_{x\in\mathcal{D}_i}\left|x - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\overline{\mathcal{D}}_i\right|^2\right]\\ &= \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\tilde{\theta}) - \mathcal{L}_{*,\mathcal{H}}\right].\end{aligned}$$

Thus, (17) holds with loss ℓ being the one-dimensional quadratic loss and mechanism $\widetilde{\mathcal{M}}$ returning $\tilde{\theta}$ instead of $\hat{\theta}$. Since $\tilde{\theta}$ is a function of $\hat{\theta}$ without access to $\mathcal{D}_1$, $\widetilde{\mathcal{M}}$ is also (ε, δ)-DP by post-processing. **Throughout the remainder of the proof, we set $d = 1$ without loss of generality.**

We consider below the RHS of (17). We have for any $\theta \in \mathbb{R}$:

$$\mathcal{L}_{\mathcal{H}}(\theta) = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\frac{1}{m}\sum_{x\in\mathcal{D}_i}|\theta - x|^2. \tag{22}$$

The above function is minimized at $\theta^*_{\mathcal{H}} \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\overline{\mathcal{D}}_i$ the average of one-way marginals $\overline{\mathcal{D}}_i \coloneqq \frac{1}{m}\sum_{x\in\mathcal{D}_i} x$.

Next, following (30), we consider **two possible cases** of honest identities, a priori indistinguishable to the algorithm. In the first case, we consider the set of honest identities $\mathcal{H}$ to be $\mathcal{H}_1 = \{1, \ldots, n-f\}$. In the second case, we consider the set of honest identities $\mathcal{H}$ to be $\mathcal{H}_2 \coloneqq \{1\} \cup \{f+2, \ldots, n\}$. As $|\mathcal{H}| = n - f$, upon invoking Definition 2.1 in both the cases, we obtain a upper bound on $\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right]$ in terms of ϱ.

First case: Consider $\mathcal{H}$ to be $\mathcal{H}_1 = \{1, \ldots, n-f\}$. Recall that $\mathcal{D}_i = \mathcal{D}_+$ for all $i \in \{2, \ldots, n-f\}$. By (22), we have for all $\theta \in \mathbb{R}$:

$$\begin{aligned}\mathcal{L}_{\mathcal{H}_1}(\theta) &= \frac{1}{|\mathcal{H}_1|}\sum_{i\in\mathcal{H}_1}\frac{1}{m}\sum_{x\in\mathcal{D}_i}|\theta - x|^2 = \frac{1}{|\mathcal{H}_1|}\frac{1}{m}\sum_{x\in\mathcal{D}_1}|\theta - x|^2 + \frac{|\mathcal{H}_1| - 1}{|\mathcal{H}_1|}\frac{1}{m}\sum_{x\in\mathcal{D}_+}|\theta - x|^2\\ &= \frac{1}{n-f}\frac{1}{m}\sum_{x\in\mathcal{D}_1}|\theta - x|^2 + (1 - \frac{1}{n-f})\left|\theta - \overline{\mathcal{D}}_+\right|^2\\ &\geq \frac{1}{n-f}\left|\theta - \overline{\mathcal{D}}_1\right|^2 + (1 - \frac{1}{n-f})\left|\theta - \overline{\mathcal{D}}_+\right|^2. && \text{(Jensen's inequality)}\end{aligned}$$

Thus, from above we obtain that

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_1}(\hat{\theta})\right] \geq \frac{1}{n-f}\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + (1 - \frac{1}{n-f})\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_+|^2\right]. \tag{23}$$

Now, recall the following bias-variance decomposition: for any $x_1, \ldots, x_n \in \mathbb{R}$ we have $\frac{1}{n}\sum_{i=1}^n |x_i - \overline{x}|^2 = \frac{1}{n}\sum_{i=1}^n |x_i|^2 - |\overline{x}|^2$ where $\overline{x} \coloneqq \frac{1}{n}\sum_{i=1}^n x_i$. Thus, from (22) we obtain that $\theta^*_{\mathcal{H}_1} = \frac{1}{|\mathcal{H}_1|}\sum_{i\in\mathcal{H}_1}\overline{\mathcal{D}}_i$. Thus, as $|x|^2 = 1$ for all $x \in \mathcal{X}$, we have

$$\begin{aligned}\mathcal{L}_{*,\mathcal{H}_1} = \mathcal{L}_{\mathcal{H}_1}(\theta^*_{\mathcal{H}_1}) &= \frac{1}{m|\mathcal{H}_1|}\sum_{i\in\mathcal{H}_1}\sum_{x\in\mathcal{D}_i}\left|\theta^*_{\mathcal{H}_1} - x\right|^2 = \frac{1}{m|\mathcal{H}_1|}\sum_{i\in\mathcal{H}_1}\sum_{x\in\mathcal{D}_i}|x|^2 - \left|\theta^*_{\mathcal{H}_1}\right|^2\\ &= 1 - \left|\theta^*_{\mathcal{H}_1}\right|^2 = 1 - \left|\frac{1}{|\mathcal{H}_1|}\sum_{i\in\mathcal{H}_1}\overline{\mathcal{D}}_i\right|^2 = 1 - \left|\frac{1}{n-f}\overline{\mathcal{D}}_1 + (1 - \frac{1}{n-f})\overline{\mathcal{D}}_+\right|^2\\ &= 1 - \left|\frac{1}{n-f}\overline{\mathcal{D}}_1 + 1 - \frac{1}{n-f}\right|^2 = 1 - \frac{1}{(n-f)^2}\left|\overline{\mathcal{D}}_1 + n - f - 1\right|^2.\end{aligned}$$

Note that, as $\mathcal{D}_1 \in \mathcal{X}^m = \{\pm 1\}^m$, we have $\overline{\mathcal{D}}_1 \in [\pm 1]$. Also, since $f < n/2$ and $n \geq 3$, we have $n - f - 2 \geq 0$. Therefore,

$\left|\overline{\mathcal{D}}_1 + n - f - 1\right|^2 \geq |n - f - 2|^2$. Substituting this in the above, we obtain that

$$\begin{aligned}
\mathcal{L}_{*,\mathcal{H}_1} &= 1 - \frac{1}{(n-f)^2}\left|\overline{\mathcal{D}}_1 + n - f - 1\right|^2 \leq 1 - \frac{1}{(n-f)^2}\left|n - f - 2\right|^2 = 1 - \left|1 - \frac{2}{n-f}\right|^2 \\
&= \frac{2}{n-f}(2 - \frac{2}{n-f}) = \frac{4}{n-f}(1 - \frac{1}{n-f}) \leq \frac{4}{n-f} \leq \frac{4(f+1)}{n-f}. \qquad (24)
\end{aligned}$$

Substituting from (23) and (24) in (17) we obtain that

$$\varrho + \frac{4(f+1)}{n-f} \geq \varrho + \mathcal{L}_{*,\mathcal{H}_1} \geq \mathbb{E}\left[\mathcal{L}_{\mathcal{H}_1}(\hat{\theta})\right] \geq \frac{1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + (1 - \frac{1}{n-f})\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_+|^2\right]. \qquad (25)$$

<u>Second case</u>: Consider $\mathcal{H}$ to be $\mathcal{H}_2 = \{1\} \cup \{f+2, \ldots, n\}$. Recall that $\mathcal{D}_i = \mathcal{D}_-$ for all $i \in \{n - f + 1, \ldots, n\}$. By (22), we have for all $\theta \in \mathbb{R}$:

$$\begin{aligned}
\mathcal{L}_{\mathcal{H}_2}(\theta) &= \frac{1}{|\mathcal{H}_2|}\sum_{i \in \mathcal{H}_2}\frac{1}{m}\sum_{x \in \mathcal{D}_i}|\theta - x|^2 \\
&= \frac{1}{|\mathcal{H}_2|}\frac{1}{m}\sum_{x \in \mathcal{D}_1}|\theta - x|^2 + \left(\frac{|\mathcal{H}_2| - 1 - f}{|\mathcal{H}_2|}\right)\frac{1}{m}\sum_{x \in \mathcal{D}_+}|\theta - x|^2 + \left(\frac{f}{|\mathcal{H}_2|}\right)\frac{1}{m}\sum_{x \in \mathcal{D}_-}|\theta - x|^2 \\
&= \left(\frac{1}{n-f}\right)\frac{1}{m}\sum_{x \in \mathcal{D}_1}|\theta - x|^2 + \left(\frac{n - 2f - 1}{n-f}\right)\left|\theta - \overline{\mathcal{D}}_+\right|^2 + \frac{f}{n-f}\left|\theta - \overline{\mathcal{D}}_-\right|^2 \\
&\geq \left(\frac{1}{n-f}\right)\frac{1}{m}\sum_{x \in \mathcal{D}_1}|\theta - x|^2 + \frac{f}{n-f}\left|\theta - \overline{\mathcal{D}}_-\right|^2 \qquad (n \geq 2f+1) \\
&\geq \frac{1}{n-f}\left|\theta - \overline{\mathcal{D}}_1\right|^2 + \frac{f}{n-f}\left|\theta - \overline{\mathcal{D}}_-\right|^2. \qquad \text{(Jensen's inequality)}
\end{aligned}$$

Substituting $\theta = \hat{\theta}$, and taking expectation yields

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}_2}(\hat{\theta})\right] \geq \frac{1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + \frac{f}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_-|^2\right]. \qquad (26)$$

Now, recall the following bias-variance decomposition: for any $x_1, \ldots, x_n \in \mathbb{R}$ we have $\frac{1}{n}\sum_{i=1}^n |x_i - \overline{x}|^2 = \frac{1}{n}\sum_{i=1}^n |x_i|^2 - |\overline{x}|^2$, where we denoted $\overline{x} := \frac{1}{n}\sum_{i=1}^n x_i$. Using this in (22), and that $\forall x \in \mathcal{X}$, $|x|^2 = 1$, we get

$$\begin{aligned}
\mathcal{L}_{*,\mathcal{H}_2} = \mathcal{L}_{\mathcal{H}_2}(\theta^*_{\mathcal{H}_2}) &= \frac{1}{m\,|\mathcal{H}_2|}\sum_{i \in \mathcal{H}_2}\sum_{x \in \mathcal{D}_i}\left|\theta^*_{\mathcal{H}_2} - x\right|^2 = \frac{1}{m\,|\mathcal{H}_2|}\sum_{i \in \mathcal{H}_2}\sum_{x \in \mathcal{D}_i}|x|^2 - \left|\theta^*_{\mathcal{H}_2}\right|^2 \\
&= 1 - \left|\theta^*_{\mathcal{H}_2}\right|^2 = 1 - \left|\frac{1}{|\mathcal{H}_2|}\sum_{i \in \mathcal{H}_2}\overline{\mathcal{D}}_i\right|^2 = 1 - \left|\frac{1}{n-f}\overline{\mathcal{D}}_1 + \frac{n - 2f - 1}{n-f}\overline{\mathcal{D}}_+ + \frac{f}{n-f}\overline{\mathcal{D}}_-\right|^2 \\
&= 1 - \left|\frac{1}{n-f}\overline{\mathcal{D}}_1 + \frac{n - 2f - 1}{n-f} - \frac{f}{n-f}\right|^2 = 1 - \left|1 + \frac{1}{n-f}\overline{\mathcal{D}}_1 - \frac{2f+1}{n-f}\right|^2 \\
&= \left(1 - 1 - \frac{1}{n-f}\overline{\mathcal{D}}_1 + \frac{2f+1}{n-f}\right)\left(1 + 1 + \frac{1}{n-f}\overline{\mathcal{D}}_1 - \frac{2f+1}{n-f}\right) \\
&= \left(\frac{2f + 1 - \overline{\mathcal{D}}_1}{n-f}\right)\left(2 - \frac{2f + 1 - \overline{\mathcal{D}}_1}{n-f}\right).
\end{aligned}$$

Note that, as $\mathcal{D}_1 \in \mathcal{X}^m = \{\pm 1\}^m$, we have $\overline{\mathcal{D}}_1 \in [\pm 1]$. This, together with $n \geq 2f + 1$, implies that both the terms in the product above are non-negative. Moreover, as $\overline{\mathcal{D}}_1 \geq -1$, the first term can be bounded by

$$\frac{2f + 1 - \overline{\mathcal{D}}_1}{n-f} \leq \frac{2(f+1)}{n-f}.$$

Similarly, as $\overline{\mathcal{D}}_1 \leq 1$, the second term can be bounded by

$$2 - \frac{2f + 1 - \overline{\mathcal{D}}_1}{n-f} \leq 2 - \frac{2f}{n-f} \leq 2.$$

Consequently, we have

$$\mathcal{L}_{*,\mathcal{H}_2} \leq \frac{4(f+1)}{n-f}. \tag{27}$$

Invoking (17) with the set of honest identities $\mathcal{H}_2$, and using the bounds shown in (26), (27) yields:

$$\varrho + \frac{4(f+1)}{n-f} \geq \varrho + \mathcal{L}_{*,\mathcal{H}_2} \geq \mathbb{E}\left[\mathcal{L}_{\mathcal{H}_2}(\hat{\theta})\right] \geq \frac{1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + \frac{f}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_-|^2\right]. \tag{28}$$

Final step: We deduce from (25), (28) that

$$\begin{aligned}
\varrho + \frac{4(f+1)}{n-f} &\geq \max\Big\{\frac{1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + (1 - \frac{1}{n-f})\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_+|^2\right], \\
&\qquad \frac{1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + \frac{f}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_-|^2\right]\Big\} \\
&= \frac{1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + \max\left\{(1 - \frac{1}{n-f})\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_+|^2\right], \frac{f}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_-|^2\right]\right\} \\
&\geq \frac{1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] + \frac{f}{n-f}\max\left\{\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_+|^2\right], \mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_-|^2\right]\right\},
\end{aligned} \tag{29}$$

where the last inequality is due to $f < \frac{n}{2}$, which implies that $1 - \frac{1}{n-f} \geq \frac{f}{n-f}$. Besides, observe that, as $\mathcal{D}_1 \in \mathcal{X}^m = \{\pm 1\}^m$, we have $\overline{\mathcal{D}}_1 \in [\pm 1]$. Recall that $\overline{\mathcal{D}}_+ = +1$ and $\overline{\mathcal{D}}_- = -1$. Thus, it holds that

$$\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] \leq \max\left(\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_+|^2\right], \mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_-|^2\right]\right). \tag{30}$$

Indeed, since $\overline{\mathcal{D}}_1 \in [\pm 1]$, we can write $\overline{\mathcal{D}}_1 = \lambda \times (+1) + (1 - \lambda) \times (-1)$ for some $\lambda \in [0, 1]$. Thus, using Jensen's inequality and then taking expectations, we have $\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right] \leq \lambda\,\mathbb{E}\left[|\hat{\theta} - 1|^2\right] + (1 - \lambda)\,\mathbb{E}\left[|\hat{\theta} + 1|^2\right] \leq \max\left(\mathbb{E}\left[|\hat{\theta} - 1|^2\right], \mathbb{E}\left[|\hat{\theta} + 1|^2\right]\right)$.

Using (30) in (29), we obtain, for every $\mathcal{D}_1 \in \mathcal{X}^m$, that

$$\varrho + \frac{4(f+1)}{n-f} \geq \frac{f+1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right]. \tag{31}$$

Before concluding, recall that $1 \leq f \leq \frac{n}{2}$, thus applying Proposition 3.2 with $G = 1$ yields

$$\varrho = \Omega\left(\frac{f}{n}\right) = \Omega\left(\frac{f+1}{n-f}\right). \tag{32}$$

Indeed, since the data universe considered in the proof includes $\{\pm \frac{1}{\sqrt{\nu d}}\}^d$, we can apply Proposition 3.2. Plugging this back in (31), we have for every $\mathcal{D}_1 \in \mathcal{X}^m$ that

$$\varrho = \Omega\left(\frac{f+1}{n-f}\,\mathbb{E}\left[|\hat{\theta} - \overline{\mathcal{D}}_1|^2\right]\right).$$

Invoking Lemma A.1. Hence, since $\varrho \leq \frac{f+1}{100(n-f)}$, we can proceed in the same way as in the proof of Proposition 3.1 to leverage Lemma A.1 (with $d = 1$) for showing

$$\frac{n-f}{f+1}\varrho = \Omega\left(\frac{\log(1/\delta)}{\varepsilon^2 m^2}\right).$$

We finally conclude the desired result by rearranging terms and ignoring absolute constants:

$$\varrho = \Omega\left(\frac{f+1}{n-f} \cdot \frac{\log(1/\delta)}{\varepsilon^2 m^2}\right).$$

□

A.5. Final Lower Bound

We prove below the final lower bound stated in Theorem 3.1.

Theorem 3.1. *Let $\mathcal{X} = \mathbb{R}^d$, $\ell = \|\cdot\|^2$, $n \geq 3$, $0 \leq f < n/2$, $m \geq 1$, and $\varepsilon, \delta \in (0,1)$. Consider arbitrary datasets $\mathcal{D}_1, \dots, \mathcal{D}_n \in \mathcal{X}^m$ such that Assumption 2.1 is satisfied with $G \geq 1$. Let $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathbb{R}^d$ be an (ε, δ)-DP distributed algorithm. Assume that $\varepsilon \leq 1/4\sqrt{2n \ln(m+1)}$, and that $2^{-m^{1-\gamma}} \leq n\delta \leq 1/8m^{1+\gamma}$ for some $\gamma \in (0,1)$. For any $\varrho \leq \frac{f+1}{100(n-f)}$, if $\mathcal{A}$ is (f, ϱ)-robust, then*

$$\varrho = \widetilde{\Omega}\left(\frac{d}{\varepsilon^2 nm^2} + \frac{f}{n} \cdot \frac{1}{\varepsilon^2 m^2} + \frac{f}{n} \cdot G^2\right).$$

Proof. The proof consists in showing that the setting we consider in the above theorem allows us to merge the lower bounds from propositions 3.1, 3.3, and 3.2. First, we remark that the case $f = 0$ corresponds to simply showing that $\varrho = \widetilde{\Omega}\left(\frac{d}{\varepsilon^2 nm^2}\right)$, which follows immediately from Proposition 3.1 directly (see Step 1 below for verifying the applicability of the proposition). In the remainder of the proof, we will assume $f > 0$ and $\eta > 0$. Let $\mathcal{H}$ denote the set of honest nodes of size $n - f$.

Step 1: To derive the first term in $\Omega\left(\frac{d}{\varepsilon^2 nm^2}\right)$, we remark that all the conditions of Proposition 3.1 on $\varepsilon, \delta, \varrho, n, m$ hold under the assumptions stated in the theorem. Consider $\mathcal{D}_1, \dots, \mathcal{D}_n \in \{\pm 1/\sqrt{8d}\}^{d \times m} \subset \mathcal{X}^m$. Note that in this case, we have

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\nabla \mathcal{L}(\theta; \mathcal{D}_i) - \nabla \mathcal{L}_{\mathcal{H}}(\theta)\|^2 \leq 1 \leq G^2.$$

Hence, $\mathcal{D}_1, \dots, \mathcal{D}_n$ is a valid collection of datasets with regard to the theorem statement. Since $\mathcal{A}$ is assumed to be (f, ϱ)-robust, it guarantees an error less than or equal to ϱ on the honest global loss $\mathcal{L}(\theta; \mathcal{D}_i, \, i \in \mathcal{H})$. Using the proof technique of Proposition 3.1, we can show that (as $f < n/2$ and $|\mathcal{H}| = n - f \leq n$)

$$\varrho = \Omega\left(\frac{d}{\varepsilon^2 |\mathcal{H}| m^2}\right) = \Omega\left(\frac{d}{\varepsilon^2 nm^2}\right). \tag{33}$$

Step 2: To derive the second term in $\Omega(\frac{f}{n} \cdot \frac{1}{\varepsilon^2 m^2})$, we remark that all conditions of Proposition 3.3 on $\varepsilon, \delta, \varrho, n, f, m$, and $\mathcal{A}$ are verified. Note also that, similar to Step 1, the datasets considered in the proof Proposition 3.2, scaled by a constant, are also valid instances with regard to the theorem statement. Using the proof technique of Proposition 3.2 we can show that (since $0 < f < n/2$, we have $f + 1 \geq f$ and $n - f \leq n$)

$$\varrho = \Omega\left(\frac{f+1}{n-f} \cdot \frac{\log(1/\delta)}{\varepsilon^2 m^2}\right) = \widetilde{\Omega}\left(\frac{f}{n} \cdot \frac{1}{\varepsilon^2 m^2}\right), \tag{34}$$

where we ignore the logarithmic term in $\widetilde{\Omega}(\cdot)$.

Step 3: To obtain the third term in $\Omega\left(\frac{f}{n} \cdot G^2\right)$, we first remark that Assumption 2.1 holds, as well as all the conditions in Proposition 3.2 on n, f, m and $\mathcal{A}$. As the input domain in Proposition 3.2 is a subset of $\mathcal{X}$, using the proof technique of Proposition 3.2 we can show that

$$\varrho = \Omega\left(\frac{f}{n} \cdot G^2\right). \tag{35}$$

Final step: Combining (33), (34), and (35) proves the theorem, i.e., we obtain that

$$\varrho = \widetilde{\Omega}\left(\max\left\{\frac{d}{\varepsilon^2 nm^2}, \frac{f}{n} \cdot \frac{1}{\varepsilon^2 m^2}, \frac{f}{n} \cdot G^2\right\}\right) = \widetilde{\Omega}\left(\frac{d}{\varepsilon^2 nm^2} + \frac{f}{n} \cdot \frac{1}{\varepsilon^2 m^2} + \frac{f}{n} \cdot G^2\right).$$

□

B. Robustness Analysis

In this section, we prove all our claims related to (f, κ)-robustness and SMEA. In Section B.1, we analyze SMEA. In Section B.2, we discuss Filter (Diakonikolas et al., 2017; Steinhardt et al., 2018), a related algorithm.

We first recall the definition of our robustness criterion:

Definition 4.1. *Let $n \geq 1$, $0 \leq f < n/2$ and $\kappa \geq 0$. An aggregation rule F is said to be (f, κ)-*robust averaging *if for any vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, and any set $S \subseteq \{1, \ldots, n\}$ of size $n - f$, the output $\hat{x} = F(x_1, \ldots, x_n)$ satisfies*

$$\left\|\hat{x} - \overline{x}_S\right\|^2 \leq \kappa \cdot \lambda_{\max}\left(\frac{1}{|S|}\sum_{i \in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right),$$

where $\overline{x}_S \coloneqq \frac{1}{|S|}\sum_{i \in S} x_i$ and $\lambda_{\max}$ denotes the maximum eigenvalue. We refer to κ as the robustness coefficient of F.

B.1. Smallest Maximum Eigenvalue Averaging (SMEA)

Given a set of n vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, the SMEA algorithm first searches for a set S^* of cardinality $n - f$ with the smallest empirical *maximum eigenvalue*, i.e.,

$$S^* \in \underset{\substack{S \subseteq \{1, \ldots, n\} \\ |S| = n - f}}{\operatorname{argmin}} \lambda_{\max}\left(\frac{1}{|S|}\sum_{i \in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right). \tag{36}$$

Then the algorithm outputs the average of the inputs in set S^*:

$$\text{SMEA}(x_1, \ldots, x_n) \coloneqq \frac{1}{|S^*|}\sum_{i \in S^*} x_i. \tag{37}$$

Proposition 5.1. *Let $f < n/2$. SMEA is (f, κ)-robust averaging with*

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

Proof. Let $n \geq 1$ and $0 \leq f < n/2$. Fix a set $S \subseteq \{1, \ldots, n\}$ such that $|S| = n - f$. Recall the definition of S^* in (36). Denote by $\overline{x}_{S^*}$ the output of SMEA defined in (37):

$$\overline{x}_{S^*} \coloneqq \frac{1}{|S^*|}\sum_{i \in S^*} x_i. \tag{38}$$

First, observe that we have

$$|S \setminus S^*| = |S^* \setminus S| = |S \cup S^*| - |S| \leq n - (n - f) = f. \tag{39}$$

From (38), we have

$$\begin{aligned}
\left\|\overline{x}_{S^*} - \overline{x}_S\right\|^2 &= \left\|\frac{1}{n-f}\sum_{i \in S^*} x_i - \frac{1}{n-f}\sum_{i \in S} x_i\right\|^2 = \left\|\frac{1}{n-f}\sum_{i \in S^* \setminus S} x_i - \frac{1}{n-f}\sum_{i \in S \setminus S^*} x_i\right\|^2 \\
&= \left\|\frac{1}{n-f}\sum_{i \in S^* \setminus S}(x_i - \overline{x}_{S^*}) - \frac{1}{n-f}\sum_{i \in S \setminus S^*}(x_i - \overline{x}_S) + \frac{|S^* \setminus S|}{n-f}(\overline{x}_{S^*} - \overline{x}_S)\right\|^2 \\
&= \left\|\frac{1}{n-f}\sum_{i \in S^* \setminus S}(x_i - \overline{x}_{S^*}) - \frac{1}{n-f}\sum_{i \in S \setminus S^*}(x_i - \overline{x}_S)\right\|^2 + \frac{|S^* \setminus S|^2}{(n-f)^2}\left\|\overline{x}_{S^*} - \overline{x}_S\right\|^2 \\
&\quad + 2\frac{|S^* \setminus S|}{n-f}\left\langle \overline{x}_{S^*} - \overline{x}_S,\, \frac{1}{n-f}\sum_{i \in S^* \setminus S}(x_i - \overline{x}_{S^*}) - \frac{1}{n-f}\sum_{i \in S \setminus S^*}(x_i - \overline{x}_S)\right\rangle.
\end{aligned}$$

However, notice that

$$\begin{aligned}
\frac{1}{n-f}\sum_{i\in S^*\setminus S}(x_i-\overline{x}_{S^*})-\frac{1}{n-f}\sum_{i\in S\setminus S^*}(x_i-\overline{x}_S) &= \frac{1}{n-f}\sum_{i\in S^*\setminus S}x_i-\frac{1}{n-f}\sum_{i\in S\setminus S^*}x_i-\frac{|S^*\setminus S|}{n-f}(\overline{x}_{S^*}-\overline{x}_S)\\
&= \frac{1}{n-f}\sum_{i\in S^*}x_i-\frac{1}{n-f}\sum_{i\in S}x_i-\frac{|S^*\setminus S|}{n-f}(\overline{x}_{S^*}-\overline{x}_S)\\
&= \left(1-\frac{|S^*\setminus S|}{n-f}\right)(\overline{x}_{S^*}-\overline{x}_S).
\end{aligned}$$

This implies that

$$\begin{aligned}
\left\|\overline{x}_{S^*}-\overline{x}_S\right\|^2 &= \left\|\frac{1}{n-f}\sum_{i\in S^*\setminus S}(x_i-\overline{x}_{S^*})-\frac{1}{n-f}\sum_{i\in S\setminus S^*}(x_i-\overline{x}_S)\right\|^2\\
&\quad+\left[\frac{|S^*\setminus S|^2}{(n-f)^2}+2\frac{|S^*\setminus S|}{n-f}\left(1-\frac{|S^*\setminus S|}{n-f}\right)\right]\left\|\overline{x}_{S^*}-\overline{x}_S\right\|^2\\
&= \left\|\frac{1}{n-f}\sum_{i\in S^*\setminus S}(x_i-\overline{x}_{S^*})-\frac{1}{n-f}\sum_{i\in S\setminus S^*}(x_i-\overline{x}_S)\right\|^2+\left[1-\left(1-\frac{|S^*\setminus S|}{n-f}\right)^2\right]\left\|\overline{x}_{S^*}-\overline{x}_S\right\|^2
\end{aligned}$$

By rearranging the terms, applying Jensen's inequality, and using the fact that $\sup_{\|v\|\leq 1}|\langle v,\,x\rangle|=\|x\|$, we obtain

$$\begin{aligned}
\left(1-\frac{|S^*\setminus S|}{n-f}\right)^2\left\|\overline{x}_{S^*}-\overline{x}_S\right\|^2 &= \left\|\frac{1}{n-f}\sum_{i\in S^*\setminus S}(x_i-\overline{x}_{S^*})-\frac{1}{n-f}\sum_{i\in S\setminus S^*}(x_i-\overline{x}_S)\right\|^2\\
&= \sup_{\|v\|\leq 1}\left|\left\langle v,\,\frac{1}{n-f}\sum_{i\in S^*\setminus S}(x_i-\overline{x}_{S^*})-\frac{1}{n-f}\sum_{i\in S\setminus S^*}(x_i-\overline{x}_S)\right\rangle\right|^2\\
&= \sup_{\|v\|\leq 1}\left|\frac{1}{n-f}\sum_{i\in S^*\setminus S}\langle v,\,x_i-\overline{x}_{S^*}\rangle-\frac{1}{n-f}\sum_{i\in S\setminus S^*}\langle v,\,x_i-\overline{x}_S\rangle\right|^2\\
&\leq \frac{|S^*\setminus S|+|S\setminus S^*|}{(n-f)^2}\sup_{\|v\|\leq 1}\left[\sum_{i\in S^*\setminus S}|\langle v,\,x_i-\overline{x}_{S^*}\rangle|^2+\sum_{i\in S\setminus S^*}|\langle v,\,x_i-\overline{x}_S\rangle|^2\right]\\
&\leq \frac{|S^*\setminus S|+|S\setminus S^*|}{(n-f)^2}\left[\sup_{\|v\|\leq 1}\sum_{i\in S^*\setminus S}|\langle v,\,x_i-\overline{x}_{S^*}\rangle|^2+\sup_{\|v\|\leq 1}\sum_{i\in S\setminus S^*}|\langle v,\,x_i-\overline{x}_S\rangle|^2\right]\\
&\leq \frac{2f}{(n-f)^2}\left[\sup_{\|v\|\leq 1}\sum_{i\in S^*\setminus S}|\langle v,\,x_i-\overline{x}_{S^*}\rangle|^2+\sup_{\|v\|\leq 1}\sum_{i\in S\setminus S^*}|\langle v,\,x_i-\overline{x}_S\rangle|^2\right], \qquad (40)
\end{aligned}$$

where the last inequality is due to $|S\setminus S^*|=|S^*\setminus S|\leq f$ shown in (39).

The first term on the RHS of (40) can be bounded by construction of S^*, and using the fact that $\sup_{\|v\|\leq 1}\langle v,\,Mv\rangle=\lambda_{\max}(M)$:

$$\begin{aligned}
\sup_{\|v\|\leq 1}\sum_{i\in S^*\setminus S}|\langle v,\,x_i-\overline{x}_{S^*}\rangle|^2 \leq \sup_{\|v\|\leq 1}\sum_{i\in S^*}|\langle v,\,x_i-\overline{x}_{S^*}\rangle|^2 &= \sup_{\|v\|\leq 1}\left\langle v,\,\sum_{i\in S^*}(x_i-\overline{x}_{S^*})(x_i-\overline{x}_{S^*})^\top v\right\rangle\\
&= \lambda_{\max}\left(\sum_{i\in S^*}(x_i-\overline{x}_{S^*})(x_i-\overline{x}_{S^*})^\top\right)\leq\lambda_{\max}\left(\sum_{i\in S}(x_i-\overline{x}_S)(x_i-\overline{x}_S)^\top\right).
\end{aligned}$$

The second term on the RHS of (40) can be bounded similarly:

$$\sup_{\|v\|\leq 1} \sum_{i\in S\setminus S^*} \left|\langle v,\, x_i - \overline{x}_S\rangle\right|^2 \leq \sup_{\|v\|\leq 1} \sum_{i\in S} \left|\langle v,\, x_i - \overline{x}_S\rangle\right|^2 = \lambda_{\max}\left(\sum_{i\in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right).$$

Plugging these two bounds back in (40), we obtain

$$\left(1 - \frac{|S^* \setminus S|}{n-f}\right)^2 \|\overline{x}_{S^*} - \overline{x}_S\|^2 \leq \frac{4f}{n-f}\frac{1}{n-f}\lambda_{\max}\left(\sum_{i\in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right).$$

Finally, since $|S^* \setminus S| \leq f$ (see (39)), we have $\left(1 - \frac{|S^*\setminus S|}{n-f}\right)^2 \geq \left(1 - \frac{f}{n-f}\right)^2 = \left(\frac{n-2f}{n-f}\right)^2$. We can therefore obtain

$$\|\overline{x}_{S^*} - \overline{x}_S\|^2 \leq \frac{4f(n-f)}{(n-2f)^2} \cdot \lambda_{\max}\left(\frac{1}{|S|}\sum_{i\in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right).$$

The proof concludes by noticing that $\frac{4f(n-f)}{(n-2f)^2} = \frac{4f}{n-f}\left(1 + \frac{f}{n-2f}\right)^2$. □

B.2. Filter Algorithm

In this section, we present the Filter algorithm (Diakonikolas et al., 2017; Steinhardt, 2018) in Algorithm 2 and discuss its robustness properties, stated in Proposition B.1, in the distributed ML context we consider. Recall that Filter was also used in (Data & Diggavi, 2021).

Algorithm 2 Filter algorithm (Diakonikolas et al., 2017; Steinhardt, 2018)

```
Input: vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, spectral norm bound $\sigma_0^2$, constant factor $\eta > 0$.
Initialize $c_1, \ldots, c_n = 1$, $\hat{\sigma}_c = +\infty$.
while True do
    Compute the empirical mean $\hat{\mu}_c = \sum_{i=1}^n c_i x_i / \sum_{i=1}^n c_i$.
    Compute the empirical covariance $\hat{\Sigma}_c = \sum_{i=1}^n c_i(x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^\top / \sum_{i=1}^n c_i$.
    Compute maximum eigenvalue $\hat{\sigma}_c^2$ of $\hat{\Sigma}_c$ and an associated eigenvector $\hat{v}_c$.
    if $\hat{\sigma}_c^2 > \eta \cdot \sigma_0^2$ then
        return $\hat{\mu}_c$
    else
        Compute weight $\tau_i = \langle \hat{v}_c,\, x_i - \hat{\mu}_c\rangle^2$.
        Update $c_i \leftarrow c_i(1 - \tau_i/\tau_{\max})$, where $\tau_{\max} = \max_{1\leq i\leq n}\tau_i$.
    end if
end while
```

In Proposition B.1, we recall the robustness guarantees of the Filter procedure (Algorithm 2). The proposition is followed by a discussion further below.

Proposition B.1. *Let $n \geq 1, 0 \leq f < n/2$, $x_1, \ldots, x_n \in \mathbb{R}^n$, and $S \subseteq [n], |S| = n - f$. Denote $\overline{x}_S \coloneqq \frac{1}{|S|}\sum_{i\in S} x_i$.*

Set the parameters

$$\sigma_0^2 \geq \lambda_{\max}\left(\frac{1}{|S|}\sum_{i\in S}(x_i - \overline{x}_S)(x_i - \overline{x}_S)^\top\right)$$

and

$$\eta = 2n(n-f)/(n-2f)^2.$$

Then, the output $\widehat{x}$ of the Filter procedure (Algorithm 2) with parameters σ_0^2 and η satisfies

$$\|\widehat{x} - \overline{x}_S\|^2 \leq \kappa \cdot \sigma_0^2,$$

with $\kappa = \frac{4fn}{(n-2f)^2} + \frac{2f}{n-f} = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$.

Proof. The proof follows directly from (Theorem 4.2, (Zhu et al., 2022)) combined with (Lemma 2.2, (Zhu et al., 2022)). □

Discussion. Note that Filter does not satisfy (f, κ)-robust averaging (see Definition 4.1) as its parameter σ_0^2 must depend on the maximum eigenvalue of the honest inputs. Indeed, such dependency is precluded by (f, κ)-robust averaging. Moreover, in our learning setting, the bound σ_0^2 potentially depends on the noise of stochastic gradients σ^2 and the heterogeneity metric G^2, which are unknown a priori. Thus, devising aggregation rules agnostic to the statistical properties of the honest inputs, like SMEA, is even more desirable in our setting.

C. Privacy Analysis

C.1. Preliminaries

We first recall definitions and useful lemmas on Differential Privacy (DP) and Rényi Differential Privacy (RDP), including the privacy amplification by subsampling (without replacement) results for RDP.

Definition C.1 (Rényi Differential Privacy, (Mironov, 2017))**.** *Let $\alpha > 1$ and $\varepsilon > 0$. A randomized algorithm $\mathcal{M}$ is (α, ε)-RDP if for any adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^m$ it holds that*

$$D_\alpha(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \leq \varepsilon,$$

where $D_\alpha(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) \coloneqq \frac{1}{\alpha-1} \log \mathbb{E}_{\theta \sim \mathcal{M}(\mathcal{D}')} \left[\left(\frac{\mathcal{M}(\mathcal{D})(\theta)}{\mathcal{M}(\mathcal{D}')(\theta)} \right)^\alpha \right]$ is the Rényi divergence of order α.

Lemma C.1 (RDP Adpative Composition, (Mironov, 2017))**.** *If $\mathcal{M}_1$ that takes the dataset as input is (α, ε_1)-RDP, and $\mathcal{M}_2$ that takes the dataset and the output of $\mathcal{M}_1$ as input is (α, ε_2)-RDP, then their composition is $(\alpha, \varepsilon_1 + \varepsilon_2)$-RDP.*

Lemma C.2 (RDP to DP conversion, (Mironov, 2017))**.** *If $\mathcal{M}$ is (α, ε)-RDP, then $\mathcal{M}$ is $(\varepsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-DP for all $\delta \in (0, 1)$.*

Definition C.2 (ℓ_2-sensitivity, (Dwork et al., 2014))**.** *The ℓ_2-sensitivity of a function $g \colon \mathcal{X}^m \to \mathbb{R}^d$ is*

$$\Delta(g) \coloneqq \sup_{\mathcal{D}, \mathcal{D}' \text{ adjacent}} \|g(\mathcal{D}) - g(\mathcal{D}')\| .$$

Lemma C.3 (RDP for Gaussian Mechanisms, (Mironov, 2017))**.** *If $g : \mathcal{X}^m \to \mathbb{R}^d$ has ℓ_2-sensitivity smaller than Δ, then the Gaussian mechanism $G_{\sigma,g} = g + \mathcal{N}(0, \sigma^2 I_d)$ is $(\alpha, \frac{\Delta^2}{2\sigma^2}\alpha)$-RDP.*

Definition C.3 (Subsampling Mechanism)**.** *Consider a dataset $\mathcal{D} \subseteq \mathcal{X}^m$, a constant $b \in [m]$, and define $r \coloneqq {}^b\!/_m$. The procedure $\mathsf{subsample}_r : \mathcal{X}^m \to \mathcal{X}^b$ selects b points at random and without replacement from $\mathcal{D}$.*

Lemma C.4 (RDP for Subsampled Mechanisms, (Wang et al., 2019a))**.** *Let $\alpha \in \mathbb{N}, \alpha \geq 2$, and $r \in (0, 1)$ the sampling parameter. If $\mathcal{M}$ is $(\alpha, \varepsilon(\alpha))$-RDP, then $\mathcal{M} \circ \mathsf{subsample}_r$ is $(\alpha, \varepsilon'(\alpha))$-RDP, with*

$$\begin{aligned} \varepsilon'(\alpha) = \frac{1}{\alpha - 1} \log \Big(1 + r^2 \binom{\alpha}{2} \min \Big\{ 4(e^{\varepsilon(2)} - 1), e^{\varepsilon(2)} \min \left\{ 2, (e^{\varepsilon(\infty)} - 1)^2 \right\} \Big\} \\ + \sum_{j=3}^{\alpha} r^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} \min \left\{ 2, (e^{\varepsilon(\infty)} - 1)^j \right\} \Big). \end{aligned} \tag{41}$$

Lemma C.5 (Real-valued RDP for Subsampled Mechanisms)**.** *Let $\alpha \in \mathbb{R}, \alpha > 1$, and $r \in (0, 1)$ the sampling parameter. If $\mathcal{M}$ is $(\alpha, \varepsilon(\alpha))$-RDP, then $\mathcal{M} \circ \mathsf{subsample}_r$ is $(\alpha, \varepsilon''(\alpha))$-RDP, with*

$$\varepsilon''(\alpha) = (1 - \alpha + \lfloor \alpha \rfloor) \frac{\lfloor \alpha \rfloor - 1}{\alpha - 1} \varepsilon'(\lfloor \alpha \rfloor) + (\alpha - \lfloor \alpha \rfloor) \frac{\lceil \alpha \rceil - 1}{\alpha - 1} \varepsilon'(\lceil \alpha \rceil),$$

where ε' is defined in Equation (41).

Proof. The result follows immediately from Corollary 10 and Remark 7 in (Wang et al., 2019a). □

C.2. Proof of Theorem 4.1

We state below the DP guarantees without approximation:

Theorem C.1. *Let $\delta \in (0, 1)$. Algorithm 1 is (ε^*, δ)-DP with*

$$\varepsilon^* = \inf_{\alpha > 1} \Big(T\varepsilon_1(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \Big),$$

where for every $\alpha > 1$,

$$\begin{cases} \varepsilon_1(\alpha) \coloneqq (1 - \alpha + \lfloor \alpha \rfloor) \frac{\lfloor \alpha \rfloor - 1}{\alpha - 1} \varepsilon'(\lfloor \alpha \rfloor) + (\alpha - \lfloor \alpha \rfloor) \frac{\lceil \alpha \rceil - 1}{\alpha - 1} \varepsilon'(\lceil \alpha \rceil), \\ \varepsilon'(\alpha) \coloneqq \frac{1}{\alpha - 1} \log \Big(1 + r^2 \binom{\alpha}{2} \min \left\{ 4(e^{\varepsilon(2)} - 1), 2e^{\varepsilon(2)} \right\} + 2 \sum_{j=3}^{\alpha} r^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} \Big), \\ \varepsilon(\alpha) \coloneqq \left(\frac{2C}{b} \right)^2 \frac{\alpha}{2\sigma_{\mathrm{DP}}^2}. \end{cases}$$

$$\begin{tikzcd}\theta_t \arrow[r, "\textbf{(I)}"] & \tilde{g}_t^{(i)} \arrow[r, "\textbf{(II)}"] & \theta_{t+1}\end{tikzcd}$$

Figure 1. (**I**): Subsampling + Gaussian mechanism, (**II**): Post-processing.

Proof. To derive the above DP guarantees, we first track the privacy loss for a single iteration of Algorithm 1 using RDP. Then we apply adaptive composition to track the end-to-end privacy loss of the algorithm. Finally, we optimize over the privacy loss for several levels of RDP to compute the noise parameter needed for DP.

Single-iteration privacy. First, we analyze a single fixed iteration $t \in \{0, \dots, T-1\}$ of Algorithm 1. To do so, we divide the analysis into two steps, i.e. Step I and Step II, as shown in Figure 1.

*Step (**I**):* This step corresponds to lines 2-6 in Algorithm 1. Recall that our definition of DP for a distribution algorithm (given in Definition 2.3) requires that the transcript of communications of each worker satisfies (centralized) (ε, δ)-DP with respect to their own data. Thus, since the workers only send their local momentum to the server, we show that for any $i \in \mathcal{H}$ computing $\tilde{g}_t^{(i)}$ from $\mathcal{D}_i$ and θ_t is RDP for any $\alpha > 1$.

Let $i \in \mathcal{H}, \alpha > 1$ and $r = {}^{b}/_{m}$. First, we show that $\Delta := \frac{2C}{b}$ is an upper bound of the ℓ_2-sensitivity of the mini-batch (clipped) averaging. To see this, consider two adjacent training sets $\mathcal{D}_i, \tilde{\mathcal{D}}_i$, the mini-batch average (after clipping) $g_t^{(i)}$ computed on mini-batch $S_t^{(i)} \subseteq \mathcal{D}_i$, and $\tilde{g}_t^{(i)}$ the analogous quantities for $\tilde{\mathcal{D}}_i$. Note that $S_t^{(i)}$ and $\tilde{S}_t^{(i)}$ differ by one element at most. Without loss of generality, let $x_* \in S_t^{(i)}, \tilde{x}_* \in \tilde{S}_t^{(i)}$ be the only two elements that differ from $S_t^{(i)}$ to $\tilde{S}_t^{(i)}$. Thanks to the triangle inequality, we have that

$$\begin{aligned}\left\| g_t^{(i)} - \tilde{g}_t^{(i)} \right\| &= \left\| \frac{1}{b} \sum_{x \in S_t^{(i)}} \textbf{Clip}\left(\nabla \ell(\theta_t, x); C\right) - \frac{1}{b} \sum_{x \in \tilde{S}_t^{(i)}} \textbf{Clip}\left(\nabla \ell(\theta_t, x); C\right) \right\| \\ &= \left\| \frac{1}{b} \textbf{Clip}\left(\nabla \ell(\theta_t, x_*); C\right) - \frac{1}{b} \textbf{Clip}\left(\nabla \ell(\theta_t, \tilde{x}_*); C\right) \right\| \\ &\leq \frac{1}{b} \left\| \textbf{Clip}\left(\nabla \ell(\theta_t, x_*); C\right) \right\| + \frac{1}{b} \left\| \textbf{Clip}\left(\nabla \ell(\theta_t, \tilde{x}_*); C\right) \right\| \\ &\leq \frac{2C}{b}.\end{aligned}$$

Thanks to the above, the sensitivity of computing the gradient $g_t^{(i)}$ when given a batch of b point $S_t^{(i)}$ is upper bounded by $\Delta = \frac{2C}{b}$. Accordingly, invoking Lemma C.3, the Gaussian mechanism used in Line 6 of Algorithm 1 is $(\alpha, \frac{\alpha \Delta^2}{2\sigma_{\mathrm{DP}}^2})$-RDP.

Furthermore, by Lemma C.5, for every $j \in \mathcal{H}$, the corresponding mechanism $\mathcal{M}_j$ taking the dataset $\mathcal{D}_j$ and θ_t as input and returning $\tilde{g}_t^{(j)}$ is $(\alpha, \varepsilon_1(\alpha))$-RDP with

$$\varepsilon_1(\alpha) := (1 - \alpha + \lfloor \alpha \rfloor) \frac{\lfloor \alpha \rfloor - 1}{\alpha - 1} \varepsilon'(\lfloor \alpha \rfloor) + (\alpha - \lfloor \alpha \rfloor) \frac{\lceil \alpha \rceil - 1}{\alpha - 1} \varepsilon'(\lceil \alpha \rceil). \tag{42}$$

Where

$$\begin{aligned}\varepsilon'(\alpha) = \frac{1}{\alpha - 1} \log \Big(1 + r^2 \binom{\alpha}{2} \min \left\{4(e^{\varepsilon(2)} - 1), e^{\varepsilon(2)} \min\left\{2, (e^{\varepsilon(\infty)} - 1)^2\right\}\right\} \\ + \sum_{j=3}^{\alpha} r^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} \min\left\{2, (e^{\varepsilon(\infty)} - 1)^j\right\}\Big),\end{aligned}$$

and $\varepsilon(\alpha) := \frac{\alpha \Delta^2}{2\sigma_{\mathrm{DP}}^2} = \left(\frac{2C}{b}\right)^2 \frac{\alpha}{2\sigma_{\mathrm{DP}}^2}$. Furthermore, since $\varepsilon(\infty) = +\infty$, we get

$$\varepsilon'(\alpha) = \frac{1}{\alpha - 1} \log \Big(1 + r^2 \binom{\alpha}{2} \min \left\{4(e^{\varepsilon(2)} - 1), 2e^{\varepsilon(2)}\right\} + 2 \sum_{j=3}^{\alpha} r^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)}\Big). \tag{43}$$

*Step (**II**):* This step consists in computing the local momentums from the noisy gradients, and then aggregating the momentums and updating the model accordingly. As this process does not have direct access to the datasets $\mathcal{D}_i, i \in \mathcal{H}$, it should be considered as a post-processing operation for Step (**I**). As RDP is preserved by post-processing (Mironov, 2017), we conclude that a single iteration of Algorithm 1 is $(\alpha, \varepsilon_1(\alpha))$-RDP with respect to each worker's data for any $\alpha > 1$, with $\varepsilon_1(\alpha)$ as defined above.

End-to-end privacy. We can now compute the end-to-end DP of our algorithm. First, invoking Lemma C.1 and the per-iteration RDP guarantee of Algorithm 1, we obtain that Algorithm 1 is $(\alpha, T\varepsilon_1(\alpha))$-RDP towards the server, for any $\alpha > 1$. Next, by Lemma C.2, we deduce that Algorithm 1 is $(\varepsilon^*(\alpha), \delta)$-DP towards the server for every $\delta \in (0,1), \alpha > 1$, with

$$\varepsilon^*(\alpha) \coloneqq T\varepsilon_1(\alpha) + \frac{\log(1/\delta)}{\alpha - 1}.$$

This implies that, for any $\delta \in (0,1)$, Algorithm 1 is (ε^*, δ)-DP with

$$\varepsilon^* \coloneqq \inf_{\alpha>1} \varepsilon^*(\alpha) = \inf_{\alpha>1} \Big(T\varepsilon_1(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \Big).$$

The above concludes the proof. □

We now prove the (closed-form) approximate DP guarantees of SAFE-DSHB in Theorem 4.1, as a corollary of Theorem C.1.

Theorem 4.1. *Consider Algorithm 1. Let $\varepsilon > 0, \delta \in (0,1)$ be such that $\varepsilon \leq \log(1/\delta)$. There exists a constant $k > 0$ such that, for a sufficiently small batch size b, when $\sigma_{\mathrm{DP}} \geq k \cdot \frac{2C}{b} \max\left\{1, \frac{b\sqrt{T\log(1/\delta)}}{m\varepsilon}\right\}$, Algorithm 1 is (ε, δ)-DP.*

Proof. Suppose that $\frac{b}{m}$ is sufficiently small. Let $\varepsilon > 0$ and $\delta \in (0,1)$ be such that $\varepsilon \leq \log(1/\delta)$. Finally consider $\Delta, \epsilon^*(\cdot), \epsilon_1(\cdot), \epsilon'(\cdot)$, and $\epsilon(\cdot)$ as defined in the statement and the proof of Theorem C.1. Below, we show that there exists $k > 0$ such that, when $\sigma_{\mathrm{DP}} \geq k \cdot {}^{2C}/_{b} \max\{1, {}^{b\sqrt{T\log(1/\delta)}}/_{m\varepsilon}\}$, Algorithm 1 ensures (ε, δ)-DP towards an honest-but-curious server. First note that, when $\sigma_{\mathrm{DP}} \geq {}^{2C}/_{b}$, we have

$$\varepsilon(2) = \frac{\Delta^2}{\sigma_{\mathrm{DP}}^2} = \frac{({}^{2C}/_{b})^2}{\sigma_{\mathrm{DP}}^2} \leq 1.$$

Since $h \coloneqq x \mapsto \frac{1}{x}(e^x - 1)$ is non-decreasing on $(0, +\infty)$, this also implies that $\frac{1}{\varepsilon(2)}(e^{\varepsilon(2)} - 1) = h(\varepsilon(2)) \leq h(1) = e - 1 \leq 2$. As a result, we have

$$\min\left\{4(e^{\varepsilon(2)} - 1), 2e^{\varepsilon(2)}\right\} \leq 4(e^{\varepsilon(2)} - 1) \leq 8\,\varepsilon(2). \tag{44}$$

Recall that

$$\varepsilon'(\alpha) = \frac{1}{\alpha - 1} \log\Big(1 + r^2 \binom{\alpha}{2} \min\left\{4(e^{\varepsilon(2)} - 1), 2e^{\varepsilon(2)}\right\} + 2\sum_{j=3}^{\alpha} r^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)}\Big). \tag{45}$$

Therefore, since we assume that $\frac{b}{m}$ is sufficiently small ($r \ll 1$), the dominating term inside the logarithm is the term in r^2. Using $\log(1 + x) \leq x$, there exists a constant k' such that

$$\begin{aligned}
\varepsilon'(\alpha) &\leq \frac{1}{\alpha - 1}\left(r^2 \binom{\alpha}{2} \min\left\{4(e^{\varepsilon(2)} - 1), 2e^{\varepsilon(2)}\right\} + 2\sum_{j=3}^{\alpha} r^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} \right)\\
&\leq \frac{k'}{\alpha - 1}\left(r^2 \binom{\alpha}{2} \min\left\{4(e^{\varepsilon(2)} - 1), 2e^{\varepsilon(2)}\right\}\right)\\
&= \frac{k'}{\alpha - 1}\mathcal{O}\Big(r^2\alpha(\alpha - 1)\min\left\{4(e^{\varepsilon(2)} - 1), 2e^{\varepsilon(2)}\right\}\Big).
\end{aligned}$$

Hence substituting from (44), we get

$$\varepsilon'(\alpha) \leq 8k'r^2\alpha\varepsilon(2) = 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\alpha.$$

This directly implies that

$$\begin{aligned}\varepsilon_1(\alpha) &= (1-\alpha+\lfloor\alpha\rfloor)\frac{\lfloor\alpha\rfloor-1}{\alpha-1}\varepsilon'(\lfloor\alpha\rfloor) + (\alpha-\lfloor\alpha\rfloor)\frac{\lceil\alpha\rceil-1}{\alpha-1}\varepsilon'(\lceil\alpha\rceil)\\ &\le 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\Big[(1-\alpha+\lfloor\alpha\rfloor)\frac{\lfloor\alpha\rfloor-1}{\alpha-1}\left\lfloor\alpha\right\rfloor + (\alpha-\lfloor\alpha\rfloor)\frac{\lceil\alpha\rceil-1}{\alpha-1}\left\lceil\alpha\right\rceil\Big].\end{aligned} \tag{46}$$

Now, recall that $\alpha-1 \le \lfloor\alpha\rfloor \le \alpha$ and $\alpha \le \lceil\alpha\rceil \le \alpha+1$. We will prove that $\varepsilon_1(\alpha) \le 32k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}$ by distinguishing two cases:

Case $\alpha \in (1,2)$: Since $\alpha > 1$, we have $\lfloor\alpha\rfloor \ge 1$ and therefore ${}^{\alpha-\lfloor\alpha\rfloor}/_{\alpha-1} \le 1$. We therefore have from Equation (46)

$$\begin{aligned}\varepsilon_1(\alpha) &\le 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\Big[(1-\alpha+\lfloor\alpha\rfloor)\frac{\lfloor\alpha\rfloor-1}{\alpha-1}\left\lfloor\alpha\right\rfloor + (\alpha-\lfloor\alpha\rfloor)\frac{\lceil\alpha\rceil-1}{\alpha-1}\left\lceil\alpha\right\rceil\Big]\\ &\le 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\Big[\underbrace{(1-\alpha+\lfloor\alpha\rfloor)}_{\le 1}\underbrace{\frac{\lfloor\alpha\rfloor-1}{\alpha-1}}_{\le 1}\left\lfloor\alpha\right\rfloor + \underbrace{(\lceil\alpha\rceil-1)}_{\le\alpha}\left\lceil\alpha\right\rceil\Big]\\ &\le 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\Big[\left\lfloor\alpha\right\rfloor + \alpha\left\lceil\alpha\right\rceil\Big] \underset{(i)}{\le} 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\Big[\alpha+2\alpha\Big] = 24k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\alpha,\end{aligned}$$

where (i) is due to $\lceil\alpha\rceil \le 2$ because $\alpha < 2$.

Case $\alpha \in [2,+\infty)$:

Since $\alpha \ge 2$, we have both $\lfloor\alpha\rfloor \le \lceil\alpha\rceil \le \alpha+1 \le 2\alpha$ and $\lfloor\alpha\rfloor - 1 \le \lceil\alpha\rceil - 1 \le 2(\alpha-1)$. Therefore, we have from Equation (46) that

$$\begin{aligned}\varepsilon_1(\alpha) &\le 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\Big[(1-\alpha+\lfloor\alpha\rfloor)\frac{\lfloor\alpha\rfloor-1}{\alpha-1}\left\lfloor\alpha\right\rfloor + (\alpha-\lfloor\alpha\rfloor)\frac{\lceil\alpha\rceil-1}{\alpha-1}\left\lceil\alpha\right\rceil\Big]\\ &\le 8k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\Big[(1-\alpha+\lfloor\alpha\rfloor)4\alpha + (\alpha-\lfloor\alpha\rfloor)4\alpha\Big] = 32k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}.\end{aligned}$$

We have now proved for every $\alpha > 1$ that $\varepsilon_1(\alpha) \le 32k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}$. This implies that

$$\varepsilon^* = \inf_{\alpha>1}\left(T\varepsilon_1(\alpha) + \frac{\log(1/\delta)}{\alpha-1}\right) \le \inf_{\alpha>1}\left(32k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\alpha T + \frac{\log(1/\delta)}{\alpha-1}\right).$$

The above (convex) optimization problem is solved for $\alpha = \alpha^* := 1 + \sigma_{\mathrm{DP}}\sqrt{\frac{\log(1/\delta)}{32k'r^2\Delta^2 T}}$. Remark that the constraint $\alpha > 1$ is satisfied at α^*. Additionally, the objective at $\alpha = \alpha^*$ is equal to

$$32k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}\alpha^* T + \frac{\log(1/\delta)}{\alpha^*-1} = 32k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}T + 2r\Delta\frac{\sqrt{32k'\,T\log(1/\delta)}}{\sigma_{\mathrm{DP}}}.$$

Therefore, using the assumption $\varepsilon \le \log(1/\delta)$, when $\sigma_{\mathrm{DP}} \ge \frac{6C\sqrt{32k'T\log(1/\delta)}}{m\varepsilon} = 3r\Delta\frac{\sqrt{32k'T\log(1/\delta)}}{\varepsilon}$, we have

$$\begin{aligned}\varepsilon^* &\le 32k'r^2\frac{\Delta^2}{\sigma_{\mathrm{DP}}^2}T + 2r\Delta\frac{\sqrt{32k'\,T\log(1/\delta)}}{\sigma_{\mathrm{DP}}}\\ &\le \frac{\varepsilon^2}{9\log(1/\delta)} + {}^{2}/_{3}\,\varepsilon \le ({}^{1}/_{9} + {}^{2}/_{3})\varepsilon \le \varepsilon.\end{aligned}$$

Recall that to derive this last inequality, we overall needed $\sigma_{\mathrm{DP}} \ge {}^{2C}/_{b} = \Delta$ and $\sigma_{\mathrm{DP}} \ge \frac{6C\sqrt{32k'T\log(1/\delta)}}{m\varepsilon} = 3r\Delta\frac{\sqrt{32k'T\log(1/\delta)}}{\varepsilon}$. Therefore, by choosing $k := \max\{1, 3\sqrt{32k'}\}$, we can now conclude that, when $\sigma_{\mathrm{DP}} \ge k\cdot{}^{2C}/_{b}\max\{1,\, {}^{b\sqrt{T\log(1/\delta)}}/_{m\varepsilon}\}$, Algorithm 1 is (ε,δ)-DP. □

D. Upper Bounds

D.1. Proof Outline

Our analysis of SAFE-DSHB (Algorithm 1), inspired from (Farhadkhani et al., 2022), consists of three elements: (i) *Momentum drift* (Lemma D.1), (ii) *Momentum deviation* (Lemma D.2), and (iii) *Descent bound* (Lemma D.3). We combine these elements to obtain the final convergence result stated in Theorem 4.2, and the matching upper bound stated in Corollary 5.1.

Notation. Recall that for each step t, for each honest worker w_i,

$$m_t^{(i)} = \beta_{t-1} m_{t-1}^{(i)} + (1 - \beta_{t-1}) \tilde{g}_t^{(i)}, \tag{47}$$

$$\tilde{g}_t^{(i)} = g_t^{(i)} + \xi_t^{(i)}; \;\; \xi_t^{(i)} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d), \tag{48}$$

where we initialize $m_0^{(i)} = 0$. As we analyze Algorithm 1 with aggregation F, we denote

$$R_t \coloneqq F\Big(m_t^{(1)}, \dots, m_t^{(n)}\Big), \tag{49}$$

$$\theta_{t+1} = \theta_t - \gamma_t R_t. \tag{50}$$

Throughout, we denote the loss function over dataset $\mathcal{D}_i$ by $\mathcal{L}_i = \mathcal{L}(\,\cdot\,; \mathcal{D}_i)$. Also, we denote by $\mathcal{P}_t$ the history from steps 0 to t. Specifically,

$$\mathcal{P}_t \coloneqq \Big\{\theta_0, \dots, \theta_t;\; m_1^{(i)}, \dots, m_{t-1}^{(i)}; i \in [n]\Big\}.$$

By convention, $\mathcal{P}_1 = \{\theta_0\}$. We denote by $\mathbb{E}_t\left[\cdot\right]$ and $\mathbb{E}\left[\cdot\right]$ the conditional expectation $\mathbb{E}\left[\cdot \,|\, \mathcal{P}_t\right]$ and the total expectation, respectively. Thus, $\mathbb{E}\left[\cdot\right] = \mathbb{E}_1\left[\cdots \mathbb{E}_T\left[\cdot\right]\right]$.

D.1.1. Momentum Drift

Along the trajectory $\theta_0, \dots, \theta_t$, the honest workers' local momentums may drift away from each other. The drift has three distinct sources: (i) noise injected by the DP mechanism, (ii) gradient dissimilarity induced by data heterogeneity, and (iii) stochasticity of the mini-batch gradients. The aforementioned drift of local momentums can be exploited by the Byzantine adversaries to maliciously bias the aggregation output.

In this section, we will control the growth of the drift Δ_t between momentums, which we define as

$$\Delta_t \coloneqq \lambda_{\max}\left(\frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} (m_t^{(i)} - \overline{m}_t)(m_t^{(i)} - \overline{m}_t)^\top\right), \tag{51}$$

where $\lambda_{\max}$ denotes the maximum eigenvalue, and $\overline{m}_t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} m_t^{(i)}$ denotes the average honest momentum. We show in Lemma D.1 below that the growth of the drift Δ_t of the momentums can be controlled by tuning the momentum coefficient β_t. The full proof can be found in Appendix D.5.2.

Lemma D.1. *Suppose that assumptions 2.2 and 2.3 hold. Consider Algorithm 1. For every $t \in \{0, \dots, T-1\}$, we have*

$$\mathbb{E}\left[\Delta_{t+1}\right] \leq \beta_t \, \mathbb{E}\left[\Delta_t\right] + 2(1 - \beta_t)^2 \left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1 + \frac{d}{n-f})\right) + (1 - \beta_t) G_{\mathrm{cov}}^2,$$

where $\overline{m}_t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} m_t^{(i)}$, $\sigma_b^2 \coloneqq 2(1 - \frac{b}{m})\frac{\sigma^2}{b}$, and $G_{\mathrm{cov}}^2 \coloneqq \sup_{\theta \in \mathbb{R}^d} \sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} \langle v,\, \nabla \mathcal{L}_i(\theta) - \nabla \mathcal{L}_{\mathcal{H}}(\theta)\rangle^2$.

The dimension factor d due to DP noise is divided by $n - f$, which would not have been possible without leveraging the Gaussian nature of the noise. This dependence will prove crucial to match our lower bound. To leverage Gaussianity, we use a concentration argument on the empirical covariance matrix of Gaussian random variables, stated in Lemma D.6.

The remaining term G_{cov}^2 of the upper bound is only due to data heterogeneity. An important distinction from (Karimireddy et al., 2022) is that G_{cov}^2 is a tighter bound on heterogeneity, compared to G^2 the bound on the average squared distance from Assumption 2.1. This is because the drift Δ_t is not an average squared distance, but rather a bound on average squared distances of every projection on the unit ball. Controlling this quantity requires a covering argument (stated in Lemma D.4).

D.1.2. MOMENTUM DEVIATION

Next, we study the momentum deviation; i.e., the distance between the average honest momentum $\overline{m}_t$ and the true gradient $\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)$ in an arbitrary step t. Specifically, we define momentum *deviation* to be

$$\delta_t \coloneqq \overline{m}_t - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta_t\right). \tag{52}$$

Also, we introduce the error between the aggregate R_t and $\overline{m}_t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} m_t^{(i)}$ the average momentum of honest workers for the case. Specifically, when defining the error

$$\epsilon_t \coloneqq R_t - \overline{m}_t, \tag{53}$$

we get the following bound on the momentum deviation in Lemma D.2, proof of which can be found in Appendix D.5.3.

Lemma D.2. *Suppose that assumptions 2.2 and 2.3 hold and that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Consider Algorithm 1. For all $t \in \{0, \dots, T-1\}$, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] \leq \beta_t^2(1+\gamma_t L)(1+4\gamma_t L)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + 4\gamma_t L(1+\gamma_t L)\beta_t^2\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] \\ + (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)} + 2\gamma_t L(1+\gamma_t L)\beta_t^2\,\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right],\end{aligned}$$

where $\overline{\sigma}^2_{\mathrm{DP}} \coloneqq 2\left(1-\frac{b}{m}\right)\frac{\sigma^2}{b} + d \cdot \sigma^2_{\mathrm{DP}}$.

D.1.3. DESCENT BOUND

Finally, we bound the progress made at each learning step in minimizing the loss $\mathcal{L}_{\mathcal{H}}$ using Algorithm 1. From (50) and (49), we obtain that, for each step t,

$$\theta_{t+1} = \theta_t - \gamma_t R_t = \theta_t - \gamma_t \overline{m}_t - \gamma_t (R_t - \overline{m}_t),$$

Furthermore, by (53), $R_t - \overline{m}_t = \epsilon_t$. Thus, for all t,

$$\theta_{t+1} = \theta_t - \gamma_t \overline{m}_t - \gamma_t \epsilon_t. \tag{54}$$

This means that Algorithm 1 can actually be treated as distributed SGD with a momentum term that is subject to perturbation proportional to ϵ_t at each step t. This perspective leads us to Lemma D.3, proof of which can be found in Appendix D.5.4.

Lemma D.3. *Assume that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Consider Algorithm 1. For any $t \in [T]$, we have*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t)\right] \leq -\frac{\gamma_t}{2}\left(1-4\gamma_t L\right)\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \gamma_t\left(1+2\gamma_t L\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \gamma_t\left(1+\gamma_t L\right)\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right].$$

Putting all of the previous lemmas together, we prove Theorem 4.2 in Section D.2. We then prove Corollary 5.1 in Section D.3, and its non-convex version in Corollary D.1 in Section D.4.

D.2. Proof of Theorem 4.2

We recall the theorem statement below for convenience. Recall that

$$\mathcal{L}_* = \inf_{\theta \in \mathbb{R}^d} \mathcal{L}_{\mathcal{H}}(\theta), \mathcal{L}_0 = \mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*, a_1 = 240, a_2 = 480, a_3 = 5760, \text{ and } a_4 = 270.$$

Theorem 4.2. *Suppose that assumptions 2.2 and 2.3 hold true, and that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Let F satisfy the condition of (f, κ)-robust averaging. We let*

$$\overline{\sigma}^2 = \frac{\sigma_b^2 + d\sigma^2_{\mathrm{DP}}}{n-f} + 4\kappa\left(\sigma_b^2 + 36\sigma^2_{\mathrm{DP}}\left(1 + \frac{d}{n-f}\right)\right),$$

where $\sigma_b^2 = 2(1-\frac{b}{m})\frac{\sigma^2}{b}$. Consider Algorithm 1 with $T \geq 1$, the learning rates γ_t and momentum coefficients β_t specified below. We prove that the following holds, where the expectation $\mathbb{E}\left[\cdot\right]$ is over the randomness of the algorithm.

1. ***Strongly convex:*** *Assume that $\mathcal{L}_{\mathcal{H}}$ is μ-strongly convex. If $\gamma_t = \frac{10}{\mu(t+a_1\frac{L}{\mu})}$ and $\beta_t = 1 - 24L\gamma_t$ then*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_T) - \mathcal{L}_*\right] \leq \frac{4a_1\kappa G^2_{\text{cov}}}{\mu} + \frac{2a_1^2 L\overline{\sigma}^2}{\mu^2 T} + \frac{2a_1^2 L^2 \mathcal{L}_0}{\mu^2 T^2}.$$

2. ***Non-convex:*** *If $\gamma = \min\left\{\frac{1}{24L},\ \frac{\sqrt{a_4\mathcal{L}_0}}{2\overline{\sigma}\sqrt{a_3 LT}}\right\}$ and $\beta_t = 1 - 24L\gamma$ then*

$$\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] \leq a_2\kappa G^2_{\text{cov}} + \frac{\sqrt{a_3 a_4 L\mathcal{L}_0}\,\overline{\sigma}}{\sqrt{T}} + \frac{a_4 L\mathcal{L}_0}{T}.$$

We prove Theorem 4.2 in the strongly convex case in Section D.2.1, and in the non-convex case in Section D.2.2.

D.2.1. Strongly Convex Case

Proof. Let Assumption 2.2 hold and assume that $\mathcal{L}_{\mathcal{H}}$ is L-smooth and μ-strongly convex, and that F is a (f,κ)-robust averaging aggregation rule. Let $t \in \{0,\dots,T-1\}$. We set the learning rate and momentum schedules to be

$$\gamma_t = \frac{10}{\mu(t + a_1\frac{L}{\mu})},\ \beta_t = 1 - 24L\gamma_t, \tag{55}$$

where $a_1 \coloneqq 240$. Note that we have

$$\gamma_t \leq \gamma_0 = \frac{10}{\mu 240\frac{L}{\mu}} = \frac{1}{24L}. \tag{56}$$

To obtain the convergence result we define the Lyapunov function to be

$$V_t \coloneqq \left(t + a_1\frac{L}{\mu}\right)^2 \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_t) - \mathcal{L}_* + \frac{z_1}{L}\left\|\delta_t\right\|^2 + \kappa\cdot\frac{z_2}{L}\Delta_t\right], \tag{57}$$

where $a_1 = 240, z_1 = \frac{1}{16}$, and $z_2 = 2$. Throughout the proof, we denote $\hat{t} \coloneqq t + a_1\frac{L}{\mu}$. Therefore, we have $\gamma_t = \frac{10}{\mu\hat{t}}$. Consider also the auxiliary sequence W_t defined as

$$W_t \coloneqq \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_t) - \mathcal{L}_* + \frac{z_1}{L}\left\|\delta_t\right\|^2 + \kappa\cdot\frac{z_2}{L}\Delta_t\right]. \tag{58}$$

Therefore, we have

$$\begin{aligned} V_{t+1} - V_t &= (\hat{t}+1)^2 W_{t+1} - \hat{t}^2 W_t = (\hat{t}+1)^2 W_{t+1} - (\hat{t}^2 + 2\hat{t} + 1)W_t + (2\hat{t}+1)W_t \\ &= (\hat{t}+1)^2(W_{t+1} - W_t) + (2\hat{t}+1)W_t. \end{aligned} \tag{59}$$

We now bound the quantity $W_{t+1} - W_t$ below.

Invoking Lemma D.1. Upon substituting from Lemma D.1, we obtain

$$\begin{aligned} \mathbb{E}\left[\kappa\cdot\frac{z_2}{L}\Delta_{t+1} - \kappa\cdot\frac{z_2}{L}\Delta_t\right] &\leq \kappa\cdot\frac{z_2}{L}\beta_t\,\mathbb{E}\left[\Delta_t\right] + 2\kappa\cdot\frac{z_2}{L}(1-\beta_t)^2\left(\sigma_b^2 + 36\sigma_{\text{DP}}^2(1 + \frac{d}{n-f})\right) + \kappa\cdot\frac{z_2}{L}(1-\beta_t)G^2_{cov} \\ &\quad - \kappa\cdot\frac{z_2}{L}\,\mathbb{E}\left[\Delta_t\right]. \end{aligned} \tag{60}$$

Invoking Lemma D.2. Upon substituting from Lemma D.2, we obtain

$$\begin{aligned} \mathbb{E}\left[\frac{z_1}{L}\left\|\delta_{t+1}\right\|^2 - \frac{z_1}{L}\left\|\delta_t\right\|^2\right] &\leq \frac{z_1}{L}\beta_t^2 c_t\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + 4z_1\gamma_t(1+\gamma_t L)\beta_t^2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \frac{z_1}{L}(1-\beta_t)^2\frac{\overline{\sigma}^2_{\text{DP}}}{n-f} \\ &\quad + 2z_1\gamma_t(1+\gamma_t L)\beta_t^2\,\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right] - \frac{z_1}{L}\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right], \end{aligned} \tag{61}$$

where we introduced the following quantity for simplicity

$$c_t = (1 + \gamma_t L)\,(1 + 4\gamma_t L) = 1 + 5\gamma_t L + 4\gamma_t^2 L^2. \tag{62}$$

Invoking Lemma D.3. Substituting from Lemma D.3, we obtain

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t)\right] \leq -\frac{\gamma_t}{2}\,(1 - 4\gamma_t L)\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \gamma_t\,(1 + 2\gamma_t L)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \gamma_t\,(1 + \gamma_t L)\,\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right]. \tag{63}$$

Substituting from (60), (61) and (63) in (58), we obtain

$$\begin{aligned} W_{t+1} - W_t &= \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t)\right] + \mathbb{E}\left[\frac{z_1}{L}\left\|\delta_{t+1}\right\|^2 - \frac{z_1}{L}\left\|\delta_t\right\|^2\right] + \mathbb{E}\left[\kappa \cdot \frac{z_2}{L}\Delta_{t+1} - \kappa \cdot \frac{z_2}{L}\Delta_t\right] \\ &\leq -\frac{\gamma_t}{2}\,(1 - 4\gamma_t L)\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \gamma_t\,(1 + 2\gamma_t L)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \gamma_t\,(1 + \gamma_t L)\,\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right] \\ &\quad + \frac{z_1}{L}\beta_t^2 c_t\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + 4z_1\gamma_t(1 + \gamma_t L)\beta_t^2\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \frac{z_1}{L}(1 - \beta_t)^2\frac{\overline{\sigma}_{\mathrm{DP}}^2}{n - f} \\ &\quad + 2z_1\gamma_t(1 + \gamma_t L)\beta_t^2\,\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right] - \frac{z_1}{L}\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] \\ &\quad + \kappa \cdot \frac{z_2}{L}\beta_t\,\mathbb{E}\left[\Delta_t\right] + 2\kappa \cdot \frac{z_2}{L}(1 - \beta_t)^2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1 + \frac{d}{n - f})\right) + \kappa \cdot \frac{z_2}{L}(1 - \beta_t)G_{cov}^2 \\ &\quad - \kappa \cdot \frac{z_2}{L}\,\mathbb{E}\left[\Delta_t\right]. \end{aligned} \tag{64}$$

Upon rearranging the R.H.S. in (64) we obtain that

$$\begin{aligned} W_{t+1} - W_t \leq &-\frac{\gamma_t}{2}\left((1 - 4\gamma_t L) - 8z_1(1 + \gamma_t L)\beta_t^2\right)\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \frac{z_1}{L}(1 - \beta_t)^2\frac{\overline{\sigma}_{\mathrm{DP}}^2}{n - f} \\ &- z_1\gamma_t\left(-\frac{1}{z_1}\,(1 + 2\gamma_t L) - \frac{1}{\gamma_t L}\beta_t^2 c_t + \frac{1}{\gamma_t L}\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \gamma_t\left(1 + \gamma_t L + 2z_1(1 + \gamma_t L)\beta_t^2\right)\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right] \\ &- \kappa \cdot \frac{z_2}{L}(1 - \beta_t)\,\mathbb{E}\left[\Delta_t\right] + 2\kappa \cdot \frac{z_2}{L}(1 - \beta_t)^2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1 + \frac{d}{n - f})\right) + \kappa \cdot \frac{z_2}{L}(1 - \beta_t)G_{cov}^2. \end{aligned} \tag{65}$$

Since we assume F to be (f, κ)-robust averaging, we can bound $\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right]$ as follows. Starting from the definition of ϵ_t, we have

$$\left\|\epsilon_t\right\|^2 = \left\|R_t - \overline{m}_t\right\|^2 = \left\|F(m_t^{(1)}, \ldots, m_t^{(n)}) - \overline{m}_t\right\|^2 \leq \kappa \cdot \lambda_{\max}\left(\frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}}(m_t^{(i)} - \overline{m}_t)(m_t^{(i)} - \overline{m}_t)^\top\right) = \kappa \cdot \Delta_t.$$

Then taking total expectations above gives the bound

$$\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right] \leq \kappa \cdot \mathbb{E}\left[\Delta_t\right].$$

Using the bound above in Equation (65), and then rearranging terms, yields

$$\begin{aligned}
W_{t+1} - W_t \leq & -\frac{\gamma_t}{2}\left((1-4\gamma_t L) - 8z_1(1+\gamma_t L)\beta_t^2\right)\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \frac{z_1}{L}(1-\beta_t)^2\frac{\overline{\sigma}_{\mathrm{DP}}^2}{n-f} \\
& - z_1\gamma_t\left(-\frac{1}{z_1}\left(1+2\gamma_t L\right) - \frac{1}{\gamma_t L}\beta_t^2 c_t + \frac{1}{\gamma_t L}\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \kappa\gamma_t\left(1+\gamma_t L + 2z_1(1+\gamma_t L)\beta_t^2\right)\mathbb{E}\left[\Delta_t\right] \\
& - \kappa \cdot \frac{z_2}{L}(1-\beta_t)\,\mathbb{E}\left[\Delta_t\right] + 2\kappa \cdot \frac{z_2}{L}(1-\beta_t)^2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right) + \kappa \cdot \frac{z_2}{L}(1-\beta_t)G_{cov}^2 \\
= & -\frac{\gamma_t}{2}\left((1-4\gamma_t L) - 8z_1(1+\gamma_t L)\beta_t^2\right)\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \frac{z_1}{L}(1-\beta_t)^2\frac{\overline{\sigma}_{\mathrm{DP}}^2}{n-f} \\
& - z_1\gamma_t\left(-\frac{1}{z_1}\left(1+2\gamma_t L\right) - \frac{1}{\gamma_t L}\beta_t^2 c_t + \frac{1}{\gamma_t L}\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right] \\
& - \kappa z_2\gamma_t\left(\frac{1}{\gamma_t L}(1-\beta_t) - \frac{1}{z_2}\left(1+\gamma_t L + 2z_1(1+\gamma_t L)\beta_t^2\right)\right)\mathbb{E}\left[\Delta_t\right] \\
& + 2\kappa \cdot \frac{z_2}{L}(1-\beta_t)^2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right) + \kappa \cdot \frac{z_2}{L}(1-\beta_t)G_{cov}^2.
\end{aligned}$$

For simplicity, we define

$$A \coloneqq \frac{1}{2}\left(1-4\gamma_t L\right) - 8z_1(1+\gamma_t L)\beta_t^2, \tag{66}$$

$$B \coloneqq -\frac{1}{z_1}\left(1+2\gamma_t L\right) - \frac{1}{\gamma_t L}\beta_t^2 c_t + \frac{1}{\gamma_t L}, \tag{67}$$

and

$$C \coloneqq \frac{1}{\gamma_t L}(1-\beta_t) - \frac{1}{z_2}\left(1+\gamma_t L + 2z_1(1+\gamma_t L)\beta_t^2\right), \tag{68}$$

Denote also

$$\overline{\sigma}^2 \coloneqq \frac{\sigma_b^2 + d\sigma_{\mathrm{DP}}^2}{n-f} + 4\kappa\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right).$$

Recall that, as $z_1 = \frac{1}{16}$ and $z_2 = 2$, and $\overline{\sigma}_{\mathrm{DP}}^2 = \sigma_b^2 + d\sigma_{\mathrm{DP}}^2$, we have

$$\overline{\sigma}^2 \geq z_1\frac{\overline{\sigma}_{\mathrm{DP}}^2}{n-f} + 2\kappa \cdot z_2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right).$$

Thus, substituting the above variables, we obtain

$$\begin{aligned}
W_{t+1} - W_t \leq & -A\gamma_t\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] - z_1 B\gamma_t\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] - \kappa \cdot z_2 C\gamma_t\,\mathbb{E}\left[\Delta_t\right] \\
& + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{z_2}{L}(1-\beta_t)G_{cov}^2.
\end{aligned} \tag{69}$$

We now analyze below the terms A, B and C on the RHS of (69).

Term A. Recall from (56) that $\gamma_t \leq \frac{1}{24L}$. Upon using this in (66), and the facts that $z_1 = \frac{1}{16}$ and $\beta_t^2 \leq 1$, we obtain that

$$A \geq \frac{1}{2}\left(1-4\gamma_t L\right) - 8z_1(1+\gamma_t L) \geq \frac{1}{2}(1-4\times\frac{1}{24}) - \frac{8}{16}(1+\frac{1}{24}) \geq \frac{1}{10}. \tag{70}$$

Term B. Substituting c_t from (62) in (67) we obtain that

$$\begin{aligned}
B &= -\frac{1}{z_1}\left(1+2\gamma_t L\right) - \frac{1}{\gamma_t L}\beta_t^2\left(1+5\gamma_t L + 4\gamma_t^2 L^2\right) + \frac{1}{\gamma_t L} \\
&= \frac{1}{\gamma_t L}\left(1-\beta^2\right) - \frac{1}{z_1}\left(1+2\gamma_t L + 5z_1\beta_t^2 + 4z_1\beta_t^2\gamma_t L\right).
\end{aligned}$$

Using the facts that $\beta_t \leq 1$ and $\gamma_t \leq \frac{1}{24L}$, and then substituting $z_1 = \frac{1}{16}$ we obtain

$$B \geq \frac{1}{\gamma_t L}(1-\beta_t^2) - 16\left(1 + \frac{2}{24} + \frac{5}{16} + \frac{4}{24 \times 16}\right) \geq \frac{1}{\gamma_t L}(1-\beta_t^2) - 23 \geq \frac{1}{\gamma_t L}(1-\beta_t) - 23 = 1. \tag{71}$$

where the last equality follows from the fact that $1 - \beta_t = 24\gamma_t L$.

Term C. Substituting $z_1 = \frac{1}{16}, z_2 = 2$ in (68), and then using the facts that $\beta_t \leq 1$ and $\gamma_t \leq \frac{1}{24L}$, we obtain

$$\begin{aligned} C &= \frac{1}{\gamma_t L}(1-\beta_t) - \frac{1}{2}\left(1 + \gamma_t L + (2 \times 16)(1+\gamma_t L)\beta_t^2\right) \geq \frac{1}{\gamma_t L}(1-\beta_t) - \frac{1}{2}\left(1 + \frac{1}{24} + 32(1 + \frac{1}{24})\right) \\ &\geq \frac{1}{\gamma_t L}(1-\beta_t) - 18 = 6, \end{aligned} \tag{72}$$

where the last equality follows from the fact that $1 - \beta_t = 24\gamma_t L$.

Combining terms A, B, and C. Finally, substituting from (70), (71), and (72) in (69) (and recalling that $z_2 = 2$) we obtain that

$$\begin{aligned} W_{t+1} - W_t \leq &-\frac{\gamma_t}{10}\mathbb{E}\left[\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\|^2\right] - z_1\gamma_t\mathbb{E}\left[\|\delta_t\|^2\right] - 6\kappa z_2 \gamma_t \mathbb{E}\left[\Delta_t\right] \\ &+ \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2. \end{aligned} \tag{73}$$

Since $\mathcal{L}_{\mathcal{H}}$ is μ-strongly convex, we have (Karimi et al., 2016) for any $\theta \in \mathbb{R}^d$ that

$$\|\nabla \mathcal{L}_{\mathcal{H}}(\theta)\|^2 \geq 2\mu(\mathcal{L}(\theta) - \mathcal{L}_*). \tag{74}$$

Plugging (74) in (73) above, and then recalling that $L \geq \mu$, yields

$$\begin{aligned} W_{t+1} - W_t &\leq -\frac{\mu\gamma_t}{5}\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_t) - \mathcal{L}_*\right] - z_1\gamma_t\mathbb{E}\left[\|\delta_t\|^2\right] - 6\kappa z_2\gamma_t\mathbb{E}\left[\Delta_t\right] \\ &\quad + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2 \\ &\leq -\frac{\mu\gamma_t}{5}\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_t) - \mathcal{L}_* + \frac{z_1}{\mu}\|\delta_t\|^2 + \kappa \cdot \frac{z_2}{\mu}\Delta_t\right] + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2 \\ &\leq -\frac{\mu\gamma_t}{5}\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_t) - \mathcal{L}_* + \frac{z_1}{L}\|\delta_t\|^2 + \kappa \cdot \frac{z_2}{L}\Delta_t\right] + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2 \\ &= -\frac{\mu\gamma_t}{5}W_t + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2. \end{aligned}$$

Upon plugging the above bound back in Equation (59), rearranging terms and substituting $1 - \beta_t = 24L\gamma_t$, we obtain

$$\begin{aligned} V_{t+1} - V_t &\leq (\hat{t}+1)^2\left[-\frac{\mu\gamma_t}{5}W_t + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2\right] + (2\hat{t}+1)W_t \\ &= -\left[(\hat{t}+1)^2\frac{\mu\gamma_t}{5} - (2\hat{t}+1)\right]W_t + \frac{(\hat{t}+1)^2}{L}(24L\gamma_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2(\hat{t}+1)^2}{L}(24L\gamma_t)G_{cov}^2. \end{aligned}$$

Recall however that $\gamma_t = \frac{10}{\mu\hat{t}}$ as $\hat{t} = t + a_1\frac{L}{\mu}$. Recall that we denote $a_1 = 24 \times 10 = 240$. Substituting γ_t above yields

$$\begin{aligned} V_{t+1} - V_t &\leq (\hat{t}+1)^2\left[-\frac{\mu\gamma_t}{5}W_t + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2\right] + (2\hat{t}+1)W_t \\ &= -\left[2\frac{(\hat{t}+1)^2}{\hat{t}} - (2\hat{t}+1)\right]W_t + a_1^2L\frac{(\hat{t}+1)^2}{\mu^2\hat{t}^2}\overline{\sigma}^2 + 2a_1\kappa \cdot \frac{(\hat{t}+1)^2}{\mu\hat{t}}G_{cov}^2. \end{aligned}$$

Observe that $2\frac{(\hat{t}+1)^2}{\hat{t}} \geq 2(\hat{t}+1) > 2\hat{t}+1$, implying that the first term above is negative:

$$V_{t+1} - V_t \leq a_1^2L\frac{(\hat{t}+1)^2}{\mu^2\hat{t}^2}\overline{\sigma}^2 + 2a_1\kappa \cdot \frac{(\hat{t}+1)^2}{\mu\hat{t}}G_{cov}^2.$$

Observe now that, as $\hat{t} = t + a_1 \frac{L}{\mu} \geq a_1 = 240$ (because $L \geq \mu$), we have $(\hat{t}+1)^2 \leq (1 + \frac{1}{240})^2 \hat{t}^2 \leq 2\hat{t}^2$. Plugging this bound in the inequality above gives

$$V_{t+1} - V_t \leq \frac{2a_1^2 L}{\mu^2} \overline{\sigma}^2 + 4a_1 \kappa \cdot \frac{\hat{t}}{\mu} G_{cov}^2.$$

Therefore, we have for every $t \in \{0, \dots, T-1\}$ that

$$V_{t+1} - V_0 = \sum_{k=0}^{t} (V_{k+1} - V_k) \leq (t+1) \frac{2a_1^2 L}{\mu^2} \overline{\sigma}^2 + \left(\sum_{k=0}^{t} \hat{k} \right) \frac{4a_1 \kappa}{\mu} G_{cov}^2.$$

Since $\sum_{k=0}^{t} \hat{k} = \sum_{k=0}^{t} (k + a_1 \frac{L}{\mu}) = \sum_{k=0}^{t} k + a_1(t+1)\frac{L}{\mu} = \frac{t(t+1)}{2} + a_1(t+1)\frac{L}{\mu}$, we obtain

$$\begin{aligned} V_{t+1} - V_0 = \sum_{k=0}^{t} (V_{k+1} - V_k) &\leq (t+1) \frac{2a_1^2 L}{\mu^2} \overline{\sigma}^2 + \left(\frac{t(t+1)}{2} + a_1(t+1)\frac{L}{\mu} \right) \frac{4a_1 \kappa}{\mu} G_{cov}^2 \\ &= (t+1) \frac{2a_1^2 L}{\mu^2} \overline{\sigma}^2 + (t+1) \left(\frac{t}{2} + a_1 \frac{L}{\mu} \right) \frac{4a_1 \kappa}{\mu} G_{cov}^2. \end{aligned}$$

However, recalling the definition (57) of V_t, we obtain

$$(t+1+a_1 \frac{L}{\mu})^2 \, \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_* \right] \leq V_{t+1} \leq V_0 + (t+1) \frac{2a_1^2 L}{\mu^2} \overline{\sigma}^2 + (t+1) \left(\frac{t}{2} + a_1 \frac{L}{\mu} \right) \frac{4a_1 \kappa}{\mu} G_{cov}^2.$$

By rearranging terms, and using the fact that $\frac{L}{\mu} \geq 1$, we then get

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_* \right] &\leq \frac{V_0}{(t+1+a_1\frac{L}{\mu})^2} + \frac{t+1}{(t+1+a_1\frac{L}{\mu})^2} \frac{2a_1^2 L \overline{\sigma}^2}{\mu^2} + \frac{(t+1)\left(\frac{t}{2} + a_1 \frac{L}{\mu} \right)}{(t+1+a_1\frac{L}{\mu})^2} \frac{4a_1 \kappa}{\mu} G_{cov}^2 \\ &\leq \frac{V_0}{(t+1+a_1\frac{L}{\mu})^2} + \frac{1}{t+1+a_1\frac{L}{\mu}} \frac{2a_1^2 L \overline{\sigma}^2}{\mu^2} + \frac{4a_1 \kappa}{\mu} G_{cov}^2. \end{aligned} \tag{75}$$

It remains to bound V_0. By definition, we have

$$V_0 = \left(a_1 \frac{L}{\mu} \right)^2 \left[\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_* + \frac{z_1}{L} \left\| \delta_0 \right\|^2 + \frac{z_2}{L} \Delta_0 \right].$$

By definition of $\overline{m}_t = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} m_t^{(i)}$ and the initializations $m_0^{(i)} = 0$ for all $i \in \mathcal{H}$, we have $\Delta_0 = \lambda_{\max}\left(\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} (m_0^{(i)} - \overline{m}_0)(m_0^{(i)} - \overline{m}_0)^\top \right) = 0$. Therefore, we have

$$V_0 = \left(a_1 \frac{L}{\mu} \right)^2 \left[\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_* + \frac{z_1}{L} \left\| \delta_0 \right\|^2 \right].$$

Moreover, by definition of δ_t in (52), we obtain that

$$\left\| \delta_0 \right\|^2 = \left\| \overline{m}_0 - \nabla \mathcal{L}_{\mathcal{H}}(\theta_0) \right\|^2 = \left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_0) \right\|^2.$$

Recall that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Thus, $\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_0) \right\|^2 \leq 2L(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*)$ (see (Nesterov et al., 2018), Theorem 2.1.5). Therefore, substituting $z_1 = \frac{1}{16}$, we have

$$V_0 \leq \left(a_1 \frac{L}{\mu} \right)^2 \left[\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_* + \frac{2L}{16L} (\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*) \right] = \leq \left(a_1 \frac{L}{\mu} \right)^2 \frac{9}{8} (\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*) \leq 2 \left(a_1 \frac{L}{\mu} \right)^2 (\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*).$$

Plugging the above bound back in Equation (75), rearranging terms, and then recalling that $a_1 \frac{L}{\mu} \geq 0$, yields

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_*\right] &\leq \frac{4a_1}{\mu}\kappa G_{cov}^2 + \frac{2a_1^2 L\overline{\sigma}^2}{\mu^2(t+1+a_1\frac{L}{\mu})} + \frac{2a_1L^2(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*)}{\mu^2(t+1+a_1\frac{L}{\mu})^2} \\
&\leq \frac{4a_1}{\mu}\kappa G_{cov}^2 + \frac{2a_1^2 L\overline{\sigma}^2}{\mu^2(t+1)} + \frac{2a_1L^2(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*)}{\mu^2(t+1)^2}.
\end{aligned}$$

Specializing the inequality above for $t = T-1$ and denoting $\mathcal{L}_0 \coloneqq \mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*$ proves the theorem:

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_T) - \mathcal{L}_*\right] \leq \frac{4a_1}{\mu}\kappa G_{cov}^2 + \frac{2a_1^2 L\overline{\sigma}^2}{\mu^2 T} + \frac{2a_1^2 L^2\mathcal{L}_0}{\mu^2 T^2}.$$

□

Remark D.1. In the proof of the strongly convex case of Theorem 4.2 above, we do not need the function $\mathcal{L}_{\mathcal{H}}$ to be μ-strongly convex. In fact, it is sufficient for $\mathcal{L}_{\mathcal{H}}$ to satisfy the μ-PL inequality stated in (74). Accordingly, our results not only apply to smooth μ-strongly convex functions, but more generally to the class of smooth μ-PL functions, which may be non-convex (Karimi et al., 2016).

D.2.2. Non-convex Case

Proof. Let Assumption 2.2 hold and assume that $\mathcal{L}_{\mathcal{H}}$ is L-smooth, and that F is a (f, κ)-robust averaging aggregation rule. Let $t \in \{0, \dots, T-1\}$. We set the learning rate and momentum to constant as follows:

$$\gamma_t = \gamma \coloneqq \min\left\{\frac{1}{24L}, \frac{\sqrt{a_4\mathcal{L}_0}}{2\overline{\sigma}\sqrt{a_3LT}}\right\}, \ \beta_t = \beta \coloneqq 1 - 24L\gamma, \tag{76}$$

where $a_1 \coloneqq 240$. Note that we have

$$\gamma_t = \gamma \leq \frac{1}{24L}. \tag{77}$$

To obtain the convergence result we define the Lyapunov function to be

$$V_t \coloneqq \mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_t) - \mathcal{L}_* + \frac{z_1}{L}\left\|\delta_t\right\|^2 + \kappa \cdot \frac{z_2}{L}\Delta_t\right], \tag{78}$$

where $z_1 = \frac{1}{16}$, and $z_2 = 2$. Note that V_t corresponds to the sequence W_t defined in Equation (58), and analyzed in Appendix D.2.1 under the assumption that $\gamma_t \leq \frac{1}{24L}$. Since the latter holds by Equation (77), we directly apply the bound obtained in Equation (73):

$$\begin{aligned}
V_{t+1} - V_t \leq & -\frac{\gamma_t}{10}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] - z_1\gamma_t\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] - 6\kappa z_2\gamma_t\,\mathbb{E}\left[\Delta_t\right] \\
& + \frac{1}{L}(1-\beta_t)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta_t)G_{cov}^2.
\end{aligned}$$

In turn, substituting $\gamma_t = \gamma, \beta_t = \beta$ and bounding the second and third terms on the RHS by zero, this implies that

$$V_{t+1} - V_t \leq -\frac{\gamma}{10}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \frac{1}{L}(1-\beta)^2\overline{\sigma}^2 + \kappa \cdot \frac{2}{L}(1-\beta)G_{cov}^2.$$

By rearranging terms and then averaging over $t \in \{0, \dots, T-1\}$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] \leq \frac{10}{\gamma T}\sum_{t=0}^{T-1}(V_t - V_{t+1}) + \frac{10}{\gamma L}(1-\beta)^2\overline{\sigma}^2 + \kappa \cdot \frac{20}{\gamma L}(1-\beta)G_{cov}^2.$$

We now substitute $\beta = 1 - 24\gamma L$. Denoting $a_3 = 10 \times 24^2 = 5760, a_2 = 20 \times 24 = 480$, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] &\leq \frac{10}{\gamma T}\sum_{t=0}^{T-1}(V_t - V_{t+1}) + \frac{(10 \times 24^2)}{\gamma L}(\gamma L)^2\overline{\sigma}^2 + \kappa \cdot \frac{(20 \times 24)}{\gamma L}(\gamma L)G_{cov}^2 \\
&= \frac{10}{\gamma T}(V_0 - V_T) + a_3\gamma L\overline{\sigma}^2 + a_2\kappa G_{cov}^2.
\end{aligned} \tag{79}$$

We now bound $V_0 - V_T$. First recall that $V_T \geq 0$ as a sum of non-negative terms (see (78)). Therefore, we have

$$V_0 - V_T \leq V_0 = \mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_* + \frac{z_1}{L} \left\| \delta_0 \right\|^2 + \frac{z_2}{L} \Delta_0.$$

By definition of $\overline{m}_t = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} m_t^{(i)}$ and the initializations $m_0^{(i)} = 0$ for all $i \in \mathcal{H}$, we have $\Delta_0 = \lambda_{\max}\Big(\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} (m_0^{(i)} - \overline{m}_0)(m_0^{(i)} - \overline{m}_0)^\top\Big) = 0$. Therefore, we have

$$V_0 = \mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_* + \frac{z_1}{L} \left\| \delta_0 \right\|^2.$$

Moreover, by definition of δ_t in (52), we obtain that

$$\left\| \delta_0 \right\|^2 = \left\| \overline{m}_0 - \nabla \mathcal{L}_{\mathcal{H}}(\theta_0) \right\|^2 = \left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_0) \right\|^2.$$

Recall that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Thus, $\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_0) \right\|^2 \leq 2L(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*)$ (see (Nesterov et al., 2018), Theorem 2.1.5). Therefore, substituting $z_1 = \frac{1}{16}$, we have

$$V_0 - V_T \leq V_0 \leq \mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_* + \frac{2L}{16L}(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*) = \frac{9}{8}(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*).$$

By plugging this bound back in (79), and denoting $a_4 \coloneqq 24 \times 10 \times (\frac{9}{8}) = 270$ and $\mathcal{L}_0 \coloneqq \mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*$, we obtain

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) \right\|^2 \right] &\leq \frac{10 \times (\frac{9}{8})}{\gamma T} (\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}_*) + a_3 \gamma L \overline{\sigma}^2 + a_2 \kappa G_{cov}^2 \\ &= \frac{a_4 \mathcal{L}_0}{24 \gamma T} + a_3 \gamma L \overline{\sigma}^2 + a_2 \kappa G_{cov}^2. \end{aligned} \tag{80}$$

Recall that by definition

$$\gamma = \min\left\{ \frac{1}{24L}, \frac{\sqrt{a_4 \mathcal{L}_0}}{2\overline{\sigma}\sqrt{a_3 L T}} \right\},$$

and thus $\frac{1}{\gamma} = \max\left\{ 24L, \frac{2}{\sqrt{a_4 \mathcal{L}_0}} \overline{\sigma} \sqrt{a_3 L T} \right\} \leq 24L + \frac{2}{\sqrt{a_4 \mathcal{L}_0}} \overline{\sigma} \sqrt{a_3 L T}$. Therefore, we have

$$\frac{a_4 \mathcal{L}_0}{24 \gamma T} \leq \frac{a_4 \mathcal{L}_0}{24 T} \left(24L + \frac{2}{\sqrt{a_4 \mathcal{L}_0}} \overline{\sigma} \sqrt{a_3 L T} \right) = \frac{a_4 L \mathcal{L}_0}{T} + \frac{\sqrt{a_3 a_4 L \mathcal{L}_0} \overline{\sigma}}{12 \sqrt{T}}.$$

Upon using the above, and that $\gamma \leq \frac{\sqrt{a_4 \mathcal{L}_0}}{2\overline{\sigma}\sqrt{a_3 L T}}$, in (80), we obtain that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) \right\|^2 \right] \leq \frac{a_4 L \mathcal{L}_0}{T} + \frac{\sqrt{a_3 a_4 L \mathcal{L}_0} \overline{\sigma}}{12 \sqrt{T}} + \frac{\sqrt{a_3 a_4 L \mathcal{L}_0} \overline{\sigma}}{2 \sqrt{T}} + a_2 \kappa G_{cov}^2 \leq a_2 \kappa G_{cov}^2 + \frac{\sqrt{a_3 a_4 L \mathcal{L}_0} \overline{\sigma}}{\sqrt{T}} + \frac{a_4 L \mathcal{L}_0}{T}.$$

Finally, recall from Algorithm 1 that $\hat{\theta}$ is chosen randomly from the set of parameter vectors $(\theta_0, \ldots, \theta_{T-1})$. Thus, $\mathbb{E}\left[\left\| \nabla \mathcal{L}_{\mathcal{H}}\left(\hat{\theta}\right) \right\|^2 \right] = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) \right\|^2 \right]$. Substituting this above proves the theorem. □

D.3. Proof of Corollary 5.1

We now state the proof of Corollary 5.1 below.

Corollary 5.1. *Consider Algorithm 1 with aggregation $F = \mathrm{SMEA}$, under the strongly convex setting of Theorem 4.2. Suppose that assumptions 2.1, 2.2, 2.3 hold, and that $n \geq (2+\eta)f$, for some absolute constant $\eta > 0$. Let $\varepsilon > 0, \delta \in (0,1)$ be such that $\varepsilon \leq \log(1/\delta)$. Then, there exists a constant $k > 0$ such that, if $\sigma_{\mathrm{DP}} = k \cdot {}^{2C}/_{b} \max\{1, {}^{b\sqrt{T \log(1/\delta)}}/_{\varepsilon m}\}$, then Algorithm 1 is (ε, δ)-DP and (f, ϱ)-robust where*

$$\varrho = \mathcal{O}\left(\frac{d \log(1/\delta)}{\varepsilon^2 n m^2} + \frac{f}{n} \cdot \frac{\log(1/\delta)}{\varepsilon^2 m^2} + \frac{f}{n} G^2 \right).$$

Proof. Assume that $\mathcal{L}_{\mathcal{H}}$ is L-smooth and μ-strongly convex. Consider Algorithm 1 with aggregation $F = \mathrm{SMEA}$, learning rate $\gamma_t = \frac{10}{\mu(t+a_1\frac{L}{\mu})}$, and momentum coefficient $\beta_t = 1 - 24L\gamma_t$. By Theorem 4.1, the condition on σ_{DP} ensures that Algorithm 1 is (ε, δ)-DP. In the remaining, we prove that Algorithm 1 is (f, ϱ)-robust as stated in the corollary.

First, note that, by Proposition 5.1, SMEA is (f, κ)-robust averaging with $\kappa = \frac{4f}{n-f}(1 + \frac{f}{n-2f})^2$. In fact, as we assume $n \geq (2+\eta)f$ where $\eta > 0$ is an absolute constant, we have

$$\kappa \leq \frac{4f}{n-f}(1+\frac{1}{\eta})^2 = \mathcal{O}(\frac{f}{n-f}). \tag{81}$$

Therefore, thanks to Theorem 4.2, we have

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_T) - \mathcal{L}_*\right] \leq 4a_1\frac{\kappa G_{cov}^2}{\mu} + \frac{2a_1^2 L\overline{\sigma}^2}{\mu^2 T} + \frac{2a_1^2 L^2\mathcal{L}_0}{\mu^2 T^2}, \tag{82}$$

where the constant a_1 is defined as in (3), and

$$\overline{\sigma}^2 := \frac{\sigma_b^2 + d\sigma_{\mathrm{DP}}^2}{n-f} + 4\kappa\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right), \quad \sigma_b^2 := 2(1-\frac{b}{m})\frac{\sigma^2}{b}.$$

We now analyze independently the terms of (82) that depend on T, i.e. the last two terms on the RHS of (82). Recall that, asymptotically in T, the condition on σ_{DP} implies

$$\sigma_{\mathrm{DP}} = k \cdot \frac{2C}{b}\max\{1, {}^{b\sqrt{T\log(1/\delta)}}/_{m\varepsilon}\} = \mathcal{O}\left(\frac{C\sqrt{T\log(1/\delta)}}{m\varepsilon}\right). \tag{83}$$

Term $\frac{2a_1^2 L\overline{\sigma}^2}{\mu^2 T}$. Recalling the expression of $\overline{\sigma}^2$, and using (83) and (81) and the facts that σ_b^2 is independent of T and $f < n-f$, we obtain

$$\begin{aligned}\overline{\sigma}^2 &= \frac{\sigma_b^2 + d\sigma_{\mathrm{DP}}^2}{n-f} + 4\kappa\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right) = \mathcal{O}\bigg(\frac{d\sigma_{\mathrm{DP}}^2}{n-f} + \frac{f}{n-f}\cdot\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\bigg)\\ &= \mathcal{O}\bigg(\frac{d\sigma_{\mathrm{DP}}^2}{n-f} + \frac{f}{n-f}\cdot\sigma_{\mathrm{DP}}^2\bigg) = \mathcal{O}\bigg(\frac{C^2 dT\log(1/\delta)}{m^2(n-f)\varepsilon^2} + \frac{f}{n-f}\cdot\frac{C^2 T\log(1/\delta)}{m^2\varepsilon^2}\bigg).\end{aligned}$$

As a result, we obtain

$$\frac{2a_1^2 L\overline{\sigma}^2}{\mu^2 T} = \mathcal{O}\bigg(\frac{C^2 d\,\log(1/\delta)}{m^2(n-f)\varepsilon^2} + \frac{f}{n-f}\cdot\frac{C^2\log(1/\delta)}{m^2\varepsilon^2}\bigg). \tag{84}$$

Term $\frac{2a_1^2 L^2\mathcal{L}_0}{\mu^2 T^2}$. This term is independent of σ_{DP} and vanishes with T.

Going back to (82), and ignoring terms vanishing in T, and using (81), we obtain

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_T) - \mathcal{L}_*\right] = \mathcal{O}\left(\frac{C^2 d\,\log(1/\delta)}{m^2(n-f)\varepsilon^2} + \frac{f}{n-f}\frac{C^2\log(1/\delta)}{m^2\varepsilon^2} + \frac{f}{n-f}G_{cov}^2\right).$$

Finally, note that $G_{cov}^2 \leq G^2$. Indeed, using the definition of G_{cov}^2 and Assumption 2.1, together with Cauchy-Schwartz, we have

$$G_{cov}^2 = \sup_{\theta\in\mathbb{R}^d}\sup_{\|v\|\leq 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v,\, \nabla\mathcal{L}_i(\theta) - \nabla\mathcal{L}_{\mathcal{H}}(\theta)\rangle^2 \leq \sup_{\theta\in\mathbb{R}^d}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\nabla\mathcal{L}_i(\theta) - \nabla\mathcal{L}_{\mathcal{H}}(\theta)\|^2 \leq G^2.$$

Using the fact above in the last inequality, together with the fact that $n - f \geq \frac{n}{2}$ (as $n > 2f$), we conclude

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_T) - \mathcal{L}_*\right] = \mathcal{O}\left(\frac{C^2 d\,\log(1/\delta)}{m^2 n\varepsilon^2} + \frac{f}{n}\frac{C^2\log(1/\delta)}{m^2\varepsilon^2} + \frac{f}{n}G^2\right).$$

Ignoring the constant C above concludes the proof. □

D.4. Upper Bound for Non-Convex Case

We now state the robustness and DP guarantees of SAFE-DSHB with SMEA in the non-convex case in Corollary D.1 below.

Corollary D.1. *Consider Algorithm 1 with aggregation $F = \mathrm{SMEA}$, under the non-convex setting of Theorem 4.2. Suppose that assumptions 2.1, 2.2, 2.3 hold, that $\mathcal{L}_{\mathcal{H}}$ is L-smooth, and that $n \geq (2+\eta)f$, for some absolute constant $\eta > 0$. Let $\varepsilon > 0, \delta \in (0,1)$ be such that $\varepsilon \leq \log(1/\delta)$. Then, there exists a constant $k > 0$ such that, if $\sigma_{\mathrm{DP}} = k \cdot {}^{2C}/{}_{b} \max\{1, {}^{b\sqrt{T\log(1/\delta)}}/{}_{\varepsilon m}\}$, then Algorithm 1 is (ε, δ)-DP and (f, ϱ)-robust, where*

$$\varrho = \mathcal{O}\left(\frac{\sqrt{d\,\log(1/\delta)}}{\varepsilon\sqrt{n}m} + \sqrt{\frac{f}{n}} \cdot \frac{\sqrt{\log(1/\delta)}}{\varepsilon m} + \frac{f}{n}G^2\right).$$

Proof. Assume that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Consider Algorithm 1 with aggregation $F = \mathrm{SMEA}$, learning rate $\gamma_t = \gamma = \min\left\{\frac{1}{24L}, \frac{\sqrt{a_4\mathcal{L}_0}}{2\overline{\sigma}\sqrt{a_3LT}}\right\}$, and momentum coefficient $\beta_t = \beta = 1 - 24L\gamma$. By Theorem 4.1, the condition on σ_{DP} ensures that Algorithm 1 is (ε, δ)-DP. In the remaining, we prove that Algorithm 1 is (f, ϱ)-robust as stated in the corollary.

First, note that, by Proposition 5.1, SMEA is (f, κ)-robust averaging with $\kappa = \frac{4f}{n-f}(1 + \frac{f}{n-2f})^2$. In fact, as we assume $n \geq (2+\eta)f$ where $\eta > 0$ is an absolute constant, we have

$$\kappa \leq \frac{4f}{n-f}(1+\frac{1}{\eta})^2 = \mathcal{O}(\frac{f}{n-f}). \tag{85}$$

Therefore, thanks to Theorem 4.2, we have

$$\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{H}}(\hat{\theta})\|^2\right] \leq a_2\kappa G_{cov}^2 + \frac{\sqrt{a_3a_4L\mathcal{L}_0}\overline{\sigma}}{\sqrt{T}} + \frac{a_4L\mathcal{L}_0}{T}, \tag{86}$$

where the constants a_1, a_2, a_3, a_4 are defined as in (3), and

$$\overline{\sigma}^2 := \frac{\sigma_b^2 + d\sigma_{\mathrm{DP}}^2}{n-f} + 4\kappa\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right), \quad \sigma_b^2 := 2(1-\frac{b}{m})\frac{\sigma^2}{b}.$$

We now analyze independently the terms of (82) that depend on T, i.e. the last two terms on the RHS of (82). Recall that, asymptotically in T, the condition on σ_{DP} implies

$$\sigma_{\mathrm{DP}} = k \cdot \frac{2C}{b}\max\{1, {}^{b\sqrt{T\log(1/\delta)}}/{}_{m\varepsilon}\} = \mathcal{O}\left(\frac{C\sqrt{T\log(1/\delta)}}{m\varepsilon}\right). \tag{87}$$

Term $\frac{\sqrt{a_3a_4L\mathcal{L}_0}\overline{\sigma}}{\sqrt{T}}$. Recalling the expression of $\overline{\sigma}^2$, and using (83) and (81) and the facts that σ_b^2 is independent of T and $f < n-f$, we obtain

$$\begin{aligned}\overline{\sigma}^2 &= \frac{\sigma_b^2 + d\sigma_{\mathrm{DP}}^2}{n-f} + 4\kappa\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right) = \mathcal{O}\left(\frac{d\sigma_{\mathrm{DP}}^2}{n-f} + \frac{f}{n-f}\cdot\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\right)\\ &= \mathcal{O}\left(\frac{d\sigma_{\mathrm{DP}}^2}{n-f} + \frac{f}{n-f}\cdot\sigma_{\mathrm{DP}}^2\right) = \mathcal{O}\left(\frac{C^2d\,T\log(1/\delta)}{m^2(n-f)\varepsilon^2} + \frac{f}{n-f}\cdot\frac{C^2T\log(1/\delta)}{m^2\varepsilon^2}\right).\end{aligned}$$

Therefore, using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, we obtain

$$\overline{\sigma} = \mathcal{O}\left(\frac{C\sqrt{d\,T\log(1/\delta)}}{m\sqrt{n-f}\varepsilon} + \sqrt{\frac{f}{n-f}}\cdot\frac{C\sqrt{T\log(1/\delta)}}{m\varepsilon}\right).$$

As a result, we obtain

$$\frac{\sqrt{a_3a_4L\mathcal{L}_0}\overline{\sigma}}{\sqrt{T}} = \mathcal{O}\left(\frac{C\sqrt{d\,\log(1/\delta)}}{m\sqrt{n-f}\varepsilon} + \sqrt{\frac{f}{n-f}}\cdot\frac{C\sqrt{\log(1/\delta)}}{m\varepsilon}\right). \tag{88}$$

Term $\frac{a_4 L \mathcal{L}_0}{T}$. This term is independent of σ_{DP} and vanishes with T.

Going back to (86), ignoring terms vanishing in T, and using (85), we obtain

$$\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{H}}(\hat{\theta})\|^2\right] = \mathcal{O}\left(\frac{C\sqrt{d\,\log\left(1/\delta\right)}}{m\sqrt{n-f}\varepsilon} + \sqrt{\frac{f}{n-f}}\cdot\frac{C\sqrt{\log\left(1/\delta\right)}}{m\varepsilon} + \frac{f}{n-f}G_{cov}^2\right).$$

Finally, note that $G_{cov}^2 \leq G^2$. Indeed, using the definition of G_{cov}^2 and Assumption 2.1, together with Cauchy-Schwartz, we have

$$G_{cov}^2 = \sup_{\theta\in\mathbb{R}^d}\sup_{\|v\|\leq 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\langle v,\, \nabla\mathcal{L}_i(\theta) - \nabla\mathcal{L}_{\mathcal{H}}(\theta)\right\rangle^2 \leq \sup_{\theta\in\mathbb{R}^d}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\nabla\mathcal{L}_i(\theta) - \nabla\mathcal{L}_{\mathcal{H}}(\theta)\right\|^2 \leq G^2.$$

Using the fact above in the last inequality, together with the fact that $n-f \geq \frac{n}{2}$ (as $n > 2f$), we conclude

$$\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{H}}(\hat{\theta})\|^2\right] = \mathcal{O}\left(\frac{C\sqrt{d\,\log\left(1/\delta\right)}}{m\sqrt{n}\varepsilon} + \sqrt{\frac{f}{n}}\cdot\frac{C\sqrt{\log\left(1/\delta\right)}}{m\varepsilon} + \frac{f}{n}G^2\right).$$

Ignoring the constant C above concludes the proof. □

Discussion. We conjecture that the non-convex upper bound can be improved as observed recently in the centralized DP setting using other variance reduction techniques (Arora et al., 2022). Nevertheless, both in the centralized and distributed settings, it remains an open question to derive tight lower bounds for non-convex problems.

D.5. Proof of Supporting Lemmas

Before proving Lemmas D.1 to D.3 in Appendices D.5.2 to D.5.4, respectively, we first present some additional results in Appendix D.5.1 below.

D.5.1. Technical Lemmas

Lemma D.4. *Let $M \in \mathbb{R}^{d\times d}$ be a random real symmetric matrix and $g\colon \mathbb{R}\to\mathbb{R}$ an increasing function. It holds that*

$$\mathbb{E}\left[\sup_{\|v\|\leq 1} g(\langle v,\, Mv\rangle)\right] \leq 9^d \cdot \sup_{\|v\|\leq 1}\mathbb{E}\left[g(2\left\langle v,\, Mv\right\rangle)\right].$$

Proof. Let $M \in \mathbb{R}^{d\times d}$ be a random real symmetric matrix and $g\colon \mathbb{R}\to\mathbb{R}$ a increasing function.

The proof follows the construction of (Section 5.2, (Vershynin, 2010)). Recall from standard covering net results (Vershynin, 2010) that we can construct $\mathcal{N}_{1/4}$ a finite $1/4$-net of the unit ball, i.e., for any vector v in the unit ball, there exists $u_v \in \mathcal{N}_{1/4}$ such that $\|u_v - v\| \leq 1/4$. Moreover, we have the bound $\left|\mathcal{N}_{1/4}\right| \leq (1+2/(1/4))^d = 9^d$. Denote by $\|M\| := \sup_{\|v\|\leq 1}\|Mv\|$ the operator norm of M. By recalling that M is symmetric, we obtain for any v in the unit ball

$$\begin{aligned}|\langle v,\, Mv\rangle - \langle u_v,\, Mu_v\rangle| &= |\langle v+u_v,\, M(v-u_v)\rangle| \leq \|v+u_v\|\,\|M(v-u_v)\| \leq (\|v\| + \|u_v\|)\,\|M(v-u_v)\| \\ &\leq 2\,\|M(v-u_v)\| \leq 2\,\|M\|\,\|v-u_v\| \leq 2\,\|M\|/4 = \|M\|/2.\end{aligned}$$

Therefore, we have $\langle v,\, Mv\rangle - \langle u_v,\, Mu_v\rangle \leq \|M\|/2$, and $\langle v,\, Mv\rangle - \|M\|/2 \leq \langle u_v,\, Mu_v\rangle \leq \sup_{u\in\mathcal{N}_{1/4}}\langle u,\, Mu\rangle$. Recall that since M is symmetric, its operator norm coincides with its maximum eigenvalue: $\|M\| = \sup_{\|v\|\leq 1}\langle v,\, Mv\rangle$. We therefore deduce that

$$\sup_{\|v\|\leq 1}\langle v,\, Mv\rangle \leq 2\cdot \sup_{v\in\mathcal{N}_{1/4}}\langle v,\, Mv\rangle.$$

Upon composing with g, which is increasing, we get

$$\sup_{\|v\|\leq 1} g(\langle v,\, Mv\rangle) = g\left(\sup_{\|v\|\leq 1}\langle v,\, Mv\rangle\right) \leq g\left(2\cdot\sup_{v\in\mathcal{N}_{1/4}}\langle v,\, Mv\rangle\right) = \sup_{v\in\mathcal{N}_{1/4}} g(2\left\langle v,\, Mv\right\rangle).$$

Upon taking expectations and applying union bound, we finally conclude

$$\mathbb{E}\left[\sup_{\|v\|\leq 1} g(\langle v,\, Mv\rangle)\right] \leq \mathbb{E}\left[\sup_{v\in\mathcal{N}_{1/4}} g(2\,\langle v,\, Mv\rangle)\right] \leq \left|\mathcal{N}_{1/4}\right| \cdot \sup_{v\in\mathcal{N}_{1/4}} \mathbb{E}\left[g(2\,\langle v,\, Mv\rangle)\right] \leq 9^d \cdot \sup_{\|v\|\leq 1} \mathbb{E}\left[g(2\,\langle v,\, Mv\rangle)\right].$$

□

Lemma D.5. *Suppose assumptions 2.2 and 2.3 hold. For any $t \in \{0, \dots, T-1\}$ and $i \in \mathcal{H}$, we have*

$$\mathbb{E}\left[\left\|\tilde{g}_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2\right] \leq 2\left(1 - \frac{b}{m}\right)\frac{\sigma^2}{b} + d \cdot \sigma_{\mathrm{DP}}^2.$$

Proof. Suppose assumptions 2.2 and 2.3 hold. Let $i \in \mathcal{H}$ and $t \in \{0, \dots, T-1\}$.

First recall from (48) that, since $\tilde{g}_t^{(i)} = g_t^{(i)} + \xi_t^{(i)}, \xi_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$, we have

$$\mathbb{E}_{\xi_t^{(i)}}\left[\left\|\tilde{g}_t^{(i)} - g_t^{(i)}\right\|^2\right] = \mathbb{E}\left[\left\|\xi_t^{(i)}\right\|^2\right] = d \cdot \sigma_{\mathrm{DP}}^2.$$

Next, we have

$$\begin{aligned}\left\|\tilde{g}_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2 &= \left\|\tilde{g}_t^{(i)} - g_t^{(i)} + g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2 \\ &= \left\|\tilde{g}_t^{(i)} - g_t^{(i)}\right\|^2 + \left\|g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2 + 2\left\langle \tilde{g}_t^{(i)} - g_t^{(i)},\, g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\rangle.\end{aligned}$$

Now taking expectation on the randomness of $\xi_t^{(i)}$ (independent of all other random variables), and since $\mathbb{E}\left[\xi_t^{(i)}\right] = 0$, we get

$$\begin{aligned}\mathbb{E}_{\xi_t^{(i)}}\left[\left\|\tilde{g}_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2\right] &= \mathbb{E}_{\xi_t^{(i)}}\left[\left\|\tilde{g}_t^{(i)} - g_t^{(i)}\right\|^2\right] + \left\|g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2 + 2\left\langle \underbrace{\mathbb{E}_{\xi_t^{(i)}}\left[\tilde{g}_t^{(i)} - g_t^{(i)}\right]}_{=\mathbb{E}\left[\xi_t^{(i)}\right]=0},\, g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\rangle \\ &= \mathbb{E}_{\xi_t^{(i)}}\left[\left\|\tilde{g}_t^{(i)} - g_t^{(i)}\right\|^2\right] + \left\|g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2.\end{aligned}$$

Upon taking total expectation, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{g}_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2\right] &= \mathbb{E}\left[\left\|\tilde{g}_t^{(i)} - g_t^{(i)}\right\|^2\right] + \mathbb{E}\left[\left\|g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2\right] + d \cdot \sigma_{\mathrm{DP}}^2. \end{aligned} \tag{89}$$

First observe that when $m = 1$, as $b \in [m]$, we must have $b = m$. Thus, the gradient is deterministic, i.e., $g_t^{(i)} = \nabla\mathcal{L}_i(\theta_t)$. Thus, the first term in the equation above is zero, and the claimed bound holds.

Else, when $m \geq 2$, recall that from Assumption 2.2, we have $\mathbb{E}_{x\sim\mathcal{U}(\mathcal{D}_i)}\left[\|\nabla_\theta \ell(\theta_t; x) - \nabla\mathcal{L}_i(\theta_t)\|^2\right] \leq \sigma^2$. From (Rice, 2006), the variance reduction due to subsampling without replacement gives

$$\mathbb{E}\left[\left\|g_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2\right] \leq \left(1 - \frac{b-1}{m-1}\right)\frac{\sigma^2}{b}.$$

Plugging this bound back in Equation (89) yields

$$\mathbb{E}\left[\left\|\tilde{g}_t^{(i)} - \nabla\mathcal{L}_i(\theta_t)\right\|^2\right] \leq \left(1 - \frac{b-1}{m-1}\right)\frac{\sigma^2}{b} + d \cdot \sigma_{\mathrm{DP}}^2.$$

By observing, as $m \geq 2$, that $1 - \frac{b-1}{m-1} = \frac{m-b}{m-1} = \frac{m}{m-1} \cdot \frac{m-b}{m} = (1 + \frac{1}{m-1})(1 - \frac{b}{m}) \leq 2(1 - \frac{b}{m})$, we obtain the final result:

$$\mathbb{E}\left[\left\|\tilde{g}_t^{(i)} - \nabla \mathcal{L}_i(\theta_t)\right\|^2\right] \leq 2\left(1 - \frac{b}{m}\right)\frac{\sigma^2}{b} + d \cdot \sigma_{\mathrm{DP}}^2.$$

□

Lemma D.6. *Let $\sigma_{\mathrm{DP}} \geq 0$ and $d, n \geq 1$. Consider $\xi^{(1)}, \ldots, \xi^{(n)}$ to be i.i.d. random variables drawn from the Gaussian distribution $\mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$. We have*

$$\mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{n} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right] \leq 36\sigma_{\mathrm{DP}}^2 \left(1 + \frac{d}{n}\right).$$

Proof. Let $\sigma_{\mathrm{DP}} \geq 0$ and $d, n \geq 1$. Consider $\xi^{(1)}, \ldots, \xi^{(n)}$ to be i.i.d. random variables drawn from the Gaussian distribution $\mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$.

If $\sigma_{\mathrm{DP}} = 0$, then $\xi^{(i)} = 0$ almost surely for every $i \in [n]$, and the remainder of the proof holds with $\sigma_{\mathrm{DP}} = 0$. Else, we assume $\sigma_{\mathrm{DP}} > 0$ in the remaining.

Thus, the law of the random variable ${}^{\xi^{(i)}}/_{\sigma_{\mathrm{DP}}}$ is $\mathcal{N}(0, I_d)$ for every $i \in [n]$. Thus, for every vector v in the unit ball, the random variable $\left\langle v,\, {}^{\xi^{(i)}}/_{\sigma_{\mathrm{DP}}} \right\rangle$ is sub-Gaussian with variance equal to 1 (see Chapter 1, (Rigollet & Hütter, 2015)). Therefore, for every $i \in [n]$ and every vector v in the unit ball, applying Theorem 2.1.1 in (Pauwels, 2020)), we have

$$\mathbb{E}\left[\exp\left(\frac{1}{8}\left\langle v,\, {}^{\xi^{(i)}}/_{\sigma_{\mathrm{DP}}} \right\rangle^2\right)\right] \leq 2.$$

As a result, by the independence of $\xi^{(i)}$'s, we obtain

$$\sup_{\|v\| \leq 1} \mathbb{E}\left[\exp\left(\frac{1}{8\sigma_{\mathrm{DP}}^2} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right)\right] = \sup_{\|v\| \leq 1} \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{1}{8}\left\langle v,\, {}^{\xi^{(i)}}/_{\sigma_{\mathrm{DP}}} \right\rangle^2\right)\right] \leq 2^n.$$

Now, observe that we can write $\sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2$ as the quadratic form $\langle v,\, Mv \rangle$, where $M := \sum_{i=1}^n \xi^{(i)} \cdot {\xi^{(i)}}^\top$ is a random real symmetric matrix. Thus, applying Lemma D.4 with the increasing function $g(\cdot) = \exp\left(\frac{1}{16\sigma_{\mathrm{DP}}^2} \times \cdot\right)$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\|v\| \leq 1} \exp\left(\frac{1}{16\sigma_{\mathrm{DP}}^2} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right)\right] &= \mathbb{E}\left[\sup_{\|v\| \leq 1} g(\langle v,\, Mv \rangle)\right] \leq 9^d \cdot \sup_{\|v\| \leq 1} \mathbb{E}\left[g(2\langle v,\, Mv \rangle)\right] \\ &= 9^d \cdot \sup_{\|v\| \leq 1} \mathbb{E}\left[\exp\left(\frac{1}{8\sigma_{\mathrm{DP}}^2} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right)\right] \leq 9^d \cdot 2^n.\end{aligned}$$

We can now use this inequality to bound the term of interest. We apply Jensen's inequality thanks to $\exp$ being convex, and we also interchange $\exp$ and $\sup$ thanks to the former being increasing:

$$\begin{aligned}\exp\left(\frac{1}{16\sigma_{\mathrm{DP}}^2} \mathbb{E}\left[\sup_{\|v\| \leq 1} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right]\right) &\leq \mathbb{E}\left[\exp\left(\frac{1}{16\sigma_{\mathrm{DP}}^2} \sup_{\|v\| \leq 1} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right)\right] \\ &= \mathbb{E}\left[\sup_{\|v\| \leq 1} \exp\left(\frac{1}{16\sigma_{\mathrm{DP}}^2} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right)\right] \leq 9^d \cdot 2^n.\end{aligned}$$

Upon taking $\ln$ and multiplying by ${}^{16\sigma_{\mathrm{DP}}^2}/_n$ on both sides, we obtain that

$$\mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{n} \sum_{i=1}^n \left\langle v,\, \xi^{(i)} \right\rangle^2\right] \leq 16\frac{\sigma_{\mathrm{DP}}^2}{n}(d \ln 9 + n \ln 2) \leq 36\frac{\sigma_{\mathrm{DP}}^2}{n}(d + n) = 36\sigma_{\mathrm{DP}}^2 \left(1 + \frac{d}{n}\right).$$

The above concludes the lemma. □

D.5.2. Proof of Lemma D.1

Lemma D.1. *Suppose that assumptions 2.2 and 2.3 hold. Consider Algorithm 1. For every $t \in \{0, \dots, T-1\}$, we have*

$$\mathbb{E}\left[\Delta_{t+1}\right] \leq \beta_t \, \mathbb{E}\left[\Delta_t\right] + 2(1-\beta_t)^2 \left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2(1 + \frac{d}{n-f})\right) + (1-\beta_t) G_{\mathrm{cov}}^2,$$

where $\overline{m}_t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} m_t^{(i)}$, $\sigma_b^2 \coloneqq 2(1 - \frac{b}{m})\frac{\sigma^2}{b}$, *and* $G_{\mathrm{cov}}^2 \coloneqq \sup_{\theta \in \mathbb{R}^d} \sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} \langle v,\, \nabla \mathcal{L}_i(\theta) - \nabla \mathcal{L}_{\mathcal{H}}(\theta) \rangle^2$.

Proof. Let $t \in \{0, \dots, T-1\}$. Suppose that Assumption 2.2 holds. Recall that the alternate definition of maximum eigenvalue implies, following the definition of Δ_t in Equation (51), that

$$\Delta_t = \lambda_{\max}\left(\frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} (m_t^{(i)} - \overline{m}_t)(m_t^{(i)} - \overline{m}_t)^\top\right) = \sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle^2.$$

We will use the latter expression above for Δ_t throughout this lemma.

For every $i \in \mathcal{H}$, by definition of $m_t^{(i)}$, given in Equation (47), we have

$$m_{t+1}^{(i)} = \beta_t m_t^{(i)} + (1-\beta_t)\tilde{g}_{t+1}^{(i)}.$$

We also denote $\overline{m}_t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} m_t^{(i)}$ and $\overline{\tilde{g}}_{t+1} \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} \tilde{g}_{t+1}^{(i)}$. Therefore, we have $\overline{m}_{t+1} = \beta_t \overline{m}_t + (1-\beta_t)\overline{\tilde{g}}_{t+1}$. As a result, we can write for every $i \in \mathcal{H}$

$$\begin{aligned}
m_{t+1}^{(i)} - \overline{m}_{t+1} &= \beta_t(m_t^{(i)} - \overline{m}_t) + (1-\beta_t)(\tilde{g}_{t+1}^{(i)} - \overline{\tilde{g}}_{t+1}) \\
&= \beta_t(m_t^{(i)} - \overline{m}_t) + (1-\beta_t)(\nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1})) \\
&\quad + (1-\beta_t)(\tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\tilde{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1})).
\end{aligned}$$

By projecting the above expression on an arbitrary vector v and then taking squares, we obtain

$$\begin{aligned}
\left\langle v,\, m_{t+1}^{(i)} - \overline{m}_{t+1} \right\rangle^2 &= \Big[\beta_t \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle + (1-\beta_t) \left\langle v,\, \nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle \\
&\quad + (1-\beta_t)\left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\tilde{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\Big]^2 \\
&= \beta_t^2 \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle^2 + (1-\beta_t)^2 \left\langle v,\, \nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle^2 \\
&\quad + (1-\beta_t)^2 \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\tilde{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle^2 \\
&\quad + 2\beta_t(1-\beta_t)\left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle \\
&\quad + 2\beta_t(1-\beta_t)\left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\tilde{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle \\
&\quad + 2\beta_t(1-\beta_t)\left\langle v,\, \nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\tilde{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle.
\end{aligned}$$

Upon averaging over $i \in \mathcal{H}$, taking the supremum over the unit ball, and then total expectations, we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\langle v,\, m_{t+1}^{(i)} - \overline{m}_{t+1} \right\rangle^2\right] &= \beta_t^2\, \mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle^2\right] \\
&+ (1-\beta_t)^2\, \mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\langle v,\, \nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle^2\right] \\
&+ (1-\beta_t)^2\, \mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle^2\right] \\
&+ 2\beta_t(1-\beta_t)\, \mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\right] \\
&+ 2\beta_t(1-\beta_t)\, \mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\right] \\
&+ 2\beta_t(1-\beta_t)\, \mathbb{E}\left[\sup_{\|v\| \leq 1} \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left\langle v,\, \nabla \mathcal{L}_i(\theta_{t+1}) - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\right].
\end{aligned} \tag{90}$$

We now show that the last two terms on the RHS of Equation (90) are non-positive. We show it for the first one, as the second one can be shown to be non-positive in the same way.

First, note that we can write the inner expression as a quadratic form. Precisely, we have for any vector v and any $i \in \mathcal{H}$ that

$$2 \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle = \langle v,\, Mv \rangle\,,$$

where we have introduced the $d \times d$ matrix $M := N + N^\top$, such that $N := \sum_{i \in \mathcal{H}} (m_t^{(i)} - \overline{m}_t)(\tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}))^\top$. By observing that M is symmetric, we can apply Lemma D.4 with g being the identity mapping:

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|v\| \leq 1} 2 \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\right] &= \mathbb{E}\left[\sup_{\|v\| \leq 1} \langle v,\, Mv \rangle\right] \\
&\leq 9^d \cdot \sup_{\|v\| \leq 1} \mathbb{E}\left[2 \langle v,\, Mv \rangle\right].
\end{aligned} \tag{91}$$

However, the last term is zero by the total law of expectation. Indeed, recall that stochastic gradients are unbiased (Assumption 2.2) and that θ_{t+1} and $m_t^{(i)}$ are deterministic when given history $\mathcal{P}_{t+1}$. This gives

$$\begin{aligned}
\mathbb{E}\left[\langle v,\, Mv \rangle\right] &= \mathbb{E}\left[2 \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}_{t+1}\left[2 \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\right]\right] \\
&= \mathbb{E}\left[2 \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \underbrace{\mathbb{E}_{t+1}\left[\tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1})\right]}_{=0} - \underbrace{\mathbb{E}_{t+1}\left[\overline{\overline{g}}_{t+1} - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right]}_{=0} \right\rangle\right] = 0.
\end{aligned}$$

Moreover, going back to Equation (91), we obtain

$$\mathbb{E}\left[\sup_{\|v\| \leq 1} 2 \sum_{i \in \mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t \right\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla \mathcal{L}_i(\theta_{t+1}) - \overline{\overline{g}}_{t+1} + \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) \right\rangle\right] \leq 9^d \cdot \sup_{\|v\| \leq 1} \mathbb{E}\left[2 \langle v,\, Mv \rangle\right] = 0.$$

As mentioned previously, we can prove in the same way that

$$\mathbb{E}\left[\sup_{\|v\|\leq 1} 2\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1}) - \overline{\widetilde{g}}_{t+1} + \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle\right] \leq 0.$$

Plugging the two previous bounds back in Equation (90), we have thus proved that

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_{t+1}^{(i)} - \overline{m}_{t+1}\right\rangle^2\right] &= \beta_t^2\, \mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2\right] \\
&\quad + (1-\beta_t)^2\, \mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle^2\right] \\
&\quad + (1-\beta_t)^2\, \mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1}) - \overline{\widetilde{g}}_{t+1} + \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle^2\right] \\
&\quad + 2\beta_t(1-\beta_t)\, \mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle\right]. \qquad (92)
\end{aligned}$$

We now bound the two last terms on the RHS of Equation (92).

First, by using the fact that $2ab \leq a^2 + b^2$, we have for any vector v that

$$\begin{aligned}
&\frac{2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle \leq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left[\left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2 + \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle^2\right] \\
&\qquad = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2 + \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle^2. \qquad (93)
\end{aligned}$$

Taking the supremum over the unit ball and then total expectations on both sides yields

$$\begin{aligned}
&2\,\mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle\right] \\
&\qquad \leq \mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2\right] + \mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle^2\right]. \qquad (94)
\end{aligned}$$

Second, recall that $\tilde{g}_{t+1}^{(i)} = g_{t+1}^{(i)} + \xi_{t+1}^{(i)}$, where $\xi_{t+1}^{(i)} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$. Denote $\overline{\xi}_{t+1} \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \xi_{t+1}^{(i)}$. Therefore, by applying Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1}) - \overline{\widetilde{g}}_{t+1} + \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle^2\right] \\
&\quad = \mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, g_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1}) - \overline{g}_{t+1} + \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) + \xi_{t+1}^{(i)} - \overline{\xi}_{t+1}\right\rangle^2\right] \\
&\quad \leq 2\,\mathbb{E}\left[\sup_{\|v\|\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left[\left\langle v,\, g_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1}) - \overline{g}_{t+1} + \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle^2 + \left\langle v,\, \xi_{t+1}^{(i)} - \overline{\xi}_{t+1}\right\rangle^2\right]\right]
\end{aligned}$$

Recall the following bias-variance decomposition: for any $x_1, \ldots, x_n \in \mathbb{R}$, denoting $\overline{x} \coloneqq \frac{1}{n}\sum_{i=1}^n x_i$, we have $\frac{1}{n}\sum_{i=1}^n (x_i -$

$\overline{x})^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \overline{x}^2 \le \sum_{i=1}^n x_i^2$. Applying this fact above yields

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1}) - \overline{\widetilde{g}}_{t+1} + \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle^2\right] \\
&\le 2\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left[\left\langle v,\, g_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1})\right\rangle^2 + \left\langle v,\, \xi_{t+1}^{(i)}\right\rangle^2\right]\right] \\
&\le 2\,\mathbb{E}\left[\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\|g_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1})\right\|^2\right] + 2\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, \xi_{t+1}^{(i)}\right\rangle^2\right],
\end{aligned} \tag{95}$$

where the last inequality is due to the Cauchy-Schwartz inequality. Recall that, by Assumption 2.2 and Lemma D.5 applied with zero privacy noise, we have for every $i \in \mathcal{H}$ that $\mathbb{E}\left[\left\|g_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1})\right\|^2\right] \le 2(1-\frac{b}{m})\frac{\sigma^2}{b} =: \sigma_b^2$. Therefore, upon averaging over $i \in \mathcal{H}$, we have

$$\mathbb{E}\left[\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\|g_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1})\right\|^2\right] \le \sigma_b^2. \tag{96}$$

We now bound the remaining (last) term on the RHS of (95). By applying Lemma D.6 to the random variables $(\xi_{t+1}^{(i)})_{i\in\mathcal{H}}$ which are drawn i.i.d. from $\mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$, we obtain

$$\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, \xi_{t+1}^{(i)}\right\rangle^2\right] \le 36\sigma_{\mathrm{DP}}^2\left(1 + \frac{d}{n-f}\right). \tag{97}$$

Plugging the bounds obtained in (96) and (97) back in (95), we get

$$\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, \tilde{g}_{t+1}^{(i)} - \nabla\mathcal{L}_i(\theta_{t+1}) - \overline{\widetilde{g}}_{t+1} + \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle^2\right] \le 2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2\left(1 + \frac{d}{n-f}\right)\right). \tag{98}$$

We use the above bounds in (94) and (98) to bound the RHS of (92), which yields

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_{t+1}^{(i)} - \overline{m}_{t+1}\right\rangle^2\right] \le \beta_t^2\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2\right] \\
&\quad + (1-\beta_t)^2\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle^2\right] + 2(1-\beta_t)^2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2\left(1 + \frac{d}{n-f})\right)\right) \\
&\quad + \beta_t(1-\beta_t)\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2 + \sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle^2\right].
\end{aligned}$$

By rearranging terms, and noticing that $\beta_t^2 + \beta_t(1-\beta_t) = \beta_t$ and $(1-\beta_t)^2 + \beta_t(1-\beta_t) = 1 - \beta_t$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_{t+1}^{(i)} - \overline{m}_{t+1}\right\rangle^2\right] \le \beta_t\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2\right] \\
&\quad + (1-\beta_t)\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta_{t+1}) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\rangle^2\right] + 2(1-\beta_t)^2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2\left(1 + \frac{d}{n-f}\right)\right).
\end{aligned}$$

Denote $G_{\mathrm{cov}}^2 \coloneqq \sup_{\theta\in\mathbb{R}^d}\sup_{\|v\|\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v,\, \nabla\mathcal{L}_i(\theta) - \nabla\mathcal{L}_{\mathcal{H}}(\theta)\rangle^2$. Then, the above bound implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_{t+1}^{(i)} - \overline{m}_{t+1}\right\rangle^2\right] \le \beta_t\,\mathbb{E}\left[\sup_{\|v\|\le 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\langle v,\, m_t^{(i)} - \overline{m}_t\right\rangle^2\right] \\
&\qquad\qquad + 2(1-\beta_t)^2\left(\sigma_b^2 + 36\sigma_{\mathrm{DP}}^2\left(1 + \frac{d}{n-f}\right)\right) + (1-\beta_t)G_{\mathrm{cov}}^2.
\end{aligned}$$

The above inequality concludes the proof. □

D.5.3. Proof of Lemma D.2

Lemma D.2. *Suppose that assumptions 2.2 and 2.3 hold and that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Consider Algorithm 1. For all $t \in \{0, \dots, T-1\}$, we have*

$$\mathbb{E}\left[\|\delta_{t+1}\|^2\right] \leq \beta_t^2(1+\gamma_t L)(1+4\gamma_t L)\,\mathbb{E}\left[\|\delta_t\|^2\right] + 4\gamma_t L(1+\gamma_t L)\beta_t^2\,\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\|^2\right]$$
$$+ (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)} + 2\gamma_t L(1+\gamma_t L)\beta_t^2\,\mathbb{E}\left[\|\epsilon_t\|^2\right],$$

where $\overline{\sigma}^2_{\mathrm{DP}} \coloneqq 2\left(1-\frac{b}{m}\right)\frac{\sigma^2}{b} + d\cdot\sigma^2_{\mathrm{DP}}$.

Proof. Let $t \in \{0, \dots, T-1\}$. Suppose that assumptions 2.2 and 2.3 hold and that $\mathcal{L}_{\mathcal{H}}$ is L-smooth.

Recall from (52) that

$$\delta_{t+1} \coloneqq \overline{m}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right).$$

Denote $\overline{\widetilde{g}}_t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\tilde{g}^{(i)}_t$. Substituting from (47) and recalling that $\overline{m}_t = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} m^{(i)}_t$, we obtain

$$\delta_{t+1} = \beta_t\,\overline{m}_t + (1-\beta_t)\,\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right).$$

Upon adding and subtracting $\beta_t\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)$ and $\beta_t\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})$ on the R.H.S. above we obtain that

$$\begin{aligned}\delta_{t+1} &= \beta_t\,\overline{m}_t - \beta_t\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) + (1-\beta_t)\,\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right) + \beta_t\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) + \beta_t\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \beta_t\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\\ &= \beta_t\left(\overline{m}_t - \nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right) + (1-\beta_t)\,\overline{\widetilde{g}}_{t+1} - (1-\beta_t)\nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right) + \beta_t\left(\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right).\end{aligned}$$

As $\overline{m}_t - \nabla\mathcal{L}_{\mathcal{H}}(\theta_t) = \delta_t$ (by (52)), from above we obtain that

$$\delta_{t+1} = \beta_t\delta_t + (1-\beta_t)\,\left(\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right)\right) + \beta_t\left(\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right).$$

Therefore,

$$\begin{aligned}\|\delta_{t+1}\|^2 =&\beta_t^2\,\|\delta_t\|^2 + (1-\beta_t)^2\left\|\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right)\right\|^2\\ &+ \beta_t^2\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\|^2 + 2\beta_t(1-\beta_t)\left\langle\delta_t,\,\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right)\right\rangle\\ &+ 2\beta_t^2\left\langle\delta_t,\,\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle + 2\beta_t(1-\beta_t)\left\langle\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right),\,\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle.\end{aligned}$$

By taking conditional expectation $\mathbb{E}_{t+1}\left[\cdot\right]$ on both sides, and recalling that δ_t, θ_{t+1} and θ_t are deterministic values when the history $\mathcal{P}_{t+1}$ is given, we obtain that

$$\begin{aligned}\mathbb{E}_{t+1}\left[\|\delta_{t+1}\|^2\right] =&\beta_t^2\,\|\delta_t\|^2 + (1-\beta_t)^2\mathbb{E}_{t+1}\left[\left\|\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right)\right\|^2\right] + \beta_t^2\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\|^2 +\\ &2\beta_t(1-\beta_t)\left\langle\delta_t,\,\mathbb{E}_{t+1}\left[\overline{\widetilde{g}}_{t+1}\right] - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right)\right\rangle + 2\beta_t^2\left\langle\delta_t,\,\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle\\ &+ 2\beta_t(1-\beta_t)\left\langle\mathbb{E}_{t+1}\left[\overline{\widetilde{g}}_{t+1}\right] - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right),\,\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle.\end{aligned}$$

Recall that $\overline{\widetilde{g}}_{t+1} \coloneqq \frac{1}{(n-f)}\sum_{j\in\mathcal{H}}\tilde{g}^{(i)}_{t+1}$. Thus, as we ignore clipping by Assumption 2.3, we have $\mathbb{E}_{t+1}\left[\overline{\widetilde{g}}_{t+1}\right] = \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})$. Using this above we obtain that

$$\begin{aligned}\mathbb{E}_{t+1}\left[\|\delta_{t+1}\|^2\right] =&\beta_t^2\,\|\delta_t\|^2 + (1-\beta_t)^2\mathbb{E}_{t+1}\left[\left\|\overline{\widetilde{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right)\right\|^2\right] + \beta_t^2\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\|^2\\ &+ 2\beta_t^2\left\langle\delta_t,\,\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle.\end{aligned}$$

Now, denote $\overline{\sigma}^2_{\mathrm{DP}} \coloneqq 2\left(1-\frac{b}{m}\right)\frac{\sigma^2}{b} + d \cdot \sigma^2_{\mathrm{DP}}$. By assumptions 2.2 and 2.3, we can invoke Lemma D.5 which implies, together with the fact that $g^{(j)}_{t+1}$'s for $j \in \mathcal{H}$ are independent, that $\mathbb{E}_{t+1}\left[\left\|\overline{\overline{g}}_{t+1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta_{t+1}\right)\right\|^2\right] \leq \frac{\overline{\sigma}^2_{\mathrm{DP}}}{n-f}$. Thus,

$$\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] \leq \beta_t^2\left\|\delta_t\right\|^2 + (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)} + \beta_t^2\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\|^2 + 2\beta_t^2\left\langle\delta_t,\, \nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle.$$

By the Cauchy-Schwartz inequality, $\left\langle\delta_t,\, \nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\rangle \leq \left\|\delta_t\right\|\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\|$. Since $\mathcal{L}_{\mathcal{H}}$ is L-smooth, we have $\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\| \leq L\left\|\theta_{t+1} - \theta_t\right\|$. Recall from (50) that $\theta_{t+1} = \theta_t - \gamma_t R_t$. Thus,$\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t) - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t+1})\right\| \leq \gamma_t L\left\|R_t\right\|$. Using this above we obtain that

$$\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] \leq \beta_t^2\left\|\delta_t\right\|^2 + (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)} + \gamma_t^2\beta_t^2L^2\left\|R_t\right\|^2 + 2\gamma_t\beta_t^2L\left\|\delta_t\right\|\left\|R_t\right\|.$$

As $2ab \leq a^2+b^2$, from above we obtain that

$$\begin{aligned}\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] &\leq \beta_t^2\left\|\delta_t\right\|^2 + (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)} + \gamma_t^2\beta_t^2L^2\left\|R_t\right\|^2 + \gamma_tL\beta_t^2\left(\left\|\delta_t\right\|^2 + \left\|R_t\right\|^2\right)\\ &= (1+\gamma_tL)\beta_t^2\left\|\delta_t\right\|^2 + (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)} + \gamma_tL(1+\gamma_tL)\beta_t^2\left\|R_t\right\|^2. \qquad (99)\end{aligned}$$

By definition of ϵ_t in (53), we have $R_t = \epsilon_t + \overline{m}_t$. Thus, owing to the triangle inequality and the fact that $2ab \leq a^2+b^2$, we have $\left\|R_t\right\|^2 \leq 2\left\|\epsilon_t\right\|^2 + 2\left\|\overline{m}_t\right\|^2$. Similarly, by definition of δ_t in (52), we have $\left\|\overline{m}_t\right\|^2 \leq 2\left\|\delta_t\right\|^2 + 2\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2$. Thus, $\left\|R_t\right\|^2 \leq 2\left\|\epsilon_t\right\|^2 + 4\left\|\delta_t\right\|^2 + 4\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2$. Using this in (99) we obtain that

$$\begin{aligned}\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] \leq\ & (1+\gamma_tL)\beta_t^2\left\|\delta_t\right\|^2 + (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)}\\ &+ 2\gamma_tL(1+\gamma_tL)\beta_t^2\left(\left\|\epsilon_t\right\|^2 + 2\left\|\delta_t\right\|^2 + 2\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right).\end{aligned}$$

By rearranging the terms on the R.H.S., we get

$$\begin{aligned}\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] \leq\ & \beta_t^2(1+\gamma_tL)\left(1+4\gamma_tL\right)\left\|\delta_t\right\|^2 + 4\gamma_tL(1+\gamma_tL)\beta_t^2\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2 + (1-\beta_t)^2\frac{\overline{\sigma}^2_{\mathrm{DP}}}{(n-f)}\\ &+ 2\gamma_tL(1+\gamma_tL)\beta_t^2\left\|\epsilon_t\right\|^2.\end{aligned}$$

The proof concludes upon taking total expectation on both sides. □

D.5.4. Proof of Lemma D.3

Lemma D.3. *Assume that $\mathcal{L}_{\mathcal{H}}$ is L-smooth. Consider Algorithm 1. For any $t \in [T]$, we have*

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t)\right] \leq -\frac{\gamma_t}{2}\left(1-4\gamma_tL\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right] + \gamma_t\left(1+2\gamma_tL\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \gamma_t\left(1+\gamma_tL\right)\mathbb{E}\left[\left\|\epsilon_t\right\|^2\right].$$

Proof. Let $t \in \{0,\ldots,T-1\}$. Assuming $\mathcal{L}_{\mathcal{H}}$ is L-smooth, we have (see Lemma 1.2.3 (Nesterov et al., 2018))

$$\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t) \leq \left\langle\theta_{t+1} - \theta_t,\, \nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\rangle + \frac{L}{2}\left\|\theta_{t+1} - \theta_t\right\|^2.$$

Substituting from (54), i.e., $\theta_{t+1} = \theta_t - \gamma_t\,\overline{m}_t - \gamma_t\epsilon_t$, we obtain that

$$\begin{aligned}\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t) &\leq -\gamma_t\left\langle\overline{m}_t,\, \nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\rangle - \gamma_t\left\langle\epsilon_t,\, \nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\rangle + \gamma_t^2\frac{L}{2}\left\|\,\overline{m}_t + \epsilon_t\right\|^2\\ &= -\gamma_t\left\langle\overline{m}_t - \nabla\mathcal{L}_{\mathcal{H}}(\theta_t) + \nabla\mathcal{L}_{\mathcal{H}}(\theta_t),\, \nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\rangle - \gamma_t\left\langle\epsilon_t,\, \nabla\mathcal{L}_{\mathcal{H}}(\theta_t)\right\rangle + \gamma_t^2\frac{L}{2}\left\|\,\overline{m}_t + \epsilon_t\right\|^2.\end{aligned}$$

By Definition (52), $\overline{m}_t - \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) = \delta_t$. Thus, from above we obtain

$$\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t) \leq -\gamma_t \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2 - \gamma_t \left\langle \delta_t, \, \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) \right\rangle - \gamma_t \left\langle \epsilon_t, \, \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) \right\rangle + \frac{1}{2}\gamma_t^2 L \left\| \overline{m}_t + \epsilon_t \right\|^2. \tag{100}$$

Now, we consider the last three terms on the R.H.S. separately. Using Cauchy-Schwartz inequality, and the fact that $2ab \leq \frac{1}{c}a^2 + cb^2$ for any $c > 0$, we obtain that (by substituting $c = 2$)

$$2 \left|\left\langle \delta_t, \, \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) \right\rangle\right| \leq 2 \left\|\delta_t\right\| \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\| \leq \frac{2}{1} \left\|\delta_t\right\|^2 + \frac{1}{2} \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2. \tag{101}$$

Similarly,

$$2 \left|\left\langle \epsilon_t, \, \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) \right\rangle\right| \leq 2 \left\|\epsilon_t\right\| \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\| \leq \frac{2}{1} \left\|\epsilon_t\right\|^2 + \frac{1}{2} \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2. \tag{102}$$

Finally, using triangle inequality and the fact that $2ab \leq a^2 + b^2$ we have

$$\begin{aligned} \left\| \overline{m}_t + \epsilon_t \right\|^2 &\leq 2 \, \left\|\overline{m}_t\right\|^2 + 2 \left\|\epsilon_t\right\|^2 = 2 \, \left\|\overline{m}_t - \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) + \nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2 + 2 \left\|\epsilon_t\right\|^2 \\ &\leq 4 \, \left\|\delta_t\right\|^2 + 4 \, \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2 + 2 \left\|\epsilon_t\right\|^2. \qquad [\text{since } \; \overline{m}_t - \nabla \mathcal{L}_{\mathcal{H}}(\theta_t) = \delta_t] \end{aligned} \tag{103}$$

Substituting from (101), (102) and (103) in (100) we obtain that

$$\begin{aligned} \mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t) \leq &- \gamma_t \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2 + \frac{1}{2}\gamma_t \left(2 \left\|\delta_t\right\|^2 + \frac{1}{2} \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right) + \frac{1}{2}\gamma_t \left(2 \left\|\epsilon_t\right\|^2 + \frac{1}{2} \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2\right) \\ &+ \frac{1}{2}\gamma_t^2 L \left(4 \, \left\|\delta_t\right\|^2 + 4 \, \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2 + 2 \left\|\epsilon_t\right\|^2\right). \end{aligned}$$

Upon rearranging the terms in the R.H.S., we obtain that

$$\mathcal{L}_{\mathcal{H}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{H}}(\theta_t) \leq -\frac{\gamma_t}{2} \left(1 - 4\gamma_t L\right) \left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_t)\right\|^2 + \gamma_t \left(1 + 2\gamma_t L\right) \left\|\delta_t\right\|^2 + \gamma_t \left(1 + \gamma_t L\right) \left\|\epsilon_t\right\|^2.$$

This concludes the proof. □

E. Experimental Evaluation

In Section E.1, we present our experimental setup. In Section E.2, we report our empirical results.

E.1. Experimental Setup

In our experiments, we test the performance of SAFE-DSHB using SMEA and Filter (Diakonikolas et al., 2017; Data & Diggavi, 2021) in the server-based architecture and in three privacy regimes.

Dataset, model architecture, and hyperparameters. We train a logistic regression model of $d = 69$ parameters on the academic *Phishing*[5] dataset. We employ the *binary cross entropy* (bce) loss as well as L2-regularization of parameter $\lambda = 10^{-4}$, making the underlying learning problem strongly convex. We train the model using a fixed learning rate $\gamma = 1$ over a total of $T = 400$ learning steps. We set the clipping threshold $C = 1$ and the batch size $b = 25$. We run all algorithms, except DSGD, with momentum $\beta = 0.99$.

Distributed setup, and privacy accounting. We consider a server-based architecture composed of $n = 7$ workers, among which $f = 3$ are adversarial. The honest workers inject a privacy noise $\sigma_{\mathrm{DP}} = \frac{2C}{b} \times \sigma_{\mathrm{NM}}$ to their gradients, where σ_{NM} is referred to as the noise multiplier. We consider three privacy regimes in our experiments; namely *low* privacy where $\sigma_{\mathrm{NM}} = 1$, *moderate* privacy where $\sigma_{\mathrm{NM}} = 2$, and *high* privacy where $\sigma_{\mathrm{NM}} = 3$. In order to estimate the privacy budgets achieved at the end of the learning, we use Opacus (Yousefpour et al., 2021), a DP library for deep learning in PyTorch (Paszke et al., 2019). Using Opacus, the aggregate privacy budgets after $T = 400$ steps of learning are $(\epsilon, \delta) = (1.14, 10^{-4})$ in the *low* privacy regime, $(\epsilon, \delta) = (0.32, 10^{-4})$ in the *moderate* privacy regime, and $(\epsilon, \delta) = (0.19, 10^{-4})$ in the *high* privacy regime.

Evaluation details and reproducibility. As a benchmark, we compare the performance of SAFE-DSHB against the DP-DSGD algorithm, i.e., the private version of the adversary-free DSGD. We test SAFE-DSHB using SMEA and Filter. These algorithms are obtained by running Algorithm 1 while replacing the aggregation method F with the robust algorithm in question, namely SMEA and Filter. Note that we run Filter with spectral norm bound $\sigma_0^2 = 0$ (see Section B.2) because it provides the best empirical results, and it cannot be set to its theoretical value since the values of data heterogeneity G^2 and stochastic gradient noise σ^2 are unknown. We run each experiment with five seeds from 1 to 5 for reproducibility. The code we use to launch the different experiments will be made available.

Adversarial attacks. In our experiments, the adversarial workers execute four state-of-the-art attacks from the robust distributed ML literature, namely A Little is Enough (ALIE) (Baruch et al., 2019), Fall of Empires (FOE) (Xie et al., 2019), Sign-flipping (SF) (Allen-Zhu et al., 2020), and Label-flipping (LF) (Allen-Zhu et al., 2020).
The first three attacks rely on the same attack primitive that we explain below, while LF is executed differently.
Let b_t be the attack vector in step t and $\tau \geq 0$ a fixed real number. In every step t, the adversarial workers send to the server the gradient $B_t = \overline{g}_t + \tau_t b_t$, where $\overline{g}_t$ is an estimation of the true gradient at step t. Experimentally, we set $\overline{g}_t = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} g_t^{(i)}$.

- **ALIE:** In this attack, $b_t = \sigma_t$, where σ_t is coordinate-wise standard deviation of $\overline{g}_t$. In our experiments on ALIE, τ_t is chosen through an extensive grid search. Essentially, in each step t, we choose the value that results in the worst adversarial vector, i.e, the vector for which the distance to $\overline{g}_t$ is the largest.
- **FOE:** In this attack, $b_t = -\overline{g}_t$. All adversarial workers thus send $(1 - \tau_t)\overline{g}_t$ in step t. Similar to *ALIE*, τ_t for *FoE* is also estimated through grid searching.
- **SF:** In this attack, $b_t = -\overline{g}_t$, and $\tau_t = 2$. All adversarial workers thus send $B_t = b_t = -\overline{g}_t$ in step t.
- **LF:** Every adversarial worker computes its gradient on flipped labels. Since the labels l for Phishing are in $\{0, 1\}$, the adversarial workers flip the labels by computing $l' = 1 - l$ on the batch, where l' is the flipped/modified label.

E.2. Experimental Results

We present our results in the *low* privacy regime in Figures 2 and 3, in the *mid* privacy regime in Figures 4 and 5, and finally in the *high* privacy regime in Figures 6 and 7. We then comment on the results below.

[5] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

Low Privacy Regime ($\sigma_{\mathrm{NM}} = 1$)**.**

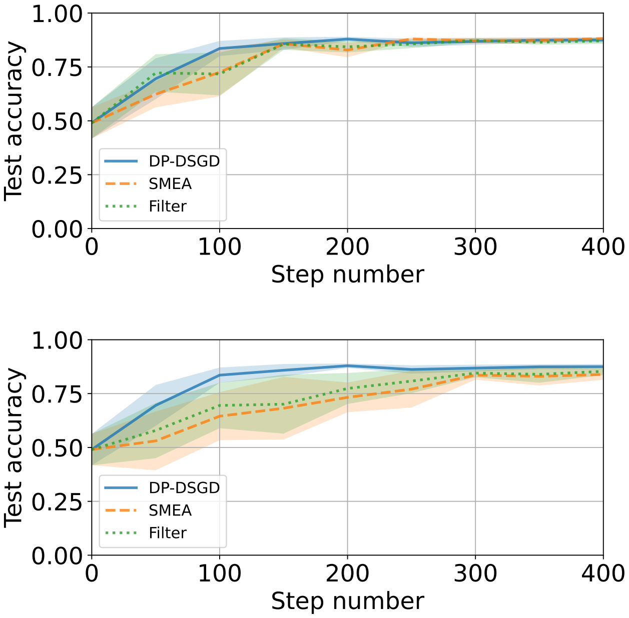

Figure 2. Test accuracy on Phishing with $f = 3$ adversarial workers among $n = 7$ workers, with $\beta = 0.99$. The adversarial workers execute the LF (*row 1, left*), SF (*row 1, right*), ALIE (*row 2, left*), and FOE (*row 2, right*) attacks. Privacy budget after $T = 400$ steps is $(\epsilon, \delta) = (1.14, 10^{-4})$.

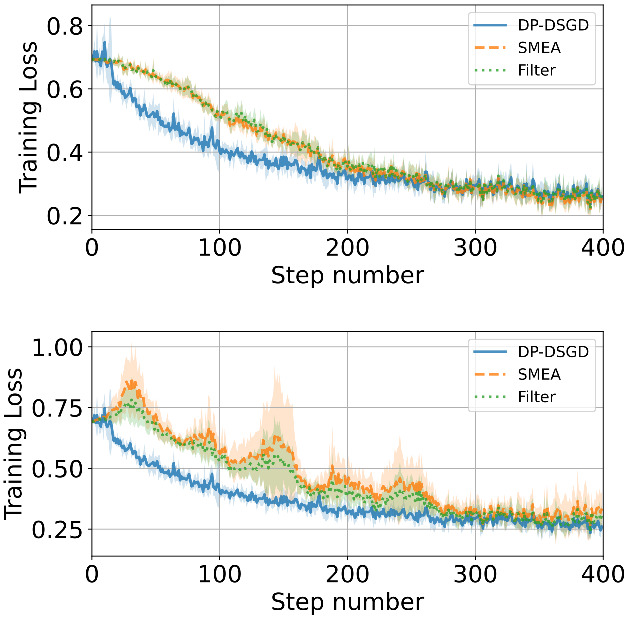

Figure 3. Training loss on Phishing with $f = 3$ adversarial workers among $n = 7$ workers, with $\beta = 0.99$. The adversarial workers execute the LF (*row 1, left*), SF (*row 1, right*), ALIE (*row 2, left*), and FOE (*row 2, right*) attacks. Privacy budget after $T = 400$ steps is $(\epsilon, \delta) = (1.14, 10^{-4})$.

Moderate Privacy Regime ($\sigma_{\mathrm{NM}} = 2$)**.**

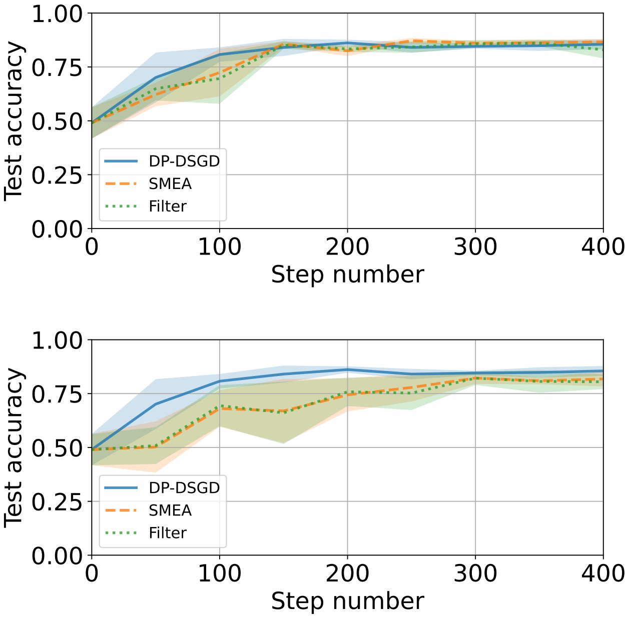

Figure 4. Test accuracy on Phishing with $f = 3$ adversarial workers among $n = 7$ workers, with $\beta = 0.99$. The adversarial workers execute the LF (*row 1, left*), SF (*row 1, right*), ALIE (*row 2, left*), and FOE (*row 2, right*) attacks. Privacy budget after $T = 400$ steps is $(\epsilon, \delta) = (0.32, 10^{-4})$.

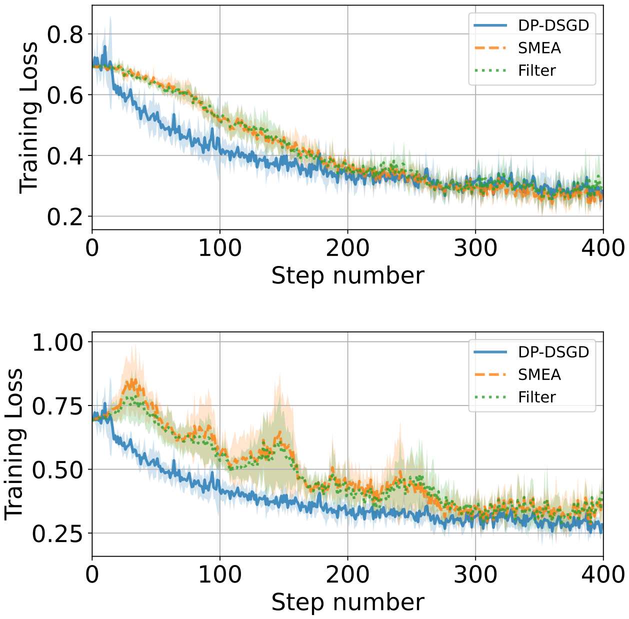

Figure 5. Training loss on Phishing with $f = 3$ adversarial workers among $n = 7$ workers, with $\beta = 0.99$. The adversarial workers execute the LF (*row 1, left*), SF (*row 1, right*), ALIE (*row 2, left*), and FOE (*row 2, right*) attacks. Privacy budget after $T = 400$ steps is $(\epsilon, \delta) = (0.32, 10^{-4})$.

High Privacy Regime ($\sigma_{\mathrm{NM}} = 3$).

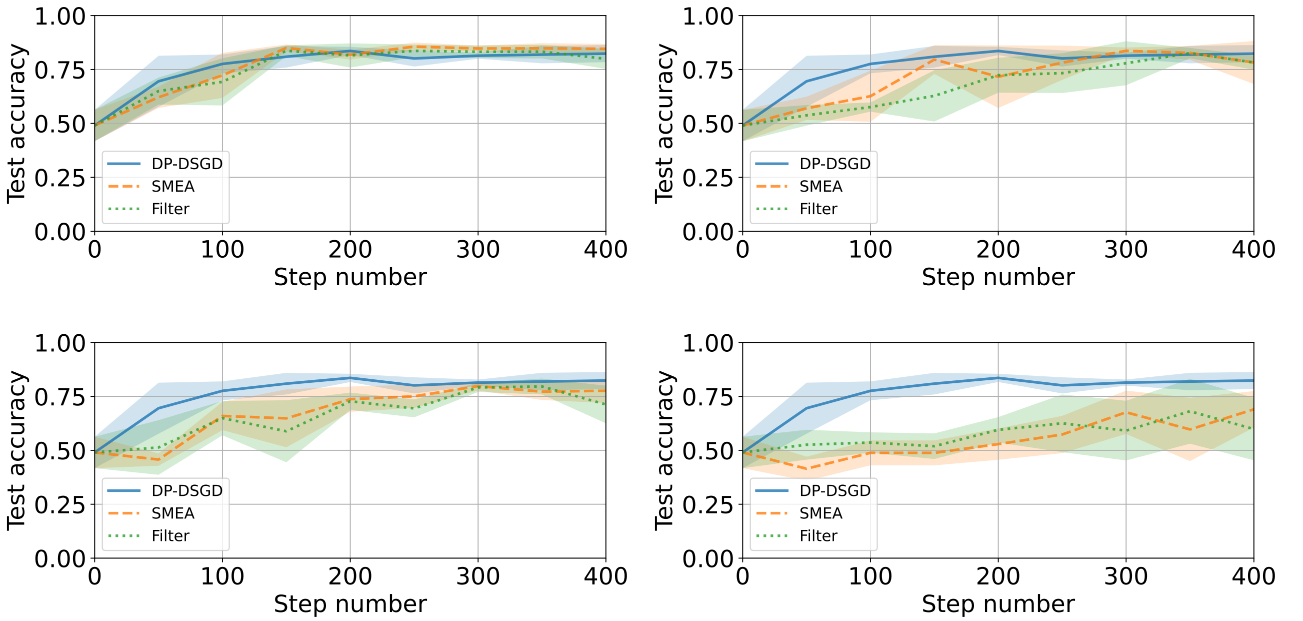

Figure 6. Test accuracy on Phishing with $f = 3$ adversarial workers among $n = 7$ workers, with $\beta = 0.99$. The adversarial workers execute the LF (*row 1, left*), SF (*row 1, right*), ALIE (*row 2, left*), and FOE (*row 2, right*) attacks. Privacy budget after $T = 400$ steps is $(\epsilon, \delta) = (0.19, 10^{-4})$.

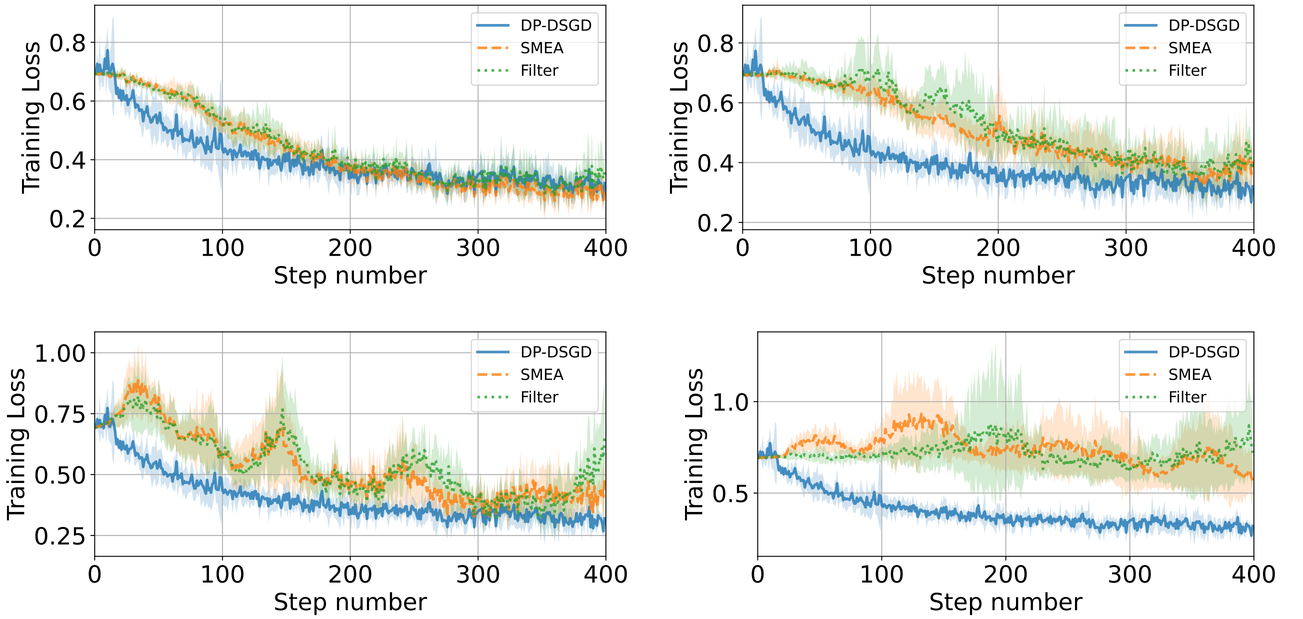

Figure 7. Training loss on Phishing with $f = 3$ adversarial workers among $n = 7$ workers, with $\beta = 0.99$. The adversarial workers execute the LF (*row 1, left*), SF (*row 1, right*), ALIE (*row 2, left*), and FOE (*row 2, right*) attacks. Privacy budget after $T = 400$ steps is $(\epsilon, \delta) = (0.19, 10^{-4})$.

Discussion. We consider four different attacks executed by the adversarial nodes, and report on the performance of the algorithms in three different privacy regimes. Our observations are twofold.

First, as expected, we see that as the privacy regime becomes more demanding, the performances of DP-DSGD and SMEA degrade both in terms of test accuracy and training loss. This confirms that the standard privacy-utility trade-off also occurs in the presence of adversarial workers. Second, we see that under all three privacy regimes, SAFE-DSHB with SMEA is able to successfully mitigate adversarial attacks while still ensuring strong levels of differential privacy. Indeed, the final accuracies reached by SAFE-DSHB with SMEA are around 80% in the *low* and *moderate* privacy regimes, and around 75% in *high* privacy (a bit lower under the FOE attack). On the other hand, the training losses are decreasing under all attacks and in all privacy regimes, sometimes asymptotically matching the curves of DP-DSGD (e.g., the LF attack in all three privacy regimes, the ALIE attack in *low* and *moderate* privacy). The same observations hold for SAFE-DSHB with Filter.