

# Byzantine-Resilient High-Dimensional Federated Learning

Deepesh Data and Suhas Diggavi  
University of California, Los Angeles

Email: {deepesh.data@gmail.com, suhas@ee.ucla.edu}

## Abstract

We study stochastic gradient descent (SGD) with local iterations in the presence of malicious/Byzantine clients, motivated by the federated learning. The clients, instead of communicating with the central server in every iteration, maintain their local models, which they update by taking several SGD iterations based on their own datasets and then communicate the net update with the server, thereby achieving communication-efficiency. Furthermore, only a subset of clients communicate with the server, and this subset may be different at different synchronization times. The Byzantine clients may collaborate and send arbitrary vectors to the server to disrupt the learning process. To combat the adversary, we employ an efficient high-dimensional robust mean estimation algorithm from Steinhardt et al. [SCV18, ITCS 2018] at the server to filter-out corrupt vectors; and to analyze the outlier-filtering procedure, we develop a novel matrix concentration result that may be of independent interest.

We provide convergence analyses for strongly-convex and non-convex smooth objectives in the heterogeneous data setting, where different clients may have different local datasets, and we do not make any probabilistic assumptions on data generation. We believe that ours is the first Byzantine-resilient algorithm and analysis with local iterations. We derive our convergence results under minimal assumptions of bounded variance for SGD and bounded gradient dissimilarity (which captures heterogeneity among local datasets). We also extend our results to the case when clients compute full-batch gradients.

## 1 Introduction

In the *federated learning* (FL) paradigm [Kon17, KMRR16, MMR<sup>+</sup>17, MSS19], several clients (e.g., mobiles devices, organizations, etc.) collaboratively learn a machine learning model, where the training process is facilitated by the data held by the participating clients (without data centralization) and is coordinated by a central server (e.g., the service provider). Due to its many advantages over the traditional centralized learning [DCM<sup>+</sup>12] (e.g., training a machine learning model without collecting the clients' data, which, in addition to reducing the communication load on the network, provides a basic level of privacy to clients' data), FL has emerged as an active area of research recently; see [K<sup>+</sup>19] for a detailed survey. Stochastic gradient descent (SGD) has become a de facto standard in optimization for training machine learning models at such a large scale [Bot10, MMR<sup>+</sup>17, K<sup>+</sup>19], where clients iteratively communicate the gradient updates with the central server, which aggregates the gradients, updates the learning model, and sends the aggregated gradient back to the clients. The promise of FL comes with its own set of challenges [K<sup>+</sup>19]: (i) optimizing with *heterogeneous* data at different clients, who may have different local datasets, which may be “non-i.i.d.”, i.e., can be thought of as being generated from different underlying distributions; (ii) slow and unreliable network connections between the server and the clients, so communication in every iteration may not be feasible; (iii) availability of only a subset of clients for training at a given time (maybe due to low connectivity, as clients may be located in different geographic locations); and (iv) robustness against the malicious/Byzantine clients who may send incorrect gradient updates to the central server to disrupt the training process. In this paper, we propose and analyze a *single* SGD algorithm that addresses all these challenges *together*. First we setup the problem, put our work in context with the related work, and then summarize our contributions.

We consider an empirical risk minimization problem, where data is stored at  $R$  clients, each having a different dataset (with no probabilistic assumption on data generation); client  $r \in [R]$  has dataset  $\mathcal{D}_r$ .

Let  $F_r : \mathbb{R}^d \rightarrow \mathbb{R}$  denote the local loss function associated with the dataset  $\mathcal{D}_r$ , which is defined as  $F_r(\mathbf{x}) \triangleq \mathbb{E}_{i \in U[n_r]}[F_{r,i}(\mathbf{x})]$ , where  $n_r = |\mathcal{D}_r|$ ,  $i$  is uniformly distributed over  $[n_r] \triangleq \{1, 2, \dots, n_r\}$ , and  $F_{r,i}(\mathbf{x})$  is the loss associated with the  $i$ 'th data point at client  $r$  with respect to (w.r.t.)  $\mathbf{x}$ . Our goal is to solve the following minimization problem:

$$\arg \min_{\mathbf{x} \in \mathbb{R}^d} \left( F(\mathbf{x}) \triangleq \frac{1}{R} \sum_{r=1}^R \mathbb{E}_{i \in U[n_r]}[F_{r,i}(\mathbf{x})] \right). \quad (1)$$

Let  $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x})$  denote a minimizer of the global loss function  $F(\mathbf{x})$ . In absence of the above-mentioned FL challenges, we can minimize (1) using distributed *vanilla* SGD, where in any iteration, server broadcasts the current model parameters to all the clients, each of them then computes a stochastic gradient from its local dataset and sends it back to the server, who aggregates the received gradients and updates the global model parameters. However, this simple solution does not satisfy the FL challenges, as *every* client communicates with the server (i.e., no sampling of clients) in *every* SGD iteration (i.e., no local iterations), and furthermore, this solution breaks down even with a single malicious client [BMGS17].

## 1.1 Related Work

Recent work has proposed variants of the above-described vanilla SGD that address *some* of the FL challenges. The algorithms in [HKMC19, HM19, KKM<sup>+</sup>19, KMR19, LHY<sup>+</sup>20, SLS<sup>+</sup>20, YYZ19, BDKD19] work under different heterogeneity assumptions but do not provide any robustness to malicious clients. On the other hand, [CSX17, BMGS17, YCRB18, AAL18, SX19, XKG19b, YCRB19] provide robustness, but with no local iterations or sampling of clients; furthermore, they assume homogeneous (either same or i.i.d.) data across all clients. A different line of work [CWCP18, RWCP19, DSD19b, DD19, DSD19a, LXC<sup>+</sup>19, GHYR19, DD20, HKJ20] provides robustness with heterogeneous data, but without local iterations or sampling of clients, which we briefly explain in the following. [CWCP18, RWCP19, DSD19b, DD19, DSD19a] use coding across datasets, which is hard to implement in FL. [LXC<sup>+</sup>19] changes the objective function and adds a regularizer term to combat the adversary. [GHYR19] effectively reduces the heterogeneous problem to a homogeneous problem by clustering, and then learning happens within each cluster having homogeneous data. [HKJ20] proposed a resampling technique that effectively adapts existing robust algorithms (which might have been designed to work with homogeneous – identical or i.i.d. – datasets) to work with heterogeneous datasets. Note that [HKJ20] provides convergence guarantees of their resampling techniques applied to only KRUM, which is the robust aggregation rule from [BMGS17].

[DD20] is the closest related work to ours, in the sense that they also proposed an SGD algorithm on heterogeneous data that uses robust mean estimation subroutines to filter out corrupt gradients and analyzed it under the same minimal assumptions as ours. We want to emphasize that [DD20] does not incorporate local iterations and sampling of clients in their algorithm and analyses, which makes our analyses fundamentally different from theirs. We had to develop new tools (a matrix concentration inequality) to analyze our algorithm, and also the convergence analyses in our paper are very different from those in literature, including that in [DD20]. Our analyses differ from those of local SGD without adversaries, as (apart from differing in other technical details) our method requires establishing two separate recurrences, one at synchronization indices and the other one for the rest of the indices. Our analyses also differ from those of SGD without local iterations and without adversaries, as local SGD causes drift in the local parameter vectors at clients in between any two synchronization indices – this drift occurs even when all clients have identical data. Note that bounding this drift is necessary for convergence but is non-trivial with heterogeneous data and without having strong assumptions. Our matrix concentration result and its analysis is also very different from that of [DD20], as we need to prove it in the presence of local iterations.

We believe that ours is the first work that combines *local iterations* with *Byzantine-resilience* for SGD.<sup>1</sup> Not only that, we also analyze our algorithm on *heterogeneous* data and allow *sampling of clients*. Note that,

<sup>1</sup>At the completion of our work, we found that [XKG19a] also analyzed SGD in the FL setting, but with the following major differences: Not only do they make bounded gradient assumption, the approximation error (even in the Byzantine-free setting) of their solution could be as large as  $\mathcal{O}(D^2 + G^2)$ , where  $G$  is the gradient bound and  $D$  is the diameter of the parameter space

apart from the notable exception of [DD20], the earlier work that provides robustness (without local iterations or sampling of clients) either assume homogeneous data across clients [CSX17, BMGS17, YCRB18, AAL18, SX19, YCRB19] or require strong assumptions, such as the bounded gradient assumption on local functions (i.e.,  $\|\nabla F_r(\mathbf{x})\| \leq G$  for some finite  $G$ ) [XKG19b]. Note that even without robustness, assuming bounded gradients is a common way to make the analysis on heterogeneous data simple [YYZ19, LHY<sup>+</sup>20], as under this assumption, we can trivially bound the heterogeneity among local datasets by  $\|\nabla F_r(\mathbf{x}) - \nabla F_s(\mathbf{x})\| \leq 2G$ ,<sup>2</sup> which makes handling heterogeneity vacuous.

## 1.2 Our Contributions

In this paper, we tackle heterogeneity assuming only that the gradient dissimilarity among local datasets is bounded (see (6)), and propose and analyze a Byzantine-resilient SGD algorithm with local iterations and sampling of clients under the bounded variance assumption for SGD (see (2)); see [Algorithm 1](#). We provide convergence analyses for strongly-convex and non-convex smooth objectives. Our convergence results are summarized below, where  $b$  is the mini-batch size for stochastic gradients,  $\sigma^2$  is the variance bound,  $\kappa^2$  captures the gradient dissimilarity,  $H$  is the number of local iterations in between any two consecutive synchronization indices,  $K$  is the number of clients sampled at synchronization times,  $\epsilon < \frac{K}{4R}$  is the fraction of Byzantine clients, and  $\epsilon'$  is any constant such that  $(\epsilon + \epsilon') \leq \frac{K}{4R}$ .

For strongly-convex objectives, our algorithm can find approximate optimal parameters within an error of  $\Gamma = \mathcal{O}\left(\frac{H\sigma^2}{be'}\left(1 + \frac{d}{K}\right)(\epsilon + \epsilon') + H\kappa^2\right)$  exponentially (in  $\frac{T}{H}$ ) fast, and for non-convex objectives, it can reach to a stationary point within the same error  $\Gamma$  with a speed of  $\frac{1}{T/H}$ . Note that the convergence rate of *vanilla* SGD (i.e., without local iterations and in Byzantine-free settings) decays exponentially (in  $T$ ) fast for strongly-convex objectives and with a speed of  $\frac{1}{T}$  for non-convex objectives, whereas, our convergence rates are affected by the number of local iterations  $H$ . This is a result of working with weak assumptions – if we work with the bounded gradient assumption, then we can also get exponential (in  $T$ ) convergence in the strongly-convex case and  $\frac{1}{T}$  convergence in the non-convex case.

In the approximation error  $\Gamma$ , the first error term  $\frac{H\sigma^2}{be'}\left(1 + \frac{d}{K}\right)(\epsilon + \epsilon')$  mainly arises because of the stochasticity in gradients due to SGD and is equal to zero if we work with full-batch gradients (which gives  $\sigma = 0$ ), and the second error term  $H\kappa^2$  arises because of heterogeneity in local datasets. Note that  $\Gamma$  only has a linear dependence on  $H$ .

We also give a simplified analysis of our algorithm with full-batch gradients for all three objectives. See [Theorem 1](#) and [Theorem 2](#) for our mini-batch SGD and full-batch GD convergence results, respectively. See a detailed discussion on the approximation error analysis and the convergence rates in [Section 2.4](#).

To tackle the malicious behavior of Byzantine clients, we borrow tools from recent advances in high-dimensional robust statistics [LRV16, SCV18, DKK<sup>+</sup>19, DK19]; in particular, we use the polynomial-time outlier-filtering procedure from [SCV18], which was developed for robust mean estimation in high dimensions. In order to use this algorithm, we develop a novel matrix concentration result (see [Theorem 3](#)) which may be of independent interest. For full-batch gradients, we give our matrix concentration result with better guarantees, which can be proved by a much simplified analysis than its mini-batch counterpart; see [Theorem 4](#).

## 1.3 Paper Organization

We describe our algorithm and state the main convergence results in [Section 2](#). We describe the core part of our algorithm, the robust accumulated gradient estimation (RAGE), and our new matrix concentration result

---

that contains the optimal parameters  $\mathbf{x}^*$  and all the local parameters  $\mathbf{x}_r^t$  ever emerged at any client  $r \in [R]$  in any iteration  $t \in [T]$ ; this, in our opinion, makes the bound vacuous. In optimization, one would ideally like to have the convergence rates depend on diameter of the parameter space with a factor that decays with the number of iterations, e.g., with  $\frac{1}{T}$  or  $\frac{1}{\sqrt{T}}$ , and also see [Theorem 1](#).

<sup>2</sup>See [KMR19] for a detailed discussion on the inappropriateness of making bounded gradient assumption in heterogeneous data settings and examine the effect of heterogeneity on convergence rates (even without robustness).

in [Section 3](#) and also prove it there. We prove our main convergence results for mini-batch SGD in [Section 4](#) and [Section 5](#) and for full-batch SGD in [Section 6](#).

## 1.4 Notation

For any  $n \in \mathbb{N}$ , we denote the set  $\{1, 2, \dots, n\}$  by  $[n]$ , and for any  $n_1, n_2 \in \mathbb{N}$  such that  $n_1 \leq n_2$ , we denote the set  $\{n_1, n_1 + 1, \dots, n_2\}$  by  $[n_1 : n_2]$ . We denote vectors by bold small letters  $\mathbf{x}, \mathbf{y}$ , etc., and matrices by bold capital letters  $\mathbf{A}, \mathbf{B}$ , etc. For any finite set  $\mathcal{K}$ , we write  $k \in_U \mathcal{K}$  to denote that  $k$  is chosen uniformly at random from  $\mathcal{K}$ . All vector norms in this paper are  $\ell_2$  norms, and for convenience, we simply denote them by  $\|\cdot\|$ . For a square matrix  $\mathbf{A}$ , we write  $\lambda_{\max}(\mathbf{A})$  to denote the largest eigenvalue of  $\mathbf{A}$ .

## 2 Problem Setup and Our Results

In this section, we state our assumptions, describe the adversary model and our algorithm, and state our main convergence results.

### 2.1 Assumptions

As mentioned in [Section 1](#), we make minimal assumptions to analyze our algorithm. Our first assumption is a standard one in SGD, which assumes bounded variance for stochastic gradients. Our second assumption is for heterogeneous data and assumes that the heterogeneity in different local datasets is bounded.

**Assumption 1** (Bounded local variances). *The stochastic gradients sampled from any local dataset have uniformly bounded variance over  $\mathbb{R}^d$ , i.e., there exists a finite  $\sigma \geq 0$ , such that*

$$\mathbb{E}_{i \in_U [n_r]} \|\nabla F_{r,i}(\mathbf{x}) - \nabla F_r(\mathbf{x})\|^2 \leq \sigma^2, \quad \forall \mathbf{x} \in \mathbb{R}^d, r \in [R]. \quad (2)$$

It will be helpful to formally define mini-batch stochastic gradients, where instead of computing stochastic gradients based on just one data point, each client selects a subset of size  $b$  uniformly at random from its own local dataset and computes the average of  $b$  gradients. For any  $\mathbf{x} \in \mathbb{R}^d, r \in [R], b \in [n_r]$ , consider the following set

$$\mathcal{F}_r^{\otimes b}(\mathbf{x}) := \left\{ \frac{1}{b} \sum_{i \in \mathcal{H}_b} \nabla F_{r,i}(\mathbf{x}) : \mathcal{H}_b \in \binom{[n_r]}{b} \right\}. \quad (3)$$

Note that  $\mathbf{g}_r(\mathbf{x}) \in_U \mathcal{F}_r^{\otimes b}(\mathbf{x})$  is a mini-batch stochastic gradient with batch size  $b$  at client  $r$ . It is not hard to see the following:

$$\mathbb{E}[\mathbf{g}_r(\mathbf{x})] = \nabla F_r(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d, r \in [R] \quad (4)$$

$$\mathbb{E} \|\mathbf{g}_r(\mathbf{x}) - \nabla F_r(\mathbf{x})\|^2 \leq \frac{\sigma^2}{b}, \quad \forall \mathbf{x} \in \mathbb{R}^d, r \in [R] \quad (5)$$

where (4) says that  $\mathbf{g}_r(\mathbf{x})$  is an unbiased gradient and (5) says that the variance of mini-batch stochastic gradients reduces by the same factor as the batch size. Though the bound in (5) goes down with  $b$ , it does not become zero when we compute full-batch gradients, which uses all  $n_r$  data points. This is because (5) only uses that the clients sample  $b$  data points *with* replacement. However, in reality, since this sampling is done *without* replacement, we can show a finer variance bound of  $\mathbb{E} \|\mathbf{g}_r(\mathbf{x}) - \nabla F_r(\mathbf{x})\|^2 \leq \frac{(n_r - b)}{b(n_r - 1)} \sigma^2$ ; see [\[Sa\]](#) for a proof. We can slightly improve our results by using this finer variance bound instead of (5) everywhere in this paper, but, for simplicity, we only use the weaker bound (5) throughout.

**Assumption 2** (Bounded gradient dissimilarity). *The difference between the local gradients  $\nabla F_r(\mathbf{x}), r \in [R]$  and the global gradient  $\nabla F(\mathbf{x}) = \frac{1}{R} \sum_{r=1}^R \nabla F_r(\mathbf{x})$  is uniformly bounded over  $\mathbb{R}^d$  for all clients, i.e., there exists a finite  $\kappa$ , such that*

$$\|\nabla F_r(\mathbf{x}) - \nabla F(\mathbf{x})\|^2 \leq \kappa^2, \quad \forall \mathbf{x} \in \mathbb{R}^d, r \in [R]. \quad (6)$$

In [Assumption 2](#),  $\kappa$  quantifies the bounded deviation between the local loss functions  $F_r, r \in [R]$  and the global loss function  $F$ ; see also [\[YJY19, LYWZ19\]](#), where this assumption has been used in heterogeneous data settings in decentralized SGD without Byzantine clients. The gradient dissimilarity bound in [\(6\)](#) can be seen as a *deterministic* condition on local datasets, under which we derive our results.

### 2.1.1 Need of [Assumption 2](#)

For any method that filters out malicious updates from the clients and work with heterogeneous (“non i.i.d.”) data, as the server does not know the identities of the adversarial clients, we need to have some regularity condition relating the datasets, and we believe [Assumption 2](#) is a natural way to model that. [Assumption 2](#) intuitively captures the heterogeneity among local datasets, without making any statistical assumptions on the data. To see the necessity of bounding heterogeneity even without adversary, note that we allow clients to perform local SGD steps, where, in between any two synchronization indices, clients compute gradients from their local datasets and update their local parameter vectors; as a result, their local parameter vectors can drift away from each other. This drift needs to be bounded for convergence analyses, and if we do not assume bounded heterogeneity, it is impossible to bound this drift. As we have discussed at the end of [Section 1.1](#), [Assumption 2](#) is much weaker than the bounded gradient assumption, which not only makes bounding the drift (and the convergence analyses) trivial, but also obscure the dependence of the convergence bounds on the heterogeneity of datasets, which is clearly brought out in our convergence results.

### 2.1.2 Bounds on $\sigma^2$ and $\kappa^2$ in the statistical heterogeneous model

Since all results (matrix concentration and convergence) in this paper are given in terms of  $\sigma^2$  and  $\kappa^2$ , to show the clear dependence of our results on the dimensionality of the problem, we can bound these quantities in the statistical *heterogeneous* data model under different distributional assumptions on local gradients. For the variance bound [\(2\)](#), it was shown in [\[DD20, Theorem 7\]](#) that if local gradients have sub-Gaussian distribution, then  $\sigma = \mathcal{O}\left(\sqrt{d \log(d)}\right)$ . For the gradient dissimilarity bound [\(6\)](#), it was shown in [\[DD20, Theorem 6\]](#) that if either the local gradients have sub-exponential distribution and each client has at least  $n = \Omega(d \log(nd))$  data points or local gradients have sub-Gaussian distribution and  $n \in \mathbb{N}$  is arbitrary, then  $\kappa \leq \kappa_{\text{mean}} + \mathcal{O}\left(\sqrt{\frac{d \log(nd)}{n}}\right)$ , where  $\kappa_{\text{mean}}$  denotes the distance of the expected local gradients from the global gradient.

## 2.2 Adversary Model

We assume that an  $\epsilon$  fraction of  $R$  clients are malicious; as we see later, we can tolerate  $\epsilon < \frac{K}{4R}$ ,<sup>3</sup> where  $K \leq R$  is the number of clients sampled at synchronization indices. The malicious clients can *collaborate* and *arbitrarily* deviate from their pre-specified programs: In any SGD iteration, instead of sending true stochastic gradients, corrupt clients may send adversarially chosen vectors (they may not even send anything if they wish, in which case, the server can treat them as *erasures* and replace them with a fixed value). Note that, in the erasure case, server knows which clients are corrupt; whereas, in the Byzantine problem, server does not have this information.

## 2.3 Main Results

Let  $\mathcal{I}_T = \{t_1, t_2, \dots, t_k, \dots\}$ , with  $t_1 = 0$ , denote the set of synchronization indices at which clients communicate their net updates with the server. Let  $H$  denote the difference between any two consecutive indices, i.e., every worker performs the same number  $H$  of local iterations between any two consecutive synchronization indices. At synchronization indices, server samples a subset of  $K$  clients (denoted by  $\mathcal{K} \subseteq [R]$ ) and sends the global model (denoted by  $\mathbf{x}$ ) to them; each client  $r \in \mathcal{K}$  updates its local model  $\mathbf{x}_r$  by taking SGD steps

<sup>3</sup>Actually, we can tolerate  $\epsilon < \frac{1}{4}$  fraction of malicious clients from the  $K$  clients that we select; so,  $\epsilon < \frac{K}{4R}$  is a worst case bound in case we sample *all* the malicious clients in a selection, which is an unlikely event.

---

**Algorithm 1** Byzantine-Resilient SGD with Local Iterations

---

- 1: **Initialize.** Set  $t := 0$ ,  $\mathbf{x}_r^0 := \mathbf{0}, \forall r \in [R]$ , and  $\mathbf{x} := \mathbf{0}$ . Here,  $\mathbf{x}$  denotes the global model and  $\mathbf{x}_r^0$  denotes the local model at client  $r$  at time 0. Fix a constant step-size  $\eta$  and a mini-batch size  $b$ .
  - 2: **while** ( $t \leq T$ ) **do**
  - 3:   Server selects an arbitrary subset of clients  $\mathcal{K} \subseteq [R]$  of size  $|\mathcal{K}| = K$  and sends  $\mathbf{x}$  to all clients in  $\mathcal{K}$ .
  - 4:   **All clients**  $r \in \mathcal{K}$  **do in parallel:**
  - 5:     Set  $\mathbf{x}_r^t = \mathbf{x}$ .
  - 6:     **while** (true) **do**
  - 7:       Take a mini-batch stochastic gradient  $\mathbf{g}_r(\mathbf{x}_r^t) \in_U \mathcal{F}_r^{\otimes b}(\mathbf{x}_r^t)$  and update the local model:
$$\mathbf{x}_r^{t+1} \leftarrow \mathbf{x}_r^t - \eta \mathbf{g}_r(\mathbf{x}_r^t); \quad t \leftarrow (t + 1).$$
  - 8:     **if** ( $t \in \mathcal{I}_T$ ) **then**
  - 9:       Let  $\tilde{\mathbf{x}}_r^t = \mathbf{x}_r^t$ , if client  $r$  is honest, otherwise  $\tilde{\mathbf{x}}_r^t$  can be an arbitrary vector in  $\mathbb{R}^d$ .
  - 10:       Send  $\tilde{\mathbf{x}}_r^t$  to the server and break the inner **while** loop.
  - 11:     **end if**
  - 12:   **end while**
  - 13:   **At Server:**
  - 14:   Receive  $\{\tilde{\mathbf{x}}_r, r \in \mathcal{K}\}$  from the clients in  $\mathcal{K}$ .
  - 15:   For every  $r \in \mathcal{K}$ , let  $\tilde{\mathbf{g}}_{r,\text{accu}} := (\tilde{\mathbf{x}}_r - \mathbf{x})/\eta$ .
  - 16:   Apply the decoding algorithm RAGE (see [Algorithm 2 in Section 3.2](#)) on  $\{\tilde{\mathbf{g}}_{r,\text{accu}}, r \in \mathcal{K}\}$ . Let
$$\hat{\mathbf{g}}_{\text{accu}} := \text{RAGE}(\tilde{\mathbf{g}}_{r,\text{accu}}, r \in \mathcal{K}).$$
  - 17:   Update the global model  $\mathbf{x} \leftarrow \mathbf{x} - \eta \hat{\mathbf{g}}_{\text{accu}}$ .
  - 18: **end while**
- 

based on its local dataset until the next synchronization time, when all clients in  $\mathcal{K}$  send their local models to the server. Note that some of these clients may be corrupt and may send arbitrary vectors. Server employs a decoding algorithm RAGE<sup>4</sup> and update the global model  $\mathbf{x}$  based on that.

**Remark 1.** *Note that the only disruption that the corrupt clients can cause in the training process is during the gradient aggregation at synchronization indices by sending adversarially chosen vectors to the server, and we give unlimited power to the adversary for that. Because of this and for the purpose of analysis, we can assume, without loss of generality, that in between the synchronization indices, the corrupt clients sample stochastic gradients and update their local parameters honestly.*

We present our Byzantine-resilient SGD algorithm with local iterations in [Algorithm 1](#).

Before we present our results, we need some definitions.

- **$L$ -smoothness:** A function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  is called  $L$ -smooth over  $\mathbb{R}^d$ , if for every  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , we have  $\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$  (this property is also known as  $L$ -Lipschitz gradients). This is also equivalent to  $F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$ .
- **$\mu$ -strong convexity:** A function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  is called  $\mu$ -strongly convex over  $\mathbb{R}^d$  for  $\mu \geq 0$ , if for every  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , we have  $F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$ .

---

<sup>4</sup>Our decoding algorithm, which we call RAGE, is the same as the robust mean estimation algorithm proposed by Steinhardt et al. [SCV18]. We gave it a different name, as we use it in a much more general FL setting of running SGD with local iterations on heterogeneous data. Note that the same algorithm has also been used in [SX19, YCRB19] in the context of Byzantine-robust *full batch* gradient descent *without* local iterations, assuming *homogeneous* i.i.d. data, whereas, we employ that algorithm in the FL setting, which makes its analysis significantly more challenging.



All convergence results in this paper only require properties of the global loss function  $F$ ; the local loss functions  $F_r, r \in [R]$  may be arbitrary. For example, in the smooth strongly-convex case, we only require  $F$  to be smooth and strongly-convex, and  $F_r, r \in [R]$  may be arbitrary. Similarly for the non-convex case.

Our convergence results are for strongly-convex and non-convex smooth objectives.

**Theorem 1** (Mini-Batch Local Stochastic Gradient Descent). *Suppose an  $\epsilon > 0$  fraction of clients are adversarially corrupt. Let  $\mathcal{K}_t$  denote the set of  $K$  clients that are active at any given time  $t \in [0 : T]$ . For a global objective function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$ , let **Algorithm 1** generate a sequence of iterates  $\{\mathbf{x}_r^t : t \in [0 : T], r \in \mathcal{K}_t\}$  when run with a fixed step-size  $\eta = \frac{1}{8HL}$ . Fix an arbitrary constant  $\epsilon' > 0$ . If  $\epsilon \leq \frac{K}{4R} - \epsilon'$ , then with probability at least  $1 - \frac{T}{H} \exp(-\frac{\epsilon'^2(1-\epsilon)K}{16})$ , the sequence of average iterates  $\{\mathbf{x}^t = \frac{1}{K} \sum_{r \in \mathcal{K}_t} \mathbf{x}_r^t : t \in [0 : T]\}$  satisfy the following convergence guarantees:*

- **Strongly-convex:** If  $F$  is  $L$ -smooth for  $L \geq 0$  and  $\mu$ -strongly convex for  $\mu > 0$ , we get:

$$\mathbb{E} \|\mathbf{x}^T - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\mu}{16HL}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{13}{\mu^2} \Gamma. \quad (7)$$

- **Non-convex:** If  $F$  is  $L$ -smooth for  $L \geq 0$ , we get:

$$\frac{1}{T} \sum_{t=0}^T \mathbb{E} \|\nabla F(\mathbf{x}^t)\|^2 \leq \frac{16HL}{T} [\mathbb{E}[F(\mathbf{x}^0)] - \mathbb{E}[F(\mathbf{x}^*)]] + \frac{9}{2} \Gamma. \quad (8)$$

In (7), (8),  $\Gamma = \left(\frac{3\Upsilon^2}{H} + \frac{11H\sigma^2}{b} + 36H\kappa^2\right)$  with  $\Upsilon^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$ , where  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'} \left(1 + \frac{4d}{3K}\right) + 28H^2\kappa^2$ , and expectation is taken over the sampling of mini-batch stochastic gradients.

We prove (7) and (8) in **Section 4** and **Section 5**, respectively. In addition to other complications arising due to handling Byzantine clients together with local iterations, our proof deviates from the standard proofs for local SGD without adversary, as we need to show two recurrences, one at synchronization indices and the other at non-synchronization indices. This is because at synchronization indices, server performs decoding to filter-out the corrupt clients, while at other indices there is no decoding, as there is no communication.

The failure probability of our algorithm is at most  $\frac{T}{H} \exp(-\frac{\epsilon'^2(1-\epsilon)K}{16})$ , which though scales linearly with  $T$ , also goes down exponentially with  $K$ . As a result, in settings such as federated learning, where number of clients could be very large (e.g., in millions) and server samples a few thousand clients, we can get a very small probability of error, even if we run our algorithm for a very long time. Note that the error probability is due to the *stochastic* sampling of gradients, and if we want a “zero” probability of error, we can run full-batch gradient descent, for which we get the following result, which we prove in **Section 6** with a much simplified analysis than that of **Theorem 1**.

**Theorem 2** (Full-Batch Local Gradient Descent). *In the same setting as that of **Theorem 1**, except for that we run **Algorithm 1** with a fixed step-size  $\eta = \frac{1}{5HL}$ , and in any iteration, instead of sampling mini-batch stochastic gradients, every honest client takes full-batch gradients from their local datasets. If  $\epsilon \leq \frac{K}{4R}$ , then with probability 1, the sequence of average iterates  $\{\mathbf{x}^t = \frac{1}{K} \sum_{r \in \mathcal{K}_t} \mathbf{x}_r^t : t \in [0 : T]\}$  satisfy the following convergence guarantees:*

- **Strongly-convex:** If  $F$  is  $L$ -smooth for  $L \geq 0$  and  $\mu$ -strongly convex for  $\mu > 0$ , we get:

$$\|\mathbf{x}^T - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\mu}{10HL}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{14}{\mu^2} \Gamma_{GD}. \quad (9)$$

- **Non-convex:** If  $F$  is  $L$ -smooth for  $L \geq 0$ , we get:

$$\frac{1}{T} \sum_{t=0}^T \|\nabla F(\mathbf{x}^t)\|^2 \leq \frac{10HL}{T} [F(\mathbf{x}^0) - F(\mathbf{x}^*)] + \frac{24}{5} \Gamma_{GD}. \quad (10)$$

In (9), (10),  $\Gamma_{GD} = \frac{2\Upsilon_{GD}^2}{H} + 25H\kappa^2$ , where  $\Upsilon_{GD} = \mathcal{O}(H\kappa\sqrt{\epsilon})$ .

## 2.4 Important Remarks About Theorem 1 and Theorem 2

**Analysis of the approximation error.** In Theorem 1, the approximation error  $\Gamma$  essentially consists of two types of error terms:  $\Gamma_1 = \mathcal{O}\left(\frac{H\sigma^2}{b\epsilon'}\left(1 + \frac{4d}{3K}\right)(\epsilon + \epsilon')\right)$  and  $\Gamma_2 = \mathcal{O}(H\kappa^2)$ , where  $\Gamma_1$  arises due to stochastic sampling of gradients and  $\Gamma_2$  arises due to dissimilarity in the local datasets. Observe that  $\Gamma_1$  decreases as we increase the batch size  $b$  of stochastic gradients and becomes zero if we take full-batch gradients (which implies  $\sigma = 0$ ), as is the case in Theorem 2. Note that both  $\Gamma_1$  and  $\Gamma_2$  have a linear dependence on the number of local iterations  $H$ . Observe that since we are working with heterogeneous datasets, the presence of gradient dissimilarity bound  $\kappa^2$  (which captures the heterogeneity) in the approximation error is inevitable, and will always show up when bounding the deviation of the true “global” gradient from the decoded one in the presence of Byzantine clients, even when  $H = 1$ .

**Convergence rates.** In the strongly-convex case, Algorithm 1 approximately finds the optimal parameters  $\mathbf{x}^*$  (within  $\Gamma$  error) with  $\left(1 - \frac{\mu}{cHL}\right)^T$  speed, where  $c = 16$  for SGD and  $c = 10$  for GD. Note that  $\left(1 - \frac{\mu}{cHL}\right)^T \leq \exp^{-\frac{\mu}{cL} \frac{T}{H}}$ , where the inequality follows from  $\left(1 - \frac{1}{x}\right)^x \leq \frac{1}{e}$ . This implies that the convergence rate in this case is exponentially fast (but in  $\frac{T}{H}$ ). In the non-convex case, Algorithm 1 reaches to a stationary point (within  $\Gamma$  error) with a speed of  $\frac{1}{T/H}$ . Note that the convergence rate of *vanilla* SGD (i.e., without local iterations and in Byzantine-free settings) is exponentially fast (in  $T$ ) for strongly-convex objectives and with a speed of  $\frac{1}{T}$  for non-convex objectives, whereas, our convergence rates are affected by the number of local iterations  $H$ . The reason for this is precisely because, under standard SGD assumptions we need  $\eta \leq \frac{1}{8HL}$  to bound the drift in local parameters across different clients; see Lemma 2. Instead, if we had assumed a stronger bounded gradient assumption (which trivially bound the heterogeneity, as explained at the end of Section 1.1), then Lemma 2 would hold for a constant step-size that does not depend on  $H$  (e.g.,  $\eta = \frac{1}{2L}$  would suffice), which would lead to an exponentially fast (in  $T$ ) convergence for strongly-convex objectives and  $\frac{1}{T}$  convergence rate for non-convex objectives.

## 3 Robust Accumulated Gradient Estimation (RAGE)

In this section, we provide our main result on robust accumulated gradient estimation (RAGE), which is the subroutine for robustly estimating the average of uncorrupted *accumulated* gradients at every synchronization index; see Footnote 4. First we setup the notation. Let Algorithm 1 generate a sequence of iterates  $\{\mathbf{x}_r^t : t \in [0 : T], r \in \mathcal{K}_t\}$  when run with a fixed step-size  $\eta$  satisfying  $\eta \leq \frac{1}{8HL}$ , where  $\mathcal{K}_t$  denotes the set of  $K$  clients that are active at time  $t \in [0 : T]$ . Take any two consecutive synchronization indices  $t_k, t_{k+1} \in \mathcal{I}_T$ . Note that  $|t_{k+1} - t_k| \leq H$ . For an honest client  $r \in \mathcal{K}_{t_k}$ , let  $\mathbf{g}_{r,\text{accu}}^{t_k, t_{k+1}} := \sum_{t=t_k}^{t_{k+1}-1} \mathbf{g}_r(\mathbf{x}_r^t)$  denote the sum of local mini-batch stochastic gradients sampled by client  $r$  between time  $t_k$  and  $t_{k+1}$ , where  $\mathbf{g}_r(\mathbf{x}_r^t) \in_U \mathcal{F}_r^{\otimes b}(\mathbf{x}_r^t)$  satisfies (4), (5). At iteration  $t_{k+1}$ , every honest client  $r \in \mathcal{K}_{t_k}$  reports its local model  $\mathbf{x}_r^{t_{k+1}}$  to the server, from which server computes  $\mathbf{g}_{r,\text{accu}}^{t_k, t_{k+1}}$  (see line 15 of Algorithm 1), whereas, the corrupt clients may report arbitrary and adversarially chosen vectors in  $\mathbb{R}^d$ . Server does not know the identity of the corrupt clients, and its goal is to produce an estimate  $\hat{\mathbf{g}}_{\text{accu}}^{t_k, t_{k+1}}$  of the average accumulated gradients from honest clients as best as possible.

To this end, first we show that there exists a large subset  $\mathcal{S} \subseteq \mathcal{K}_{t_k}$  of accumulated gradients from honest clients that are concentrated around their average, i.e., have bounded empirical covariance. Once we have shown that, then we will use the polynomial-time outlier-filtering algorithm from [SCV18] to estimate the average of the accumulated gradients in  $\mathcal{S}$ . Our main result on RAGE is as follows:

**Theorem 3** (Robust Accumulated Gradient Estimation). *Suppose an  $\epsilon$  fraction of  $K$  clients that communicate with the server are corrupt. In the setting described above, suppose we are given  $K \leq R$  accumulated gradients  $\tilde{\mathbf{g}}_{r,\text{accu}}^{t_k, t_{k+1}}, r \in \mathcal{K}_{t_k}$  in  $\mathbb{R}^d$ , where  $\tilde{\mathbf{g}}_{r,\text{accu}}^{t_k, t_{k+1}} = \mathbf{g}_{r,\text{accu}}^{t_k, t_{k+1}}$  if the  $r$ 'th client is honest, otherwise can be arbitrary. For any constant  $\epsilon' > 0$ , if  $(\epsilon + \epsilon') \leq \frac{1}{4}$ , then we have:*

1. **Matrix concentration:** *With probability  $1 - \exp(-\frac{\epsilon'^2(1-\epsilon)K}{16})$ , there exists a subset  $\mathcal{S} \subseteq \mathcal{K}_{t_k}$  of*



uncorrupted *gradients* of size  $(1 - (\epsilon + \epsilon'))K \geq \frac{3K}{4}$ , such that

$$\lambda_{\max} \left( \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\mathbf{g}_i - \mathbf{g}_S)(\mathbf{g}_i - \mathbf{g}_S)^T \right) \leq \frac{25H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right) + 28H^2\kappa^2, \quad (11)$$

where, for  $i \in \mathcal{S}$ ,  $\mathbf{g}_i = \mathbf{g}_{i,\text{accu}}^{t_k, t_{k+1}}$ ,  $\mathbf{g}_S = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{g}_{i,\text{accu}}^{t_k, t_{k+1}}$ ; and  $\lambda_{\max}$  denotes the largest eigenvalue.

**2. Outlier-filtering algorithm:** We can find an estimate  $\hat{\mathbf{g}}$  of  $\mathbf{g}_S$  in polynomial-time with probability 1, such that  $\|\hat{\mathbf{g}} - \mathbf{g}_S\| \leq \mathcal{O}(\sigma_0 \sqrt{\epsilon + \epsilon'})$ , where  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right) + 28H^2\kappa^2$ .

Proving the matrix concentration bound stated in the first part of [Theorem 3](#) is non-trivial and we prove it separately in [Section 3.1](#). For the second part, we use the polynomial-time outlier-filtering procedure of [\[SCV18\]](#), which is a robust mean estimation algorithm, that takes a collection of vectors as input, out of which an unknown large subset (at least a  $\frac{3}{4}$ -fraction) is promised to be well-concentrated around its sample mean (i.e., has a bounded covariance), and outputs an estimate of the sample mean of the vectors in that subset. For completeness, we describe this procedure in [Section 3.2](#) and refer the reader to [\[DD20, Appendices E, F\]](#) for more details.

Note that the same filtering procedure has also been used in [\[SX19, YCRB19\]](#) in the context of Byzantine-robust *full batch* gradient descent *without* local iterations for minimizing the population risk, assuming *homogeneous* i.i.d. data. Our setting is very different from theirs, as we minimize the empirical risk by mini-batch *stochastic* gradient descent *with* local iterations on *heterogeneous* data. They also derived a matrix-concentration result, whose need arises because they minimize the population risk, whereas, we need a matrix concentration bound because we use SGD. On top of that our setting is much more complicated than theirs, as clients have heterogeneous data and do not communicate with the server in every iteration. As a result, as opposed to their matrix concentration bound (which they proved assuming sub-exponential/sub-Gaussian distribution on local gradients and also assuming i.i.d. data across clients), our matrix concentration result is of a very different nature, and we use entirely different tools to derive that.

### 3.1 Matrix Concentration

Now we prove the first part of [Theorem 3](#). For that, we need to show an existence of a subset  $\mathcal{S}$  of the  $K$  accumulated gradients (out of which an  $\epsilon < \frac{1}{4}$  fraction is corrupted) that has good concentration, as quantified by the matrix concentration bound in (11). To prove this, we use a separate matrix concentration result stated in the following lemma from [\[DD20\]](#).

**Lemma 1** (Lemma 1 in [\[DD20\]](#)). *Suppose there are  $m$  independent distributions  $p_1, p_2, \dots, p_m$  in  $\mathbb{R}^d$  such that  $\mathbb{E}_{\mathbf{y} \sim p_i}[\mathbf{y}] = \boldsymbol{\mu}_i, i \in [m]$  and each  $p_i$  has a bounded variance in all directions, i.e.,  $\mathbb{E}_{\mathbf{y} \sim p_i}[(\langle \mathbf{y} - \boldsymbol{\mu}_i, \mathbf{v} \rangle)^2] \leq \sigma_{p_i}^2, \forall \mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\| = 1$ . Take any  $\epsilon' > 0$ . Then, given  $m$  independent samples  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ , where  $\mathbf{y}_i \sim p_i$ , with probability  $1 - \exp(-\epsilon'^2 m/16)$ , there is a subset  $\mathcal{S}$  of  $(1 - \epsilon')m$  points such that*

$$\lambda_{\max} \left( \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T \right) \leq \frac{4\sigma_{p_{\max}}^2}{\epsilon'} \left( 1 + \frac{d}{(1 - \epsilon')m} \right), \quad \text{where } \sigma_{p_{\max}}^2 = \max_{i \in [m]} \sigma_{p_i}^2.$$

Now we prove the first part of [Theorem 3](#) with the help of [Lemma 1](#).

Let  $t_k, t_{k+1} \in \mathcal{I}_T$  be any two consecutive synchronization indices. For  $i \in \mathcal{K}_{t_k}$  corresponding to an honest client, let  $Y_i^{t_k}, Y_i^{t_k+1}, \dots, Y_i^{t_{k+1}-1}$  be a sequence of  $(t_{k+1} - t_k) \leq H$  (dependent) random variables, where, for any  $t \in [t_k : t_{k+1} - 1]$ , the random variable  $Y_i^t$  is distributed as

$$Y_i^t \sim \text{Unif} \left( \mathcal{F}_i^{\otimes b}(\mathbf{x}_i^t(\mathbf{x}_i^{t_k}, Y_i^{t_k}, \dots, Y_i^{t-1})) \right). \quad (12)$$

Here,  $Y_i^t$  is a random variable that corresponds to the stochastic sampling of mini-batch gradients from the set  $\mathcal{F}_i^{\otimes b}(\mathbf{x}_i^t(\mathbf{x}_i^{t_k}, Y_i^{t_k}, \dots, Y_i^{t-1}))$ , which itself depends on the local parameters  $\mathbf{x}_i^{t_k}$  (which is a deterministic

quantity) at the last synchronization index and the past realizations of  $Y_i^{t_k}, \dots, Y_i^{t-1}$ . This is because the evolution of local parameters  $\mathbf{x}_i^t$  depends on  $\mathbf{x}_i^{t_k}$  and the choice of gradients in between time indices  $t_k$  and  $t-1$ . Now define  $Y_i := \sum_{t=t_k}^{t_{k+1}-1} Y_i^t$ ; and let  $p_i$  be the distribution of  $Y_i$ . This is the distribution  $p_i$  we will take when using [Lemma 1](#).

**Claim 1.** *For any honest client  $i \in \mathcal{K}_{t_k}$ , we have  $\mathbb{E}\|Y_i - \mathbb{E}[Y_i]\|^2 \leq \frac{H^2\sigma^2}{b}$ , where expectation is taken over sampling stochastic gradients by client  $i$  between synchronization indices  $t_k$  and  $t_{k+1}$ .*

[Claim 1](#) is proved in [Appendix A](#).

It is easy to see that the hypothesis of [Lemma 1](#) is satisfied with  $\boldsymbol{\mu}_i = \mathbb{E}[Y_i]$ ,  $\sigma_{p_i}^2 = \frac{H^2\sigma^2}{b}$  for all honest clients  $i \in \mathcal{K}_{t_k}$  (note that  $p_i$  is the distribution of  $Y_i$ ):

$$\mathbb{E}_{\mathbf{y}_i \sim p_i} [\langle \mathbf{y}_i - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle^2] \stackrel{(d)}{\leq} \mathbb{E}[\|\mathbf{y}_i - \mathbb{E}_{\mathbf{y}_i \sim p_i}[\mathbf{y}_i]\|^2] \cdot \|\mathbf{v}\|^2 \stackrel{(e)}{\leq} \frac{H^2\sigma^2}{b},$$

where (d) follows from the Cauchy-Schwarz inequality and (e) follows from [Claim 1](#) and  $\|\mathbf{v}\| \leq 1$ .

We are given  $K$  different (summations of  $H$ ) gradients, out of which at least  $(1-\epsilon)K$  are according to the correct distribution. By considering only the uncorrupted gradients (i.e., taking  $m = (1-\epsilon)K$ ), we have from [Lemma 1](#) that there exists a subset  $\mathcal{S} \subseteq \mathcal{K}_{t_k}$  of  $K$  gradients of size  $(1-\epsilon')(1-\epsilon)K \geq (1-(\epsilon+\epsilon'))K \geq \frac{3K}{4}$  (where in the last inequality we used  $(\epsilon+\epsilon') \leq \frac{1}{4}$ ) that satisfies

$$\lambda_{\max} \left( \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\mathbf{y}_i - \mathbb{E}[\mathbf{y}_i]) (\mathbf{y}_i - \mathbb{E}[\mathbf{y}_i])^T \right) \leq \frac{4H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right). \quad (13)$$

Note that (13) bounds the deviation of the points in  $\mathcal{S}$  from their respective means  $\mathbb{E}[\mathbf{y}_i]$ . However, in (11), we need to bound the deviation of the points in  $\mathcal{S}$  from their sample mean  $\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{y}_i$ . As it turns out, due to our use of local iterations, bounding this requires a substantial amount of technical work, which we do in the rest of this subsection.

From the alternate definition of the largest eigenvalue of symmetric matrices  $\mathbf{A} \in \mathbb{R}^{d \times d}$ , we have

$$\lambda_{\max}(\mathbf{A}) = \sup_{\mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{A} \mathbf{v}. \quad (14)$$

Applying this with  $\mathbf{A} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\mathbf{y}_i - \mathbb{E}[\mathbf{y}_i]) (\mathbf{y}_i - \mathbb{E}[\mathbf{y}_i])^T$ , we can equivalently write (13) as

$$\sup_{\mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|=1} \left( \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \langle \mathbf{y}_i - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle^2 \right) \leq \hat{\sigma}_0^2 := \frac{4H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right). \quad (15)$$

Define  $\mathbf{y}_S := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{y}_i$  to be the sample mean of the points in  $\mathcal{S}$ . Take an arbitrary  $\mathbf{v} \in \mathbb{R}^d$  such that  $\|\mathbf{v}\| = 1$ .

$$\begin{aligned} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \langle \mathbf{y}_i - \mathbf{y}_S, \mathbf{v} \rangle^2 &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} [\langle \mathbf{y}_i - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle + \langle \mathbb{E}[\mathbf{y}_i] - \mathbf{y}_S, \mathbf{v} \rangle]^2 \\ &\leq \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \langle \mathbf{y}_i - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \langle \mathbb{E}[\mathbf{y}_i] - \mathbf{y}_S, \mathbf{v} \rangle^2 \quad (\text{using } (a+b)^2 \leq 2a^2 + 2b^2) \end{aligned}$$

Using (15) to bound the first term, we get

$$\begin{aligned} &\leq 2\hat{\sigma}_0^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \mathbb{E}[\mathbf{y}_i] - \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \mathbf{y}_j, \mathbf{v} \right\rangle^2 = 2\hat{\sigma}_0^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left[ \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \langle \mathbf{y}_j - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle \right]^2 \\ &\leq 2\hat{\sigma}_0^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \langle \mathbf{y}_j - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle^2 \quad (\text{using the Jensen's inequality}) \end{aligned}$$

$$\begin{aligned}
&= 2\hat{\sigma}_0^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} [\langle \mathbf{y}_j - \mathbb{E}[\mathbf{y}_j], \mathbf{v} \rangle + \langle \mathbb{E}[\mathbf{y}_j] - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle]^2 \\
&\leq 2\hat{\sigma}_0^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{2}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \langle \mathbf{y}_j - \mathbb{E}[\mathbf{y}_j], \mathbf{v} \rangle^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{2}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \langle \mathbb{E}[\mathbf{y}_j] - \mathbb{E}[\mathbf{y}_i], \mathbf{v} \rangle^2 \\
&\quad \text{(using } (a+b)^2 \leq 2a^2 + 2b^2) \\
&\leq 2\hat{\sigma}_0^2 + \frac{4}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \langle \mathbf{y}_j - \mathbb{E}[\mathbf{y}_j], \mathbf{v} \rangle^2 + \frac{4}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \|\mathbb{E}[\mathbf{y}_j] - \mathbb{E}[\mathbf{y}_i]\|^2 \\
&\quad \text{(using the Cauchy-Schwarz inequality and that } \|\mathbf{v}\| \leq 1) \\
&\leq 6\hat{\sigma}_0^2 + \frac{4}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \|\mathbb{E}[\mathbf{y}_j] - \mathbb{E}[\mathbf{y}_i]\|^2 \tag{16}
\end{aligned}$$

**Claim 2.** For any  $r, s \in \mathcal{K}_{t_k}$ , we have

$$\|\mathbb{E}[\mathbf{y}_r] - \mathbb{E}[\mathbf{y}_s]\|^2 \leq H \sum_{t=t_k}^{t_{k+1}-1} (6\kappa^2 + 3L^2 \mathbb{E}\|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2), \tag{17}$$

where expectations in  $\mathbb{E}[\mathbf{y}_r]$  and  $\mathbb{E}[\mathbf{y}_s]$  are taken over sampling stochastic gradients between the synchronization indices  $t_k, \dots, t_{k+1}$  by client  $r$  and client  $s$ , respectively.

*Proof.* Note that we can equivalently write  $\mathbb{E}[\mathbf{y}_r] = \mathbb{E}[Y_r]$  and  $\mathbb{E}[\mathbf{y}_s] = \mathbb{E}[Y_s]$ .

$$\begin{aligned}
\|\mathbb{E}[Y_r] - \mathbb{E}[Y_s]\|^2 &= \|\mathbb{E}[Y_r] - \mathbb{E}[Y_s]\|^2 = \left\| \sum_{t=t_k}^{t_{k+1}-1} (\mathbb{E}[Y_r^t] - \mathbb{E}[Y_s^t]) \right\|^2 \\
&\leq (t_{k+1} - t_k) \sum_{t=t_k}^{t_{k+1}-1} \|\mathbb{E}[Y_r^t] - \mathbb{E}[Y_s^t]\|^2 \tag{18}
\end{aligned}$$

By definition of  $Y_s^t$  from (12), we have  $Y_s^t \sim \text{Unif}\left(\mathcal{F}_s^{\otimes b}(\mathbf{x}_s^t(\mathbf{x}_s^{t_k}, Y_s^{t_k}, \dots, Y_s^{t-1}))\right)$ , which implies using (4) that  $\mathbb{E}[Y_s^t] = \mathbb{E}[\nabla F_s(\mathbf{x}_s^t(\mathbf{x}_s^{t_k}, Y_s^{t_k}, \dots, Y_s^{t-1}))]$ , where on the RHS, expectation is taken over  $(Y_s^{t_k}, \dots, Y_s^{t-1})$ . To make the notation less cluttered, in the following, for any  $s \in \mathcal{K}_{t_k}$ , we write  $\mathbf{x}_s^t$  to denote  $\mathbf{x}_s^t(\mathbf{x}_s^{t_k}, Y_s^{t_k}, \dots, Y_s^{t-1})$  with the understanding that expectation is always taken over the sampling of stochastic gradients between  $t_k$  and  $t_{k+1}$ . With these substitutions, the  $t$ 'th term from (19) can be written as:

$$\begin{aligned}
\|\mathbb{E}[Y_r^t] - \mathbb{E}[Y_s^t]\|^2 &= \|\mathbb{E}[\nabla F_r(\mathbf{x}_r^t) - \nabla F_s(\mathbf{x}_s^t)]\|^2 \\
&\stackrel{(a)}{\leq} \mathbb{E}\|\nabla F_r(\mathbf{x}_r^t) - \nabla F_s(\mathbf{x}_s^t)\|^2 \tag{19} \\
&\stackrel{(b)}{\leq} 3\mathbb{E}\|\nabla F_r(\mathbf{x}_r^t) - \nabla F(\mathbf{x}_r^t)\|^2 + 3\mathbb{E}\|\nabla F_s(\mathbf{x}_s^t) - \nabla F(\mathbf{x}_s^t)\|^2 \\
&\quad + 3\mathbb{E}\|\nabla F(\mathbf{x}_r^t) - \nabla F(\mathbf{x}_s^t)\|^2 \\
&\stackrel{(c)}{\leq} 6\kappa^2 + 3L^2 \mathbb{E}\|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2. \tag{20}
\end{aligned}$$

Here, (a) and (b) both follow from the Jensen's inequality. (c) used the gradient dissimilarity bound from (6) to bound the first two terms<sup>5</sup> and  $L$ -Lipschitzness of  $\nabla F$  to bound the last term.

Substituting the bound from (20) back in (18) and using  $(t_{k+1} - t_k) \leq H$  proves Claim 2.  $\square$

<sup>5</sup>Note that though  $\mathbf{x}_r^t$ 's are random quantities, we can still bound  $\mathbb{E}\|\nabla F_r(\mathbf{x}_r^t) - \nabla F_s(\mathbf{x}_s^t)\|^2 \leq \kappa^2$  because the gradient dissimilarity bound (6) holds uniformly over the entire domain.

Using the bound from (17) in (16) gives

$$\begin{aligned} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \langle \mathbf{y}_i - \mathbf{y}_S, \mathbf{v} \rangle^2 &\leq 6\hat{\sigma}_0^2 + \frac{4}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} H \sum_{t=t_k}^{t_{k+1}-1} (6\kappa^2 + 3L^2 \mathbb{E} \|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2) \\ &= 6\hat{\sigma}_0^2 + 24H^2\kappa^2 + \frac{12HL^2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{E} \|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2 \end{aligned} \quad (21)$$

Now we bound the last term of (21), which is the drift in local parameters at different clients in between any two synchronization indices.

**Lemma 2.** *For any  $r, s \in \mathcal{K}_{t_k}$ , if  $\eta \leq \frac{1}{8HL}$ , we have*

$$\sum_{t=t_k}^{t_{k+1}-1} \mathbb{E} \|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2 \leq 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right), \quad (22)$$

where expectation is taken over sampling stochastic gradients at clients  $r, s$  between the synchronization indices  $t_k$  and  $t_{k+1}$ .

*Proof.* For any  $t \in [t_k : t_{k+1} - 1]$  and  $r, s \in \mathcal{K}_{t_k}$ , define  $D_{r,s}^t = \mathbb{E} \|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2$ . Note that at synchronization time  $t_k$ , all clients in the active set  $\mathcal{K}_{t_k}$  have the same parameters, i.e.,  $\mathbf{x}_r^{t_k} = \mathbf{x}^{t_k}$  for every  $r \in \mathcal{K}_{t_k}$ .

$$\begin{aligned} D_{r,s}^t &= \mathbb{E} \|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2 = \mathbb{E} \left\| \left( \mathbf{x}_r^{t_k} - \eta \sum_{j=t_k}^{t-1} \mathbf{g}_r(\mathbf{x}_r^j) \right) - \left( \mathbf{x}_s^{t_k} - \eta \sum_{j=t_k}^{t-1} \mathbf{g}_s(\mathbf{x}_s^j) \right) \right\|^2 \\ &= \eta^2 \mathbb{E} \left\| \sum_{j=t_k}^{t-1} (\mathbf{g}_r(\mathbf{x}_r^j) - \mathbf{g}_s(\mathbf{x}_s^j)) \right\|^2 \quad (\text{Since } \mathbf{x}_r^{t_k} = \mathbf{x}^{t_k}, \forall r \in \mathcal{K}_{t_k}) \\ &\leq \eta^2 (t - t_k) \sum_{j=t_k}^{t-1} \mathbb{E} \|\mathbf{g}_r(\mathbf{x}_r^j) - \mathbf{g}_s(\mathbf{x}_s^j)\|^2 \\ &\leq \eta^2 H \sum_{j=t_k}^{t-1} \left( 3\mathbb{E} \|\mathbf{g}_r(\mathbf{x}_r^j) - \nabla F_r(\mathbf{x}_r^j)\|^2 + 3\mathbb{E} \|\mathbf{g}_s(\mathbf{x}_s^j) - \nabla F_s(\mathbf{x}_s^j)\|^2 \right. \\ &\quad \left. + 3\mathbb{E} \|\nabla F_r(\mathbf{x}_r^j) - \nabla F_s(\mathbf{x}_s^j)\|^2 \right) \end{aligned} \quad (23)$$

To bound the first and the second terms we use the variance bound from (5).<sup>6</sup> We can bound the third term in the same way as we bounded it in (19) and obtained (20). This gives

$$\begin{aligned} D_{r,s}^t &\leq \eta^2 H \sum_{j=t_k}^{t-1} \left( \frac{6\sigma^2}{b} + 18\kappa^2 + 9L^2 \mathbb{E} \|\mathbf{x}_r^j - \mathbf{x}_s^j\|^2 \right) \\ &\leq \frac{6H^2\sigma^2\eta^2}{b} + 18H^2\eta^2\kappa^2 + 9L^2H\eta^2 \sum_{j=t_k}^{t-1} D_{r,s}^j \quad (\text{Since } D_{r,s}^j = \mathbb{E} \|\mathbf{x}_r^j - \mathbf{x}_s^j\|^2) \end{aligned}$$

Taking summation from  $t = t_k$  to  $t_{k+1} - 1$  gives

$$\sum_{t=t_k}^{t_{k+1}-1} D_{r,s}^t \leq \sum_{t=t_k}^{t_{k+1}-1} \left( \frac{6H^2\sigma^2\eta^2}{b} + 18H^2\eta^2\kappa^2 + 9L^2H\eta^2 \sum_{j=t_k}^{t-1} D_{r,s}^j \right)$$

---

<sup>6</sup>Note that  $\mathbf{x}_r^j$ 's are random quantities, however, since the variance bound (5) holds uniformly over the entire domain,  $\mathbb{E} \|\mathbf{g}_r(\mathbf{x}_r^j) - \nabla F_r(\mathbf{x}_r^j)\|^2 \leq \frac{\sigma^2}{b}$  holds for a random  $\mathbf{x}_r^j \in \mathbb{R}^d$ .

$$\leq \frac{6H^3\sigma^2\eta^2}{b} + 18H^3\eta^2\kappa^2 + 9L^2H^2\eta^2 \sum_{t=t_k}^{t_{k+1}-1} D_{r,s}^t.$$

After rearranging terms, we get

$$(1 - 9L^2H^2\eta^2) \sum_{t=t_k}^{t_{k+1}-1} D_{r,s}^t \leq \frac{6H^3\sigma^2\eta^2}{b} + 18H^3\eta^2\kappa^2. \quad (24)$$

If we take  $\eta \leq \frac{1}{8HL}$ , we get  $(1 - 9\eta^2L^2H^2) \geq \frac{6}{7}$ . Substituting this in the LHS of (24) yields  $\sum_{t=t_k}^{t_{k+1}-1} D_{r,s}^t \leq \frac{7H^3\sigma^2\eta^2}{b} + 21H^3\eta^2\kappa^2$ , which proves [Lemma 2](#).  $\square$

Substituting the bound from (22) for the last term in (21) gives

$$\begin{aligned} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \langle \mathbf{y}_i - \mathbf{y}_S, \mathbf{v} \rangle^2 &\leq 6\hat{\sigma}_0^2 + 24H^2\kappa^2 + \frac{12HL^2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \left( 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \right) \\ &= 6\hat{\sigma}_0^2 + 24H^2\kappa^2 + 84H^4L^2\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \\ &\leq 6\hat{\sigma}_0^2 + 28H^2\kappa^2 + \frac{21H^2\sigma^2}{16b} \quad (\text{Using } \eta \leq \frac{1}{8LH}) \\ &\leq \frac{24H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right) + \frac{21H^2\sigma^2}{16b} + 28H^2\kappa^2 \quad (\text{Since } \hat{\sigma}_0^2 = \frac{4H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right)) \\ &\leq \frac{25H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right) + 28H^2\kappa^2. \end{aligned} \quad (25)$$

In the last inequality we used  $\frac{21}{16} \leq \frac{1}{\epsilon'} \leq \frac{1}{\epsilon'} \left( 1 + \frac{4d}{3K} \right)$ , where the first inequality follows because  $\epsilon' \leq \frac{1}{4}$ . Note that (25) holds for every unit vector  $\mathbf{v} \in \mathbb{R}^d$ . Using this and substituting  $\mathbf{g}_{i,\text{accu}}^{t_k,t_{k+1}} = \mathbf{y}_i, \mathbf{g}_{S,\text{accu}}^{t_k,t_{k+1}} = \mathbf{y}_S$  in (25), we get

$$\sup_{\mathbf{v} \in \mathbb{R}^d: \|\mathbf{v}\|=1} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \mathbf{g}_{i,\text{accu}}^{t_k,t_{k+1}} - \mathbf{g}_{S,\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 \leq \frac{25H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right) + 28H^2\kappa^2.$$

This, in view of the alternate definition of the largest eigenvalue given in (14), is equivalent to (11), which proves the first part of [Theorem 3](#).

### 3.2 Proof of the Second Part of [Theorem 3](#)

In this section, we describe the procedure for robust mean estimation in high dimensions from [\[SCV18\]](#) that we use in the second part of [Theorem 3](#) to filter-out corrupt vectors and compute an estimate of the average of uncorrupted accumulated gradients. We refer the reader to [\[DD20, Section 4\]](#) to get an intuition on why filtering-out corrupt gradients (even when  $H = 1$ , i.e., without local iterations) is difficult in high dimensions.

We describe the procedure in [Algorithm 2](#) and refer the reader to [\[DD20, Appendix E\]](#) to get an intuition behind [Algorithm 2](#) and its running-time analysis. Though our algorithm for robust accumulated gradient estimation (RAGE) is the same as the one proposed by Steinhardt et al. [\[SCV18\]](#) for high-dimensional robust mean estimation, we give it a different name, as we are applying the procedure in a much more general federated learning setting; see [Footnote 4](#).

For simplicity, we reorder the received gradient indices from  $1, 2, \dots, K$ . Now, the proof of the second part of [Theorem 3](#) follows from [\[SCV18, Proposition 16\]](#), which we state below for completeness.

---

**Algorithm 2** Robust Accumulated Gradient Estimation (RAGE) [SCV18]

---

- 1: **Initialize.**  $c_i := 1, i \in [K]$ ,  $\alpha := (1 - \tilde{\epsilon}) \geq 3/4$ ,  $\mathcal{A} := \{1, 2, \dots, K\}$ ;  $\mathbf{G} := [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_K] \in \mathbb{R}^{d \times K}$ .
- 2: **while** true **do**
- 3:   Let  $\mathbf{W}^* \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{A}|}$  and  $\mathbf{Y}^* \in \mathbb{R}^{d \times d}$  be the minimizer/maximizer of the saddle point problem:

$$\max_{\substack{\mathbf{Y} \succeq \mathbf{0}, \\ \text{tr}(\mathbf{Y}) \leq 1}} \min_{\substack{0 \leq W_{ji} \leq \frac{4-\alpha}{\alpha(2+\alpha)R}, \\ \sum_{j \in \mathcal{A}} W_{ji} = 1, \forall i \in \mathcal{A}}} \Phi(\mathbf{W}, \mathbf{Y}), \quad (26)$$

where the cost function  $\Phi(\mathbf{W}, \mathbf{Y})$  is defined as

$$\Phi(\mathbf{W}, \mathbf{Y}) := \sum_{i \in \mathcal{A}} c_i (\mathbf{g}_i - \mathbf{G}_{\mathcal{A}} \mathbf{w}_i)^T \mathbf{Y} (\mathbf{g}_i - \mathbf{G}_{\mathcal{A}} \mathbf{w}_i), \quad (27)$$

To avoid cluttered notation, we index the  $|\mathcal{A}|$  rows/columns of  $\mathbf{W}$  by the elements of  $\mathcal{A}$ ;  $\mathbf{G}_{\mathcal{A}}$  denotes the restriction of  $\mathbf{G}$  to the columns in  $\mathcal{A}$ ; for  $i \in \mathcal{A}$ ,  $\mathbf{w}_i$  denotes the column of  $\mathbf{W}$  indexed by  $i$ .

- 4:   For  $i \in \mathcal{A}$ , let

$$\tau_i = (\mathbf{g}_i - \mathbf{G}_{\mathcal{A}} \mathbf{w}_i^*)^T \mathbf{Y}^* (\mathbf{g}_i - \mathbf{G}_{\mathcal{A}} \mathbf{w}_i^*) \quad (28)$$

- 5:   **if**  $\sum_{i \in \mathcal{A}} c_i \tau_i > 4R\sigma_0^2$  **then**
  - 6:     For  $i \in \mathcal{A}$ ,  $c_i \leftarrow \left(1 - \frac{\tau_i}{\tau_{\max}}\right) c_i$ , where  $\tau_{\max} = \max_{j \in \mathcal{A}} \tau_j$ .
  - 7:     For all  $i$  with  $c_i < \frac{1}{2}$ , remove  $i$  from  $\mathcal{A}$ .
  - 8:   **else**
  - 9:     Break **while**-loop
  - 10: **end if**
  - 11: **end while**
  - 12: **return**  $\hat{\mathbf{g}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbf{g}_i$ .
- 

**Lemma 3** (Proposition 16 in [SCV18]). *Suppose we are given  $K$  arbitrary vectors  $\mathbf{g}_1, \dots, \mathbf{g}_K \in \mathbb{R}^d$  with the promise that there exists a subset  $\mathcal{S}$  of these  $K$  vectors such that  $|\mathcal{S}| = (1 - \tilde{\epsilon})K$  for some  $\tilde{\epsilon} > 0$  and  $\mathcal{S}$  satisfies  $\lambda_{\max} \left( \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\mathbf{g}_i - \mathbf{g}_S)(\mathbf{g}_i - \mathbf{g}_S)^T \right) \leq \sigma_0^2$ , where  $\mathbf{g}_S = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{g}_i$  denotes the sample mean of the vectors in  $\mathcal{S}$ . Then, if  $\tilde{\epsilon} \leq \frac{1}{4}$ , **Algorithm 2** can find an estimate  $\hat{\mathbf{g}}$  of  $\mathbf{g}_S$  in polynomial-time, such that  $\|\hat{\mathbf{g}} - \mathbf{g}_S\| \leq \mathcal{O}(\sigma_0 \sqrt{\tilde{\epsilon}})$ .*

Note that **Lemma 3** takes arbitrary vectors as inputs, which are not required to have been generated from a probability distribution.

We refer the reader to [DD20, Appendix F] for a comprehensive proof of **Lemma 3**. To analyze the running time complexity of **Algorithm 2**, first note that (26) can be solved by computing the singular value decomposition (SVD) of a certain  $d \times K$  matrix (see [SCV18, Appendix F] for more details), and second, that **Algorithm 2** removes at least one vector in each iteration of the while loop. So, in the worst case, **Algorithm 2** requires  $\mathcal{O}(dK^2 \min\{d, K\})$  time to execute; see [DD20, Appendix E] for more details on the running time analysis of **Algorithm 2**. Note that this running time does not depend on the total number  $R$  of clients (which may be in millions), and only depends on  $K$ , which is the number of clients selected by the server at synchronization indices. In federated learning,  $R$  may be in millions, but  $K$  is typically a small number, in 1000's.

This completes the proof of the second part of **Theorem 3**.



## 4 Convergence Proof of the Strongly-Convex Part of Theorem 1

At any iteration  $t \in [T]$ , let  $\mathcal{K}_t \subseteq [R]$  denote the set of clients that are active at time  $t$ . Let  $\mathbf{x}^t := \frac{1}{K} \sum_{r \in \mathcal{K}_t} \mathbf{x}_r^t$  denote the average parameter vector of the clients in the active set  $\mathcal{K}_t$ . Note that, for any  $t_i \in \mathcal{I}_T$ , the clients in  $\mathcal{K}_{t_i}$  remain active at all  $t \in [t_i : t_{i+1} - 1]$ .

In the following, we denote the decoded gradient at the server at any synchronization time  $t_{i+1}$  by  $\hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}}$ , which is an estimate of the average of the accumulated gradients between time  $t_i$  and  $t_{i+1}$  of the honest clients in  $\mathcal{K}_{t_i}$ , as in Theorem 3. From Algorithm 1, we can write the parameter update rule for the global model at the synchronization indices as:

$$\mathbf{x}^{t_{i+1}} = \mathbf{x}^{t_i} - \eta \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}}.$$

Note that at any synchronization index  $t_i \in \mathcal{I}_T$ , when the server selects a subset  $\mathcal{K}_{t_i}$  of clients and sends the global parameter vector  $\mathbf{x}^{t_i}$ , all clients in  $\mathcal{K}_{t_i}$  set their local model parameters to be equal to the global model parameters, i.e.,  $\mathbf{x}_r^{t_i} = \mathbf{x}^{t_i}$  holds for every  $r \in \mathcal{K}_{t_i}$ .

First we derive a recurrence relation for the synchronization indices and then for non-synchronization indices. Consider the  $(i+1)$ 'st synchronization index  $t_{i+1} \in \mathcal{I}_T$ . We have

$$\begin{aligned} \mathbf{x}^{t_{i+1}} &= \mathbf{x}^{t_i} - \eta \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} \\ &= \mathbf{x}^{t_i} - \eta \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) - \eta \left( \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right) \end{aligned}$$

For simplicity of notation, define  $\mathcal{E} \triangleq \left( \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right)$ . Substituting this in the above and using  $\mathbf{x}^{t_i} = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbf{x}_r^{t_i}$  gives

$$\begin{aligned} \mathbf{x}^{t_{i+1}} &= \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbf{x}_r^{t_i} - \eta \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) - \eta \mathcal{E} \\ &= \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \left( \mathbf{x}_r^{t_i} - \eta \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right) - \eta \mathcal{E} \\ &= \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\mathbf{x}_r^{t_{i+1}-1} - \eta \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) - \eta \mathcal{E} \\ &= \mathbf{x}^{t_{i+1}-1} - \eta \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \nabla F_r(\mathbf{x}_r^{t_{i+1}-1}) - \eta \mathcal{E} \\ &= \mathbf{x}^{t_{i+1}-1} - \eta \nabla F(\mathbf{x}^{t_{i+1}-1}) + \eta \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) - \eta \mathcal{E} \end{aligned} \quad (29)$$

Subtracting  $\mathbf{x}^*$  from both sides gives:

$$\mathbf{x}^{t_{i+1}} - \mathbf{x}^* = \underbrace{\mathbf{x}^{t_{i+1}-1} - \mathbf{x}^* - \eta \nabla F(\mathbf{x}^{t_{i+1}-1})}_{=: \mathbf{u}} + \eta \underbrace{\frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) - \eta \mathcal{E}}_{=: \mathbf{v}} \quad (30)$$

This gives  $\mathbf{x}^{t_{i+1}} - \mathbf{x}^* = \mathbf{u} + \eta(\mathbf{v} - \mathcal{E})$ . Taking norm on both sides and then squaring gives

$$\|\mathbf{x}^{t_{i+1}} - \mathbf{x}^*\|^2 = \|\mathbf{u}\|^2 + \eta^2 \|\mathbf{v} - \mathcal{E}\|^2 + 2\eta \langle \mathbf{u}, \mathbf{v} - \mathcal{E} \rangle \quad (31)$$

Now we use a simple but powerful trick on inner-products together with the inequality  $2\langle \mathbf{a}, \mathbf{b} \rangle \leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$  and get:

$$2\eta\langle \mathbf{u}, \mathbf{v} - \mathcal{E} \rangle = 2\left\langle \sqrt{\frac{\eta\mu}{2}}\mathbf{u}, \sqrt{\frac{2\eta}{\mu}}(\mathbf{v} - \mathcal{E}) \right\rangle \leq \frac{\eta\mu}{2}\|\mathbf{u}\|^2 + \frac{2\eta}{\mu}\|\mathbf{v} - \mathcal{E}\|^2 \quad (32)$$

Substituting this back into (31) gives

$$\begin{aligned} \|\mathbf{x}^{t_{i+1}} - \mathbf{x}^*\|^2 &\leq \left(1 + \frac{\eta\mu}{2}\right)\|\mathbf{u}\|^2 + \eta\left(\eta + \frac{2}{\mu}\right)\|\mathbf{v} - \mathcal{E}\|^2 \\ &\leq \left(1 + \frac{\eta\mu}{2}\right)\|\mathbf{u}\|^2 + 2\eta\left(\eta + \frac{2}{\mu}\right)\|\mathbf{v}\|^2 + 2\eta\left(\eta + \frac{2}{\mu}\right)\|\mathcal{E}\|^2 \end{aligned}$$

Substituting the values of  $\mathbf{u}, \mathbf{v}, \mathcal{E}$  and taking expectation w.r.t. the stochastic sampling of gradients by clients in  $\mathcal{K}_{t_i}$  between iterations  $t_i$  and  $t_{i+1}$  (while conditioning on the past) gives:

$$\begin{aligned} \mathbb{E}\|\mathbf{x}^{t_{i+1}} - \mathbf{x}^*\|^2 &\leq \left(1 + \frac{\mu\eta}{2}\right)\mathbb{E}\|\mathbf{x}^{t_{i+1}-1} - \eta\nabla F(\mathbf{x}^{t_{i+1}-1}) - \mathbf{x}^*\|^2 \\ &\quad + 2\eta\left(\eta + \frac{2}{\mu}\right)\mathbb{E}\left\|\frac{1}{K}\sum_{r \in \mathcal{K}_{t_i}}(\nabla F(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1}))\right\|^2 \\ &\quad + 2\eta\left(\eta + \frac{2}{\mu}\right)\mathbb{E}\left\|\hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K}\sum_{r \in \mathcal{K}_{t_i}}\sum_{t=t_i}^{t_{i+1}-1}\nabla F_r(\mathbf{x}_r^t)\right\|^2 \end{aligned} \quad (33)$$

Now we bound each of the three terms on the RHS of (33) separately in Claim 3, Claim 4, and Claim 5 below. We prove these claims in Appendix B.

**Claim 3.** For  $\eta < \frac{1}{L}$ , we have

$$\mathbb{E}\|\mathbf{x}^{t_{i+1}-1} - \eta\nabla F(\mathbf{x}^{t_{i+1}-1}) - \mathbf{x}^*\|^2 \leq (1 - \mu\eta)\mathbb{E}\|\mathbf{x}^{t_{i+1}-1} - \mathbf{x}^*\|^2. \quad (34)$$

**Claim 4.** For  $\eta \leq \frac{1}{8HL}$ , we have

$$\mathbb{E}\left\|\frac{1}{K}\sum_{r \in \mathcal{K}_{t_i}}(\nabla F_r(\mathbf{x}_r^{t_{i+1}-1}) - \nabla F(\mathbf{x}^{t_{i+1}-1}))\right\|^2 \leq 2\kappa^2 + \frac{7H}{32}\left(\frac{\sigma^2}{b} + 3\kappa^2\right). \quad (35)$$

**Claim 5.** If  $\eta \leq \frac{1}{8HL}$ , then with probability at least  $1 - \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$ , we have

$$\mathbb{E}\left\|\hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K}\sum_{r \in \mathcal{K}_{t_i}}\sum_{t=t_i}^{t_{i+1}-1}\nabla F_r(\mathbf{x}_r^t)\right\|^2 \leq 3\Upsilon^2 + \frac{8H^2\sigma^2}{b} + 30H^2\kappa^2, \quad (36)$$

where  $\Upsilon^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$  and  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'}\left(1 + \frac{4d}{3K}\right) + 28H^2\kappa^2$ .

Substituting the bounds from (34), (35), (36) into (33) and using  $(1 + \frac{\mu\eta}{2})(1 - \mu\eta) \leq (1 - \frac{\mu\eta}{2})$  for the first term gives

$$\begin{aligned} \mathbb{E}\|\mathbf{x}^{t_{i+1}} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta}{2}\right)\mathbb{E}\|\mathbf{x}^{t_{i+1}-1} - \mathbf{x}^*\|^2 + 2\eta\left(\eta + \frac{2}{\mu}\right)\left(2\kappa^2 + \frac{7H}{32}\left(\frac{\sigma^2}{b} + 3\kappa^2\right)\right) \\ &\quad + 2\eta\left(\eta + \frac{2}{\mu}\right)\left(3\Upsilon^2 + \frac{8H^2\sigma^2}{b} + 30H^2\kappa^2\right) \end{aligned}$$

$$\leq \left(1 - \frac{\mu\eta}{2}\right) \mathbb{E} \|\mathbf{x}^{t_{i+1}-1} - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left(3\Upsilon^2 + \frac{9H^2\sigma^2}{b} + 33H^2\kappa^2\right), \quad (37)$$

where  $\Upsilon^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$  and  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'} (1 + \frac{4d}{3K}) + 28H^2\kappa^2$ . In the last inequality (37) we used  $\eta \leq \frac{1}{8LH} \leq \frac{1}{L} \leq \frac{1}{\mu}$ , which implies  $(\eta + \frac{2}{\mu}) \leq \frac{3}{\mu}$ . Note that (37) holds with probability at least  $1 - \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$ .

Note that the above recurrence in (37) holds only at the synchronization indices  $t_i \in \mathcal{I}_T$  for  $i = 1, 2, 3, \dots$ . However, in order to establish a recurrence that we can use to prove convergence, we need to show a recurrence relation for all  $t$ . Now we give a recurrence at non-synchronization indices.

Take an arbitrary  $t \in [T]$  and let  $t_i \in \mathcal{I}_T$  be such that  $t \in [t_i : t_{i+1} - 1]$ ; when  $H \geq 2$ , such  $t$ 's exist. Note that  $\mathbf{x}^t = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbf{x}_r^t$ . We have

$$\begin{aligned} \mathbf{x}^{t+1} &= \mathbf{x}^t - \eta \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbf{g}_r(\mathbf{x}_r^t) \\ &= \mathbf{x}^t - \eta \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \nabla F_r(\mathbf{x}_r^t) - \eta \left( \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbf{g}_r(\mathbf{x}_r^t) - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \nabla F_r(\mathbf{x}_r^t) \right) \\ &= \mathbf{x}^t - \eta \nabla F(\mathbf{x}^t) + \frac{\eta}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^t) - \nabla F_r(\mathbf{x}_r^t)) - \frac{\eta}{K} \sum_{r \in \mathcal{K}_{t_i}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)) \end{aligned} \quad (38)$$

Now, subtracting  $\mathbf{x}^*$  from both sides and following the same steps that we used to go from (30) to (33), we get (in the following, expectation is taken w.r.t. the stochastic sampling of gradients at the  $t$ 'th iteration while conditioning on the past):

$$\begin{aligned} \mathbb{E} \|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 + \frac{\mu\eta}{2}\right) \mathbb{E} \|\mathbf{x}^t - \mathbf{x}^* - \eta \nabla F(\mathbf{x}^t)\|^2 \\ &\quad + 2\eta \left(\eta + \frac{2}{\mu}\right) \mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \\ &\quad + 2\eta \left(\eta + \frac{2}{\mu}\right) \mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \end{aligned} \quad (39)$$

We can bound the first and the second terms on the RHS of (39) using (34) and (35), respectively, as  $\mathbb{E} \|\mathbf{x}^t - \eta \nabla F(\mathbf{x}^t) - \mathbf{x}^*\|^2 \leq (1 - \mu\eta) \mathbb{E} \|\mathbf{x}^t - \mathbf{x}^*\|^2$  and  $\mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \leq 2\kappa^2 + \frac{7H}{32} \left(\frac{\sigma^2}{b} + 3\kappa^2\right)$ . To bound the third term on the RHS of (39), we use the fact that variance of the sum of independent random variables is equal to the sum of the variances and that clients sample stochastic gradients  $\mathbf{g}_r(\mathbf{x}_r^t)$  independent of each other; using this fact and (5), we have  $\mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \leq \frac{\sigma^2}{bK}$ . Substituting these in (39) and using  $(1 + \frac{\mu\eta}{2})(1 - \mu\eta) \leq (1 - \frac{\mu\eta}{2})$  for the first term and  $(\eta + \frac{2}{\mu}) \leq \frac{3}{\mu}$  (which follows because  $\eta \leq \frac{1}{8HL} \leq \frac{1}{L} \leq \frac{1}{\mu}$ ) give

$$\begin{aligned} \mathbb{E} \|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta}{2}\right) \mathbb{E} \|\mathbf{x}^t - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left(2\kappa^2 + \frac{7H}{32} \left(\frac{\sigma^2}{b} + 3\kappa^2\right) + \frac{\sigma^2}{bK}\right) \\ &\leq \left(1 - \frac{\mu\eta}{2}\right) \mathbb{E} \|\mathbf{x}^t - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left(3H\kappa^2 + \frac{2H\sigma^2}{b}\right) \end{aligned} \quad (40)$$

Note that (40) holds with probability 1.

Now we have a recurrence at the synchronization indices given in (37) and at non-synchronization indices given in (40). Let  $\alpha = (1 - \frac{\mu\eta}{2})$ ,  $\beta_1 = \left(3\Upsilon^2 + \frac{9H^2\sigma^2}{b} + 33H^2\kappa^2\right)$ , and  $\beta_2 = \left(3H\kappa^2 + \frac{2H\sigma^2}{b}\right)$ . Substituting

these and using (37) for the synchronization indices and (40) for the rest of the indices, we get:

$$\mathbb{E} \|\mathbf{x}^T - \mathbf{x}^*\|^2 \leq \alpha^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left( \sum_{i=0}^{T/H} \sum_{j=1}^{H-1} \alpha^{iH+j} \beta_2 + \sum_{i=0}^{T/H} \alpha^{iH} \beta_1 \right) \quad (41)$$

$$\begin{aligned} &\leq \alpha^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left( \sum_{i=0}^{\infty} \alpha^i \beta_2 + \sum_{i=0}^{\infty} \alpha^{iH} \beta_1 \right) \\ &= \alpha^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left( \frac{1}{1-\alpha} \beta_2 + \frac{1}{1-\alpha^H} \beta_1 \right) \end{aligned} \quad (42)$$

Since  $\alpha = (1 - \frac{\mu\eta}{2})$ , we have  $\alpha^H = (1 - \frac{\mu\eta}{2})^H \stackrel{(a)}{\leq} \exp(-\frac{\mu\eta H}{2}) \stackrel{(b)}{\leq} 1 - \frac{\mu\eta H}{2} + \left(\frac{\mu\eta H}{2}\right)^2 \stackrel{(c)}{\leq} 1 - \frac{\mu\eta H}{2} + \frac{1}{16} \frac{\mu\eta H}{2} = 1 - \frac{15}{16} \frac{\mu\eta H}{2}$ . In (a) we used the inequality  $(1 - \frac{1}{x})^x \leq \frac{1}{e}$  which holds for any  $x > 0$ ; in (b) we used  $\exp(-x) \leq 1 - x + x^2$  which holds for any  $x \geq 0$ ; in (c) we used  $\eta \leq \frac{1}{8HL}$  and  $\mu \leq L$ , which together imply  $\frac{\mu\eta H}{2} \leq \frac{1}{16}$ . Substituting these in (42) gives

$$\begin{aligned} \mathbb{E} \|\mathbf{x}^T - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta}{2}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left( \frac{2}{\mu\eta} \beta_2 + \frac{32}{15\mu\eta H} \beta_1 \right) \\ &\leq \left(1 - \frac{\mu\eta}{2}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6 \times 32}{15\mu^2} \left( \frac{15}{16} \beta_2 + \frac{1}{H} \beta_1 \right) \\ &\leq \left(1 - \frac{\mu\eta}{2}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{13}{\mu^2} \left( \frac{3Y^2}{H} + \frac{11H\sigma^2}{b} + 36H\kappa^2 \right) \end{aligned} \quad (43)$$

Note that the last term on the RHS of (43) is independent of  $\eta$ , which together with the dependence of  $\eta$  on the first term implies that bigger the  $\eta$ , faster the convergence. Since we need  $\eta \leq \frac{1}{8HL}$  for Claim 4 and Claim 5 to hold, we choose  $\eta = \frac{1}{8HL}$ . Substituting this in (43) yields the convergence rate (7) of Theorem 1.

**Error probability analysis.** Note that (37) holds with probability at least  $1 - \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$  and (40) holds with probability 1. Since to arrive at (41) (which leads to our final bound (43)), we used (37)  $\frac{T}{H}$  times and (40)  $(T - \frac{T}{H})$  times; as a consequence, by union bound, we have that (43) holds with probability at least  $1 - \frac{T}{H} \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$ , which is at least  $(1 - \delta)$ , for any  $\delta > 0$ , provided we run our algorithm for at most  $T \leq \delta H \exp\left(\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$  iterations.

This concludes the proof of the strongly-convex part of Theorem 1.

## 5 Convergence Proof of the Non-Convex Part of Theorem 1

Let  $\mathcal{K}_t \subseteq [R]$  denote the subset of clients of size  $|\mathcal{K}_t| = K$  sampled at the  $t$ 'th iteration. For any  $t \in [t_i : t_{i+1}-1]$ , let  $\mathbf{x}^t = \frac{1}{K} \sum_{k \in \mathcal{K}_t} \mathbf{x}_k^t$  denote the average of the local parameters of clients in the sampling set  $\mathcal{K}_t$ .

Similar to the proof given in Section 4, here also, first we derive a recurrence for the synchronization indices and then for non-synchronization indices. For the synchronization indices  $t_1, t_2, \dots, t_k, \dots \in \mathcal{I}_T$ , from (29), we have

$$\mathbf{x}^{t_{i+1}} = \mathbf{x}^{t_{i+1}-1} - \eta \nabla F(\mathbf{x}^{t_{i+1}-1}) + \eta C \quad (44)$$

where

$$C = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) - \left( \widehat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right). \quad (45)$$

Now, using the definition of  $L$ -smoothness in (44), we have

$$\begin{aligned}
F(\mathbf{x}^{t_{i+1}}) &\leq F(\mathbf{x}^{t_{i+1}-1}) + \langle \nabla F(\mathbf{x}^{t_{i+1}-1}), \mathbf{x}^{t_{i+1}} - \mathbf{x}^{t_{i+1}-1} \rangle + \frac{L}{2} \|\mathbf{x}^{t_{i+1}} - \mathbf{x}^{t_{i+1}-1}\|^2 \\
&= F(\mathbf{x}^{t_{i+1}-1}) - \eta \langle \nabla F(\mathbf{x}^{t_{i+1}-1}), \nabla F(\mathbf{x}^{t_{i+1}-1}) - C \rangle + \frac{\eta^2 L}{2} \|\nabla F(\mathbf{x}^{t_{i+1}-1}) - C\|^2 \\
&= F(\mathbf{x}^{t_{i+1}-1}) - \eta \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \eta \langle \nabla F(\mathbf{x}^{t_{i+1}-1}), C \rangle + \frac{\eta^2 L}{2} \|\nabla F(\mathbf{x}^{t_{i+1}-1}) - C\|^2 \\
&\stackrel{(a)}{\leq} F(\mathbf{x}^{t_{i+1}-1}) - \eta \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \eta \left( \frac{\|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2}{4} + \|C\|^2 \right) \\
&\quad + \frac{\eta^2 L}{2} \|\nabla F(\mathbf{x}^{t_{i+1}-1}) - C\|^2 \\
&\stackrel{(b)}{\leq} F(\mathbf{x}^{t_{i+1}-1}) - \frac{3\eta}{4} \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \eta \|C\|^2 + \eta^2 L \left( \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \|C\|^2 \right) \\
&= F(\mathbf{x}^{t_{i+1}-1}) - \eta \left( \frac{3}{4} - \eta L \right) \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \eta (1 + \eta L) \|C\|^2
\end{aligned} \tag{46}$$

In (a), we used the inequality  $2\langle \mathbf{a}, \mathbf{b} \rangle \leq \tau \|\mathbf{a}\|^2 + \frac{1}{\tau} \|\mathbf{b}\|^2$ , which holds for every  $\tau > 0$ , and we used  $\tau = \frac{1}{2}$  in (a). In (b), we used the inequality  $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$ . For  $\eta \leq \frac{1}{8HL} \leq \frac{1}{8L}$ , we have  $(\frac{3}{4} - \eta L) \geq \frac{1}{2}$  and  $(1 + \eta L) \leq \frac{9}{8}$ . Substituting these in (46) and taking expectation w.r.t. the stochastic sampling of gradients at clients in  $\mathcal{K}_{t_i}$  between iterations  $t_i$  and  $t_{i+1}$  (while conditioning on the past) gives:

$$\mathbb{E}[F(\mathbf{x}^{t_{i+1}})] \leq \mathbb{E}[F(\mathbf{x}^{t_{i+1}-1})] - \frac{\eta}{2} \mathbb{E} \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \frac{9\eta}{8} \mathbb{E} \|C\|^2. \tag{47}$$

Now we bound  $\mathbb{E} \|C\|^2$ . Substituting the value of  $C$  from (45) gives:

$$\begin{aligned}
\mathbb{E} \|C\|^2 &\leq 2\mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) \right\|^2 + 2\mathbb{E} \left\| \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right\|^2 \\
&\leq 2 \left( 2\kappa^2 + \frac{7H}{32} \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \right) + 2 \left( 3\Upsilon^2 + \frac{8H^2\sigma^2}{b} + 30H^2\kappa^2 \right) \\
&\leq 2 \left( 3\Upsilon^2 + \frac{9H^2\sigma^2}{b} + 33H^2\kappa^2 \right)
\end{aligned} \tag{48}$$

Here, the first inequality used  $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$  and the second inequality used the bounds from (35) and (36).

Substituting the bound from (48) into (47) gives

$$\mathbb{E}[F(\mathbf{x}^{t_{i+1}})] \leq \mathbb{E}[F(\mathbf{x}^{t_{i+1}-1})] - \frac{\eta}{2} \mathbb{E} \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \frac{9\eta}{4} \left( 3\Upsilon^2 + \frac{9H^2\sigma^2}{b} + 33H^2\kappa^2 \right) \tag{49}$$

where  $\Upsilon^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$  and  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'} (1 + \frac{4d}{3K}) + 28H^2\kappa^2$ . Note that (49) holds with probability at least  $1 - \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$ .

Note that the above recurrence in (49) holds only at the synchronization indices  $t_i \in \mathcal{I}_T$  for  $i = 1, 2, 3, \dots$ . Now we give a recurrence at non-synchronization indices.

We have done a similar calculation in the strongly-convex part of Theorem 1 in Section 4. Take an arbitrary  $t \in [T]$  and let  $t_i \in \mathcal{I}_T$  be such that  $t \in [t_i : t_{i+1} - 1]$ ; when  $H \geq 2$ , such  $t$ 's exist. Note that  $\mathbf{x}^t = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbf{x}_r^t$ .

From (38), we have  $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta \nabla F(\mathbf{x}^t) + \eta D$ , where

$$D = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^t) - \nabla F_r(\mathbf{x}_r^t)) - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)).$$

Using  $L$ -smoothness of  $F$ , and then performing similar algebraic manipulations that we used in order to arrive at (47), we get:

$$\mathbb{E}[F(\mathbf{x}^{t+1})] \leq \mathbb{E}[F(\mathbf{x}^t)] - \frac{\eta}{2} \mathbb{E} \|\nabla F(\mathbf{x}^t)\|^2 + \frac{9\eta}{8} \mathbb{E} \|D\|^2 \quad (50)$$

Now we bound  $\mathbb{E} \|D\|^2$ :

$$\begin{aligned} \mathbb{E} \|D\|^2 &\leq 2\mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 + 2\mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \\ &\leq 2 \left( 2\kappa^2 + \frac{7H}{32} \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) + \frac{\sigma^2}{bK} \right) \\ &\leq 2 \left( 3H\kappa^2 + \frac{2H\sigma^2}{b} \right) \end{aligned} \quad (51)$$

Here, the second inequality used the same bounds on both the quantities on the RHS of the first inequality that we used to go from (39) to (40).

Substituting the bound on  $\mathbb{E} \|D\|^2$  from (51) into (50) gives

$$\mathbb{E}[F(\mathbf{x}^{t+1})] \leq \mathbb{E}[F(\mathbf{x}^t)] - \frac{\eta}{2} \mathbb{E} \|\nabla F(\mathbf{x}^t)\|^2 + \frac{9\eta}{4} \left( 3H\kappa^2 + \frac{2H\sigma^2}{b} \right) \quad (52)$$

Note that (52) holds with probability 1.

Now we have a recurrence at synchronization indices given in (49) and at non-synchronization indices given in (52). Adding (49) and (52) from  $t = 0$  to  $T$  (use (49) for the synchronization indices and (52) for the rest of the indices) gives:

$$\begin{aligned} \sum_{t=0}^T \mathbb{E}[F(\mathbf{x}^{t+1})] &\leq \sum_{t=0}^T \mathbb{E}[F(\mathbf{x}^t)] - \frac{\eta}{2} \sum_{t=0}^T \mathbb{E} \|\nabla F(\mathbf{x}^t)\|^2 + \frac{9\eta}{4} \left[ \frac{T}{H} \left( 3\Upsilon^2 + \frac{9H^2\sigma^2}{b} + 33H^2\kappa^2 \right) \right. \\ &\quad \left. + \left( T - \frac{T}{H} \right) \left( 3H\kappa^2 + \frac{2H\sigma^2}{b} \right) \right] \end{aligned} \quad (53)$$

We can simplify the constant term in the RHS of (53) as follows:

$$\begin{aligned} &\frac{1}{H} \left( 3\Upsilon^2 + \frac{9H^2\sigma^2}{b} + 33H^2\kappa^2 \right) + \left( 1 - \frac{1}{H} \right) \left( 3H\kappa^2 + \frac{2H\sigma^2}{b} \right) \\ &\leq \frac{1}{H} \left( 3\Upsilon^2 + \frac{9H^2\sigma^2}{b} + 33H^2\kappa^2 \right) + \left( 3H\kappa^2 + \frac{2H\sigma^2}{b} \right) \\ &\leq \frac{3\Upsilon^2}{H} + \frac{11H\sigma^2}{b} + 36H\kappa^2 \end{aligned}$$

Substituting this in (53) and then rearranging, we get:

$$\frac{1}{T} \sum_{t=0}^T \mathbb{E} \|\nabla F(\mathbf{x}^t)\|^2 \leq \frac{2}{\eta T} [\mathbb{E}[F(\mathbf{x}^0)] - \mathbb{E}[F(\mathbf{x}^{T+1})]] + \frac{9}{2} \left( \frac{3\Upsilon^2}{H} + \frac{11H\sigma^2}{b} + 36H\kappa^2 \right) \quad (54)$$



Note that the last term in (54) is a constant. So, it would be best to take the step-size  $\eta$  to be as large as possible such that it satisfies  $\eta \leq \frac{1}{8HL}$ . We take  $\eta = \frac{1}{8HL}$ . Substituting this in (54) and using  $F(\mathbf{x}^{T+1}) \geq F(\mathbf{x}^*)$  gives

$$\frac{1}{T} \sum_{t=0}^T \mathbb{E} \|\nabla F(\mathbf{x}^t)\|^2 \leq \frac{16HL}{T} [\mathbb{E}[F(\mathbf{x}^0)] - \mathbb{E}[F(\mathbf{x}^*)]] + \frac{9}{2} \left( \frac{3\Upsilon^2}{H} + \frac{11H\sigma^2}{b} + 36H\kappa^2 \right), \quad (55)$$

where  $\Upsilon^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$  and  $\sigma_0^2 = \frac{25H^2\sigma^2}{be'} (1 + \frac{4d}{3K}) + 28H^2\kappa^2$ . Note that (55) is the convergence rate (8) in Theorem 1.

**Error probability analysis.** Note that (49) holds with probability at least  $1 - \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$  and (52) holds with probability 1. Since to arrive at (53) (which leads to our final bound (55)), we used (49)  $\frac{T}{H}$  times and (52)  $(T - \frac{T}{H})$  times; as a consequence, by union bound, we have that (55) holds with probability at least  $1 - \frac{T}{H} \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$ , which is at least  $(1 - \delta)$ , for any  $\delta > 0$ , provided we run our algorithm for at most  $T \leq \delta H \exp(\frac{\epsilon'^2(1-\epsilon)K}{16})$  iterations.

This concludes the proof of the non-convex part of Theorem 1.

## 6 Convergence Proof of Theorem 2

In this section, we focus on the case when in each local iteration clients compute *full-batch* gradients (instead of computing mini-batch stochastic gradients) in Algorithm 1 and prove Theorem 2. Note that the robust accumulated gradient estimation (RAGE) result of Theorem 3 (which is for stochastic gradients) is one of the main ingredients behind the convergence analyses of Theorem 1. So, in order to prove Theorem 2, first we need to show a RAGE result for full-batch gradients. Note that we can obtain such a result by substituting  $\sigma = 0$  in both the parts of Theorem 3; however, this would give a loose bound on the approximation error in the second part. In the following, we get a tighter bound (both for RAGE and the convergence rates in Theorem 2) by working directly with full-batch gradients. To get a RAGE result for full-batch gradients, we do a much simplified analysis than what we did before to prove Theorem 3, and the resulting result is stated and proved below in Theorem 4.

Note that, in order to prove Theorem 3, we showed an existence of a subset  $\mathcal{S}$  of honest clients (from the set  $\mathcal{K}$  of clients who communicate with the server) from whom the accumulated stochastic gradients are well-concentrated, as stated in form of a matrix concentration bound (11) in the first part of Theorem 3. It turns out that for full-batch gradients, an analogous result can be proven directly (as there is no randomness due to stochastic gradients); and below we provide such a result. Note that Theorem 3 is a probabilistic statement, where we show that with high probability, there exists a large subset  $\mathcal{S} \subseteq \mathcal{K}$  of honest clients whose stochastic accumulated gradients are well-concentrated. In contrast, in the following result, we can deterministically take the set of *all* honest clients in  $\mathcal{K}$  to be that subset for which we can directly show the concentration.

First we setup the notation to state our main result on RAGE for full-batch gradients. Let  $\mathcal{K}_t \subseteq [R]$  denote the subset of clients of size  $K$  that are active at any time  $t \in [0 : T]$ . Let Algorithm 1 generate a sequence of iterates  $\{\mathbf{x}_r^t : t \in [0 : T], r \in \mathcal{K}_t\}$  when run with a fixed step-size  $\eta$  satisfying  $\eta \leq \frac{1}{5HL}$  while minimizing a global objective function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$ , where in any iteration, instead of sampling mini-batch stochastic gradients, every honest client takes full-batch gradients from their local datasets. Take any two consecutive synchronization indices  $t_k, t_{k+1} \in \mathcal{I}_T$ . Note that  $|t_{k+1} - t_k| \leq H$ . For an honest client  $r \in \mathcal{K}_{t_k}$ , let  $\nabla F_{r, \text{accu}}^{t_k, t_{k+1}} := \sum_{t=t_k}^{t_{k+1}-1} \nabla F_r(\mathbf{x}_r^t)$  denote the sum of local full-batch gradients taken by client  $r$  between time  $t_k$  and  $t_{k+1}$ . Note that at iteration  $t_{k+1}$ , every honest client  $r \in \mathcal{K}_{t_k}$  reports its local parameters  $\mathbf{x}_r^{t_{k+1}}$  to the server, from which server can compute  $\nabla F_{r, \text{accu}}^{t_k, t_{k+1}}$ , whereas, corrupt clients may report arbitrary and adversarially chosen vectors in  $\mathbb{R}^d$ . The goal of the server is to produce an estimate  $\widehat{\nabla F}_{\text{accu}}^{t_k, t_{k+1}}$  of the average accumulated gradients from honest clients as best as possible.

**Theorem 4** (Robust Accumulated Gradient Estimation for Full-Batch Gradient Descent). *Suppose an  $\epsilon$  fraction of clients who communicate with the server are corrupt. In the setting and notation described above, suppose we are given  $K \leq R$  accumulated full-batch gradients  $\nabla \tilde{F}_{r,\text{accu}}^{t_k,t_{k+1}}, r \in \mathcal{K}_{t_k}$  in  $\mathbb{R}^d$ , where  $\nabla \tilde{F}_{r,\text{accu}}^{t_k,t_{k+1}} = \nabla F_{r,\text{accu}}^{t_k,t_{k+1}}$  if the  $r$ 'th client is honest, otherwise can be arbitrary. Let  $\mathcal{S} \subseteq \mathcal{K}_{t_k}$  be the subset of all honest clients in  $\mathcal{K}_{t_k}$  and  $\nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla F_{i,\text{accu}}^{t_k,t_{k+1}}$  be the sample average of uncorrupted full-batch gradients. If  $\epsilon \leq \frac{1}{4}$ , then with probability 1, we can find an estimate  $\nabla \hat{F}_{\text{accu}}^{t_k,t_{k+1}}$  of  $\nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}}$  in polynomial-time, such that  $\left\| \nabla \hat{F}_{\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}} \right\| \leq \mathcal{O}(H\kappa\sqrt{\epsilon})$ .*

*Proof.* First we prove that

$$\lambda_{\max} \left( \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left( \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}} \right) \left( \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}} \right)^T \right) \leq 11H^2\kappa^2. \quad (56)$$

In view of the alternate characterization the largest eigenvalue given in (14), this is equivalent to showing

$$\sup_{\mathbf{v} \in \mathbb{R}^d: \|\mathbf{v}\|=1} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 \leq 11H^2\kappa^2, \quad (57)$$

which we prove below. Define  $F_{\text{accu}}^{t_k,t_{k+1}} := \sum_{t=t_k}^{t_{k+1}-1} F(\mathbf{x}^t)$ , where  $\mathbf{x}^t = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_k}} \mathbf{x}_r^t$  for any  $t \in [t_k : t_{k+1} - 1]$ . Take an arbitrary unit vector  $\mathbf{v} \in \mathbb{R}^d$ .

$$\begin{aligned} & \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 \\ &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left[ \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}} + \nabla F_{\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle \right]^2 \\ &\leq \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 \\ &\quad \text{(Using } \|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2) \\ &= \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 + 2 \left\langle \nabla F_{\mathcal{S},\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 \\ &= \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 + 2 \left[ \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle \right]^2 \\ &\leq \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 + \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 \\ &= \frac{4}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}}, \mathbf{v} \right\rangle^2 \\ &\leq \frac{4}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\| \nabla F_{i,\text{accu}}^{t_k,t_{k+1}} - \nabla F_{\text{accu}}^{t_k,t_{k+1}} \right\|^2 \\ &\quad \text{(Using Cauchy-Schwarz inequality } \langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\| \|\mathbf{v}\| \text{ and that } \|\mathbf{v}\| = 1) \\ &= \frac{4}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\| \sum_{t=t_k}^{t_{k+1}-1} (\nabla F_i(\mathbf{x}_i^t) - \nabla F(\mathbf{x}^t)) \right\|^2 \quad \text{(Since } F_{\text{accu}}^{t_k,t_{k+1}} = \sum_{t=t_k}^{t_{k+1}-1} F(\mathbf{x}^t)) \\ &\leq \frac{4}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (t_{k+1} - t_k) \sum_{t=t_k}^{t_{k+1}-1} \left\| \nabla F_i(\mathbf{x}_i^t) - \nabla F(\mathbf{x}^t) \right\|^2 \quad \text{(Using Jensen's inequality)} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{4H}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{t=t_k}^{t_{k+1}-1} \left( 2 \|\nabla F_i(\mathbf{x}_i^t) - \nabla F(\mathbf{x}_i^t)\|^2 + 2 \|\nabla F(\mathbf{x}_i^t) - \nabla F(\mathbf{x}^t)\|^2 \right) \\
&\stackrel{(a)}{\leq} \frac{4H}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{t=t_k}^{t_{k+1}-1} \left( 2\kappa^2 + 2L^2 \|\mathbf{x}_i^t - \mathbf{x}^t\|^2 \right) \\
&\leq 8H^2\kappa^2 + 8HL^2 \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\| \mathbf{x}_i^t - \frac{1}{K} \sum_{j \in \mathcal{K}_{t_k}} \mathbf{x}_j^t \right\|^2 \quad (\text{Since } \mathbf{x}^t = \frac{1}{K} \sum_{j \in \mathcal{K}_{t_k}} \mathbf{x}_j^t) \\
&\leq 8H^2\kappa^2 + 8HL^2 \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{1}{K} \sum_{j \in \mathcal{K}_{t_k}} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2 \tag{58}
\end{aligned}$$

The last inequality follows from the Jensen's inequality. In (a) we used (6) to bound  $\|\nabla F_i(\mathbf{x}_i^t) - \nabla F(\mathbf{x}_i^t)\|^2 \leq \kappa^2$  and  $L$ -Lipschitz gradient property of  $F$  to bound  $\|\nabla F(\mathbf{x}_i^t) - \nabla F(\mathbf{x}^t)\| \leq L\|\mathbf{x}_i^t - \mathbf{x}^t\|$ .

Now we bound the last term of (58).

**Lemma 4.** For any  $r, s \in \mathcal{K}_{t_k}$ , if  $\eta \leq \frac{1}{5HL}$ , we have

$$\sum_{t=t_k}^{t_{k+1}-1} \|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2 \leq 7\eta^2 H^3 \kappa^2. \tag{59}$$

*Proof.* Note that we have shown a similar result (but, in expectation) in Lemma 2 (on page 12), which is for stochastic gradients. We will simplify that proof to prove Lemma 4, which is for full-batch deterministic gradients.

Take an arbitrary  $t \in [t_k : t_{k+1} - 1]$ . Following the proof of Lemma 2 until (23) and removing the factor of 3 inside the summation (the factor of 3 appeared because we applied the Jensen's inequality earlier to separate the deterministic gradient term and the stochastic gradient terms) would give

$$\|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2 \leq \eta^2 H \sum_{j=t_k}^{t-1} \|\nabla F_r(\mathbf{x}_r^j) - \nabla F_s(\mathbf{x}_s^j)\|^2. \tag{60}$$

Following the remaining proof of Lemma 2 from (23) until the end and substituting  $\sigma = 0$  gives the desired result.  $\square$

Substituting the bound from (59) into (58) gives

$$\begin{aligned}
\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \left\langle \nabla F_{i,\text{accu}}^{t_k, t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k, t_{k+1}}, \mathbf{v} \right\rangle^2 &\leq 8H^2\kappa^2 + 56H^4L^2\eta^2\kappa^2 \\
&\leq 8H^2\kappa^2 + \frac{56}{25}H^2\kappa^2 \quad (\text{Substituting } \eta \leq \frac{1}{5HL}) \\
&\leq 11H^2\kappa^2. \tag{61}
\end{aligned}$$

Note that (61) holds for an arbitrary unit vector  $\mathbf{v} \in \mathbb{R}^d$ , implying that (57) holds true. Since (57) and (56) are equivalent, we have thus shown (56).

Now apply the second part of Theorem 3 with  $\mathcal{S}$  being the set of all honest clients, and  $\mathbf{g}_{i,\text{accu}}^{t_k, t_{k+1}} = \nabla F_{i,\text{accu}}^{t_k, t_{k+1}}$ ,  $\mathbf{g}_{\mathcal{S},\text{accu}}^{t_k, t_{k+1}} = \nabla F_{\mathcal{S},\text{accu}}^{t_k, t_{k+1}}$ ,  $\hat{\mathbf{g}}_{\text{accu}}^{t_k, t_{k+1}} = \nabla \hat{F}_{\text{accu}}^{t_k, t_{k+1}}$ ,  $\epsilon' = 0$ , and  $\sigma_0^2 = 11H^2\kappa^2$ . We would get that we can find an estimate  $\nabla \hat{F}_{\text{accu}}^{t_k, t_{k+1}}$  of  $\nabla F_{\mathcal{S},\text{accu}}^{t_k, t_{k+1}}$  in polynomial-time, such that  $\|\nabla \hat{F}_{\text{accu}}^{t_k, t_{k+1}} - \nabla F_{\mathcal{S},\text{accu}}^{t_k, t_{k+1}}\| \leq \mathcal{O}(H\kappa\sqrt{\epsilon})$  holds with probability 1.  $\square$

Theorem 2 can be proved with appropriate modifications in the proof of Theorem 1, and we prove it in Appendix C.

## Acknowledgement

This work was supported by the NSF grants #1740047, #1514731, and by the UC-NL grant LFR-18-548554.

## References

- [AAL18] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Neural Information Processing Systems (NeurIPS)*, pages 4618–4628, 2018.
- [BDKD19] Debraj Basu, Deepesh Data, Can Karakus, and Suhas N. Diggavi. Qsparse-local-sgd: Distributed SGD with quantization, sparsification and local computations. In *NeurIPS*, pages 14668–14679, 2019.
- [BMGS17] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NIPS*, pages 119–129, 2017.
- [Bot10] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186, 2010.
- [CSX17] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *POMACS*, 1(2):44:1–44:25, 2017.
- [CWCP18] Lingjiao Chen, Hongyi Wang, Zachary B. Charles, and Dimitris S. Papailiopoulos. DRACO: byzantine-resilient distributed training via redundant gradients. In *ICML*, pages 902–911, 2018.
- [DCM<sup>+</sup>12] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc’Aurelio Ranzato, Andrew W. Senior, Paul A. Tucker, Ke Yang, and Andrew Y. Ng. Large scale distributed deep networks. In *Neural Information Processing Systems (NIPS)*, pages 1232–1240, 2012.
- [DD19] Deepesh Data and Suhas N. Diggavi. Byzantine-tolerant distributed coordinate descent. In *ISIT*, pages 2724–2728, 2019.
- [DD20] Deepesh Data and Suhas N. Diggavi. Byzantine-resilient SGD in high dimensions on heterogeneous data. *CoRR*, abs/2005.07866, 2020.
- [DK19] Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.
- [DKK<sup>+</sup>19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019. Preliminary version appeared in FOCS 2016.
- [DSD19a] Deepesh Data, Linqi Song, and Suhas N. Diggavi. Data encoding for byzantine-resilient distributed optimization. *CoRR*, abs/1907.02664, 2019.
- [DSD19b] Deepesh Data, Linqi Song, and Suhas N. Diggavi. Data encoding methods for byzantine-resilient distributed optimization. In *ISIT*, pages 2719–2723, 2019.
- [GHYR19] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *CoRR*, abs/1906.06629, 2019.
- [HKJ20] Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via resampling. *CoRR*, abs/2006.09365, 2020.

- [HKMC19] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck R. Cadambe. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Neural Information Processing Systems (NeurIPS)*, pages 11080–11092, 2019.
- [HM19] Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *CoRR*, abs/1910.14425, 2019.
- [K<sup>+</sup>19] Peter Kairouz et al. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019.
- [KKM<sup>+</sup>19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for on-device federated learning. *CoRR*, abs/1910.06378, 2019.
- [KMR19] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. *CoRR*, abs/1909.04746, 2019. To appear in AISTATS 2020.
- [KMRR16] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.
- [Kon17] Jakub Konečný. Stochastic, distributed and federated optimization for machine learning. *CoRR*, abs/1707.01155, 2017.
- [LHY<sup>+</sup>20] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations (ICLR)*, 2020.
- [LRV16] Kevin A. Lai, Anup B. Rao, and Santosh S. Vempala. Agnostic estimation of mean and covariance. In *FOCS*, pages 665–674, 2016.
- [LXC<sup>+</sup>19] Liping Li, Wei Xu, Tianyi Chen, Georgios B. Giannakis, and Qing Ling. RSA: byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Conference on Artificial Intelligence (AAAI)*, pages 1544–1551, 2019.
- [LYWZ19] Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication efficient decentralized training with multiple local updates. *CoRR*, abs/1910.09126, 2019.
- [MMR<sup>+</sup>17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.
- [MSS19] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning (ICML)*, pages 4615–4625, 2019.
- [RWCP19] Shashank Rajput, Hongyi Wang, Zachary B. Charles, and Dimitris S. Papailiopoulos. DETOX: A redundancy-based framework for faster and more robust gradient aggregation. In *NeurIPS*, pages 10320–10330, 2019.
- [Sa] Christopher De Sa. Simple techniques for improving sgd. CS6787 Lecture 2 – Fall 2017.
- [SCV18] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *ITCS*, pages 45:1–45:21, 2018.
- [SLS<sup>+</sup>20] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Conference on Machine Learning and Systems (MLSys)*, 2020.

- [SX19] Lili Su and Jiaming Xu. Securing distributed gradient descent in high dimensional statistical learning. *POMACS*, 3(1):12:1–12:41, 2019.
- [XKG19a] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. SLSGD: secure and efficient distributed on-device machine learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD, Proceedings, Part II*, pages 213–228, 2019.
- [XKG19b] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International Conference on Machine Learning (ICML)*, pages 6893–6901, 2019.
- [YCRB18] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *ICML*, pages 5636–5645, 2018.
- [YCRB19] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Defending against saddle point attack in byzantine-robust distributed learning. In *ICML*, pages 7074–7084, 2019.
- [YJY19] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *ICML*, pages 7184–7193, 2019.
- [YYZ19] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Conference on Artificial Intelligence (AAAI)*, pages 5693–5700, 2019.

## A Omitted Details from Section 3.1

In this section, we prove [Claim 1](#).

**Claim** (Restating [Claim 1](#)). *For any honest client  $i \in \mathcal{K}_{t_k}$ , we have  $\mathbb{E}\|Y_i - \mathbb{E}[Y_i]\|^2 \leq \frac{H^2\sigma^2}{b}$ , where expectation is taken over sampling stochastic gradients by client  $i$  between the synchronization indices  $t_k$  and  $t_{k+1}$ .*

*Proof.* Take an arbitrary honest client  $i \in \mathcal{K}_{t_k}$ .

$$\mathbb{E}\|Y_i - \mathbb{E}[Y_i]\|^2 = \mathbb{E}\left\|\sum_{t=t_k}^{t_{k+1}-1} (Y_i^t - \mathbb{E}[Y_i^t])\right\|^2 \stackrel{(a)}{\leq} (t_{k+1} - t_k) \sum_{t=t_k}^{t_{k+1}-1} \mathbb{E}\|Y_i^t - \mathbb{E}[Y_i^t]\|^2 \stackrel{(b)}{\leq} \frac{H^2\sigma^2}{b},$$

where (a) follows from the Jensen’s inequality; in (b) we used  $(t_{k+1} - t_k) \leq H$  and that  $\mathbb{E}\|Y_i^t - \mathbb{E}[Y_i^t]\|^2 \leq \frac{\sigma^2}{b}$  for all  $j \in [H]$ , which follows from the explanation below:

$$\begin{aligned} \mathbb{E}\|Y_i^t - \mathbb{E}[Y_i^t]\|^2 &= \sum_{\mathbf{y}_i^{t_k}, \dots, \mathbf{y}_i^{t-1}} \Pr\left[Y_i^j = \mathbf{y}_i^j, j \in [t_k : t-1]\right] \\ &\quad \times \mathbb{E}\left[\|Y_i^t - \mathbb{E}[Y_i^t]\|^2 \mid Y_i^j = \mathbf{y}_i^j, j \in [t_k : t-1]\right] \\ &\stackrel{(c)}{\leq} \sum_{\mathbf{y}_i^{t_k}, \dots, \mathbf{y}_i^{t-1}} \Pr\left[Y_i^j = \mathbf{y}_i^j, j \in [t_k : t-1]\right] \cdot \frac{\sigma^2}{b} \\ &= \frac{\sigma^2}{b}. \end{aligned}$$

Note that  $Y_i^t \sim \text{Unif}\left(\mathcal{F}_i^{\otimes b}(\mathbf{x}_i^t(\mathbf{x}_i^{t_k}, Y_i^{t_k}, \dots, Y_i^{t-1}))\right)$ . So, when we fix the values  $Y_i^{t_k} = \mathbf{y}_i^{t_k}, \dots, Y_i^{t-1} = \mathbf{y}_i^{t-1}$ , the parameter vector  $\mathbf{x}_i^t(\mathbf{x}_i^{t_k}, Y_i^{t_k}, \dots, Y_i^{t-1})$  becomes a deterministic quantity. Now we can use the variance bound (5) in order to bound  $\mathbb{E}\left[\|Y_i^t - \mathbb{E}[Y_i^t]\|^2 \mid Y_i^j = \mathbf{y}_i^j, j \in [t_k : t-1]\right] \leq \frac{\sigma^2}{b}$ . This is what we used in (c).  $\square$



## B Omitted Details from Section 4

In this section, we prove [Claim 3](#), [Claim 4](#), and [Claim 5](#).

**Claim** (Restating [Claim 3](#)). *For  $\eta < \frac{1}{L}$ , we have*

$$\mathbb{E} \left\| \mathbf{x}^{t_{i+1}-1} - \eta \nabla F(\mathbf{x}^{t_{i+1}-1}) - \mathbf{x}^* \right\|^2 \leq (1 - \mu\eta) \mathbb{E} \left\| \mathbf{x}^{t_{i+1}-1} - \mathbf{x}^* \right\|^2.$$

*Proof.* Expand the LHS.

$$\begin{aligned} \mathbb{E} \left\| \mathbf{x}^{t_{i+1}-1} - \mathbf{x}^* - \eta \nabla F(\mathbf{x}^{t_{i+1}-1}) \right\|^2 &= \mathbb{E} \left\| \mathbf{x}^{t_{i+1}-1} - \mathbf{x}^* \right\|^2 + \eta^2 \mathbb{E} \left\| \nabla F(\mathbf{x}^{t_{i+1}-1}) \right\|^2 \\ &\quad + 2\eta \mathbb{E} \langle \mathbf{x}^* - \mathbf{x}^{t_{i+1}-1}, \nabla F(\mathbf{x}^{t_{i+1}-1}) \rangle \end{aligned} \quad (62)$$

We can bound the second term on the RHS using  $L$ -smoothness of  $F$ , which implies that  $\|\nabla F(\mathbf{x})\|^2 \leq 2L(F(\mathbf{x}) - F(\mathbf{x}^*))$  holds for every  $\mathbf{x} \in \mathbb{R}^d$ ; see [Fact 1](#) on page 30. We can bound the third term on the RHS using  $\mu$ -strong convexity of  $F$  as follows:  $\langle \mathbf{x}^* - \mathbf{x}^{t_{i+1}-1}, \nabla F(\mathbf{x}^{t_{i+1}-1}) \rangle \leq F(\mathbf{x}^*) - F(\mathbf{x}^{t_{i+1}-1}) - \frac{\mu}{2} \|\mathbf{x}^{t_{i+1}-1} - \mathbf{x}^*\|^2$ . Substituting these back in (62) gives:

$$\begin{aligned} \mathbb{E} \left\| \mathbf{x}^{t_{i+1}-1} - \mathbf{x}^* - \eta \nabla F(\mathbf{x}^{t_{i+1}-1}) \right\|^2 &\leq (1 - \mu\eta) \mathbb{E} \left\| \mathbf{x}^{t_{i+1}-1} - \mathbf{x}^* \right\|^2 \\ &\quad - 2\eta(1 - \eta L) \mathbb{E} (F(\mathbf{x}^{t_{i+1}-1}) - F(\mathbf{x}^*)) \end{aligned} \quad (63)$$

Since  $\eta < \frac{1}{L}$ , we have  $(1 - \eta L) > 0$ . We also have  $F(\mathbf{x}^{t_{i+1}-1}) \geq F(\mathbf{x}^*)$ . Using these together, we can ignore the last term in the RHS of (63). This proves [Claim 3](#).  $\square$

**Claim** (Restating [Claim 4](#)). *For  $\eta \leq \frac{1}{8HL}$ , we have*

$$\mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F_r(\mathbf{x}_r^{t_{i+1}-1}) - \nabla F(\mathbf{x}^{t_{i+1}-1})) \right\|^2 \leq 2\kappa^2 + \frac{7H}{32} \left( \frac{\sigma^2}{b} + 3\kappa^2 \right).$$

*Proof.* By definition, we have  $\mathbf{x}^{t_{i+1}-1} = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbf{x}_r^{t_{i+1}-1}$ .

$$\begin{aligned} &\mathbb{E} \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F_r(\mathbf{x}_r^{t_{i+1}-1}) - \nabla F(\mathbf{x}^{t_{i+1}-1})) \right\|^2 \leq \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \nabla F_r(\mathbf{x}_r^{t_{i+1}-1}) - \nabla F(\mathbf{x}^{t_{i+1}-1}) \right\|^2 \\ &\leq \frac{2}{K} \sum_{r \in \mathcal{K}_{t_i}} \left( \mathbb{E} \left\| \nabla F_r(\mathbf{x}_r^{t_{i+1}-1}) - \nabla F(\mathbf{x}_r^{t_{i+1}-1}) \right\|^2 + \mathbb{E} \left\| \nabla F(\mathbf{x}_r^{t_{i+1}-1}) - \nabla F(\mathbf{x}^{t_{i+1}-1}) \right\|^2 \right) \\ &\stackrel{(a)}{\leq} \frac{2}{K} \sum_{r \in \mathcal{K}_{t_i}} \left( \kappa^2 + L^2 \mathbb{E} \left\| \mathbf{x}_r^{t_{i+1}-1} - \mathbf{x}^{t_{i+1}-1} \right\|^2 \right) \\ &= 2\kappa^2 + \frac{2L^2}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \mathbf{x}_r^{t_{i+1}-1} - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbf{x}_s^{t_{i+1}-1} \right\|^2 \\ &\leq 2\kappa^2 + \frac{2L^2}{K} \sum_{r \in \mathcal{K}_{t_i}} \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \mathbf{x}_r^{t_{i+1}-1} - \mathbf{x}_s^{t_{i+1}-1} \right\|^2 \\ &\stackrel{(b)}{\leq} 2\kappa^2 + \frac{2L^2}{K} \sum_{r \in \mathcal{K}_{t_i}} \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \left( 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \right) \\ &= 2\kappa^2 + 14L^2H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \stackrel{(c)}{\leq} 2\kappa^2 + \frac{7H}{32} \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \end{aligned} \quad (64)$$

In (a) we used the gradient dissimilarity bound from (6) to bound the first term and  $L$ -Lipschitz gradient property of  $F$  to bound the second term. For (b), note that we have already bounded  $\sum_{t=t_i}^{t_{i+1}-1} \mathbb{E} \|\mathbf{x}_r^t - \mathbf{x}_s^t\|^2 \leq 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right)$  in (22) in Lemma 2. Since each term in the summation is trivially bounded by the same quantity, which we used in (b) to bound  $\mathbb{E} \left\| \mathbf{x}_r^{t_{i+1}-1} - \mathbf{x}_s^{t_{i+1}-1} \right\|^2 \leq 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right)$ . In (c) we used  $\eta \leq \frac{1}{8HL}$ .  $\square$

**Claim** (Restating Claim 5). *If  $\eta \leq \frac{1}{8HL}$ , then with probability at least  $1 - \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$ , we have*

$$\mathbb{E} \left\| \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right\|^2 \leq 3\mathcal{R}^2 + \frac{8H^2\sigma^2}{b} + 30H^2\kappa^2,$$

where  $\mathcal{R}^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$  and  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'} \left(1 + \frac{4d}{3K}\right) + 28H^2\kappa^2$ .

*Proof.* Let  $\mathcal{S} \subseteq \mathcal{K}_{t_i}$  denote the subset of honest clients of size  $(1 - (\epsilon + \epsilon'))K$ , whose average accumulated gradient between time  $t_i$  and  $t_{i+1}$  that server approximates at time  $t_{i+1}$  in Theorem 3. Let the average accumulated gradient be denoted by  $\mathbf{g}_{\mathcal{S}, \text{accu}}^{t_i, t_{i+1}} = \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \mathbf{g}_{r, \text{accu}}^{t_i, t_{i+1}}$ , where  $\mathbf{g}_{r, \text{accu}}^{t_i, t_{i+1}} = \sum_{t=t_i}^{t_{i+1}-1} \mathbf{g}_r(\mathbf{x}_r^t)$ , and server approximates it by  $\hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}}$ . Note that  $\mathcal{S}$  exists with probability at least  $1 - \exp\left(-\frac{\epsilon'^2(1-\epsilon)K}{16}\right)$ . To make the notation less cluttered, for every  $r \in \mathcal{K}_{t_i}$ , define  $\nabla F_r^{t_i, t_{i+1}} := \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t)$ .

$$\begin{aligned} \mathbb{E} \left\| \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \nabla F_r^{t_i, t_{i+1}} \right\|^2 &\leq 3\mathbb{E} \left\| \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \mathbf{g}_{r, \text{accu}}^{t_i, t_{i+1}} \right\|^2 \\ &\quad + 3\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \mathbf{g}_{r, \text{accu}}^{t_i, t_{i+1}} - \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F_r^{t_i, t_{i+1}} \right\|^2 \\ &\quad + 3\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F_r^{t_i, t_{i+1}} - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F_s^{t_i, t_{i+1}} \right\|^2 \end{aligned} \quad (65)$$

Now we bound each term on the RHS of (65).

**Bounding the first term on the RHS of (65).** We can bound this using the second part of Theorem 3 as follows (note that given the first part of Theorem 3 is satisfied, the second part provides deterministic approximation guarantees, which implies that it also holds in expectation):

$$\mathbb{E} \left\| \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \mathbf{g}_{r, \text{accu}}^{t_i, t_{i+1}} \right\|^2 \leq \mathcal{R}^2, \quad (66)$$

where  $\mathcal{R}^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$  and  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'} \left(1 + \frac{4d}{3K}\right) + 28H^2\kappa^2$ .

**Bounding the second term on the RHS of (65).** We can bound this using the variance bound (5).

$$\begin{aligned} \mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} (\mathbf{g}_{r, \text{accu}}^{t_i, t_{i+1}} - \nabla F_r^{t_i, t_{i+1}}) \right\|^2 &= \mathbb{E} \left\| \sum_{t=t_i}^{t_{i+1}-1} \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \\ &\stackrel{(a)}{\leq} (t_{i+1} - t_i) \sum_{t=t_i}^{t_{i+1}-1} \mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{\leq} H \sum_{t=t_i}^{t_{i+1}-1} \frac{1}{|\mathcal{S}|^2} \mathbb{E} \left\| \sum_{r \in \mathcal{S}} (\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \\
&\stackrel{(c)}{=} H \sum_{t=t_i}^{t_{i+1}-1} \frac{1}{|\mathcal{S}|^2} \sum_{r \in \mathcal{S}} \mathbb{E} \|\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)\|^2 \\
&\stackrel{(d)}{\leq} H \sum_{t=t_i}^{t_{i+1}-1} \frac{1}{|\mathcal{S}|} \frac{\sigma^2}{b} \stackrel{(e)}{\leq} \frac{4H^2\sigma^2}{3bK}.
\end{aligned} \tag{67}$$

In (a) we used the Jensen's inequality. In (b) used  $|t_{i+1} - t_i| \leq H$ . In (c) we used (4) (which states that  $\mathbb{E}[\mathbf{g}_r(\mathbf{x})] = \nabla F_r(\mathbf{x})$  holds for every honest client  $r \in [R]$  and  $\mathbf{x} \in \mathbb{R}^d$ ) together with that the stochastic gradients at different clients are sampled independently, and then we used the fact that the variance of independent random variables is equal to the sum of the variances. Note that  $\text{Var}(\mathbf{g}_r(\mathbf{x}_r^t)) = \mathbb{E} \|\mathbf{g}_r(\mathbf{x}_r^t) - \nabla F_r(\mathbf{x}_r^t)\|^2$ . In (d) we used the variance bound (5). In (e) we used  $|\mathcal{S}| \geq (1 - (\epsilon + \epsilon'))K \geq \frac{3K}{4}$ , where the last inequality uses  $(\epsilon + \epsilon') \leq \frac{1}{4}$ .

**Bounding the third term on the RHS of (65).**

$$\begin{aligned}
\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F_r^{t_i, t_{i+1}} - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F_s^{t_i, t_{i+1}} \right\|^2 &= \mathbb{E} \left\| \sum_{t=t_i}^{t_{i+1}-1} \left( \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F_r(\mathbf{x}_r^t) - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F_s(\mathbf{x}_s^t) \right) \right\|^2 \\
&\stackrel{(a)}{\leq} H \sum_{t=t_i}^{t_{i+1}-1} \mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F_r(\mathbf{x}_r^t) - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F_s(\mathbf{x}_s^t) \right\|^2
\end{aligned} \tag{68}$$

In (a), first we used the Jensen's inequality and then substituted  $|t_{i+1} - t_i| \leq H$ . In order to bound (68), it suffices to bound  $\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F_r(\mathbf{x}_r^t) - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F_s(\mathbf{x}_s^t) \right\|^2$  for every  $t \in [t_i : t_{i+1} - 1]$ . We bound this in the following. Take an arbitrary  $t \in [t_i : t_{i+1} - 1]$ .

$$\begin{aligned}
&\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F_r(\mathbf{x}_r^t) - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F_s(\mathbf{x}_s^t) \right\|^2 \leq 3\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} (\nabla F_r(\mathbf{x}_r^t) - \nabla F(\mathbf{x}_r^t)) \right\|^2 \\
&\quad + 3\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F(\mathbf{x}_r^t) - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F(\mathbf{x}_s^t) \right\|^2 + 3\mathbb{E} \left\| \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}_s^t) - \nabla F_s(\mathbf{x}_s^t)) \right\|^2 \\
&\leq \frac{3}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \mathbb{E} \|\nabla F_r(\mathbf{x}_r^t) - \nabla F(\mathbf{x}_r^t)\|^2 + \frac{3}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbb{E} \|\nabla F(\mathbf{x}_s^t) - \nabla F_r(\mathbf{x}_r^t)\|^2 \\
&\quad + 3\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} (\nabla F(\mathbf{x}_r^t) - \nabla F(\mathbf{x}^t)) - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}_s^t) - \nabla F(\mathbf{x}^t)) \right\|^2 \\
&\leq 3\kappa^2 + 3\kappa^2 + 6\mathbb{E} \left\| \frac{1}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \nabla F(\mathbf{x}_r^t) - \nabla F(\mathbf{x}^t) \right\|^2 + 6\mathbb{E} \left\| \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}_s^t) - \nabla F(\mathbf{x}^t)) \right\|^2 \\
&\leq 6\kappa^2 + \frac{6}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} \mathbb{E} \|\nabla F(\mathbf{x}_r^t) - \nabla F(\mathbf{x}^t)\|^2 + \frac{6}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbb{E} \|\nabla F(\mathbf{x}_s^t) - \nabla F(\mathbf{x}^t)\|^2 \\
&\leq 6\kappa^2 + \frac{6}{|\mathcal{S}|} \sum_{r \in \mathcal{S}} L^2 \mathbb{E} \|\mathbf{x}_r^t - \mathbf{x}^t\|^2 + \frac{6}{K} \sum_{s \in \mathcal{K}_{t_i}} L^2 \mathbb{E} \|\mathbf{x}_s^t - \mathbf{x}^t\|^2
\end{aligned}$$

$$\begin{aligned}
&= 6\kappa^2 + \frac{6L^2}{|S|} \sum_{r \in S} \mathbb{E} \left\| \mathbf{x}_r^t - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbf{x}_s^t \right\|^2 + \frac{6L^2}{K} \sum_{r \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \mathbf{x}_r^t - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbf{x}_s^t \right\|^2 \\
&\leq 6\kappa^2 + \frac{6L^2}{|S|} \sum_{r \in S} \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \mathbf{x}_r^t - \mathbf{x}_s^t \right\|^2 + \frac{6L^2}{K} \sum_{r \in \mathcal{K}_{t_i}} \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \mathbf{x}_r^t - \mathbf{x}_s^t \right\|^2
\end{aligned}$$

Substituting this back in (68) gives:

$$\begin{aligned}
&\mathbb{E} \left\| \frac{1}{|S|} \sum_{r \in S} \nabla F_r^{t_i, t_{i+1}} - \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \nabla F_s^{t_i, t_{i+1}} \right\|^2 \leq H \sum_{t=t_i}^{t_{i+1}-1} 6\kappa^2 \\
&\quad + H \sum_{t=t_i}^{t_{i+1}-1} \left( \frac{6L^2}{|S|} \sum_{r \in S} \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \mathbf{x}_r^t - \mathbf{x}_s^t \right\|^2 + \frac{6L^2}{K} \sum_{r \in \mathcal{K}_{t_i}} \frac{1}{K} \sum_{s \in \mathcal{K}_{t_i}} \mathbb{E} \left\| \mathbf{x}_r^t - \mathbf{x}_s^t \right\|^2 \right) \\
&\stackrel{(a)}{\leq} 6H^2\kappa^2 + 6HL^2 \left( 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \right) + 6HL^2 \left( 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \right) \\
&= 6H^2\kappa^2 + 84L^2H^4\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right) \\
&\stackrel{(b)}{\leq} 10H^2\kappa^2 + \frac{21H^2\sigma^2}{16b}. \tag{69}
\end{aligned}$$

In (a) we used  $t_{i+1} - t_i \leq H$  and the bound  $\sum_{t=t_i}^{t_{i+1}-1} \mathbb{E} \left\| \mathbf{x}_r^t - \mathbf{x}_s^t \right\|^2 \leq 7H^3\eta^2 \left( \frac{\sigma^2}{b} + 3\kappa^2 \right)$ , which holds when  $\eta \leq \frac{1}{8HL}$ ; we have already shown this in (22) in Lemma 2. In (b) we used  $\eta \leq \frac{1}{8HL}$ .

Substituting the bounds from (66), (67), (69) into (65) gives

$$\begin{aligned}
\mathbb{E} \left\| \hat{\mathbf{g}}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \nabla F_r^{t_i, t_{i+1}} \right\|^2 &\leq 3\Upsilon^2 + \frac{4H^2\sigma^2}{bK} + 3 \left( 10H^2\kappa^2 + \frac{21H^2\sigma^2}{16b} \right) \\
&\leq 3\Upsilon^2 + \frac{4H^2\sigma^2}{bK} + 30H^2\kappa^2 + \frac{4H^2\sigma^2}{b} \\
&= 3\Upsilon^2 + \frac{8H^2\sigma^2}{b} + 30H^2\kappa^2,
\end{aligned}$$

where  $\Upsilon^2 = \mathcal{O}(\sigma_0^2(\epsilon + \epsilon'))$  and  $\sigma_0^2 = \frac{25H^2\sigma^2}{b\epsilon'} \left( 1 + \frac{4d}{3K} \right) + 28H^2\kappa^2$ .

This completes the proof of Claim 5.  $\square$

**Fact 1.** Let  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  be an  $L$ -smooth function with a global minimizer  $\mathbf{x}^*$ . Then, for every  $\mathbf{x} \in \mathbb{R}^d$ , we have

$$\|\nabla F(\mathbf{x})\|^2 \leq 2L(F(\mathbf{x}) - F(\mathbf{x}^*)).$$

*Proof.* By definition of  $L$ -smoothness, we have  $F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2$  holds for every  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ . Fix an arbitrary  $\mathbf{x} \in \mathbb{R}^d$  and take infimum over  $\mathbf{y}$  on both sides:

$$\begin{aligned}
\inf_{\mathbf{y}} F(\mathbf{y}) &\leq \inf_{\mathbf{y}} \left( F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2 \right) \\
&\stackrel{(a)}{=} \inf_{\mathbf{v}: \|\mathbf{v}\|=1} \inf_t \left( F(\mathbf{x}) + t \langle \nabla F(\mathbf{x}), \mathbf{v} \rangle + \frac{Lt^2}{2} \right) \\
&\stackrel{(b)}{=} \inf_{\mathbf{v}: \|\mathbf{v}\|=1} \left( F(\mathbf{x}) - \frac{1}{2L} \langle \nabla F(\mathbf{x}), \mathbf{v} \rangle^2 \right)
\end{aligned}$$

$$\stackrel{(c)}{=} \left( F(\mathbf{x}) - \frac{1}{2L} \|\nabla F(\mathbf{x})\|^2 \right)$$

The value of  $t$  that minimizes the RHS of (a) is  $t = -\frac{1}{L} \langle \nabla F(\mathbf{x}), \mathbf{v} \rangle$ , this implies (b); (c) follows from the Cauchy-Schwarz inequality:  $\langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\| \|\mathbf{v}\|$ , where equality is achieved whenever  $\mathbf{u} = \mathbf{v}$ . Now, substituting  $\inf_{\mathbf{y}} F(\mathbf{y}) = F(\mathbf{x}^*)$  yields the result.  $\square$

## C Omitted Details from Section 6

In this section, we prove [Theorem 2](#). This can be proved along the lines of the proof of [Theorem 1](#). Here we only write what changes in those proofs. We prove the strongly-convex and non-convex parts of [Theorem 2](#) in [Appendix C.1](#) and [Appendix C.2](#), respectively.

### C.1 Strongly-convex

Let  $\mathcal{K}_t \subseteq [R]$  denote the subset of clients of size  $|\mathcal{K}_t| = K$  that are active at the  $t$ 'th iteration. For any  $t \in [t_i : t_{i+1} - 1]$ , let  $\mathbf{x}^t = \frac{1}{K} \sum_{k \in \mathcal{K}_{t_i}} \mathbf{x}_k^t$  denote the average of the local parameters of clients in the sampling set  $\mathcal{K}_{t_i}$ .

Following the proof of the strongly-convex part of [Theorem 1](#) given in [Section 4](#) until (33) gives

$$\begin{aligned} \|\mathbf{x}^{t_{i+1}} - \mathbf{x}^*\|^2 &\leq \left(1 + \frac{\mu\eta}{2}\right) \|\mathbf{x}^{t_{i+1}-1} - \eta \nabla F(\mathbf{x}^{t_{i+1}-1}) - \mathbf{x}^*\|^2 \\ &\quad + 2\eta \left(\eta + \frac{2}{\mu}\right) \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) \right\|^2 \\ &\quad + 2\eta \left(\eta + \frac{2}{\mu}\right) \left\| \hat{F}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right\|^2 \end{aligned} \quad (70)$$

We have already bounded the first term in [Claim 3](#) (on page 16) by

$$\|\mathbf{x}^{t_{i+1}} - \eta \nabla F(\mathbf{x}^{t_{i+1}-1}) - \mathbf{x}^*\|^2 \leq (1 - \eta\mu) \|\mathbf{x}^{t_{i+1}-1} - \mathbf{x}^*\|^2. \quad (71)$$

In order to bound the second term, we follow the proof of [Claim 4](#) exactly until (64), and then to bound  $\left\| \mathbf{x}_r^{t_{i+1}-1} - \mathbf{x}_s^{t_{i+1}-1} \right\|^2$  for every  $r, s \in \mathcal{K}_{t_i}$ , we use the bound from (59) in [Lemma 4](#) and use  $\eta \leq \frac{1}{5HL}$ , which gives

$$\left\| \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F_r(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) \right\|^2 \leq 3H\kappa^2. \quad (72)$$

To bound the third term in the RHS of (70), we can simplify the proof of [Claim 5](#): Firstly, note that with full-batch gradients, the variance  $\sigma^2$  becomes zero; secondly, as shown in [Theorem 4](#), the robust estimation of accumulated gradients holds with probability 1. Following the proof of [Claim 5](#) with these changes and using  $\eta \leq \frac{1}{5HL}$ , we get

$$\left\| \hat{F}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t) \right\|^2 \leq 2\gamma_{\text{GD}}^2 + 20H^2\kappa^2, \quad (73)$$

where  $\mathcal{R}_{\text{GD}} = \mathcal{O}(H\kappa\sqrt{\epsilon})$ . Substituting all these bounds from (71)-(73) into (70) and simplifying further using  $(1 + \frac{\mu\eta}{2})(1 - \mu\eta) \leq (1 - \frac{\mu\eta}{2})$  and  $(\eta + \frac{2}{\mu}) \leq \frac{3}{\mu}$  gives

$$\|\mathbf{x}^{t_{i+1}} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\mu\eta}{2}\right) \|\mathbf{x}^{t_{i+1}-1} - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} (2\mathcal{R}_{\text{GD}}^2 + 23H^2\kappa^2) \quad (74)$$

Note that (74) gives a recurrence at the synchronization indices. Now we give a recurrence at non-synchronization indices. Take an arbitrary  $t \in [T]$  and let  $t_i \in \mathcal{I}_T$  be such that  $t \in [t_i : t_{i+1} - 1]$ ; when  $H \geq 2$ , such  $t$ 's exist. Following the steps that we used to arrive at (39), we get the following (note that the last term on the RHS of (39) is zero, as  $\mathbf{g}_r(\mathbf{x}_r^t) = \nabla F_r(\mathbf{x}_r^t)$  holds for every  $r \in [R]$  and  $t \in [T]$ ; this will also save us the factor of 2 in the previous term as we don't have to use the Jensen's inequality to get to (39)):

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \leq \left(1 + \frac{\mu\eta}{2}\right) \|\mathbf{x}^t - \mathbf{x}^* - \eta \nabla F(\mathbf{x}^t)\|^2 + \eta \left(\eta + \frac{2}{\mu}\right) \left\| \frac{1}{K} \sum_{r \in \mathcal{K}_t} (\nabla F(\mathbf{x}^t) - \nabla F_r(\mathbf{x}_r^t)) \right\|^2 \quad (75)$$

Substituting the bounds from (71) and (72) into (75) and simplifying the coefficients as above, we get

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\mu\eta}{2}\right) \|\mathbf{x}^t - \mathbf{x}^*\|^2 + \frac{3\eta}{\mu} (3H\kappa^2) \quad (76)$$

Now we have a recurrence at the synchronization indices given in (74) and at non-synchronization indices given in (76). Let  $\alpha = (1 - \frac{\mu\eta}{2})$ ,  $\beta_1 = (2\mathcal{R}_{\text{GD}}^2 + 23H^2\kappa^2)$ , and  $\beta_2 = (\frac{3}{2}H\kappa^2)$ . Following the same steps that we used to arrive at (42) gives:

$$\|\mathbf{x}^T - \mathbf{x}^*\|^2 \leq \alpha^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left( \frac{1}{1 - \alpha} \beta_2 + \frac{1}{1 - \alpha^H} \beta_1 \right) \quad (77)$$

Since  $\alpha = (1 - \frac{\mu\eta}{2})$ , we have  $\alpha^H = (1 - \frac{\mu\eta}{2})^H \stackrel{(a)}{\leq} \exp(-\frac{\mu\eta H}{2}) \stackrel{(b)}{\leq} 1 - \frac{\mu\eta H}{2} + \left(\frac{\mu\eta H}{2}\right)^2 \stackrel{(c)}{\leq} 1 - \frac{\mu\eta H}{2} + \frac{1}{10} \frac{\mu\eta H}{2} = 1 - \frac{9}{10} \frac{\mu\eta H}{2}$ . In (a) we used the inequality  $(1 - \frac{1}{x})^x \leq \frac{1}{e}$  which holds for any  $x > 0$ ; in (b) we used  $\exp(-x) \leq 1 - x + x^2$  which holds for any  $x \geq 0$ ; in (c) we used  $\eta \leq \frac{1}{5HL}$  and  $\mu \leq L$ , which imply  $\frac{\mu\eta H}{2} \leq \frac{1}{10}$ . Substituting these in (77) gives

$$\begin{aligned} \|\mathbf{x}^T - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta}{2}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6\eta}{\mu} \left( \frac{2}{\mu\eta} \beta_2 + \frac{20}{9\mu\eta H} \beta_1 \right) \\ &\leq \left(1 - \frac{\mu\eta}{2}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{6 \times 20}{9\mu^2} \left( \frac{9}{10} \beta_2 + \frac{1}{H} \beta_1 \right) \\ &\leq \left(1 - \frac{\mu\eta}{2}\right)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{14}{\mu^2} \left( \frac{2\mathcal{R}_{\text{GD}}^2}{H} + 25H\kappa^2 \right), \end{aligned} \quad (78)$$

where  $\mathcal{R}_{\text{GD}} = \mathcal{O}(H\kappa\sqrt{\epsilon})$ . Substituting the value of  $\eta = \frac{1}{5HL}$  yields the convergence rate (9) in the strongly-convex part of Theorem 2. Note that (78) holds with probability 1.

## C.2 Non-convex

Following the proof of the non-convex part of Theorem 1 given in Section 5 until (47) and using  $\eta \leq \frac{1}{5HL}$  gives:

$$F(\mathbf{x}^{t_{i+1}}) \leq F(\mathbf{x}^{t_{i+1}-1}) - \frac{\eta}{2} \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \frac{6\eta}{5} \|C\|^2, \quad (79)$$

where  $C = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^{t_{i+1}-1}) - \nabla F_r(\mathbf{x}_r^{t_{i+1}-1})) - (\hat{F}_{\text{accu}}^{t_i, t_{i+1}} - \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} \sum_{t=t_i}^{t_{i+1}-1} \nabla F_r(\mathbf{x}_r^t))$ .

Using the bounds from (72) and (73), together with the Jensen's inequality, we can bound  $\|C\|^2$  as follows:

$$\|C\|^2 \leq 2(3H\kappa^2) + 2(2\Upsilon_{\text{GD}}^2 + 20H^2\kappa^2) \leq 2(2\Upsilon_{\text{GD}}^2 + 23H^2\kappa^2) \quad (80)$$

Substituting the bound from (80) into (79) gives:

$$F(\mathbf{x}^{t_{i+1}}) \leq F(\mathbf{x}^{t_{i+1}-1}) - \frac{\eta}{2} \|\nabla F(\mathbf{x}^{t_{i+1}-1})\|^2 + \frac{12\eta}{5} (2\Upsilon_{\text{GD}}^2 + 23H^2\kappa^2), \quad (81)$$

where  $\Upsilon_{\text{GD}} = \mathcal{O}(H\kappa\sqrt{\epsilon})$ .

Note that above recurrence in (81) holds only at the synchronization indices. Now we give a recurrence at non-synchronization indices.

We have done a similar calculations in the non-convex part of [Theorem 1](#) in [Section 5](#).

Take an arbitrary  $t \in [T]$  and let  $t_i \in \mathcal{I}_T$  be such that  $t \in [t_i : t_{i+1} - 1]$ ; when  $H \geq 2$ , such  $t$ 's exist. Following the same steps until (50) and using  $\eta \leq \frac{1}{5HL}$  gives:

$$F(\mathbf{x}^{t+1}) \leq F(\mathbf{x}^t) - \frac{\eta}{2} \|\nabla F(\mathbf{x}^t)\|^2 + \frac{6\eta}{5} \|D\|^2, \quad (82)$$

where  $D = \frac{1}{K} \sum_{r \in \mathcal{K}_{t_i}} (\nabla F(\mathbf{x}^t) - \nabla F_r(\mathbf{x}_r^t))$ .

Using the bound from (72), we have  $\|D\|^2 \leq 3H\kappa^2$ . Substituting this in (82) gives:

$$F(\mathbf{x}^{t+1}) \leq F(\mathbf{x}^t) - \frac{\eta}{2} \|\nabla F(\mathbf{x}^t)\|^2 + \frac{6\eta}{5} (3H\kappa^2) \quad (83)$$

Now we have a recurrence at the synchronization indices given in (81) and at non-synchronization indices given in (83). Adding (81) and (83) from  $t = 0$  to  $T$  (use (81) for the synchronization indices and (83) for the rest of the indices) gives:

$$\sum_{t=0}^T F(\mathbf{x}^{t+1}) \leq \sum_{t=0}^T F(\mathbf{x}^t) - \frac{\eta}{2} \sum_{t=0}^T \|\nabla F(\mathbf{x}^t)\|^2 + \frac{12\eta}{5} \left[ \frac{T}{H} (2\Upsilon_{\text{GD}}^2 + 23H^2\kappa^2) + \left(T - \frac{T}{H}\right) \left(\frac{3}{2}H\kappa^2\right) \right] \quad (84)$$

After rearranging and simplifying the last constant terms, we get:

$$\frac{1}{T} \sum_{t=0}^T \|\nabla F(\mathbf{x}^t)\|^2 \leq \frac{2}{\eta T} [F(\mathbf{x}^0) - F(\mathbf{x}^{T+1})] + \frac{24}{5} \left( \frac{2\Upsilon_{\text{GD}}^2}{H} + 25H\kappa^2 \right) \quad (85)$$

Note that the last term in (85) is a constant. So, it would be best to take the step-size  $\eta$  to be as large as possible such that it satisfies  $\eta \leq \frac{1}{5HL}$ . We take  $\eta = \frac{1}{5HL}$ . Substituting this in (85) and using  $F(\mathbf{x}^{T+1}) \geq F(\mathbf{x}^*)$  gives

$$\frac{1}{T} \sum_{t=0}^T \|\nabla F(\mathbf{x}^t)\|^2 \leq \frac{10HL}{T} [F(\mathbf{x}^0) - F(\mathbf{x}^*)] + \frac{24}{5} \left( \frac{2\Upsilon_{\text{GD}}^2}{H} + 25H\kappa^2 \right), \quad (86)$$

where  $\Upsilon_{\text{GD}} = \mathcal{O}(H\kappa\sqrt{\epsilon})$ . This yields the convergence rate (10) in the non-convex part of [Theorem 2](#). Note that (86) holds with probability 1.

This concludes the proof of [Theorem 2](#).