Journal of Machine Learning Research 27 (2026) 1-44 Submitted 10/24; Revised 3/26; Published 4/26

On the Relevance of Byzantine Robust Optimization Against Data Poisoning

Sadegh Farhadkhani SADEGH.FARHADKHANI@EPFL.CH
School of Computer and Communication Sciences
EPFL
Lausanne, Switzerland

Rachid Guerraoui RACHID.GUERRAOUI@EPFL.CH
School of Computer and Communication Sciences
EPFL
Lausanne, Switzerland

Nirupam Gupta NIGU@DI.KU.DK
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark

Rafael Pinot PINOT@LPSM.PARIS
Sorbonne Université and Université Paris Cité, CNRS,
Laboratoire de Probabilités, Statistique et Modélisation
Paris, France

Editor: Sebastian Stich

Abstract

The success of machine learning (ML) has been intimately linked with the availability of large amounts of data, typically collected from heterogeneous sources and processed on vast networks of computing devices (also called *workers*). Beyond accuracy, the use of ML in critical domains such as healthcare and autonomous driving calls for robustness against *data poisoning* and some *faulty workers*. The problem of *Byzantine ML* formalizes these robustness issues by considering a distributed ML environment in which workers (storing a portion of the global dataset) can deviate arbitrarily from the prescribed algorithm. Although the problem has attracted a lot of attention from a theoretical point of view, its practical importance for addressing realistic faults (where the behavior of any worker is locally constrained) remains unclear. It has been argued that the seemingly weaker threat model where only workers' local datasets get poisoned is more reasonable. We prove that, while tolerating a wider range of faulty behaviors, Byzantine ML yields solutions that are, in a precise sense, optimal even under the weaker data poisoning threat model. Then, we study a generic data poisoning model wherein some workers have *fully-poisonous local data*, i.e., their datasets are entirely corruptible, and the remainders have *partially-poisonous local data*, i.e., only a fraction of their local datasets is corruptible. We prove that Byzantine-robust schemes yield optimal solutions against both these forms of data poisoning, and that the former is more harmful when workers have *heterogeneous* local data.

Keywords: Federated learning, Byzantine failure, Data Poisoning

1. Introduction

Learning a model using several machines over their collective data is appealing. The motivation behind this *distributed* machine learning (ML) scheme (a.k.a. *federated learning* (Kairouz et al., 2021)) is usually efficiency. Another motivation is data protection, as each machine retains control over its local data. The distributed ML problem can be precisely stated as follows in a standard *server-based system* comprising n machines (referred as *workers*), represented by set $[n] \coloneqq \{1, \ldots, n\}$, and a server. Each worker i has access to a common data space $\mathcal{X}$ through a local distribution $\mathcal{D}^{(i)}$. A model parameterized by $\theta \in \mathbb{R}^d$ incurs a loss for each data point $x \in \mathcal{X}$ measured by a real-valued *loss function* $q : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$. Then, for each worker $i \in [n]$, the *local loss function* is given by

$$Q^{(i)}(\theta) \coloneqq \mathbb{E}_{x \sim \mathcal{D}^{(i)}} \left[q(\theta, x) \right]. \tag{1}$$

The server aims to compute a model parameter $\theta^* \in \mathbb{R}^d$ minimizing the *global loss function*

$$Q(\theta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} Q^{(i)}(\theta). \tag{2}$$

We assume that the gradient of the loss function $q(\theta, x)$ with respect to θ, denoted by $\nabla q(\theta, x)$, exists and is continuous at all $\theta \in \mathbb{R}^d$ and $x \in \mathcal{X}$, which is standard in machine learning (Bottou et al., 2018).

1.1 Distributed ML with D(S)GD and Associated Threats

Minimizing the global average loss is typically achieved using a first-order distributed method such as the celebrated Distributed Gradient Descent (or DGD) and its *stochastic* variant DSGD (Konečnỳ et al., 2016).[1] At each iteration $t \geq 0$, the server maintains a model θ_t, which is broadcast to all the workers. Then, each worker i sends back to the server an update vector that is either their *local gradient* $\nabla Q^{(i)}(\theta_t)$ in the case of DGD or an unbiased stochastic estimate $g_t^{(i)}$ of their local gradient in the case of DSGD. Finally, the server updates the current model θ_t using the *average* of the local updates sent by the workers. When all the workers are *honest*, i.e, correctly follow the instructions of the server, the above iterative procedure provably converges to a parameter θ^* that is either a minimum or a stationary point of the global loss function depending on whether the function is convex or non-convex, respectively.

Threats to Distributed ML. DSGD (or DGD) is however extremely vulnerable to misbehaving workers that can deviate from the instructions given by the server. Such misbehavior could result from either inadvertent software/hardware bugs, incorrect data, or malicious players controlling part of the system. Typically, misbehaving workers are modeled by considering an *adversary* that corrupts a fraction of the workers, whose identity is a priori unknown (Guerraoui et al., 2023). The corruptions induced by the adversary can be characterized by two threat models: *Byzantine failure* (Feng et al., 2015; Su and Vaidya, 2016; Blanchard et al., 2017) and *data poisoning* (Shejwalkar and Houmansadr, 2021; Farhadkhani et al., 2022b).

1. For more details on these distributed methods, refer the book by Bertsekas and Tsitsiklis (2015).

1. **Byzantine failure.** In this particular threat model, we assume that a corrupted worker can deviate arbitrarily from its prescribed algorithm (Lamport et al., 1982). In the context of DSGD, a Byzantine worker can send (arbitrary) malicious vectors for its local gradients to the server (Baruch et al., 2019; Xie et al., 2019).

2. **Data poisoning.** In this particular threat model, we assume that a corrupted worker follows the prescribed algorithm correctly but its local dataset can be poisoned (Mahloujifar et al., 2019). In the context of DSGD, while the gradients sent by a worker i need not be arbitrary, they can correspond to another data distribution $\widetilde{\mathcal{D}}^{(i)}$ that differs from the true data distribution $\mathcal{D}^{(i)}$ (but is fixed for all iterations). The data poisoning threat can be further classified into two sub-cases: *fully-poisonous local data* and *partially-poisonous local data*. Suppose that worker i is corrupted by an adversary. In the case of fully-poisonous local data, the entire local dataset of worker i can be poisoned, i.e., $\widetilde{\mathcal{D}}^{(i)}$ is truly arbitrary. In the case of partially-poisonous local data, only a fraction (of unknown identity) of worker i's local distribution is corrupted.

Note that the Byzantine failure threat model, subsumes the data poisoning threat model. Nevertheless, the latter has received more attention in the past mainly due to its relevance in the conventional *centralized* ML paradigm (Charikar et al., 2017; Diakonikolas et al., 2019; Prasad et al., 2020). Although the defenses proposed for data poisoning can be extended to the Byzantine threat model in distributed ML (Chen et al., 2017; Yin et al., 2018), they rely upon *data homogeneity*, i.e., the honest workers are assumed to have identical local data distributions (Diakonikolas and Kane, 2022). In general distributed ML however the workers have *heterogeneous data*, i.e., their local data distributions are distinct (El Mhamdi et al., 2021; Data and Diggavi, 2021; Karimireddy et al., 2022; Farhadkhani et al., 2023).

1.2 Byzantine failure vs data poisoning

Given the heterogeneous nature of workers' data in distributed ML, it seems reasonable to seek novel solutions to data poisoning. But what about Byzantine failures? One could argue that a truly arbitrary behavior is largely fictitious and unlikely to be realized in practice (Shejwalkar et al., 2022). Indeed, each worker of a distributed system is typically restricted to very limited local information and cannot possibly be omniscient, unlike what is assumed in the Byzantine threat model (Karimireddy et al., 2022; Farhadkhani et al., 2022a). Somehow, the cost of defending against Byzantine workers might not be justifiable compared to the cost for defending against data poisoning. But what is that cost difference anyway? The motivation of this work is to address that question, and equivalently:

Is defending against Byzantine failure an overkill with respect to data poisoning?

We answer this question negatively in the context of a large class of ML problems. We prove (perhaps surprisingly) that, although the Byzantine failure threat model is strictly stronger, the best learning guarantees that a first-order distributed algorithm, such as DSGD, can achieve under this threat model are optimal even in the weaker data poisoning threat model. Furthermore, we precisely characterize the impact on the learning due to both fully-poisonous and partially-poisonous local data. We show that in real-world applications when workers' datasets are heterogeneous, (Kairouz et al., 2021), fully-poisonous local data is a stronger adversarial setting. Our contributions are summarized as follows.

1.3 Main results

Solution to Byzantine failure is tight with respect to data poisoning. We consider the class of ML problems that can be solved by optimizing L-Lipschitz smooth loss functions satisfying the μ-PL inequality, where the local gradients (of honest workers) have bounded covariance trace of σ^2. These conditions are satisfied in many cases (Bottou et al., 2018). We further assume that the global *gradient dissimilarity* that characterizes data heterogeneity is bounded by ζ^2, which is essential to tackling misbehaving workers of either type (Karimireddy et al., 2022; Allouah et al., 2023a). We assume that the total number of fully corrupted workers is bounded by f. Note that the case of $f \geq n/2$ is trivial as the learning error can be arbitrarily large (Lemma 1 in (Liu et al., 2021)): we thus assume $f < n/2$ in all our results.

1. **Lower bound under data poisoning.** We first characterize the suboptimality gap (or error) of a stochastic first-order distributed algorithm under data poisoning. Specifically, we show that with f workers with corrupted data the error is in $\Omega\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu}\right)$. Moreover, the convergence rate (a.k.a. iteration complexity) to realize an ε-approximation of this error is in

$$\Omega\left(\frac{1+f}{n} \cdot \frac{\sigma^2}{\mu\varepsilon} + \frac{L}{\mu} \cdot \log \frac{Q_0}{\varepsilon}\right) \ , \tag{3}$$

   where Q_0 is the initial error of the algorithm. These lower bounds characterize how *good* and *fast* we can learn using n workers when f of the workers suffer from local data poisoning.

2. **Matching upper bound under Byzantine failure.** We then consider the Byzantine-robust adaptation of DSGD, incorporating distributed *Polyak's momentum* and *coordinate wise trimmed mean* (Farhadkhani et al., 2022a). We show that, despite the presence of f Byzantine corrupted workers, this algorithm achieves an error in $\mathcal{O}\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu} + \varepsilon\right)$ with a convergence rate of

$$\mathcal{O}\left(\frac{1+f}{n} \cdot \frac{K\sigma^2}{\mu\varepsilon} + \frac{L}{\mu} \cdot \log \frac{Q_0}{\varepsilon}\right) \ , \tag{4}$$

   where $K \coloneqq \frac{L}{\mu}$ is the *condition number* of the average loss function for the honest workers. Hence, when $K \in \mathcal{O}(1)$, we get a matching upper bound to the lower bound in the data poisoning threat model (which automatically also applies to the Byzantine failure threat model). To the best of our knowledge, this is the first tight analysis (up to multiplicative factor K) of Byzantine robustness in terms the convergence rate and the asymptotic error of a stochastic first-order method (see also Section 1.6).

Partially-poisonous vs fully-poisonous local data. We then consider a scenario where in addition to having f out of n workers with fully-poisonous local datasets, each worker can have partially-poisonous local data. Specifically, we assume that each worker i has b number of corruptible data points out of m total data points. Note that in this

particular case, for each worker i the distribution $\mathcal{D}^{(i)}$ is given by the uniform distribution over the $m - b$ incorruptible local data points. We prove that the optimization error is in

$$\Theta\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu} + \frac{b}{m} \cdot \frac{\sigma^2}{\mu}\right) \quad .$$

We show that the above error, which is optimal in general, can be achieved using a *Byzantine-robust* first-order method with an exponential convergence rate (i.e., logarithmic iteration complexity). Hence, demonstrating the tightness of Byzantine-robust schemes even against data poisoning at the local level. Moreover, as the error resulting from partially-poisonous local data is independent of the heterogeneity factor ζ, this result also shows that in practical distributed ML applications, where dataset heterogeneity among workers is often significant (Karimireddy et al., 2020), fully-poisonous local data alone (i.e., $\frac{f}{n} = \delta > 0$, and $b = 0$) is a stronger adversarial setting than partially-poisonous local data alone (i.e., $\frac{b}{m} = \delta > 0$, and $f = 0$), when considering the same fraction δ of corrupted data points in the system. In short, this means that it is stronger to corrupt a few workers completely than all partially.

1.4 Key elements of our proof

Our proof for the lower bound in the homogeneous case, i.e., the first term in (3), involves an extension of Huber's general contamination model (Huber, 1964; Diakonikolas and Kane, 2022). Specifically, we consider a special distributed ML problem of mean estimation where each worker samples data points from a common distribution $\mathcal{D}$, and the goal for the server is to compute the true mean of $\mathcal{D}$ in the case when f out of n workers can sample data points from arbitrary distributions. We show that solving this problem using a robust implementation of DSGD with T iterations reduces to robust mean estimation using n *batches* of T i.i.d. data points from $\mathcal{D}$ with f batches being arbitrarily corrupted. To derive the lower bound due to heterogeneity, we consider the mean estimation problems with workers sampling data points from two distinct Dirac delta distributions with means $\frac{\zeta}{\mu}\sqrt{\frac{n}{f}}$ apart. We conclude the result by considering two indistinguishable executions, exploiting the anonymity of corrupted workers. Details can be found in Section 3.

The more challenging part of our analysis lies in proving a tight upper bound in the Byzantine setting. To prove the matching upper bound, we consider a Byzantine-robust adaptation of DSGD, originally proposed by Farhadkhani et al. (2022a), that uses Polyak's momentum operation at workers' end and replaces the averaging at the server by coordinate-wise trimmed mean. Although this algorithm has been shown to guarantee a tight asymptotic error under the Byzantine threat model (Allouah et al., 2023a), its convergence rate remained loose for the specific class of PL functions that we consider. To overcome the shortcoming, we consider a scheduled diminishing step sizes, generalizing the results on the tightness of SGD (Stich, 2019; Khaled and Richtárik, 2023).

The caveat of varying step sizes however is that it leads to *dynamic* momentum coefficient, if we are to obtain a tight convergence rate in the presence of Byzantine failures. This renders the existing proof techniques for analyzing the convergence of this particular class of algorithms inapplicable (see the works of Karimireddy et al. (2021); Farhadkhani et al. (2022a); Allouah et al. (2023a)), mainly because we can no longer obtain a uniform bound on

the *momentum drifts*. To remedy this, we design a novel *time-invariant* Lyapunov function that includes an additive term of appropriately scaled momentum drift (see Section 5).

1.5 Conjecture on the tightness of the upper bound

Our upper bound (in (4)) holds for any smooth PL loss function. Our lower bound (in (3)) however is derived by considering a quadratic loss function that is strongly convex[2] with condition number $K = 1$, which renders our overall analysis loose in terms of K. We however conjecture our upper bound to be tight (even in the condition number) for the class of loss functions we consider. Indeed, if we assume $f = 0$, our upper bound matches the best known result for the class of smooth PL loss functions (Karimi et al., 2016). Moreover, while we are not aware of any lower bound in stochastic optimization that is specific to the PL functions, it was recently shown by Yue et al. (2022) that, in the non-stochastic case, the dependence of the lower bound on the condition number is indeed different for strongly convex and PL functions. Accordingly, we believe that obtaining a tight result in terms of the condition number K would involve demonstrating that, for general PL loss functions, the convergence rate of a stochastic first-order method is in $\Omega\left(\frac{1+f}{n} \cdot \frac{K\sigma^2}{\mu\varepsilon} + \frac{L}{\mu} \cdot \log \frac{Q_0}{\varepsilon}\right)$.

1.6 Other related work

On the convergence rates. As we mentioned in Section 1.4, most existing works (with the exception of (Gorbunov et al., 2023)) that provide convergence results for robust DSGD focus on achieving tight asymptotic error but offer loose convergence rates. In fact, many of these results rely on constant step sizes and momentum coefficients, which yield a *uniform bound* on the drift between the local momentums (e.g., Lemma 1 of Farhadkhani et al. (2022a), Lemma 8 of Karimireddy et al. (2022), and Lemma 6 of Allouah et al. (2023a)). However, obtaining a tight convergence rate for PL functions calls for diminishing step sizes (Stich, 2019; Khaled and Richtárik, 2023). As the momentum coefficients are coupled with the step sizes, for Byzantine robustness, using diminishing step sizes leads to dynamic momentum coefficients. Accordingly, we can only obtain a recursive bound on the momentum drift, which makes the analysis more intricate. While Allouah et al. (2023b) address this challenge using a *time-variant* Lyapunov function (see Appendix D.2.1 in the work of Allouah et al. (2023b)), the resulting convergence analysis is also loose as it features a *sublinear* convergence rate even when honest workers compute exact local gradients, i.e., $\sigma = 0$. To the best of our knowledge, the current state-of-the-art result on the convergence rates of Byzantine resilient stochastic first-order algorithms is presented by Gorbunov et al. (2023). However, their approach alternates between full-gradient and stochastic gradient steps, where the full gradient is computed with probability p, and the stochastic gradient with probability $1 - p$. The parameter p influences both the breakdown point (in $\mathcal{O}(p)$) and the asymptotic error (in $\mathcal{O}(1/p)$) of the algorithm. As p approaches zero, the results become progressively less sharp. Thus, it remains unclear how this analysis applies to standard first-order methods like robust DSGD, as a very small p would be required for the algorithm to be practical.

On the comparison between Byzantine resilience and poisoning. Another work that provides a comparison between the Byzantine failure and the data poisoning

2. Strong convex functions constitute a subclass of PL functions.

threats in distributed ML is the work of Alistarh et al. (2018). However, there are several notable distinctions. First, Alistarh et al. (2018) consider the i.i.d. case where all honest workers sample data points from the same distribution. Second, the lower and upper bounds of Alistarh et al. (2018) are not obtained under exactly the same assumptions. The lower bound (Theorem 5.5 of Alistarh et al. (2018)) is derived by considering a Gaussian data distribution, whereas the upper bound relies on the assumption that the distribution of the stochastic gradients has a uniformly bounded support, which is not the case for a Gaussian distribution. We remark that the bounded-support assumption considerably weakens the Byzantine failure threat model as it ensures that the pairwise distances between honest local gradients are bounded. Until now, it remained unclear whether a tight upper bound could be obtained without restricting the Byzantine adversary, and under standard learning assumptions.

1.7 Paper organization

Section 2 presents the problem statement. Section 3 presents the lower bound under the (fully-poisonous) data poisoning threat model. Section 4 presents the matching upper bound under Byzantine failure. Section 5 presents an outline of our upper bound proof, specifically the analysis of the algorithm. Section 6 introduces the case of partially-poisonous local data and compare it with the fully-poisonous case. Section 7 provides concluding remarks and a discussion on open problems. Detailed proofs are deferred to appendices A and B.

2. Problem Statement and Assumptions

We consider a server-based system architecture with n workers and a central server. The workers only communicate with the server and there is no communication between workers. We assume that at most f out of n workers may be faulty, either as per Byzantine failure or fully-poisonous local data. We denote by $\mathcal{H}$ the set of $n-f$ honest workers, and let $Q^{(\mathcal{H})}(\theta)$ denote their average loss, i.e.,

$$Q^{(\mathcal{H})}(\theta) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} Q^{(i)}(\theta), \quad \forall \theta \in \mathbb{R}^d \ . \tag{5}$$

We assume that $Q^{(\mathcal{H})}$ admits a minimum, i.e., $\exists \theta^* \in \mathbb{R}^d$ such that for all $\theta \in \mathbb{R}^d$, $Q^{(\mathcal{H})}(\theta) \geq Q^{(\mathcal{H})}(\theta^*)$. We let $Q^* := Q^{(\mathcal{H})}(\theta^*)$. Furthermore, we consider the class of smooth loss functions satisfying the Polyak-Łojasiewicz (PL) inequality, which is more general than strong convexity (Bottou et al., 2018) and can indeed be satisfied by some non-convex functions (Karimi et al., 2016).

Assumption 1 (Smoothness) *There exists $L < \infty$ such that for all $i \in [n]$ and $\theta, \theta' \in \mathbb{R}^d$,*

$$\left\| \nabla Q^{(i)}(\theta) - \nabla Q^{(i)}(\theta') \right\| \leq L \left\| \theta' - \theta \right\| \ .$$

Assumption 2 (PL-condition) *There exists $\mu \geq 0$ such that for all $\theta \in \mathbb{R}^d$,*

$$\left\| \nabla Q^{(\mathcal{H})}(\theta) \right\|^2 \geq 2\mu \left(Q^{(\mathcal{H})}(\theta) - Q^* \right) \ .$$

As stated below, we also assume that the stochastic gradients computed by the honest workers have a bounded local covariance trace. This assumption is standard for analyzing the convergence of stochastic first-order methods (Tang et al., 2018). For all $i \in \mathcal{H}$, by definition of $Q^{(i)}$, and the assumption that $\nabla q(\theta, x)$ is continuous in θ and x, we have $\mathbb{E}_{x\sim\mathcal{D}^{(i)}}\left[\nabla q(\theta, x)\right] = \nabla Q^{(i)}(\theta)$.

Assumption 3 (Stochasticity) *There exists $\sigma < \infty$ such that for all $i \in \mathcal{H}$ and $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}_{x\sim\mathcal{D}^{(i)}}\left[\left\|\nabla q(\theta, x) - \nabla Q^{(i)}(\theta)\right\|^2\right] \leq \sigma^2 \ .$$

Lastly, as stated below, we assume the local gradients of the honest workers to have bounded diversity (or *heterogeneity*) over the parameter space. Without this assumption we cannot obtain meaningful guarantees in the threat models we consider, as shown by Karimireddy et al. (2022).

Assumption 4 (Heterogeneity) *There exists $\zeta < \infty$ such that for all $\theta \in \mathbb{R}^d$,*

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\nabla Q^{(i)}(\theta) - \nabla Q^{(\mathcal{H})}(\theta)\right\|^2 \leq \zeta^2 \ .$$

3. Lower Bound with Data Poisoning (fully-poisonous local data)

We characterize here the limitation of iterative stochastic first-order distributed algorithms in the data poisoning model. Specifically, we consider a generic randomized distributed algorithm Π that executes in T iterations. We define an execution of Π as follows. The server begins by choosing an initial parameter vector θ_0. In each iteration $t \geq 0$, the server maintains a parameter vector $\theta_t \in \mathbb{R}^d$ that is broadcast to the workers. Each honest worker i then samples one data point $x_t^{(i)}$ from its local distribution $\mathcal{D}^{(i)}$, computes a gradient $g_t^{(i)} = \nabla q\left(\theta_t, x_t^{(i)}\right)$, and sends back to the server a message

$$\text{msg}_t^{(i)} = \Psi_t\left((\theta_\tau)_{0\leq\tau\leq t},\, (g_\tau^{(i)})_{0\leq\tau\leq t}\right) \ ,$$

where $\Psi_t : \mathbb{R}^{d\times t} \times \mathbb{R}^{d\times t} \to \mathbb{R}^d$. A faulty worker j with a fully-poisonous dataset behaves exactly like an honest worker, except it samples its data point from an arbitrary distribution $\widetilde{\mathcal{D}}^{(j)}$ instead of its true local distribution $\mathcal{D}^{(j)}$. The faulty workers fix their poisonous data distributions before the execution of the algorithm, i.e., for a faulty worker j the distribution $\widetilde{\mathcal{D}}^{(j)}$ does not change during the execution of Π. The server then proceeds to update the current parameter vector θ_t to θ_{t+1}. At the completion of the T-th iteration, the server outputs $\hat{\theta}$. Note that this generic formulation includes the class of first-order optimization methods such as D-SGD and distributed momentum (Polyak, 1964; Farhadkhani et al., 2022a). We obtain a lower bound on the sub-optimality of Π, presented in Theorem 1, when there are at most $f < n/2$ faulty workers. The lower-bound is agnostic to the functions $\{\Psi_t\}_{t\in\{0,\ldots,T-1\}}$ that the workers implement to generate their messages, or the methods that the server implements to update its parameter vectors and generate the output.

Theorem 1 *Suppose assumptions 1, 2, 3, and 4 hold true. Let* $Q_0 := Q^{(\mathcal{H})}(\theta_0) - Q^*$. *Consider algorithm* Π *as described above. If there exists* $A \geq 0$ *such that* $\mathbb{E}_\Pi\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \leq A$, *then* $A \in \Omega\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu}\right)$, *where* $\mathbb{E}_\Pi[\cdot]$ *denotes the expectation over the randomness in* Π. *Moreover, we can guarantee that* $\mathbb{E}_\Pi\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \in \mathcal{O}\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu} + \varepsilon\right)$ *only if*

$$T \in \Omega\left(\frac{1+f}{n} \cdot \frac{\sigma^2}{\mu\varepsilon} + \frac{L}{\mu} \cdot \log\frac{Q_0}{\varepsilon}\right).$$

Proof [Proof sketch] We present here a sketch of our proof, and defer the formal proof to Appendix A. We prove the theorem for the scalar domain, i.e., $d = 1$, $\mathcal{X} \in \mathbb{R}$, and a quadratic loss, i.e., $q(\theta, x) = \frac{\mu}{4}(\theta - x)^2$. As the lower bound is established using the squared Euclidean norm, the proof applies directly to $d > 1$ since the instances used in the proof are still valid in a 1-dimensional subspace. We consider two separate cases, where the first case obtains the non-vanishing error term and the second case lower bounds the convergence rate.

First case. In this case, using the idea in Theorem III of Karimireddy et al. (2022) we derive a lower bound on the error when honest workers may have non-identical data distributions, which is the non-vanishing error term in Theorem 1. We partition the set of workers into $S = \{1, \ldots, n-f\}$ and $\hat{S} = \{n-f+1, \ldots, n\}$, and consider the following Dirac distributions:

$$\text{Distribution} \quad \mathcal{D}^{(i)} : \begin{cases} x = 0 \text{ , w.p. } 1 \text{ ;} & i \in S \\ x = \frac{2\zeta}{\mu}\sqrt{\frac{n-f}{f}} \text{ , w.p. } 1 \text{ ;} & i \in \hat{S} \end{cases}$$

We consider two valid executions of Π with different identities for the honest workers. In Execution 1, $\mathcal{H} = S$ and in Execution 2, $\mathcal{H} = \{1, \ldots, n-2f\} \cup \hat{S}$. As the guarantee of algorithm Π must hold true in both these executions, upon simply applying the condition on the loss function $Q^{(\mathcal{H})}(\hat{\theta})$ in both executions, we conclude that $\varepsilon \in \Omega\left(f/n \cdot \zeta^2/\mu\right)$.

Second case. In this case, we consider homogeneity, i.e., let $\mathcal{D}^{(i)} = \mathcal{D}$ for all $i \in \mathcal{H}$. Recall that in each execution of Π each worker computes a batch of T stochastic gradients, and f out of these n batches may be corrupted. Thus, upon extending the Huber's contamination model (see e.g. (Diakonikolas and Kane, 2022)) to batch sampling, we can show that it is impossible for Π to tell whether the honest workers send stochastic gradients corresponding to distribution $\mathcal{D}$ or another distribution $\mathcal{D}'$, both satisfying Assumption 3, if $\text{TV}\left(\mathcal{D}^T, \mathcal{D}'^T\right) \leq \frac{2f}{n}$.[3] We realize this scenario by the following instances:

$$\text{Distribution} \quad \mathcal{D} : \quad x = 0 \text{ , } \quad \text{w.p. } 1.$$

$$\text{Distribution} \quad \mathcal{D}' : \quad x = \begin{cases} \frac{2\sigma}{\mu}\sqrt{\frac{Tn}{2f}} & \text{, } \quad \text{w.p. } \frac{2f}{nT} \\ 0 & \text{, } \quad \text{w.p. } 1 - \frac{2f}{nT} \end{cases}$$

3. TV represents the *total variation distance* between two probability measures (Gibbs and Su, 2002).

As $\left(\mathbb{E}_{x\sim\mathcal{D}'}[x] - \mathbb{E}_{x\sim\mathcal{D}}[x]\right)^2 = \left(\frac{2\sigma}{\mu}\sqrt{\frac{2f}{nT}} - 0\right)^2 = \frac{f}{n}\cdot\frac{8\sigma^2}{\mu^2 T}$, for the considered quadratic loss function $q(\theta, x) = \frac{\mu}{4}(\theta - x)^2$, we conclude that

$$\mathbb{E}_{\Pi}\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \in \Omega\left(\frac{f}{n}\cdot\frac{\sigma^2}{\mu T} + \frac{1}{n}\cdot\frac{\sigma^2}{\mu T}\right) \ ,$$

which means to get an ε-approximate solution, we must have

$$T \in \Omega\left(\frac{f+1}{n}\cdot\frac{\sigma^2}{\mu\varepsilon}\right).$$

While the first term in the argument of Ω above comes from the fact that we cannot distinguish between the two valid distributions $\mathcal{D}$ and $\mathcal{D}'$, the second term is due to the classical lower bound on the minimax statistical error considering Gaussian distributions (Wu, 2017), i.e., the worst-case squared-error incurred in estimating the mean of a distribution with variance σ^2 from at most nT i.i.d. samples. Finally, in the case where $\sigma = 0$ and $\zeta = 0$, i.e., all the honest workers send the same gradient vector, we have the lower bound of $\Omega(L/\mu \cdot \log Q_0/\varepsilon)$, shown by Yue et al. (2023).

We conclude by composing the bounds obtained in the different cases. ■

Remark 2 *The lower bound shown in Theorem 1 extends to the* mini-batch *gradients setting, wherein honest workers sample multiple number of data points b from their local distribution (sampled independently), by replacing σ^2 by σ^2/b. The proof for this extension remains almost identical to that of Theorem 1, except that in the second case each worker now computes a batch of bT stochastic gradients in each execution of Π. Hence, replacing T with bT in the proof suffices.*

4. Upper Bound with Byzantine failure

We present here a matching upper bound (up to a multiplicative factor of condition number) for Theorem 1, considering a Byzantine adversary. We first describe the algorithm we consider, and then present its convergence guarantee.

4.1 Algorithm Description

The algorithm follows the skeleton of DSGD and imparts robustness to the learning procedure by applying a momentum operation at the workers' level and a trimmed mean operation at the server (instead of averaging), as described in Algorithm 1. Essentially, in each iteration $t \geq 0$, each honest worker i computes a stochastic gradient

$$g_t^{(i)} := \nabla q\left(\theta_t, x^{(i)}\right), \quad \text{where } x^{(i)} \sim \mathcal{D}^{(i)} \ , \tag{6}$$

and returns a *Polyak's momentum* of its stochastic gradients, denoted by $m_t^{(i)}$ and defined:

$$m_t^{(i)} = \beta_t m_{t-1}^{(i)} + (1-\beta_t) g_t^{(i)} \ , \tag{7}$$

Algorithm 1: DSGD with distributed momentum and trimmed mean aggregation

Input : $T \geq 2$, $(\gamma_0, \ldots, \gamma_{T-1})$ and $(\beta_0, \ldots, \beta_{T-1})$.

Server chooses arbitrarily $\theta_0 \in \mathbb{R}^d$. Each **honest worker** i sets $m_{-1}^{(i)} = 0$.
for $t = 0$ *to* $T-1$ **do**
  Server broadcasts θ_t to all workers;
  for *each* **honest worker** i *(in parallel)* **do**
    Compute a stochastic gradient $g_t^{(i)}$, as defined in (6);
    Send to the server the momentum $m_t^{(i)}$, as defined in (7);
  end
  % A corrupted worker i may send an arbitrary value for $m_t^{(i)}$ to the server.
  Server updates the parameter vector $\theta_{t+1} = \theta_t - \gamma_t \mathrm{TM}^{(f)}\left(m_t^{(1)}, \ldots, m_t^{(n)}\right)$;
end

Output: $\hat{\theta} = \theta_T$

where $\beta_t \in [0, 1)$ is the *momentum coefficient*, and $m_{-1}^{(i)} = 0$ by convention. The server updates its current parameter vector θ_t by aggregating the workers' momentums using *coordinate-wise trimmed mean* (TM), defined below. Hereafter, for any $z \in \mathbb{R}^d$ and $k \in [d]$, we denote by $[z]_k$ the k-th coordinate of z. Then, given n input vectors $z_1, \ldots, z_n \in \mathbb{R}^d$, for all $k \in [d]$, we denote by τ_k the permutation on $[n]$ that sorts the k-th coordinates of the input vectors in non-decreasing order, i.e., $[z_{\tau_k(1)}]_k \leq [z_{\tau_k(2)}]_k \leq \ldots \leq [z_{\tau_k(n)}]_k$. Then, the trimmed mean of $z_1, \ldots, z_n$, with trimming parameter f is a vector in $\mathbb{R}^d$ whose k-th coordinate is defined as follows,

$$\left[\mathrm{TM}^{(f)}(z_1, \ldots, z_n)\right]_k \coloneqq \frac{1}{n-2f} \sum_{j \in [f+1, n-f]} [z_{\tau_k(j)}]_k \ .$$

Why coordinate-wise trimmed mean? Our choice of TM is motivated from prior work Allouah et al. (2023a) that showed its optimality (in the case of distributed gradient descent) under the bounded gradient heterogeneity condition (i.e., Assumption 4). However, TM need not be optimal in a setting wherein the gradients' stochasticity or heterogeneity conditions, i.e., Assumptions 3 or 4, respectively, are relaxed from covariance matrices with bounded traces to bounded spectral norms. In the latter case, high-dimensional robust aggregators using spectral filtering, e.g., see Diakonikolas et al. (2017), are usually the optimal choices. For more insights on this subtle but important distinction between bounded trace and bounded spectral norm we refer the reader to Allouah et al. (2023b).

4.2 Formal Statement

Theorem 3 below establishes the convergence of Algorithm 1, with a Byzantine adversary, assuming a scheduled decreasing step sizes and increasing momentum coefficients. Note that the algorithm is oblivious to the identity of faulty workers that may send arbitrary values to the server. We denote by $\mathbb{E}\,[\cdot]$ the expectation on the randomness of the algorithm, formally defined in Appendix B.

Theorem 3 *Suppose assumptions 1, 2, 3, and 4 hold true. Consider Algorithm 1 with $T \geq 2$ and the following two options for the scheduled step sizes and momentum coefficients.*

- ***Option 1:*** *If $T \leq \frac{54L}{\mu}$, then, $\forall t \in \{0, \ldots, T-1\}$, set $\gamma_t = \frac{1}{18L}$, and $\beta_t = 0$.*
- ***Option 2:*** *If $T > \frac{54L}{\mu}$, then, $\forall t \in \{0, \ldots, T-1\}$, set*

$$\gamma_t = \frac{1}{18L + \left[\frac{\mu}{6}(t - t_0 + 1)\right]^+} \ , \quad \text{and} \quad \beta_t = 1 - 18L\gamma_{t-1} \ .$$

Where $t_0 = \left\lceil \frac{T}{2} \right\rceil$, $\gamma_{-1} = \frac{1}{18L}$ and $[\cdot]^+ := \max\{0, \cdot\}$.

Then, the following holds true

$$\mathbb{E}\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \leq \frac{7}{6} Q_0 \cdot e^{-\frac{T}{108K}} + \left(\lambda + \frac{1}{n-f}\right) \cdot \frac{4374K\sigma^2}{T\mu} + \frac{9\lambda\zeta^2}{2\mu} \ ,$$

where $Q_0 := Q^{(\mathcal{H})}(\theta_0) - Q^*$, $\lambda = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$, and $K = \frac{L}{\mu}$.

Using Theorem 3, we can derive a matching upper bound for Theorem 1 when $K \in \mathcal{O}(1)$. Specifically, ignoring the constants, we obtain the following corollary.

Corollary 4 *Suppose $n \geq (2+\nu)f$ for some constant $\nu > 0$. Under the conditions stated in Theorem 3, Algorithm 1 guarantees that*

$$\mathbb{E}\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \in \mathcal{O}\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu} + \varepsilon\right) \ ,$$

with an iteration complexity in

$$T \in \mathcal{O}\left(\frac{1+f}{n} \cdot \frac{K\sigma^2}{\mu\varepsilon} + \frac{L}{\mu} \cdot \log\frac{Q_0}{\varepsilon}\right) \ .$$

5. Roadmap to Proving Theorem 3

We present here the key steps involved in proving Theorem 3. Our proof is based on a new Lyapunov function, denoted by V_t. We first motivate the design of V_t, and define it formally. We then analyze the growth of V_t along the trajectory of Algorithm 1. Lastly, we show the convergence of the sequence $(V_t)_{t=0}^{T-1}$ for the specified diminishing step sizes, thereby proving our result.

Analyzing the growth of the loss function. We analyze the growth of the loss function $Q^{(\mathcal{H})}(\theta_t)$ along the trajectory of Algorithm 1. For any $t \geq 0$ we denote the average momentum of the honest workers as $\overline{m}_t \coloneqq \frac{1}{(n-f)} \sum_{i \in \mathcal{H}} m_t^{(i)}$. Combining the result of Allouah et al. (2023a) on the robustness of TM with the standard decomposition of the loss function under smoothness assumption (Bottou et al., 2018), we get the following bound on the growth of the loss function.

Lemma 5 *Suppose Assumption 1 holds true. Consider Algorithm 1 with $T \geq 2$, and $\gamma_t \leq 1/L$ for all $t \in \{0, \ldots, T-1\}$. Let λ be as defined in Lemma 15. Then, for all t, the following holds true*

$$\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t)\right] \leq -\frac{\gamma_t}{2}\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right] + \gamma_t \frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + \gamma_t\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t) - \overline{m}_t\right\|^2\right] \ ,$$

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

From Lemma 5, we obtained a bound on the growth of the loss function $Q^{(\mathcal{H})}$ during the learning procedure. This lemma highlights the importance of two key quantities: (i) the *deviation* of the average momentum, and (ii) the *drift* of each worker i from the average momentum.

Incorporating the drift and deviation in the Lyapunov function. In the remaining, for any $t \geq 0$, we denote the deviation and the drift of each worker i as

$$\delta_t \coloneqq \overline{m}_t - \nabla Q^{(\mathcal{H})}(\theta_t) \quad \text{and} \quad \Delta m_t^{(i)} \coloneqq m_t^{(i)} - \overline{m}_t, \forall i \in \mathcal{H} \ . \tag{8}$$

Due to the time-varying step size and momentum coefficient in Algorithm 1, it is difficult to derive a uniform bound (i.e., a bound that holds true for any $t \geq 0$) on the second and third terms in the right hand side of Lemma 5. Accordingly, we cannot simply and directly analyze the variation of $\mathbb{E}\left[Q^{(\mathcal{H})}\left(\theta_t\right) - Q^*\right]$ with t. Instead, we have to incorporate the drift and the deviation in the analysis. Specifically, we define the following Lyapunov function for our problem.

$$V_t \coloneqq \mathbb{E}\left[Q^{(\mathcal{H})}\left(\theta_t\right) - Q^* + \rho\left\|\delta_t\right\|^2 + \rho\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\left\|\Delta m_t^{(i)}\right\|^2\right] \ , \tag{9}$$

where $\rho = \frac{1}{12L}$. Then, by definition of V_t, we have $\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_T) - Q^*\right] \leq V_T$. Hence an upper bound on V_T gives us an upper bound on $\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_T) - Q^*\right]$. With this Lyapunov function at hand, we can construct the proof by following three critical steps: (i) determining a recursive bound on the Lyapunov function V_t, (ii) choosing a desirable sequence $(\gamma_0, \ldots, \gamma_{T-1})$ to obtain tight convergence rate, and (iii) combining (i) and (ii) to derive the final bound on $\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_T) - Q^*\right]$.

Recursive bound on V_t. We first derive a recursive bound for each of the terms in the V_t. In doing so, we start by showing in Lemma 6 that the average drift over the honest workers' momentum is controlled by β_t, the gradient diversity ζ^2 and the gradient stochasticity σ^2.

Lemma 6 *Suppose assumptions 3, and 4 hold true, and consider Algorithm 1 with $T \geq 2$. Then, for any $t \in \{0, \ldots, T-1\}$, the following holds true*

$$\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\Delta m_t^{(i)}\right\|^2\right] \leq \beta_t\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\Delta m_{t-1}^{(i)}\right\|^2\right] + (1-\beta_t)\zeta^2 + (1-\beta_t)^2\sigma^2 \ .$$

Next, we study the deviation δ_t of the average momentum $\overline{m}_t$ from the true gradient $\nabla Q^{(\mathcal{H})}(\theta_t)$. We obtain in Lemma 7 an upper bound on the growth of the deviation over the steps $t \in \{0, \dots, T-1\}$.

Lemma 7 *Suppose assumptions 1, 3, and 4 hold true, consider Algorithm 1 with $T \geq 2$, and λ as defined in Lemma 5. Then for any $t \in \{0, \dots, T-1\}$, the following holds true*

$$\mathbb{E}\left[\|\delta_{t+1}\|^2\right] \leq \beta_{t+1}^2 \left(1 + 4\gamma_t L + 3\gamma_t^2 L^2\right) \mathbb{E}\left[\|\delta_t\|^2\right] + (1-\beta_{t+1})^2 \frac{\sigma^2}{n-f} + 3\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L)\left(\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}} \mathbb{E}\left[\left\|\Delta m_t^{(i)}\right\|^2\right] + \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right]\right) \ .$$

Finally, combining Lemmas 6 and 7 with Lemma 5, we can derive a proper recursive bound on V_t, as presented in Lemma 8 below.

Lemma 8 *Suppose assumptions 1, 2, 3, and 4 hold true. Consider Algorithm 1 with $T \geq 2$ and a set of parameters such that $t \in \{0, \cdots, T\}$, $\gamma_t \leq \frac{1}{18L}$, and $1 - \beta_{t+1} = 18\gamma_t L$. Finally, let $(V_t)_{t\geq 0}$ be as defined in (9) and λ as defined in Lemma 5. Then the following holds true*

$$V_{t+1} \leq \left(1 - \frac{\mu\gamma_t}{3}\right) V_t + 27L\left(\lambda + \frac{1}{n-f}\right)\sigma^2\gamma_t^2 + \frac{3}{2}\lambda\zeta^2\gamma_t \ .$$

Choice of the step sizes $(\gamma_0, \dots, \gamma_{T-1})$. To obtain a tight convergence rate (and avoid logarithmic terms), we need to carefully choose the sequence of the step sizes $(\gamma_0, \dots, \gamma_{T-1})$ we use, as recently pointed out by Stich (2019). Specifically, following the recent advancement on this matter (Khaled and Richtárik, 2023), we design a generic scheduling technique, described in Lemma 9 below.

Lemma 9 *Let a, b, c, d be positive real values with $a < b$, and let $T \geq 2$ be a positive integer. Let $(\gamma_0, \dots, \gamma_{T-1})$ and $(r_0, \dots, r_T)$ be real valued sequences such that for all $t \in \{0, \dots, T-1\}$,*

$$r_{t+1} \leq (1 - a\gamma_t) r_t + c\gamma_t^2 + d\gamma_t \ .$$

Consider the following two cases:

- ***Case 1:*** *$T \leq b/a$ and $\gamma_t = 1/b$, $\forall t \in \{0, \dots, T-1\}$.*
- ***Case 2:*** *$T > b/a$, and $\gamma_t = \frac{1}{b + \left[\frac{a}{2}(t - t_0 + 1)\right]^+}$, $\forall t \in \{0, \dots, T-1\}$, where $t_0 = \lceil T/2 \rceil$.*

In both Case 1 and Case 2, we have: $r_T \leq r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{18c}{a^2 T} + \frac{3d}{a}$.

Final step for the proof sketch of Theorem 3. Lastly, we apply Lemma 9 to the recursion of Lemma 8, with $a = \frac{\mu}{3}$, $b = 18L$, $c = 27L\left(\lambda + \frac{1}{n-f}\right)\sigma^2$ and $d = \frac{3}{2}\lambda\zeta^2$, and obtain that

$$V_T \leq V_0 \exp\left(-\frac{\mu T}{108L}\right) + \frac{4374L\left(\lambda + \frac{1}{n-f}\right)\sigma^2}{T\mu^2} + \frac{9\lambda\zeta^2}{2\mu} \ .$$

As $\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_T) - Q^*\right] \leq V_T$, we conclude the proof by showing that $V_0 \leq \frac{7}{6}\left(Q^{(\mathcal{H})}(\theta_0) - Q^*\right)$.

6. Partially-Poisonous Local Data

A standard assumption in robust distributed ML literature that we have also made so far is that each worker is either *entirely corrupted* or honest. If a worker is honest then it is assumed that all of its data points are sampled correctly and that it always follows the prescribed algorithm. However, in practice, we might have some corrupted data points among the data points available to all the workers. In particular, instead of considering a fraction of corrupted workers, we may assume a fraction of the data points available to all workers are poisonous (or incorrectly sampled). To address a general data poisoning setting, in this section, we consider both worker-level and global-level data corruptions. Specifically, we assume that the datasets of up to f out of n workers are fully corruptible and that the datasets of remaining $n-f$ workers is partially corruptible. To characterize the impact of these two types of corruptions, we focus on empirical loss minimization where each worker i has a dataset $\mathcal{S}^{(i)}$ of m data points.[4] We assume that b out of m data points of each worker can be arbitrarily corrupted. We let $\mathcal{D}^{(i)}$ denote the uniform distribution over the remaining $m-b$ incorruptible data points. By (1), we have

$$Q^{(i)}(\theta) \coloneqq \mathbb{E}_{x\sim\mathcal{D}^{(i)}}\left[q(\theta,x)\right] = \frac{1}{m-b}\sum_{x\in\mathcal{S}_h^{(i)}} q(\theta,x), \tag{10}$$

where $\mathcal{S}_h^{(i)} \coloneqq \textsc{supp}\left(\mathcal{D}^{(i)}\right)$ is the set of honest data points of worker i. This general data poisoning model encompasses various scenarios. For instance, setting $n=1$ and $f=0$ corresponds to the centralized poisoning problem, where a portion of a large dataset is corrupted. Furthermore, $b=0$ corresponds to the case where some of the workers are always correct which is the scenario often studied in the Byzantine ML literature that we considered in the previous sections. In the rest of this section, we prove matching upper and lower bounds on the learning error in the above setting.

Remark 10 *For the simplicity of presentation, we only consider the data poisoning threat. However, our upper bound holds even for the stronger Byzantine failure threat model, where recall that when a worker is corrupted, it can send an arbitrary vector for its gradient.*

6.1 Lower Bound

Theorem 11 *Consider the average empirical loss* (5) *with individual loss functions as defined in* (10)*. Suppose assumptions 1, 2, 3, and 4. For any algorithm* Π *outputting a model* $\hat{\theta}_\Pi$*, we have*

$$Q^{(\mathcal{H})}\left(\hat{\theta}_\Pi\right) - Q^* \in \Omega\left(\frac{f}{n}\cdot\frac{\zeta^2}{\mu} + \frac{b}{m}\cdot\frac{\sigma^2}{\mu}\right).$$

Proof We prove the theorem for the scalar domain, i.e., $d=1$, $\mathcal{X}\in\mathbb{R}$, and a quadratic loss, i.e., $q(\theta,x)=\frac{\mu}{4}(\theta-x)^2$. The proof for the first term, i.e., $\Omega\left(\frac{f}{n}\cdot\frac{\zeta^2}{\mu}\right)$, follows from the second case in the proof of Theorem 1. The second term, i.e., $\Omega\left(\frac{b}{m}\cdot\frac{\sigma^2}{\mu}\right)$ term, also follows from the arguments made in the proof of Theorem 1. We provide key differences below.

4. A solution to the empirical loss minimization is a $\mathcal{O}\left(\frac{\sigma^2}{m}\right)$ approximate solution to the statistical loss.

Suppose that $f = 0$, i.e., there is no worker with full-poisonous data in the system. Also, suppose that all the workers have identical local datasets, i.e., $\zeta = 0$. Since, having multiple copies of the same dataset does not provide any additional information, the problem reduces to the case with a single worker possessing a dataset denoted as $\mathcal{S}^{(1)}$ such that the honest data points $\mathcal{S}_h^{(1)}$ satisfy Assumption 3. Consider a quadratic loss function $q(\theta, x) = \frac{\mu}{4}(\theta - x)^2$ with gradient $\nabla q(\theta, x) = \frac{\mu}{2}(\theta - x)$. This loss satisfies assumptions 1 and 2. For any $j \in [m]$, let $x^{(1,j)}$ be the j-th data point in $\mathcal{S}^{(1)}$. Now suppose that $x^{(1,j)} = 0$ for $1 \leq j \leq m - b$, and $x^{(1,j)} = \frac{2\sigma}{\mu}\sqrt{\frac{m-b}{b}}$ for $m - b + 1 \leq j \leq m$. Consider the following two cases:

Case 1: $\mathcal{S}_h^{(1)} \coloneqq \left\{x^{(1,j)} : 1 \leq j \leq m - b\right\}$.

Case 2: $\mathcal{S}_h^{(1)} \coloneqq \left\{x^{(1,j)} : b + 1 \leq j \leq m\right\}$.

In case 1, we have

$$\mathbb{E}_{x\sim\mathcal{D}^{(1)}}\left[\left(\nabla q\left(\theta, x\right) - \nabla Q^{(1)}(\theta)\right)^2\right] = \frac{1}{m-b}\sum_{x\in\mathcal{S}_h^{(1)}}\left(\nabla q\left(\theta, x\right) - \nabla Q^{(1)}(\theta)\right)^2 = 0 \leq \sigma^2 \ .$$

In case 2, using the same technique as in Execution 2 of the second case in the proof of Theorem 1, we have

$$\mathbb{E}_{x\sim\mathcal{D}^{(1)}}\left[\left(\nabla q\left(\theta, x\right) - \nabla Q^{(1)}(\theta)\right)^2\right] = \frac{1}{m-b}\sum_{x\in\mathcal{S}_h^{(1)}}\left(\nabla q\left(\theta, x\right) - \nabla Q^{(1)}(\theta)\right)^2 = \sigma^2 \ .$$

Therefore, in both cases, Assumption 3 is satisfied.

Now, suppose that algorithm Π provides an ε-approximation guarantee on the learning error. Specifically, in both cases, we have

$$Q^{(\mathcal{H})}\left(\hat{\theta}_\Pi\right) - Q^* \leq \varepsilon.$$

This implies that (refer the first case in the proof of Theorem 1),

$$\frac{\mu}{4}\left(\hat{\theta}_\Pi\right)^2 \leq \varepsilon \quad \text{and} \quad \frac{\mu}{4}\left(\hat{\theta}_\Pi - \frac{2\sigma}{\mu}\sqrt{\frac{b}{m-b}}\right)^2 \leq \varepsilon.$$

Thus, applying Jensen's inequality, we obtain that

$$\varepsilon \geq \frac{\mu}{16}\left(\frac{2\sigma}{\mu}\sqrt{\frac{b}{m-b}}\right)^2.$$

The above implies that $\varepsilon \in \Omega\left(\frac{b}{m}\cdot\frac{\sigma^2}{\mu}\right)$. This concludes the proof. ■

In contrast to the case of fully-poisonous local data (cf., Section 3), in the case of partially-poisonous local data the impact of σ^2 on the error does not vanish with T. Intuitively this comes from the fact that each worker can be seen as a server trying to aggregate

information from m sub-workers out of which b are Byzantine. These sub-workers, each holding a single data point, creates a local heterogeneity characterized by σ^2. Hence, the learning error is limited by the heterogeneity of point-wise gradients σ^2 and the fraction of poisoned data points b/m, analogously to ζ^2 and f/n, respectively, in the lower bound derived in Section 3.

6.2 Upper Bound

In this section, we establish an upper bound that matches the lower bound presented in Theorem 11 by considering Algorithm 2. Notably, Algorithm 2 exhibits three key distinctions when compared to Algorithm 1. Firstly, Algorithm 2 operates deterministically; at each iteration, every worker computes the gradient over its entire dataset, in contrast to the stochastic nature of Algorithm 1. Secondly, in addition to the global aggregation functions performed by the server, each worker in Algorithm 2 incorporates a locally applied trimmed mean aggregation function. This function serves to filter out outliers, ensuring the robustness of the local updates. Finally, Algorithm 2 does not require local momentum (owing to its deterministic nature), and the model is updated using robustified gradient vectors. The following theorem shows the convergence of Algorithm 2. The proof can be found in Appendix C.

Algorithm 2: DGD with local and global trimmed mean aggregations

Input : $T \geq 2$, step size $\gamma > 0$.

Server chooses arbitrarily $\theta_0 \in \mathbb{R}^d$.

for $t = 0$ *to* $T-1$ **do**

  Server broadcasts θ_t to all workers;

  for *each* **honest worker** i *(in parallel)* **do**

    for each data point $x \in \mathcal{S}^{(i)}$, compute $\nabla q\left(\theta^{(t)}, x\right)$, compute

$$G_t^{(i)} := \mathrm{TM}^{(b)}\left(\nabla q\left(\theta^{(t)}, x\right), \forall x \in \mathcal{S}^{(i)}\right),$$

    and send $G_t^{(i)}$ to the server.

  end

  % A corrupted worker i may send an arbitrary vector to the server.

  Server updates the parameter vector $\theta_{t+1} = \theta_t - \gamma \mathrm{TM}^{(f)}\left(G_t^{(1)}, \ldots, G_t^{(n)}\right)$;

end

Output: $\hat{\theta} = \theta_T$

Theorem 12 *Consider the average empirical loss* (5) *with individual loss functions as defined in* (10)*. Suppose assumptions 1, 2, 3, and 4. Consider Algorithm 2 with* $\gamma = 1/L$*. Then, we have*

$$Q^{(\mathcal{H})}(\theta_T) - Q^* \leq \exp\left(-\frac{\mu}{L}T\right)\left(Q^{(\mathcal{H})}(\theta_0) - Q^*\right) + \frac{1}{\mu}(\lambda'\sigma^2 + 3\lambda\lambda'\sigma^2 + 3\lambda\zeta^2),$$

where $\lambda = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right)$ *and* $\lambda' = \frac{6b}{m-2b}\left(1 + \frac{b}{m-2b}\right)$.

Note that $\lambda' \in \mathcal{O}\left(\frac{b}{m}\right)$ and $\lambda \in \mathcal{O}\left(\frac{f}{n}\right)$. Hence, we obtain the following corollary of Theorem 12.

Corollary 13 *Under the same conditions as in Theorem 12, Algorithm 2 outputs θ_T such that*

$$Q^{(\mathcal{H})}(\theta_T) - Q^* \in \mathcal{O}\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu} + \frac{b}{m} \cdot \frac{\sigma^2}{\mu} + \varepsilon\right),$$

as long as $T \in \mathcal{O}\left(\frac{L}{\mu} \cdot \log \frac{Q_0}{\varepsilon}\right)$.

7. Concluding Remarks & Open Problems

We have shown that the Byzantine failure threat model is not an overkill for addressing the more practical threat model of data poisoning. Specifically, we have shown that state-of-the-art solutions to the *Byzantine ML* problem, such as the ones proposed in Farhadkhani et al. (2022a); Karimireddy et al. (2022); Allouah et al. (2023a); Gorbunov et al. (2023), provide optimal protection against data poisoning attacks. Although our result applies to ML problems that are solvable by optimizing over Polyak-Łojasiewicz (PL) loss functions, we believe that our deductions hold true even for a larger set of functions that do not necessarily satisfy the PL inequality. This constitutes an interesting future research direction. Furthermore, we have also shown that Byzantine robustness schemes yield tight solutions in both partial-poisonous and full-poisonous local data settings.

Note that we have only considered *untargeted* attacks in both the Byzantine failure and the data poisoning threat models. An interesting future direction would be to consider *targeted* attacks, wherein corrupted workers do not necessarily attempt to maximize the learning error, but rather act strategically to manipulate the learning into converging to a target region in the model space that performs poorly on specific types of inputs (i.e., has high generalization errors), e.g., see (Dai et al., 2019; Wang et al., 2020; Zhao et al., 2020; Truong et al., 2020; Severi et al., 2021). While a recent work has attempted to compare Byzantine failure and data poisoning in the context of targeted attacks (Farhadkhani et al., 2022b), the findings only applicable to conventional ML methods that do not incorporate any robustness properties. Our proof techniques could be used to obtain a principled comparison between the two threat models in the targeted attacks scenario.

8. Acknowledgments

The authors are thankful to the editor and anonymous reviewers for their constructive feedback that helped improve the manuscript. Rafael is partially supported by the French National Research Agency and the French Ministry of Research and Higher Education. The joint work of Nirupam and Rafael is also supported in part by the CNRS through the International Emerging Action (IEA) program and the French National Research Agency through the ANR TuLIP. This work has also been partly supported by the Swiss National Science Foundation.

References

D. Alistarh, Z. Allen-Zhu, and J. Li. Byzantine stochastic gradient descent. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

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

Y. Allouah, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. On the privacy-robustness-utility trilemma in distributed learning. In *International Conference on Machine Learning*, number 202, 2023b.

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

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

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

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

M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 2017.

Y. Chen, L. Su, and J. Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2), 2017.

J. Dai, C. Chen, and Y. Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7, 2019.

D. Data and S. Diggavi. Byzantine-resilient high-dimensional SGD with local iterations on heterogeneous data. In *International Conference on Machine Learning*. PMLR, 2021.

I. Diakonikolas and D. M. Kane. *Algorithmic High-Dimensional Robust Statistics*. 2022.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 999–1008. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/diakonikolas17a.html`.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*. PMLR, 2019.

E. M. El Mhamdi, S. Farhadkhani, R. Guerraoui, A. Guirguis, L. N. Hoang, and S. Rouault. Collaborative learning in the jungle (decentralized, Byzantine, heterogeneous, asynchronous and nonconvex learning). In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

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

S. Farhadkhani, R. Guerraoui, O. Villemaud, et al. An equivalence between data poisoning and Byzantine gradient attacks. In *International Conference on Machine Learning*. PMLR, 2022b.

S. Farhadkhani, R. Guerraoui, N. Gupta, L.-N. Hoang, R. Pinot, and J. Stephan. Robust collaborative learning with linear gradient overhead. In *International Conference on Machine Learning*, pages 9761–9813. PMLR, 2023.

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

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3), 2002.

E. Gorbunov, S. Horváth, P. Richtárik, and G. Gidel. Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=pfuqQQCB34`.

R. Guerraoui, N. Gupta, and R. Pinot. Byzantine machine learning: A primer. *ACM Computing Surveys*, 2023.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1), 1964. doi: 10.1214/aoms/1177703732. URL `https://doi.org/10.1214/aoms/1177703732`.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in

federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2), 2021. ISSN 1935-8237. doi: 10.1561/2200000083.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46128-1.

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

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

A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=AU4qHN2VkS`. Survey Certification.

J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

L. Lamport, R. Shostak, and M. Pease. The Byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3), July 1982. ISSN 0164-0925. doi: 10.1145/357172.357176.

S. Liu, N. Gupta, and N. H. Vaidya. Approximate Byzantine fault-tolerance in distributed optimization. In *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, PODC'21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385480. doi: 10.1145/3465084.3467902. URL `https://arxiv.org/pdf/2101.09337.pdf`.

S. Mahloujifar, M. Mahmoody, and A. Mohammed. Data poisoning attacks in multi-party learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*. PMLR, 2019.

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

A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3), 2020.

G. Severi, J. Meyer, S. Coull, and A. Oprea. Explanation-guided backdoor poisoning attacks against malware classifiers. In M. Bailey and R. Greenstadt, editors, *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*. USENIX Association, 2021.

V. Shejwalkar and A. Houmansadr. Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.

V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022.

S. U. Stich. Unified optimal analysis of the (stochastic) gradient method, 2019. URL `https://arxiv.org/abs/1907.04232`.

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

H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu. D^2: Decentralized training over decentralized data. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018. URL `http://proceedings.mlr.press/v80/tang18a.html`.

L. Truong, C. Jones, B. Hutchinson, A. August, B. Praggastis, R. Jasper, N. Nichols, and A. Tuor. Systematic evaluation of backdoor data poisoning attacks on image classifiers. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. Computer Vision Foundation / IEEE, 2020.

H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Y. Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*, 16, 2017.

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

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

P. Yue, C. Fang, and Z. Lin. On the lower bound of minimizing Polyak-Lojasiewicz functions, 2022. URL `https://arxiv.org/abs/2212.13551`.

P. Yue, C. Fang, and Z. Lin. On the lower bound of minimizing polyak-łojasiewicz functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2948–2968. PMLR, 2023.

S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y. Jiang. Clean-label backdoor attacks on video recognition models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020.

Appendix

Appendix A. Proof of Theorem 1

Remark 14 *Note that, to prove Theorem 1, we focus on the special case where $d = 1$. As the lower bound is established using the squared Euclidean norm, the proof applies directly to $d > 1$ since the instances used in the proof are still valid in a 1-dimensional subspace. Moreover, as we later prove in Corollary 4, this lower bound is tight as it is matched by Algorithm 1 for an arbitrary d. Note, however, that despite the explicit absence of the dimension d in the asymptotic error and the convergence rate, the impact of dimension d is implicit through σ^2, i.e., the bound stated in Assumption 3 on the* covariance trace *of the local stochastic noise. Indeed, when the variance of noise in each coordinate of the stochastic gradients might be as large as some real value ς^2, we have $\sigma^2 = d \cdot \varsigma^2$.*

To prove Theorem 1, we need to show that for any $T > 0$, and any algorithm Π, we must have[5]

$$\mathbb{E}_{\Pi}\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \in \Omega\left(\frac{f+1}{n} \cdot \frac{\sigma^2}{\mu T} + \frac{f}{n} \cdot \frac{\zeta^2}{\mu} + e^{-\frac{T}{K}}\right).$$

We assume that the output $\hat{\theta}$ of algorithm Π satisfies the condition: $\mathbb{E}_{\Pi}\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \leq A$ for $A > 0$. To obtain a lower bound on A, we consider a setting where $d = 1$, $\mathcal{X} \subseteq \mathbb{R}$ and the loss function $q(\theta, x) = \frac{\mu}{4}(\theta - x)^2$ where $0 < \mu < \infty$. We consider two separate cases, each with different instances of data distributions subject to assumptions 1, 2, 3, and 4.

In the first case, we consider heterogeneous distributions for honest workers, i.e., $\zeta \geq 0$. In this particular case, we adapt the proof of Theorem III of Karimireddy et al. (2022) to show that

$$A \in \Omega\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu}\right). \tag{11}$$

In the second case, we assume $\mathcal{D}^{(i)} = \mathcal{D}$ for all $i \in \mathcal{H}$, i.e., $\zeta = 0$ in Assumption 4. In this particular case, we develop upon the indistinguishability of valid distributions in the general contamination model (shown in Proposition 1.7 of Diakonikolas and Kane (2022)) to show that

$$A \in \Omega\left(\frac{f+1}{n} \cdot \frac{\sigma^2}{\mu T}\right). \tag{12}$$

As $\hat{\theta}$ should satisfy the bound in both cases, **the proof concludes upon combining** (12) **and** (11)**.** and the recently discovered $\Omega(e^{-\frac{T}{K}})$ lower bound for first-order deterministic algorithms (Yue et al., 2023) in the vanilla (non-Byzantine) setting.[6]

First Case. In this case, we obtain a bound on the error when honest workers may have non-identical data distributions. Our derivation follows from the proof of Theorem III of

5. Here we ignore the absolute constant in the exponent as it corresponds to a constant multiplied by the logarithmic term in Theorem 1.
6. Follows from the fact that $\max\{a, b, c\} \geq \frac{1}{3}(a + b + c)$.

Karimireddy et al. (2022). We partition the set of workers into $S = \{1, \ldots, n-f\}$ and $\hat{S} = \{n-f+1, \ldots, n\}$. We consider the following Dirac distributions of data.

$$\text{Distribution} \quad \mathcal{D}^{(i)} : \begin{cases} x = 0 \ \text{ with probability } \ 1 \ ; & i \in S \\ x = \frac{2\zeta}{\mu}\sqrt{\frac{n-f}{f}} \ \text{ with probability } \ 1 \ ; & i \in \hat{S} \end{cases} .$$

Next, we consider two valid executions of Π with different identities for honest workers. In Execution 1, $\mathcal{H} = S$ and in Execution 2, $\mathcal{H} = \{1, \ldots, n-2f\} \cup \hat{S}$. It is easy to verify (using similar steps as in the first case) that assumptions 1, 2 and 3 are satisfied in either executions. We show below that Assumption 4 is also satisfied in the two executions. Hence, validating both the executions. Recall that we assume $f < \frac{n}{2}$.

Note that $Q^{(i)}(\theta) \coloneqq \frac{\mu}{4}\theta^2$ for all $i \in S$, and $Q^{(i)}(\theta) \coloneqq \frac{\mu}{4}\left(\theta - \frac{2\zeta}{\mu}\sqrt{\frac{n-f}{f}}\right)^2$ for all $i \in \hat{S}$. In **Execution 1**, as $\mathcal{H} = S$, it is easy to see that

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\nabla Q^{(i)}(\theta) - \nabla Q^{(\mathcal{H})}(\theta)\right\|^2 = \frac{1}{|S|}\sum_{i\in S}\left(\frac{\mu}{2}\theta - \frac{1}{|S|}\sum_{j\in S}\frac{\mu}{2}\theta\right)^2 = 0 \leq \zeta^2. \tag{13}$$

In **Execution 2**, as $\mathcal{H} = \{1, \ldots, n-2f\} \cup \hat{S}$, we have

$$Q^{(\mathcal{H})}(\theta) = \frac{\mu(n-2f)}{4(n-f)}\theta^2 + \frac{\mu f}{4(n-f)}\left(\theta - \frac{2\zeta}{\mu\sqrt{\frac{n-f}{f}}}\right)^2 = \frac{\mu}{4}\left(\theta - \frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right)^2 + \frac{n-2f}{n-f}\cdot\frac{\zeta^2}{\mu}.$$

Therefore,

$$\nabla Q^{(\mathcal{H})}(\theta) = \frac{\mu}{2}\left(\theta - \frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right).$$

Thus,

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\nabla Q^{(i)}(\theta) - \nabla Q^{(\mathcal{H})}(\theta)\right\|^2 = \frac{1}{n-f}\sum_{i=1}^{n-2f}\left(\frac{\mu}{2}\theta - \frac{\mu}{2}\left(\theta - \frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right)\right)^2$$
$$+ \frac{1}{n-f}\sum_{i\in\hat{S}}\left(\frac{\mu}{2}\left(\theta - \frac{2\zeta}{\mu}\sqrt{\frac{n-f}{f}}\right) - \frac{\mu}{2}\left(\theta - \frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right)\right)^2$$

Upon simplifying the RHS above we obtain that

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\nabla Q^{(i)}(\theta) - \nabla Q^{(\mathcal{H})}(\theta)\right\|^2 = \frac{f(n-2f)}{(n-f)^2}\zeta^2 + \frac{(n-2f)^2}{(n-f)^2}\zeta^2 = \frac{n-2f}{n-f}\zeta^2 \leq \zeta^2. \tag{14}$$

Thus, due to (13) and (14), Assumption 4 is also satisfied in both executions.

Recall that in each execution of algorithm Π the output $\hat{\theta}$ satisfies the condition: $\mathbb{E}_{\Pi}\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \leq A$. Thus, from Execution 1, as $Q^* = 0$ and $Q^{(\mathcal{H})}(\theta) \coloneqq \frac{\mu}{4}\theta^2$, we have

$$\frac{\mu}{4}\,\mathbb{E}_{\Pi}\left[\hat{\theta}^2\right] \leq A. \tag{15}$$

Similarly, from Execution 2, as $Q^* = \frac{n-2f}{n-f}\cdot\frac{\zeta^2}{\mu}$ and $Q^{(\mathcal{H})}(\theta) \coloneqq \frac{\mu}{4}\left(\theta - \frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right)^2 + \frac{n-2f}{n-f}\cdot\frac{\zeta^2}{\mu}$, we obtain that

$$\frac{\mu}{4}\,\mathbb{E}_{\Pi}\left[\left(\hat{\theta} - \frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right)^2\right] \leq A \tag{16}$$

From Jensen's inequality, as $a^2 \leq (a-b+b)^2 \leq 2(a-b)^2 + 2b^2$, we have

$$\left(\frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right)^2 \leq 2\left(\frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}} - \hat{\theta}\right)^2 + 2\hat{\theta}^2. \tag{17}$$

Upon substituting from (15) and (16) in the above, we obtain that

$$\left(\frac{2\zeta}{\mu}\sqrt{\frac{f}{n-f}}\right)^2 \leq \frac{16}{\mu}\,A.$$

From above, we obtain that $A \geq \frac{f}{n-f}\cdot\frac{\zeta^2}{4\mu} \geq \frac{f}{n}\cdot\frac{\zeta^2}{4\mu}$, which implies (11), i.e.,

$$A \in \Omega\left(\frac{f}{n}\cdot\frac{\zeta^2}{\mu}\right)$$

This completes the proof of Theorem 1.

Second Case. Let $\mathcal{D}^{(i)} = \mathcal{D}$ for all $i \in \mathcal{H}$, where distribution $\mathcal{D}$ satisfies the following:

$$\mathbb{E}_{x\sim\mathcal{D}}[x] < \infty, \quad \text{and} \quad \mathbb{E}_{x\sim\mathcal{D}}\left[\left(x - \mathbb{E}_{x\sim\mathcal{D}}[x]\right)^2\right] \leq \frac{4\sigma^2}{\mu^2}.$$

By definition of $Q^{(i)}(\theta)$, we obtain that for all i,

$$Q^{(i)}(\theta) = \frac{\mu}{4}\,\mathbb{E}_{x\sim\mathcal{D}}\left[(\theta - x)^2\right], \quad \text{and thus,} \quad \nabla Q^{(i)}(\theta) = \frac{\mu}{2}\,(\theta - \mathbb{E}_{x\sim\mathcal{D}}[x]). \tag{18}$$

Thus, Assumption 1 holds true, i.e., $\nabla Q^{(i)}(\theta)$ is Lipschitz continuous, with $L = \mu$. Assumption 3 holds true due to the following:

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{D}}\left[\left(\nabla Q^{(i)}(\theta) - \nabla q(\theta, x)\right)^2\right] &= \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\frac{\mu}{2}\,(\theta - \mathbb{E}_{x\sim\mathcal{D}}[x]) - \frac{\mu}{2}\,(\theta - x)\right)^2\right] \\ &= \frac{\mu^2}{4}\,\mathbb{E}_{x\sim\mathcal{D}}\left[\left(x - \mathbb{E}_{x\sim\mathcal{D}}[x]\right)^2\right] = \sigma^2.\end{aligned}$$

From (18), we obtain that

$$Q^{(\mathcal{H})}(\theta) \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} Q^{(i)}(\theta) = \frac{\mu}{4}\mathbb{E}_{x\sim\mathcal{D}}\left[(\theta - x)^2\right]. \tag{19}$$

Thus, $Q^{(i)}$ and $Q^{(\mathcal{H})}$ are identical in this case, and Assumption 4 holds true trivially for $\zeta = 0$. From above we obtain that $\theta^* \coloneqq \arg\min_{\theta\in\mathbb{R}^d} Q^{(\mathcal{H})}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}[x]$, and thereby,

$$Q^* = Q^{(\mathcal{H})}(\theta^*) = \frac{\mu}{4}\,\mathbb{E}_{x\sim\mathcal{D}}\left[(x - \mathbb{E}_{x\sim\mathcal{D}}[x])^2\right]. \tag{20}$$

From (19) and (20) we obtain that

$$Q^{(\mathcal{H})}(\theta) - Q^* = \frac{\mu}{4}\,(\theta - \mathbb{E}_{x\sim\mathcal{D}}[x])^2. \tag{21}$$

Thus,

$$\left(\nabla Q^{(\mathcal{H})}(\theta)\right)^2 = \frac{\mu^2}{4}\,(\theta - \mathbb{E}_{x\sim\mathcal{D}}[x])^2 = \mu\left(Q^{(\mathcal{H})}(\theta) - Q^*\right).$$

Therefore, Assumption 2 also holds true.

We show that the accuracy of Algorithm Π reduces to that of an algorithm for estimating the mean of $\mathcal{D}$ by processing n batches of T points; $n-f$ batches sampled from $\mathcal{D}^T$ but the remainder f batches may contain arbitrary points. From (21) we obtain that

$$\mathbb{E}_\Pi\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] = \frac{\mu}{4}\mathbb{E}_\Pi\left[\left(\hat{\theta} - \mathbb{E}_{x\sim\mathcal{D}}[x]\right)^2\right].$$

Recall that we assume that $\mathbb{E}_\Pi\left[Q^{(\mathcal{H})}\left(\hat{\theta}\right) - Q^*\right] \leq A$. Thus, from above we have

$$A \geq \frac{\mu}{4}\,\mathbb{E}_\Pi\left[\left(\hat{\theta} - \mathbb{E}_{x\sim\mathcal{D}}[x]\right)^2\right]. \tag{22}$$

The above implies that Algorithm Π can estimate the mean of distribution $\mathcal{D}$ within a sqaured-error of $4A/\mu$. Recall that in algorithm Π, each honest worker $i \in \mathcal{H}$ computes T local stochastic gradients $\{\nabla q(\theta_t, x_t^{(i)})\ ;\ t = 1, \ldots, T\}$ where each element in the set of observations $X^{(i)} \coloneqq \{x_t^{(i)}\ ;\ t = 1, \ldots, T\}$ is i.i.d. from the distribution $\mathcal{D}^{(i)} = \mathcal{D}$. Recall that $\nabla q(\theta, x) = \frac{\mu}{2}(\theta - x)$. Therefore, given the value of μ, the set of parameter vectors $\{\theta_t\ ;\ t \in [T]\}$, we can recover the collection of random observations $\left\{X^{(i)}\ ;\ i \in \mathcal{H}\right\}$ where $X^{(i)} \sim \mathcal{D}^T$. Hence, it is obvious that the squared error for the mean estimation of $\mathcal{D}$ obtained upon executing Π cannot be smaller than that of an *optimal* (possibly randomized) robust mean estimator Π_{mean} that takes in as inputs n sets of random values $X^{(1)}, \ldots, X^{(n)}$ such that $X^{(i)} \sim \mathcal{D}^T$ for all $i \in \mathcal{H}$ and X_i for $i \in [n] \setminus \mathcal{H}$ may be an arbitrarily tuple of T points. Specifically, let $\hat{x} = \Pi_{mean}\left(X^{(1)}, \ldots, X^{(n)}\right)$, then

$$\frac{4A}{\mu} \geq \mathbb{E}_{\Pi_{mean}}\left[(\hat{x} - \mathbb{E}_{x\sim\mathcal{D}}[x])^2\right]. \tag{23}$$

We obtain in the following a lower bound on the squared-error $(\hat{x} - \mathbb{E}_{x\sim\mathcal{D}}[x])^2$ reasoning by indistinguishability of correct distributions under Huber's contamination model. Suppose there exists a distribution D' such that the variance of $\mathcal{D}'$ is also upper bounded by $\frac{4\sigma^2}{\mu^2}$ (same as that for $\mathcal{D}$) and $\mathrm{TV}(\mathcal{D}^T, \mathcal{D}'^T) \leq \frac{2f}{n}$. Then, by virtue of Proposition 1.7 of Diakonikolas and Kane (2022), no algorithm can reliably distinguish whether the sets of observations $\left\{X^{(i)} \ ; \ i \in \mathcal{H}\right\}$ were generated from $\mathcal{D}^T$ or $\mathcal{D}'^T$. Therefore,

$$\mathbb{E}_{\Pi_{mean}}\left[(\hat{x} - \mathbb{E}_{x\sim\mathcal{D}}[x])^2\right] \geq \frac{1}{4}\left(\mathbb{E}_{x\sim\mathcal{D}}[x] - \mathbb{E}_{x\sim\mathcal{D}'}[x]\right)^2. \tag{24}$$

We construct the following valid distributions $\mathcal{D}$ and $\mathcal{D}'$ to obtain a lower bound for the RHS in (24).

$$\text{Distribution } \mathcal{D}: \quad x = 0 \quad \text{with probability } 1.$$

$$\text{Distribution } \mathcal{D}': \quad x = \begin{cases} \frac{2\sigma}{\mu}\sqrt{\frac{Tn}{2f}} & \text{with probability } \frac{2f}{nT} \\ 0 & \text{with probability } 1 - \frac{2f}{nT} \end{cases}.$$

Validity of $\mathcal{D}$ and $\mathcal{D}'$. Note that $\mathbb{E}_{x\sim\mathcal{D}}[x] = 0$, and variance $\mathbb{E}_{x\sim\mathcal{D}}\left[(x - \mathbb{E}_{x\sim\mathcal{D}}[x])^2\right] = 0 \leq \frac{4\sigma^2}{\mu^2}$. Similarly, $\mathbb{E}_{x\sim\mathcal{D}'}[x] = \frac{2\sigma}{\mu}\sqrt{\frac{2f}{nT}}$ and variance $\mathbb{E}_{x\sim\mathcal{D}'}\left[(x - \mathbb{E}_{x\sim\mathcal{D}}[x])^2\right] = \frac{4\sigma^2}{\mu^2}(1 - \frac{2f}{nT}) \leq \frac{4\sigma^2}{\mu^2}$. Let 0^T denote a T-tuple with all elements equal to 0. If $X \sim \mathcal{D}'^T$ then

$$\Pr(X = 0^T) = \left(1 - \frac{2f}{nT}\right)^T.$$

As $\left(1 - \frac{2f}{nT}\right)^T \geq 1 - \frac{2f}{n}$, from above we obtain that $\mathrm{TV}(\mathcal{D}^T, \mathcal{D}'^T) = 1 - \left(1 - \frac{2f}{nT}\right)^T \leq \frac{2f}{n}$. Therefore, $\mathcal{D}$ and $\mathcal{D}'$ are indistinguishable.

Substituting the mean values of $\mathcal{D}$ and $\mathcal{D}'$ in (24) we obtain that

$$\mathbb{E}_{\Pi_{mean}}\left[(\hat{x} - \mathbb{E}_{x\sim\mathcal{D}}[x])^2\right] \geq \frac{1}{4}\left(\frac{2\sigma}{\mu}\sqrt{\frac{2f}{nT}}\right)^2 = \frac{2\sigma^2}{\mu^2}\cdot\frac{f}{nT}.$$

Substituting from above in (23) we have

$$\frac{4A}{\mu} \in \Omega\left(\frac{f}{n}\cdot\frac{\sigma^2}{\mu^2 T}\right). \tag{25}$$

As we have at most nT samples drawn from distribution $\mathcal{D}$, by the classical lower bound on statistical error rate (see Section 3.2 of Wu (2017)), we also have

$$\frac{4A}{\mu} \in \Omega\left(\frac{1}{n}\cdot\frac{\sigma^2}{\mu^2 T}\right). \tag{26}$$

Finally, combining (25) and (26) we obtain (12), i.e.,

$$A \in \Omega\left(\frac{f+1}{n}\cdot\frac{\sigma^2}{\mu T}\right).$$

Appendix B. Deferred Proofs for Theorem 3

Before proving a few simple lemmas that will be used in the subsequent proofs, let us introduce some useful notations.

Notation: We denote by $\mathcal{P}_t$ the history of nodes from steps 0 to t. Specifically, we define

$$\mathcal{P}_t \coloneqq \left\{\theta_0, \ldots, \theta_t;\ m_1^{(i)}, \ldots, m_{t-1}^{(i)}; i = 1, \ldots, n\right\}.$$

By convention, $\mathcal{P}_0 = \{\theta_0\}$. Furthermore, we denote by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{P}_t]$ the conditional expectation given the history $\mathcal{P}_t$, and by $\mathbb{E}[\cdot]$ the total expectation over the randomness of the algorithm; thus, $\mathbb{E}[\cdot] := \mathbb{E}_0[\cdots \mathbb{E}_T[\cdot]]$. Also denote by

$$R_t \coloneqq \mathrm{TM}\left(m_t^{(1)}, \ldots, m_t^{(n)}\right) \ , \tag{27}$$

the output of trimmed mean operation.

B.1 Preliminary Lemmas

Note that by decomposing the update rule computed by the server at step t, we can treat Algorithm 1 as DSGD with a momentum term and a bias $\gamma_t (R_t - \overline{m}_t)$. Specifically, we have

$$\theta_{t+1} = \theta_t - \gamma_t R_t = \theta_t - \gamma_t \overline{m}_t - \gamma_t (R_t - \overline{m}_t) \ . \tag{28}$$

The key to better understand the bias term in (28) is the analysis of TM, that attempts to robustly estimate the average of the honest momentums at every step. Using a recent result in (Allouah et al., 2023a), we can actually bound the bias $(R_t - \overline{m}_t)$ from above by the spread of honest nodes' momentums. Specifically, we have the following lemma.

Lemma 15 (Proposition 2 in (Allouah et al., 2023a)) *Let $n > 2f$. Consider Algorithm 1 and R_t as defined in (27). For any $t \geq 0$, we have*

$$\|R_t - \overline{m}_t\|^2 \leq \frac{\lambda}{n-f} \sum_{i \in \mathcal{H}} \left\|m_t^{(i)} - \overline{m}_t\right\|^2, \quad \textit{with} \quad \lambda = \frac{6f}{n-2f}\left(1 + \frac{f}{n-2f}\right) \ .$$

We also prove two useful lemmas.

Lemma 16 *Suppose Assumption 1, i.e., $Q^{(\mathcal{H})}$ is Lipschitz smooth with coefficient L. We denote $Q^* = \min_{\theta \in \mathbb{R}^d} Q^{(\mathcal{H})}(\theta)$. For all $\theta \in \mathbb{R}^d$, we have*

$$\|\nabla Q^{(\mathcal{H})}(\theta)\|^2 \leq 2L(Q^{(\mathcal{H})}(\theta) - Q^*) \ .$$

Proof As $Q^{(\mathcal{H})} \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} Q^{(i)}$, Assumption 1 implies that for all θ and θ',

$$\left\|\nabla Q^{(\mathcal{H})}(\theta) - \nabla Q^{(\mathcal{H})}(\theta')\right\| \leq L\left\|\theta - \theta'\right\|.$$

Thus, from Lipschitz inequality, for all $\theta, \theta' \in \mathbb{R}^d$ (Bottou et al., 2018),

$$Q^{(\mathcal{H})}(\theta') \leq Q^{(\mathcal{H})}(\theta) + \langle \nabla Q^{(\mathcal{H})}(\theta), \theta' - \theta \rangle + \frac{L}{2}\|\theta' - \theta\|^2 \ .$$

Consider an arbitrary $\theta \in \mathbb{R}^d$, and let $\theta' = \theta - \frac{1}{L}\nabla Q^{(\mathcal{H})}(\theta)$. Thus, from above we obtain that

$$\begin{aligned} Q^{(\mathcal{H})}\left(\theta - \frac{1}{L}\nabla Q^{(\mathcal{H})}(\theta)\right) &\leq Q^{(\mathcal{H})}(\theta) - \frac{1}{L}\|\nabla Q^{(\mathcal{H})}(\theta)\|^2 + \frac{1}{2L}\|\nabla Q^{(\mathcal{H})}(\theta)\|^2 \\ &= Q^{(\mathcal{H})}(\theta) - \frac{1}{2L}\|\nabla Q^{(\mathcal{H})}(\theta)\|^2 \ . \end{aligned}$$

As $Q^* = \min_{\mathbb{R}^d} Q^{(\mathcal{H})}$, we have

$$Q^* \leq Q^{(\mathcal{H})}\left(\theta - \frac{1}{L}\nabla Q^{(\mathcal{H})}(\theta)\right) \leq Q^{(\mathcal{H})}(\theta) - \frac{1}{2L}\|\nabla Q^{(\mathcal{H})}(\theta)\|^2 \ .$$

Rearranging the terms we obtain that

$$\|\nabla Q^{(\mathcal{H})}(\theta)\|^2 \leq 2L(Q^{(\mathcal{H})}(\theta) - Q^*) \ .$$

■

Lemma 17 *Consider an arbitrary non-empty set $S \subseteq \{1, \ldots, n\}$. For any set of $|S|$ real-valued vectors $\{x^{(i)}\}_{i\in S}$, we obtain that*

$$\frac{1}{|S|}\sum_{i\in S}\left\|x^{(i)} - \bar{x}\right\|^2 = \frac{1}{2\,|S|^2}\sum_{i,j\in S}\left\|x^{(i)} - x^{(j)}\right\|^2, \quad \textit{where} \quad \bar{x} = \frac{1}{|S|}\sum_{i\in S} x^{(i)} \ .$$

Proof

$$\begin{aligned} \frac{1}{|S|^2}\sum_{i,j\in S}\left\|x^{(i)} - x^{(j)}\right\|^2 &= \frac{1}{|S|^2}\sum_{i,j\in S}\left\|(x^{(i)} - \bar{x}) - (x^{(j)} - \bar{x})\right\|^2 \\ &= \frac{1}{|S|^2}\sum_{i,j\in S}\left[\left\|x^{(i)} - \bar{x}\right\|^2 + \left\|x^{(j)} - \bar{x}\right\|^2 + 2\left\langle x^{(i)} - \bar{x},\, x^{(j)} - \bar{x}\right\rangle\right] \\ &= \frac{2}{|S|}\sum_{i,j\in S}\left\|x^{(i)} - \bar{x}\right\|^2 + \frac{2}{|S|^2}\sum_{i\in S}\left\langle x^{(i)} - \bar{x},\, \sum_{j\in S}(x^{(j)} - \bar{x})\right\rangle \ . \end{aligned}$$

As $\sum_{j\in S}(x^{(j)} - \bar{x}) = 0$, from above we obtain that

$$\frac{1}{|S|^2}\sum_{i,j\in S}\left\|x^{(i)} - x^{(j)}\right\|^2 = \frac{2}{|S|}\sum_{i,j\in S}\left\|x^{(i)} - \bar{x}\right\|^2 \ .$$

■

B.2 Proof of the lemmas provided in the main paper

Lemma 5. *Suppose Assumption 1 holds true. Consider Algorithm 1 with $T \geq 2$, and $\gamma_t \leq 1/L$ for all $t \in \{0, \ldots, T-1\}$. Let λ be as defined in Lemma 15. Then, for all t, the following holds true*

$$\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t)\right] \leq -\frac{\gamma_t}{2}\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right] + \gamma_t \frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + \gamma_t\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t) - \overline{m}_t\right\|^2\right] \quad .$$

Proof Consider an arbitrary step t. Note that Assumption 1 implies the Lipschitz continuity of $\nabla Q^{(\mathcal{H})}(\theta)$ with coefficient L. Thus, we have

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

Substituting from Algorithm 1, $\theta_{t+1} = \theta_t - \gamma_t R_t$, we obtain that

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

Using the fact that $2\langle a,\, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, we obtain that

$$\begin{aligned} Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t) &\leq -\frac{\gamma_t}{2}\left\|R_t\right\|^2 - \frac{\gamma_t}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{\gamma_t}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{L\gamma_t^2}{2}\left\|R_t\right\|^2 \\ &= \left(\frac{L\gamma_t^2}{2} - \frac{\gamma_t}{2}\right)\left\|R_t\right\|^2 - \frac{\gamma_t}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{\gamma_t}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \quad . \end{aligned}$$

As $\gamma_t \leq 1/L$, we obtain that

$$\begin{aligned} Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t) &\leq -\frac{\gamma_t}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{\gamma_t}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \\ &\leq -\frac{\gamma_t}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \gamma_t\left\|R_t - \overline{m}_t\right\|^2 + \gamma_t\left\|\nabla Q^{(\mathcal{H})}(\theta_t) - \overline{m}_t\right\|^2 \quad . \end{aligned}$$

Using Lemma 15, we then obtain that

$$Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t) \leq -\frac{\gamma_t}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \gamma_t\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\left\|m_t^{(i)} - \overline{m}_t\right\|^2 + \gamma_t\left\|\nabla Q^{(\mathcal{H})}(\theta_t) - \overline{m}_t\right\|^2 \quad .$$

Taking the total expectation from both sides we then have

$$\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t)\right] \leq -\frac{\gamma_t}{2}\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right] + \gamma_t \frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + \gamma_t\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t) - \overline{m}_t\right\|^2\right] \quad ,$$

which is the desired result. ■

Lemma 6. *Suppose assumptions 3, and 4 hold true, and consider Algorithm 1 with $T \geq 2$. Then, for any $t \in \{0, \ldots, T-1\}$, the following holds true*

$$\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\Delta m_t^{(i)}\right\|^2\right] \leq \beta_t \frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\Delta m_{t-1}^{(i)}\right\|^2\right] + (1-\beta_t)\zeta^2 + (1-\beta_t)^2\sigma^2 \ .$$

Proof Consider two arbitrary correct nodes i and j. By the definition of the momentum vector from (7), we obtain that

$$\begin{aligned} m_t^{(i)} - m_t^{(j)} &= \beta_t(m_{t-1}^{(i)} - m_{t-1}^{(j)}) + (1-\beta_t)(g_t^{(i)} - g_t^{(j)}) \\ &= \beta_t(m_{t-1}^{(i)} - m_{t-1}^{(j)}) + (1-\beta_t)\left(\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right) \\ &\quad + (1-\beta_t)\left(g_t^{(i)} - \nabla Q^{(i)}(\theta_t) - g_t^{(j)} + \nabla Q^{(j)}(\theta_t)\right) \ . \end{aligned}$$

Taking the squared norm from both sides, we obtain that

$$\begin{aligned} &\left\|m_t^{(i)} - m_t^{(j)}\right\|^2 = \left\|\beta_t(m_{t-1}^{(i)} - m_{t-1}^{(j)}) + (1-\beta_t)\left(\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right)\right\|^2 \\ &+ \left\|(1-\beta_t)\left(g_t^{(i)} - \nabla Q^{(i)}(\theta_t) - g_t^{(j)} + \nabla Q^{(j)}(\theta_t)\right)\right\|^2 \\ &+ \left\langle \beta_t(m_{t-1}^{(i)} - m_{t-1}^{(j)}) + (1-\beta_t)\left(\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right), (1-\beta_t)\left(g_t^{(i)} - \nabla Q^{(i)}(\theta_t) - g_t^{(j)} + \nabla Q^{(j)}(\theta_t)\right)\right\rangle \ . \end{aligned}$$

Taking the conditional expectation $\mathbb{E}_t[\cdot]$ from both sides and noting that $\mathbb{E}_t\left[g_t^{(i)}\right] = \nabla Q^{(i)}(\theta_t)$ and $\mathbb{E}_t\left[g_t^{(j)}\right] = \nabla Q^{(j)}(\theta_t)$, we obtain that

$$\begin{aligned} \mathbb{E}_t\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right] &= \left\|\beta_t(m_{t-1}^{(i)} - m_{t-1}^{(j)}) + (1-\beta_t)\left(\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right)\right\|^2 \\ &\quad + (1-\beta_t)^2\mathbb{E}_t\left[\left\|g_t^{(i)} - \nabla Q^{(i)}(\theta_t)\right\|^2\right] + (1-\beta_t)^2\mathbb{E}_t\left[\left\|g_t^{(j)} - \nabla Q^{(j)}(\theta_t)\right\|^2\right] \ . \end{aligned}$$

Using Assumption 3, we then obtain that

$$\mathbb{E}_t\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right] \leq \left\|\beta_t(m_{t-1}^{(i)} - m_{t-1}^{(j)}) + (1-\beta_t)\left(\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right)\right\|^2 + 2(1-\beta_t)^2\sigma^2 \ .$$

By Jensen's inequality, we then have

$$\mathbb{E}_t\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right] \leq \beta_t\left\|m_{t-1}^{(i)} - m_{t-1}^{(j)}\right\|^2 + (1-\beta_t)\left\|\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right\|^2 + 2(1-\beta_t)^2\sigma^2 \ .$$

Taking total expectation and averaging over all possible $i, j \in \mathcal{H}$, we then obtain that

$$\begin{aligned} \frac{1}{(n-f)^2}\sum_{i,j\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right] &\leq \beta_t\frac{1}{(n-f)^2}\sum_{i,j\in\mathcal{H}}\mathbb{E}\left[\left\|m_{t-1}^{(i)} - m_{t-1}^{(j)}\right\|^2\right] \\ &+ (1-\beta_t)\frac{1}{(n-f)^2}\sum_{i,j\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right\|^2\right] + 2(1-\beta_t)^2\sigma^2 \ . \end{aligned}$$

Using Lemma 17, we then obtain that

$$\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)}-\overline{m}_t\right\|^2\right]\leq\beta_t\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_{t-1}^{(i)}-\overline{m}_{t-1}\right\|^2\right]$$
$$+(1-\beta_t)\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla Q^{(i)}(\theta_t)-\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right]+(1-\beta_t)^2\sigma^2\ .$$

By Assumption 4, we then obtain that

$$\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)}-\overline{m}_t\right\|^2\right]\leq\beta_t\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_{t-1}^{(i)}-\overline{m}_{t-1}\right\|^2\right]+(1-\beta_t)\zeta^2+(1-\beta_t)^2\sigma^2\ .$$

This is the desired result. ■

Lemma 7. *Suppose assumptions 1, 3, and 4 hold true, consider Algorithm 1 with $T\geq 2$, and λ as defined in Lemma 5. Then for any $t\in\{0,\ldots,T-1\}$, the following holds true*

$$\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right]\leq\beta_{t+1}^2\left(1+4\gamma_tL+3\gamma_t^2L^2\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right]+(1-\beta_{t+1})^2\frac{\sigma^2}{n-f}$$
$$+3\beta_{t+1}^2(\gamma_t^2L^2+\gamma_tL)\left(\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\Delta m_t^{(i)}\right\|^2\right]+\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right]\right)\ .$$

Proof We recall that at any round $t\in\{0,\ldots,T-1\}$ and any worker $i\in\mathcal{H}$, the momentum $m_t^{(i)}$ is computed as follows

$$m_t^{(i)}=\beta_tm_{t-1}^{(i)}+(1-\beta_t)g_t^{(i)}\ .$$

Hence, we have

$$\delta_{t+1}=\overline{m}_{t+1}-\nabla Q^{(\mathcal{H})}(\theta_{t+1})=\beta_{t+1}\overline{m}_t+(1-\beta_{t+1})\overline{g}_{t+1}-\nabla Q^{(\mathcal{H})}(\theta_{t+1})\ .$$

where for any $t\in\{0,\ldots,T-1\}$, $\overline{m}_t\coloneqq\frac{1}{(n-f)}\sum_{i\in\mathcal{H}}m_t^{(i)}$ and $\overline{g}_t\coloneqq\frac{1}{(n-f)}\sum_{i\in\mathcal{H}}g_t^{(i)}$.

Adding and subtracting $\beta_{t+1}\nabla Q^{(\mathcal{H})}(\theta_t)$, we obtain that

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

Now by Assumption 3, we have $\mathbb{E}_{t+1}\left[\overline{g}_{t+1}\right]=\nabla Q^{(\mathcal{H})}(\theta_{t+1})$ and $\mathbb{E}_{t+1}\left[\left\|\overline{g}_{t+1}-\nabla Q^{(\mathcal{H})}(\theta_{t+1})\right\|^2\right]\leq\frac{\sigma^2}{n-f}$. Therefore,

$$\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right]\leq\beta_{t+1}^2\left\|\delta_t+\nabla Q^{(\mathcal{H})}(\theta_t)-\nabla Q^{(\mathcal{H})}(\theta_{t+1})\right\|^2+(1-\beta_{t+1})^2\frac{\sigma^2}{n-f}\ .$$

Now as $(a+b)^2 \leq (1+c)a^2 + (1+1/c)b^2$ for any c, we obtain that

$$\mathbb{E}_{t+1}\left[\|\delta_{t+1}\|^2\right] \leq \beta_{t+1}^2(1+\gamma_t L)\|\delta_t\|^2 + \beta_{t+1}^2(1+\frac{1}{\gamma_t L})\left\|\nabla Q^{(\mathcal{H})}(\theta_t) - \nabla Q^{(\mathcal{H})}(\theta_{t+1})\right\|^2 + (1-\beta_{t+1})^2\frac{\sigma^2}{n-f} \ .$$

From Assumption 1, we have $\left\|\nabla Q^{(\mathcal{H})}(\theta_t) - \nabla Q^{(\mathcal{H})}(\theta_{t+1})\right\| \leq L\left\|\theta_t - \theta_{t+1}\right\|$. Using this above, we obtain that

$$\mathbb{E}_{t+1}\left[\|\delta_{t+1}\|^2\right] \leq \beta_{t+1}^2(1+\gamma_t L)\|\delta_t\|^2 + \beta_{t+1}^2\left(1+\frac{1}{\gamma_t L}\right)L^2\left\|\theta_t - \theta_{t+1}\right\|^2 + (1-\beta_{t+1})^2\frac{\sigma^2}{n-f} \ . \tag{29}$$

Now recall that $\theta_t - \theta_{t+1} = \gamma_t R_t$. Therefore,

$$\begin{aligned}
\left\|\theta_t - \theta_{t+1}\right\|^2 &= \gamma_t^2\|R_t\|^2 \\
&= \gamma_t^2\left\|R_t - \overline{m}_t + \overline{m}_t - \nabla Q^{(\mathcal{H})}(\theta_t) + \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \\
&\leq 3\gamma_t^2\|R_t - \overline{m}_t\|^2 + 3\gamma_t^2\left\|\overline{m}_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + 3\gamma_t^2\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \\
&\leq 3\gamma_t^2\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\left\|m_t^{(i)} - \overline{m}_t\right\|^2 + 3\gamma_t^2\left\|\overline{m}_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + 3\gamma_t^2\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \ ,
\end{aligned}$$

where in the last inequality we used 15. Combining this with (29), we obtain that

$$\begin{aligned}
\mathbb{E}_{t+1}\left[\|\delta_{t+1}\|^2\right] &\leq \beta_{t+1}^2(1+\gamma_t L)\|\delta_t\|^2 + (1-\beta_{t+1})^2\frac{\sigma^2}{n-f} \\
&+ \beta_{t+1}^2(1+\frac{1}{\gamma_t L})L^2\left(3\gamma_t^2\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\left\|m_t^{(i)} - \overline{m}_t\right\|^2 + 3\gamma_t^2\|\delta_t\|^2 + 3\gamma_t^2\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right) \ .
\end{aligned}$$

Rearranging the terms and taking the total expectation, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\|\delta_{t+1}\|^2\right] &\leq \beta_{t+1}^2(1+4\gamma_t L + 3\gamma_t^2 L^2)\,\mathbb{E}\left[\|\delta_t\|^2\right] + (1-\beta_{t+1})^2\frac{\sigma^2}{n-f} \\
&+ 3\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L)\left(\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right]\right) \ .
\end{aligned}$$

This is the desired result. ■

Lemma 8. *Suppose assumptions 1, 2, 3, and 4 hold true. Consider Algorithm 1 with $T \geq 2$ and a set of parameters such that $t \in \{0, \cdots, T\}$, $\gamma_t \leq \frac{1}{18L}$, and $1-\beta_{t+1} = 18\gamma_t L$. Finally, let $(V_t)_{t\geq 0}$ be as defined in (9) and λ as defined in Lemma 15. Then the following holds true*

$$V_{t+1} \leq \left(1 - \frac{\mu\gamma_t}{3}\right)V_t + 27L\left(\lambda + \frac{1}{n-f}\right)\sigma^2\gamma_t^2 + \frac{3}{2}\lambda\zeta^2\gamma_t \ .$$

Proof Consider an arbitrary $t \in \{0, \dots, T-1\}$. Combining Lemmas 6, 7, and 5, we obtain that

$$\begin{aligned}
V_{t+1} &= \mathbb{E}\left[Q^{(\mathcal{H})}\left(\theta_{t+1}\right)\right] - Q^* + \rho\,\mathbb{E}\left[\|\delta_{t+1}\|^2\right] + \rho\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_{t+1}^{(i)} - \overline{m}_{t+1}\right\|^2\right] \\
&\leq \mathbb{E}\left[Q^{(\mathcal{H})}\left(\theta_t\right)\right] - Q^* - \frac{\gamma_t}{2}\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right] + \gamma_t\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] \\
&\quad + \gamma_t\,\mathbb{E}\left[\|\delta_t\|^2\right] + \rho\beta_{t+1}^2(1 + 4\gamma_t L + 3\gamma_t^2 L^2)\,\mathbb{E}\left[\|\delta_t\|^2\right] + \rho(1-\beta_{t+1})^2\frac{\sigma^2}{n-f} \\
&\quad + 3\rho\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L)\left(\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right]\right) \\
&\quad + \rho\lambda\beta_{t+1}\frac{1}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + \rho\lambda(1-\beta_{t+1})\zeta^2 + \rho\lambda(1-\beta_{t+1})^2\sigma^2 \ .
\end{aligned}$$

Re-arranging the terms, we obtain that

$$\begin{aligned}
V_{t+1} &\leq \mathbb{E}\left[Q^{(\mathcal{H})}\left(\theta_t\right)\right] - Q^* + \left(-\frac{\gamma_t}{2} + 3\rho\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L)\right)\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right] \\
&\quad + \left(\gamma_t + 3\rho\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L) + \rho\beta_{t+1}\right)\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] \\
&\quad + \left(\gamma_t + \rho\beta_{t+1}^2(1 + 4\gamma_t L + 3\gamma_t^2 L^2)\right)\mathbb{E}\left[\|\delta_t\|^2\right] \\
&\quad + \rho\lambda(1-\beta_{t+1})\zeta^2 + \rho\lambda(1-\beta_{t+1})^2\sigma^2 + \rho(1-\beta_{t+1})^2\frac{\sigma^2}{n-f} \ . \qquad (30)
\end{aligned}$$

We denote,

$$\begin{aligned}
A &\coloneqq -\frac{\gamma_t}{2} + 3\rho\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L), \\
B &\coloneqq \gamma_t + 3\rho\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L) + \rho\beta_{t+1}, \\
C &\coloneqq \gamma_t + \rho\beta_{t+1}^2(1 + 4\gamma_t L + 3\gamma_t^2 L^2) \\
D &\coloneqq \rho\lambda(1-\beta_{t+1})\zeta^2 + \rho\lambda(1-\beta_{t+1})^2\sigma^2 + \rho(1-\beta_{t+1})^2\frac{\sigma^2}{n-f} \ .
\end{aligned}$$

Substituting from above in (30) we obtain that

$$\begin{aligned}
V_{t+1} &\leq \mathbb{E}\left[Q^{(\mathcal{H})}\left(\theta_t\right)\right] - Q^* + A\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right] \\
&\quad + B\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + C\,\mathbb{E}\left[\|\delta_t\|^2\right] + D \ .
\end{aligned}$$

Now, we separately analyse the terms A, B, C and D below by using the following,

$$\rho = \frac{1}{12L},\ \gamma_t \leq \frac{1}{18L},\ \text{ and } \ 1 - \beta_{t+1} = 18\gamma_t L \ . \qquad (31)$$

Note that the condition on γ_t above follows

Term A. Using the facts that $\rho = 1/12L$, $\gamma_t \leq 1/18L \leq 1/3L$ and that $\beta_{t+1}^2 < 1$, we obtain that

$$\begin{aligned} A = -\frac{\gamma_t}{2} + 3\rho\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L) &\leq -\frac{\gamma_t}{2} + 3\rho(\gamma_t^2 L^2 + \gamma_t L) \\ &\leq -\frac{\gamma_t}{2} + \frac{1}{4L}(\frac{\gamma_t L}{3} + \gamma_t L) = -\frac{\gamma_t}{6} \ . \end{aligned} \tag{32}$$

Term B. We obtain that

$$B = \gamma_t + 3\rho\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L) + \rho\beta_{t+1} = \rho(12\gamma_t L + 3\beta_{t+1}^2(\gamma_t^2 L^2 + \gamma_t L) + \beta_{t+1}) \ .$$

Noting that $\beta_{t+1} \leq 1$, $\beta_{t+1} = 1 - 18\gamma_t L$ and $\gamma_t \leq 1/18L \leq 1/12L$ we obtain that

$$B \leq \rho\left(12\gamma_t L + 3\gamma_t L + \frac{\gamma_t L}{4} + (1 - 18\gamma_t L)\right) \leq \rho\left(1 - \frac{11\gamma_t L}{4}\right) \leq \rho\left(1 - \frac{\gamma_t L}{3}\right) \leq \rho\left(1 - \frac{\mu\gamma_t}{3}\right) \ ,$$

where in the last inequality we used $\mu \leq L$.

Term C. Using the facts that $\beta_{t+1} < 1$ and $\rho = 1/12L$, we obtain that

$$\begin{aligned} C \coloneqq \gamma_t + \rho\beta_{t+1}^2(1 + 4\gamma_t L + 3\gamma_t^2 L^2) &\leq \rho\left(\frac{\gamma_t}{\rho} + \beta_{t+1} + 4\gamma_t L + 3\gamma_t^2 L^2\right) \\ &= \rho\left(12\gamma_t L + \beta_{t+1} + 4\gamma_t L + 3\gamma_t^2 L^2\right) \ . \end{aligned}$$

Using the fact $\gamma_t \leq 1/18L \leq 1/12L$ we then have

$$C \leq \rho\left(16\gamma_t L + \frac{\gamma_t L}{4} + (1 - 18\gamma_t L)\right) \leq \rho\left(1 - \frac{7\gamma_t L}{4}\right) \leq \rho\left(1 - \frac{\gamma_t L}{3}\right) \leq \rho\left(1 - \frac{\mu\gamma_t}{3}\right) \ , \tag{33}$$

Term D.

$$\begin{aligned} D &= \rho\lambda(1 - \beta_{t+1})\zeta^2 + \rho\lambda(1 - \beta_{t+1})^2\sigma^2 + \rho(1 - \beta_{t+1})^2\frac{\sigma^2}{n - f} \\ &= \frac{3}{2}\gamma_t\lambda\zeta^2 + 27\gamma_t^2 L\left(\lambda + \frac{1}{n - f}\right)\sigma^2 \ . \end{aligned}$$

Combining all, we obtain that

$$\begin{aligned} V_{t+1} \leq \mathbb{E}\left[Q^{(\mathcal{H})}(\theta_t)\right] - Q^* - \frac{\gamma_t}{6}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2\right] + \left(1 - \frac{\mu\gamma_t}{3}\right)\rho\,\mathbb{E}\left[\|\delta_t\|^2\right] \\ + \left(1 - \frac{\mu\gamma_t}{3}\right)\rho\frac{\lambda}{n - f}\sum_{i \in \mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right] + \frac{3}{2}\gamma_t\lambda\zeta^2 + 27\gamma_t^2 L\left(\lambda + \frac{1}{n - f}\right)\sigma^2 \ . \end{aligned}$$

Recall from Assumption 2 that $\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \geq 2\mu\left(Q^{(\mathcal{H})}(\theta_t) - Q^*\right)$. Therefore,

$$V_{t+1} \leq \left(1 - \frac{\mu\gamma_t}{3}\right)\left(\mathbb{E}\left[Q^{(\mathcal{H})}\left(\theta_t\right)\right] - Q^* + \rho\,\mathbb{E}\left[\|\delta_t\|^2\right] + \rho\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|m_t^{(i)} - \overline{m}_t\right\|^2\right]\right)$$
$$+ \frac{3}{2}\gamma_t\lambda\zeta^2 + 27\gamma_t^2 L\left(\lambda + \frac{1}{n-f}\right)\sigma^2$$
$$= \left(1 - \frac{\mu\gamma_t}{3}\right)V_t + 27L\left(\lambda + \frac{1}{n-f}\right)\sigma^2\gamma_t^2 + \frac{3}{2}\lambda\zeta^2\gamma_t \ .$$

■

Lemma 9. *Let a, b, c, d be positive real values with $a < b$, and let $T \geq 2$ be a positive integer. Let $(\gamma_0, \ldots, \gamma_{T-1})$ and $(r_0, \ldots, r_T)$ be real valued sequences such that for all $t \in \{0, \ldots, T-1\}$,*

$$r_{t+1} \leq (1 - a\gamma_t)r_t + c\gamma_t^2 + d\gamma_t \ . \tag{34}$$

Consider the following two cases:

- ***Case 1:*** *$T \leq b/a$ and $\gamma_t = 1/b,\ \forall t \in \{0, \ldots, T-1\}$.*
- ***Case 2:*** *$T > b/a$ and for $s = 2b/a$ and $t_0 = \lceil T/2 \rceil$,*

$$\gamma_t = \begin{cases} \frac{1}{b} & , \quad \text{if } t < t_0 \\ \frac{2}{a(s+t-t_0+1)} & , \quad \text{otherwise} \end{cases} \ .$$

In both Case 1 and Case 2, we have

$$r_T \leq r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{18c}{a^2 T} + \frac{3d}{a} \ . \tag{35}$$

Proof Our technique closely follows that of the proof of Lemma 3 in Khaled and Richtárik (2023), which itself build upon the analysis presented in Stich (2019).

Case 1. Here, $T \leq \frac{b}{a}$ and $\gamma_t = \gamma = 1/b,\ \forall t \in \{0, \ldots, T-1\}$. Thus, as $a < b$, note that $(1 - a\gamma) \in (0,\, 1)$. Then, by applying recursion on (34) we obtain that for all $t \in [T]$,

$$r_{t+1} \leq (1-a\gamma)^{t+1} r_0 + \gamma^2 c\sum_{\tau=0}^{t}(1-a\gamma)^\tau + \gamma d\sum_{\tau=0}^{t}(1-a\gamma)^\tau \leq (1-a\gamma)^{t+1} r_0 + \frac{\gamma c}{a} + \frac{d}{a} \ ,$$

where the last inequation comes from the fact that $\sum_{\tau=0}^{t}(1-a\gamma)^\tau \leq \sum_{\tau=0}^{\infty}(1-a\gamma)^\tau = \frac{1}{1-(1-a\gamma)}$. As $(1-x) \leq \exp(-x)$ for all $x \geq 0$, the above implies that for all $t \in [T]$,

$$r_{t+1} \leq r_0 \exp\left(-a\gamma(t+1)\right) + \frac{\gamma c}{a} + \frac{d}{a} \ .$$

Substituting $\gamma = 1/b$ in the above, we obtain that for all $t \in \{0, \ldots, T-1\}$,

$$r_{t+1} \leq r_0 \exp\left(-\frac{a(t+1)}{b}\right) + \frac{c}{ab} + \frac{d}{a} \ . \tag{36}$$

Recall that in this particular case, we assume $Ta \leq b$. Thus, $\frac{1}{b} \leq \frac{1}{Ta}$ and we obtain that for all $t \in \{0, \ldots, T-1\}$,

$$r_{t+1} \leq r_0 \exp\left(-\frac{a(t+1)}{b}\right) + \frac{c}{a^2 T} + \frac{d}{a} \ .$$

Substituting $t = (T-1)$ in the above yields

$$r_T \leq r_0 \exp\left(-\frac{aT}{b}\right) + \frac{c}{a^2 T} + \frac{d}{a} \ .$$

As $T > T/2$ and $a, c, d > 0$, we have,

$$r_T \leq r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{18c}{a^2 T} + \frac{3d}{a} \ .$$

Case 2. $T > b/a$ and for $s = 2b/a$ and $t_0 = \lceil T/2 \rceil$,

$$\gamma_t = \begin{cases} \frac{1}{b} & , \quad \text{if } t < t_0 \\ \\ \frac{2}{a(s+t-t_0+1)} & , \quad \text{otherwise} \end{cases} .$$

First, we consider the sub-case when $t < t_0$. As $\gamma_t = \gamma = 1/b$ for all $t < t_0$, (36) holds true for any $t < t_0$. Thus, upon substituting $t = t_0 - 1$ in (36) we obtain that

$$r_{t_0} \leq r_0 \exp\left(-\frac{at_0}{b}\right) + \frac{c}{ab} + \frac{d}{a} \ .$$

As $t_0 \geq \frac{T}{2}$, the above implies that

$$r_{t_0} \leq r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{c}{ab} + \frac{d}{a} \ . \tag{37}$$

Next, we consider the sub-case when $t_0 \leq t \leq T-1$. For an arbitrary such t, upon substituting $\gamma_t = \frac{2}{a(s+t-t_0+1)}$ in (34) we obtain that

$$\begin{aligned} r_{t+1} &\leq (1 - a\gamma_t) r_t + c\,\gamma_t^2 + d\gamma_t = \left(1 - \frac{2}{s+t-t_0+1}\right) r_t + \frac{4c}{a^2(s+t-t_0+1)^2} + \frac{2d}{a(s+t-t_0+1)} \\ &= \left(\frac{s+t-t_0-1}{s+t-t_0+1}\right) r_t + \frac{4c}{a^2(s+t-t_0+1)^2} + \frac{2d}{a(s+t-t_0+1)} \ . \end{aligned}$$

Multiplying both sides above by $(s+t-t_0+1)^2$ we obtain that

$$\begin{aligned} (s+t-t_0+1)^2 r_{t+1} &\leq (s+t-t_0-1)(s+t-t_0+1) r_t + \frac{4c}{a^2} + \frac{2d}{a}(s+t-t_0+1) \\ &= ((s+t-t_0)^2 - 1) r_t + \frac{4c}{a^2} + \frac{2d}{a}(s+t-t_0+1) \\ &\leq (s+t-t_0)^2 r_t + \frac{4c}{a^2} + \frac{2d}{a}(s+t-t_0+1) \ . \end{aligned}$$

By rewriting $(s+t-t_0+1)$ as $(s+t+1-t_0)$ in the above, we have

$$(s+t+1-t_0)^2 r_{t+1} \leq (s+t-t_0)^2 r_t + \frac{4c}{a^2} + \frac{2d}{a}(s+t+1-t_0) \ .$$

Recall that t above is an arbitrary integer in $[t_0, T-1]$. Thus, the inequality holds true for all $t \in [t_0, T-1]$. Therefore, upon summing both the sides over all $t \in [t_0, T-1]$, we have

$$\sum_{t=t_0}^{T-1}(s+t+1-t_0)^2 r_{t+1} \leq \sum_{t=t_0}^{T-1}(s+t-t_0)^2 r_t + \sum_{t=t_0}^{T-1}\frac{4c}{a^2} + \frac{2d}{a}\sum_{t=t_0}^{T-1}(s+t+1-t_0) \ .$$

Upon expanding the LHS and the first-term in the RHS we obtain that

$$\begin{aligned}(s+T-t_0)^2 \, r_T &\leq s^2 r_{t_0} + \sum_{t=t_0}^{T-1}\frac{4c}{a^2} + \frac{2d}{a}\sum_{t=t_0}^{T-1}(s+t+1-t_0) \\ &= s^2 r_{t_0} + \frac{4c}{a^2}(T-t_0) + \frac{d}{a}(T-t_0)(T-t_0+1+2s) \ .\end{aligned}$$

Therefore,

$$r_T \leq \frac{s^2}{(s+T-t_0)^2}\, r_{t_0} + \frac{4c(T-t_0)}{a^2(s+T-t_0)^2} + \frac{d(T-t_0)(T-t_0+1+2s)}{a(s+T-t_0)^2} \ .$$

As $T-t_0 \leq s+T-t_0$ and $T-t_0+1+2s \leq 2(s+T-t_0)$, from above we obtain that

$$r_T \leq \frac{s^2}{(s+T-t_0)^2}\, r_{t_0} + \frac{4c}{a^2(T-t_0)} + \frac{2d}{a} \ .$$

As $t_0 \leq \frac{2T}{3}$, we have $T-t_0 \geq \frac{T}{3}$. Using this above we obtain that

$$r_T \leq \frac{s^2}{(s+T-t_0)^2} r_{t_0} + \frac{12c}{a^2 T} + \frac{2d}{a} \ .$$

Substituting from (37) in the above, we obtain that

$$r_T \leq \frac{s^2}{(s+T-t_0)^2}\left(r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{c}{ab} + \frac{d}{a}\right) + \frac{12c}{a^2 T} + \frac{2d}{a} \ .$$

As $s \leq s+T-t_0$, the above implies that

$$r_T \leq \frac{s}{s+T-t_0}\left(\frac{c}{ab}\right) + r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{d}{a} + \frac{12c}{a^2 T} + \frac{2d}{a} \ .$$

Using the fact that $s+T-t_0 \geq \frac{T}{3}$ above we have

$$r_T \leq \frac{3s}{T}\left(\frac{c}{ab}\right) + r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{d}{a} + \frac{12c}{a^2 T} + \frac{2d}{a} \ .$$

Substituting $s = \frac{2b}{a}$ proves (35), i.e., we obtain that

$$r_T \leq r_0 \exp\left(-\frac{aT}{2b}\right) + \frac{18c}{a^2 T} + \frac{3d}{a} \ .$$

■

B.3 Final step to prove Theorem 3

Proof We now apply Lemma 9 to the recursion of Lemma 8, for $a = \frac{\mu}{3}$, $b = 18L$, $c = 27L\left(\lambda + \frac{1}{n-f}\right)\sigma^2$ and $d = \frac{3}{2}\lambda\zeta^2$. Choosing the learning rates as specified in Lemma 9, we then obtain that

$$V_T \leq \exp(-\frac{\mu T}{108L})V_0 + \frac{4374L\left(\lambda + \frac{1}{n-f}\right)\sigma^2}{T\mu^2} + \frac{9\lambda\zeta^2}{2\mu} \ . \tag{38}$$

As $m_0^{(i)} = 0$ for all $i \in \mathcal{H}$, we have

$$\frac{1}{n-f}\sum_{i\in\mathcal{H}}\left\|m_0^{(i)} - \overline{m}_0\right\|^2 = 0 \ ,$$

and

$$\|\delta_0\|^2 = \left\|\nabla Q^{(\mathcal{H})}(\theta_0) - \overline{m}_0\right\|^2 = \left\|\nabla Q^{(\mathcal{H})}(\theta_0)\right\|^2 \leq 2L\left(Q^{(\mathcal{H})}(\theta_0) - Q^*\right) \ ,$$

where in the last inequality we used Lemma 16. Thus,

$$V_0 = Q^{(\mathcal{H})}(\theta_0) - Q^* + \frac{1}{12L}\|\delta_0\|^2 + \frac{1}{12L}\frac{\lambda}{n-f}\sum_{i\in\mathcal{H}}\left\|m_0^{(i)} - \overline{m}_0\right\|^2 \leq \frac{7}{6}\left(Q^{(\mathcal{H})}(\theta_0) - Q^*\right) \ .$$

Combining this with (38), we obtain that

$$V_T \leq \frac{7}{6}\left(Q^{(\mathcal{H})}(\theta_0) - Q^*\right)\cdot\exp(-\frac{\mu T}{108L}) + \frac{4374L\left(\lambda + \frac{1}{n-f}\right)\sigma^2}{T\mu^2} + \frac{9\lambda\zeta^2}{2\mu} \ .$$

By the definition of V_t in (9), we have $\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_T) - Q^*\right] \leq V_T$. Therefore,

$$\mathbb{E}\left[Q^{(\mathcal{H})}(\theta_T) - Q^*\right] \leq \frac{7}{6}\left(Q^{(\mathcal{H})}(\theta_0) - Q^*\right)\cdot\exp(-\frac{\mu T}{108L}) + \frac{4374L\left(\lambda + \frac{1}{n-f}\right)\sigma^2}{T\mu^2} + \frac{9\lambda\zeta^2}{2\mu} \ .$$

This is the desired result. ■

B.4 Proof of Corollary 4

As $n \geq (2+\nu)f$, we have

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

Rearranging the terms we have

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

Therefore,

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

As $\nu > 0$ is a constant, we have

$$\lambda = \frac{6f}{n-2f}\left(1+\frac{f}{n-2f}\right) \leq \frac{2+\nu}{\nu} \cdot \frac{6f}{n}\left(1+\frac{2+\nu}{\nu} \cdot \frac{f}{n}\right) \in \mathcal{O}\left(\frac{f}{n}\right) \ . \tag{39}$$

Theorem 3 then implies that

$$Q^{(\mathcal{H})}(\theta_T) - Q^* \in \mathcal{O}\left(Q_0 \cdot \exp(-\frac{\mu T}{108L}) + \frac{L\left(\lambda + \frac{1}{n-f}\right)\sigma^2}{T\mu^2} + \frac{\lambda\zeta^2}{\mu}\right) \ ,$$

Combining this with (39), and noting that $\frac{1}{n-f} \leq \frac{2}{n}$, we have

$$Q^{(\mathcal{H})}(\theta_T) - Q^* \in \mathcal{O}\left(Q_0 \cdot \exp(-\frac{\mu T}{108L}) + \frac{L\sigma^2}{\mu^2 T}\left(\frac{1}{n}+\frac{f}{n}\right) + \frac{f}{n}\frac{\zeta^2}{\mu}\right) \ . \tag{40}$$

Now note that as $T \to \infty$, the first two terms converge to 0. More precisely, for any $\varepsilon > 0$, setting

$$T = \max\left\{\frac{2L\sigma^2}{\mu^2\varepsilon}\left(\frac{f+1}{n}\right), 108\frac{L}{\mu}\log\frac{2Q_0}{\varepsilon}\right\} \leq \frac{2L\sigma^2}{\mu^2\varepsilon}\left(\frac{f+1}{n}\right) + 108\frac{L}{\mu}\log\frac{2Q_0}{\varepsilon},$$

we obtain that

$$Q_0 \cdot \exp(-\frac{\mu T}{108L}) + \frac{L\sigma^2}{\mu^2 T}\left(\frac{1}{n}+\frac{f}{n}\right) \leq \varepsilon.$$

Combing this with (40), we have

$$Q^{(\mathcal{H})}(\theta_T) - Q^* \in \mathcal{O}\left(\frac{f}{n} \cdot \frac{\zeta^2}{\mu} + \varepsilon\right) \ ,$$

for

$$T \in \mathcal{O}\left(\frac{L\sigma^2}{\mu^2\varepsilon}\left(\frac{f+1}{n}\right) + \frac{L}{\mu}\log\frac{Q_0}{\varepsilon}\right),$$

which is the desired result.

Appendix C. Proof of Theorem 12

Let us denote $R_t \coloneqq \mathrm{TM}^{(f)}\left(G_t^{(1)}, \ldots, G_t^{(n)}\right)$ and $\bar{G}_t \coloneqq \sum_{i\in\mathcal{H}} G_t^{(i)}$. By Proposition 2 of (Allouah et al., 2023a), we have

$$\left\|R_t - \bar{G}_t\right\|^2 \leq \lambda \frac{1}{n-f}\sum_{i\in\mathcal{H}}\left\|G_t^{(i)} - \bar{G}_t\right\|^2, \quad \text{where} \quad \lambda = \frac{6f}{n-2f}\left(1+\frac{f}{n-2f}\right) \ . \tag{41}$$

Similarly, for each $i \in \mathcal{H}$ and $\theta_t \in \mathbb{R}^d$, we have

$$\left\|G_t^{(i)} - \nabla Q^{(i)}(\theta_t)\right\|^2 \leq \lambda' \frac{1}{m-b} \sum_{j \in \mathcal{S}_h^{(i)}} \left\|\nabla q(x^{(i,j)}, \theta_t) - \nabla Q^{(i)}(\theta_t)\right\|^2,$$

where $\lambda' = \frac{6b}{m-2b}\left(1 + \frac{m}{m-2b}\right)$. Therefore, by Assumption 3, we have

$$\left\|G_t^{(i)} - \nabla Q^{(i)}(\theta_t)\right\|^2 \leq \lambda' \sigma^2. \tag{42}$$

We now prove a few useful lemmas.

Lemma 18 *Suppose Assumption 1. Consider Algorithm 2 with $T \geq 2$, and $\gamma \leq 1/L$. Then, for all $t \in \{0, \ldots, T-1\}$, the following holds true:*

$$Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t) \leq -\frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{\gamma}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2.$$

Proof Consider an arbitrary step t. Note that Assumption 1 implies L-Lipschitz continuity of $\nabla Q^{(\mathcal{H})}(\theta)$. Thus, we have

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

Substituting from Algorithm 2, $\theta_{t+1} = \theta_t - \gamma R_t$, we obtain that

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

Using the fact that $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, we obtain that

$$\begin{aligned} Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t) &\leq -\frac{\gamma}{2}\|R_t\|^2 - \frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{\gamma}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{L\gamma^2}{2}\|R_t\|^2 \\ &= \left(\frac{L\gamma^2}{2} - \frac{\gamma}{2}\right)\|R_t\|^2 - \frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{\gamma}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \quad . \end{aligned}$$

As $\gamma \leq \frac{1}{L}$, we have $\left(\frac{L\gamma^2}{2} - \frac{\gamma}{2}\right) \leq 0$ in the above, thereby proving the lemma. ■

Lemma 19 *Suppose assumptions 3, and 4 hold true. Consider Algorithm 2. For all $t \in \{0, \ldots, T-1\}$, the following holds true:*

$$\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \leq 2\lambda'\sigma^2 + 6\lambda\lambda'\sigma^2 + 6\lambda\zeta^2.$$

Proof From the triangle and the Jensen's inequalities, we obtain that

$$\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 = \left\|R_t - \bar{G}_t + \bar{G}_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \leq 2\left\|R_t - \bar{G}_t\right\|^2 + 2\left\|\bar{G}_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2.$$

From Jensen's inequality, we have

$$\left\|\bar{G}_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 = \left\|\frac{1}{n-f}\sum_{i\in\mathcal{H}}(G_t^{(i)} - \nabla Q^{(i)}(\theta_t))\right\|^2 \leq \frac{1}{n-f}\sum_{i\in\mathcal{H}}\left\|G_t^{(i)} - \nabla Q^{(i)}(\theta_t)\right\|^2 \leq \lambda'\sigma^2.$$

Moreover, we have

$$\begin{aligned}
&\left\|R_t - \bar{G}_t\right\|^2 \leq \lambda\frac{1}{n-f}\sum_{i\in\mathcal{H}}\left\|G_t^{(i)} - \bar{G}_t\right\|^2 \\
&= \lambda\frac{1}{2(n-f)^2}\sum_{i,j\in\mathcal{H}}\left\|G_t^{(i)} - G_t^{(j)}\right\|^2 \\
&\leq \lambda\frac{1}{2(n-f)^2}\sum_{i,j\in\mathcal{H}}\left\|G_t^{(i)} - \nabla Q^{(i)}(\theta_t) + \nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t) + \nabla Q^{(j)}(\theta_t) - G_t^{(j)}\right\|^2 \\
&\leq \lambda\frac{3}{2(n-f)^2}\sum_{i,j\in\mathcal{H}}\left(\left\|G_t^{(i)} - \nabla Q^{(i)}(\theta_t)\right\|^2 + \left\|\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right\|^2 + \left\|\nabla Q^{(j)}(\theta_t) - G_t^{(j)}\right\|^2\right) \\
&= \lambda\frac{3}{n-f}\sum_{i\in\mathcal{H}}\left\|G_t^{(i)} - \nabla Q^{(i)}(\theta_t)\right\|^2 + \lambda\frac{3}{2(n-f)^2}\sum_{i,j\in\mathcal{H}}\left\|\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right\|^2.
\end{aligned}$$

From (42), for all $i \in \mathcal{H}$, we have $\left\|G_t^{(i)} - \nabla Q^{(i)}(\theta_t)\right\|^2 \leq \lambda'\sigma^2$. Furthermore, by Assumption 4, $\frac{1}{2(n-f)^2}\sum_{i,j\in\mathcal{H}}\left\|\nabla Q^{(i)}(\theta_t) - \nabla Q^{(j)}(\theta_t)\right\|^2 = \frac{1}{n-f}\sum_{i\in\mathcal{H}}\left\|\nabla Q^{(i)}(\theta_t) - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \leq \zeta^2$. Thus, from above we obtain that

$$\left\|R_t - \bar{G}_t\right\|^2 \leq 3\lambda\lambda'\sigma^2 + 3\lambda\zeta^2.$$

Combining the above we obtain that

$$\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \leq 2\lambda'\sigma^2 + 6\lambda\lambda'\sigma^2 + 6\lambda\zeta^2.$$

■

Proof [Back to the proof of Theorem 12] Using the fact that the loss function satisfies the PL condition, from Lemma 18, we obtain that

$$\begin{aligned}
Q^{(\mathcal{H})}(\theta_{t+1}) - Q^{(\mathcal{H})}(\theta_t) &\leq -\frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 + \frac{\gamma}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \\
&\leq -\mu\gamma(Q^{(\mathcal{H})}(\theta_t) - Q^*) + \frac{\gamma}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2.
\end{aligned}$$

Therefore, substituting from Lemma 19 in the above, we obtain that

$$\begin{aligned}
Q^{(\mathcal{H})}(\theta_{t+1}) - Q^* &\leq (1-\mu\gamma)(Q^{(\mathcal{H})}(\theta_t) - Q^*) + \frac{\gamma}{2}\left\|R_t - \nabla Q^{(\mathcal{H})}(\theta_t)\right\|^2 \\
&\leq (1-\mu\gamma)(Q^{(\mathcal{H})}(\theta_t) - Q^*) + \gamma(\lambda'\sigma^2 + 3\lambda\lambda'\sigma^2 + 3\lambda\zeta^2).
\end{aligned}$$

Recall that the above holds true for any $t \in \{0, \ldots, T-1\}$. As $\mu \leq L$, we have $1 - \mu\gamma = 1 - \frac{\mu}{L} \in [0, 1)$. Thus, substituting $\gamma = 1/L$ and applying the inequality recursively, we obtain that

$$Q^{(\mathcal{H})}(\theta_T) - Q^* \leq \left(1 - \frac{\mu}{L}\right)^T \left(Q^{(\mathcal{H})}(\theta_0) - Q^*\right) + \frac{1}{\mu}(\lambda'\sigma^2 + 3\lambda\lambda'\sigma^2 + 3\lambda\zeta^2).$$

As $\left(1 - \frac{\mu}{L}\right)^T \leq \exp\left(-\frac{\mu}{L}T\right)$, the above proves the theorem. ■