Robust Collaborative Learning with Linear Gradient Overhead

Sadegh Farhadkhani [1] **Rachid Guerraoui** [1] **Nirupam Gupta** [1]
Lê Nguyên Hoang [2 3] **Rafael Pinot** [1] **John Stephan** [1]

Abstract

Collaborative learning algorithms, such as *distributed SGD* (or D-SGD), are prone to faulty machines that may deviate from their prescribed algorithm because of software or hardware bugs, poisoned data or malicious behaviors. While many solutions have been proposed to enhance the robustness of D-SGD to such machines, previous works either resort to strong assumptions (*trusted* server, *homogeneous* data, specific noise model) or impose a gradient computational cost that is several orders of magnitude higher than that of D-SGD. We present MoNNA, a new algorithm that (a) is provably robust under standard assumptions and (b) has a gradient computation overhead that is linear in the fraction of faulty machines, which is conjectured to be tight. Essentially, MoNNA uses *Polyak's momentum* of local gradients for *local updates* and *nearest-neighbor averaging (NNA)* for *global mixing*, respectively. While MoNNA is rather simple to implement, its analysis has been more challenging and relies on two key elements that may be of independent interest. Specifically, we introduce the mixing criterion of (α, λ)-*reduction* to analyze the *non-linear mixing* of non-faulty machines, and present a way to control the tension between the momentum and the model *drifts*. We validate our theory by experiments on image classification and make our code available at `https://github.com/LPD-EPFL/robust-collaborative-learning`.

1. Introduction

Collaborative learning allows multiple machines (or *nodes*), each with a local dataset, to learn local models that offer a high accuracy on the union of their local datasets (Boyd et al., 2011). This paradigm facilitates the training of complex models over a large volume of data, while addressing concerns on data locality and ownership. The general task of collaborative learning can be formulated as follows. Consider a *parameter space* $\mathbb{R}^d$, a *data space* $\mathcal{X}$ and a *loss function* $q : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$. Given a parameter $\theta \in \mathbb{R}^d$, a data point $x \in \mathcal{X}$ incurs a loss of value $q(\theta, x)$. The system comprises n nodes. Each node i samples data from distribution $\mathcal{D}_i$, and thus has a *local loss function* $Q^{(i)}(\theta) \coloneqq \mathbb{E}_{x \sim \mathcal{D}_i}[q(\theta, x)]$. The goal for each node i is to compute $\theta_*^{(i)}$ minimizing the *global average loss*, i.e.,

$$\theta_*^{(i)} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{j=1}^{n} Q^{(j)}(\theta) \,. \tag{1}$$

Collaborative learning with D-SGD. The most standard way of solving the optimization problem (1) is through the use of the celebrated distributed SGD (D-SGD) method (Tang et al., 2018; Koloskova et al., 2020). Each node maintains a local parameter, approximating a solution of the optimization problem (1), which is updated iteratively in two phases. In the first phase, also called *the local phase*, each node updates its current parameter *partially* using a stochastic estimate of its local loss function's gradient. In the second phase, also called *the coordination phase*, the nodes exchange their partially updated parameters with each other over a network, and then each node replaces its current parameter by the *average* of all the partially updated parameters. While the former is essential for reducing the local loss functions, the latter yields reduction in the global average loss function. Alternately, as is the case in *federated learning* (Kairouz et al., 2021), the nodes may rely on a *trusted* coordinator (called the *server*) to execute the coordination phase involving the averaging operation.

Robustness issue. D-SGD is not very robust: a handful of faulty nodes, deviating from their prescribed algorithm, may prevent the remaining non-faulty (or *correct*) nodes from computing a valid solution (Su & Vaidya, 2016). Such behavior may result from software and hardware bugs, poisoned data, or malicious adversaries controlling part of the network. We consider a setting where at most f (out of n) nodes in the system are faulty and assume that these can

[1]EPFL, Lausanne, Switzerland. [2]Tournesol, [3]Calicarpa, Switzerland. Correspondence to: Sadegh Farhadkhani <sadegh.farhadkhani@epfl.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Table 1. Comparison of MoNNA with other prominent schemes for robust collaborative learning including BRIDGE (Fang et al., 2022), BTARD (Gorbunov et al., 2022), SCC (He et al., 2022), and LEARN (El Mhamdi et al., 2021a). **S** - Stochastic gradients, **H** - Heterogeneous (a.k.a., non-iid) datasets, **A** - Asynchronous communication, and $\mathbf{f}/\mathbf{n}$ - tolerable fraction of faulty nodes.

Method	Loss Function	S	H	A	Communication	$\mathbf{f}/\mathbf{n}$	Gradient Complexity
BRIDGE	Locally strongly convex	×	×	×	Sparse	$< \frac{1}{2}$	$\times^{(*)}$
BTARD	Non-convex	✓	×	×	Pairwise	$< \frac{1}{10}$	$\mathcal{O}\left(\frac{1}{n\epsilon^2}\right)^{(**)}$
SCC	Non-convex	✓	✓	×	Sparse	$\leq \frac{1}{10240}$	$\mathcal{O}\left(\left(\frac{1}{n}+\frac{f}{n}\right)\frac{1}{\epsilon^2}\right)$
LEARN	Non-convex	✓	✓	✓	Pairwise	$< \frac{1}{3}$ or $< \frac{1}{6}$	$\mathcal{O}\left(\frac{1}{\epsilon^5}\right)$
MoNNA	Non-convex	✓	✓	✓	Pairwise	$\leq \frac{1}{11}$ $< \frac{1}{5}$	$\mathcal{O}\left(\left(\frac{1}{n}+\frac{f}{n}\right)\frac{1}{\epsilon^2}\right)$ $\mathcal{O}\left(\frac{(1+f)^2}{n\epsilon^2}\right)$
Gradient complexity for non-convex losses with a trusted server (Karimireddy et al., 2022):							$\mathcal{O}\left(\left(\frac{1}{n}+\frac{f}{n}\right)\frac{1}{\epsilon^2}\right)$

$^{(*)}$ As of yet, no finite time convergence rate is known for BRIDGE.

$^{(**)}$ The leading term in the convergence rate of BTARD (Gorbunov et al., 2022) is identical to that of D-SGD without faults and does not introduce any overhead. The reason is that it considers a weaker adversarial model with public datasets, where each node can access the entire training data and validate the computations done by other nodes, and thereby check any faults.

behave arbitrarily[1] (either by accident or intent). In this case, the original optimization problem (1) is rendered vacuous. A more reasonable goal is to minimize the average loss function for the correct nodes (Gupta & Vaidya, 2020). However, there is a fundamental limit on achieving this goal, because faulty nodes may behave as correct nodes with outlying local data distributions (Liu et al., 2021; Karimireddy et al., 2022). Thus, the ultimate goal of *robustness* reduces to designing an algorithm that enables all correct nodes to compute a *tight* approximation of a minimum of the average correct loss (El Mhamdi et al., 2021a; He et al., 2022).

1.1. Prior Work

The problem of robustness in collaborative learning has received significant attention in recent years (Yang et al., 2020; Liu, 2021; Bouhata & Moumen, 2022). Most previous works focused on server-based coordination (i.e., the nodes have access to a server that is assumed fault-free) (El Mhamdi et al., 2018; Damaskinos et al., 2018; Chen et al., 2017; Yin et al., 2018; Karimireddy et al., 2021; 2022; Farhadkhani et al., 2022). This server constitutes a *single point of failure*, which greatly compromises the security of the learning procedure. It is therefore appealing to consider a scenario in which the nodes collaborate by communicating directly, without relying on a central server.

The absence of a central authority, combined with asynchronous communication (Cachin et al., 2011) and faulty nodes, lead to a non-trivial *drift* between the local parameters maintained by the correct nodes. Controlling this drift is key to learning an accurate model by the correct nodes, and constitutes a major challenge. Prior attempts to address this issue, including (Fang et al., 2022; Yang & Bajwa, 2019; El Mhamdi et al., 2021a; Guo et al., 2021; He et al., 2022; Gorbunov et al., 2022), rely on strong assumptions such as *homogeneous* data (Fang et al., 2022; Yang & Bajwa, 2019; Guo et al., 2021; Gorbunov et al., 2022), *strong convexity* (Gupta et al., 2021; Yang & Bajwa, 2019), a precise gradient noise modeling, and an extremely small fraction of faulty nodes as in the parallel work of He et al. (2022); or impose orders of magnitude larger gradient overhead compared to D-SGD (El Mhamdi et al., 2021a). These shortcomings limit the practicality of the state-of-the-art methods.

1.2. Contributions

We take an important step towards making robust collaborative learning more realizable. Specifically, we present an adaptation of D-SGD, named MoNNA, which to the best of our knowledge, is the first collaborative learning algorithm that is provably robust under assumptions that are standard in analyzing D-SGD (Lian et al., 2017; Tang et al., 2018). Moreover, the gradient computational overhead imposed by MoNNA, compared to D-SGD, only grows linearly in the fraction of faulty nodes, which is conjectured to be tight (Karimireddy et al., 2021). We compare MoNNA with the most relevant related approaches in Table 1.

[1] In distributed computing, such faulty nodes are also commonly referred to as Byzantine (Lamport et al., 1982).

Overview of MoNNA. In the local phase, unlike D-SGD, each correct node uses the Polyak's momentum (Polyak, 1964) of its local stochastic gradients to partially update its current local parameter. The use of local momentum amortizes the dependence on local variance in the error due to faulty nodes. In the coordination phase, instead of simply averaging the received partial updates, each correct node aggregates them using a *robust aggregation rule* we call nearest neighbor averaging (NNA). In NNA, as the name suggests, a node eliminates the f parameters it receives that are the farthest from its own and then averages the rest. This filtering aims to reduce the drift between correct nodes' local parameters, by mitigating the influence of arbitrary parameters that may be sent by the faulty nodes. While MoNNA has been rather simple to implement, its analysis has been more challenging, involving two elements that may be of independent interest to the distributed optimization community at large: namely, (i) the mixing criterion of (α, λ)-*reduction*, and (ii) the control of local parameters' drift under (α, λ)-reduction mixing when incorporating Polyak's momentum. We discuss these elements below, after the summary of our theoretical results.

Theoretical results. We assume at most f out of n nodes may be faulty and behave arbitrarily. We denote by $\mathcal{C}$ the set of correct nodes and $Q^{(\mathcal{C})}(\theta)$ their average loss, i.e.,

$$Q^{(\mathcal{C})}(\theta) := \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} Q^{(i)}(\theta). \tag{2}$$

We consider the class of Lipschitz smooth non-convex loss functions, and assume local stochastic gradients (of correct nodes) to satisfy standard properties in the context of D-SGD (Tang et al., 2018), i.e., bounded local variance of σ^2 and bounded global diversity of ζ^2. We show that if $n \geq 11f$, then upon executing T iterations of MoNNA, each correct node i returns a local parameter $\hat{\theta}^{(i)}$ such that $\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq \epsilon$ where

$$\epsilon \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{T}\left(\frac{1+f}{n}\right)} + \frac{f}{n}\zeta^2\right). \tag{3}$$

Recall that the number of iterations T equals the total number of gradients computed by each correct node. Hence, the gradient complexity of MoNNA is $1 + f$ times that of D-SGD, i.e., the gradient overhead is linear in the fraction of faulty nodes. Note that the non-vanishing error of $(f/n)\zeta^2$ is a fundamental lower bound in the presence of faulty nodes due to diversity in local distributions (Karimireddy et al., 2022). We also show that, by increasing gradient complexity by a factor f, MoNNA is robust to $n/5$ faulty nodes.

(α, λ)-Reduction mixing. In the presence of faulty nodes, it is impossible to ensure *linear* mixing of correct updates in the coordination phase. We can no longer rely on the linear mixing criterion of double stochasticity with a bounded spectral gap, usually assumed in the case of D-SGD (Tsitsiklis et al., 1986; Xiao & Boyd, 2004; Tang et al., 2018; Koloskova et al., 2020). To circumvent this limitation, we introduce a new mixing criterion of (α, λ)-*reduction* that extends the classical mixing criterion to analyze *robust mixing* schemes, such as NNA, that may be non-linear and even non-continuous. Parameters α and λ are positive real values quantifying the levels of contraction and centering, respectively, over the set of correct updates. We prove that the use of (α, λ)-reduction, with $\alpha < 1$ and $\lambda < \infty$, in the coordination phase of D-SGD enables each correct node i to return $\hat{\theta}^{(i)}$ such that $\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq \epsilon$, where

$$\epsilon \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{nT}} + \frac{\lambda}{(1-\alpha)^2}\left(\zeta^2 + \sigma^2\right)\right). \tag{4}$$

Assuming asynchronous pairwise communication between nodes and $f \leq \frac{n}{11}$, we prove that NNA satisfies (α, λ)-reduction with $\alpha \leq 0.988 < 1$ and $\lambda \in \Theta\left(f/n\right)$. The key to proving this result is that, unlike standard aggregation rules (Farhadkhani et al., 2022), NNA makes each correct node pivot the aggregation around their own local parameter. Substituting these values of α and λ in (4) yields the error in (3), plus an additional non-vanishing term of $(f/n)\sigma^2$. We show that this term vanishes at the rate of $\sqrt{1/T}$ when using local momentum, as specified in MoNNA. Hence, reducing the error term to (3). Proving this reduction however requires a novel technique for controlling drift.

Controlling drift under momentum. The second key element underlying our analysis pertains to the use of Polyak's momentum for local updates in MoNNA. We prove that momentum eliminates the non-vanishing error due to the local variance σ^2 in the optimization error (4), and thereby matches the lower bound. While such observation has been made in the case of server-based coordination (El Mhamdi et al., 2021b; Karimireddy et al., 2021; Farhadkhani et al., 2022), it is not immediate in our setting because of the cross-coupling of the momentum drift and the drift between correct nodes' models. By carefully analyzing this coupling, we obtain uniform bounds on both model and momentum drifts. We then adapt the *Lyapunov function* (a.k.a. potential function) to account for the model drift.

Empirical evaluation. We evaluate MoNNA on two benchmark image classification tasks. We consider a distributed asynchronous system including $n/5$ faulty nodes executing four different attacks. MoNNA significantly outperforms state-of-the-art robust collaborative learning algorithms in all adversarial settings, and almost matches the performance of D-SGD in terms of learning accuracy.

1.3. Paper Organization

We formalize robust collaborative learning in Section 2. Section 3 presents our algorithm as well as its convergence and robustness. Section 4 discusses the key elements of our convergence analysis. Section 5 presents our empirical evaluation. We discuss future research directions in Section 6. Due to space limitations, full proofs and some auxiliary empirical results are deferred to the appendices.

2. Problem Statement

We consider a set of n nodes, $[n] = \{1, \dots, n\}$, out of which at most $f < n/3$ may behave arbitrarily. We refer to such nodes as *faulty*. The identity of faulty nodes is a priori unknown to the remaining correct, i.e., non-faulty nodes. We assume that the nodes interact with each other using the following communication model.

Communication model. We assume a *pairwise* communication scheme where nodes exchange messages with each other over a network. The messages however need not arrive in a timely manner, i.e., communication is *asynchronous*. A correct node cannot wait to receive messages from all the other nodes since it can be indefinitely stalled by a single faulty node that chooses not to send any message. This amplifies the challenge of robustness. A simple adaption of server based solutions (El Mhamdi et al., 2018; Damaskinos et al., 2018; Chen et al., 2017; Yin et al., 2018; Karimireddy et al., 2022; Farhadkhani et al., 2022) cannot prevent the local models at the correct nodes from *drifting* apart, rending their local gradients useless for the others.

Robustness. We consider an arbitrary set $\mathcal{C}$ comprising $n - f$ correct nodes. We denote by $Q^{(\mathcal{C})}(\theta)$ the average loss function of the nodes in $\mathcal{C}$, defined in (2). The ideal objective of *robust* learning is to design an algorithm that allows the correct nodes, under the above communication model, to minimize $Q^{(\mathcal{C})}(\theta)$, despite the presence of faulty nodes. Solving this problem however is NP-hard in general, as the loss function need not be convex (Boyd et al., 2004). Thus, a more realizable goal is finding a *critical point* of $Q^{(\mathcal{C})}(\theta)$, assuming the point-wise loss function $q(\theta, x)$ to be differentiable in θ. In our context, we formally define the problem of robustness through the notion of *resilience*.

Definition 1. *An algorithm is said to be* (f, ϵ)-resilient *if it enables each correct node* $i \in \mathcal{C}$ *to compute* $\hat{\theta}^{(i)}$ *such that*

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq \epsilon,$$

despite the presence of f *faulty nodes, where the expectation* $\mathbb{E}[\cdot]$ *is taken over the randomness of the algorithm.*

We assume the gradients of the loss functions to be Lipschitz smooth and the variance of the local stochastic gradients to be bounded. These assumptions are classical to the analysis of stochastic first-order methods, and hold true for many learning problems (Bottou et al., 2018). Note that, by definition of $Q^{(i)}$, we have $\nabla Q^{(i)}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_i}[\nabla q(\theta, x)]$.

Assumption 1 (Lipschitz smoothness)**.** *There exists* $L < \infty$ *such that for all* $i \in \mathcal{C}$ *and all* $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\left\|\nabla Q^{(i)}(\theta_1) - \nabla Q^{(i)}(\theta_2)\right\| \leq L \left\|\theta_1 - \theta_2\right\|.$$

Assumption 2 (Bounded variance)**.** *There exists* $\sigma < \infty$ *such that for all* $i \in \mathcal{C}$*, and all* $\theta \in \mathbb{R}^d$,

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

Additionally, we assume that the diversity amongst the gradients of local loss functions is bounded, as stated below. We note that this assumption is standard in *heterogeneous settings*, i.e., when nodes have different data distributions (Lian et al., 2018; Tang et al., 2018), especially when addressing the problem of resilience (Data & Diggavi, 2021).

Assumption 3 (ζ-heterogeneous)**.** *There exists* $\zeta < \infty$ *such that for all* $\theta \in \mathbb{R}^d$,

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

In particular, the heterogeneity bound ζ can be derived based on the closeness of the underlying local data distributions at the nodes (Fallah et al., 2020).

Lower bound. Under heterogeneity, it is generally impossible to achieve (f, ϵ)-resilience for any arbitrary value of ϵ (El Mhamdi et al., 2021a; Karimireddy et al., 2022). Specifically, we have the following lower bound.

Lemma 1 (Theorem III, Karimireddy et al. (2022))**.** *Suppose assumptions 1, 2, and 3 hold true. If an algorithm is* (f, ϵ)*-resilient, then*

$$\epsilon \in \Omega\left(\frac{f}{n}\zeta^2\right).$$

3. MoNNA

We describe below our algorithm, MoNNA, its (per step) computational costs and its robustness properties.

3.1. Description

MoNNA enhances D-SGD (Tang et al., 2018) by incorporating a momentum (Mo) operation (Polyak, 1964) as well as a new mixing scheme named nearest neighbor averaging (NNA). We summarize below the key elements of the local phase and the coordination phase in an iteration t where

each correct node i maintains a local model $\theta_t^{(i)}$. The initial models for the correct nodes are assumed identical, i.e., each correct node i chooses an initial model $\theta_0^{(i)}$ such that $\theta_0^{(i)} := \theta_0 \in \mathbb{R}^d$. Complete execution of MoNNA is presented in Algorithm 1.

Local phase. Each correct node i samples a data point $x \sim \mathcal{D}_i$ and computes a local stochastic gradient

$$g_t^{(i)} = \nabla q\left(\theta_t^{(i)}, x\right). \tag{5}$$

Then node i updates its current local momentum as follows

$$m_t^{(i)} = \beta m_{t-1}^{(i)} + (1-\beta) g_t^{(i)}, \tag{6}$$

where $\beta \in [0,\, 1)$ is called the *momentum coefficient*, and $m_{-1}^{(i)} = 0$ by convention. Lastly, node i *partially* updates it current model $\theta_t^{(i)}$ by computing $\theta_{t+1/2}^{(i)} := \theta_t^{(i)} - \gamma\, m_t^{(i)}$.

Coordination phase. Each correct node i initializes a new vector $x_0^{(i)} = \theta_{t+1/2}^{(i)}$. The coordination phase is composed of $K \geq 1$ rounds. In each round $k \in K$, the following interaction and mixing schemes are executed.

(a) Interaction. The nodes exchange their respective vectors $\{x_{k-1}^{(i)},\, i \in n\}$ with each other using *signed echo broadcast* (Cachin et al., 2011).[2] Recall that a faulty node j may choose to send either an arbitrary value for its vector $x_{k-1}^{(j)}$ or no message at all. Hence, to avoid getting stalled, a correct node i only waits to receive $n-f-1$ messages before moving to the mixing step.

(b) Mixing. Each correct node i updates its vector $x_{k-1}^{(i)}$ to $x_k^{(i)}$ by aggregating the $n-f-1$ vectors it receives with its own, using *nearest neighbor averaging* (NNA). For $n-f$ vectors $z^{(0)}, \ldots,\, z^{(n-f-1)}$ in $\mathbb{R}^d$, this aggregation is defined to be

$$\text{NNA}\left(z^{(0)}; \left\{z^{(i)}\right\}_{i=1}^{n-f-1}\right) := \frac{1}{n-2f} \sum_{i=0}^{n-2f-1} z^{(\tau(i))},$$

where τ is a permutation on $\{1, \ldots, n-f-1\}$ such that $\left\|z^{(0)} - z^{(\tau(1))}\right\| \leq \ldots \leq \left\|z^{(0)} - z^{(\tau(n-f-1))}\right\|$.

3.2. Computation & Communication Costs

Computing a local momentum as per (6) is equivalent in terms of complexity to computing a single local gradient. Thus, the computational cost of the local phase in MoNNA is the same as that of D-SGD. Second, the coordination phase in MoNNA comprises K rounds in which each correct node computes the output of NNA. This involves computing $n-f-1$ distances in $\mathbb{R}^d$ and sorting them to obtain τ. The former is in $\mathcal{O}(nd)$ and the latter can be done using a sorting algorithm, e.g., *quicksort* (Cormen et al., 2022), in $\mathcal{O}(n \log n)$. Hence, the total computation cost for the coordination phase of MoNNA is in $\mathcal{O}(n(d + \log n) K)$, compared to $\mathcal{O}(nd)$ for D-SGD. Similarly, the communication cost of MoNNA is in $\mathcal{O}(nK)$, which is a factor K more than that of D-SGD.

Constant K is however usually relatively small compared to the standard costs of D-SGD. Indeed, in the main result of the paper (Theorem 1) when $n \geq 11f$, we set $K = 1$. Therefore, in this case, the computational and communication complexity of MoNNA is $\mathcal{O}(n(d + \log n))$ and $\mathcal{O}(n)$, respectively, which almost matches the $\mathcal{O}(nd)$ computational and $\mathcal{O}(n)$ communication complexity of D-SGD. Furthermore, to improve the robustness of MoNNA to $n > 5f$, we set $K = \mathcal{O}(\log n)$ which adds a $\log n$ overhead to the communication and computational costs (see Corollary 2 in the Appendix), but remains reasonable compared to other existing solutions such as (El Mhamdi et al., 2021a).

[2] For pedagogical reasons, we defer the implementation details of signed echo broadcast (SEB) to Appendix C. Essentially, SEB prevents a faulty node from sending mismatching messages to different correct nodes.

Algorithm 1 **MoNNA** as executed by a correct node i

Initialization: Initial model $\theta_0^{(i)} := \theta_0 \in \mathbb{R}^d$, initial momentum $m_{-1}^{(i)} = 0$, momentum coefficient $\beta \in [0,\, 1)$, total iterations T, learning rate γ, number of coordination rounds K, and threshold f on the number of faulty nodes.

For each **iteration** $t = 1, \ldots,\, T$, do the following:

Local phase:

(a) Update local momentum $m_t^{(i)} = \beta m_{t-1}^{(i)} + (1-\beta) g_t^{(i)}$ where $g_t^{(i)}$ is defined as in (5).

(b) Partially update local model $\theta_{t+1/2}^{(i)} := \theta_t^{(i)} - \gamma\, m_t^{(i)}$.

Coordination phase: Initialize vector $x_0^{(i)} := \theta_{t+1/2}^{(i)}$ and execute the following K rounds.

(a) In each **round** $k = 1, \ldots,\, K$, do the following

i. Initialize $\mathcal{R}_k^{(i)} = \emptyset$.

ii. Broadcast vector $x_{k-1}^{(i)}$ to the other nodes. *(A faulty node j may send an arbitrary value for $x_{k-1}^{(j)}$)*

iii. **While** $|\mathcal{R}_k^{(i)}| < n - f - 1$ **do**: Upon receiving a vector from node j, update $\mathcal{R}_k^{(i)} = \mathcal{R}_k^{(i)} \cup \{j\}$.

iv. Compute $x_k^{(i)} = \text{NNA}\left(x_{k-1}^{(i)}; \left\{x_{k-1}^{(j)} \mid j \in \mathcal{R}_k^{(i)}\right\}\right)$.

(b) Update local model $\theta_{t+1}^{(i)} = x_K^{(i)}$.

Output: $\hat{\theta}^{(i)} \sim \mathcal{U}\left\{\theta_0^{(i)}, \ldots,\, \theta_{T-1}^{(i)}\right\}$.

3.3. Convergence & Robustness

We now present our main theoretical result demonstrating the finite time convergence of MoNNA. Essentially, we analyze Algorithm 1 under assumptions 1, 2, and 3, and upon assuming a sufficiently small learning rate γ. When $n \geq 11f$, it suffices to perform one round per coordination phase, i.e., set $K = 1$. We now state our main theorem[3], upon introducing the following notation:

$$Q^* = \min_{\theta \in \mathbb{R}^d} Q^{(\mathcal{C})}(\theta), \quad \overline{\theta}_t := \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \theta_t^{(i)}.$$

Theorem 1. *Suppose that assumptions 1, 2 and 3 hold true, and that $n \geq 11f$. Let us denote*

$$\alpha = \frac{9.88f}{n-f} \leq 0.988, \quad \lambda = \frac{9f}{n-f},$$
$$c_0 := 12\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right), \quad c_1 := \frac{18\alpha(1+\alpha)}{(1-\alpha)^2},$$
$$c_2 := 72L\left(\frac{3}{n-f} + 2c_1 + \frac{9\lambda}{2}(2c_1+3)\right),$$
$$c_3 := 6\left(6c_1 + \frac{9\lambda}{2}(4c_1+9)\right) \text{ and } c_4 := \frac{9nc_0c_1}{c_2}.$$

Consider Algorithm 1 with $K = 1$, $\gamma = \min\left\{\frac{1}{12L}, \frac{1}{L}\sqrt{\frac{2}{3c_1}}, \sqrt{\frac{c_0}{c_2LT\sigma^2}}\right\}$, and $\beta = \sqrt{1-12\gamma L}$. Then, for all $T \geq 1$ and $i \in \mathcal{C}$, we have

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq 2\sqrt{\frac{c_0c_2L\sigma^2}{T}} + \frac{12Lc_0}{T}$$
$$+ \frac{Lc_0}{T}\sqrt{\frac{3c_1}{2}} + \frac{36}{T}\left(\frac{\sigma^2}{n-f}\right) + \frac{c_4L}{T}\left(1 + \frac{\zeta^2}{\sigma^2}\right) + c_3\zeta^2.$$

Using Theorem 1, we can show that Algorithm 1 guarantees (f, ϵ)-resilience. Specifically, upon ignoring the higher-order terms in T, we obtain the following corollary.

Corollary 1. *Under the conditions stated in Theorem 1, Algorithm 1 guarantees (f, ϵ)-resilience where*

$$\epsilon \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{T}\left(\frac{1+f}{n}\right)} + \frac{f}{n}\zeta^2\right).$$

Linear gradient overhead. In the fault-free setting, i.e., when $f = 0$, the convergence result shown in Corollary 1 reduces to that of the conventional D-SGD (Tang et al., 2018). However, when $f > 0$, MoNNA induces an overhead on the number of gradients computed per correct nodes compared to D-SGD. Specifically, correct nodes in MoNNA compute $(1+f)$ times more gradients than in the fault-free case (to obtain a comparable error), which is linear in f. We believe this gradient overhead to be tight, as conjectured in (Karimireddy et al., 2021). While MoNNA only imposes a linear overhead under the assumption that $f \leq n/11$, it can tolerate a larger fraction of faulty nodes, i.e., arbitrarily close to $n/5$, by imposing a quadratic gradient overhead in f (see Corollary 2 in Appendix A.4).

[3]The dependence of c_4 on n comes from the fact that we provide the convergence guarantee for any honest node i. This is stronger than the prior work (Koloskova et al., 2020; He et al., 2022) where the convergence guarantee is often given for the average of the local models $\overline{\theta}_t$.

4. Convergence Analysis

We now explain the key elements that we build upon to prove the convergence guarantee stated in Theorem 1. Essentially, we first introduce the mixing criterion of (α, λ)-reduction and, then, show how to control the drift in local updates.

4.1. (α, λ)-reduction

To handle the non-linear mixing of correct momentums, we introduce the notion of (α, λ)-reduction. This notion can be seen as a relaxation of the classical linear mixing criterion of double stochasticity with a bounded spectral gap. We show that (α, λ)-reduction is sufficient to maintain tight convergence guarantees, while it can be satisfied by non-linear and even non-continuous schemes such as NNA.

Definition 2 (*(α, λ)-reduction*)**.** *Consider a coordination phase Ψ. For correct nodes $i \in \mathcal{C}$ initiating the coordination phase with vectors $\left\{z^{(i)}, i \in \mathcal{C}\right\}$, we denote by $\left\{y^{(i)}, i \in \mathcal{C}\right\}$ the vectors obtained by these nodes upon the completion of Ψ. Then, Ψ is said to guarantee (α, λ)-*reduction *if, for any $\left\{z^{(i)}, i \in \mathcal{C}\right\}$, the following holds true:*

$$\textit{i)} \quad \frac{1}{|\mathcal{C}|}\sum_{i \in \mathcal{C}}\left\|y^{(i)} - \bar{y}\right\|^2 \leq \alpha\frac{1}{|\mathcal{C}|}\sum_{i \in \mathcal{C}}\left\|z^{(i)} - \bar{z}\right\|^2$$
$$\textit{ii)} \quad \|\bar{y} - \bar{z}\|^2 \leq \lambda\frac{1}{|\mathcal{C}|}\sum_{i \in \mathcal{C}}\left\|z^{(i)} - \bar{z}\right\|^2$$

where $\bar{z}$ and $\bar{y}$ denote the vector averages of $\left\{z^{(i)}, i \in \mathcal{C}\right\}$ and $\left\{y^{(i)}, i \in \mathcal{C}\right\}$, respectively.

In MoNNA, each correct node i initializes the coordination phase with vector $z^{(i)} = x_0^{(i)}$ and outputs $y^{(i)} = x_K^{(i)}$ at its completion. In Appendix A.5, we show that the coordination phase of Algorithm 1 satisfies (α, λ)-reduction for $\lambda \in \Theta(f/n)$, $\alpha < 1$, and $\alpha \in \Theta(f/n)$ when $n \geq 11f$. Additionally, when $n > 5f$, we have $\lambda \in \Theta\left(f^2/n\right)$, $\alpha < 1$, and $\alpha \in \Theta(f/n)$ (shown in Appendix A.4).

4.2. D-SGD with (α, λ)-reduction

To better understand the utility of (α, λ)-reduction, we first provide a convergence guarantee for MoNNA *without* mo-

mentum (i.e., $\beta = 0$), while assuming the communication phase to satisfy (α, λ)-reduction. Alternately, this algorithm reduces to D-SGD with (α, λ)-reduction mixing.

Proposition 1. *Consider Algorithm 1 with $\beta = 0$. Suppose that assumptions 1, 2 and 3 hold true, and that the coordination phase satisfies (α, λ)-reduction for $\alpha < 1$. Then there exists a constant learning rate γ for which each correct node $i \in \mathcal{C}$ returns $\hat{\theta}^{(i)}$ such that*

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{nT}} + \frac{\lambda\left(\sigma^2 + \zeta^2\right)}{(1-\alpha)^2}\right).$$

Notably, when all the nodes are correct (i.e., $f = 0$), then the coordination phase of Algorithm 1 simply computes the average and satisfies (α, λ)-reduction for $\alpha = \lambda = 0$. Then, we recover the classical convergence guarantee of D-SGD without faulty nodes, i.e., $\mathcal{O}\left(\sqrt{\sigma^2/nT}\right)$. However, in the presence of faulty nodes (when $\lambda \in \Theta\left(f/n\right)$, and $\alpha < 1$), Proposition 1 shows an asymptotic error of $\frac{f}{n}\sigma^2 + \frac{f}{n}\zeta^2$. While the term depending on ζ^2 is a fundamental lower bound as per Lemma 1, the one depending on σ^2 can be alleviated through the use of momentum, as we show next.

4.3. Polyak's Momentum

Although momentum has been shown to be beneficial in the particular case of server-based coordination (Karimireddy et al., 2021; Farhadkhani et al., 2022), extending the existing analyses to our setting is not straightforward. The main bottleneck is the non-trivial drift that occurs between the local parameters maintained by correct nodes, i.e., $\sum_{i\in\mathcal{C}} \left\|\theta_t^{(i)} - \bar{\theta}_t\right\|^2$. When momentum is applied, this drift gets coupled with the drift between their momentum vectors, i.e., $\sum_{i\in\mathcal{C}} \left\|m_t^{(i)} - \overline{m}_t\right\|^2$. Indeed, an elementary analysis of this coupling suggests the possibility of uncontrolled growth of the two drifts: a high model drift increases the momentum drift and vice versa. We devise a refined analysis, showing that an appropriate learning rate, which is a function of parameter α in (α, λ)-reduction, ensures uniform bounds proportional to $(1-\beta)\sigma^2$ for both model and momentum drifts (Lemma 3 in Appendix A). This suggests that we can diminish the dependence on σ^2 by choosing a large momentum parameter close to 1. However, an arbitrarily large momentum parameter increases the bias, i.e.,

$$\delta_t := \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left(m_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right),$$

which in turn negatively impacts the convergence. To prove that the positive effect of momentum (i.e., reduction in drift) outweighs the negative effect (i.e., increase in bias), we define our Lyapunov function (or *potential* function) to be

$$V_t := \mathbb{E}\left[Q^{(\mathcal{C})}\left(\bar{\theta}_t\right) - Q^* + \frac{1}{4L}\left\|\delta_t\right\|^2\right].$$

The above Lyapunov function is inspired from the existing momentum literature (Cutkosky & Orabona, 2019; Farhadkhani et al., 2022), but adapted to address the model drift.

5. Empirical Evaluation

We compare the performance of MoNNA to D-SGD and several state-of-the-art algorithms from the literature. In particular, we consider three methods, namely BRIDGE (Fang et al., 2022), SCC (He et al., 2022), and LEARN (El Mhamdi et al., 2021a)[4]. We consider two classical image classification tasks, on which we test the robustness of MoNNA under four attacks. Every experiment is repeated five times using seeds 1 to 5 for reproducibility. Our code will be made available online.

5.1. Experimental Setup

Datasets. We use the MNIST (LeCun & Cortes, 2010) and CIFAR-10 (Krizhevsky et al., 2009) datasets, pre-processed as in (Baruch et al., 2019) and (El Mhamdi et al., 2021b). Refer to Appendix D.1 for more details on pre-possessing.

Model architecture and hyperparameters. For MNIST, we consider a convolutional neural network (CNN) with two convolutional layers followed by two fully-connected layers. The model is trained using a learning rate $\gamma = 0.75$ for $T = 600$ iterations. We use a total number of nodes $n = 26$, out of which $f = {}^{(26-1)}/_{5} = 5$ are faulty. For CIFAR-10, we use a CNN with four convolutional layers and two fully-connected layers. Furthermore, we set $\gamma = 0.5$ and $T = 2000$ iterations. The distributed system in this case consists of $n = 16$ nodes, out of which $f = {}^{(16-1)}/_{5} = 3$ are faulty. A detailed presentation of the entire experimental setup can be found in Appendix D.2.

Heterogeneity. In order to simulate data heterogeneity, the correct nodes sample from the original datasets using a Dirichlet distribution of parameter α, as done in (Hsu et al., 2019). We evaluate our algorithm on MNIST with $\alpha \in \{1, 5\}$, and on CIFAR-10 with $\alpha = 5$. A pictorial representation of the resulting heterogeneity as a function of α can be found in Appendix D.3.

Attacks and asynchrony. We consider four state-of-the-art attacks performed by the faulty nodes, namely *a little is enough (ALIE)* (Baruch et al., 2019), *fall of empires (FOE)* (Xie et al., 2019), *sign-flipping (SF)* (Allen-Zhu et al., 2020), and *label-flipping (LF)* (Allen-Zhu et al., 2020).

[4]We do not implement BTARD (Gorbunov et al., 2022) due to its weaker adversarial model and assumption of a public data pool.

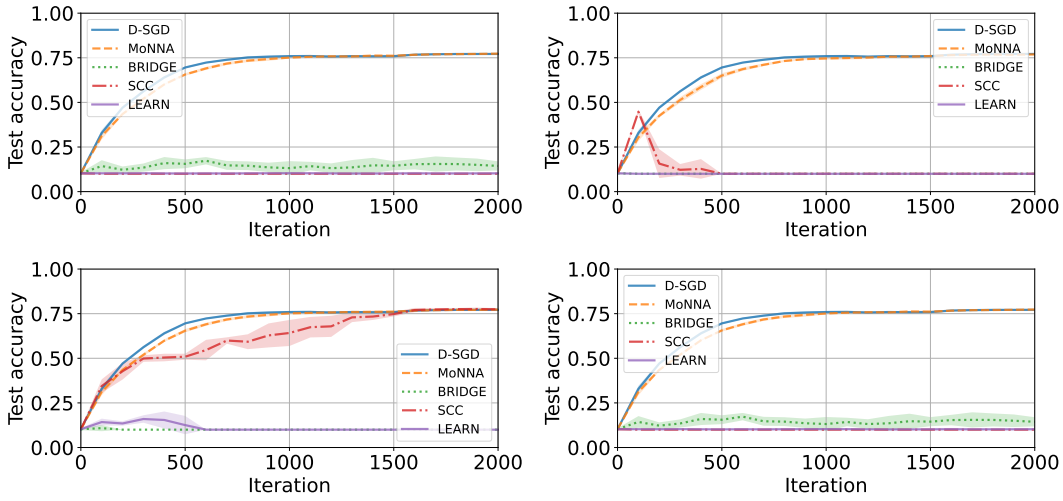

Figure 1. Learning accuracies achieved on CIFAR-10 with $\alpha = 5$ by D-SGD, MoNNA, BRIDGE, SCC, and LEARN. There are $n = 16$ nodes out of which $f = 3$ are faulty. The faulty nodes execute the *FOE* (row 1, left), *ALIE* (row 1, right), *LF* (row 2, left), and *SF* (row 2, right) attacks. All algorithms except LEARN compute 100,000 gradients, while LEARN computes 2,001,000 gradients.

These attacks are explained in detail in Appendix D.4. In order to emulate the ill-effects of asynchrony, we ensure that the correct nodes receive first the messages of the faulty nodes. Put differently, to construct any correct node i's set of $n - f - 1$ first received messages $\mathcal{R}_k^{(i)}$, we first insert the messages sent by faulty nodes. We then complete the set by randomly sampling the remaining messages from the correct vectors.

Evaluation details. To serve as a benchmark for MoNNA, we run D-SGD in a non-adversarial environment (i.e., without faults), and with momentum $\beta = 0.99$. We also execute MoNNA with $\beta = 0.99$ and $K = 1$ (i.e., one coordination round per iteration), and report on its performance in four adversarial settings. Furthermore, we execute SCC with $\beta = 0.9$ (fine-tuned as in (He et al., 2022)), and LEARN without momentum as prescribed in (El Mhamdi et al., 2021a). We use the same number of iterations T for all algorithms and compare their learning accuracies and computational workloads per node.

5.2. Experimental Results

Figure 1 showcases the performance of MoNNA compared to the other algorithms. For space limitations and better readability, we only show in Figure 1 our results on CIFAR-10. The remaining results on MNIST are deferred to Appendix E, and convey the same observations made hereafter.

Figure 1 clearly shows the empirical superiority of MoNNA in four adversarial settings. Indeed, MoNNA is the only solution that performs consistently well under all the four attacks, almost matching the performance of D-SGD without faults. Its closest rival among the considered techniques is SCC. Even then, while SCC displays comparable performance to MoNNA under the LF attack, its learning capabilities drop significantly when tested against the remaining three attacks, especially FOE and SF that make the accuracy of SCC fall to 10%. Moreover, BRIDGE and LEARN also present poor performances under all four attacks, their final accuracies stagnating at around 10%.

Note also that the number of gradients each node computes when using LEARN is 20 times more than that in MoNNA. The inflated computational costs associated to LEARN are explained by the dynamic sampling technique the algorithm implements, whereby the batch-size is gradually augmented across iterations. MoNNA, BRIDGE, and SCC compute the same number of gradients per node as D-SGD since they all share a constant batch-size during the entire learning. In summary, our results show the empirical superiority of our algorithm in adversarial settings since MoNNA matches the performance of fault-free D-SGD, both in terms of learning accuracy and computational complexity.

Remark. Although our empirical evaluation conveys a poor performance of BRIDGE and LEARN, it is important to note that we do not contradict the previous findings on these methods reported in (Fang et al., 2022) and (El Mhamdi et al., 2021a), respectively, that consider very weak attack models. In short, (Fang et al., 2022) report on an evaluation of BRIDGE assuming faulty nodes that only send random vectors, instead of executing state-of-the-art attacks. On the other hand, (El Mhamdi et al., 2021a) report on an evaluation of LEARN in a fault-less system, i.e., without any attack. Additionally, as opposed to MoNNA and SCC, these techniques do not use local momentum, which has

been recently recognized as a key ingredient in the robustness of distributed learning algorithms (Farhadkhani et al., 2022; Karimireddy et al., 2021). We further comment on the necessity of momentum in Appendix E.2.

6. Concluding Remarks

We present MoNNA, a novel collaborative learning algorithm that is provably robust under standard learning assumptions. We show that MoNNA has a linear gradient computation overhead in the fraction of faulty machines. One of our main contributions is the introduction of the new mixing criterion of (α, λ)-reduction, allowing us to obtain tight convergence guarantees. Following prior works on robust collaborative learning (El Mhamdi et al., 2021a; Gorbunov et al., 2022), we studied this criterion under the pairwise communication scheme, which comes with a high communication overhead. We aim to study (α, λ)-reduction under sparse communication networks or client sub-sampling schemes to reduce the communication overhead and further improve the practical applicability of our method.

Acknowledgments

This work has been supported in part by the Swiss National Science Foundation (SNSF) projects 200021-200477 and 200021-182542.

References

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

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

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

Bouhata, D. and Moumen, H. Byzantine fault tolerance in distributed machine learning: a survey. *arXiv preprint arXiv:2205.02572*, 2022.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Cachin, C., Guerraoui, R., and Rodrigues, L. *Introduction to reliable and secure distributed programming*. Springer Science & Business Media, 2011.

Chen, Y., Su, L., and Xu, J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to algorithms*. 2022.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Damaskinos, G., El Mhamdi, E. M., Guerraoui, R., Patra, R., and Taziki, M. Asynchronous Byzantine machine learning (the case of SGD). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 10–15 Jul 2018.

Data, D. and Diggavi, S. Byzantine-resilient sgd in high dimensions on heterogeneous data. In *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021.

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

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

El Mhamdi, E. M., Guerraoui, R., and Rouault, S. Distributed momentum for byzantine-resilient stochastic gradient descent. In *9th International Conference on Learning Representations, ICLR 2021, Vienna, Austria, May 4–8, 2021*. OpenReview.net, 2021b.

Fallah, A., Mokhtari, A., and Ozdaglar, A. E. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Fang, C., Yang, Z., and Bajwa, W. U. Bridge: Byzantine-resilient decentralized gradient descent. *IEEE Transactions on Signal and Information Processing over Networks*, 8:610–626, 2022.

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

Gorbunov, E., Borzunov, A., Diskin, M., and Ryabinin, M. Secure distributed training at scale. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Guo, S., Zhang, T., Yu, H., Xie, X., Ma, L., Xiang, T., and Liu, Y. Byzantine-resilient decentralized stochastic gradient descent. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

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

Gupta, N., Doan, T. T., and Vaidya, N. H. Byzantine fault-tolerance in decentralized optimization under 2f-redundancy. In *2021 American Control Conference (ACC)*, pp. 3632–3637. IEEE, 2021.

He, L., Karimireddy, S. P., and Jaggi, M. Byzantine-robust decentralized learning via self-centered clipping. *CoRR*, abs/2202.01545, 2022.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237.

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

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 2020.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research). 2009.

Lamport, L., Shostak, R., and Pease, M. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4 (3):382–401, July 1982. ISSN 0164-0925.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Lian, X., Zhang, W., Zhang, C., and Liu, J. Asynchronous decentralized parallel stochastic gradient descent. In *ICML*, pp. 3049–3058, 2018.

Liu, S. A survey on fault-tolerance in distributed optimization and machine learning. *arXiv preprint arXiv:2106.08545*, 2021.

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

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

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

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

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. D^2: Decentralized training over decentralized data. In Dy, J. G. and Krause, A. (eds.), *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*, pp. 4855–4863. PMLR, 2018.

Tsitsiklis, J., Bertsekas, D., and Athans, M. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9):803–812, 1986.

Xiao, L. and Boyd, S. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1), 2004.

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

Yang, Z. and Bajwa, W. U. ByRDiE: Byzantine-resilient distributed coordinate descent for decentralized learning. *IEEE Transactions on Signal and Information Processing over Networks*, 5(4):611–627, 2019.

Yang, Z., Gang, A., and Bajwa, W. U. Adversary-resilient distributed and decentralized statistical inference and machine learning: An overview of recent advances under the Byzantine threat model. *IEEE Signal Processing Magazine*, 37(3), 2020.

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

Appendix

Organization

The appendices are organized as follows:

- Appendix A proves the convergence of MoNNA for $n > 11f$ (Theorem 1) and $n > 5f$ (Corollary 2).
- Appendix B analyzes D-SGD ($\beta = 0$) under (α, λ)-reduction (proof of Proposition 1).
- Appendix C explains Signed Echo Broadcast which supports the reliability of our communication model.
- Appendix D provides additional information on our experimental setup.
- Appendix E provides some additional experimental results.

A. Convergence Proof for MoNNA

In this section, we derive convergence guarantees of MoNNA. First, we prove the main result of the paper, Theorem 1, for $n > 11f$. Then, in Section A.4, we show that MoNNA can actually tolerate a larger fraction of faulty nodes (i.e., $n > 5f$) but with a slightly worse convergence rate.

We first present below the skeleton of our main proof.

A.1. Skeleton of the Proof of Theorem 1

Our proof comprises 4 key steps, listed as follows.

Step-I: Demonstrating that the coordination phase of Algorithm 1 satisfies (α, λ)-reduction.

Step-II: Analyzing the *parameter drift* and the *momentum drift*.

Step-III: Analyzing the *momentum deviation* from the true gradient.

Step-IV: Studying the *growth* of loss function $Q^{(\mathcal{C})}$.

To present the technical details, we introduce the following notation.

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

$$\mathcal{P}_t \coloneqq \left\{\theta_0^{(i)}, \ldots, \theta_t^{(i)};\ m_0^{(i)}, \ldots, m_{t-1}^{(i)};\ i = 1, \ldots, n\right\}.$$

By convention, $\mathcal{P}_0 = \{\theta_0^{(i)};\ i = 1, \ldots, n\}$. Furthermore, we denote by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \,|\, \mathcal{P}_t]$ the conditional expectation given the history $\mathcal{P}_t$, and by $\mathbb{E}[\cdot]$ the total expectation over the randomness of the algorithm; thus, $\mathbb{E}[\cdot] := \mathbb{E}_0[\cdots \mathbb{E}_T[\cdot]]$. We recall that $\mathcal{C}$ denotes the set of correct nodes, and that $|\mathcal{C}| = n - f$. For an arbitrary t, we denote by $\Gamma(*_t)$ the variance of respective correct nodes' local values, denoted by $*_t$, i.e., $\Gamma(*_t) = \frac{1}{n-f}\sum_{i\in\mathcal{C}} \left\|*_t^{(i)} - \bar{*}_t\right\|^2$, where $\bar{*}_t = \frac{1}{n-f}\sum_{i\in\mathcal{C}} *_t^{(i)}$ is the average of correct values. For instance,

$$\Gamma(\theta_t) = \frac{1}{n-f}\sum_{i\in\mathcal{C}} \left\|\theta_t^{(i)} - \bar{\theta}_t\right\|^2,\ \Gamma(m_t) = \frac{1}{n-f}\sum_{i\in\mathcal{C}} \left\|m_t^{(i)} - \bar{m}_t\right\|^2$$
$$\text{and } \Gamma(g_t) = \frac{1}{n-f}\sum_{i\in\mathcal{C}} \left\|g_t^{(i)} - \bar{g}_t\right\|^2.$$

We present below technical summaries of the aforementioned 4 steps.

Step-I: Coordination phase with NNA and satisfies (α, λ)-reduction

Recall the definition of (α, λ)-reduction from Definition 2. Here we show that this condition is satisfied by the coordination phase of Algorithm 1. Recall that in each iteration of Algorithm 1, each node i initializes the coordination phase with input $z^{(i)} = x_0^{(i)}$ and obtains the output vector $y^{(i)} = x_K^{(i)}$ at its completion. Therefore, using the $\Gamma(.)$ notation, to obtain (α, λ)-reduction, we need to prove the following two conditions

$$\Gamma(x_K) \leq \alpha \Gamma(x_0) \quad \text{and} \quad \|\bar{x}_0 - \bar{x}_K\|^2 \leq \lambda \Gamma(x_0).$$

Here, we prove that these conditions are satisfied when $n \geq 11f$. In Section A.4, Lemma 6 proves that the conditions are also satisfied when $n > 5f$, but with different values for α and λ.

Lemma 2. *Suppose that $n \geq 11f$. For any $K \geq 1$, the coordination phase of Algorithm 1 guarantees (α, λ)-reduction for*

$$\alpha = \left(\frac{9.88f}{n-f}\right)^K \quad \text{and} \quad \lambda = \frac{9f}{n-f} \cdot \min\left\{K, \frac{1}{(1-\sqrt{\alpha})^2}\right\}.$$

The proof of the lemma is provided in Section A.5.

Step-II: Parameter drift and the momentum drift

Second, we note that, at any step t, neither the momentums $m_t^{(i)}$ nor the parameters $\theta_t^{(i)}$ of the correct nodes are guaranteed to stay close to each other even when the stochastic gradients $g_t^{(i)}$ come from a common gradient oracle. Yet, given our lemmas 2, and 6, we show in Lemma 3 below that the *drift* both between the correct nodes' momentums and between their parameters can be controlled by cleverly parametrizing the momentum coefficient β. Hence, we can guarantee approximate agreement on both the parameters and the momentums of the correct nodes.

Lemma 3. *Suppose that assumptions 1, 2, and 3 hold true. Consider Algorithm 1 with $\gamma \leq \frac{1-\alpha}{L\sqrt{27\alpha(1+\alpha)}}$, and $\beta > 0$. Suppose that the coordination phase satisfies (α, λ)-reduction for $\alpha < 1$. For each $t \in [T]$, we obtain that*

$$\mathbb{E}\left[\Gamma(\theta_t)\right] \leq E(\alpha)\gamma^2\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right),$$

and

$$\mathbb{E}\left[\Gamma(m_t)\right] \leq 3\sigma^2\left(\frac{1-\beta}{1+\beta}\right) + 9\zeta^2 + 9L^2\gamma^2 E(\alpha)\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right),$$

where

$$E(\alpha) := \frac{18\alpha(1+\alpha)}{(1-\alpha)^2}.$$

Step-III: Momentum deviation.

Next, we study the *deviation* of the average correct momentum $\overline{m}_t$ from the average of the true gradients $\overline{\nabla Q}_t$, at step t. Let us denote by

$$\overline{\nabla Q}_t := \frac{1}{n-f}\sum_{i \in \mathcal{C}} \nabla Q^{(i)}(\theta_t^{(i)}),$$

the average of the true local gradient vectors at nodes' local models. We define

$$\delta_t := \overline{m}_t - \overline{\nabla Q}_t.$$

We now have the following lemma.

Lemma 4. *Suppose that assumptions 1 and 2 hold true. Consider Algorithm 1. For all $t \in [T]$, we obtain that*

$$\begin{aligned}\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] \leq \beta^2(1+4L\gamma)\left(1+\frac{9}{8}L\gamma\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{3}{4}\beta^2 L\gamma(1+4L\gamma)\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \\ + 9\beta^2L^2\left(1+\frac{1}{4\gamma L}\right)\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1}-\overline{\theta}_{t+1/2}\right\|^2\right]\right) \\ + \frac{9}{4}\beta^2 L\gamma\left(1+4L\gamma\right)L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{(1-\beta)^2\sigma^2}{n-f}.\end{aligned}$$

Step-IV: Growth function.

Finally, we analyze the growth of loss function $Q^{(\mathcal{C})}$ computed at the average parameter of the correct nodes $\overline{\theta}_t$ along the trajectory of Algorithm 1. Let us denote by $\overline{\theta}_t \coloneqq {}^1\!/_{n-f}\sum_{i\in\mathcal{C}}\theta_t^{(i)}$ the average parameter of the correct nodes at step t. Then we obtain the following lemma.

Lemma 5. *Suppose that assumptions 1 and 2 hold true. Consider Algorithm 1 with $\gamma \leq 1/L$. For each $t \in [T]$, we obtain that*

$$\begin{aligned}\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] \leq -\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + \frac{3\gamma}{2}\mathbb{E}\left[\left\|\delta_t\right\|^2\right] \\ + \frac{3}{2\gamma}\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right] + \frac{3\gamma}{2}L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right].\end{aligned}$$

This means that Algorithm 1 can actually be treated as DSGD with an additional error term which is proportional to the coupled drift of the momentums and the parameters at each step t.

Combining steps I, II, III and IV

To obtain, our final convergence result, as stated in Theorem 1, we combine these elements. Note however that the deviation term in Lemma 5 cannot be readily treated with a standard convergence analysis. To address this issue, we devise a new Lyapunov function

$$V_t \coloneqq \mathbb{E}\left[Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - Q^* + \frac{1}{4L}\left\|\delta_t\right\|^2\right]. \tag{7}$$

By analyzing the growth of V_t along the steps of Algorithm 1, we prove Theorem 1 as follows.

A.2. Proof for Theorem 1

Recall that $\mathcal{C}$ denotes the set of correct nodes, and that $|\mathcal{C}| = n - f$. Consider the Lyapunov function V_t defined in (7). Consider an arbitrary $t \in [T]$. From Lemma 5 and Lemma 4 we obtain that

$$\begin{aligned}V_{t+1} - V_t &= \mathbb{E}\left[Q^{(\mathcal{C})}\left(\overline{\theta}_{t+1}\right) - Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right] + \frac{1}{4L}\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2 - \left\|\delta_t\right\|^2\right] \\ &\leq -\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + \frac{3\gamma}{2}\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{3}{2\gamma}\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right] + \frac{3\gamma}{2}L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] \\ &\quad + \frac{1}{4L}\beta^2(1+4L\gamma)\left(1+\frac{9}{8}L\gamma\right)\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{3}{16}\beta^2\gamma(1+4L\gamma)\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \\ &\quad + \frac{9}{4}\beta^2 L\left(1+\frac{1}{4\gamma L}\right)\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1}-\overline{\theta}_{t+1/2}\right\|^2\right]\right) \\ &\quad + \frac{9}{16}\beta^2\gamma\left(1+4L\gamma\right)L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{1}{4L}\frac{(1-\beta)^2\sigma^2}{n-f} - \frac{1}{4L}\mathbb{E}\left[\left\|\delta_t\right\|^2\right].\end{aligned}$$

Upon re-arranging the terms on the R.H.S. we obtain that

$$\begin{aligned} V_{t+1} - V_t \leq & -\gamma \left(\frac{1}{2} - \frac{3}{16}\beta^2(1+4L\gamma)\right) \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \\ & + \left(\frac{3\gamma}{2} + \frac{1}{4L}\beta^2(1+4L\gamma)\left(1+\frac{9}{8}L\gamma\right) - \frac{1}{4L}\right) \mathbb{E}\left[\|\delta_t\|^2\right] \\ & + \frac{9}{4}\beta^2 L\left(1+\frac{1}{4\gamma L}\right)\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right]\right) \\ & + \frac{9}{16}\beta^2\gamma\left(1+4L\gamma\right)L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{1}{4L}\frac{(1-\beta)^2\sigma^2}{n-f} \\ & + \frac{3}{2\gamma}\,\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}\right\|^2\right] + \frac{3\gamma}{2}L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]. \end{aligned} \tag{8}$$

We denote,

$$\begin{aligned} A &\coloneqq \frac{1}{2} - \frac{3}{16}\beta^2(1+4L\gamma), \\ B &\coloneqq \frac{3\gamma}{2} + \frac{1}{4L}\beta^2(1+4L\gamma)\left(1+\frac{9}{8}L\gamma\right) - \frac{1}{4L}, \text{ and} \\ C &\coloneqq \frac{9}{4}\beta^2 L\left(1+\frac{1}{4\gamma L}\right)\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right]\right) \\ &\quad + \frac{9}{16}\beta^2\gamma\left(1+4L\gamma\right)L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{3}{2\gamma}\,\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}\right\|^2\right] + \frac{3\gamma}{2}L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]. \end{aligned}$$

Substituting from above in (8) we obtain that

$$V_{t+1} - V_t \leq -\gamma A\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + B\,\mathbb{E}\left[\|\delta_t\|^2\right] + C + \frac{1}{4L}(1-\beta)^2\frac{\sigma^2}{(n-f)}. \tag{9}$$

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

$$\gamma \leq \frac{1}{12L}, \text{ and } 1-\beta^2 = 12\gamma L. \tag{10}$$

Term A. Using the facts that $\gamma \leq 1/12L$ and that $\beta^2 < 1$, we obtain that

$$A = \frac{1}{2} - \frac{3}{16}\beta^2(1+4L\gamma) \geq \frac{1}{2} - \frac{3}{16}\left(1+\frac{4}{12}\right) = \frac{1}{4}. \tag{11}$$

Term B. As $1-\beta^2 = 12\gamma L$ and $\beta^2 < 1$, we obtain that

$$B = \frac{3\gamma}{2} - \frac{1}{4L}\left(1-\beta^2\right) + \frac{1}{4L}\beta^2\left(\frac{41}{8}L\gamma + \frac{9}{2}L^2\gamma^2\right) \leq \frac{3\gamma}{2} - 3\gamma + \frac{1}{4L}\left(\frac{41}{8}L\gamma + \frac{9}{2}L^2\gamma^2\right).$$

As $\gamma \leq 1/12L$, from above we obtain that

$$B \leq \frac{3\gamma}{2} - 3\gamma + \frac{\gamma}{4}\left(\frac{41}{8} + \frac{9}{2}L\gamma\right) \leq -\frac{3\gamma}{2} + \frac{\gamma}{4}\left(\frac{41}{8} + \frac{9}{24}\right) \leq 0. \tag{12}$$

Term C. Using the fact that $\beta^2 \leq 1$, we obtain that

$$\begin{aligned}
C &= \frac{9}{4}\beta^2 L\left(1+\frac{1}{4\gamma L}\right)\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right]+\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]+\mathbb{E}\left[\left\|\overline{\theta}_{t+1}-\overline{\theta}_{t+1/2}\right\|^2\right]\right)\\
&+\frac{9}{16}\beta^2\gamma\left(1+4L\gamma\right)L^2\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]+\frac{3}{2\gamma}\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right]+\frac{3\gamma}{2}L^2\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]\\
&\leq \mathbb{E}\left[\Gamma\left(\theta_t\right)\right]\left(\frac{9L}{16\gamma L}(1+4\gamma L)+\frac{9}{16}\gamma L^2\left(1+4L\gamma\right)+\frac{3\gamma}{2}L^2\right)\\
&+\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right]\frac{9L}{16\gamma L}(1+4\gamma L)+\left(\frac{3}{2\gamma}+\frac{9L}{16\gamma L}(1+4\gamma L)\right)\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right].
\end{aligned}$$

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

$$\begin{aligned}
C &\leq \mathbb{E}\left[\Gamma\left(\theta_t\right)\right]\left(\frac{9}{16\gamma}\left(\frac{4}{3}\right)+\frac{9}{16}\left(\frac{1}{144\gamma}\right)\left(\frac{4}{3}\right)+\frac{3}{288\gamma}\right)\\
&+\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right]\frac{9}{16\gamma}\left(\frac{4}{3}\right)+\left(\frac{3}{2\gamma}+\frac{9}{16\gamma}\left(\frac{4}{3}\right)\right)\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right]\\
&\leq \frac{1}{\gamma}\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]+\frac{1}{\gamma}\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right]+\frac{9}{4\gamma}\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right]. \qquad (13)
\end{aligned}$$

From Lemma 2, we have

$$\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right] \leq \lambda\,\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right].$$

From Algorithm 1, we have for all $i \in \mathcal{C}$, $\theta_{t+1/2}^{(i)} = \theta_t^{(i)} - \gamma m_t^{(i)}$. Therefore, by definition of $\Gamma(\cdot)$, $\Gamma\left(\theta_{t+1/2}\right) \leq 2\Gamma\left(\theta_t\right) + 2\gamma^2\Gamma\left(m_t\right)$. Thus, from above we obtain that

$$\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2}-\overline{\theta}_{t+1}\right\|^2\right] \leq \lambda\left(2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]+2\gamma^2\,\mathbb{E}\left[\Gamma\left(m_t\right)\right]\right)$$

Substituting from above in (13) we obtain that

$$C \leq \frac{1}{\gamma}\left(1+\frac{9\lambda}{2}\right)\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]+\frac{1}{\gamma}\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right]+\frac{9\lambda\gamma}{2}\mathbb{E}\left[\Gamma\left(m_t\right)\right].$$

By invoking Lemma 3, we obtain from above that

$$\begin{aligned}
C &\leq \left(2+\frac{9\lambda}{2}\right)E(\alpha)\gamma\left(\sigma^2\left(\frac{1-\beta}{1+\beta}\right)+3\zeta^2\right)\\
&+\frac{9\lambda\gamma}{2}\left(3\sigma^2\left(\frac{1-\beta}{1+\beta}\right)+9\zeta^2+9L^2\gamma^2E(\alpha)\left(\sigma^2\left(\frac{1-\beta}{1+\beta}\right)+3\zeta^2\right)\right).
\end{aligned}$$

Upon re-arranging the terms, and using the facts that $\gamma \leq 1/12L$, we obtain that

$$\begin{aligned}
C &\leq \gamma\zeta^2\left(6E(\alpha)+\frac{9\lambda}{2}\left(3E(\alpha)+9+\frac{3E(\alpha)}{16}\right)\right)+\gamma\sigma^2\left(\frac{1-\beta}{1+\beta}\right)\left(2E(\alpha)+\frac{9\lambda}{2}\left(E(\alpha)+3+\frac{E(\alpha)}{16}\right)\right)\\
&\leq \gamma\zeta^2\left(6E(\alpha)+\frac{9\lambda}{2}\left(4E(\alpha)+9\right)\right)+\gamma\sigma^2\left(\frac{1-\beta}{1+\beta}\right)\left(2E(\alpha)+\frac{9\lambda}{2}\left(2E(\alpha)+3\right)\right). \qquad (14)
\end{aligned}$$

Note that

$$\frac{1-\beta}{1+\beta} = \frac{1-\beta^2}{(1+\beta)^2} \leq 1-\beta^2 = 12\gamma L.$$

Substituting from above in (14) we obtain that

$$C \leq \gamma\zeta^2 \left(6E(\alpha) + \frac{9\lambda}{2}\left(4E(\alpha)+9\right)\right) + 12\gamma^2\sigma^2 L\left(2E(\alpha) + \frac{9\lambda}{2}\left(2E(\alpha)+3\right)\right). \tag{15}$$

Combining A, B and C. Substituting from (11) and (12) in (9), we obtain that

$$\begin{aligned} V_{t+1} - V_t &\leq -\gamma A\, \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + B\, \mathbb{E}\left[\left\|\delta_t\right\|^2\right] + C + \frac{1}{4L}(1-\beta)^2\frac{\sigma^2}{(n-f)} \\ &\leq -\frac{\gamma}{4}\, \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + C + (1-\beta)^2\frac{\sigma^2}{4L(n-f)}. \end{aligned}$$

Note that, as $\beta \in (0,\, 1)$, $1-\beta = (1-\beta^2)/(1+\beta) \leq 1-\beta^2$. Using this above we obtain that

$$V_{t+1} - V_t \leq -\frac{\gamma}{4}\, \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + C + (1-\beta^2)^2\frac{\sigma^2}{4L(n-f)}.$$

Recall that $1-\beta^2 = 12\gamma L$. Therefore,

$$V_{t+1} - V_t \leq -\frac{\gamma}{4}\, \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + C + 36\gamma^2 L\frac{\sigma^2}{n-f}.$$

This implies that

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \leq \left(V_t - V_{t+1}\right)\frac{4}{\gamma} + \frac{4}{\gamma}C + 144\gamma L\frac{\sigma^2}{n-f}. \tag{16}$$

By taking the average on both sides from $t=0$ to $T-1$, we obtain that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \leq \left(V_0 - V_T\right)\frac{4}{\gamma T} + \frac{4}{\gamma}C + 144\gamma L\frac{\sigma^2}{n-f}. \tag{17}$$

Analysis on V_t. Recall that $Q^* = \inf_\theta Q^{(\mathcal{C})}(\theta)$. Note that for any t,

$$V_t = \mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_t) - Q^*\right] + \frac{1}{4L}\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] \geq \mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_t) - Q^*\right] \geq 0.$$

Thus, $V_T \geq 0$. Using this in (17) we obtain that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \leq \left(V_0\right)\frac{4}{\gamma T} + \frac{4}{\gamma}C + 144\gamma L\frac{\sigma^2}{n-f}. \tag{18}$$

Recall that

$$V_0 = \mathbb{E}\left[Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^* + \frac{1}{4L}\left\|\delta_0\right\|^2\right].$$

Recall that, by Definition (52) of δ_t, we have $\delta_0 = \overline{m}_0 - \overline{\nabla Q}_0$. Thus, under Assumption 2, we have

$$\begin{aligned} \mathbb{E}\left[\left\|\delta_0\right\|^2\right] = \mathbb{E}\left[\left\|(1-\beta)\overline{g}_0 - \overline{\nabla Q}_0\right\|^2\right] &\leq 2(1-\beta)^2\,\mathbb{E}\left[\left\|\overline{g}_0 - \overline{\nabla Q}_0\right\|^2\right] + 2\beta^2\left\|\overline{\nabla Q}_0\right\|^2 \\ &\leq 2(1-\beta)^2\left(\frac{\sigma^2}{n-f}\right) + 2\beta^2\left\|\overline{\nabla Q}_0\right\|^2. \end{aligned}$$

Recall, from Algorithm 1, that for each correct node i, the initial model $\theta_0^{(i)}$ is identical, denoted by θ_0. Therefore, we have $\overline{\nabla Q}_0 = \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_0\right)$. Substituting this in the above we obtain that

$$\mathbb{E}\left[\left\|\delta_0\right\|^2\right] \leq 2(1-\beta)^2\left(\frac{\sigma^2}{n-f}\right) + 2\beta^2\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_0\right)\right\|^2. \tag{19}$$

Recall that $1-\beta^2 = 12\gamma L$. Thus, $(1-\beta)^2 \leq (1-\beta^2)^2/(1+\beta)^2 \leq (1-\beta^2)^2 = 144\gamma^2 L^2$. Substituting this in (19), and using the fact that $\beta^2 < 1$, we obtain that

$$\mathbb{E}\left[\|\delta_0\|^2\right] \leq 288\gamma^2 L^2 \left(\frac{\sigma^2}{n-f}\right) + 2\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_0\right)\right\|^2.$$

Therefore,

$$\begin{aligned} V_0 &\leq Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^* + \frac{1}{4L}\left(288\gamma^2 L^2 \left(\frac{\sigma^2}{n-f}\right) + 2\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_0\right)\right\|^2\right) \\ &= Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^* + 72\gamma^2 L\left(\frac{\sigma^2}{n-f}\right) + \frac{1}{2L}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_0\right)\right\|^2. \end{aligned}$$

Note that, as $Q^{(\mathcal{C})}$ is L-smooth (see Remark 1), $\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_0\right)\right\|^2 \leq 2L\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right)$. Using this in the above yields

$$V_0 \leq 2(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*) + 72\gamma^2 L\left(\frac{\sigma^2}{n-f}\right). \tag{20}$$

where in the last inequality we used Remark 1 below. Substituting from above in (18) we obtain that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \leq \frac{8\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right)}{\gamma T} + \frac{288\gamma L}{T}\left(\frac{\sigma^2}{n-f}\right) + \frac{4}{\gamma}C + 144\gamma L\frac{\sigma^2}{n-f}. \tag{21}$$

Now, note that for any correct node $i \in \mathcal{C}$, we have

$$\begin{aligned} \mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] &\leq \frac{3}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 3\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \\ &\leq \frac{3}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 3L^2\,\mathbb{E}\left[\left\|\overline{\theta}_t - \theta_t^{(i)}\right\|^2\right] \\ &\leq \frac{3}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 3L^2\sum_{j\in\mathcal{C}}\mathbb{E}\left[\left\|\overline{\theta}_t - \theta_t^{(j)}\right\|^2\right] \\ &\leq \frac{3}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 3L^2 n\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]. \end{aligned}$$

where the second inequality follows from Assumption 1. Combining this with (21), we obtain that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \leq \frac{12\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right)}{\gamma T} + \frac{432\gamma L}{T}\left(\frac{\sigma^2}{n-f}\right) + \frac{6}{\gamma}C + 216\gamma L\frac{\sigma^2}{n-f} + 3L^2 n\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right].$$

Substituting C from (15) in above and using the bound on $\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]$ from Lemma 3, we obtain that

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] &\leq \frac{12\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right)}{\gamma T} + \frac{432\gamma L}{T}\left(\frac{\sigma^2}{n-f}\right) + 216\gamma L\frac{\sigma^2}{n-f} \\ &\quad + 6\zeta^2\left(6E(\alpha) + \frac{9\lambda}{2}\left(4E(\alpha)+9\right)\right) + 72\gamma\sigma^2 L\left(2E(\alpha) + \frac{9\lambda}{2}\left(2E(\alpha)+3\right)\right) \\ &\quad + 3L^2 nE(\alpha)\gamma^2\left(2\sigma^2\left(\frac{1-\beta}{1+\beta}\right) + 3\zeta^2\right) \\ &\leq \frac{12\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right)}{\gamma T} + \frac{432\gamma L}{T}\left(\frac{\sigma^2}{n-f}\right) \\ &\quad + 72\gamma L\sigma^2\left(\frac{3}{n-f} + 2E(\alpha) + \frac{9\lambda}{2}\left(2E(\alpha)+3\right)\right) \\ &\quad + 6\zeta^2\left(6E(\alpha) + \frac{9\lambda}{2}\left(4E(\alpha)+9\right)\right) \\ &\quad + 3L^2 nE(\alpha)\gamma^2\left(2\sigma^2 + 3\zeta^2\right) \end{aligned}$$

We now define

$$c_0 := 12\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right), \quad c_1 := E(\alpha) = \frac{18\alpha(1+\alpha)}{(1-\alpha)^2},$$
$$c_2 := 72L\left(\frac{3}{n-f} + 2c_1 + \frac{9\lambda}{2}\left(2c_1+3\right)\right), \text{ and } c_3 := 6\left(6c_1 + \frac{9\lambda}{2}\left(4c_1+9\right)\right).$$

Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \leq \frac{c_0}{\gamma T} + c_2\gamma L\sigma^2 + \frac{432\gamma L}{T}\left(\frac{\sigma^2}{n-f}\right) + c_3\zeta^2 + 9c_1n\gamma^2L^2(\sigma^2+\zeta^2).$$

Now recall that

$$\gamma = \min\left\{\frac{1}{12L}, \frac{1}{L}\sqrt{\frac{2}{3c_1}}, \sqrt{\frac{c_0}{c_2LT\sigma^2}}\right\},$$

and thus

$$\frac{1}{\gamma} = \max\left\{12L, L\sqrt{\frac{3c_1}{2}}, \sqrt{\frac{c_2LT\sigma^2}{c_0}}\right\} \leq 12L + L\sqrt{\frac{3c_1}{2}} + \sqrt{\frac{c_2LT\sigma^2}{c_0}}.$$

Therefore,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \leq 2\sqrt{\frac{c_0c_2L\sigma^2}{T}} + \frac{12Lc_0}{T} + \frac{Lc_0}{T}\sqrt{\frac{3c_1}{2}} + \frac{36}{T}\left(\frac{\sigma^2}{n-f}\right) + c_3\zeta^2 + 9c_1nL(1+\frac{\zeta^2}{\sigma^2})\frac{c_0}{c_2T}$$

Denoting $c_4 := 9c_1nc_0/c_2$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \leq 2\sqrt{\frac{c_0c_2L\sigma^2}{T}} + \frac{12Lc_0}{T} + \frac{Lc_0}{T}\sqrt{\frac{3c_1}{2}} + \frac{36}{T}\left(\frac{\sigma^2}{n-f}\right) + (1+\frac{\zeta^2}{\sigma^2})\frac{c_4L}{T} + c_3\zeta^2. \quad (22)$$

As $\hat{\theta}^{(i)} \sim \mathcal{U}\{\theta_0^{(i)}, \ldots, \theta_{T-1}^{(i)}\}$, we get

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right].$$

Substituting from above in (22) concludes the proof.

Remark 1. *If a function Q is Lipschitz smooth, with coefficient L, then for all $x \in \mathbb{R}^d$, $\|\nabla Q(x)\|^2 \leq 2L\left(Q(x) - Q^*\right)$ where Q^* denotes the minimum value of Q. Proof of this fact is as follows.*

Proof. By the Lipschitzness of ∇Q, for any $x, y \in \mathbb{R}^d$, we have (see Lemma 1.2.3 (Nesterov et al., 2018))

$$Q(y) \leq Q(x) + \langle \nabla Q(x), y - x\rangle + \frac{L}{2}\|y-x\|^2.$$

Let x be an arbitrary vector in $\mathbb{R}^d$, and $y = x - \frac{1}{L}\nabla Q(x)$. Thus, from the above we obtain that

$$Q\left(x - \frac{1}{L}\nabla Q(x)\right) \leq Q(x) - \frac{1}{L}\|\nabla Q(x)\|^2 + \frac{1}{2L}\|\nabla Q(x)\|^2 = Q(x) - \frac{1}{2L}\|\nabla Q(x)\|^2.$$

As Q^* is the minimum value of Q, we have

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

Rearranging the terms we obtain that

$$\|\nabla Q(x)\|^2 \leq 2L(Q(x) - Q^*).$$

Recall that x in the above can be any vector in $\mathbb{R}^d$. The above completes the proof. □

A.3. Proof of Corollary 1

Proof. Note that ignoring the higher order terms in the bound of Theorem 1, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \in \mathcal{O}\left(\sqrt{\frac{c_0 c_2 \sigma^2}{T}} + c_3\zeta^2\right).$$

Now note also that in Theorem 1 for $n \geq 11f$, we have $\alpha \leq 0.988 < 1$. This implies that $\frac{1+\alpha}{(1-\alpha)^2} \in \mathcal{O}(1)$. Therefore,

$$c_1 = \frac{18\alpha(1+\alpha)}{(1-\alpha)^2} \in \mathcal{O}(\alpha).$$

Next, by noting that $n \geq 2f$, we obtain that

$$c_2 = 72L\left(\frac{3}{n-f} + 3c_1 + \frac{9\lambda}{2}(2c_1+3)\right) \in \mathcal{O}\left(\frac{1}{n} + \alpha + \lambda\right),$$

and

$$c_3 = 7\left(6c_1 + \frac{9\lambda}{2}(4c_1+9)\right) \in \mathcal{O}\left(\alpha+\lambda\right).$$

Finally, note that c_0 is a constant depending on the initial model and thus $c_0 \in \mathcal{O}(1)$. Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{T}\left(\frac{1}{n}+\alpha+\lambda\right)} + (\alpha+\lambda)\zeta^2\right).$$

Now note that, we have $\alpha = \frac{9.88f}{n-f} \in \mathcal{O}(\frac{f}{n})$, and $\lambda = \frac{9f}{n-f} \in \mathcal{O}(\frac{f}{n})$. Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{T}\left(\frac{1+f}{n}\right)} + \frac{f}{n}\zeta^2\right).$$

This completes the proof. □

A.4. Convergence of MoNNA for $n > 5f$

Note that Theorem 1 is stated for the case where $n \geq 11f$. This comes from the fact that we need the number of correct nodes to be sufficiently large to guarantee (α, λ)-reduction as stated in Lemma 2. However, by setting $K \in \mathcal{O}(\log(n))$ we can still guarantee (α, λ)-reduction for $n > 5f$ as stated in the following Lemma. The proof of this Lemma is given in Section A.9.

Lemma 6. *Suppose that there exists $\delta > 0$ such that $n \geq (5+\delta)f$. For $K = \frac{\log(8(n-f))}{2\log(\frac{3+\delta}{\delta})} \in \mathcal{O}(\log(n))$, the coordination phase of Algorithm 1 guarantees (α, λ)-reduction for*

$$\alpha = \frac{2f}{n-f} \leq \frac{1}{2} \quad \textit{and} \quad \lambda = \left(\frac{3+\delta}{\delta}\right)^2 \frac{(8f)^2}{n-f}.$$

Replacing Lemma 2 by Lemma 6, and following the same steps as the proof of Theorem 1 we can show the following result which is essentially a convergence proof for MoNNA while tolerating a larger fraction of faulty nodes ($n > 5f$).

Corollary 2. *Suppose that assumptions 1, 2 and 3 hold true. Suppose also that there exists $\delta > 0$ such that $n \geq (5+\delta)f$. Denote*

$$\alpha = \frac{2f}{n-f}, \quad \lambda = \left(\frac{3+\delta}{\delta}\right)^2 \frac{(8f)^2}{n-f} \quad c_0 := 12\left(Q^{(\mathcal{C})}\left(\overline{\theta}_0\right) - Q^*\right),$$
$$c_1 := \frac{18\alpha(1+\alpha)}{(1-\alpha)^2}, \quad c_2 := 72L\left(\frac{3}{n-f} + 3c_1 + \frac{9\lambda}{2}\left(2c_1+3\right)\right), \text{ and } c_3 := 7\left(6c_1 + \frac{9\lambda}{2}\left(4c_1+9\right)\right).$$

Consider Algorithm 1 with $K = \frac{\log(8(n-f))}{2\log(\frac{3+\delta}{\delta})} \in \mathcal{O}(\log(n))$, $\gamma = \min\left\{\frac{1}{12L}, \frac{1}{L}\sqrt{\frac{2}{3c_1}}, \sqrt{\frac{c_0}{c_2 LT\sigma^2}}\right\}$, and $\beta = \sqrt{1-12\gamma L}$. Then, for all $T \geq 1$, we obtain that

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq 2\sqrt{\frac{c_0 c_2 L\sigma^2}{T}} + \frac{12Lc_0}{T} + \frac{Lc_0}{T}\sqrt{\frac{3c_1}{2}} + \frac{36}{T}\left(\frac{\sigma^2}{n-f}\right) + c_3\zeta^2.$$

Remark 2. *Ignoring higher order terms and following the same reasoning as that of proof of Corollary 1, we can show that for $n > 5f$,* MoNNA *guarantees (f,ϵ)-resilience for*

$$\epsilon \in \mathcal{O}\left(\sqrt{\frac{(1+f)^2}{nT}} + \frac{f^2}{n}\zeta^2\right).$$

A.5. Proof of Lemma 2

Throughout the proof, we make use of the following notation.

Notation: Recall that $\mathcal{R}_k^{(i)}$ is the set of indices received by node i at coordination round k. Let $\psi_k^{(i)} : [n-f-1] \to \mathcal{R}_k^{(i)}$ be a bijection that sorts the elements in $\mathcal{R}_k^{(i)}$ based on the distance of their corresponding vector to $x_{k-1}^{(i)}$, i.e.,

$$\left\|x_{k-1}^{\left(\psi_k^{(i)}(1)\right)} - x_{k-1}^{(i)}\right\| \leq \ldots \leq \left\|x_{k-1}^{\left(\psi_k^{(i)}(n-f-1)\right)} - x_{k-1}^{(i)}\right\|.$$

We then denote by

$$\mathcal{S}_k^{(i)} := \left\{\psi_k^{(i)}(j) : j \in [n-2f-1]\right\} \cup \{i\}, \tag{23}$$

the set of indices of the vectors selected by the NNA function. From (3.3), we then have

$$x_k^{(i)} = \text{NNA}\left(x_{k-1}^{(i)}; \left\{x_{k-1}^{(j)} \mid j \in \mathcal{R}_k^{(i)}\right\}\right) = \frac{1}{n-2f}\sum_{j\in\mathcal{S}_k^{(i)}} x_{k-1}^{(j)}. \tag{24}$$

We first prove the following few useful lemmas.

Lemma 7. *For any set $\{x^{(i)}\}_{i\in S}$ of $|S|$ vectors, we have*

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

Proof.

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

Now as $\sum_{j\in S}(x^{(j)}-\bar{x})=0$, we have

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

□

Lemma 8. *For any pair of correct nodes p, q and coordination round $k\in[K]$ we obtain that*

$$\left|\mathcal{S}_k^{(q)}\setminus\mathcal{S}_k^{(p)}\right|=\left|\mathcal{S}_k^{(p)}\setminus\mathcal{S}_k^{(q)}\right|\leq 2f. \tag{25}$$

Proof. Consider an arbitrary pair of correct nodes p, q, and coordination round k. By definition of set $\mathcal{S}_k^{(i)}$ for all $i\in\mathcal{C}$ in (23) we obtain that

$$\left|\mathcal{S}_k^{(q)}\setminus\mathcal{S}_k^{(p)}\right|=\left|\mathcal{S}_k^{(q)}\cup\mathcal{S}_k^{(p)}\right|-\left|\mathcal{S}_k^{(p)}\right|\leq n-(n-2f)=2f.$$

□

Lemma 9. *If $n\geq 11f$ then for each coordination round $k\in[K]$ we obtain that*

$$\Gamma\left(x_k\right)\leq\alpha\Gamma\left(x_{k-1}\right)\quad\text{for}\quad\alpha=\frac{9.88f}{n-f}.$$

Proof. Consider two arbitrary correct nodes p and q in $\mathcal{C}$, and an arbitrary $k\in[K]$. We first introduce some sets that will be used later in the proof. We denote by F_p the set of faulty nodes whose local parameters are selected by p (using the NNA rule) but not by q in the k-th coordination round, i.e.,

$$F_p\coloneqq\left\{i\in[n]\setminus\mathcal{C}\,\middle|\,i\in\mathcal{S}_k^{(p)}\setminus\mathcal{S}_k^{(q)}\right\}.$$

Similarly, $F_q\coloneqq\left\{i\in[n]\setminus\mathcal{C}\,\middle|\,i\in\mathcal{S}_k^{(q)}\setminus\mathcal{S}_k^{(p)}\right\}$. Recall that by Lemma 8 we have $\left|\mathcal{S}_k^{(p)}\setminus\mathcal{S}_k^{(q)}\right|\leq 2f$. We consider an arbitrary subset $\mathcal{C}_p$ comprising correct nodes selected by node p in round k such that $|\mathcal{C}_p|+|F_p|=2f$ and $\mathcal{S}_k^{(p)}\setminus\mathcal{S}_k^{(q)}\subseteq\mathcal{C}_p$, i.e.,

$$\mathcal{C}_p\coloneqq\left\{i\in\mathcal{C}\cap\mathcal{S}_k^{(p)}\,\middle|\,|\mathcal{C}_p|+|F_p|=2f,\,\mathcal{S}_k^{(p)}\setminus\mathcal{S}_k^{(q)}\subseteq\mathcal{C}_p\right\}.$$

Similarly, $\mathcal{C}_q\coloneqq\left\{i\in\mathcal{C}\cap\mathcal{S}_k^{(q)}\,\middle|\,|\mathcal{C}_q|+|F_q|=2f,\,\mathcal{S}_k^{(q)}\setminus\mathcal{S}_k^{(p)}\subseteq\mathcal{C}_q\right\}$. We let $f_p\coloneqq|F_p|$ and $f_q\coloneqq|F_q|$. Note that $f_p+f_q\leq f$. We sort the nodes in $\mathcal{C}_p$ based on the distance of their vectors to $x_{k-1}^{(q)}$ (with ties broken arbitrarily). Let $\mathcal{C}_p[i]$ denote the i-th element in $\mathcal{C}_p$ after the sorting. Thus, we have $\left\|x_{k-1}^{(q)}-x_{k-1}^{(\mathcal{C}_p[i])}\right\|\leq\left\|x_{k-1}^{(q)}-x_{k-1}^{(\mathcal{C}_p[i+1])}\right\|$. We do the similar operation on $\mathcal{C}_q$.

By definition of NNA (24), we obtain that

$$\begin{aligned}
\left\|x_k^{(p)}-x_k^{(q)}\right\|&=\left\|\frac{1}{n-2f}\sum_{j\in\mathcal{S}_k^{(p)}}x_{k-1}^{(j)}-\frac{1}{n-2f}\sum_{j\in\mathcal{S}_k^{(q)}}x_{k-1}^{(j)}\right\|\\
&=\frac{1}{n-2f}\left\|\sum_{j\in F_p}x_{k-1}^{(j)}+\sum_{j\in\mathcal{C}_p}x_{k-1}^{(j)}-\sum_{j\in F_q}x_{k-1}^{(j)}-\sum_{j\in\mathcal{C}_q}x_{k-1}^{(j)}\right\|\\
&=\frac{1}{n-2f}\left\|\left(\sum_{j\in F_p}x_{k-1}^{(j)}-\sum_{j\in[f_p]}x_{k-1}^{(\mathcal{C}_q[j])}\right)+\left(\sum_{j\in[f+1,2f-f_p]}x_{k-1}^{(\mathcal{C}_p[j])}-\sum_{j\in[f_p+1,f]}x_{k-1}^{(\mathcal{C}_q[j])}\right)\right.\\
&\quad\left.-\left(\sum_{j\in F_q}x_{k-1}^{(j)}-\sum_{j\in[f_q]}x_{k-1}^{(\mathcal{C}_p[j])}\right)-\left(\sum_{j\in[f+1,2f-f_q]}x_{k-1}^{(\mathcal{C}_q[j])}-\sum_{j\in[f_q+1,f]}x_{k-1}^{(\mathcal{C}_p[j])}\right)\right\|.
\end{aligned}$$

Therefore,

$$\left\|x_k^{(p)} - x_k^{(q)}\right\| = \frac{1}{n-2f}\Bigg\|\left(\sum_{j\in F_p}(x_{k-1}^{(j)} - x_{k-1}^{(p)}) - \sum_{j\in[f_p]}(x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)})\right)$$
$$+\left(\sum_{j\in[f+1,2f-f_p]}(x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(p)}) - \sum_{j\in[f_p+1,f]}(x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)})\right)$$
$$-\left(\sum_{j\in F_q}(x_{k-1}^{(j)} - x_{k-1}^{(q)}) - \sum_{j\in[f_q]}(x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)})\right)$$
$$-\left(\sum_{j\in[f+1,2f-f_q]}(x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(q)}) - \sum_{j\in[f_q+1,f]}(x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)})\right)\Bigg\|.$$

Using triangle inequality above we obtain that

$$\left\|x_k^{(p)} - x_k^{(q)}\right\| \leq \frac{1}{n-2f}\left[\left(\sum_{j\in F_p}\left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\| + \sum_{j\in[f_p]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|\right)\right.$$
$$+\left(\sum_{j\in[f+1,2f-f_p]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(p)}\right\| + \sum_{j\in[f_p+1,f]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|\right)$$
$$+\left(\sum_{j\in F_q}\left\|x_{k-1}^{(j)} - x_{k-1}^{(q)}\right\| + \sum_{j\in[f_q]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)}\right\|\right)$$
$$\left.+\left(\sum_{j\in[f+1,2f-f_q]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(q)}\right\| + \sum_{j\in[f_q+1,f]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)}\right\|\right)\right]$$

As the right hand side above is a summation over $4f$ terms, we obtain that

$$\left\|x_k^{(p)} - x_k^{(q)}\right\|^2 \leq \frac{4f}{(n-2f)^2}\left[\left(\sum_{j\in F_p}\left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2 + \sum_{j\in[f_p]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|^2\right)\right.$$
$$+\left(\sum_{j\in[f+1,2f-f_p]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(p)}\right\|^2 + \sum_{j\in[f_p+1,f]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|^2\right)$$
$$+\left(\sum_{j\in F_q}\left\|x_{k-1}^{(j)} - x_{k-1}^{(q)}\right\|^2 + \sum_{j\in[f_q]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)}\right\|^2\right)$$
$$\left.+\left(\sum_{j\in[f+1,2f-f_q]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(q)}\right\|^2 + \sum_{j\in[f_q+1,f]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)}\right\|^2\right)\right]$$
$$\leq \frac{4f}{(n-2f)^2}\left[\sum_{j\in F_p}\left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2 + \sum_{j\in[f]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|^2 + \sum_{j\in[f+1,2f-f_p]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(p)}\right\|^2\right.$$
$$\left.+\sum_{j\in F_q}\left\|x_{k-1}^{(j)} - x_{k-1}^{(q)}\right\|^2 + \sum_{j\in[f]}\left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)}\right\|^2 + \sum_{j\in[f+1,2f-f_q]}\left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(q)}\right\|^2\right]. \tag{26}$$

Note that $\mathcal{S}_k^{(p)}$ contains at least f_p faulty nodes. Thus there are at most $n-2f-f_p$ correct nodes in $\mathcal{S}_k^{(p)}$. This implies that there are at least $f+f_p$ correct nodes that are not selected by node p. We define $\mathcal{C}'_p$ to be a subset of $f+f_p$ correct nodes

not in $\mathcal{S}_k^{(p)}$ that are farthest from $x_{k-1}^{(p)}$. We sort the nodes in $\mathcal{C}'_p$ such that $\left\|x_{k-1}^{(p)} - x_{k-1}^{(\mathcal{C}'_p[i])}\right\| \leq \left\|x_{k-1}^{(p)} - x_{k-1}^{(\mathcal{C}'_p[i+1])}\right\|$ for $i = 1, \ldots, f + f_p - 1$. Note that for each faulty node in $\mathcal{S}_k^{(p)}$ there is a correct node in set $\mathcal{R}_k^{(p)} \setminus \mathcal{S}_k^{(p)}$. Thus, by definition of $\mathcal{S}_k^{(p)}$ in (23) we obtain that

$$\sum_{j \in F_p} \left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2 \leq \sum_{j \in [f+1, f+f_p]} \left\|x_{k-1}^{(\mathcal{C}'_p[j])} - x_{k-1}^{(p)}\right\|^2 \tag{27}$$

By definition of $\mathcal{C}'_p$, for each $j \in [f]$, $\left\|x_{k-1}^{(p)} - x_{k-1}^{(\mathcal{C}_q[j])}\right\| \leq \left\|x_{k-1}^{(p)} - x_{k-1}^{(\mathcal{C}'_p[j])}\right\|$. Thus,

$$\sum_{j \in [f]} \left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|^2 \leq \sum_{j \in [f]} \left\|x_{k-1}^{(\mathcal{C}'_p[j])} - x_{k-1}^{(p)}\right\|^2. \tag{28}$$

From (27) and (28) we obtain that

$$\sum_{j \in F_p} \left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2 + \sum_{j \in [f]} \left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|^2 \leq \sum_{j \in \mathcal{C}'_p} \left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2.$$

Therefore, we have

$$\sum_{j \in F_p} \left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2 + \sum_{j \in [f]} \left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(p)}\right\|^2 + \sum_{j \in [f+1, 2f-f_p]} \left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(p)}\right\|^2 \leq \sum_{j \in \mathcal{C}} \left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2. \tag{29}$$

Similarly,

$$\sum_{j \in F_q} \left\|x_{k-1}^{(j)} - x_{k-1}^{(q)}\right\|^2 + \sum_{j \in [f]} \left\|x_{k-1}^{(\mathcal{C}_p[j])} - x_{k-1}^{(q)}\right\|^2 + \sum_{j \in [f+1, 2f-f_q]} \left\|x_{k-1}^{(\mathcal{C}_q[j])} - x_{k-1}^{(q)}\right\|^2 \leq \sum_{j \in \mathcal{C}} \left\|x_{k-1}^{(j)} - x_{k-1}^{(q)}\right\|^2. \tag{30}$$

Substituting from (29) and (30) in (26) we obtain that

$$\left\|x_k^{(p)} - x_k^{(q)}\right\|^2 \leq \frac{4f}{(n-2f)^2} \left[\sum_{j \in \mathcal{C}} \left\|x_{k-1}^{(j)} - x_{k-1}^{(p)}\right\|^2 + \sum_{j \in \mathcal{C}} \left\|x_{k-1}^{(j)} - x_{k-1}^{(q)}\right\|^2\right].$$

As the above holds true for an arbitrary pair of correct nodes p and q, by averaging over all such possible pairs we obtain that

$$\frac{1}{(n-f)^2} \sum_{p,q \in \mathcal{C}} \left\|x_k^{(p)} - x_k^{(q)}\right\|^2 \leq \frac{8f(n-f)}{(n-2f)^2} \frac{1}{(n-f)^2} \sum_{p,q \in \mathcal{C}} \left\|x_{k-1}^{(p)} - x_{k-1}^{(q)}\right\|^2.$$

Recall the notation $\Gamma(\cdot)$. The above implies that

$$\Gamma(x_k) \leq \frac{8f(n-f)}{(n-2f)^2} \Gamma(x_{k-1}).$$

As $n \geq 11f$, $\frac{(n-f)^2}{(n-2f)^2} \leq \frac{100}{81}$. Using this above proves the lemma, i.e.,

$$\Gamma(x_k) \leq \frac{800f}{81(n-f)} \Gamma(x_{k-1}) \leq \frac{9.88f}{n-f} \Gamma(x_{k-1}).$$

□

We now present below the proof of Lemma 2. For convenience, let us recall the lemma below.

Lemma 2. *Suppose that $n \geq 11f$. For any $K \geq 1$, the coordination phase of Algorithm 1 guarantees (α, λ)-reduction for*

$$\alpha = \left(\frac{9.88f}{n-f}\right)^K \quad \text{and} \quad \lambda = \frac{9f}{n-f} \cdot \min\left\{K, \frac{1}{(1-\sqrt{\alpha})^2}\right\}.$$

Proof. The first condition of (α, λ)-reduction stated in Definition 2, i.e., $\Gamma\left(x_K\right) \leq \alpha\Gamma\left(x_0\right)$, follows trivially from Lemma 9 for the stated value of α. We show below the second condition, i.e., $\left\|\bar{x}_0 - \bar{x}_K\right\|^2 \leq \lambda\Gamma\left(x_0\right)$, for the stated λ.

For doing so, we first consider an arbitrary round $k \in [K]$. For each correct node i, by definition of NNA operator in (24) we have that

$$\begin{aligned} x_k^{(i)} - \bar{x}_{k-1} &= \frac{1}{n-2f} \sum_{j \in \mathcal{S}_k^{(i)}} x_{k-1}^{(j)} - \frac{1}{n-f} \sum_{j \in \mathcal{C}} x_{k-1}^{(j)} \\ &= \frac{1}{n-2f} \sum_{j \in \mathcal{S}_k^{(i)}} \left(x_{k-1}^{(j)} - x_{k-1}^{(i)}\right) - \frac{1}{n-f} \sum_{j \in \mathcal{C}} \left(x_{k-1}^{(j)} - x_{k-1}^{(i)}\right). \end{aligned}$$

Upon decomposing the right hand side we obtain that

$$\begin{aligned} x_k^{(i)} - \bar{x}_{k-1} = \left(\frac{1}{n-2f} - \frac{1}{n-f}\right) \sum_{j \in \mathcal{S}_k^{(i)} \cap \mathcal{C}} \left(x_{k-1}^{(j)} - x_{k-1}^{(i)}\right) + \frac{1}{n-2f} \sum_{j \in \mathcal{S}_k^{(i)} \setminus \mathcal{C}} \left(x_{k-1}^{(j)} - x_{k-1}^{(i)}\right) \\ - \frac{1}{n-f} \sum_{j \in \mathcal{C} \setminus \mathcal{S}_k^{(i)}} \left(x_{k-1}^{(j)} - x_{k-1}^{(i)}\right). \end{aligned}$$

Thus,

$$\begin{aligned} x_k^{(i)} - \bar{x}_{k-1} = \frac{1}{(n-f)(n-2f)} \Bigg(f \sum_{j \in \mathcal{S}_k^{(i)} \cap \mathcal{C}} (x_{k-1}^{(j)} - x_{k-1}^{(i)}) + (n-f) \sum_{j \in \mathcal{S}_k^{(i)} \setminus \mathcal{C}} (x_{k-1}^{(j)} - x_{k-1}^{(i)}) \\ -(n-2f) \sum_{j \in \mathcal{C} \setminus \mathcal{S}_k^{(i)}} (x_{k-1}^{(j)} - x_{k-1}^{(i)}) \Bigg). \end{aligned}$$

By taking norm on both sides and then applying the triangle inequality we obtain that

$$\begin{aligned} \left\|x_k^{(i)} - \bar{x}_{k-1}\right\| \leq \frac{1}{(n-f)(n-2f)} \Bigg(f \sum_{j \in \mathcal{S}_k^{(i)} \cap \mathcal{C}} \left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\| \\ + (n-f) \sum_{j \in \mathcal{S}_k^{(i)} \setminus \mathcal{C}} \left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\| + (n-2f) \sum_{j \in \mathcal{C} \setminus \mathcal{S}_k^{(i)}} \left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\| \Bigg). \end{aligned} \tag{31}$$

Now let $v := \left|\mathcal{S}_k^{(i)} \cap \mathcal{C}\right|$. We then have $v = \left|\mathcal{S}_k^{(i)}\right| + |\mathcal{C}| - \left|\mathcal{S}_k^{(i)} \cup \mathcal{C}\right| \geq n - 2f + n - f - n = n - 3f$. Also, $\left|\mathcal{S}_k^{(i)} \setminus \mathcal{C}\right| = n - 2f - v$ and $\left|\mathcal{C} \setminus \mathcal{S}_k^{(i)}\right| = n - f - v$. There for the number $A(v)$ of items that are added in (31) is

$$A(v) = fv + (n-2f-v)(n-f) + (n-2f)(n-f-v) = 2(n-2f)(n-f-v), \tag{32}$$

which is decreasing in v. There the maximum of $A(v)$ is reached for $v = n - 3f$ and we have $A(v) \leq 4f(n-2f)$.

Therefore, (31) yields

$$\begin{aligned}
\left\|x_k^{(i)} - \bar{x}_{k-1}\right\|^2 &\leq \frac{4f(n-2f)}{(n-f)^2(n-2f)^2}\left(f\sum_{j\in\mathcal{S}_k^{(i)}\cap\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2\right.\\
&\quad\left.+\,(n-f)\sum_{j\in\mathcal{S}_k^{(i)}\setminus\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2 + (n-2f)\sum_{j\in\mathcal{C}\setminus\mathcal{S}_k^{(i)}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2\right)\\
&\leq \frac{4f(n-2f)}{(n-f)^2(n-2f)^2}\left(f\sum_{j\in\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2\right.\\
&\quad\left.+\,(n-f)\sum_{j\in\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2 + (n-2f)\sum_{j\in\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2\right)\\
&\leq \frac{4f(n-2f)(2n-2f)}{(n-f)^2(n-2f)^2}\sum_{j\in\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2\\
&= \frac{8f}{(n-f)(n-2f)}\sum_{j\in\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2. \qquad (33)
\end{aligned}$$

But now note that

$$\begin{aligned}
\left\|\bar{x}_k - \bar{x}_{k-1}\right\|^2 &= \left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}} x_k^{(i)} - \bar{x}_{k-1}\right\|^2\\
&\leq \frac{1}{n-f}\sum_{i\in\mathcal{C}}\left\|x_k^{(i)} - \bar{x}_{k-1}\right\|^2.
\end{aligned}$$

Combining above with (33) then yields

$$\left\|\bar{x}_k - \bar{x}_{k-1}\right\|^2 \leq \frac{8f}{n-2f}\cdot\frac{1}{(n-f)^2}\sum_{i,j\in\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|^2.$$

Using the notation $\Gamma(\cdot)$, we then have

$$\left\|\bar{x}_k - \bar{x}_{k-1}\right\|^2 \leq \frac{8f}{n-2f}\Gamma\left(x_{k-1}\right) \leq \frac{8f\alpha^{k-1}}{n-2f}\Gamma\left(x_0\right), \qquad (34)$$

where in the second inequality we used Lemma 9. Now note that

$$\begin{aligned}
\left\|\bar{x}_K - \bar{x}_0\right\|^2 &= \left\|\sum_{k\in[K]}\left(\bar{x}_k - \bar{x}_{k-1}\right)\right\|^2\\
&= \left\langle\sum_{k\in[K]}\left(\bar{x}_k - \bar{x}_{k-1}\right), \sum_{k\in[K]}\left(\bar{x}_k - \bar{x}_{k-1}\right)\right\rangle\\
&= \sum_{k,l\in[K]}\langle\bar{x}_k - \bar{x}_{k-1},\, \bar{x}_l - \bar{x}_{l-1}\rangle.
\end{aligned}$$

By the Cauchy–Schwarz inequality we then have

$$\left\|\bar{x}_K - \bar{x}_0\right\|^2 \leq \sum_{k,l\in[K]}\sqrt{\left\|\bar{x}_k - \bar{x}_{k-1}\right\|^2\cdot\left\|\bar{x}_l - \bar{x}_{l-1}\right\|^2}.$$

Combining this with 34, we obtain that

$$\begin{aligned}\left\|\bar{x}_K - \bar{x}_0\right\|^2 &\leq \frac{8f}{n-2f}\Gamma\left(x_0\right) \sum_{k,l\in[K]} \sqrt{\alpha^{k-1}\alpha^{l-1}} \\ &= \frac{8f}{n-2f}\Gamma\left(x_0\right) \sum_{k\in[K]} \left(\sqrt{\alpha}\right)^{k-1} \sum_{l\in[K]} \left(\sqrt{\alpha}\right)^{l-1}.\end{aligned}$$

Now since $\alpha < 1$ we have $\sum_{k\in[K]} \left(\sqrt{\alpha}\right)^{k-1} \leq K$, and thus

$$\left\|\bar{x}_K - \bar{x}_0\right\|^2 \leq \frac{8fK^2}{n-2f}\Gamma\left(x_0\right). \tag{35}$$

Moreover, we have

$$\sum_{k\in[K]} \left(\sqrt{\alpha}\right)^{k-1} \leq \sum_{k=1}^{\infty} \left(\sqrt{\alpha}\right)^{k-1} = \frac{1}{1-\sqrt{\alpha}},$$

and thus

$$\left\|\bar{x}_K - \bar{x}_0\right\|^2 \leq \frac{8f}{n-2f} \cdot \frac{1}{(1-\sqrt{\alpha})^2}\Gamma\left(x_0\right). \tag{36}$$

Combining (35) and (36) and noting that $\frac{8f}{n-2f} \leq \frac{9f}{n-f}$ for $n \geq 11f$ proves the lemma. □

A.6. Proof of Lemma 3

We recall the lemma below.

Lemma 3. *Suppose that assumptions 1, 2, and 3 hold true. Consider Algorithm 1 with $\gamma \leq \frac{1-\alpha}{L\sqrt{27\alpha(1+\alpha)}}$, and $\beta > 0$. Suppose that the coordination phase satisfies (α, λ)-reduction for $\alpha < 1$. For each $t \in [T]$, we obtain that*

$$\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] \leq E(\alpha)\gamma^2\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right),$$

and

$$\mathbb{E}\left[\Gamma\left(m_t\right)\right] \leq 3\sigma^2\left(\frac{1-\beta}{1+\beta}\right) + 9\zeta^2 + 9L^2\gamma^2E(\alpha)\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right),$$

where

$$E(\alpha) := \frac{18\alpha(1+\alpha)}{(1-\alpha)^2}.$$

Proof. Consider an arbitrary step $t \in [T]$. The proof comprises 3 steps.

Step i. In this step, we analyse the growth of $\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]$. From Algorithm 1 recall that for all $i \in \mathcal{C}$, we have $\theta_{t+1/2}^{(i)} = \theta_t^{(i)} - \gamma m_t^{(i)}$. As $(x+y)^2 \leq (1+c)x^2 + (1+1/c)y^2$ for any $c > 0$, we obtain for all $i,\, j \in \mathcal{C}$ that

$$\begin{aligned}\mathbb{E}\left[\left\|\theta_{t+1/2}^{(i)} - \theta_{t+1/2}^{(j)}\right\|^2\right] &\leq \mathbb{E}\left[\left\|\theta_t^{(i)} - \theta_t^{(j)} - \gamma\left(m_t^{(i)} - m_t^{(j)}\right)\right\|^2\right] \\ &\leq (1+c)\,\mathbb{E}\left[\left\|\theta_t^{(i)} - \theta_t^{(j)}\right\|^2\right] + \left(1+\frac{1}{c}\right)\gamma^2\,\mathbb{E}\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right].\end{aligned}$$

Thus, by definition of notation $\Gamma(*_t)$ and using Lemma 7, we have

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right] \leq (1+c)\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \left(1+\frac{1}{c}\right)\gamma^2\,\mathbb{E}\left[\Gamma\left(m_t\right)\right]. \tag{37}$$

Recall that, the coordination phase of Algorithm 1 satisfies (α,λ)-reduction. Thus, for all t, we have $\Gamma\left(\theta_{t+1}\right) \leq \alpha\Gamma\left(\theta_{t+1/2}\right)$. Substituting from above we obtain that

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq (1+c)\alpha\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \left(1+\frac{1}{c}\right)\alpha\gamma^2\,\mathbb{E}\left[\Gamma\left(m_t\right)\right]. \tag{38}$$

Step ii. In this step, we analyse the growth of $\mathbb{E}\left[\Gamma\left(m_t\right)\right]$. From the definition of momentum in (6), we obtain for all $i\,j \in \mathcal{C}$ that

$$\begin{aligned}
\mathbb{E}\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right] &= \mathbb{E}\left[\left\|(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\left(g_s^{(i)} - g_s^{(j)}\right)\right\|^2\right] \\
&= (1-\beta)^2\,\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right) + \nabla Q^{(i)}\left(\theta_s^{(i)}\right) - \nabla Q^{(j)}\left(\theta_s^{(j)}\right) + \nabla Q^{(j)}\left(\theta_s^{(j)}\right) - g_s^{(j)}\right)\right\|^2\right].
\end{aligned}$$

Using the fact that $(x+y+z)^2 \leq 3x^2+3y^2+3z^2$, from above we obtain that

$$\begin{aligned}
\mathbb{E}\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right] \leq\ & 3(1-\beta)^2\,\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right] \\
&+ 3(1-\beta)^2\,\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}\left(g_s^{(j)} - \nabla Q^{(j)}\left(\theta_s^{(j)}\right)\right)\right\|^2\right] \\
&+ 3(1-\beta)^2\,\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}\left(\nabla Q^{(i)}\left(\theta_s^{(i)}\right) - \nabla Q^{(j)}\left(\theta_s^{(j)}\right)\right)\right\|^2\right].
\end{aligned} \tag{39}$$

Consider an arbitrary $i \in \mathcal{C}$, and denote

$$A_t \coloneqq \mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right]. \tag{40}$$

Note that

$$\begin{aligned}
A_t &= \mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right) + \left(g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right)\right\|^2\right].
\end{aligned}$$

From above we obtain that

$$\begin{aligned}
A_t = \mathbb{E}&\left[\left\|\sum_{s=1}^{t-1}\beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right] + \mathbb{E}\left[\left\|g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\|^2\right] \\
&+ \mathbb{E}\left[\left\langle\sum_{s=1}^{t-1}\beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right),\, g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\rangle\right].
\end{aligned}$$

Recall that in the above, $\mathbb{E}[\cdot] = \mathbb{E}_1[\ldots \mathbb{E}_t[\cdot]]$. Thus, due to Assumption 2, we have $\mathbb{E}\left[\left\|g_t^{(i)} - \nabla Q^{(i)} Q\left(\theta_t^{(i)}\right)\right\|^2\right] \leq \sigma^2$. Using this above we obtain that

$$A_t \leq \mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right] + \sigma^2$$
$$+ \mathbb{E}\left[\left\langle \sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right), g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\rangle\right]. \tag{41}$$

Also, by tower rule we have

$$\mathbb{E}\left[\left\langle \sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right), g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\rangle\right] =$$
$$\mathbb{E}_1\left[\ldots \mathbb{E}_t\left[\left\langle \sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right), g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\rangle\right]\right].$$

By the definition of conditional expectation $\mathbb{E}_t[\cdot]$, we have

$$\mathbb{E}_t\left[\left\langle \sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right), g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\rangle\right] =$$
$$\left\langle \sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right), \mathbb{E}_t\left[g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right]\right\rangle.$$

By Assumption 2, we obtain that $\mathbb{E}_t\left[g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right] = \nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(i)}\left(\theta_t^{(i)}\right) = 0$. Using this above implies that

$$\mathbb{E}\left[\left\langle \sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right), g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\rangle\right] = 0.$$

Substituting from above in (41) we obtain that

$$\begin{aligned} A_t &\leq \mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right] + \sigma^2 \\ &= \beta^2\, \mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \beta^{t-1-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right] + \sigma^2 \\ &= \beta^2 A_{t-1} + \sigma^2. \end{aligned}$$

Note from the definition of A_t in (40) that, under Assumption 2, $A_1 \leq \sigma^2$. Thus, from above we obtain that

$$A_t \coloneqq \mathbb{E}\left[\left\|\sum_{s=1}^{t} \beta^{t-s}\left(g_s^{(i)} - \nabla Q^{(i)}\left(\theta_s^{(i)}\right)\right)\right\|^2\right] \leq \sigma^2 \sum_{s=0}^{t-1} \beta^{2s} \leq \frac{\sigma^2}{1-\beta^2}$$

Substituting from above in (39), computing the pair-wise average, and using Lemma 7, we obtain that

$$\mathbb{E}[\Gamma(m_t)] = \frac{1}{2(n-f)^2} \sum_{i,j \in \mathcal{C}} \mathbb{E}\left[\left\|m_t^{(i)} - m_t^{(j)}\right\|^2\right]$$
$$\leq 3(1-\beta)^2 \frac{\sigma^2}{1-\beta^2} + \frac{3(1-\beta)^2}{2(n-f)^2} \sum_{i,j \in \mathcal{C}} \mathbb{E}\left[\left\|\sum_{s=1}^{t} \beta^{t-s}\left(\nabla Q^{(i)}\left(\theta_s^{(i)}\right) - \nabla Q^{(j)}\left(\theta_s^{(j)}\right)\right)\right\|^2\right]. \tag{42}$$

Let us denote

$$C_t \coloneqq \frac{1}{(n-f)^2} \sum_{i,j \in \mathcal{C}} \mathbb{E}\left[\left\|\sum_{s=1}^{t} \beta^{t-s}\left(\nabla Q^{(i)}\left(\theta_s^{(i)}\right) - \nabla Q^{(j)}\left(\theta_s^{(j)}\right)\right)\right\|^2\right]. \tag{43}$$

Now note that

$$C_t = \frac{1}{(n-f)^2} \sum_{i,j \in \mathcal{C}} \mathbb{E}\left[\left\|\beta \sum_{s=1}^{t-1} \beta^{t-1-s}\left(\nabla Q^{(i)}\left(\theta_s^{(i)}\right) - \nabla Q^{(j)}\left(\theta_s^{(j)}\right)\right) + \left(\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(j)}\left(\theta_t^{(j)}\right)\right)\right\|^2\right]$$

By Jensen's inequality, we obtain that

$$\begin{aligned}
C_t &\leq \frac{1}{(n-f)^2} \sum_{i,j \in \mathcal{C}} \beta\, \mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \beta^{t-1-s}\left(\nabla Q^{(i)}\left(\theta_s^{(i)}\right) - \nabla Q^{(j)}\left(\theta_s^{(j)}\right)\right)\right\|^2\right] \\
&\quad + \frac{1}{(n-f)^2} \sum_{i,j \in \mathcal{C}} (1-\beta)\, \mathbb{E}\left[\left\|\frac{1}{1-\beta}\left(\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(j)}\left(\theta_t^{(j)}\right)\right)\right\|^2\right] \\
&= \beta C_{t-1} + \frac{1}{(n-f)^2(1-\beta)} \sum_{i,j \in \mathcal{C}} \mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(j)}\left(\theta_t^{(j)}\right)\right\|^2\right].
\end{aligned}$$

Now using the fact that $(x+y+z)^2 \leq 3x^2 + 3y^2 + 3z^2$, we obtain that

$$\begin{aligned}
&\left\|\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(j)}\left(\theta_t^{(j)}\right)\right\|^2 \\
&= \left\|\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(i)}\left(\overline{\theta}_t\right) + \nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(j)}\left(\overline{\theta}_t\right) + \nabla Q^{(j)}\left(\overline{\theta}_t\right) - \nabla Q^{(j)}\left(\theta_t^{(j)}\right)\right\|^2 \\
&\leq 3\left\|\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2 + 3\left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(j)}\left(\overline{\theta}_t\right)\right\|^2 + 3\left\|\nabla Q^{(j)}\left(\overline{\theta}_t\right) - \nabla Q^{(j)}\left(\theta_t^{(j)}\right)\right\|^2.
\end{aligned}$$

By Assumption 1, we have that $\left\|\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2 \leq L^2\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2$ and $\left\|\nabla Q^{(j)}\left(\theta_t^{(j)}\right) - \nabla Q^{(j)}\left(\overline{\theta}_t\right)\right\|^2 \leq L^2\left\|\theta_t^{(j)} - \overline{\theta}_t\right\|^2$. Using this above we obtain that

$$\begin{aligned}
C_t &\leq \beta C_{t-1} + \frac{3L^2}{(n-f)^2(1-\beta)} \sum_{i,j \in \mathcal{C}} \left(\mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] + \mathbb{E}\left[\left\|\theta_t^{(j)} - \overline{\theta}_t\right\|^2\right]\right) \\
&\quad + \frac{3}{(n-f)^2(1-\beta)} \sum_{i,j \in \mathcal{C}} \left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(j)}\left(\overline{\theta}_t\right)\right\|^2.
\end{aligned} \tag{44}$$

Now by Lemma 7 and Assumption 3, we have

$$\frac{1}{(n-f)^2} \sum_{i,j \in \mathcal{C}} \left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(j)}\left(\overline{\theta}_t\right)\right\|^2 = \frac{2}{n-f} \sum_{i \in \mathcal{C}} \left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 \leq 2\zeta^2.$$

Combining above with (44), we obtain that

$$\begin{aligned}
C_t &\leq \beta C_{t-1} + \frac{6L^2}{(n-f)(1-\beta)} \sum_{i \in \mathcal{C}} \mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] + \frac{12\zeta^2}{1-\beta} \\
&= \beta C_{t-1} + \frac{6L^2}{(1-\beta)}\, \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{6\zeta^2}{1-\beta}.
\end{aligned} \tag{45}$$

Now note that by definition, $C_0 = 0$. Thus, from above we obtain that

$$\begin{aligned} C_t &\leq \frac{6L^2}{1-\beta} \sum_{s=1}^{t} \beta^{t-s} \Gamma\left(\theta_s\right) + \frac{6\zeta^2}{1-\beta} \sum_{s=1}^{t} \beta^{t-s} \\ &\leq \frac{6L^2}{1-\beta} \sum_{s=1}^{t} \beta^{t-s} \Gamma\left(\theta_s\right) + \frac{6\zeta^2}{1-\beta} \sum_{s=0}^{\infty} \beta^{s} \\ &= \frac{6L^2}{1-\beta} \sum_{s=1}^{t} \beta^{t-s} \Gamma\left(\theta_s\right) + \frac{6\zeta^2}{(1-\beta)^2}. \end{aligned}$$

Substituting from above in (42) we obtain that

$$\mathbb{E}\left[\Gamma\left(m_t\right)\right] \leq 3\sigma^2 \left(\frac{1-\beta}{1+\beta}\right) + 9(1-\beta)L^2 \sum_{s=1}^{t} \beta^{t-s} \mathbb{E}\left[\Gamma\left(\theta_s\right)\right] + 9\zeta^2. \tag{46}$$

Recall from (38) that

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq (1+c)\alpha \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \left(1 + \frac{1}{c}\right) \alpha\gamma^2 \mathbb{E}\left[\Gamma\left(m_t\right)\right]. \tag{47}$$

In the next and the final step we use the results derived in (46) and (47) above to conclude the proof.

Step iii. Now in (47) we let

$$c = \frac{1-\alpha}{2\alpha} > 0. \tag{48}$$

Substituting this in (47) we obtain that

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq \frac{1+\alpha}{2} \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \left(\frac{1+\alpha}{1-\alpha}\right) \alpha\gamma^2 \mathbb{E}\left[\Gamma\left(m_t\right)\right].$$

Substituting from (46) above we obtain that

$$\begin{aligned} \mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] &\leq \left(\frac{1+\alpha}{2}\right) \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] \\ &\quad + \frac{\alpha(1+\alpha)}{1-\alpha}\gamma^2 \left(3\sigma^2 \left(\frac{1-\beta}{1+\beta}\right) + 9\zeta^2 + 9(1-\beta)L^2 \sum_{s=1}^{t} \beta^{t-s} \mathbb{E}\left[\Gamma\left(\theta_s\right)\right]\right) \\ &= \left(\frac{1+\alpha}{2}\right) \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{\alpha(1+\alpha)}{1-\alpha} \left(3\sigma^2 \frac{1-\beta}{1+\beta} + 9\zeta^2\right) \\ &\quad + 9(1-\beta)\gamma^2 L^2 \left(\frac{\alpha(1+\alpha)}{1-\alpha}\right) \sum_{s=1}^{t} \beta^{t-s} \mathbb{E}\left[\Gamma\left(\theta_s\right)\right]. \end{aligned} \tag{49}$$

As we assume that $\gamma \leq \frac{1-\alpha}{L\sqrt{27\alpha(1+\alpha)}}$, we have

$$\gamma^2 L^2 \leq \frac{(1-\alpha)^2}{27\alpha\left(1+\alpha\right)}.$$

Using the above in (49), we obtain that

$$\begin{aligned} \mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] &\leq \left(\frac{1+\alpha}{2}\right) \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{\alpha(1+\alpha)\gamma^2}{1-\alpha} \left(\frac{1-\beta}{1+\beta} 3\sigma^2 + 9\zeta^2\right) \\ &\quad + \left(\frac{1-\alpha}{3}\right) (1-\beta) \sum_{s=1}^{t} \beta^{t-s} \mathbb{E}\left[\Gamma\left(\theta_s\right)\right]. \end{aligned}$$

For convenience, we denote

$$D = \frac{\alpha(1+\alpha)\gamma^2}{1-\alpha}\left(3\sigma^2\frac{1-\beta}{1+\beta} + 9\zeta^2\right).$$

Thus,

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq \left(\frac{1+\alpha}{2}\right)\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + D + \left(\frac{1-\alpha}{3}\right)(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\,\mathbb{E}\left[\Gamma\left(\theta_s\right)\right]. \tag{50}$$

As (50) above holds true for an arbitrary $t \in [T]$, we reason below by mathematical induction that for all t,

$$\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] \leq \frac{6D}{1-\alpha}. \tag{51}$$

First, note that, as $\theta_1^{(i)} = \theta_1^{(j)}$ for all $i, j \in \mathcal{C}$, the above is trivially true for $t = 1$. Second, let us assume that (51) is true for all $t \leq k$. Then, from (50) we obtain that

$$\begin{aligned}\mathbb{E}\left[\Gamma\left(\theta_{k+1}\right)\right] &\leq \left(\frac{1+\alpha}{2}\right)\frac{6D}{1-\alpha} + D + \left(\frac{1-\alpha}{3}\right)(1-\beta)\sum_{s=1}^{k}\beta^{k-s}\frac{6D}{1-\alpha}\\ &= \left(\frac{1+\alpha}{2}\right)\frac{6D}{1-\alpha} + D + 2D = \frac{6D}{1-\alpha}.\end{aligned}$$

Thus, (51) holds true for $k+1$. Therefore, for all $t \in [T]$, we have

$$\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] \leq \frac{\alpha(1+\alpha)\gamma^2}{(1-\alpha)^2}\left(18\sigma^2\frac{1-\beta}{1+\beta} + 54\zeta^2\right)$$

We now define $E(\alpha) := \frac{18\alpha(1+\alpha)}{(1-\alpha)^2}$. We then have

$$\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] \leq E(\alpha)\gamma^2\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right).$$

Combining this with (46), we then obtain

$$\begin{aligned}\mathbb{E}\left[\Gamma\left(m_t\right)\right] &\leq 3\sigma^2\left(\frac{1-\beta}{1+\beta}\right) + 9(1-\beta)L^2\sum_{s=1}^{t}\beta^{t-s}\,\mathbb{E}\left[\Gamma\left(\theta_s\right)\right] + 9\zeta^2\\ &\leq 3\sigma^2\left(\frac{1-\beta}{1+\beta}\right) + 9\zeta^2 + 9(1-\beta)L^2\sum_{s=1}^{t}\beta^{t-s}\left(E(\alpha)\gamma^2\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right)\right)\\ &\leq 3\sigma^2\left(\frac{1-\beta}{1+\beta}\right) + 9\zeta^2 + 9(1-\beta)L^2\sum_{s=0}^{\infty}\beta^{s}\left(E(\alpha)\gamma^2\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right)\right)\\ &= 3\sigma^2\left(\frac{1-\beta}{1+\beta}\right) + 9\zeta^2 + 9L^2\gamma^2E(\alpha)\left(\sigma^2\frac{1-\beta}{1+\beta} + 3\zeta^2\right)\end{aligned}$$

□

A.7. Proof of Lemma 4

Recall that

$$\overline{\nabla Q}_t \coloneqq \frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}(\theta_t^{(i)}),$$

and that

$$\delta_t \coloneqq \overline{m}_t - \overline{\nabla Q}_t. \tag{52}$$

Also, we recall the lemma below.

Lemma 4. *Suppose that assumptions 1 and 2 hold true. Consider Algorithm 1. For all $t \in [T]$, we obtain that*

$$\begin{aligned}\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] \leq{}& \beta^2(1+4L\gamma)(1+\frac{9}{8}L\gamma)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{3}{4}\beta^2 L\gamma(1+4L\gamma)\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \\ &+ 9\beta^2L^2\left(1+\frac{1}{4\gamma L}\right)\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1}-\overline{\theta}_{t+1/2}\right\|^2\right]\right) \\ &+ \frac{9}{4}\beta^2L\gamma\left(1+4L\gamma\right)L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{(1-\beta)^2\sigma^2}{n-f}.\end{aligned}$$

Proof. Consider an arbitrary step $t \geq 1$. Recall from Algorithm 1 that

$$\overline{m}_{t+1} := \beta\overline{m}_t + (1-\beta)\overline{g}_{t+1}.$$

Combining this with (52) we obtain that

$$\delta_{t+1} = \beta\overline{m}_t + (1-\beta)\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}.$$

Adding and subtracting $\beta\overline{\nabla Q}_t$ and $\beta\overline{\nabla Q}_{t+1}$ on the R.H.S. we then obtain that

$$\delta_{t+1} = \beta\left(\overline{m}_t - \overline{\nabla Q}_t\right) + \beta\left(\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right) + (1-\beta)\left(\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right),$$

which yields

$$\begin{aligned}\left\|\delta_{t+1}\right\|^2 = {}& \beta^2\left\|\overline{m}_t - \overline{\nabla Q}_t\right\|^2 + \beta^2\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2 + (1-\beta)^2\left\|\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right\|^2 \\ &+ 2\beta^2\left\langle\overline{m}_t - \overline{\nabla Q}_t,\, \overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\rangle + 2\beta(1-\beta)\left\langle\overline{g}_{t+1} - \overline{\nabla Q}_{t+1},\, \overline{m}_t - \overline{\nabla Q}_t\right\rangle \\ &+ 2\beta(1-\beta)\left\langle\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1},\, \overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right\rangle.\end{aligned}$$

Applying the conditional expectation $\mathbb{E}_{t+1}[\cdot]$ on both sides, and noting that $\overline{m}_t$, $\overline{\nabla Q}_t$, and $\overline{\nabla Q}_{t+1}$ are deterministic values when given $\mathcal{P}_{t+1}$, we obtain that

$$\begin{aligned}\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] = {}& \beta^2\mathbb{E}_{t+1}\left[\left\|\delta_t\right\|^2\right] + \beta^2\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2 + (1-\beta)^2\mathbb{E}_{t+1}\left[\left\|\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right\|^2\right] \\ &+ 2\beta^2\left\langle\delta_t,\, \overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\rangle + 2\beta(1-\beta)\left\langle\mathbb{E}_{t+1}\left[\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right],\, \overline{m}_t - \overline{\nabla Q}_t\right\rangle \\ &+ 2\beta(1-\beta)\left\langle\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1},\, \mathbb{E}_{t+1}\left[\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right]\right\rangle.\end{aligned}$$

Owing to Assumption 2, $\mathbb{E}_{t+1}\left[g_{t+1}^{(i)}\right] = \nabla Q^{(i)}\left(\theta_{t+1}^{(i)}\right)$ for all $i \in \mathcal{C}$. Thus, we have $\mathbb{E}_{t+1}\left[\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right] = 0$. Therefore,

$$\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] = \beta^2\left\|\delta_t\right\|^2 + \beta^2\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2 + (1-\beta)^2\mathbb{E}_{t+1}\left[\left\|\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right\|^2\right] + 2\beta^2\left\langle\delta_t,\, \overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\rangle. \tag{53}$$

Note that

$$\begin{aligned}(1-\beta)^2\mathbb{E}_{t+1}\left[\left\|\overline{g}_{t+1} - \overline{\nabla Q}_{t+1}\right\|^2\right] &= (1-\beta)^2\mathbb{E}_{t+1}\left[\left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}}\left(g_{t+1}^{(i)} - \nabla Q^{(i)}\left(\theta_{t+1}^{(i)}\right)\right)\right\|^2\right] \\ &= \frac{(1-\beta)^2}{(n-f)^2}\sum_{i,j\in\mathcal{C}}\mathbb{E}_{t+1}\left[\left\langle g_{t+1}^{(i)} - \nabla Q^{(i)}\left(\theta_{t+1}^{(i)}\right),\, g_{t+1}^{(j)} - \nabla Q^{(j)}\left(\theta_{t+1}^{(j)}\right)\right\rangle\right] \\ &\overset{(a)}{=} \frac{(1-\beta)^2}{(n-f)^2}\sum_{i\in\mathcal{C}}\mathbb{E}_{t+1}\left[\left\|g_{t+1}^{(i)} - \nabla Q^{(i)}\left(\theta_{t+1}^{(i)}\right)\right\|^2\right] \overset{(b)}{\leq} \frac{(1-\beta)^2\sigma^2}{n-f},\end{aligned} \tag{54}$$

where (a) uses the facts that the gradient estimations are independent and $\mathbb{E}_{t+1}\left[g_{t+1}^{(i)}\right] - \nabla Q^{(i)}\left(\theta_{t+1}^{(i)}\right) = 0$, and (b) is due to Assumption 2. Substituting from (54) in (53) we obtain that (upon applying Cauchy-Schwartz inequality)

$$\begin{aligned}\mathbb{E}_{t+1}\left[\left\|\delta_{t+1}\right\|^2\right] &\leq \beta^2\left\|\delta_t\right\|^2 + \beta^2\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2 + 2\beta^2\left\langle \delta_t,\, \overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\rangle + \frac{(1-\beta)^2\sigma^2}{n-f}\\ &\leq \beta^2\left\|\delta_t\right\|^2 + \beta^2\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2 + 2\beta^2\left\|\delta_t\right\|\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\| + \frac{(1-\beta)^2\sigma^2}{n-f}.\end{aligned}$$

Upon taking total expectation on both sides above we obtain that

$$\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] \leq \beta^2\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \beta^2\,\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2\right] + 2\beta^2\,\mathbb{E}\left[\left\|\delta_t\right\|\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|\right] + \frac{(1-\beta)^2\sigma^2}{n-f}.$$

As $2xy \leq cx^2 + \frac{y^2}{c}$ for all $c > 0$, by substituting $c = 4\gamma L$ we obtain that $2\left\|\delta_t\right\|\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\| \leq 4\gamma L\left\|\delta_t\right\|^2 + \frac{1}{4\gamma L}\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2$. Using this above we obtain that

$$\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] \leq \beta^2(1+4\gamma L)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \beta^2\left(1+\frac{1}{4\gamma L}\right)\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2\right] + \frac{(1-\beta)^2\sigma^2}{n-f}. \tag{55}$$

Note that as $\overline{\nabla Q}_t \coloneqq \frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}\left(\theta_t^{(i)}\right)$, by Assumption 1 we obtain that

$$\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\| = \frac{1}{n-f}\sum_{i\in\mathcal{C}}\left\|\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(i)}\left(\theta_{t+1}^{(i)}\right)\right\| \leq \frac{L}{n-f}\sum_{i\in\mathcal{C}}\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\|.$$

This implies that

$$\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2 \leq L^2\left(\frac{1}{n-f}\sum_{i\in\mathcal{C}}\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\|\right)^2 \leq \frac{L^2}{n-f}\sum_{i\in\mathcal{C}}\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\|^2. \tag{56}$$

By triangle inequality we obtain that

$$\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\| \leq \left\|\theta_{t+1}^{(i)} - \overline{\theta}_{t+1}\right\| + \left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\| + \left\|\overline{\theta}_{t+1/2} - \theta_t^{(i)}\right\|.$$

Recall that $\overline{\theta}_{t+1/2} = \overline{\theta}_t - \gamma\overline{m}_t$. Thus, $\left\|\overline{\theta}_{t+1/2} - \theta_t^{(i)}\right\| \leq \left\|\overline{\theta}_t - \theta_t^{(i)}\right\| + \gamma\left\|\overline{m}_t\right\|$ and

$$\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\| \leq \left\|\theta_{t+1}^{(i)} - \overline{\theta}_{t+1}\right\| + \left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\| + \left\|\overline{\theta}_t - \theta_t^{(i)}\right\| + \gamma\left\|\overline{m}_t\right\|.$$

As $(x+y)^2 \leq (1+c)x^2 + (1+\frac{1}{c})y^2$, taking square on both sides for $c = 2$ we obtain that

$$\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\|^2 \leq 3\left(\left\|\theta_{t+1}^{(i)} - \overline{\theta}_{t+1}\right\| + \left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\| + \left\|\overline{\theta}_t - \theta_t^{(i)}\right\|\right)^2 + \frac{3}{2}\gamma^2\left\|\overline{m}_t\right\|^2.$$

And thus

$$\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\|^2 \leq 9\left\|\theta_{t+1}^{(i)} - \overline{\theta}_{t+1}\right\|^2 + 9\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2 + 9\left\|\overline{\theta}_t - \theta_t^{(i)}\right\|^2 + \frac{3}{2}\gamma^2\left\|\overline{m}_t\right\|^2. \tag{57}$$

Note that for any t, by definition of $\overline{\theta}_t$ we have

$$\left\|\overline{\theta}_t - \theta_t^{(i)}\right\|^2 = \left\|\frac{1}{n-f}\sum_{j\in\mathcal{C}}\left(\theta_t^{(j)} - \theta_t^{(i)}\right)\right\|^2 \leq \left(\frac{1}{n-f}\sum_{j\in\mathcal{C}}\left\|\theta_t^{(j)} - \theta_t^{(i)}\right\|\right)^2 \leq \frac{1}{n-f}\sum_{j\in\mathcal{C}}\left\|\theta_t^{(j)} - \theta_t^{(i)}\right\|^2.$$

Substituting from above in (57) we obtain that

$$\left\|\theta_t^{(i)} - \theta_{t+1}^{(i)}\right\|^2 \leq \frac{9}{n-f}\left(\sum_{j\in\mathcal{C}}\left\|\theta_{t+1}^{(j)} - \theta_{t+1}^{(i)}\right\|^2 + \sum_{j\in\mathcal{C}}\left\|\theta_t^{(j)} - \theta_t^{(i)}\right\|^2\right) + 9\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2 + \frac{3}{2}\gamma^2\left\|\overline{m}_t\right\|^2.$$

Substituting from above in (56) we obtain that

$$\begin{aligned}\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2 \leq{}& \frac{9L^2}{(n-f)^2}\left(\sum_{i,\,j\in\mathcal{C}}\left\|\theta_{t+1}^{(j)} - \theta_{t+1}^{(i)}\right\|^2 + \sum_{i,\,j\in\mathcal{C}}\left\|\theta_t^{(j)} - \theta_t^{(i)}\right\|^2\right)\\ &+ \frac{9L^2}{n-f}\sum_{i\in\mathcal{C}}\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2 + \frac{3}{2}\frac{L^2\gamma^2}{n-f}\sum_{i\in\mathcal{C}}\left\|\overline{m}_t\right\|^2.\end{aligned}$$

Taking total expectation on both sides above, and using the notation $\Gamma\left(*_t\right)$, we obtain that

$$\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2\right] \leq 9L^2\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right]\right) + 9L^2\,\mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right] + \frac{3}{2}L^2\gamma^2\,\mathbb{E}\left[\left\|\overline{m}_t\right\|^2\right].$$

By definition of e_t we have $\left\|\overline{m}_t\right\| \leq \left\|\overline{m}_t - \overline{\nabla Q}_t\right\| + \left\|\overline{\nabla Q}_t\right\| = \left\|\delta_t\right\| + \left\|\overline{\nabla Q}_t\right\|$. Therefore, $\left\|\overline{m}_t\right\|^2 \leq 3\left\|\delta_t\right\|^2 + \frac{3}{2}\left\|\overline{\nabla Q}_t\right\|^2$. Using this above we obtain that

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \overline{\nabla Q}_{t+1}\right\|^2\right] \leq{}& 9L^2\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right]\right)\\ &+ \frac{3}{2}L^2\gamma^2\left(3\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{3}{2}\,\mathbb{E}\left[\left\|\overline{\nabla Q}_t\right\|^2\right]\right).\end{aligned} \tag{58}$$

Substituting from (58) above in (55) we obtain that

$$\begin{aligned}\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] &\leq \beta^2(1+4\gamma L)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \beta^2\left(1+\frac{1}{4\gamma L}\right)\frac{9}{4}L^2\gamma^2\left(2\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \mathbb{E}\left[\left\|\overline{\nabla Q}_t\right\|^2\right]\right)\\ &\quad+ \beta^2\left(1+\frac{1}{4\gamma L}\right)9L^2\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right]\right) + \frac{(1-\beta)^2\sigma^2}{n-f}\\ &= \beta^2(1+4L\gamma)(1+\frac{9}{8}L\gamma)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{9}{16}\beta^2L\gamma(1+4L\gamma)\,\mathbb{E}\left[\left\|\overline{\nabla Q}_t\right\|^2\right]\\ &\quad+ \beta^2\left(1+\frac{1}{4\gamma L}\right)9L^2\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right]\right) + \frac{(1-\beta)^2\sigma^2}{n-f}.\end{aligned}$$

Now note also that $\left\|\overline{\nabla Q}_t\right\|^2 \leq 4\left\|\overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \frac{4}{3}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2$; thus

$$\begin{aligned}\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] &\leq \beta^2(1+4L\gamma)(1+\frac{9}{8}L\gamma)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{3}{4}\beta^2L\gamma(1+4L\gamma)\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right]\\ &\quad+ \beta^2\left(1+\frac{1}{4\gamma L}\right)9L^2\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right]\right) + \frac{(1-\beta)^2\sigma^2}{n-f}\\ &\quad+ \frac{9}{4}\beta^2L\gamma\left(1+4L\gamma\right)\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right].\end{aligned} \tag{59}$$

But now note that

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}(\theta_t^{(i)}) - \frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}}\left(\nabla Q^{(i)}(\theta_t^{(i)}) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right)\right\|^2\right] \\
&\leq \frac{1}{n-f}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\nabla Q^{(i)}(\theta_t^{(i)}) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2\right] \\
&\leq \frac{L^2}{n-f}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] \leq L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right].
\end{aligned}$$

Combining this with (59), we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\delta_{t+1}\right\|^2\right] \leq{}& \beta^2(1+4L\gamma)(1+\frac{9}{8}L\gamma)\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + \frac{3}{4}\beta^2 L\gamma(1+4L\gamma)\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \\
&+ 9\beta^2L^2\left(1+\frac{1}{4\gamma L}\right)\left(\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] + \mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \mathbb{E}\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right]\right) \\
&+ \frac{9}{4}\beta^2 L\gamma\left(1+4L\gamma\right)L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{(1-\beta)^2\sigma^2}{n-f},
\end{aligned}$$

which is the lemma. □

A.8. Proof of Lemma 5

We recall the lemma below. Also, recall that $\overline{\theta}_t \coloneqq {}^1\!/_{n-f}\sum_{i\in\mathcal{C}}\theta_t^{(i)}$ and $Q^{(\mathcal{C})}(\theta) = {}^1\!/_{n-f}\sum_{i\in\mathcal{C}}Q^{(i)}(\theta)$.

Lemma 5. *Suppose that assumptions 1 and 2 hold true. Consider Algorithm 1 with $\gamma \leq 1/L$. For each $t \in [T]$, we obtain that*

$$\begin{aligned}
\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] \leq{}& -\frac{\gamma}{2}\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + \frac{3\gamma}{2}\,\mathbb{E}\left[\left\|\delta_t\right\|^2\right] \\
&+ \frac{3}{2\gamma}\,\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}\right\|^2\right] + \frac{3\gamma}{2}L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right].
\end{aligned}$$

Proof. Consider an arbitrary $t \in [T]$. We define $G_t \coloneqq \frac{\overline{\theta}_t - \overline{\theta}_{t+1}}{\gamma}$, the step taken by the average of local models at iteration t. By the smoothness of the loss function (Assumption 1), we have

$$\begin{aligned}
Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t) &\leq \left\langle \overline{\theta}_{t+1} - \overline{\theta}_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \frac{L}{2}\left\|\overline{\theta}_{t+1} - \overline{\theta}_t\right\|^2 \\
&= -\gamma\left\langle G_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \frac{L\gamma^2}{2}\left\|G_t\right\|^2.
\end{aligned}$$

Using the fact that $\gamma \leq 1/L$, we obtain that

$$Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t) \leq -\gamma\left\langle G_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \frac{\gamma}{2}\left\|G_t\right\|^2.$$

Now note that $-\langle x,\, y\rangle + \frac{\|x\|^2}{2} = -\frac{\|y\|^2}{2} + \frac{\|x-y\|^2}{2}$; thus

$$\begin{aligned}
Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t) &\leq -\frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \frac{\gamma}{2}\left\|G_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 \\
&\leq -\frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \frac{\gamma}{2}\left\|\frac{\overline{\theta}_t - \overline{\theta}_{t+1}}{\gamma} - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 \\
&= -\frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \frac{\gamma}{2}\left\|\frac{\overline{\theta}_t - \overline{\theta}_{t+1/2} + \overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}}{\gamma} - \overline{\nabla Q}_t + \overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 \\
&= -\frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \frac{\gamma}{2}\left\|\overline{m}_t + \frac{\overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}}{\gamma} - \overline{\nabla Q}_t + \overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 \\
&\leq -\frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \frac{3\gamma}{2}\left\|\overline{m}_t - \overline{\nabla Q}_t\right\|^2 + \frac{3}{2\gamma}\left\|\overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}\right\|^2 + \frac{3\gamma}{2}\left\|\overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2,
\end{aligned}$$

where $\overline{\nabla Q}_t = \frac{1}{n-f}\sum_{i\in\mathcal{C}} \nabla Q^{(i)}\left(\theta_t^{(i)}\right)$. Now recall from Lemma 4 that we define $\delta_t = \overline{m}_t - \overline{\nabla Q}_t$. Thus, taking the expectation from both sides of the above, we obtain that

$$\begin{aligned}
\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] \leq -\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + \frac{3\gamma}{2}\mathbb{E}\left[\|\delta_t\|^2\right] + \frac{3}{2\gamma}\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}\right\|^2\right] \\
+ \frac{3\gamma}{2}\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right]. \qquad (60)
\end{aligned}$$

Now note that

$$\begin{aligned}
\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}(\theta_t^{(i)}) - \frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}}\left(\nabla Q^{(i)}(\theta_t^{(i)}) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right)\right\|^2\right] \\
&\leq \frac{1}{n-f}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\nabla Q^{(i)}(\theta_t^{(i)}) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2\right] \\
&\leq \frac{L^2}{n-f}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] \leq L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right].
\end{aligned}$$

Combining this with (60), we obtain that

$$\begin{aligned}
\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] \leq -\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + \frac{3\gamma}{2}\mathbb{E}\left[\|\delta_t\|^2\right] \\
+ \frac{3}{2\gamma}\mathbb{E}\left[\left\|\overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}\right\|^2\right] + \frac{3\gamma}{2}L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right].
\end{aligned}$$

which is the lemma. □

A.9. Proof of Lemma 6

In this section, we prove that if $n > 5f$, then $K \in \mathcal{O}(\log(n))$ coordination rounds is enough to guarantee (α, λ)-reduction.

Lemma 10. *Consider the coordination phase of Algorithm 1. Suppose that there exists $\delta > 0$ such that $n \geq (5+\delta)f$. For any $k \geq 1$ we have*

$$\max_{i,j\in\mathcal{C}}\left\|x_k^{(i)} - x_k^{(j)}\right\| \leq \left(\frac{3}{n-2f}\right)^k \max_{i,j\in\mathcal{C}}\left\|x_0^{(i)} - x_0^{(j)}\right\|.$$

Proof. Consider an arbitrary round $k \in [K]$. Consider two correct nodes $p, q \in \mathcal{C}$. We then have

$$\left\|x_k^{(p)} - x_k^{(q)}\right\| = \left\|\frac{1}{n-2f}\sum_{j\in\mathcal{S}_k^{(p)}} x_{k-1}^{(j)} - \frac{1}{n-2f}\sum_{j\in\mathcal{S}_k^{(q)}} x_{k-1}^{(j)}\right\|$$
$$= \frac{1}{n-2f}\left\|\sum_{j\in\mathcal{S}_k^{(p)}\setminus\mathcal{S}_k^{(q)}} x_{k-1}^{(j)} - \sum_{j\in\mathcal{S}_k^{(q)}\setminus\mathcal{S}_k^{(p)}} x_{k-1}^{(j)}\right\|. \tag{61}$$

Now similar to Lemma 9, we define:

$$F_p := \left\{i : i \notin \mathcal{C}, i \in \mathcal{S}_k^{(p)}, i \notin \mathcal{S}_k^{(q)}\right\}.$$

$$F_q := \left\{i : i \notin \mathcal{C}, i \in \mathcal{S}_k^{(q)}, i \notin \mathcal{S}_k^{(p)}\right\}.$$

$$\mathcal{C}_p := \left\{i : i \in \mathcal{C}, i \in \mathcal{S}_k^{(p)}, i \notin \mathcal{S}_k^{(q)}\right\}.$$

$$\mathcal{C}_q := \left\{i : i \in \mathcal{C}, i \in \mathcal{S}_k^{(q)}, i \notin \mathcal{S}_k^{(p)}\right\}.$$

We also define $f_p := |F_p|$ and $f_q := |F_q|$. We also order these four sets such that, e.g,. $F_p[i]$ refers to a unique element in set F_p. Without loss of generality, we assume $|\mathcal{C}_p| \geq f_q$ and $|\mathcal{C}_q| \geq f_p$ [5]. Now from (61), we obtain that

$$(n-2f)\left\|x_k^{(p)} - x_k^{(q)}\right\| = \left\|\sum_{j\in F_p} x_{k-1}^{(j)} + \sum_{j\in\mathcal{C}_p} x_{k-1}^{(j)} - \sum_{j\in F_q} x_{k-1}^{(j)} - \sum_{j\in\mathcal{C}_q} x_{k-1}^{(j)}\right\|$$
$$= \left\|\sum_{j\in[f_p]}\left(x_{k-1}^{(F_p[j])} - x_{k-1}^{(\mathcal{C}_q[j])}\right) - \sum_{j\in[f_q]}\left(x_{k-1}^{(F_q[j])} - x_{k-1}^{(\mathcal{C}_p[j])}\right) + \sum_{j\in[|\mathcal{C}_p|-f_q]}\left(x_{k-1}^{(\mathcal{C}_p[f_q+j])} - x_{k-1}^{(\mathcal{C}_q[f_p+j])}\right)\right\|.$$

By triangle inequality, we then have

$$(n-2f)\left\|x_k^{(p)} - x_k^{(q)}\right\| \leq \sum_{j\in[f_p]}\left\|x_{k-1}^{(F_p[j])} - x_{k-1}^{(\mathcal{C}_q[j])}\right\| + \sum_{j\in[f_q]}\left\|x_{k-1}^{(F_q[j])} - x_{k-1}^{(\mathcal{C}_p[j])}\right\|$$
$$+ \sum_{j\in[|\mathcal{C}_p|-f_q]}\left\|x_{k-1}^{(\mathcal{C}_p[f_q+j])} - x_{k-1}^{(\mathcal{C}_q[f_p+j])}\right\|$$
$$\leq \sum_{j\in[f_p]}\left\|x_{k-1}^{(F_p[j])} - x_{k-1}^{(p)}\right\| + \left\|x_{k-1}^{(p)} - x_{k-1}^{(\mathcal{C}_q[j])}\right\|$$
$$+ \sum_{j\in[f_q]}\left\|x_{k-1}^{(F_q[j])} - x_{k-1}^{(q)}\right\| + \left\|x_{k-1}^{(q)} - x_{k-1}^{(\mathcal{C}_p[j])}\right\|$$
$$+ \sum_{j\in[|\mathcal{C}_p|-f_q]}\left\|x_{k-1}^{(\mathcal{C}_p[f_q+j])} - x_{k-1}^{(\mathcal{C}_q[f_p+j])}\right\|.$$

Now note that for each faulty node $j^* \in \mathcal{S}_k^{(p)}$, there is at least one correct vector received by node p and filtered out by the

[5] Otherwise we add sufficiently many correct vectors from $\mathcal{S}_k^{(p)} \cap \mathcal{S}_k^{(q)}$ to both $\mathcal{C}_p$ and $\mathcal{C}_q$ such that $|\mathcal{C}_p| \geq f_q$ and $|\mathcal{C}_q| \geq f_p$.

NNA function. Therefore, we must have $\left\|x_{k-1}^{(j^*)} - x_{k-1}^{(p)}\right\| \leq \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\|$. Thus,

$$\begin{aligned}
(n-2f)\left\|x_k^{(p)} - x_k^{(q)}\right\| &\leq \sum_{j\in[f_p]} \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| + \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| \\
&\quad + \sum_{j\in[f_q]} \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| + \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| \\
&\quad + \sum_{j\in[|\mathcal{C}_p|-f_q]} \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| \\
&\leq (2(f_p+f_q) + (|\mathcal{C}_p| - f_q)) \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| \\
&= \left(2(f_p+f_q) + (\left|\mathcal{S}_k^{(p)} \setminus \mathcal{S}_k^{(q)}\right| - f_p - f_q)\right) \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\|.
\end{aligned}$$

Now recall from Lemma 8 that $\left|\mathcal{S}_k^{(q)} \setminus \mathcal{S}_k^{(p)}\right| \leq 2f$. Therefore,

$$\begin{aligned}
(n-2f)\left\|x_k^{(p)} - x_k^{(q)}\right\| &\leq (f_p + f_q + 2f) \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| \\
&\leq 3f \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\|.
\end{aligned}$$

Equivalently,

$$\begin{aligned}
\left\|x_k^{(p)} - x_k^{(q)}\right\| &\leq (f_p + f_q + 2f) \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\| \\
&\leq \frac{3f}{n-2f} \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\|.
\end{aligned}$$

As the above inequality holds for any choice of p and q, we obtain that

$$\max_{i,j\in\mathcal{C}} \left\|x_k^{(i)} - x_k^{(j)}\right\| \leq \frac{3f}{n-2f} \max_{i,j\in\mathcal{C}} \left\|x_{k-1}^{(i)} - x_{k-1}^{(j)}\right\|.$$

As the above holds for any $k \geq 1$, we obtain that

$$\max_{i,j\in\mathcal{C}} \left\|x_k^{(i)} - x_k^{(j)}\right\| \leq \left(\frac{3f}{n-2f}\right)^k \max_{i,j\in\mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|.$$

This is the lemma. □

Lemma 11. *Consider the coordination phase of Algorithm 1. Assume there exists $\delta > 0$ such that $n \geq (5+\delta)f$. For $K = \frac{\log(8(n-f))}{2\log(\frac{3+v}{v})} \in \mathcal{O}(\log(n))$, we have*

$$\Gamma\left(x_K\right) \leq \frac{2f}{n-f}\Gamma\left(x_0\right).$$

Proof. For $k = K$ in Lemma 10, we have

$$\max_{i,j\in\mathcal{C}} \left\|x_K^{(i)} - x_K^{(j)}\right\| \leq \left(\frac{3f}{n-2f}\right)^K \max_{i,j\in\mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|.$$

Now squaring both sides, we obtain that

$$\max_{i,j\in\mathcal{C}} \left\|x_K^{(i)} - x_K^{(j)}\right\|^2 \leq \left(\frac{3f}{n-2f}\right)^{2K} \max_{i,j\in\mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|^2.$$

Now as we assume $n \geq (5+\delta)f$, we have

$$\frac{3f}{n-2f} \leq \frac{3f}{(3+\delta)f} = \frac{3}{3+\delta} < 1.$$

We then have

$$\max_{i,j\in\mathcal{C}} \left\|x_K^{(i)} - x_K^{(j)}\right\|^2 \leq \left(\frac{3}{3+\delta}\right)^{2K-1} \left(\frac{3f}{n-2f}\right) \max_{i,j\in\mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|^2.$$

Now note that as $n > 5f$, we have $4(n-2f) \geq 3(n-f)$, thus,

$$\max_{i,j\in\mathcal{C}} \left\|x_K^{(i)} - x_K^{(j)}\right\|^2 \leq \left(\frac{3}{3+\delta}\right)^{2K-1} \left(\frac{4f}{n-f}\right) \max_{i,j\in\mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|^2. \tag{62}$$

Now note that

$$\begin{aligned}\Gamma\left(x_K\right) &= \frac{1}{(n-f)^2} \sum_{i,\, j\in\mathcal{C}} \left\|x_K^{(i)} - x_K^{(j)}\right\|^2 \\ &\leq \frac{1}{(n-f)^2} \sum_{i,\, j\in\mathcal{C}} \max_{i,j\in\mathcal{C}} \left\|x_K^{(i)} - x_K^{(j)}\right\|^2 = \max_{i,j\in\mathcal{C}} \left\|x_K^{(i)} - x_K^{(j)}\right\|^2. \end{aligned} \tag{63}$$

Note also that

$$\begin{aligned}\max_{i,j\in\mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|^2 &\leq 4\max_{i\in\mathcal{C}} \left\|x_0^{(i)} - \bar{x}_0\right\|^2 \\ &\leq 4\sum_{i\in\mathcal{C}} \left\|x_0^{(i)} - \bar{x}_0\right\|^2 \\ &\leq 4(n-f)\frac{1}{n-f}\sum_{i\in\mathcal{C}} \left\|x_0^{(i)} - \bar{x}_0\right\|^2 = 4(n-f)\Gamma\left(x_0\right). \end{aligned} \tag{64}$$

Combining (62), (63), and (64), we then obtain that

$$\Gamma\left(x_K\right) \leq \left(\frac{3}{3+\delta}\right)^{2K-1} \left(\frac{2f}{n-f}\right) 8(n-f)\Gamma\left(x_0\right).$$

Setting $K = \left\lceil \frac{\log(8(n-f))}{2\log(\frac{3+\delta}{3})} \right\rceil + 1$, we then obtain that

$$\Gamma\left(x_K\right) \leq \frac{2f}{n-f}\Gamma\left(x_0\right).$$

This is what we wanted. □

We recall Lemma 6 below.

Lemma 6. *Suppose that there exists $\delta > 0$ such that $n \geq (5+\delta)f$. For $K = \frac{\log(8(n-f))}{2\log(\frac{3+\delta}{\delta})} \in \mathcal{O}(\log(n))$, the coordination phase of Algorithm 1 guarantees (α, λ)-reduction for*

$$\alpha = \frac{2f}{n-f} \leq \frac{1}{2} \quad \textit{and} \quad \lambda = \left(\frac{3+\delta}{\delta}\right)^2 \frac{(8f)^2}{n-f}.$$

Proof. The first inequality is already proved by Lemma 11. Here we prove the second inequality. Consider an arbitrary round $k \in [K]$. For a correct node $i \in \mathcal{C}$ we have

$$\left\|x_k^{(i)} - \bar{x}_{k-1}\right\| = \left\|\frac{1}{n-2f}\sum_{j\in\mathcal{S}_k^{(i)}} x_{k-1}^{(j)} - \frac{1}{n-f}\sum_{j\in\mathcal{C}} x_{k-1}^{(j)}\right\|$$

$$= \left\|\left(\frac{1}{n-2f} - \frac{1}{n-f}\right)\sum_{j\in\mathcal{S}_k^{(i)}\cap\mathcal{C}} x_{k-1}^{(j)} + \frac{1}{n-2f}\sum_{j\in\mathcal{S}_k^{(i)}\setminus\mathcal{C}} x_{k-1}^{(j)} - \frac{1}{n-f}\sum_{j\in\mathcal{C}\setminus\mathcal{S}_k^{(i)}} x_{k-1}^{(j)}\right\|$$

$$= \frac{1}{(n-f)(n-2f)}\left\|f\sum_{j\in\mathcal{S}_k^{(i)}\cap\mathcal{C}}(x_{k-1}^{(j)} - x_{k-1}^{(i)}) + (n-f)\sum_{j\in\mathcal{S}_k^{(i)}\setminus\mathcal{C}}(x_{k-1}^{(j)} - x_{k-1}^{(i)}) - (n-2f)\sum_{j\in\mathcal{C}\setminus\mathcal{S}_k^{(i)}}(x_{k-1}^{(j)} - x_{k-1}^{(i)})\right\|.$$

By the triangle inequality, we then have

$$(n-f)(n-2f)\left\|x_k^{(i)} - \bar{x}_{k-1}\right\| \le f\sum_{j\in\mathcal{S}_k^{(i)}\cap\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\| + (n-f)\sum_{j\in\mathcal{S}_k^{(i)}\setminus\mathcal{C}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\| + (n-2f)\sum_{j\in\mathcal{C}\setminus\mathcal{S}_k^{(i)}}\left\|x_{k-1}^{(j)} - x_{k-1}^{(i)}\right\|.$$

Now note that for any faulty node $j^* \in \mathcal{S}_k^{(i)}$ there is at least one correct i^* such that $\left\|x_{k-1}^{(j^*)} - x_{k-1}^{(i)}\right\| \le \left\|x_{k-1}^{(i^*)} - x_{k-1}^{(i)}\right\|$ (as otherwise i^* would have been selected instead of j^*). Therefore, we must have $\left\|x_{k-1}^{(j^*)} - x_{k-1}^{(i)}\right\| \le \max_{p,q\in\mathcal{C}}\left\|x_{k-1}^{(p)} - x_{k-1}^{(q)}\right\|$. And clearly for any correct node i^* we have $\left\|x_{k-1}^{(i^*)} - x_{k-1}^{(i)}\right\| \le \max_{p,q\in\mathcal{C}}\left\|x_{k-1}^{(p)} - x_{k-1}^{(q)}\right\|$. Therefore,

$$\left\|x_k^{(i)} - \bar{x}_{k-1}\right\| \le \frac{f\left|\mathcal{S}_k^{(i)}\cap\mathcal{C}\right| + (n-f)\left|\mathcal{S}_k^{(i)}\setminus\mathcal{C}\right| + (n-2f)\left|\mathcal{C}\setminus\mathcal{S}_k^{(i)}\right|}{(n-f)(n-2f)}\max_{p,q\in\mathcal{C}}\left\|x_{k-1}^{(p)} - x_{k-1}^{(q)}\right\|.$$

Now let $v := \left|\mathcal{S}_k^{(i)}\cap\mathcal{C}\right|$. We then have $v = \left|\mathcal{S}_k^{(i)}\right| + |\mathcal{C}| - \left|\mathcal{S}_k^{(i)}\cup\mathcal{C}\right| \ge n - 2f + n - f - n = n - 3f$. Also, $\left|\mathcal{S}_k^{(i)}\setminus\mathcal{C}\right| = n - 2f - v$ and $\left|\mathcal{C}\setminus\mathcal{S}_k^{(i)}\right| = n - f - v$. Now we define

$$A(v) := fv + (n-2f-v)(n-f) + (n-2f)(n-f-v) = 2(n-2f)(n-f-v), \tag{65}$$

which is decreasing in v. Then the maximum of $A(v)$ is reached for $v = n - 3f$ and we have $A(v) \le 4f(n-2f)$. Therefore,

$$\left\|x_k^{(i)} - \bar{x}_{k-1}\right\| \le \frac{4f}{n-f}\max_{p,q\in\mathcal{C}}\left\|x_{k-1}^{(p)} - x_{k-1}^{(q)}\right\|.$$

Also, note that

$$\|\bar{x}_k - \bar{x}_{k-1}\| = \left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}} x_k^{(i)} - \bar{x}_{k-1}\right\| \le \frac{1}{n-f}\sum_{i\in\mathcal{C}}\left\|x_k^{(i)} - \bar{x}_{k-1}\right\|$$
$$\le \frac{1}{n-f}\sum_{i\in\mathcal{C}}\frac{4f}{n-f}\max_{p,q\in\mathcal{C}}\left\|x_{k-1}^{(p)} - x_{k-1}^{(q)}\right\| = \frac{4f}{n-f}\max_{p,q\in\mathcal{C}}\left\|x_{k-1}^{(p)} - x_{k-1}^{(q)}\right\|.$$

Now applying Lemma 10, we obtain that

$$\|\bar{x}_k - \bar{x}_{k-1}\| \le \left(\frac{3f}{n-2f}\right)^{k-1}\frac{4f}{n-f}\max_{i,j\in\mathcal{C}}\left\|x_0^{(i)} - x_0^{(j)}\right\|$$

Now as $n \geq (5+\delta)f$, we obtain that

$$\|\bar{x}_k - \bar{x}_{k-1}\| \leq \left(\frac{3}{3+\delta}\right)^{k-1} \frac{4f}{n-f} \max_{i,j \in \mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|$$

Now note that by triangle inequality we have

$$\begin{aligned}
\|\bar{x}_K - \bar{x}_0\| &\leq \sum_{k\in[K]} \|\bar{x}_k - \bar{x}_{k-1}\| \leq \sum_{k\in[K]} \left(\frac{3}{3+\delta}\right)^{k-1} \frac{4f}{n-f} \max_{i,j \in \mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\| \\
&\leq \sum_{k=1}^{\infty} \left(\frac{3}{3+\delta}\right)^{k-1} \frac{4f}{n-f} \max_{i,j \in \mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\| = \frac{1}{1-\frac{3}{3+\delta}} \frac{4f}{n-f} \max_{i,j \in \mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\| \\
&= \frac{3+\delta}{\delta} \frac{4f}{n-f} \max_{i,j \in \mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|.
\end{aligned}$$

Squaring both sides, we then have

$$\|\bar{x}_K - \bar{x}_0\|^2 \leq \left(\frac{3+\delta}{\delta}\right)^2 \left(\frac{4f}{n-f}\right)^2 \max_{i,j \in \mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|^2. \tag{66}$$

Now note that

$$\begin{aligned}
\max_{i,j \in \mathcal{C}} \left\|x_0^{(i)} - x_0^{(j)}\right\|^2 &\leq 4 \max_{i \in \mathcal{C}} \left\|x_0^{(i)} - \bar{x}_0\right\|^2 \leq 4 \sum_{i \in \mathcal{C}} \left\|x_0^{(i)} - \bar{x}_0\right\|^2 \\
&\leq 4(n-f) \frac{1}{n-f} \sum_{i \in \mathcal{C}} \left\|x_0^{(i)} - \bar{x}_0\right\|^2 = 4(n-f)\Gamma\left(x_0\right). \qquad (67)
\end{aligned}$$

Combining above with (66), we obtain that

$$\|\bar{x}_K - \bar{x}_0\|^2 \leq \left(\frac{3+\delta}{\delta}\right)^2 \frac{(8f)^2}{n-f} \Gamma\left(x_0\right).$$

□

B. D-SGD without Momentum under (α, λ)-reduction

In this section, we prove Proposition 1 which is convergence guarantee for D-SGD (i.e., setting $\beta = 0$ in Algorithm 1) under (α, λ)-reduction. First let us recall Proposition 1 (with more details).

Proposition 1. *Consider Algorithm 1 with $\beta = 0$. Suppose assumptions 1, 2 and 3 hold true and that the coordination phase satisfies (α, λ)-reduction for $\alpha < 1$. Define*

$$c_0 := 6\left(Q^{(\mathcal{C})}(\bar{\theta}_0) - Q^*\right) \text{ and } c_1 = \frac{6\sqrt{1+\alpha}}{1-\alpha}.$$

Set $\gamma = \min\{\frac{1}{2L}, \frac{1}{c_1 L}, \sqrt{\frac{nc_0}{9LT\sigma^2}}\}$. For any correct node $i \in \mathcal{C}$, Algorithm 1 then guarantees

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] &\leq 6\sqrt{\frac{c_0 L\sigma^2}{nT}} + \frac{4c_0 c_1^2 Ln\alpha}{9T\sigma^2}\left(3\sigma^2 + \zeta^2\right) + \frac{(2+c_1)Lc_0}{T} + 20c_1^2\lambda\left(3\sigma^2 + \zeta^2\right) \\
&\in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{nT}} + \frac{\lambda}{(1-\alpha)^2}\left(\sigma^2 + \zeta^2\right)\right)
\end{aligned}$$

Before proving the proposition, we first prove a few useful lemmas.

Lemma 12. *Suppose that assumptions 1 and 2 hold true. Consider Algorithm 1 with $\beta = 0$ and $\gamma \leq 1/2L$. For each $t \in [T]$, we obtain that*

$$\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] \leq -\frac{\gamma}{4}\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + L\gamma^2\frac{\sigma^2}{n-f} + \frac{\gamma L^2}{2}\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{5\lambda}{\gamma}\,\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right].$$

Proof. Consider an arbitrary $t \in [T]$. We define $G_t := \frac{\overline{\theta}_t - \overline{\theta}_{t+1}}{\gamma}$, the step taken by the average of local models at iteration t. Also, define $R_t := \overline{g}_t - G_t$. By the smoothness of the loss function (Assumption 1), we have

$$\begin{aligned}
Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t) &\leq \left\langle \overline{\theta}_{t+1} - \overline{\theta}_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \frac{L}{2}\left\|\overline{\theta}_{t+1} - \overline{\theta}_t\right\|^2 \\
&= -\gamma\left\langle G_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \frac{L\gamma^2}{2}\left\|G_t\right\|^2 \\
&= -\gamma\left\langle \overline{g}_t + R_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \frac{L\gamma^2}{2}\left\|\overline{g}_t + R_t\right\|^2.
\end{aligned}$$

Now denoting $\overline{\nabla Q}_t = \frac{1}{n-f}\sum_{i\in\mathcal{C}} \nabla Q^{(i)}\left(\theta_t^{(i)}\right)$, and taking the conditional expectation, we have

$$\begin{aligned}
\mathbb{E}_t\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1})\right] - Q^{(\mathcal{C})}(\overline{\theta}_t) &\leq -\gamma\left\langle \mathbb{E}_t\left[\overline{g}_t + R_t\right],\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \frac{L\gamma^2}{2}\mathbb{E}_t\left[\left\|\overline{g}_t + R_t\right\|^2\right] \\
&\overset{(a)}{\leq} -\gamma\left\langle \overline{\nabla Q}_t + \mathbb{E}_t\left[R_t\right],\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + L\gamma^2\mathbb{E}_t\left[\left\|\overline{g}_t\right\|^2\right] + L\gamma^2\mathbb{E}_t\left[\left\|R_t\right\|^2\right] \\
&\overset{(b)}{\leq} -\gamma\left\langle \overline{\nabla Q}_t + \mathbb{E}_t\left[R_t\right],\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + L\gamma^2\mathbb{E}_t\left[\left\|\overline{\nabla Q}_t\right\|^2\right] + L\gamma^2\frac{\sigma^2}{n-f} + L\gamma^2\mathbb{E}_t\left[\left\|R_t\right\|^2\right]
\end{aligned}$$

where (a) uses Young's inequality and (b) is based on the facts that by Assumption 2, we have $\mathbb{E}_t\left[\overline{g}_t\right] = \overline{\nabla Q}_t$, and $\mathbb{E}_t\left[\left\|\overline{g}_t - \overline{\nabla Q}_t\right\|^2\right] \leq \frac{\sigma^2}{n-f}$. We then obtain that

$$\begin{aligned}
\mathbb{E}_t\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1})\right] - Q^{(\mathcal{C})}(\overline{\theta}_t) &\leq -\gamma\left\langle \overline{\nabla Q}_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + L\gamma^2\mathbb{E}_t\left[\left\|\overline{\nabla Q}_t\right\|^2\right] \\
&\quad + L\gamma^2\frac{\sigma^2}{n-f} + L\gamma^2\mathbb{E}_t\left[\left\|R_t\right\|^2\right] - \gamma\left\langle \mathbb{E}_t\left[R_t\right],\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle. \qquad (68)
\end{aligned}$$

Now note that

$$-\left\langle \mathbb{E}_t\left[R_t\right],\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle \leq 4\left\|\mathbb{E}_t\left[R_t\right]\right\|^2 + \frac{1}{4}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 \leq 4\mathbb{E}_t\left[\left\|R_t\right\|^2\right] + \frac{1}{4}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2, \qquad (69)$$

where the second inequality uses Jensen's inequality. Also, using the fact that $\gamma \leq \frac{1}{2L}$, we have

$$\begin{aligned}
-\gamma\left\langle \overline{\nabla Q}_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + L\gamma^2\left\|\overline{\nabla Q}_t\right\|^2 &\leq \frac{\gamma}{2}\left(-2\left\langle \overline{\nabla Q}_t,\, \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\rangle + \left\|\overline{\nabla Q}_t\right\|^2\right) \\
&= \frac{\gamma}{2}\left(-\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \overline{\nabla Q}_t\right\|^2\right) \qquad (70)
\end{aligned}$$

Combining (68), (69), and (70), we obtain that

$$\mathbb{E}_t\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1})\right] - Q^{(\mathcal{C})}(\overline{\theta}_t) \leq -\frac{\gamma}{4}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2 + \frac{\gamma}{2}\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \overline{\nabla Q}_t\right\|^2 + L\gamma^2\frac{\sigma^2}{n-f} + (4\gamma + L\gamma^2)\mathbb{E}_t\left[\left\|R_t\right\|^2\right]. \qquad (71)$$

Now note that

$$R_t = \overline{g}_t - G_t = \overline{g}_t - \frac{\overline{\theta}_t - \overline{\theta}_{t+1/2} + \overline{\theta}_{t+1/2} - \overline{\theta}_{t+1}}{\gamma} = \frac{\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}}{\gamma}.$$

Thus, using the definition of (α, λ)-reduction, we obtain that

$$\mathbb{E}_t\left[\|R_t\|^2\right] = \frac{1}{\gamma^2}\mathbb{E}_t\left[\left\|\overline{\theta}_{t+1} - \overline{\theta}_{t+1/2}\right\|^2\right] \leq \frac{1}{\gamma^2}\lambda\mathbb{E}_t\left[\Gamma\left(\theta_{t+1/2}\right)\right]$$

Combining this with (71) taking total expectation, and using the fact that $\gamma \leq \frac{1}{L}$, we obtain that

$$\begin{aligned}\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] \leq -\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \overline{\nabla Q}_t\right\|^2\right] \\ + L\gamma^2\frac{\sigma^2}{n-f} + \frac{5\lambda}{\gamma}\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right].\end{aligned}$$

Now note that

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{\nabla Q}_t - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}(\theta_t^{(i)}) - \frac{1}{n-f}\sum_{i\in\mathcal{C}}\nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{n-f}\sum_{i\in\mathcal{C}}\left(\nabla Q^{(i)}(\theta_t^{(i)}) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right)\right\|^2\right] \\ &\leq \frac{1}{n-f}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\nabla Q^{(i)}(\theta_t^{(i)}) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2\right] \\ &\leq \frac{L^2}{n-f}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] \leq L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right].\end{aligned}$$

Therefore,

$$\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] \leq -\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + L\gamma^2\frac{\sigma^2}{n-f} + \frac{\gamma L^2}{2}\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{5\lambda}{\gamma}\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right].$$

This is the lemma. □

Lemma 13. *Suppose that assumptions 1, 2, and 3 hold true. Consider Algorithm 1 with $\gamma \leq \frac{1}{6L}\frac{1-\alpha}{\sqrt{\alpha(1+\alpha)}}$, and $\beta = 0$. Suppose that the coordination phase satisfies (α, λ)-reduction for $\alpha < 1$. For each $t \in [T]$, we obtain that*

$$\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] \leq \frac{6\alpha(1+\alpha)}{(1-\alpha)^2}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right),$$

and

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right] \leq \frac{6(1+\alpha)}{(1-\alpha)^2}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right).$$

Proof. First, we analyze the growth of $\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]$. From Algorithm 1 (for $\beta = 0$) recall that for all $i \in \mathcal{C}$, we have $\theta_{t+1/2}^{(i)} = \theta_t^{(i)} - \gamma g_t^{(i)}$. As $(x+y)^2 \leq (1+c)x^2 + (1+1/c)y^2$ for any $c > 0$, we obtain for all $i,\, j \in \mathcal{C}$ that

$$\begin{aligned}\mathbb{E}\left[\left\|\theta_{t+1/2}^{(i)} - \theta_{t+1/2}^{(j)}\right\|^2\right] &\leq \mathbb{E}\left[\left\|\theta_t^{(i)} - \theta_t^{(j)} - \gamma\left(g_t^{(i)} - g_t^{(j)}\right)\right\|^2\right] \\ &\leq (1+c)\,\mathbb{E}\left[\left\|\theta_t^{(i)} - \theta_t^{(j)}\right\|^2\right] + \left(1 + \frac{1}{c}\right)\gamma^2\,\mathbb{E}\left[\left\|g_t^{(i)} - g_t^{(j)}\right\|^2\right].\end{aligned}$$

Thus, by definition of notation $\Gamma\left(*_t\right)$ and by Lemma 7, we have

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right] \leq (1+c)\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \left(1 + \frac{1}{c}\right)\gamma^2\,\mathbb{E}\left[\Gamma\left(g_t\right)\right]. \tag{72}$$

Recall that, the coordination phase of Algorithm 1 satisfies (α, λ)-reduction. Thus, for all t, we have $\Gamma\left(\theta_{t+1}\right) \leq \alpha\Gamma\left(\theta_{t+1/2}\right)$. Substituting from above we obtain that

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq (1+c)\alpha\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \left(1+\frac{1}{c}\right)\alpha\gamma^2\,\mathbb{E}\left[\Gamma\left(g_t\right)\right].$$

For $c = \frac{1-\alpha}{2\alpha}$, we obtain that

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq \frac{1+\alpha}{2}\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{\alpha(1+\alpha)}{1-\alpha}\gamma^2\,\mathbb{E}\left[\Gamma\left(g_t\right)\right]. \tag{73}$$

Now note that for any $i \in \mathcal{C}$ we have

$$g_t^{(i)} - \overline{g}_t = g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right) + \nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(i)}\left(\overline{\theta}_t\right) + \nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) + \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \overline{g}_t.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\left\|g_t^{(i)} - \overline{g}_t\right\|^2\right] \leq\; & 4\,\mathbb{E}\left[\left\|g_t^{(i)} - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right\|^2\right] + 4\,\mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - \nabla Q^{(i)}\left(\overline{\theta}_t\right)\right\|^2\right] \\ & + 4\,\mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 4\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \overline{g}_t\right\|^2\right].\end{aligned}$$

Using assumptions 1, and 2, we obtain that

$$\begin{aligned}\mathbb{E}\left[\left\|g_t^{(i)} - \overline{g}_t\right\|^2\right] \leq\; & 4\sigma^2 + 4L^2\,\mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] \\ & + 4\,\mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 4\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \overline{g}_t\right\|^2\right].\end{aligned} \tag{74}$$

Now note that

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \overline{g}_t\right\|^2\right] &= \frac{1}{(n-f)^2}\,\mathbb{E}\left[\left\|\sum_{i\in\mathcal{C}}\left(\nabla Q^{(i)}\left(\overline{\theta}_t\right) - g_t^{(i)}\right)\right\|^2\right] \\ &\leq \frac{2}{(n-f)^2}\,\mathbb{E}\left[\left\|\sum_{i\in\mathcal{C}}\left(\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(i)}\left(\theta_t^{(i)}\right)\right)\right\|^2\right] \\ &\quad + \frac{2}{(n-f)^2}\,\mathbb{E}\left[\left\|\sum_{i\in\mathcal{C}}\left(\nabla Q^{(i)}\left(\theta_t^{(i)}\right) - g_t^{(i)}\right)\right\|^2\right] \\ &\leq \frac{2L^2}{n-f}\,\mathbb{E}\left[\sum_{i\in\mathcal{C}}\left\|\overline{\theta}_t - \theta_t^{(i)}\right\|^2\right] + \frac{2}{n-f}\sigma^2 \\ &= 2L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{2}{n-f}\sigma^2.\end{aligned}$$

Combining this with (74), we obtain that

$$\mathbb{E}\left[\left\|g_t^{(i)} - \overline{g}_t\right\|^2\right] \leq 4\sigma^2 + 4L^2\,\mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] + 4\,\mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 8L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{8}{n-f}\sigma^2$$

As i above is an arbitrary node in $\mathcal{C}$, the above holds true for all $i \in \mathcal{C}$. Averaging over all $i \in \mathcal{C}$ on both sides yields

$$\begin{aligned}\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|g_t^{(i)} - \overline{g}_t\right\|^2\right] \leq\; & 4\sigma^2 + 4L^2\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\theta_t^{(i)} - \overline{\theta}_t\right\|^2\right] + 4\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \\ & + 8L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{8}{n-f}\sigma^2.\end{aligned}$$

Recall, from Section A.1, the notation $\Gamma\left(*_t\right)$, i.e., $\Gamma\left(*_t\right) = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left\|*_t^{(i)} - \bar{*}_t\right\|^2$. Using this above, we get

$$\mathbb{E}\left[\Gamma\left(g_t\right)\right] \leq 4\sigma^2 + 4L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + 4\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 8L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{8}{n-f}\sigma^2.$$

From Assumption 3, we have $\frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\mathbb{E}\left[\left\|\nabla Q^{(i)}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \leq \zeta^2$. Using this above, we obtain that

$$\mathbb{E}\left[\Gamma\left(g_t\right)\right] \leq \left(4 + \frac{8}{n-f}\right)\sigma^2 + 12L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + 4\zeta^2. \tag{75}$$

Combining this with (73), we obtain that

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq \left(\frac{1+\alpha}{2} + 12\frac{\alpha(1+\alpha)}{1-\alpha}\gamma^2L^2\right)\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{\alpha(1+\alpha)}{1-\alpha}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right).$$

For $\gamma \leq \frac{1}{6L}\frac{1-\alpha}{\sqrt{\alpha(1+\alpha)}}$, we have

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1}\right)\right] \leq \frac{5+\alpha}{6}\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{\alpha(1+\alpha)}{1-\alpha}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right).$$

Unrolling the recursion, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] &\leq \frac{\alpha(1+\alpha)}{1-\alpha}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right)\sum_{s=0}^{t}\left(\frac{5+\alpha}{6}\right)^s \\
&\leq \frac{\alpha(1+\alpha)}{1-\alpha}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right)\sum_{s=0}^{\infty}\left(\frac{5+\alpha}{6}\right)^s \\
&= \frac{6\alpha(1+\alpha)}{(1-\alpha)^2}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right).
\end{aligned} \tag{76}$$

Combining (72), (75), and (76), we also obtain that

$$\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right] \leq \frac{6(1+\alpha)}{(1-\alpha)^2}\gamma^2\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right).$$

This is the desired result. □

Lemma 14. *Consider Algorithm 1 with $\beta = 0$. Define*

$$c_0 := 6\left(Q^{(\mathcal{C})}(\overline{\theta}_0) - Q^*\right) \text{ and } c_1 = \frac{6\sqrt{1+\alpha}}{1-\alpha}.$$

Suppose that assumptions 1, 2 and 3 hold true, and that the coordination phase satisfies (α, λ)-reduction for $\alpha < 1$. Suppose also that $\gamma \leq \min\{\frac{1}{2L}, \frac{1}{c_1 L}\}$. Then, for any correct node i, and any $T \geq 1$, we have

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq \frac{c_0}{\gamma T} + 9L\gamma\frac{\sigma^2}{n} + 4c_1^2 n\alpha L^2\gamma^2\left(3\sigma^2 + \zeta^2\right) + 20c_1^2\lambda\left(3\sigma^2 + \zeta^2\right).$$

Proof. Combining Lemma 12 and Lemma 13, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] &\leq -\frac{4}{\gamma}\,\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] + 4L\gamma\frac{\sigma^2}{n-f} + 2L^2\,\mathbb{E}\left[\Gamma\left(\theta_t\right)\right] + \frac{20\lambda}{\gamma^2}\,\mathbb{E}\left[\Gamma\left(\theta_{t+1/2}\right)\right] \\
&\leq -\frac{4}{\gamma}\,\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] + 4L\gamma\frac{\sigma^2}{n-f} \\
&\quad + (2L^2\alpha\gamma^2 + 20\lambda)\frac{6(1+\alpha)}{(1-\alpha)^2}\left(\left(4 + \frac{8}{n-f}\right)\sigma^2 + 4\zeta^2\right) \\
&\leq -\frac{4}{\gamma}\,\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_{t+1}) - Q^{(\mathcal{C})}(\overline{\theta}_t)\right] + 4L\gamma\frac{\sigma^2}{n-f} + (2L^2\gamma^2\alpha + 20\lambda)\frac{6(1+\alpha)}{(1-\alpha)^2}\left(12\sigma^2 + 4\zeta^2\right),
\end{aligned}$$

where in the last inequality, we used the fact that $n - f \geq 1$. Averaging over $t = 0, \ldots, T-1$, we obtain that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] \leq \frac{4}{\gamma T}\mathbb{E}\left[Q^{(\mathcal{C})}(\overline{\theta}_0) - Q^{(\mathcal{C})}(\overline{\theta}_T)\right] + 4L\gamma\frac{\sigma^2}{n-f} + (2L^2\gamma^2\alpha + 20\lambda)\frac{6(1+\alpha)}{(1-\alpha)^2}\left(12\sigma^2 + 4\zeta^2\right)$$
$$\leq \frac{4}{\gamma T}\left(Q^{(\mathcal{C})}(\overline{\theta}_0) - Q^*\right) + 6L\gamma\frac{\sigma^2}{n} + (2L^2\gamma^2\alpha + 20\lambda)\frac{6(1+\alpha)}{(1-\alpha)^2}\left(12\sigma^2 + 4\zeta^2\right), \tag{77}$$

where we used the fact that $Q^{(\mathcal{C})}(\overline{\theta}_T) \geq Q^*$, and $n \geq 5f$. Now note that for any correct node $i \in \mathcal{C}$, we have

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \leq \frac{3}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 3\,\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right) - \nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right]$$
$$\leq \frac{3}{2}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\overline{\theta}_t\right)\right\|^2\right] + 3L^2\,\mathbb{E}\left[\left\|\overline{\theta}_t - \theta_t^{(i)}\right\|^2\right],$$

where the second inequality follows from Assumption 1. Combining this with (77) and using the bound on $\mathbb{E}\left[\Gamma\left(\theta_t\right)\right]$ from Lemma 13, we obtain that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \leq \frac{6}{\gamma T}\left(Q^{(\mathcal{C})}(\overline{\theta}_0) - Q^*\right) + 9L\gamma\frac{\sigma^2}{n} + (4L^2\gamma^2 n\alpha + 20\lambda)\frac{9(1+\alpha)}{(1-\alpha)^2}\left(12\sigma^2 + 4\zeta^2\right),$$

where we used the fact that $n \geq 1$. Defining $c_0 := 6\left(Q^{(\mathcal{C})}(\overline{\theta}_0) - Q^*\right)$, $c_1^2 = \frac{36(1+\alpha)}{(1-\alpha)^2}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right] \leq \frac{c_0}{\gamma T} + 9L\gamma\frac{\sigma^2}{n} + 4c_1^2 n\alpha L^2\gamma^2\left(3\sigma^2 + \zeta^2\right) + 20c_1^2\lambda\left(3\sigma^2 + \zeta^2\right).$$

As $\hat{\theta}^{(i)} \sim \mathcal{U}\{\theta_0^{(i)}, \ldots, \theta_{T-1}^{(i)}\}$, we have

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\theta_t^{(i)}\right)\right\|^2\right].$$

This proves the desired result. □

Proof of Proposition 1 then straightforwardly follows.

Proof of Proposition 1. Recall from Lemma 14 that

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq \frac{c_0}{\gamma T} + 9L\gamma\frac{\sigma^2}{n} + 4c_1^2 n\alpha L^2\gamma^2\left(3\sigma^2 + \zeta^2\right) + 20c_1^2\lambda\left(3\sigma^2 + \zeta^2\right).$$

As $\gamma = \min\{\frac{1}{2L}, \frac{1}{c_1 L}, \sqrt{\frac{nc_0}{9LT\sigma^2}}\}$, we have

$$\frac{1}{\gamma} \leq \max\{2L, c_1 L, \sqrt{\frac{9LT\sigma^2}{nc_0}}\} \leq 2L + c_1 L + \sqrt{\frac{9LT\sigma^2}{nc_0}}.$$

Therefore,

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \leq 6\sqrt{\frac{c_0 L\sigma^2}{nT}} + \frac{4c_0 c_1^2 Ln\alpha}{9T\sigma^2}\left(3\sigma^2 + \zeta^2\right) + \frac{(2+c_1)Lc_0}{T} + 20c_1^2\lambda\left(3\sigma^2 + \zeta^2\right).$$

Now ignoring the non-dominant $\frac{1}{T}$ terms and noting $c_0 \in \mathcal{O}(1)$ and $c_1 \in \mathcal{O}\left(\frac{1}{1-\alpha}\right)$, we obtain that

$$\mathbb{E}\left[\left\|\nabla Q^{(\mathcal{C})}\left(\hat{\theta}^{(i)}\right)\right\|^2\right] \in \mathcal{O}\left(\sqrt{\frac{\sigma^2}{nT}} + \frac{\lambda}{(1-\alpha)^2}\left(\sigma^2 + \zeta^2\right)\right)$$

which is the desired result. □

C. Signed Echo Broadcast (SEB)

SEB is composed of four rounds of communication. First, in round SEND, each sender node i sends its message m_i to all the other nodes j, which contains an identifier and a signature by node i. In the context of MoNNA, the message is a vector $x_{k-1}^{(i)}$, with an identifier of the form (i, t, k), where t is the SGD iteration and k is the coordination round. Second, in round ECHO, upon receiving m_i, each recipient node j verifies that m_i is the first message with a valid signature from node i and with identifier (i, t, k). If so, node j signs m_i with its private key pk_j, thereby obtaining a *signature* s_{ij} of message m_i which node j sends to node i. Otherwise, m_i is ignored. Third, in round FINAL, upon receiving at least $\frac{n+f}{2} - 1$ valid signatures s_{ij} , the sender node i sends the set S_i of received signatures s_{ij} to all other nodes. Fourth and finally, in round ACCEPT, upon receiving the set S_i, each recipient node j verifies the signatures of the set, and if they are valid, node j accepts m_i and terminates the protocol. For $n > 3f$, SEB guarantees *validity*, even under asynchrony, i.e., any correct node's message is eventually delivered to any other correct nodes. It also guarantees *consistency*, i.e., two different correct nodes cannot deliver different messages from a faulty node (Cachin et al., 2011, Section 3.10.4). The message complexity of this protocol is linear in the total number of nodes, i.e., $\mathcal{O}(n)$. Therefore, it does not affect the communication complexity of MoNNA.

D. Additional Information on the Experimental Setup

D.1. Dataset Pre-Processing

MNIST receives an input image normalization of mean 0.1307 and standard deviation 0.3081. Furthermore, the images of CIFAR-10 are horizontally flipped, and per channel normalization is also applied with means 0.4914, 0.4822, 0.4465 and standard deviations 0.2023, 0.1994, 0.2010.

D.2. Model Architecture and Detailed Experimental Setup

In order to present the detailed architecture of the models used, we adopt the following compact notation introduced as done, e.g., in (El Mhamdi et al., 2021b).

L(#outputs) represents a **fully-connected linear layer**, R stands for **ReLU activation**, S stands for **log-softmax**, C(#channels) represents a *fully-connected 2D-convolutional layer* (kernel size 5, padding 0, stride 1), M stands for. **2D-maxpool** (kernel size 2), B stands for **batch-normalization**, and D represents **dropout** (with fixed probability 0.25).

The architecture of the models, as well as other details on the experimental setup, are presented in Table 2. Note that CNN stands for convolutional neural network, and NLL refers to the negative log likelihood loss.

D.3. Data Heterogeneity

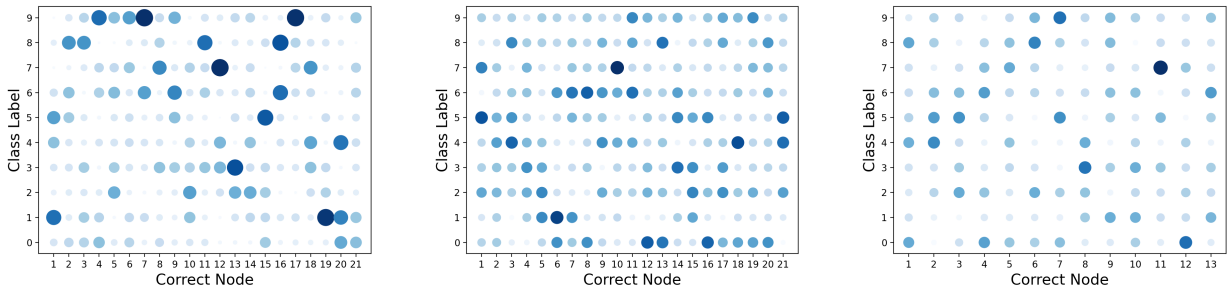

Figure 2. Distribution of class labels across correct nodes when sampling from a Dirichlet distribution of parameter α. *Left*: MNIST with $\alpha = 1$, *Middle*: MNIST with $\alpha = 5$, *Right*: CIFAR-10 with $\alpha = 5$.

We simulate data heterogeneity in the correct nodes' datasets by making the nodes sample from MNIST and CIFAR-10 using a **Dirichlet** distribution of parameter $\alpha > 0$. A smaller α implies a more heterogeneous setting (i.e., the more probable it is for nodes to sample datapoints from only one class). For MNIST, we choose $\alpha \in \{1, 5\}$, while we set $\alpha = 5$ on the more difficult task CIFAR-10. The corresponding distributions of class labels across correct nodes are shown in Figure 2.

Dataset	MNIST	CIFAR-10
Data heterogeneity	$\alpha \in \{0.5, 1, 5\}$	$\alpha = 5$
Model type	CNN	CNN
Model architecture	C(20)-R-M-C(20)-R-M-L(500)-R-L(10)-S	(3,32×32)-C(64)-R-B-C(64)-R-B-M-D-C(128)-R-B-C(128)-R-B-M-D-L(128)-R-D-L(10)-S
Loss	NLL	NLL
ℓ_2-regularization	10^{-4}	10^{-2}
Learning rate	$\gamma = 0.75$	$\gamma = 0.5$
Batch size	$b = 25$	$b = 50$
Momentum	$\beta = 0.99$ (except for SCC: $\beta = 0.9$)	$\beta = 0.99$ (except for SCC: $\beta = 0.9$)
Number of nodes	$n = 26$	$n = 16$
Number of faults	$f = 5$	$f = 3$
Number of Iterations	$T = 600$	$T = 2000$

Table 2. Detailed experimental setting of Section 5

D.4. Attacks

We use four state-of-the-art *gradient-based* attacks, namely *fall of empires (FOE)* (Xie et al., 2019), *a little is enough (ALIE)* (Baruch et al., 2019), *sign-flipping (SF)* (Allen-Zhu et al., 2020), and *label-flipping (LF)* (Allen-Zhu et al., 2020). Since these attacks are originally designed on gradients, we first modify them to be executed on parameter vectors. The first three adapted attacks (FoE, ALIE, and SF) rely on the following key notion. Let a_t be the attack vector in iteration t and let ζ_t be a fixed non-negative real number. In every iteration t, all faulty nodes broadcast the same vector $\overline{\theta_t} + \zeta_t a_t$ to all other nodes, where $\overline{\theta_t}$ is the average of the parameter vectors of the correct nodes in iteration t. Each attack among the first three follows the general scheme we just described, with the following particularities.

(a) **ALIE.** In this attack, $a_t = -\sigma_t$, where σ_t is the opposite vector of the coordinate-wise standard deviation of $\overline{\theta_t}$. In our experiments on ALIE, ζ_t is chosen through an extensive grid search. Essentially, in each iteration t, we choose the value that results in the worst faulty vector, i.e, the vector for which the distance to $\overline{\theta_t}$ is the largest.

(b) **FOE.** In this attack, $a_t = -\overline{\theta_t}$. All faulty nodes thus send $(1 - \zeta_t)\overline{\theta_t}$ in iteration t. Similar to *ALIE*, ζ_t for *FoE* is also estimated through grid searching.

(c) **SF.** In this attack, $a_t = -\overline{\theta_t}$ and $\zeta_t = 2$. All faulty nodes thus send $-\overline{\theta_t}$ in iteration t.

(d) **LF.** Under the LF attack, all faulty nodes send the same vector $\hat{\theta}_t$ in iteration t, where $\hat{\theta}_t$ is the average of the correct parameter vectors but computed on *flipped* labels. In order to do so, in each iteration t, the faulty nodes compute the gradients of the correct nodes on flipped labels. Since the labels for MNIST and CIFAR-10 are in $\{0, 1, ..., 9\}$, the labels are flipped such that $l' = 9 - l$ for every training datapoint, where l is the original label and l' is the flipped label. Each faulty node then averages all *flipped* parameter vectors to get $\hat{\theta}_t$.

D.5. Computing Infrastructure

D.5.1. Software Dependencies:

Python 3.8.10 has been used to run our scripts. Besides the standard libraries associated with Python 3.8.10, our scripts use the following libraries:

Library	Version
numpy	1.19.1
torch	1.6.0
torchvision	0.7.0
pandas	1.1.0
matplotlib	3.0.2
PIL	7.2.0
requests	2.21.0
urllib3	1.24.1
chardet	3.0.4
certifi	2018.08.24
idna	2.6
six	1.15.0
pytz	2020.1
dateutil	2.6.1
pyparsing	2.2.0
cycler	0.10.0
kiwisolver	1.0.1
cffi	1.13.2

Some dependencies are essential, while others are optional (e.g., only used to process the results and produce the plots).Furthermore, our code has been tested on the following OS: Ubuntu 20.04.4 LTS (GNU/Linux 5.4.0-121-generic x86_64).

D.5.2. HARDWARE DEPENDENCIES:

We list below the hardware components used:

- 1 Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz
- 2 Nvidia GeForce GTX 1080 Ti
- 64 GB of RAM

E. Additional Experimental Results

E.1. Remaining Plots on MNIST

We complete the missing results from Figure 1 in the main paper by presenting in Figures 3 and 4 the performance of MoNNA on MNIST with $\alpha = 1$ and 5, respectively.

As observed in Section 5.2, MoNNA is the only considered algorithm that provides consistently good performances when tested on MNIST in two heterogeneity regimes and in the presence of faulty nodes. Indeed, under all attacks, MoNNA almost matches the performance of D-SGD in terms of learning accuracy (as well as computational workload per node). While SCC showcases satisfactory results under LF, the FOE, ALIE, and SF attacks prevent the model from learning. Similar observations hold for BRIDGE and LEARN which are completely unable to learn, with their final accuracies stagnating at around 10%.

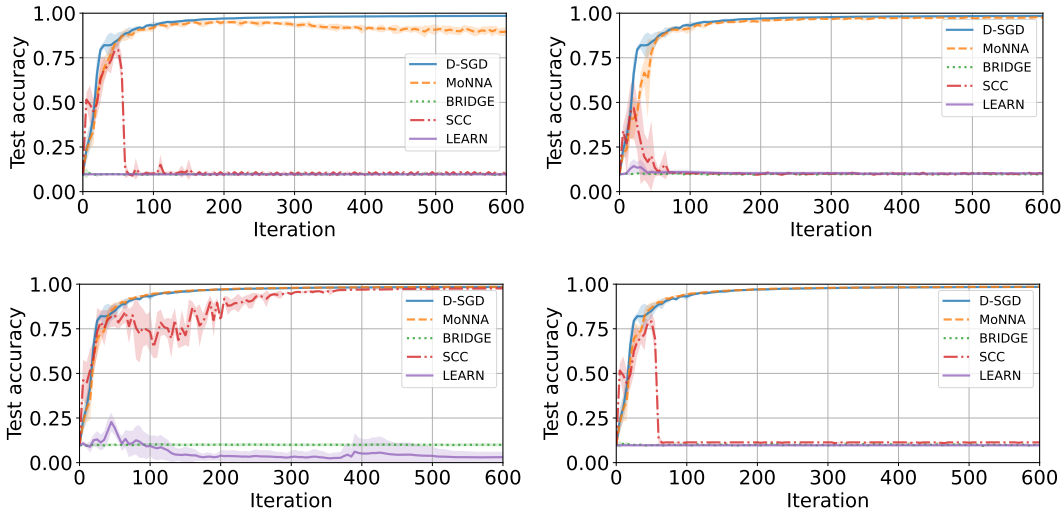

Figure 3. Learning accuracies achieved on MNIST with $\alpha = 1$ by D-SGD, MoNNA, BRIDGE, SCC, and LEARN. There are $n = 26$ nodes among which $f = 5$ are faulty. The faulty nodes execute the *FOE* (row 1, left), *ALIE* (row 1, right), *LF* (row 2, left), and *SF* (row 2, right) attacks. All algorithms except LEARN compute 15,000 gradients, while LEARN computes 180,300 gradients.

E.2. Necessity of Momentum and NNA

Our theoretical results indicate that the two key ingredients of MoNNA, namely Polyak's momentum and NNA, are sufficient to guarantee convergence in adversarial settings. In this section, we empirically measure their necessity by comparing our algorithm to momentum-less solutions as well as other algorithms that do not use NNA. In particular, in addition to D-SGD, MoNNA, and SCC, we run our experiments on two prominent aggregation algorithms from the literature, namely geometric median (GM) (Chen et al., 2017) and coordinate-wise trimmed mean (CWTM) (Yin et al., 2018). We also execute both GM and CWTM with momentum $\beta = 0.99$ (referred to as MoGM and MoCWTM, respectively). Additionally, we also compare MoNNA to its momentum-less variant, namely NNA. We report on these results on the CIFAR-10 and MNIST datasets in Figures 5 and 6, respectively.

Our observations are twofold. First, it is clear from Figures 5 and 6 that momentum plays a crucial role in ensuring the robustness of MoNNA. Indeed, momentum-less MoNNA (i.e., simply NNA) is completely unable to learn under all attacks, showcasing a very low accuracy constant at 10% throughout the entire learning. However, as previously mentioned, MoNNA drastically mitigates these attacks. Indeed, the model steadily increases in accuracy to finally reach 95% on MNIST and 75% on CIFAR-10 under all four attacks. Moreover, the importance of momentum is also further corroborated by the equally poor performances of CWTM and GM.

Second, we show the critical importance of the NNA scheme when defending against faulty nodes. Although much more resilient than its momentum-less counterpart (especially on MNIST), MoCWTM remains largely vulnerable to attacks which are able to completely hinder its learning on CIFAR-10. Indeed, even though the accuracy increases under FOE and SF, it plateaus at 50%, which is 25% less than the accuracy obtained with MoNNA on CIFAR-10. Additionally, ALIE completely annihilates the performance of MoCWTM, with a final accuracy close to 10%. The same observation holds for LF. Furthermore, while SCC and MoGM showcase good results under LF and ALIE respectively, the other attacks completely degrade their performances on both CIFAR-10 and MNIST. The **worst case performances** of MoNNA's rivals are thus very poor (unlike MoNNA which performs well in all cases). We argue that one should carefully examine this fundamental metric when evaluating the robustness of aggregation techniques, as the same algorithm can simultaneously greatly defend against some attacks but perform poorly against others.

This entire analysis demonstrates the superiority of our solution and suggests that momentum and NNA might be two necessary components in practice to ensure the robustness of distributed asynchronous systems.

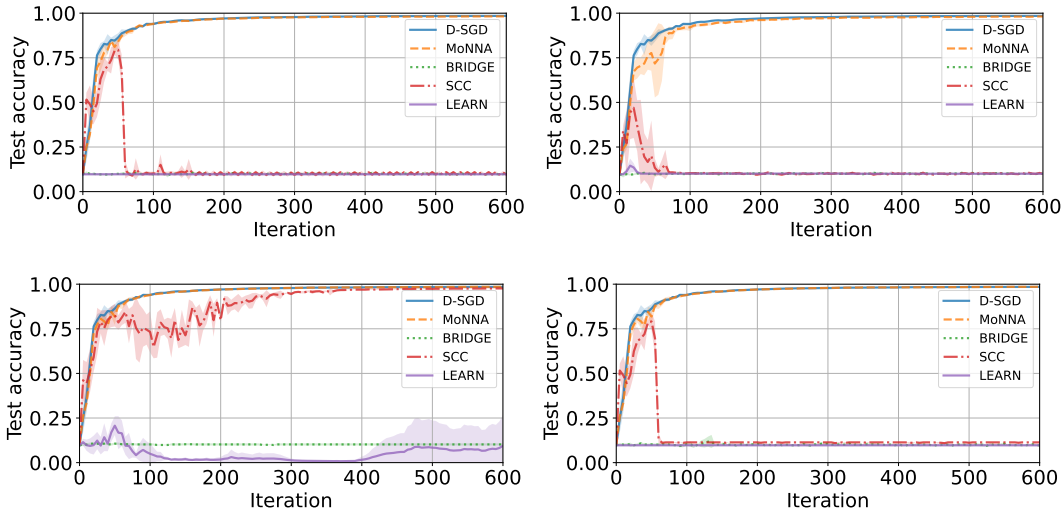

Figure 4. Learning accuracies achieved on MNIST with $\alpha = 5$ by D-SGD, MoNNA, BRIDGE, SCC, and LEARN. There are $n = 26$ nodes among which $f = 5$ are faulty. The faulty nodes execute the *FOE* (row 1, left), *ALIE* (row 1, right), *LF* (row 2, left), and *SF* (row 2, right) attacks. All algorithms except LEARN compute 15,000 gradients, while LEARN computes 180,300 gradients.

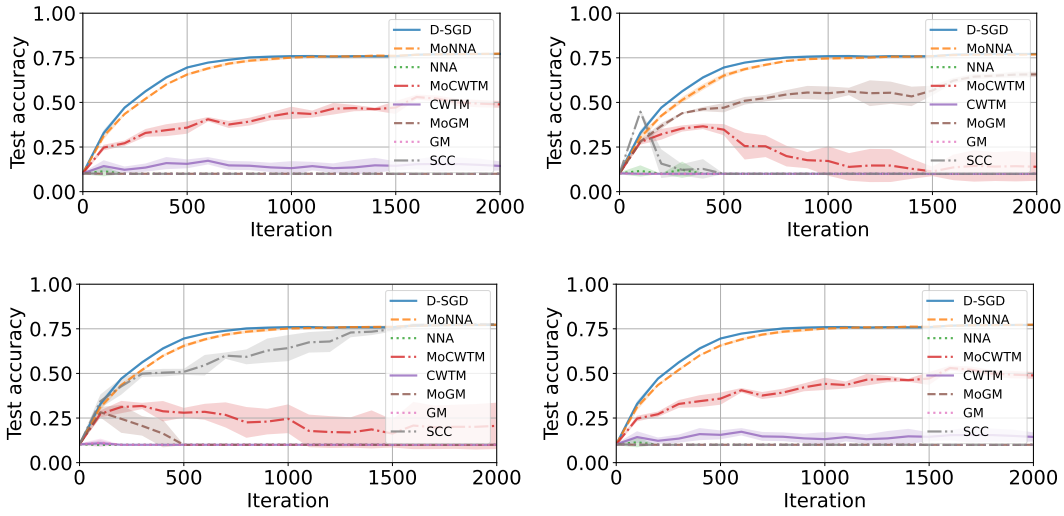

Figure 5. Comparison of the learning accuracies achieved by various aggregation algorithms on CIFAR-10 with $\alpha = 5$, notably including D-SGD, MoNNA, GM, MoGM (i.e., GM with $\beta = 0.99$), CWTM, MoCWTM (i.e., CWTM with $\beta = 0.99$), and SCC. There are $n = 16$ nodes among which $f = 3$ are faulty. The faulty nodes execute the *FOE* (row 1, left), *ALIE* (row 1, right), *LF* (row 2, left), and *SF* (row 2, right) attacks.

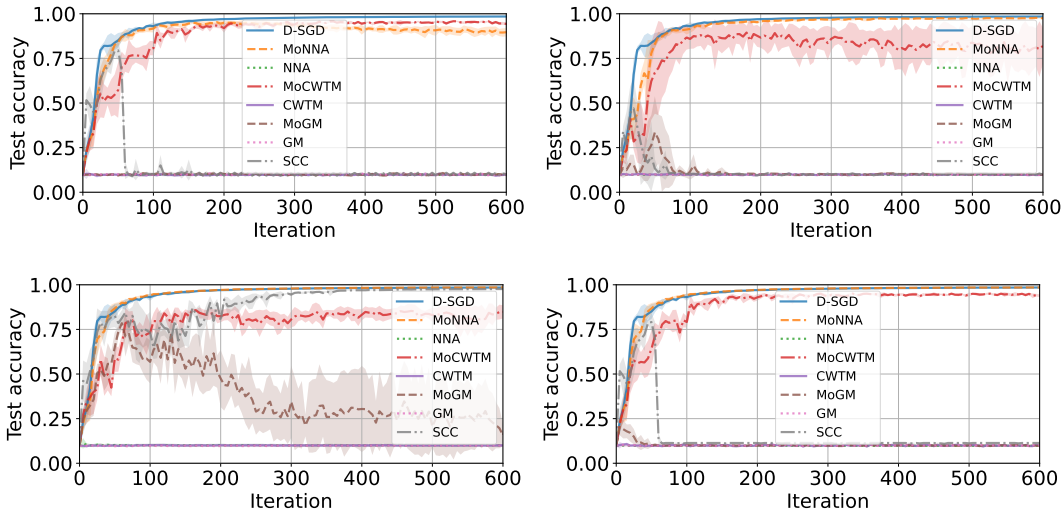

Figure 6. Comparison of the learning accuracies achieved by various algorithms on MNIST with $\alpha = 1$, notably including D-SGD, MoNNA, GM, MoGM (i.e., GM with $\beta = 0.99$), CWTM, MoCWTM (i.e., CWTM with $\beta = 0.99$), and SCC. There are $n = 26$ nodes, among which $f = 5$ are faulty. The faulty nodes execute *FOE* (row 1, left), *ALIE* (row 1, right), *LF* (row 2, left), and *SF* (row 2, right).