



**HAL**  
open science

# Achieving Optimal Breakdown for Byzantine Robust Gossip

Renaud Gaucher, Aymeric Dieuleveut, Hadrien Hendrikx

► **To cite this version:**

Renaud Gaucher, Aymeric Dieuleveut, Hadrien Hendrikx. Achieving Optimal Breakdown for Byzantine Robust Gossip. 2024. hal-04830823

**HAL Id: hal-04830823**

**<https://hal.science/hal-04830823v1>**

Preprint submitted on 11 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# ACHIEVING OPTIMAL BREAKDOWN FOR BYZANTINE ROBUST GOSSIP

**Renaud Gaucher**

Centre de mathématiques appliquées  
École polytechnique, Institut Polytechnique de Paris  
Palaiseau France  
renaud.gaucher@polytechnique.edu

**Aymeric Dieuleveut**

Centre de mathématiques appliquées  
École polytechnique, Institut Polytechnique de Paris  
Palaiseau France

**Hadrien Hendrikx**

Centre Inria de l'Univ. Grenoble Alpes  
CNRS, LJK  
Grenoble, France

## ABSTRACT

Distributed approaches have many computational benefits, but they are vulnerable to attacks from a subset of devices transmitting incorrect information. This paper investigates Byzantine-resilient algorithms in a decentralized setting, where devices communicate directly with one another. We investigate the notion of *breakdown point*, and show an upper bound on the number of adversaries that decentralized algorithms can tolerate. We introduce  $CG^+$ , an algorithm at the intersection of ClippedGossip and NNA, two popular approaches for robust decentralized learning.  $CG^+$  meets our upper bound, and thus obtains optimal robustness guarantees, whereas neither of the existing two does. We provide experimental evidence for this gap by presenting an attack tailored to sparse graphs which breaks NNA but against which  $CG^+$  is robust.

## 1 INTRODUCTION

Distributed machine learning, in which the training process is performed on multiple computing units (or nodes), responds to the increasingly distributed nature of data, its sensitivity, and the rising computational cost of optimizing models. While most distributed architectures rely on coordination from a central server, some communication networks favor peer-to-peer exchanges, and global coordination can be costly in these cases. Besides, the decentralized setting has several other perks: it mitigates the communication bottleneck and failure risk at the main server, and provides additional privacy guarantees since agents only have a local view of the system (Cyffers et al., 2022). However, distributing optimization over a large number of devices introduces new security issues: software may be faulty, local data may be corrupted, and nodes can be hacked or even controlled by a hostile party. Such issues are modeled as *Byzantine* node failures (Lamport et al., 1982), defined as omniscient adversaries able to collude with each other.

Standard distributed learning methods are known to be vulnerable to Byzantine attacks (Blanchard et al., 2017), which has led to significant efforts in the development of robust distributed learning algorithms. From the first works tackling Byzantine-robust SGD (Blanchard et al., 2017; Yin et al., 2018; Alistarh et al., 2018; El-Mhamdi et al., 2020), methods have been developed to tackle stochastic noise using Polyak momentum (Karimireddy et al., 2021; Farhadkhani et al., 2022a) and mixing strategies to handle heterogeneous loss functions (Karimireddy et al., 2020; Allouah et al., 2023). In parallel to these robust algorithms, efficient attacks have been developed to challenge Byzantine-robust algorithms (Baruch et al., 2019; Xie et al., 2020). To bridge the gap between algorithm performance and achievable accuracy in the Byzantine setting, tight lower bounds have been constructed for the heterogeneous setting (Karimireddy et al., 2020; Allouah et al., 2024). Yet, all these works rely on a trusted central server to coordinate training.

In contrast, the decentralized case has been less explored. In particular, it is still unclear how many Byzantine nodes can be tolerated over a given communication network before aggregation protocols fail. In fact, the network is often assumed to be fully connected (El-Mhamdi et al., 2021; Farhadkhani et al., 2023), and most papers that go beyond this assumption by addressing sparse graphs either do not give clear convergence rates or give weak guarantees on the asymptotic error (Peng et al., 2021; Fang et al., 2022; Wu et al., 2023). While criteria for using SGD with generic robust decentralized aggregation rules have been proposed (Wu et al., 2023; Farhadkhani et al., 2023), Decentralized SGD suffers from the same flaws. For instance, while  $(\alpha, \lambda)$ -reduction (Farhadkhani et al., 2023) is agnostic to the communication network, NNA, the associated robust communication scheme, is only introduced and analyzed for fully-connected networks. ClippedGossip (He et al., 2022), which consists in clipping the standard *gossip averaging* updates (Boyd et al., 2006) used for decentralized mean estimation, is designed for sparse networks. Unfortunately, its clipping threshold requires inaccessible information, and the theory only applies for a small fraction of Byzantine failures.

Our work revisits both the ClippedGossip and NNA frameworks to solve the aforementioned shortcomings. To do so, we carefully study the decentralized mean estimation problem. This seemingly simple problem retains most of the difficulty of handling Byzantine nodes while allowing us to derive strong convergence and robustness guarantees. We then tackle general (smooth non-convex) optimization problems through a reduction. Our contributions are the following.

**1 - New upper bound on the breakdown point of robust algorithms.** We show that in general, Byzantine robust algorithms fail arbitrarily if the number of Byzantine neighbors per node exceeds a given threshold. This threshold is expressed in terms of spectral quantities of the graph (the *algebraic connectivity*), and recovers usual ones for fully-connected topologies.

**2 - Practical robust gossip algorithm with optimal breakdown.** We propose  $CG^+$ , a practical robust aggregation procedure at the intersection of ClippedGossip and NNA. We show that  $CG^+$  is robust as soon as the number of Byzantine neighbors per agent does not exceed the upper bound on the breakdown point indicated in the previous paragraph (up to an additive factor of 2). This indicates that our upper bound is tight, and that  $CG^+$  obtains optimal breakdown. We also extend the existing NNA aggregation method to sparse graphs, but do not achieve optimal breakdown in this case.

**3 - Efficient decentralized optimization beyond averaging.** Using appropriate reductions (Farhadkhani et al., 2023), we show that Decentralized SGD on top of  $CG^+$  obtains state-of-the-art convergence guarantees, including when comparing with specialized approaches on fully-connected networks. We experimentally demonstrate the effectiveness of our proposed algorithm.

**4 - Stronger attacks for decentralized systems.** We propose a theoretically grounded attack specifically designed to challenge decentralized algorithms. In this setting, honest nodes hold different parameters, and Byzantine nodes can force them to diverge in opposite directions by declaring distinct messages to each of the honest nodes. Our attack, called *Spectral Heterogeneity* (SpH), uses the eigenvectors of the graph’s Laplacian matrix to identify which attack has to be made on each honest node to disrupt the communication.

We summarize our contributions in Table 1, comparing our algorithm to He et al. (2022) and Farhadkhani et al. (2022b). The remainder of the paper is organized as follows. Section 2 introduces the notion of Byzantine robust decentralized optimization. Section 3 presents the upper bound on the breakdown point for sparse graphs. Section 4 introduces  $CG^+$ , shows it has optimal breakdown and evaluates its robustness both for one-step aggregation and for D-SGD on top of  $CG^+$ . Finally, Section 5 presents the *Spectral Heterogeneity* attack designed for disrupting decentralized robust optimization scheme, as well as an experimental evaluation of several robust aggregation schemes against various attacks (including ours).

## 2 BACKGROUND

### 2.1 DECENTRALIZED OPTIMIZATION.

We consider a system composed of  $m$  computing units that communicate synchronously through a communication network, which is represented as an undirected graph  $\mathcal{G}$ . We denote by  $\mathcal{H}$  the set of honest nodes, and  $\mathcal{B}$  the (unknown) set of Byzantine nodes. Each unit  $i$  holds a local parameter  $\mathbf{x}_i \in \mathbb{R}^d$ , a local loss function  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ , and can communicate with its neighbors in the graph  $\mathcal{G}$ .

Method	Scheme	Threshold	Breakdown (decentralized)	
			Existing	New
NNA	Trimming	Adaptive and computable	No general result	$8b > \mu_{\min}$
Clipped Gossip	Clipping	Impractical	$O(\gamma)$ from optimal	
CG <sup>+</sup>	Clipping	Adaptive and computable		Optimal <sup>†</sup> ( $2(b+1) > \mu_{\min}$ )

Table 1: Summary of our results and comparison to previous work. NNA only had results for the fully-connected topology. †: up to an additive factor of 2.

We denote the set of neighbors of node  $i$  by  $n(i)$  and by  $n_{\mathcal{H}}(i)$  (resp.  $n_{\mathcal{B}}(i)$ ) the set of honest (resp. Byzantine) ones.

We study decentralized algorithms for solving

$$\arg \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ f_{\mathcal{H}}(\mathbf{x}) := \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} f_i(\mathbf{x}) \right\}. \quad (1)$$

Due to the averaging nature of Equation (1), centralized algorithms for solving this problem rely on global averaging of the gradients computed at each node. In the decentralized setting, we rely on (local) inexact averaging instead.

**Gossip Communication.** Standard decentralized optimization algorithms typically rely on the so-called *gossip* communication protocol (Boyd et al., 2006; Nedic & Ozdaglar, 2009; Scaman et al., 2017; Kovalev et al., 2020). The gossip protocol consists in updating parameters of any node  $i$  with a linear combination of the parameters of its neighbors, with updates of the form  $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \eta \sum_{j=1}^m w_{ij} \mathbf{x}_j^t$ , where  $\eta \geq 0$  denotes a communication step-size. The matrix of the weights  $\mathbf{W} = (w_{ij})_{i,j}$  is called the *gossip matrix*, and naturally defines the communication graph  $\mathcal{G}$ , in the sense that  $w_{ij} = 0$  if nodes  $i$  and  $j$  are not neighbors. By considering the matrix of honest parameters  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)^T$ , the gossip update is also conveniently written as

$$\mathbf{X}^{t+1} = \mathbf{X}^t - \eta \mathbf{W} \mathbf{X}^t. \quad (2)$$

In this work, we instantiate gossip algorithms by using as gossip matrix the Laplacian matrix of the graph defined as  $\mathbf{W} = \mathbf{D} - \mathbf{A}$ , where  $\mathbf{D}$  is the diagonal matrix of the degrees and  $\mathbf{A}$  is the adjacency matrix of the graph. The Laplacian matrix is symmetric non-negative. We denote by  $\mu_{\max}(\mathcal{G}_{\mathcal{H}})$  and  $\mu_2(\mathcal{G}_{\mathcal{H}})$  the largest and smallest non-zero eigenvalues of the Laplacian matrix  $\mathbf{W}_{\mathcal{H}}$  of the *honest subgraph*  $\mathcal{G}_{\mathcal{H}}$ , and by  $\gamma = \mu_2(\mathcal{G}_{\mathcal{H}})/\mu_{\max}(\mathcal{G}_{\mathcal{H}})$  its *spectral gap*. Spectral properties of the gossip matrix are known to characterize the convergence of gossip optimization methods. For instance, in the absence of Byzantine nodes, the plain gossip update with step-size  $\eta \leq \mu_{\max}(\mathcal{G})^{-1}$  leads to a linear convergence of the nodes parameter values to the average of the initial parameters:  $\|\mathbf{X}^t - \bar{\mathbf{X}}^0\|^2 \leq (1 - \eta \mu_2(\mathcal{G}))^t \|\mathbf{X}^0 - \bar{\mathbf{X}}^0\|^2$ , for  $\bar{\mathbf{X}}^0$  the matrix with columns  $m^{-1} \sum_{j=1}^m \mathbf{x}_j^0$ .

**Robustness Issue.** Gossip communication relies on updating nodes parameters by performing non-robust local averaging. As such, similarly to the centralized case, any Byzantine neighbor of node  $i$  can drive the update to any desired value (Blanchard et al., 2017). Then, the poisoned information spreads through gossip communications.

## 2.2 BYZANTINE ROBUST OPTIMIZATION.

**Threat model.** We consider Byzantine nodes to be omniscient adversaries, able to collude and to send distinct values to each of their neighbors. As such, the exact number  $|\mathcal{B}|$  of Byzantine nodes, a common metric to quantify robustness in centralized (or fully-connected) settings, does not matter, and we instead bound the number of Byzantine neighbors that each honest node has, i.e.  $\{|n_{\mathcal{B}}(i)|; i \in \mathcal{H}\}$ , as it is standard in decentralized analyses (He et al., 2022; Wu et al., 2023).

Similarly, the total number of honest nodes does not provide relevant information anymore, as the results depend on *how* they are linked, i.e., the topology of the honest subgraph. Therefore, for sparse topologies, we need to make an assumption about some property of the graph related to its topology, instead of simply the number honest neighbors. In the remainder of this paper, we consider spectral

properties of the Laplacian of the honest subgraph as a relevant quantity for robustness analyses. Yet, we emphasize that our results on the breakdown depend on the spectral properties of *the honest subgraph*, meaning that for a given graph, these properties change depending on the location of Byzantine nodes. We introduce the following class of graphs to take this dependence into account.

**Definition 1.** For any  $\mu_{\min} \geq 0$  and  $b \in \mathbb{N}$ , we define the class of graphs

$$\Gamma_{\mu_{\min}, b} = \left\{ \mathcal{G} \text{ s.t. } \mu_2(\mathcal{G}_{\mathcal{H}}) \geq \mu_{\min} \text{ and } \max_{i \in \mathcal{H}} |n_{\mathcal{B}}(i)| \leq b \right\}.$$

In other words, we introduce a subset of all possible graphs, partitioning in terms of (i) their algebraic connectivity, that is restricted to be larger than a minimal value  $\mu_{\min}$ , and (ii) the maximal number of Byzantine neighbors of a honest node, that is restricted to be smaller than  $b$ .

One should read Definition 1 as a sparse graph extension of the standard “*there are at most  $b$  byzantine nodes and at least  $|\mathcal{H}|$  honest ones*”, which now involves the relative positions of Byzantine and honest nodes in the graph. For given  $b$  and  $\mu_{\min} \geq 0$ , depending on the location of the Byzantine nodes, a given graph topology can either fall in  $\Gamma_{\mu_{\min}, b}$  (if Byzantine nodes are “well-spread”), or not (if they are adversarially chosen).

**Approximate Average Consensus.** The average consensus problem consists in finding the average of  $m$  vectors locally held by nodes. Average consensus is a specific case of Equation (1) obtained by taking  $f_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}_i\|^2$ . Because of adversarial attacks, some bias is introduced by Byzantine nodes during aggregation steps, so only an *approximate* solution of the average of honest nodes vector  $\bar{\mathbf{y}}_{\mathcal{H}} := |\mathcal{H}|^{-1} \sum_{i \in \mathcal{H}} \mathbf{y}_i$  can be expected. Thus, in the centralized setting, the variance between honest nodes is reduced to 0 (consensus is reached in one step), at the cost of some bias. In the decentralized setting (gossip), exact consensus is not possible: in this case, we aim for *reducing the variance*, still at the cost of some bias. The guarantees we give in Section 4 exactly reflect this trade-off: they quantify how much variance reduction is obtained at each step, and at the cost of what bias. Note that variance reduction here is to be understood as how different the parameters of the various nodes are, and is not directly linked with the variance of stochastic gradients. We now introduce the  $\alpha$ -robustness of a communication algorithm on a graph.

**Definition 2** ( $\alpha$ -robustness on  $\mathcal{G}$ ). For any  $\alpha < 1$ , a communication algorithm  $A$  is  $\alpha$ -robust on a graph  $G$  if from any initial local parameters  $\{\mathbf{x}_i; i \in \mathcal{H}\}$ , it allows any honest node  $i$  to compute a vector  $\hat{\mathbf{x}}_i$  such that

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\hat{\mathbf{x}}_i - \bar{\mathbf{x}}_{\mathcal{H}}\|^2 \leq \alpha \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{H}}\|^2.$$

Imposing  $\alpha < 1$  means that we would like to be closer to the initial solution after the aggregation step than before (the variance reduction needs to be larger than the bias we introduce). Note that  $\alpha = 1$  can trivially be achieved by not communicating at all. Remark that the  $\alpha$ -robustness of an algorithm on a graph  $\mathcal{G}$  means that a *single step* of the algorithm strictly reduces the average quadratic error. However, it does not mean that multiple steps would result in a geometric decrease, indeed, we cannot simply use induction as  $\bar{\hat{\mathbf{x}}}_{\mathcal{H}} \neq \bar{\mathbf{x}}_{\mathcal{H}}$ . In the following, we show that  $\alpha$ -robustness on all graphs in  $\Gamma_{\mu_{\min}, b}$  cannot be achieved for all values of  $\mu_{\min}$  and  $b$ .

### 3 FUNDAMENTAL LIMITS OF DECENTRALIZED COMMUNICATION SCHEMES

In this section we provide an upper bound on the number of Byzantine neighbors that can be tolerated by any algorithm running on a communication network in which the honest subgraph has a given algebraic connectivity.

**Theorem 1.** For any  $\mu_{\min} \geq 0$ ,  $b \geq 0$ , if  $\mu_{\min} \leq 2b$ , then for any  $H \geq 2b$ , there exists  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$  with  $H$  honest nodes such that no communication algorithm can be  $\alpha$ -robust on  $\mathcal{G}$ .

*Sketch of proof.* The proof relies on considering a specific graph  $G_{H, b}$ , decomposed as three cliques of  $H/2$  nodes, for an even  $H$  such that  $H/2 \geq b$ . We choose that nodes in any of the three cliques is neighbor to exactly  $b = k$  nodes in each of the two other cliques, in circular order. Finally, we assume that two of the three cliques are honest, and the third one is composed of Byzantine nodes.

First,  $G_{H,b} \in \Gamma_{\mu_{\min},b}$ . Indeed, the Laplacian matrix of the honest subgraph has an algebraic connectivity of  $2b$ , i.e.  $\mu_2((G_{H,b})_{\mathcal{H}}) = 2b$ , and  $2b \geq \mu_{\min}$  by our assumption. Moreover, each honest node has exactly  $b$ -Byzantine neighbors. Second, we show that no algorithm can be  $\alpha$ -robust on  $G$ . Indeed, for any of the two honest cliques, the Byzantine clique is indistinguishable from the other honest clique. As such, Byzantine nodes can send parameter values such that honest nodes cannot hope to improve the global error in general. We refer the reader to Appendix B.1 for details.  $\square$

It follows from Theorem 1 that, when aiming at obtaining a theoretical guarantee that quantifies the robustness of the honest graph through  $\mu_{\min}$ , we must have  $\mu_{\min} > 2b$ . In the specific case of a fully connected graph, where  $\mu_{\min} = |\mathcal{H}|$ , this condition boils down to requiring  $|\mathcal{H}| > 2|\mathcal{B}|$ , which is aligned with common robustness criteria for distributed system (Lamport et al., 1982; Vaidya et al., 2012; El-Mhamdi et al., 2021). As we show in Section 4, this upper bound on  $b$  is tight in the sense that there exists an  $\alpha$ -robust communication scheme over all  $G \in \Gamma_{\mu_{\min},b}$  as soon as  $2b + 2 < \mu_{\min}$ .

**Algebraic connectivity as a robustness criterion.** Note that Theorem 1 does not imply that all aggregation methods fail as long as  $2b \geq \mu_{\min}$ , but rather than since there exists a graph for which it is the case, one cannot prove that an aggregation method works with  $b > \mu_{\min}/2$  Byzantine nodes for all graphs. Yet, one can still prove breakdown points using other graph-related quantities (which might lead to tolerating  $b > \mu_{\min}/2$  Byzantine nodes for some graph architectures), or restricting the graph topologies considered. This gap is standard in the optimization literature, since convergence results often involve the spectral properties of the gossip matrix, whereas they do not naturally appear when proving lower bounds on the number of iterations required to reach a certain accuracy. For instance, the convergence rates of decentralized optimization algorithms depend on the (square root of the) *spectral gap* of the Laplacian matrix of the communication graph (Scaman et al., 2017; Kovalev et al., 2020), whereas iteration lower bounds are proven in terms of diameter. Yet, these decentralized algorithms are termed as optimal as their guarantees match the lower bound on the path graph.

**The Approximate Consensus Problem and dimension-dependent breakdown points.** The design of algorithms aiming at finding the average of parameters within a communication network is related to the *approximate consensus problem* (ACP) (Dolev et al., 1986). In the standard ACP problem, nodes need to converge to the same value while remaining within the *convex hull* of initial parameters. Yet, communication-optimal methods for this problem are memory and computationally expensive (Fekete, 1986). More recently, the work of LeBlanc et al. (2013) aims at designing communication schemes that only use local information with computationally efficient aggregation rules. They show that standard robustness criterion of *connectivity* (Sundaram & Hadjicostis, 2010) does not properly reflect the robustness of a network for such methods. To mitigate this issue, they introduce the notion of  $r$ -robust networks. Vaidya (2014) generalizes this result by proving that the ACP cannot be solved using an algorithm with *iterative communication* on a system of  $m$  nodes with  $b$  Byzantine failures in dimension  $d$  when  $m \leq (d + 2)b + 1$ . This dependence on the dimension - intractable for ML usage - derives from the requirement of staying within the *convex hull* of initial parameters for solving ACP: staying in the convex hull of initial parameters is increasingly difficult as the dimension increases. On the contrary, our definition of  $\alpha$ -robustness only requires the algorithms to improve the average squared distance to the target value. This relaxation of the consensus requirement allows us to prove *dimension independent* breakdown point. Linking LeBlanc et al. (2013)'s robustness criterion with algebraic connectivity is an interesting direction for future work.

## 4 CLIPPED GOSSIP +

In this section, we introduce a robust gossip scheme derived from ClippedGossip (He et al., 2022), but with a well-chosen and practical clipping threshold that makes it closely related to NNA (Farhadkhani et al., 2023). This scheme verifies two key properties: **(i)** a contraction property (bias-variance tradeoff); and **(ii)** it has an optimal breakdown point up to a small additive constant. Our one-step bias-variance characterizations can directly be plugged in results from Farhadkhani et al. (2023) to show state-of-the-art guarantees for D-SGD on top of  $\text{CG}^+$ .

### 4.1 THE ALGORITHM

**Making Gossip Robust.** ClippedGossip (He et al., 2022) is a robust aggregation algorithm, in which each node projects the parameters declared by its neighbors on a ball centered at its own

parameter before performing a local averaging step. Given a “communication step-size”  $\eta > 0$ , it writes

$$\text{ClippedGossip}_{\tau_i}(\mathbf{x}_i; (\mathbf{x}_j)_{j \in n(i)}) := \mathbf{x}_i + \eta \sum_{j \in n(i)} \text{Clip}(\mathbf{x}_j - \mathbf{x}_i; \tau_i). \quad (\text{CG})$$

The clipping threshold should be set with extreme care, as it is instrumental in limiting malicious nodes’ influence, but can also slow the algorithm down, or introduce bias. With a constant clipping threshold, a constant bias is added at each step, and so one needs to limit the number of aggregation step to avoid infinite drift. A better approach is to choose the clipping threshold adaptively, depending on the pairwise distances with neighbors.

This is what He et al. (2022) do, introducing the following adaptive threshold:  $\tau_i^t := (\frac{1}{(|\mathcal{H}|-b)^b} \sum_{j \in n_{\mathcal{H}}(i)} \mathbb{E} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2^2)^{1/2}$ . Yet, this threshold leads to far from optimal robustness guarantees (with a breakdown least  $O(\gamma)$  from the optimal one), and cannot be computed in reasonable practical settings: not only does each node need to know the variance of the noise of its neighbors, but *it also requires to know which nodes are honest*. This breaks the fundamental assumption of not knowing the identity (honest or Byzantine) of the nodes. While an efficient rule of thumb is proposed to circumvent this, it is not supported by theory.

**Nearest Neighbors Averaging (NNA).** Another baseline for robust decentralized averaging is NNA (Farhadkhani et al., 2023), proposed for *fully-connected communication graphs*. In this rule, each node gathers the  $n(i)$  parameters from its neighbors, drops the  $b$  furthest from its own, and averages the  $n(i) - b$  remaining ones (trimmed mean). While Farhadkhani et al. (2023) only introduce and analyze it for the fully-connected setting, we argue that this rule naturally extends to sparse graphs. Indeed, NNA can be viewed as performing update CG, but where the  $x \mapsto \text{Clip}(x, \tau)$  operator is replaced by  $x \mapsto x \mathbb{1}(\|x\| \leq \tau_i)$  where  $\mathbb{1}$  is the indicator function and  $\tau_i$  is chosen as the  $b + 1$ -th furthest value (and a well-chosen  $\eta = (n(i) - b + 1)^{-1}$ ). *The largest updates are dropped instead of clipped*. We give an analysis for this sparse graphs version of NNA and show in Theorem 3 that dropping large updates is too brutal, and does not lead to an optimal breakdown. In the remainder of this paper, we may simply call this method NNA, although it generally refers to the sparse graphs extension that we introduce. Note that in Farhadkhani et al. (2023), a subset of the nodes is allowed not to respond to account for messages loss, or Byzantine nodes deciding not to send messages. We do not consider such a variation here.

**CG<sup>+</sup>: the best of both worlds.** We now go back to the ClippedGossip framework (of which NNA is a variant, as discussed above), and introduce the following clipping rule:

1. Each honest node  $i$  computes the norm of the differences of its parameter with the one of its neighbors:  $S_i^t = \{\|\mathbf{x}_i^t - \mathbf{x}_j^t\|; j \in n(i)\}$ .
2. Define as clipping threshold the  $b + 1$  largest value of the set  $S_i^t$ :  $\tau_{i,t}^{\text{CG}^+} := Q_{\frac{|n(i)|-b+1}{|n(i)|}}(S_i^t)$ .

This threshold can then be used in (CG) to obtain CG<sup>+</sup>. Note that this threshold is close to the one discussed for NNA, using the local number of neighbors  $n(i)$  instead of  $n$ , but CG<sup>+</sup> uses the clipping operator instead of  $x \mapsto x \mathbb{1}(\|x\| \leq \tau_i)$ . Therefore, CG<sup>+</sup> is an interesting midpoint between ClippedGossip and NNA: it can be viewed either as performing the sparse graph extension of NNA but with the clipping operator, or performing ClippedGossip but with NNA-type thresholds.

Here,  $b$  is a parameter of the algorithm, which corresponds to the number of Byzantine nodes that we would like to be robust to. We do not need to know the exact number of Byzantine nodes, but simply need to specify to how many we would like to be robust.

**Gossip with CG<sup>+</sup> has only a linear computational overhead.** The first step of computing  $\tau_{i,t}^{\text{CG}^+}$  requires that node  $i$  performs  $\mathcal{O}(d \cdot |n(i)|)$  computations, and the third one can be done in  $\mathcal{O}(|n(i)|)$  on average using Quickselect, so that CG<sup>+</sup> only requires  $\mathcal{O}(d \cdot |n(i)|)$  computations. Consequently, in average, CG<sup>+</sup> has the same linear complexity with respect to  $d$  and  $|n(i)|$  as simple averaging.

## 4.2 CONVERGENCE RESULTS

As briefly discussed in the introduction, **the goal of communicating is to reduce the variance, which comes at the price of bias**. This is unavoidable, since communicating allows nodes to

inject wrong information which biases the system. We now tightly quantify how much a single step of  $\text{CG}^+$  reduces the variance, and how much bias is injected in the process. Let us denote  $\text{Var}_{\mathcal{H}}(\mathbf{x}) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{H}}\|^2$ , the variance of honest nodes.

**Theorem 2.** *Let  $b$  and  $\mu_{\min}$  be such that  $2(b+1) \leq \mu_{\min}$ , and let  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ . Then, assuming  $\eta \leq \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$ , the output  $\mathbf{y} = \text{CG}^+(\mathbf{x})$  (obtained by one step of  $\text{CG}^+$  on  $\mathcal{G}$  from  $\mathbf{x}$ ) verifies:*

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{y}_i - \bar{\mathbf{x}}_{\mathcal{H}}\|^2 \leq (1 - \eta(\mu_{\min} - 2(b+1))) \text{Var}_{\mathcal{H}}(\mathbf{x}) \quad (3)$$

$$\|\bar{\mathbf{y}}_{\mathcal{H}} - \bar{\mathbf{x}}_{\mathcal{H}}\|^2 \leq 2\eta(b+1) \text{Var}_{\mathcal{H}}(\mathbf{x}). \quad (4)$$

In particular,  $\text{CG}^+$  is  $(1 - \eta(\mu_{\min} - 2(b+1)))$ -robust on all  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ .

Note that Equation (3) cannot be chained directly, but it can if we notice that  $\text{Var}_{\mathcal{H}}(\mathbf{y}) \leq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{y}_i - \bar{\mathbf{x}}_{\mathcal{H}}\|^2$ . We refer the reader to Appendix B for the proof. While the bound on parameter  $\eta$  depend on the honest subgraph,  $\mu_{\max}(\mathcal{G}_{\mathcal{H}}) \leq \mu_{\max}(\mathcal{G})$ , so  $\eta$  can be set conservatively by evaluating  $\mu_{\max}$  on the whole graph.

**Tight breakdown point.** Theorem 2 shows that the upper bound on the breakdown point from Section 3 is tight, since it shows that if  $2(b+1) < \mu_{\min}$ ,  $\text{CG}^+$  is  $\alpha$ -robust on  $\mathcal{G}$  for any  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ , while Theorem 1 says that this is impossible as soon as  $2b \geq \mu_{\min}$ . Note that while  $\alpha$ -robustness is guaranteed for the whole class, the value of  $\alpha$  will depend on the actual graph within the class (through  $\gamma$ ). The only gap left is when  $\mu_{\min} \in \{2b+1, 2b+2\}$ . This is a significant improvement over He et al. (2022), who obtain an equivalent result, but where the  $\mu_{\min} - 2(b+1)$  factor is essentially replaced by  $\mu_{\min} - c\sqrt{b\mu_{\max}}$  (for regular graphs for instance), where  $c > 0$  is a constant factor. This means that they obtain a breakdown of  $b \leq c^2\gamma\mu_{\min}$ , and so they lose non-negligible constant factors as well as a full  $\gamma$  factor, which rapidly shrinks with the size (and connectivity) of the graph. In other words, our guarantees are comparable when the number of Byzantine agents is small, but theirs collapse significantly before the actual breakdown point, whereas ours gracefully loosen until it is actually impossible to guarantee anything.

**Chaining aggregation steps.** When low variance levels are required, it is necessary to perform several aggregation steps one after the other. This contrasts with the centralized setting, in which the variance can be brought to zero in one step. While the variance reduces at a linear rate, the bias accumulates as more robust aggregation steps are performed. We provide bounds for  $t$  aggregation steps in the following Corollary.

**Corollary 1.** *Let  $b$  and  $\mu_{\min}$  be such that  $2(b+1) \leq \mu_{\min}$ , let  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ , and denote  $\delta = \frac{2(b+1)}{\mu_{\min}}$  and  $\gamma = \mu_{\min}/\mu_{\max}(\mathcal{G}_{\mathcal{H}})$ . Then, let  $(\mathbf{x}^t)_{t \geq 0}$  be obtained from any  $\mathbf{x}^0$  through  $\mathbf{x}^{t+1} = \text{CG}^+(\mathbf{x}^t)$ , with  $\eta = \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$ . We have that for any  $t \geq 0$ ,*

$$\text{Var}_{\mathcal{H}}(\mathbf{x}^t) \leq (1 - \gamma(1 - \delta))^t \text{Var}_{\mathcal{H}}(\mathbf{x}^0), \quad (5)$$

$$\|\bar{\mathbf{x}}_{\mathcal{H}}^t - \bar{\mathbf{x}}_{\mathcal{H}}^0\| \leq \frac{\sqrt{\gamma\delta} (1 - [1 - \gamma(1 - \delta)]^{t/2})}{1 - \sqrt{1 - \gamma(1 - \delta)}} \sqrt{\text{Var}_{\mathcal{H}}(\mathbf{x}^0)}. \quad (6)$$

When  $t \rightarrow \infty$ , we have that  $\text{Var}_{\mathcal{H}}(\mathbf{x}^t) \rightarrow 0$  (so, consensus is reached) and:

$$\|\bar{\mathbf{x}}_{\mathcal{H}}^t - \bar{\mathbf{x}}_{\mathcal{H}}^0\|^2 \leq \frac{\gamma\delta}{(1 - \sqrt{1 - \gamma(1 - \delta)})^2} \text{Var}_{\mathcal{H}}(\mathbf{x}^0) \leq \frac{4\delta}{\gamma(1 - \delta)^2} \text{Var}_{\mathcal{H}}(\mathbf{x}^0). \quad (7)$$

The proof of this Corollary is given in Appendix B.3. Equation (7) shows that while one-step convergence results ensure that the total L2 error (bias plus variance) decreases, we can be in a situation in which the total L2 distance increases after several  $\text{CG}^+$  steps because of bias accumulation. This happens when the factor multiplying the variance in Equation (7) is larger than 1, which essentially happens when  $\gamma \ll \delta$ . Yet, despite this bias, the output of the robust aggregation procedure are (arbitrarily) close to consensus, which can be desirable.

**Dependence on the parameters.** As expected, the bias increases with the amount of Byzantine corruption (through  $\delta$ ), and decreases as the graph becomes more connected (i.e,  $\gamma \rightarrow 1$ ). One can then use parameter  $\eta$  (up to its maximum value) to control the bias-variance trade-off of the aggregation procedure.



Another important method in the decentralized setting is the adaptation of NNA (Farhadkhani et al., 2023) discussed in Section 4.1, which consists in choosing the same  $\tau_i$  as in  $\text{CG}^+$ , but *dropping* updates such as  $\|x_j - x_i\| > \tau_i$  instead of clipping them to  $\tau_i$ . We extend our analysis to show that NNA is robust over sparse graphs as well, as shown in the following result.

**Theorem 3** (NNA breakdown for sparse graphs.). *NNA also verifies the guarantees of Corollary 1, where  $\delta$  is replaced by  $\tilde{\delta} = 8b/\mu_{\min}$ .*

The proof of this result can be found in Corollary 4. Farhadkhani et al. (2023) provide a guarantee for NNA, but only in the specific case of fully-connected graphs. We both improve the constants in the fully-connected case, and generalize this result to arbitrary sparse graphs.

Note that  $\tilde{\delta} > \delta$  for  $b > 1$ , so NNA is worse than  $\text{CG}^+$ . This worse breakdown is not an artifact of the analysis, as we verify in Section 5.3, where we show experiments in which NNA breaks before  $\text{CG}^+$  does.

### 4.3 BYZANTINE ROBUST DISTRIBUTED SGD ON GRAPHS

We now give convergence results for a D-SGD-type algorithm which uses  $\text{CG}^+$  for decentralized robust aggregation. Several works on Byzantine-robust SGD abstract away the aggregation procedure through some contraction properties (Karimireddy et al., 2021; Wu et al., 2023; Farhadkhani et al., 2023), so that the global D-SGD result follows from the robustness of the averaging procedure. Corollary 2 builds on the reduction from Farhadkhani et al. (2023), since their requirements on the aggregation procedure exactly matches the guarantees of Theorem 2. We consider Problem 1, where we assume that each local function  $f_i$  is a risk computed using a loss  $\ell$  on a data distribution  $\mathcal{D}_i$ , i.e  $f_i(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}_i} [\nabla \ell(\mathbf{x}, \xi)]$ . We propose to solve Problem 1 using decentralized stochastic gradient descent over a communication network  $\mathcal{G}$ . Robustness to Byzantine nodes is obtained using  $\text{CG}^+$  as the aggregation rule, coupled with Polyak momentum to reduce the stochastic noise.

---

#### Algorithm 1 Byzantine-Resilient Decentralized SGD with $\text{CG}^+$

---

**Input:** Initial model  $\mathbf{x}_i^0 \in \mathbb{R}^d$ , local loss functions  $f_i$ , initial momentum  $\mathbf{m}_i^0 = 0$ , momentum coefficient  $\beta = 0$ , learning rate  $\rho$ , communication step size  $\eta = \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$ , assumption on Byzantine local corruption  $b$ .

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

**for**  $i \in \mathcal{H}$  **in parallel do**

    Compute a noisy oracle of the gradient:  $\mathbf{g}_i^t = \nabla f_i(\mathbf{x}_i^t) + \xi_i^t$ .

    Update the local momentum:  $\mathbf{m}_i^t = \beta \mathbf{m}_i^{t-1} + (1 - \beta) \mathbf{g}_i^t$ .

    Make an optimization step:  $\mathbf{x}_i^{t+1/2} = \mathbf{x}_i^t - \rho \mathbf{m}_i^t$ .

    Communicate parameters  $\mathbf{x}_i^{t+1/2}$  with neighbors  $n(i)$ . Receive their parameters.

    Update the model using the gossip scheme:  $\mathbf{x}_i^{t+1} = \text{CG}^+(\mathbf{x}_i^{t+1/2}; \{\mathbf{x}_j^{t+1/2}; j \in n(i)\})$ .

---

To ensure the convergence of this algorithm, we make the following standard assumptions.

**Assumption 1.** *Objective functions regularity.*

1. (**Smoothness**) There exists  $L \geq 0$ , s.t.  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ ,  $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ .
2. (**Bounded noise**) There exists  $\sigma \geq 0$  s.t.  $\forall \mathbf{x} \in \mathbb{R}^d$ ,  $\mathbb{E}[\|\nabla \ell(\mathbf{x}, \xi) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$ .
3. (**Heterogeneity**) There exist  $\zeta \geq 0$  s.t.  $\forall \mathbf{x} \in \mathbb{R}^d$ ,  $\frac{1}{\mathcal{H}} \sum_{i \in \mathcal{H}} \|\nabla f_i(\mathbf{x}) - \nabla f_{\mathcal{H}}(\mathbf{x})\|^2 \leq \zeta^2$ .

Under these assumptions, we can prove the following corollary.

**Corollary 2.** *Let  $b$  and  $\mu_{\min}$  be such that  $2(b + 1) < \mu_{\min}$ , and let  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ . Suppose that Assumption 1 holds. Then, for all  $i \in \mathcal{H}$ , the iterates produced by Algorithm 1 on  $\mathcal{G}$  with  $\eta \leq 1/\mu_{\max}(\mathcal{G})$  and learning rate  $\rho = \mathcal{O}(1/\sqrt{T})$  (depending also on problem parameters such as  $L, \gamma$  or  $\delta$ ), verify as  $T$  increases:*

$$\sum_{t=1}^T \mathbb{E} \left[ \|\nabla f_{\mathcal{H}}(\mathbf{x}_i^t)\|^2 \right] \in \mathcal{O} \left( \frac{L\sigma}{\gamma(1-\delta)\sqrt{T}} + \frac{\zeta^2}{\gamma^2(1-\delta)^2} \right).$$

If we perform  $\tilde{\mathcal{O}}(\gamma^{-1}(1-\delta)^{-1})$  steps of  $\text{CG}^+$  between each gradient computation, we obtain:

$$\sum_{t=1}^T \mathbb{E} \left[ \|\nabla f_{\mathcal{H}}(\mathbf{x}_i^t)\|^2 \right] \in \mathcal{O} \left( \frac{L\sigma}{\sqrt{T}} \sqrt{\frac{1}{|\mathcal{H}|} + \frac{\delta}{\gamma(1-\delta)^2} + \frac{\delta\zeta^2}{\gamma(1-\delta)}} \right).$$

As shown above, the guarantees improve when performing more aggregation steps between gradients computations. Yet, the communication cost also increases significantly in that case. This corollary is obtained by combining our Theorem 2 with Theorem 1 of Farhadkhani et al. (2023), which only requires that the robust aggregation satisfies an  $(\alpha, \lambda)$ -reduction property. Our Theorem 2 ensures that  $\text{CG}^+$  satisfies it with  $\alpha = 1 - \gamma(1 - \delta)$  and  $\lambda = \gamma\delta$ . The multiple communication steps case corresponds to  $\alpha \approx 0$  and  $\lambda = 4\delta/[\gamma(1 - \delta)^2]$ . A detailed proof can be found in Appendix B.4.

## 5 DESIGNING DECENTRALIZED ATTACKS.

Similarly to robust aggregation methods, most proposed distributed attacks focus on centralized communication networks. For instance, Baruch et al. (2019) and Xie et al. (2020) propose that all Byzantine units send the same vector to the server to poison the update. However, the decentralized setting offers an additional surface for attacks: all nodes have different parameters, and Byzantine nodes can leverage this heterogeneity to disrupt learning even further. We show how to leverage this in two different ways, and then compare robust algorithms against these attacks.

### 5.1 CHANGING THE CENTER OF THE ATTACK

Farhadkhani et al. (2023) implement existing centralized attacks in a decentralized setting by making all Byzantine nodes declare the same attack vector to their neighbors, namely,  $\bar{\mathbf{x}}_{\mathcal{H}}^t + a$  where  $a$  is the attack direction. We make Byzantine nodes declare the parameter  $\mathbf{x}_i^t + a$  to node  $i$  instead. In other words, Byzantine nodes take the parameter of the honest node they attack as the reference point for the attack instead of the average of parameters. This reduces the likelihood of updates being clipped, but still pushes the overall system in the same direction.

### 5.2 TOPOLOGY-AWARE ATTACK

We now introduce the *Spectral Heterogeneity* (SP) attack, which is specifically designed to drive honest nodes away farther apart. To this end, we leverage the matrix formulation of the gossip communication, thus denoting by  $\mathbf{X}_{\mathcal{H}} = (\mathbf{x}_1, \dots, \mathbf{x}_{|\mathcal{H}|})^T \in \mathbb{R}^{|\mathcal{H}| \times d}$  the matrix of honest parameters. To design attacks on gossip-based robust aggregation mechanisms, we model communication as a perturbation of a gossip scheme, as it is the case for  $\text{CG}^+$  (cf Lemma 5), and for NNA.

$$\mathbf{X}_{\mathcal{H}}^{t+1} = (\mathbf{I}_{\mathcal{H}} - \eta \mathbf{W}_{\mathcal{H}}) \mathbf{X}_{\mathcal{H}}^t + \eta \mathbf{E}^t. \quad (8)$$

By doing so, we omit the impact of the robust aggregation rule, and only consider the error term due to Byzantine nodes, i.e, we assume that  $[\mathbf{E}^t]_i = \zeta_i^t \mathbf{a}_i^t$  for any honest node  $i$ , where  $\zeta_i^t$  is a scaling factor of the attack, and  $\mathbf{a}_i^t$  is the direction of attack on node  $i$ . We will see that this leads to powerful attacks even when taking the defense mechanism into account.

**Dissensus Attack.** To disrupt the aggregation procedure, Byzantine agents might aim at maximizing the variance of the honest parameters. A natural notion of variance in a decentralized setting is the average of pairwise differences of the neighbors parameters, which corresponds to  $\|\mathbf{X}_{\mathcal{H}}\|_{\mathbf{W}_{\mathcal{H}}}^2$ . Finding  $\mathbf{a}_i^t$  such that these pairwise differences are maximized at  $t + 1$  writes

$$\arg \max_{[\mathbf{E}^t]_i = \zeta_i^t \mathbf{a}_i^t} \|(\mathbf{I}_{\mathcal{H}} - \eta \mathbf{W}_{\mathcal{H}}) \mathbf{X}_{\mathcal{H}}^t + \eta \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 = \arg \max_{[\mathbf{E}^t]_i = \zeta_i^t \mathbf{a}_i^t} 2\eta \langle \mathbf{W}_{\mathcal{H}} \mathbf{X}_{\mathcal{H}}^t, \mathbf{E}^t \rangle + o(\eta^2).$$

Hence, maximizing the heterogeneity at time  $t + 1$  suggests to take  $\mathbf{a}_i^t = [\mathbf{W}_{\mathcal{H}}^t \mathbf{X}_{\mathcal{H}}^t]_i = \sum_{j \in n_{\mathcal{H}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t)$ . This choice of  $\mathbf{a}_i^t$  corresponds to the *Dissensus* attack proposed in He et al. (2022). However, as gossip communication is usually operated for a large number a communication rounds, maximizing only the pairwise differences at the next step is a short-sighted approach.

**Spectral Heterogeneity Attack.** Taking into account the fact that several communication rounds are performed over iterations, Byzantine nodes should aim at time  $t$  at maximizing for any  $s \geq 0$  the

pairwise differences at time  $t + s$ , i.e, finding

$$\arg \max_{[\mathbf{E}^t]_i = \zeta_i^t \mathbf{a}_i^t} 2\eta \langle \mathbf{W}_{\mathcal{H}}(\mathbf{I}_{\mathcal{H}} - \eta \mathbf{W}_{\mathcal{H}})^{2s+1} \mathbf{X}_{\mathcal{H}}^t, \mathbf{E}^t \rangle + o(\eta^2).$$

Considering the asymptotic  $s \rightarrow +\infty$  leads to approximating  $\mathbf{W}_{\mathcal{H}}(\mathbf{I}_{\mathcal{H}} - \eta \mathbf{W}_{\mathcal{H}})^{2s}$  as a projection on its eigenspace associated with the largest eigenvalue of  $\mathbf{W}_{\mathcal{H}}(\mathbf{I}_{\mathcal{H}} - \eta \mathbf{W}_{\mathcal{H}})^{2s}$ . This eigenspace corresponds to the space spanned by the eigenvector of  $\mathbf{W}_{\mathcal{H}}$  associated with the smallest non-zero eigenvalue of  $\mathbf{W}_{\mathcal{H}}$ , i.e  $\mu_{\min}$ . This eigenvector (denoted  $e_{fied}$ ) is commonly referred to as the Fiedler vector of the graph. Its coordinates essentially sort the nodes of the graph with the two farthest nodes associated with the largest and smallest value. Hence the signs of the values in the Fiedler vector are typically used to partition the graph into two components. Our **Spectral Heterogeneity** attack consists in taking  $\mathbf{a}_i^t = [e_{fied} e_{fied}^T \mathbf{X}_{\mathcal{H}}^t]_i$ , which essentially leads Byzantine nodes to cut the graph into two by pushing honest nodes in either plus or minus  $e_{fied}^T \mathbf{X}_{\mathcal{H}}^t$ .

### 5.3 EXPERIMENTAL EVALUATION

Our experimental setting is the same as Farhadkhani et al. (2023), from which we used the implementation. Namely, we train a CNN with two convolutional layers followed by two fully-connected layers, and nodes draw samples from the original dataset using a Dirichlet distribution ( $\alpha = 5$ ). The main differences are that (i) we changed the part corresponding to ClippedGossip since the clipping threshold was not properly implemented, (ii) we implemented the  $CG^+$  clipping rule, (iii) each Byzantine attack is centered on the parameter of the attacked node instead of the average of honest nodes, as described earlier, (iv) we implemented support for arbitrary sparse graphs and our spectral heterogeneity attack. The rest of the setting is detailed in Figure 1’s caption.

We observe that all robust aggregation rules are robust against attacks that do not exploit the graph topology. Clipping-based attacks initially struggle against ALIE, but eventually converge to the right value. On the other hand, MoNNA fails to learn against our Spectral Heterogeneity attack, demonstrating both the efficiency of the attack and that it has a worse breakdown point than  $CG^+$  (which obtains an optimal one), and so fails quicker when approaching the max theoretical breakdown point. While ClippedGossip performs on par with  $CG^+$  overall, we insist on the fact that we used their rule of thumb clipping rule, which is not theoretically grounded, and thus might fail against other attacks. More details (including the link to the code repository) are given in Appendix A.

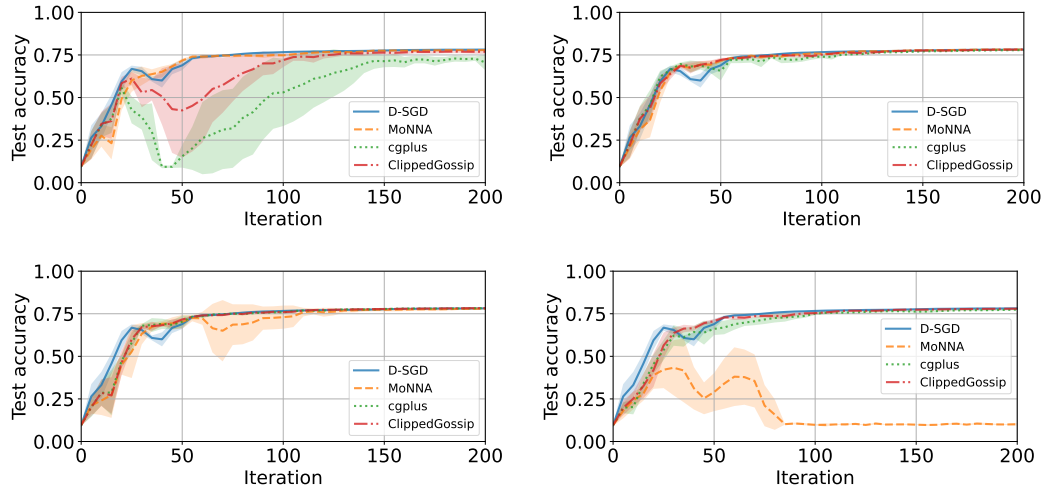


Figure 1: Test accuracies achieved by D-SGD, MoNNA,  $CG^+$  and ClippedGossip on MNIST against 4 attacks. There are  $|\mathcal{H}| = 26$  honest workers, each is neighbor to  $b = 6$  Byzantine nodes and  $\mu_2(\mathcal{G}) = 16$ . The communication graph consists of two fully connected cliques of 13 honest nodes, each honest node is connected to 8 nodes in the other clique. Byzantine nodes execute *ALIE* (row 1 left), *FOE* (row 1 right), *Dissensus* (row 2 left) and *Spectral Heterogeneity* (row 2 right). D-SGD is used as a reference and is thus not attacked by Byzantine nodes.

---

## 6 CONCLUSION

This paper revisits robust averaging over sparse communication graphs. We provide an upper bound on the optimal breakdown point, and then introduce  $CG^+$ , a midpoint between NNA and ClippedGossip, which meets this optimal breakdown (unlike the two other). Our experiments show that NNA indeed fails before the optimal breakdown point. To obtain this result, we introduced a new *Spectral Heterogeneity* attack that exploits the graph topology for sparse graphs. Now that we have precisely quantified the impact of the topology, an interesting future direction is the precise characterization of robustness when the constraint on the number of neighbors cannot be met globally, but local convergence can be obtained by considering that honest nodes with too many Byzantine neighbors are Byzantine themselves. Conversely, this opens up questions of which nodes should an attacker corrupt to maximize its influence for a specific graph, under the light of our results.

## REFERENCES

- Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. *Advances in neural information processing systems*, 31, 2018.
- Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1232–1300. PMLR, 2023.
- Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and Geovani Rizk. Robust distributed learning: tight error bounds and breakdown point under data heterogeneity. *Advances in Neural Information Processing Systems*, 36, 2024.
- Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.
- Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.
- Edwige Cyffers, Mathieu Even, Aurélien Bellet, and Laurent Massoulié. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. *Advances in Neural Information Processing Systems*, 35:15889–15902, 2022.
- Danny Dolev, Nancy A Lynch, Shlomit S Pinter, Eugene W Stark, and William E Weihl. Reaching approximate agreement in the presence of faults. *Journal of the ACM (JACM)*, 33(3):499–516, 1986.
- El-Mahdi El-Mhamdi, Rachid Guerraoui, Arsany Guirguis, Lê Nguyễn Hoang, and Sébastien Rouault. Genuinely distributed byzantine machine learning. In *Proceedings of the 39th Symposium on Principles of Distributed Computing*, pp. 355–364, 2020.
- El Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyễn Hoang, and Sébastien Rouault. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in Neural Information Processing Systems*, 34:25044–25057, 2021.
- Cheng Fang, Zhixiong Yang, and Waheed U Bajwa. Bridge: Byzantine-resilient decentralized gradient descent. *IEEE Transactions on Signal and Information Processing over Networks*, 8: 610–626, 2022.
- Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pp. 6246–6283. PMLR, 2022a.

- 
- Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pp. 6246–6283. PMLR, 2022b.
- Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. Robust collaborative learning with linear gradient overhead. In *International Conference on Machine Learning*, pp. 9761–9813. PMLR, 2023.
- Alan David Fekete. Asymptotically optimal algorithms for approximate agreement. In *Proceedings of the fifth annual ACM symposium on Principles of distributed computing*, pp. 73–87, 1986.
- Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via clippedgossip. *arXiv preprint arXiv:2202.01545*, 2022.
- Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. *arXiv preprint arXiv:2006.09365*, 2020.
- Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pp. 5311–5319. PMLR, 2021.
- Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33:18342–18352, 2020.
- Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.
- Heath J LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31(4):766–781, 2013.
- Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- Jie Peng, Weiyu Li, and Qing Ling. Byzantine-robust decentralized stochastic optimization over static and time-varying networks. *Signal Processing*, 183:108020, 2021.
- Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pp. 3027–3036. PMLR, 2017.
- Shreyas Sundaram and Christoforos N Hadjicostis. Distributed function calculation via linear iterative strategies in the presence of malicious agents. *IEEE Transactions on Automatic Control*, 56(7):1495–1508, 2010.
- Nitin H Vaidya. Iterative byzantine vector consensus in incomplete graphs. In *Distributed Computing and Networking: 15th International Conference, ICDCN 2014, Coimbatore, India, January 4-7, 2014. Proceedings 15*, pp. 14–28. Springer, 2014.
- Nitin H Vaidya, Lewis Tseng, and Guanfeng Liang. Iterative approximate byzantine consensus in arbitrary directed graphs. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing*, pp. 365–374, 2012.
- Zhaoxian Wu, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE transactions on signal processing*, 2023.
- Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pp. 261–270. PMLR, 2020.
- Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR, 2018.

---

# Appendices

## A DESCRIPTION OF THE EXPERIMENTS

Our experimental setting is built on top of the code provided by Farhadkhani et al. (2023), with the following differences:

- Attacks are designed through a linear search, but the reference point taken is the parameter of the attacked node instead of the average of all parameters. Hence, each honest node receives different messages from Byzantine nodes.
- The aggregation is performed using a gossip update in the form of Equation (CG) with  $\eta = \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$  to adapt the communication to sparse communication networks.
- Instead of considering a constant clipping threshold for ClippedGossip of He et al. (2022), as done in the experiments of (Farhadkhani et al., 2023), we use the adaptive clipping rule suggested in He et al. (2022).

The code used to run the experiments in the paper can be accessed at the following link: [https://anonymous.4open.science/r/clipped\\_gossip\\_plus\\_ICLR/](https://anonymous.4open.science/r/clipped_gossip_plus_ICLR/)

### A.1 DESCRIPTION OF THE ATTACKS

In our experiments, we consider the *Dissensus* and *Spectral Heterogeneity* attacks, and two other state-of-the-art attacks developed for the federated SGD setting: *Fall of Empire* (FOE) from Xie et al. (2020) and *A little is enough* (ALIE) from Baruch et al. (2019). Consistently with their original setting, these attacks rely on all Byzantine nodes declaring the same parameter  $\mathbf{a}^t$ . As we suggest in Section 5.3, we adapt them to the decentralized setting: Byzantine nodes declare to the honest node  $i$  having the parameter  $\mathbf{x}_i^t + \zeta_i^t \mathbf{a}_i^t$ , where  $\zeta_i^t$  is the scaling of the attack defined using a linear search and  $\mathbf{a}_i^t$  is the direction of the attack. In the case of Dissensus and Spectral Heterogeneity,  $\mathbf{a}_i^t$  is defined as described in Section 5.2. In the case of FOE and ALIE  $\mathbf{a}_i^t$  is defined as follows:

- **ALIE.** The Byzantine nodes compute the mean of the honest parameters  $\overline{\mathbf{x}}^t$  and the coordinate-wise standard deviation  $\sigma^t$ . Then they declare the parameter  $\mathbf{a}_i^t = \sigma^t$ .
- **FOE.** The Byzantine nodes declare  $\mathbf{a}_i^t = -\overline{\mathbf{x}}^t$ .

## B PROOFS

### B.1 PROOF OF THEOREM 1 - UPPER BOUND ON THE BREAKDOWN POINT.

Let  $\mu_{\min}, b$  be such that  $\mu_{\min} \leq 2b$ . Let  $H$  be an even number larger than  $2b$ .

To prove Theorem 1, we consider a communication network  $G_{H,b}$  composed of three cliques of  $m = |H|/2$  nodes  $C_1, C_2$  and  $C_3$ . Each node in  $C_i$  is additionally connected to exactly  $b$  nodes in  $C_{i+1 \bmod 3}$  and to  $b$  nodes in  $C_{i-1 \bmod 3}$ . Moreover, those connections are assumed to be *in circular order*, i.e., for any  $j \in [m]$ , node  $j$  in  $C_i$  is connected to nodes  $j, \dots, j+b \bmod m$  in  $C_{i+1 \bmod 3}$  and  $C_{i-1 \bmod 3}$ . If we assume that honest nodes can have up to  $b$  Byzantine neighbors, then any of the three cliques can be composed of Byzantine adversaries.

The proof then goes as follows: we first show by contradiction that no  $\alpha$ -robust algorithm is possible in this setting, and then that  $b = 2\mu_{\min}$  for this specific graph, so that  $G_{H,b} \in \Gamma_{\mu_{\min}, b}$ . To show the contradiction, we first assume that there exists an  $\alpha$ -robust algorithm on  $G_{H,b}$  and then:

- We show that if all nodes within one clique hold the same parameter  $\mathbf{x}^t$ , and receive this parameter from nodes of either of the two other cliques, then they cannot change their parameter.
- We consider a setting where the two honest cliques holds different parameters, and we conclude that Byzantine nodes can force all honest nodes to keep their initial parameter at all times. This shows that in the considered setting,  $\alpha < 1$  is impossible.

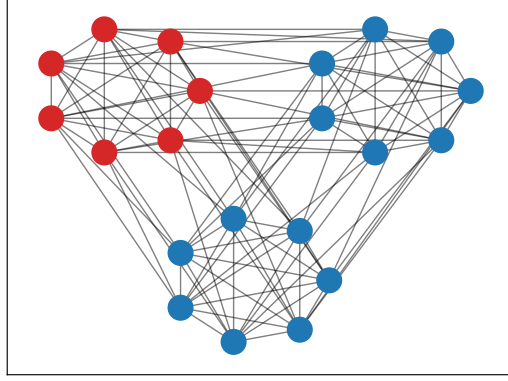


Figure 2: Topology of  $G_{H,b}$  in the lower bound: two cliques are honest, one is Byzantine.

### B.1.1 NO ALGORITHM CAN BE $\alpha$ -ROBUST ON $G_{H,b}$ .

**Lemma 1.** Consider three clique  $C_1$ ,  $C_2$  and  $C_3$  of  $m$  nodes. And say  $\mathcal{G}$  is the graph composed by these three cliques, and that any node within one of the three cliques is connected to exactly  $b \leq m$  nodes in each of the other two cliques. Assume one of these cliques is made of Byzantine nodes, then no communication algorithm is  $\alpha$ -robust on  $G_{H,b}$ .

*Proof.*

**Part I.** Assume that there exists an algorithm  $A$  that is  $\alpha$ -robust on  $G_{h,b}$ . We denote  $\hat{x}_i$  the output from node  $i$  after running  $A$ . We consider the following setting: nodes in one clique, say  $C_1$ , are honest and hold the same parameter  $x^0$ . Nodes in another clique, say  $C_2$ , declare the parameter  $x^0$  as well, while nodes in  $C_3$  declare another parameter. We show that all nodes  $i \in C_1$  must output the parameter  $\hat{x}_i = x$ .

As a matter of fact, from the point of view of nodes in  $C_1$ , it is impossible to distinguish between these two settings:

- Setting I:  $C_2$  is honest, and  $C_3$  is Byzantine.
- Setting II:  $C_2$  is Byzantine, and  $C_3$  is honest.

Consequently, nodes in  $C_1$  act in the same way in both settings. Furthermore, in Setting I, nodes of  $C_2$  are honest, and nodes in  $C_1$  and  $C_2$  have the initial same parameter; hence, the initial error is 0. Yet the  $\alpha$  criterion writes

$$\sum_{i \in \mathcal{H}} \|\hat{x}_i - \bar{x}_{\mathcal{H}}\|^2 \leq \alpha \sum_{i \in \mathcal{H}} \|x_i - \bar{x}_{\mathcal{H}}\|^2 = 0.$$

It follows that for any node  $i$  in  $C_1$ ,  $\hat{x}_i = x$ , *i.e.*, nodes do not change their parameters.

**Part II.** Consider the setting where  $C_1$  and  $C_2$  are honest, while  $C_3$  is Byzantine, and that nodes  $C_1$  hold the parameter  $x$ , while node in  $C_2$  hold the parameter  $y \neq x$ .

As Byzantine nodes can declare different values to their different neighbors, nodes in  $C_3$  can declare to nodes in  $C_1$  that they hold the value  $x$ , and to nodes in  $C_2$  that they hold the value  $y$ . Following Part I, nodes in  $C_1$  and in  $C_2$  cannot update their parameter, ( $\hat{x}_i = x_i$ ). In particular:

$$\sum_{i \in \mathcal{H}} \|\hat{x}_i - \bar{x}_{\mathcal{H}}\|^2 \leq \alpha \sum_{i \in \mathcal{H}} \|x_i - \bar{x}_{\mathcal{H}}\|^2 = \alpha \sum_{i \in \mathcal{H}} \|\hat{x}_i - \bar{x}_{\mathcal{H}}\|^2,$$

*i.e.*  $\alpha \geq 1$  since  $\sum_{i \in \mathcal{H}} \|\hat{x}_i - \bar{x}_{\mathcal{H}}\|^2 > 0$ , which means that Algorithm  $A$  is not  $\alpha$ -robust on  $G_{H,b}$ . □

### B.1.2 EIGENVALUES OF THE CONSIDERED GRAPH.

To conclude our lower bound, we only need to show that on the considered graph, the smallest non-zero eigenvalue of the honest subgraph is equal to  $2b$ . This corresponds to the following lemma.

**Lemma 2.** *Let  $\mathcal{G}$  be a graph defined as two cliques  $C_1$  and  $C_2$  of  $m$  nodes, with connections between  $C_1$  and  $C_2$  such that any node in  $C_1$  is connected to exactly  $0 \leq b \leq m$  nodes in  $C_2$ . Then the smallest non zero eigenvalue of the Laplacian matrix of  $\mathcal{G}$  is equal to  $\mu_2(\mathcal{G}_{\mathcal{H}}) = 2b$ .*

*Proof.* Let  $M$  be a circulant matrix defined as  $M = \sum_{q=0}^{b-1} J^q$ , where  $J$  denotes the permutation

$$J := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & & & 1 \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

The Laplacian matrix of  $\mathcal{G}$  can be written as:

$$W_{\mathcal{G}} = \begin{pmatrix} mI_m - \mathbf{1}_m \mathbf{1}_m^T & 0 \\ 0 & mI_m - \mathbf{1}_m \mathbf{1}_m^T \end{pmatrix} + \begin{pmatrix} bI_m & -M \\ -M^T & bI_m \end{pmatrix}$$

Hence

$$W_{\mathcal{G}} = (b+m)I_{2m} - \begin{pmatrix} -\mathbf{1}_m \mathbf{1}_m^T & 0 \\ 0 & -\mathbf{1}_m \mathbf{1}_m^T \end{pmatrix} - \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}. \quad (9)$$

This matrix decomposition allows to have the eigenvalues of the the matrix  $W_{\mathcal{G}}$ .

**Lemma 3.** *The eigenvalues of  $W_{\mathcal{G}}$  are  $\{0, 2b\} \cup \{b+m \pm |\sum_{q=0}^{b-1} \omega^{pq}|; p \in \{1, \dots, m-1\}\}$  where  $\omega := \exp(\frac{2i\pi}{m})$ .*

To prove the Lemma 3, we first need to following result.

**Lemma 4.** *If  $A$  is a symmetric matrix in  $\mathbb{R}^{2m \times 2m}$ , which can be decomposed as  $A = \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}$ , where  $M \in \mathbb{R}^{m \times m}$  is a matrix with complex eigenvalues  $\mu_0, \dots, \mu_{m-1}$ .*

*Then the eigenvalues of  $A$  are  $\{\pm|\mu_q|; q = 0 \dots m-1\}$ .*

*Proof of Lemma 4.* NB: In this specific proof, we denote by  $\bar{D}$  the matrix of complex conjugate of elements in  $D$ .

Lemma 4 follows from

$$\begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix} = \begin{pmatrix} 0 & U^*DU \\ U^T D \bar{U} & 0 \end{pmatrix} \stackrel{\bar{A}=A}{=} \begin{pmatrix} 0 & U^*DU \\ U^* \bar{D}U & 0 \end{pmatrix}.$$

Hence

$$A = \begin{pmatrix} U^* & 0 \\ 0 & U^* \end{pmatrix} \begin{pmatrix} 0 & D \\ \bar{D} & 0 \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix}.$$

A simple calculus (using that  $D$  is diagonal) yields that all eigenvalues of  $\begin{pmatrix} 0 & D \\ \bar{D} & 0 \end{pmatrix}$  are  $\{\pm|D_q|; q = 0 \dots m-1\}$ .  $\square$

*Proof of Lemma 3.* We start from the decomposition of Equation (9) :

$$W_{\mathcal{G}} = (b+m)I_{2m} - \begin{pmatrix} -\mathbf{1}_m \mathbf{1}_m^T & 0 \\ 0 & -\mathbf{1}_m \mathbf{1}_m^T \end{pmatrix} - \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}.$$



We first notice that  $(\mathbf{1}_m^T, \mathbf{1}_m^T)^T$  and  $(\mathbf{1}_m^T, -\mathbf{1}_m^T)^T$  are the only two eigenvectors of  $\begin{pmatrix} -\mathbf{1}_m \mathbf{1}_m^T & 0 \\ 0 & -\mathbf{1}_m \mathbf{1}_m^T \end{pmatrix}$  associated with non zero eigenvalues. These are eigenvectors of  $\begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}$  as well, and as such are eigenvector of  $W_G$  of eigenvalues 0 and  $2b$ . Furthermore the three matrices of Equation (9) can be diagonalized in the same orthogonal basis.

The matrix  $M$  is a circulant matrix, so it can be diagonalized in  $\mathbb{C}$ . The eigenvalues are  $\{\mu_q = \sum_{p=0}^{b-1} \omega^{pq}; q \in \{0, \dots, m-1\}\}$ , where  $\omega := \exp(\frac{2i\pi}{m})$ . The eigenvector associated with  $\mu_q$  is  $x_q = (1, \omega^q, \dots, \omega^{(m-1)q})^T$ . As such, with  $U = (x_0, \dots, x_{m-1})$  and  $D = \text{Diag}(\mu_0, \dots, \mu_{m-1})$ ,  $M$  writes:

$$M = U^* D U.$$

Considering Lemma 4, the eigenvalue of  $\begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}$  are  $\{\pm|\mu_q|; q = 0, \dots, m-1\}$ , considering that  $q = 0$  corresponds to the eigenvalues  $+b$  and  $-b$ , hence the eigenvectors  $(\mathbf{1}_m^T, \mathbf{1}_m^T)^T$  and  $(\mathbf{1}_m^T, -\mathbf{1}_m^T)^T$ , we deduce that the eigenvalues of  $W_G$  are  $\{\pm|\mu_q|; q = 0 \dots m-1\}$ .  $\square$

### End of the proof of Lemma 2.

To prove Lemma 2, considering the decomposition of Equation (9), we only have to show that  $m-b$  is always the second largest eigenvalue of the matrix

$$B := \begin{pmatrix} \mathbf{1}_m \mathbf{1}_m^T & 0 \\ 0 & \mathbf{1}_m \mathbf{1}_m^T \end{pmatrix} + \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}.$$

First, considering Lemma 3, the eigenvalues of  $B$  are  $\{m+b, m-b\} \cup \{\pm|\mu_p|; p \in \{1, \dots, m-1\}\}$  with  $\mu_p = \sum_{q=0}^{b-1} \omega^{pq}$ . As such showing that  $|\mu_p| \leq m-b$  if  $p \in \{1, \dots, m-1\}$  yields the result.

As  $\omega^{mp} = \omega^{0p} = 1$ , we have that  $\sum_{q=0}^{m-1} \omega^{pq}(1 - \omega^p) = 0$ . Hence, for  $p \in \{1, \dots, m-1\}$ , as  $\omega^p \neq 1$ ,

$$\sum_{q=0}^{m-1} \omega^{pq} = 0 \implies \mu_p = \sum_{q=0}^{b-1} \omega^{pq} = - \sum_{q=b}^{m-1} \omega^{pq}.$$

It follows from  $|\omega| = 1$  that for  $p \in \{1, \dots, m-1\}$ ,  $|\mu_p| \leq m-b$ .  $\square$

## B.2 CONVERGENCE OF CG<sup>+</sup>

We first prove Theorem 2, and then use it to derive convergence for the Byzantine-robust decentralized optimization framework. We recall that all nodes follow the following update scheme .

$$\begin{cases} \mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \eta \sum_{j \in n(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t) & \text{if } i \in \mathcal{H} \\ \mathbf{x}_i^{t+1} = * & \text{if } i \in \mathcal{B} \end{cases} \quad (10)$$

Before proving Theorem 2, we introduce the following notations:

- The matrix of honest parameters  $\mathbf{X}_{\mathcal{H}}^t := \begin{pmatrix} (\mathbf{x}_1^t)^T \\ \vdots \\ (\mathbf{x}_{|\mathcal{H}|}^t)^T \end{pmatrix} \in \mathbb{R}^{|\mathcal{H}| \times d}$ .
- The error due to clipping and Byzantine corruption:

$$\forall i \in \mathcal{H}, \quad [\mathbf{E}^t]_i := \sum_{j \in n_{\mathcal{H}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t - \text{Clip}(\mathbf{x}_i^t - \mathbf{x}_j^t; \tau_i^t)) + \sum_{j \in n_{\mathcal{B}}(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t)$$

**Lemma 5.** Equation (10) writes

$$\mathbf{X}_{\mathcal{H}}^{t+1} = (\mathbf{I}_{\mathcal{H}} - \eta \mathbf{W}_{\mathcal{H}}) \mathbf{X}_{\mathcal{H}}^t + \eta \mathbf{E}^t.$$

*Proof.* Let  $i \in \mathcal{H}$ . We decompose the update due to the gossip scheme and consider the error term coming from both clipping and the influence of Byzantine nodes.

$$\begin{aligned}
\mathbf{x}_i^{t+1} &= \mathbf{x}_i^t + \eta \sum_{j \in n(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t) \\
&= \mathbf{x}_i^t - \eta \sum_{j \in n_{\mathcal{H}}(i)} \text{Clip}(\mathbf{x}_i^t - \mathbf{x}_j^t; \tau_i^t) + \eta \sum_{j \in n_{\mathcal{B}}(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t) \\
\mathbf{x}_i^{t+1} &= \mathbf{x}_i^t - \eta \sum_{j \in n_{\mathcal{H}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t) \\
&\quad + \eta \sum_{j \in n_{\mathcal{H}}(i)} [(\mathbf{x}_i^t - \mathbf{x}_j^t) - \text{Clip}(\mathbf{x}_i^t - \mathbf{x}_j^t; \tau_i^t)] + \eta \sum_{j \in n_{\mathcal{B}}(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t)
\end{aligned}$$

Finally, the proof is concluded by remarking that  $[W_{\mathcal{H}} \mathbf{X}_{\mathcal{H}}^t]_i = \sum_{j \in n_{\mathcal{H}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t)$ .  $\square$

We begin by controlling the norm of the error term  $\|\mathbf{E}^t\|_2^2$ .

**Lemma 6** (Control of the error, case of local clipping). *The error due to clipping and Byzantine nodes is controlled by the heterogeneity as measured by the gossip matrix:*

$$\|\mathbf{E}^t\|_2^2 \leq 2(b+1) \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 = (b+1) \sum_{i \in \mathcal{H}, j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2$$

*Proof.* We recall that

$$\forall i \in \mathcal{H}, \quad [\mathbf{E}^t]_i := \sum_{j \in n_{\mathcal{H}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t - \text{Clip}(\mathbf{x}_i^t - \mathbf{x}_j^t; \tau_i^t)) + \sum_{j \in n_{\mathcal{B}}(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t)$$

By applying the triangle inequality, and by denoting  $(a)_+ = \max(a, 0)$ , we get

$$\begin{aligned}
\|\mathbf{E}^t\|^2 &= \sum_{i \in \mathcal{H}} \left\| \sum_{j \in n_{\mathcal{H}}(i)} \mathbf{x}_i^t - \mathbf{x}_j^t - \text{Clip}(\mathbf{x}_i^t - \mathbf{x}_j^t; \tau_i^t) + \sum_{j \in n_{\mathcal{B}}(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t) \right\|_2^2 \\
&\leq \sum_{i \in \mathcal{H}} \left( \sum_{j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t - \text{Clip}(\mathbf{x}_i^t - \mathbf{x}_j^t; \tau_i^t)\|_2 + \sum_{j \in n_{\mathcal{B}}(i)} \|\text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t)\|_2 \right)^2 \\
&\leq \sum_{i \in \mathcal{H}} \left( \sum_{j \in n_{\mathcal{B}}(i)} (\|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2 - \tau_i^t)_+ + b\tau_i^t \right)^2.
\end{aligned}$$

Plugging in the fact that  $\tau_i^t$  is the  $b+1$  largest value of  $\{\|\mathbf{x}_i^t - \mathbf{x}_j^t\|; j \in n(i)\}$ , we can control each term

$$\sum_{j \in n_{\mathcal{B}}(i)} (\|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2 - \tau_i^t)_+ + b\tau_i^t.$$

To do so, we apply the technical Lemma 7 below, where we denote  $a_1 \geq \dots \geq a_{|n_{\mathcal{H}}(i)|}$  the sorted values within  $\{\|\mathbf{x}_i^t - \mathbf{x}_j^t\|; j \in n_{\mathcal{H}}(i)\}$ , and  $k$  is the number of clipped honest neighbors of node  $i$ . As  $b+1 \geq b$ , at least one honest node is clipped, and  $\tau_i^t \leq a_k$ . Lemma 7 ensures that

$$\|\mathbf{E}^t\|^2 \leq \sum_{i \in \mathcal{H}} \left( \sum_{\substack{j \in n_{\mathcal{H}}(i) \\ b+1 \text{ largest}}} \|\mathbf{x}_i^t - \mathbf{x}_j^t\| \right)^2.$$

Which, using Cauchy-Schwarz inequality, yields

$$\|\mathbf{E}^t\|^2 \leq \sum_{i \in \mathcal{H}} (b+1) \sum_{j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2.$$

The result finally follows by noting that  $2\|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 = \sum_{i \in \mathcal{H}, j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2$ . Indeed, considering that  $\mathcal{G}_{\mathcal{H}}$  is an undirected graph,  $i \in n_{\mathcal{H}}(j) \iff j \in n_{\mathcal{H}}(i)$  and we have:

$$\begin{aligned} \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 &= \langle \mathbf{X}_{\mathcal{H}}, \mathbf{W}_{\mathcal{H}} \mathbf{X}_{\mathcal{H}} \rangle \\ &= \sum_{i \in \mathcal{H}} \left\langle \mathbf{x}_i^t, \sum_{j \in n_{\mathcal{H}}(i)} \mathbf{x}_i^t - \mathbf{x}_j^t \right\rangle \\ &= \sum_{i \in \mathcal{H}} \sum_{j \in n_{\mathcal{H}}(i)} \langle \mathbf{x}_i^t, \mathbf{x}_i^t - \mathbf{x}_j^t \rangle \\ &= \frac{1}{2} \sum_{i \in \mathcal{H}} \sum_{j \in n_{\mathcal{H}}(i)} \langle \mathbf{x}_i^t - \mathbf{x}_j^t, \mathbf{x}_i^t - \mathbf{x}_j^t \rangle \\ \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 &= \frac{1}{2} \sum_{i \in \mathcal{H}, j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2^2 \end{aligned}$$

□

**Lemma 7** (Technical lemma for controlling  $\mathbf{E}^t$  with  $\text{CG}^+$ ). *let  $a_1 \geq \dots \geq a_n \geq 0$  be  $n$  sorted non-negative real numbers, and  $\text{err}(k) := \sum_{i=1}^k (a_i - a_k) + ba_k$ . Recall that  $k$  denotes the number of honest neighbors clipped and we assume that  $b + 1$  neighbors are clipped. It follows that  $k \in \{1, \dots, b + 1\}$  and*

$$\max_{k=1, \dots, b+1} \left\{ \text{err}(k) = \sum_{i=1}^k (a_i - a_k) + ba_k \right\} \leq \sum_{i=1}^{b+1} a_i$$

*Proof.* The err function writes

$$\text{err}(k) = \sum_{i=1}^k a_i + (b - k)a_k.$$

- If  $b \leq k$ , then  $\text{err}(k) \leq \sum_{i=1}^k a_i$ .
- If  $b > k$ , then

$$\text{err}(k) = \sum_{i=1}^b a_i - \sum_{i=k+1}^b \underbrace{(a_i - a_k)}_{\geq 0}.$$

It follows that

$$\text{err}(k) \leq \sum_{i=1}^{\max(k, b)} a_i.$$

Hence, as  $k \leq b + 1$ , we have that

$$\text{err}(k) \leq \sum_{i=0}^{b+1} a_i$$

□

Now that we control the error term, we can conclude the proof of Theorem 2 using standard optimization arguments. Before proving this theorem, we prove the following one, from which Corollary 1 is direct.

**Theorem 4.** *Let  $b$  and  $\mu_{\min}$  be such that  $2(b + 1) \leq \mu_{\min}$ , and let  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ . Then, assuming  $\eta \leq \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$ , the output  $\mathbf{y} = \text{CG}^+(\mathbf{x})$  (obtained by one step of  $\text{CG}^+$  on  $\mathcal{G}$  from  $\mathbf{x}$ ) verifies:*

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}_{\mathcal{H}}^{t+1}\|^2 \leq (1 - \eta(\mu_{\min} - 2(b + 1))) \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^t - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \quad (11)$$

$$\|\bar{\mathbf{x}}_{\mathcal{H}}^{t+1} - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \leq \eta \frac{2(b+1)}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^t - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \quad (12)$$

*Proof. Part I: Equation (12).*

Equation (12) is a direct consequence of Lemma 6. Indeed applying  $\mathbf{P}_{1_{\mathcal{H}}} := \frac{1}{|\mathcal{H}|} \mathbf{1}_{\mathcal{H}} \mathbf{1}_{\mathcal{H}}^T$  - the orthogonal projection on the kernel of  $\mathbf{W}_{\mathcal{H}}$  - on Lemma 5 results in

$$\mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^{t+1} = \mathbf{P}_{1_{\mathcal{H}}} (\mathbf{I}_{\mathcal{H}} - \eta \mathbf{W}_{\mathcal{H}}) \mathbf{X}_{\mathcal{H}}^t + \eta \mathbf{P}_{1_{\mathcal{H}}} \mathbf{E}^t = \mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^t + \eta \mathbf{P}_{1_{\mathcal{H}}} \mathbf{E}^t.$$

Taking the norm yields

$$\|\mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^t\|^2 = \eta^2 \|\mathbf{P}_{1_{\mathcal{H}}} \mathbf{E}^t\|^2 \leq \eta^2 \|\mathbf{E}^t\|^2. \quad (13)$$

We now apply Lemma 6, and use that  $\mu_{\max}(\mathcal{G}_{\mathcal{H}})$  is the largest eigenvalue of  $\mathbf{W}_{\mathcal{H}}$ . It gives

$$\begin{aligned} \|\mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^t\|^2 &\leq \eta^2 2(b+1) \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 \\ &\leq \mu_{\max}(\mathcal{G}_{\mathcal{H}}) \eta^2 2(b+1) \|(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}}) \mathbf{X}_{\mathcal{H}}^t\|^2 \end{aligned}$$

Finally, Equation (12) derives from  $[\mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^t]_{i \in \mathcal{H}} = [\sum_{j \in \mathcal{H}} \mathbf{x}_j^t]_{i \in \mathcal{H}} = [\bar{\mathbf{x}}_{\mathcal{H}}^t]_{i \in \mathcal{H}}$  and  $\eta \mu_{\max}(\mathcal{G}_{\mathcal{H}}) \leq 1$ .

**Part II: Equation (11).**

To prove Equation (11), we consider the objective function  $\|(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}}) \mathbf{X}_{\mathcal{H}}^t\|^2$ . We denote by  $\mathbf{W}_{\mathcal{H}}^\dagger$  the Moore-Penrose pseudo inverse of  $\mathbf{W}_{\mathcal{H}}$ . We begin by applying Lemma 5.

$$\begin{aligned} \|(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}}) \mathbf{X}_{\mathcal{H}}^{t+1}\|^2 &= \|\mathbf{X}_{\mathcal{H}}^t - \eta \mathbf{W}_{\mathcal{H}} \mathbf{X}_{\mathcal{H}}^t + \eta \mathbf{E}^t\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 \\ &= \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 - 2\eta \langle \mathbf{X}_{\mathcal{H}}^t, \mathbf{W}_{\mathcal{H}} \mathbf{X}_{\mathcal{H}}^t - \mathbf{E}^t \rangle_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})} \\ &\quad + \eta^2 \|\mathbf{W}_{\mathcal{H}} \mathbf{X}_{\mathcal{H}}^t - \mathbf{E}^t\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 \\ &= \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 - 2\eta \langle \mathbf{X}_{\mathcal{H}}^t, \mathbf{X}_{\mathcal{H}}^t - \mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t \rangle_{\mathbf{W}_{\mathcal{H}}} \\ &\quad + \eta^2 \|\mathbf{X}_{\mathcal{H}}^t - \mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}^2}. \end{aligned}$$

Applying  $2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$  leads to

$$\begin{aligned} \|\mathbf{X}_{\mathcal{H}}^{t+1}\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 - \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 &= -\eta \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 - \eta \|\mathbf{X}_{\mathcal{H}}^t - \mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta \|\mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 \\ &\quad + \eta^2 \|\mathbf{X}_{\mathcal{H}}^t - \mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}^2} \\ &= -\eta \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta \|\mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}^\dagger}^2 \\ &\quad - \eta \|\mathbf{X}_{\mathcal{H}}^t - \mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta^2 \|\mathbf{X}_{\mathcal{H}}^t - \mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}^2}. \end{aligned} \quad (14)$$

We now apply that  $\mu_{\max}(\mathcal{G}_{\mathcal{H}})$  (resp.  $\mu_2(\mathcal{G}_{\mathcal{H}})$ ) is the largest (resp. smallest) non-zero eigenvalue of  $\mathbf{W}_{\mathcal{H}}$ .

$$\begin{aligned} \|\mathbf{X}_{\mathcal{H}}^{t+1}\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 - \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})}^2 &\leq -\eta \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta \frac{1}{\mu_2(\mathcal{G}_{\mathcal{H}})} \|\mathbf{E}^t\|^2 \\ &\quad - \eta (1 - \mu_{\max}(\mathcal{G}_{\mathcal{H}}) \eta) \|\mathbf{X}_{\mathcal{H}}^t - \mathbf{W}_{\mathcal{H}}^\dagger \mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}}^2. \end{aligned}$$

Eventually Lemma 6 with the assumption  $\eta \leq 1/\mu_{\max}(\mathcal{G}_{\mathcal{H}})$  yield the result

$$\begin{aligned}\|\mathbf{X}_{\mathcal{H}}^{t+1}\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 &\leq \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 - \eta \left(1 - \frac{2(b+1)}{\mu_2(\mathcal{G}_{\mathcal{H}})}\right) \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 \\ \|\mathbf{X}_{\mathcal{H}}^{t+1}\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 &\leq \left(1 - \eta\mu_2(\mathcal{G}_{\mathcal{H}}) \left(1 - \frac{2(b+1)}{\mu_2(\mathcal{G}_{\mathcal{H}})}\right)\right) \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2.\end{aligned}$$

□

To obtain Theorem 4, we note that we can actually control the one-step variation of the *MSE* using  $\alpha$  only, thus strengthening the first inequality. We rewrite the first part of Theorem 2 below for completeness.

**Corollary 3.** *Let  $b$  and  $\mu_{\min}$  be such that  $2(b+1) \leq \mu_{\min}$ , and let  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ . Then, assuming  $\eta \leq \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$ , the output  $\mathbf{y} = \text{CG}^+(\mathbf{x})$  (obtained by one step of  $\text{CG}^+$  on  $\mathcal{G}$  from  $\mathbf{x}$ ) verifies:*

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \leq (1 - \eta(\mu_{\min} - 2(b+1))) \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^t - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2$$

*Proof.* We consider Equation (13) and Equation (14), which write

$$\|\mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^t\|^2 = \eta^2 \|\mathbf{P}_{1_{\mathcal{H}}}\mathbf{E}^t\|^2.$$

$$\|\mathbf{X}_{\mathcal{H}}^{t+1}\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 - \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 \leq -\eta \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta \|\mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}^\dagger}^2.$$

It follows from the bias - variance decomposition of the MSE

$$\|\mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^t\|^2 = \|(\mathbf{I}_{\mathcal{H}} - \mathbf{P}_{1_{\mathcal{H}}})\mathbf{X}_{\mathcal{H}}^{t+1}\|^2 + \|\mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^t\|^2$$

that

$$\begin{aligned}\|\mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^t\|^2 - \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 &\leq -\eta \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta \|\mathbf{E}^t\|_{\mathbf{W}_{\mathcal{H}}^\dagger}^2 + \eta^2 \|\mathbf{P}_{1_{\mathcal{H}}}\mathbf{E}^t\|^2 \\ &\leq -\eta \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta \frac{1}{\mu_2(\mathcal{G}_{\mathcal{H}})} \|\mathbf{E}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 + \eta^2 \|\mathbf{P}_{1_{\mathcal{H}}}\mathbf{E}^t\|^2\end{aligned}$$

As  $\eta \leq \frac{1}{\mu_{\max}(\mathcal{G}_{\mathcal{H}})} \leq \frac{1}{\mu_2(\mathcal{G}_{\mathcal{H}})}$ , we eventually get

$$\begin{aligned}\|\mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^t\|^2 &\leq \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 - \eta \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 + \eta \frac{1}{\mu_2(\mathcal{G}_{\mathcal{H}})} \|\mathbf{E}^t\|^2 \\ &\leq \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2 - \eta \left(1 - \frac{2(b+1)}{\mu_2(\mathcal{G}_{\mathcal{H}})}\right) \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 \\ &\leq \left(1 - \eta\mu_2(\mathcal{G}_{\mathcal{H}}) \left(1 - \frac{2(b+1)}{\mu_2(\mathcal{G}_{\mathcal{H}})}\right)\right) \|\mathbf{X}_{\mathcal{H}}^t\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2\end{aligned}$$

□

### B.3 CONSEQUENCES

A direct consequence of the above results is Corollary 1, as we show below.

*Proof.* Using the  $(\alpha, \lambda)$  reduction notations, we have:

$$\begin{cases} \alpha = 1 - \gamma(1 - \delta) \\ \lambda = \gamma\delta \end{cases}$$

We denote here the drift increment  $d_{t+1} = \|\mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^{t+1} - \mathbf{P}_{1_{\mathcal{H}}}\mathbf{X}_{\mathcal{H}}^t\|$  and the variance at time  $t$  as  $\sigma_t^2 = \|\mathbf{X}_{\mathcal{H}}^{t+1}\|_{(\mathbf{I}_{\mathcal{H}}-\mathbf{P}_{1_{\mathcal{H}}})}^2$ .

Corollary 3 ensures that

$$\sigma_{t+1}^2 + d_{t+1}^2 \leq \alpha\sigma_t^2.$$

Hence, we have  $\sigma_{t+1}^2 + d_{t+1}^2 \leq \alpha \sigma_t^2$ , and so  $\sigma_{t+1}^2 \leq \alpha \sigma_t^2$ , which implies that  $\sigma_t \leq \alpha^{t/2} \sigma_0$ . This proves the first part of the result. Using this, we write that Theorem 4 ensures that

$$d_{t+1} \leq \sqrt{\lambda} \sigma_t \leq \sqrt{\lambda} \beta^t \sigma_0,$$

leading to:

$$\sum_{t=1}^T d_t \leq \sqrt{\lambda} \sum_{t=0}^{T-1} \sigma_t \leq \sqrt{\lambda} \sum_{t=0}^{T-1} \alpha^{t/2} \sigma_0 \leq \frac{\sqrt{\lambda}(1 - \alpha^{T/2})}{1 - \alpha^{1/2}} \sigma_0,$$

which proves the second part. The last inequality is obtained by writing.

$$\|\mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^T - \mathbf{P}_{1_{\mathcal{H}}} \mathbf{X}_{\mathcal{H}}^0\| \leq \sum_{t=1}^T d_t \leq \frac{\sqrt{\lambda}}{1 - \sqrt{\alpha}} \sigma_0$$

Then, we use that  $0 \leq \frac{1}{1 - \sqrt{1-x}} \leq \frac{2}{x}$  for  $x \geq 0$ , with  $x = \gamma(1 - \delta)$ .

□

### B.3.1 CASE OF NNA ON SPARSE GRAPHS.

The previous CG<sup>+</sup> analysis can actually be performed exactly in the same way for a sparse version of nearest neighbors averaging. Consider that, instead of clipping the  $b + 1$  farthest neighbors, each honest node  $i$  removes the  $b$  farthest neighbors. Then all results are the same with a factor  $8b$  instead of a factor  $2(b + 1)$ . And we have the following result.

**Corollary 4.** *Let  $b$  and  $\mu_{\min}$  be such that  $8b \leq \mu_{\min}$ , and let  $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$ . Then, assuming  $\eta \leq \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$ , the output  $\mathbf{y} = \text{NNA}(\mathbf{x})$  (obtained by one step of NNA on  $\mathcal{G}$  from  $\mathbf{x}$ ) verifies:*

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \leq (1 - \eta(\mu_{\min} - 8b)) \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^t - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \quad (15)$$

$$\|\bar{\mathbf{x}}_{\mathcal{H}}^{t+1} - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \leq \eta \frac{8b}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\mathbf{x}_i^t - \bar{\mathbf{x}}_{\mathcal{H}}^t\|^2 \quad (16)$$

To prove Corollary 4, we only need to change Lemma 6 to adapt if for controlling the error due to NNA. The proof hinges on the following lemma, which we state and prove first, then we will prove the equivalent of Lemma 6 in the case of NNA.

**Lemma 8** (NNA: Technical lemma for controlling  $\mathbf{E}^t$ ). *Let  $a_1 \geq \dots \geq a_n \geq 0$ , and  $\text{err}(k) := \sum_{i=1}^k a_i + k a_k$ , where  $k \leq b \leq n$ . Then*

$$\text{err}(k) \leq 2 \sum_{i=1}^b a_i.$$

*Proof.* We write

$$\text{err}(k) = 2 \sum_{i=1}^k a_i - \sum_{i=1}^k (a_i - a_k) \leq 2 \sum_{i=1}^b a_i$$

Where we used that  $a_i - a_k \geq 0$  for  $i \leq k$ . □

**Lemma 9** (NNA: Control of the error). *The error due to removing honest nodes and due to Byzantine nodes is controlled by the heterogeneity as measured by the gossip matrix.*

$$\|\mathbf{E}^t\|_2^2 \leq 8b \|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 = 4b \sum_{i \in \mathcal{H}, j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2$$

*Proof.* In this setting the error term writes

$$\forall i \in \mathcal{H}, \quad [\mathbf{E}^t]_i := \sum_{j \in n_{\mathcal{H}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t - (\mathbf{x}_i^t - \mathbf{x}_j^t) \mathbf{1}_{j \text{ not removed}}) + \sum_{j \in n_{\mathcal{B}}(i)} \text{Clip}(\mathbf{x}_j^t - \mathbf{x}_i^t; \tau_i^t)$$

Applying the triangle inequality, we get

$$\begin{aligned} \|\mathbf{E}^t\|^2 &= \sum_{i \in \mathcal{H}} \left\| \sum_{j \in n_{\mathcal{H}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t) \mathbf{1}_{j \text{ removed}} + \sum_{j \in n_{\mathcal{B}}(i)} (\mathbf{x}_i^t - \mathbf{x}_j^t) \mathbf{1}_{j \text{ not removed}} \right\|_2^2 \\ &\leq \sum_{i \in \mathcal{H}} \left( \sum_{j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\| \mathbf{1}_{j \text{ removed}} + \sum_{j \in n_{\mathcal{B}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2 \mathbf{1}_{j \text{ not removed}} \right)^2 \end{aligned}$$

By considering that node  $i \in \mathcal{H}$  removes the  $b + 1$  largest value within  $\{\|\mathbf{x}_i^t - \mathbf{x}_j^t\|; j \in n(i)\}$ , we can consider for any choice of the Byzantine nodes, that  $k$  honest neighbors are removed, and  $b - k$  Byzantine neighbors are removed. As such, we can use Lemma 9, where we denote  $a_1 \geq \dots \geq a_{|n_{\mathcal{H}}(i)|}$  the sorted values within  $\{\|\mathbf{x}_i^t - \mathbf{x}_j^t\|; j \in n_{\mathcal{H}}(i)\}$ .

Lemma 7 ensures that

$$\|\mathbf{E}^t\|^2 \leq \sum_{i \in \mathcal{H}} \left( 2 \sum_{\substack{j \in n_{\mathcal{H}}(i) \\ b \text{ largest}}} \|\mathbf{x}_i^t - \mathbf{x}_j^t\| \right)^2.$$

Which, using Cauchy-Schwarz inequality, yields

$$\|\mathbf{E}^t\|^2 \leq \sum_{i \in \mathcal{H}} 4b \sum_{j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2.$$

The final result derives from the fact that  $2\|\mathbf{X}_{\mathcal{H}}^t\|_{\mathbf{W}_{\mathcal{H}}}^2 = \sum_{i \in \mathcal{H}, j \in n_{\mathcal{H}}(i)} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2$ .  $\square$

#### B.4 PROOFS FOR $D - \text{SGD}$

*Proof of Corollary 2.* This proof hinges on the fact that the proof of Farhadkhani et al. (2023, Theorem 1) does not actually require that communication is performed using NNA, but simply that the aggregation procedure respects  $(\alpha, \lambda)$ -reduction, which they prove in their Lemma 2. Then, all subsequent results invoke this Lemma instead of the specific aggregation procedure.  $\text{CG}^+$  also satisfies  $(\alpha, \lambda)$ -reduction, as we prove in Theorem 2. We can then use the bounds on the errors out of the box.

Then, as  $T$  grows, and ignoring constant factors, only the first and last terms in their Theorem 3 remain, leading to:

$$\sum_{t=1}^T \mathbb{E} \left[ \|\nabla f_{\mathcal{H}}(\mathbf{x}_i^t)\|^2 \right] = \mathcal{O} \left( \frac{L\sigma}{\sqrt{T}} (1 + C) + \zeta^2 C \right), \quad (17)$$

where  $C = c_1 + \lambda + \lambda c_1$ , with  $c_1 = \alpha(1 + \alpha)/(1 - \alpha)^2$ . Note that we give  $\mathcal{O}()$  versions of the Theorems for simplicity, but Farhadkhani et al. (2023, Theorem 1) allows to derive precise upper bounds for any  $T \geq 1$ .

**One-step derivations.** The one-step result is obtained by taking the values of  $\alpha = 1 - \gamma(1 - \delta)$  and  $\lambda = \gamma\delta$ , and considering  $\gamma < 1$  (otherwise, the guarantees are essentially the same as in Farhadkhani et al. (2023)). More specifically:

$$c_1 = \frac{(1 - \gamma(1 - \delta))(2 - \gamma(1 - \delta))}{\gamma^2(1 - \delta)^2} = \mathcal{O} \left( \frac{1}{\gamma^2(1 - \delta)^2} \right). \quad (18)$$

Meanwhile,  $\lambda = \gamma\delta \leq 1$ , so that  $C = \mathcal{O}(c_1)$ , leading to the result.

**Multi-step derivations.** In the previous case, we see that  $C$  is dominated by the  $c_1$  term since  $c_1 \gg \lambda$ . In particular, the guarantees would increase if we were able to trade-off some  $\alpha$  for some  $\lambda$ , which is possible by using multiple communications steps. This is what we do, and take enough steps that  $c_1 \ll \lambda$  (i.e.,  $\alpha \approx 0$ ), so that  $C \approx \lambda$ . Following Corollary 1, this requires  $\tilde{\mathcal{O}}(\gamma^{-1}(1 - \delta)^{-1})$  steps, where logarithmic factors are hidden in the  $\tilde{\mathcal{O}}$  notation. We then plug the multi-step  $\lambda$  value from Corollary 1 to obtain the result.  $\square$