Unified Breakdown Analysis for Byzantine Robust Gossip

Renaud Gaucher [1 2] **Aymeric Dieuleveut** [1] **Hadrien Hendrikx** [2]

Abstract

In decentralized machine learning, different devices communicate in a peer-to-peer manner to collaboratively learn from each other's data. Such approaches are vulnerable to misbehaving (or Byzantine) devices. We introduce F–RG, a general framework for building robust decentralized algorithms with guarantees arising from robust-sum-like aggregation rules F. We then investigate the notion of *breakdown point*, and show an upper bound on the number of adversaries that decentralized algorithms can tolerate. We introduce a practical robust aggregation rule, coined $\mathrm{CS}_{\mathrm{ours}}$, such that $\mathrm{CS}_{\mathrm{ours}}$–RG has a near-optimal breakdown. Other choices of aggregation rules lead to existing algorithms such as ClippedGossip or NNA. We give experimental evidence to validate the effectiveness of $\mathrm{CS}_{\mathrm{ours}}$–RG and highlight the gap with NNA, in particular against a novel attack tailored to decentralized communications.

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. We investigate the decentralized paradigm of distributed learning, in which nodes communicate in a peer-to-peer manner within a communication network, in opposition to distributed architectures relying on a central server that coordinates all units.

Distributing the training 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 these Byzantine attacks (Blanchard et al., 2017). This has led to significant efforts in the development of robust distributed learning algorithms. From the first works on 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., 2022) and mixing strategies to handle heterogeneous loss functions (Karimireddy et al., 2023; 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). Recently, (Allouah et al., 2024a) have used pre-aggregation adaptive clipping to improve the robustness. 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., 2023; Allouah et al., 2024b). Yet, all these works rely on a trusted central server to coordinate the training.

In contrast, the decentralized case has been less explored. Especially when the communication network, abstracted as a graph where vertices represent computing units that communicate through edges, is not fully connected (a.k.a. sparse). For instance, understanding *how many Byzantine nodes can be tolerated over an arbitrary network before a communication protocol fails* is an open question (that we address in this paper). Indeed, the network is often assumed to be fully connected (El-Mhamdi et al., 2021; Farhadkhani et al., 2023). If the work of (Fang et al., 2022) addresses the case of sparse networks, they only consider homogeneous losses. Closer to our setting, He et al. (2023); Wu et al. (2023) tackle Byzantine optimization on sparse networks with heterogeneous losses, however, they only achieve suboptimal robustness, as their guarantees vanish with either the number of nodes or the connectivity of the network. Moreover, the communication scheme proposed in He et al. (2023) relies on inaccessible information. Similarly, there is a lack of attacks designed to challenge decentralized optimization. Up to our knowledge, only He et al. (2023) propose one, named *Dissensus*. Still, a few previous works (Wu et al., 2023; Farhadkhani et al., 2023) have investigated generic criteria to go from communication schemes to robust

[1]Centre de mathématiques appliquées, École polytechnique, Institut Polytechnique de Paris, Palaiseau France [2]Centre Inria de l'Univ. Grenoble Alpes, CNRS, LJK, Grenoble, France. Correspondence to: Renaud Gaucher <renaud.gaucher@polytechnique.edu>.

distributed SGD frameworks.

Our work proposes a generic method to design algorithms that 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. Our solution relies on robust adaptations of the *gossip* communication, a popular scheme for decentralized communication. We then tackle general (smooth non-convex) optimization problems through a reduction. Our contributions are the following:

1 - Unifying algorithmic framework. We develop a generic method (the RG method) to construct and analyze robust communication algorithms. It is based on the decentralized application of robust aggregation rules. This RG method recovers NNA (Farhadkhani et al., 2022) and ClippedGossip (He et al., 2023) in specific cases. We use the RG method to build CS_{ours}–RG, a novel communication algorithm that is both practical and adapted to a sparse communication network.

2 - Tight theoretical guarantees. We show that RG provides robust convergence guarantees as soon as the underlying aggregation rule verifies (b, ρ)–robustness, a new robustness criterion that we introduce, and the weight of Byzantine nodes is not too large (with respect to the *algebraic connectivity*, a spectral property of the communication graph). We also show the converse result, that is, no robustness guarantees can be obtained if the weight of Byzantine nodes exceeds this threshold. Our bounds match each other for specific aggregation rules. These results generalize the standard fully-connected breakdown point of $1/3$ of Byzantines nodes to arbitrary sparse networks.

Besides these core contributions, we introduce a theoretically grounded attack, called *Spectral Heterogeneity*, specifically designed to challenge decentralized algorithms by leveraging spectral properties of the communication graph.

The remainder of the paper is organized as follows. Section 2 formalizes the problem of Byzantine-robust decentralized optimization. Section 3 presents the robust gossip framework as well as the main convergence guarantees. Then, Section 4 instanciates the general framework with several rules (and the associated robustness guarantees), as well as gives guarantees for a D-SGD algorithm based on RG. Finally, Section 5 presents the *Spectral Heterogeneity* attack as well as an experimental evaluation of several robust aggregation schemes against various attacks.

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 $\boldsymbol{x}_i \in \mathbb{R}^d$, a local loss function $f_i : \mathbb{R}^d \to \mathbb{R}$, and can communicate with its neighbors in the graph $\mathcal{G}$. 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

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\min} \left\{ f_{\mathcal{H}}(\boldsymbol{x}) := \frac{1}{|\mathcal{H}|} \sum_{i\in\mathcal{H}} f_i(\boldsymbol{x}) \right\}. \tag{1}$$

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 performing local (node-wise) inexact averaging steps 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 of updating the parameter of a node i through a linear combination of the parameters of its neighbors, with updates of the form $\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t - \eta \sum_{j=1}^m w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t)$, where w_{ij} is a weight associated with the edge (ij) of the graph and $\eta \geq 0$ will denote a communication step-size. By denoting $\boldsymbol{W}$ the Laplacian matrix associated with the weighted graph $\mathcal{G}$, i.e $\boldsymbol{W}_{ij} = -w_{ij}$ if $i \neq j$ and $\boldsymbol{W}_{ii} = \sum_{j\in n(i)} w_{ij}$, and denoting the matrix of parameters as $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_m)^T$, then the gossip update conveniently writes:

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta \boldsymbol{W}\boldsymbol{X}^t. \tag{2}$$

The Laplacian matrix (a.k.a. gossip matrix), is symmetric non-negative. When the graph is connected, its kernel is restricted to the line of the constant vectors, i.e. $\text{span}(1, \dots, 1)^T$. In the following, we will always consider that the graph $\mathcal{G}$ is weighted, so that a unique Laplacian matrix is associated with each graph $\mathcal{G}$. Moreover, 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 $\boldsymbol{W}_{\mathcal{H}}$ 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, Equation (2) with step-size $\eta \leq \mu_{\max}(\mathcal{G})^{-1}$ leads to a linear convergence of the nodes parameter values towards the average of the initial parameters: $\|\boldsymbol{X}^t - \overline{\boldsymbol{X}}^0\|^2 \leq$

$(1-\eta\mu_2(\mathcal{G}))^t\|\boldsymbol{X}^0-\overline{\boldsymbol{X}}^0\|^2$, for $\overline{\boldsymbol{X}}^0$ the matrix with columns $m^{-1}\sum_{j=1}^m \boldsymbol{x}_j^0$.

Robustness Issue. Gossip communication relies on updating the 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 decentralized optimization.

In this section, we describe the threat model and the robustness criterion that we consider.

Threat model. We consider Byzantine nodes to be omniscient adversaries, able to collude and send distinct values to each of their neighbors. We measure their influence by considering the weight of Byzantine nodes in the neighborhood of each honest node, $(b(i) := \sum_{j\in n_{\mathcal{B}}(i)} w_{ij})_{i\in\mathcal{H}}$ (as in LeBlanc et al. (2013); He et al. (2023)), instead of the total number of Byzantine nodes $|\mathcal{B}|$ as done for the centralized or fully-connected setting. Similarly, we denote by $h(i) = \sum_{j\in n_{\mathcal{H}}(i)} w_{ij}$ the weights of the honest nodes adjacent to the node i.

In the case of an arbitrary communication network, the number of honest nodes does not provide enough information by itself, as the results depend on *how* the nodes are linked, i.e., the topology of the honest subgraph. Therefore, in the case of sparse topologies, it is necessary to consider a property of the graph related to its structure rather than relying solely on the number of honest nodes. We will show that spectral properties of the Laplacian of honest subgraph are relevant quantities for robustness analyses and introduce the following class of graphs.

Definition 2.1. For any $\mu_{\min} \geq 0$ and $b \geq 0$, we define the class of weighted graphs

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

In other words, we introduce a subset of all possible graphs, partitioning in terms of (i) the second smallest eigenvalue of the honest subgraph, (a.k.a. the *algebraic connectivity*), that is restricted to be larger than a minimal value $\mu_{\min}$, and (ii) the maximal weight of Byzantines in the neighborhood of an honest node, which is restricted to be smaller than b.

Note that if all edges are equally weighted ($w_{ij} = \omega$), then $b(i) = |n_{\mathcal{B}}(i)|\cdot\omega$, and (ii) boils down to upper bounding the number of Byzantines in the neighborhood of honest nodes by $f := b\cdot\omega^{-1}$. Hence, Definition 2.1 is an extension of the standard *"there are at most f byzantine nodes among the $|\mathcal{B}|+|\mathcal{H}|$ nodes"* to the setting of arbitrary connected graphs. We point out that this class of graphs depends on the spectral properties *of the honest subgraph*, meaning that for a given communication network, these properties change depending on the location of Byzantine failures, and the associated graph can either fall within $\Gamma_{\mu_{\min},b}$ if Byzantines are "well spread", or not, e.g. if they are adversarially chosen.

Approximate Average Consensus. The average consensus problem consists in finding the average of vectors locally held by honest nodes $(\boldsymbol{x}_i)_{i\in\mathcal{H}}$. It is a specific case of Equation (1) obtained by considering $f_i(\boldsymbol{x}) = \|\boldsymbol{x}-\boldsymbol{x}_i\|^2$. Due to adversarial attacks only an *approximate* estimate of the average of honest nodes vector $\overline{\boldsymbol{x}}_{\mathcal{H}} := |\mathcal{H}|^{-1}\sum_{i\in\mathcal{H}}\boldsymbol{x}_i$ can be reached (Karimireddy et al., 2023). We now introduce the notion of r–robustness, which measures the performance of a robust communication algorithm.

Definition 2.2 (r-robustness on $\mathcal{G}$.)**.** Let $r < 1$. A communication algorithm Alg is r-robust on a graph $\mathcal{G}$ if from any initial local parameters $(\boldsymbol{x}_i)_{i\in\mathcal{H}} \in (\mathbb{R}^d)^{\mathcal{H}}$, it allows any honest node i to compute a vector $\boldsymbol{x}_i^+$ such that

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\boldsymbol{x}_i^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 \leq r\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2.$$

Imposing $r < 1$ means the honest nodes' parameters have to be strictly closer (on average) to the initial mean after the communication than before. It thus requires that the reduction of the *variance* of nodes parameters $\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}) - \mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^+)$ is larger than the *bias* $\|\overline{\boldsymbol{x}}_{\mathcal{H}}^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2$ introduced by the Byzantines, with $\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}) := |\mathcal{H}|^{-1}\sum_{i\in\mathcal{H}}\|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2$. Remark that the r-robustness of an algorithm on a graph $\mathcal{G}$ states that a *single use* of the algorithm strictly reduces the average quadratic error. However, it does not mean that multiple uses would result in a geometric decrease; indeed, we cannot simply use induction as $\overline{\boldsymbol{x}}_{\mathcal{H}}^+ \neq \overline{\boldsymbol{x}}_{\mathcal{H}}$.

3. The Robust Gossip Framework

We now introduce a generic framework, which relies on two key building blocks: (i) a generic update form, depending on an aggregation function $\mathrm{F} : (\mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}})^{\mathrm{n}} \to \mathbb{R}^{\mathrm{d}}$, and (ii) a set of those aggregation functions F, referred to as *robust summation* functions. We then show that the combination of these two blocks, *i.e.*, the generic update used with a robust summation function F, leads to a robust decentralized algorithm, hence the name: Robust Gossip.

We finally show that this framework leads to (near-)optimal robustness guarantees for well-chosen F.

3.1. The Robust Gossip Method: adding robust differences instead of averaging robustly.

We call *aggregation rule* a function $\mathrm{F} : (\mathbb{R}_+ \times \mathbb{R}^{\mathrm{d}})^{\mathrm{n}} \to \mathbb{R}^{\mathrm{d}}$, meant to aggregate a set of vectors $(z_i)_{i\in[n]} \in (\mathbb{R}^d)^n$ with weights $(w_i)_{i\in[n]} \in \mathbb{R}_+^n$. For $\eta > 0$ a communication step-size, the associated robust gossip algorithm (coined F–RG)

consists, for any honest node $i \in \mathcal{H}$ of the communication network $\mathcal{G}$, in updating its parameter $\boldsymbol{x}_i$ using:

$$\boldsymbol{x}_i^+ := \boldsymbol{x}_i - \eta F\left((w_{ij}, \boldsymbol{x}_i - \boldsymbol{x}_j)_{j\in n(i)}\right). \qquad \text{(F–RG)}$$

Crucially, each node applies the (robust) aggregation rule F to $(\omega_j, \boldsymbol{z}_j)_{j\in n(i)} := (w_{ij}, \boldsymbol{x}_i - \boldsymbol{x}_j)_{j\in n(i)}$, i.e. to *the differences of its parameter with those of its neighbors*, and uses this estimate to update its parameter. Thus, Equation (F–RG) recovers the standard gossip update from Equation (2) if F is the weighted sum operator (which is unfortunately not robust). Hence, we will look for aggregation functions that are robust versions of the weighted sum.

In contrast, directly averaging *the parameters* of the neighbors, even with a robust aggregation rule, would highly suffer from heterogeneity. Indeed, this leads to a *biased estimate* of the mean of initial parameters for sparse communication graphs. Extra assumptions, such as homogeneity of the local objectives, are thus needed to alleviate this problem (Fang et al., 2022).

Meanwhile, the RG method uses intrinsically decentralized updates, allowing for tight convergence guarantees in sparse communication graphs with heterogeneous local objectives. The strength of the RG framework is to turn any robust summation into a robust gossip algorithm. As such, one can focus on the design of robust aggregation functions without worrying about the decentralized aspect.

3.2. Robust Summation Functions.

We now specify the conditions that aggregation rules must meet to be considered suitable robust summation functions.

Definition 3.1 ((b,ρ)–robust summation)**.** Let $b, \rho \geq 0$. An aggregation rule $F : (\mathbb{R}_+ \times \mathbb{R}^d)^n \to \mathbb{R}^d$ is a (b,ρ)–*robust summation* if, for any vectors $(\boldsymbol{z}_i)_{i\in[n]} \in (\mathbb{R}^d)^n$, any weights $(\omega_i)_{i\in[n]} \in \mathbb{R}_+^n$ and any $S \subset [n]$ such that $\sum_{i\in\overline{S}} \omega_i \leq b$ (where $\overline{S} := [n]\backslash S$),

$$\left\| F\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big) - \sum_{i\in S} \omega_i \boldsymbol{z}_i \right\|^2 \leq \rho b \sum_{i\in S} \omega_i \|\boldsymbol{z}_i\|^2.$$

In (F–RG), S is the set of honest neighbors $n_{\mathcal{H}}(i)$, while $\overline{S}$ is the set of Byzantine neighbors $n_{\mathcal{B}}(i)$. We will exhibit several (b,ρ)–robust summation rules, including a practical rule for $\rho = 2$, in Section 4.

Remark 3.2. The latter definition differs from (f,κ)–*robustness* (Allouah et al., 2023) since (i) we perform a weighted sum instead of a plain average, and (ii) we upper bound the error using the *second moment* of vectors within S *instead of their variance*. Therefore, if F is (f,κ)–robust, then F is a (b,ρ)-robust summation for constant weights $\omega_i = 1/(n-f)$ with $b = f/(n-f)$ and $\rho = \kappa/b$.

3.3. Convergence of RG under (b,ρ)-robustness.

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 write the main result of this paper, which shows how (b,ρ)–robustness enables to tightly quantify how much a single step of F–RG reduces the variance, and how much bias is injected. Recall that $\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}) = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2$ is the variance of honest nodes.

Theorem 3.3. *Let F be a (b,ρ)–robust summation, b and $\mu_{\min}$ be s.t. $2\rho b \leq \mu_{\min}$, and $\mathcal{G} \in \Gamma_{\mu_{\min},b}$. Then, assuming $\eta \leq \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$, the output $(\boldsymbol{x}_i^+)_{i\in\mathcal{H}}$ of* F–RG *verifies:*

$$\begin{cases} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 \leq (1 - \eta\,(\mu_{\min} - 2\rho b))\,\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}), \\ \|\overline{\boldsymbol{x}}_{\mathcal{H}}^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 \leq 2\rho b\,\eta\,\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}). \end{cases}$$

Thus, F–RG *is $(1-\eta\,(\mu_{\min} - 2\rho b))$–robust for $\mathcal{G} \in \Gamma_{\mu_{\min},b}$.*

While the bound on η depends on the honest subgraph, as $\mu_{\max}(\mathcal{G}_{\mathcal{H}}) \leq \mu_{\max}(\mathcal{G})$, η can be set conservatively by evaluating $\mu_{\max}$ on the whole graph. Note that while r–robustness is guaranteed for the whole class $\Gamma_{\mu_{\min},b}$, the value of r will depend on the actual graph within the class.

Chaining aggregation steps. When low variance levels are required, it is necessary to perform several robust gossip 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 multiple aggregation steps are performed. We provide bounds for t steps of RG in the following Corollary.

Corollary 3.4. *Let F be a (b,ρ)–robust summation, let b and $\mu_{\min}$ be such that $2\rho b \leq \mu_{\min}$, and let $\mathcal{G} \in \Gamma_{\mu_{\min},b}$. We denote $\delta = \frac{2\rho b}{\mu_{\min}}$ and $\gamma = \mu_{\min}/\mu_{\max}(\mathcal{G}_{\mathcal{H}})$. Then, for $(\boldsymbol{x}_i^t)_{i\in\mathcal{G},\, t\geq 0}$ obtained from any $(\boldsymbol{x}_i^0)_{i\in\mathcal{G}}$ through $(\boldsymbol{x}_i^{t+1})_{i\in\mathcal{G}} = \text{F–RG}\big((\boldsymbol{x}^t)_{i\in\mathcal{G}}\big)$, with $\eta = \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$,*

$$\begin{cases} \mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^t) \leq (1 - \gamma(1-\delta))^t\,\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^0), \\ \|\overline{\boldsymbol{x}}_{\mathcal{H}}^t - \overline{\boldsymbol{x}}_{\mathcal{H}}^0\| \leq \frac{\sqrt{\gamma\delta}\left(1 - [1-\gamma(1-\delta)]^{t/2}\right)}{1 - \sqrt{1-\gamma(1-\delta)}} \sqrt{\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^0)}. \end{cases}$$

Consensus is thus reached, as $\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^t) \to_{t\to\infty} 0$, and

$$\|\overline{\boldsymbol{x}}_{\mathcal{H}}^t - \overline{\boldsymbol{x}}_{\mathcal{H}}^0\|^2 \leq \frac{4\delta}{\gamma(1-\delta)^2}\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^0). \qquad (3)$$

While the total L2 error (bias plus variance) is guaranteed to decrease after a single step by Theorem 3.3, it may increase if several F–RG steps are performed because of bias accumulation. This happens when the factor multiplying the variance in Equation (3) is larger than 1, which essentially

means $\gamma \leq \delta$. Despite this bias, the output of the resulting robust aggregation procedure is (arbitrarily) close to consensus, which can be desirable. Proofs of Theorem 3.3 and Corollary 3.4 are respectively given in Appendices C.1 and C.2.

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

3.4. Fundamental limits for decentralized communication schemes

We now provide an *upper bound* on the weight 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 3.5. *Let $\mu_{\min} \geq 0$, $b \geq 0$ be such that $\mu_{\min} \leq 2b$. Then for any $h \geq 0$ and any algorithm* Alg*, there exists a graph $\mathcal{G} \in \Gamma_{\mu_{\min},b}$ in which all honest nodes have a weight of honest neighbors $h(i)$ larger than h, and such that for any $r < 1$,* Alg *is not r–robust on $\mathcal{G}$.*

We refer the reader to Appendix D for the proof details. It follows from Theorem 3.5 that when a theoretical guarantee quantifies the robustness of an algorithm on a graph through $\mu_{\min}$, we must have $2b < \mu_{\min}$. Interestingly, this upper bound on the breakdown point is *independent of the total weight of honest neighbors*.

For a fully-connected graph with uniform weights ω, we have $\mu_2(\mathcal{G}_{\mathcal{H}}) = \omega|\mathcal{H}|$ and $b(i) = \omega|\mathcal{B}|$. Then, the previous condition boils down to $|\mathcal{H}| > 2|\mathcal{B}|$, i.e. there is less than $1/3$ of Byzantine nodes, which recovers the known necessary robustness condition in distributed system (Lamport et al., 1982; Vaidya et al., 2012; El-Mhamdi et al., 2021).

Near-optimal breakdown point. Theorem 3.5 states that uniformly ensuring r–robustness on $\Gamma_{\mu_{\min},b}$ is impossible as soon as $2b \geq \mu_{\min}$, and we know that update (F–RG) is robust as soon as $2\rho b < \mu_{\min}$. Therefore (F–RG) achieves the optimal breakdown if a $(b, 1)$-robust aggregation rule is used. Such rules exist, as shown in Section 4, but are unfortunately not practical, as they require prior knowledge on the Byzantine nodes. It is an open question whether $\rho = 1$ can be achieved using a practical rule.

Nevertheless, we propose a practical robust summation rule that achieves $\rho = 2$, thus robust if $4b < \mu_{\min}$. This is a significant improvement over existing works. For example, in He et al. (2023), the $4b < \mu_{\min}$ condition is essentially replaced by $cb \leq \gamma\mu_{\min}$ (*e.g.*, for regular graphs), where $c > 0$ is a large constant, and $\gamma = {\mu_2(\mathcal{G}_{\mathcal{H}})}/{\mu_{\max}(\mathcal{G}_{\mathcal{H}})}$ the graph's spectral gap. This gap rapidly shrinks with the size and the lack of connectivity of the graph, making the condition orders of magnitude worse for large sparse graph. In Wu et al. (2023), the breakdown condition is $8b\sqrt{|\mathcal{H}|} \leq \mu_{\min}$, which means that the robustness guarantee decreases when the number of honest nodes increases. For instance, only a ${1}/{(9|\mathcal{H}|^{1/2})}$ fraction of Byzantine nodes is tolerated for a fully-connected network, whereas we tolerate up to ${1}/{5}$.

We conclude this section by two remarks on potential alternative characterizations of the breakdown point.

Remark 3.6 (On algebraic connectivity). Theorem 3.5 does not imply that a given communication algorithm systematically fails as soon as $2b \geq \mu_2(\mathcal{G})$, but rather that since there exists a graph for which it is the case, one can not have an r–robust algorithm with an assumption based on $\mu_2(\mathcal{G})$ and b looser than $\mu_2(\mathcal{G}) \geq 2b$. Yet, one can still prove breakdown points using other graph-related quantities, which might lead to tolerating $b > {\mu_{\min}}/{2}$ Byzantine nodes for specific graph architectures. This gap is standard in the decentralized optimization literature, where optimal algorithms depend on the (square root of the) *spectral gap* of the gossip matrix (Scaman et al., 2017; Kovalev et al., 2020), whereas iteration lower bounds are proven in terms of diameter of the communication graph.

Remark 3.7 (Dimension-dependent breakdown points). The Approximate Average Consensus problem (Section 2.2) is related to the *Approximate Consensus Problem* (ACP) (Dolev et al., 1986), in which the nodes must converge to the same value while *remaining within the convex hull* of the initial parameters. This is a harder problem, since the ACP cannot be solved using iterative communication on a system of m nodes with f Byzantine failures in dimension d if $m \leq (d+2)f + 1$ (Vaidya, 2014). This dependence on the dimension d is prohibitive for ML applications. On the contrary, our definition of r–robustness only requires the algorithm to improve the average squared distance to the target value, which is enough for D-SGD to converge, and enables us to prove *dimension-independent* breakdown. Yet, it would be interesting to link their notion of r-robust networks (LeBlanc et al., 2013) with algebraic connectivity.

4. From the general framework to practical decentralized algorithms

In this section, we first define robust summation rules and link our general framework with existing decentralized robust algorithms in Section 4.1. Then we prove convergence for Decentralized-SGD based on F–RG, in Section 4.2.

4.1. Examples of (b, ρ)-robust rules

Several methods have been proved to be (f, κ)-robust, including the Coordinate-Wise Trimmed Mean (CWTM) (Yin et al., 2018), the Coordinate-wise Median (CWM) (Yin

et al., 2018), the Geometric Median (GM) (Yin et al., 2018; Pillutla et al., 2022) and Krum (Blanchard et al., 2017). It follows from Remark 3.2 that they are also (b, ρ)–robust summation for constant weights.

Hinging on the fact that (b, ρ)–robust summation is weaker than (f, κ)–robustness, we now introduce new robust aggregation methods with tighter guarantees. In the following, we consider $(\omega_i, \boldsymbol{z}_i)_{i\in[n]} \in (\mathbb{R}_+ \times \mathbb{R}^d)^n$ and $S \subset [n]$ such that $\sum_{\overline{S}} \omega_i \leq b$, and introduce several specific robust summation functions.

Geometric Trimmed Sum (GTS). Assume w.l.o.g. that $(\|\boldsymbol{z}_i\|)_{i\in[n]}$ are sorted, i.e. $\|\boldsymbol{z}_1\| \leq \ldots \leq \|\boldsymbol{z}_n\|$, and we denote as $k^*(b) := \max\{k \in [n]; \sum_{i\geq k} \omega_i \geq b\}$ the index of the largest vector which has at least a weight b of vectors larger than it. Note that when weights sum to 1, k^* is a quantile function. (GTS) computes $\tilde{\omega}_{k^*(b)} := \sum_{i\geq k^*(b)} \omega_i - b$, and outputs

$$\mathrm{GTS}\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big) = \tilde{\omega}_{k^*(b)} \boldsymbol{z}_{k^*} + \sum_{i<k^*(b)} \omega_i \boldsymbol{z}_i.$$

In the simpler case where the weights are 1 and $b \in [n]$, GTS consists in discarding the b largest vectors within $(\boldsymbol{z}_i)_{i\in[n]}$ and summing the rest.

Clipped Sum (CS). Given a threshold function $\tau : (\mathbb{R}_+ \times \mathbb{R}^d)^n \mapsto \mathbb{R}_+$, output the mean of clipped vectors:

$$\mathrm{CS}\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}; \tau\big) := \frac{1}{n}\sum_{i=1}^n \omega_i \,\mathrm{Clip}\big(\boldsymbol{z}_i; \tau\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big)\big),$$

$$\text{where } \forall \boldsymbol{z} \in \mathbb{R}^d, \tilde{\tau} \in \mathbb{R}_+, \mathrm{Clip}(\boldsymbol{z}; \tilde{\tau}) := \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \min(\|\boldsymbol{z}\|, \tilde{\tau}).$$

We now introduce the threshold function, which gives its name to the associated clipping sum rule.[1]

Practical Clipping ($\mathrm{CS}_{\text{ours}}$). Let $\mathrm{CS}_{\text{ours}} := \mathrm{CS}\big(\,\cdot\,; \tau_{\text{ours}}\big)$, where

$$\tau_{\text{ours}}\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big) := \max\Big\{\tau \geq 0 : \sum_{i=1}^n \omega_i \mathbf{1}_{\|\boldsymbol{z}_i\|\geq\tau} \geq 2b\Big\}.$$

This rule corresponds, in the specific case of unitary weights $\omega_i = 1$ and $b \in [n]$, to defining the clipping threshold as the $2b^{th}$ largest value within $\|\boldsymbol{z}_1\|, \ldots, \|\boldsymbol{z}_n\|$ (*i.e.*, $\|z_{k^*(2b)}\|$). We now provide a robustness guarantee for these two aggregation rules in the following theorem.

Theorem 4.1. *Let $b \geq 0$, then* GTS *is (b, ρ)–robust with $\rho = 4$, and CS_{ours} is (b, ρ)–robust with $\rho = 2$.*

[1] Note that Clipped Sum cannot be a (b, ρ)–robust summation when the threshold function is a fixed constant $\tau \geq 0$: the clipping threshold must be adaptive to the input vectors.

Remark 4.2 (Concurrent work). Allouah et al. (2024a) study the influence of an adaptive clipping scheme, named Adaptive Robust Clipping (ARC), with the same adaptive clipping threshold as $\mathrm{CS}_{\text{ours}}$. However, they use it *before* any aggregation function (f, κ)–robust F, and only analyze the robustness of F $\circ$ ARC, making their approach orthogonal to ours. Moreover, their focus is on the federated case.

Next, we define *oracle* (or.) threshold functions. Those require the knowledge of the set S to be computed, which eventually corresponds to being able to identify which node is honest and which node is Byzantine. Even if this is not a reasonable assumption in practice, we use it to showcase that the optimality gap is due to the choice of threshold, and not the global rule itself. Additionally, the guarantees from He et al. (2023) rely on such assumptions[2].

Oracle Clipping. Let $\mathrm{CS}^{\text{or.}}_{\text{ours}} := \mathrm{CS}\big(\,\cdot\,; \tau^{\text{or.}}_{\text{ours}}\big)$, where $\tau^{\text{or.}}_{\text{ours}}\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big) = \max\big\{\tau \geq 0 : \sum_{i\in S} \omega_i \mathbf{1}_{\|\boldsymbol{z}_i\|\geq\tau} \geq b\big\}$.

Oracle clipping (He et al., 2023). $\mathrm{CS}^{\text{or.}}_{\text{He}} := \mathrm{CS}\big(\,\cdot\,; \tau^{\text{or.}}_{\text{He}}\big)$, where $\tau^{\text{or.}}_{\text{He}}\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big) = \sqrt{\frac{1}{b}\sum_{i\in S} \omega_i \|\boldsymbol{z}_i\|^2}$.

As shown below, $\mathrm{CS}^{\text{or.}}_{\text{ours}}$ leads to an optimal breakdown point. On the contrary, $\mathrm{CS}^{\text{or.}}_{\text{He}}$ only achieves $\rho = 4$, despite its oracle aspect. Proofs of Theorems 4.1 and 4.3 are given in Appendix E.

Theorem 4.3. *Let $b \geq 0$, then $\mathrm{CS}^{or.}_{ours}$ is (b, ρ)–robust with $\rho = 1$, and $\mathrm{CS}^{or.}_{He}$ is (b, ρ)–robust with $\rho = 4$.*

When instantiated in specific settings, our framework gives tight convergence guarantees (improving on the existing ones) for existing algorithms.

Proposition 4.4. (F–RG) *recovers existing algorithms.*

1. GTS–RG*, on a fully connected communication network $\mathcal{G}$ with constant weight, corresponds to Nearest Neighbors Averaging (Farhadkhani et al., 2022,* NNA*).*
2. $\mathrm{CS}^{or.}_{He}$–RG *recovers* ClippedGossip *(He et al., 2023).*

Robustness to asynchronous communication. Apart from sending erroneous values, Byzantine nodes can also perturb learning by never sending any message. In order for the communication protocol to continue in such settings, nodes cannot wait to receive messages from all their neighbors, since they could be stalled indefinitely by faulty nodes that would choose not to send messages. One can still build robust gossip algorithms by using a slightly modified (b, ρ)–robust summation rule.

Proposition 4.5. *Let $F : (\mathbb{R}_+ \times \mathbb{R}^d)^n \to \mathbb{R}^d$ be a (b, ρ)–robust summation. Let $F_{asyn.}$ be the rule which applies F on all inputs $(w_i, \boldsymbol{x}_i)_{i\in[n]}$ apart from an arbitrary subset $S_{delayed} \subset [n]$ of weight b, i.e $\sum_{i\in S_{delayed}} w_i = b$. Then*

[2] In their experiments, they propose a practical (i.e. non-oracle) threshold function that is not supported by theory.

$F_{asyn.}$ is a $(b, 2\rho)$–robust summation rule.

4.2. Byzantine robust Distributed SGD on graphs

We now give convergence results for a D-SGD-type algorithm that uses (F–RG) 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 4.7 builds on the reduction from Farhadkhani et al. (2023), since their requirements on the aggregation procedure exactly match the guarantees of Theorem 3.3. We consider Problem 1, where we assume that each local function f_i is a risk computed using a loss ℓ on a data distribution $\mathcal{D}_i$, i.e $f_i(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{D}_i}[\nabla\ell(\boldsymbol{x}, \boldsymbol{\xi})]$. We solve Problem 1 using D-SGD over a communication network $\mathcal{G}$. Robustness to Byzantine nodes is obtained using (F–RG) as the aggregation rule, coupled with Polyak momentum to reduce the stochastic noise.

Algorithm 1 Byzantine-Resilient D-SGD with F–RG

Input: Initial model $\boldsymbol{x}_i^0 \in \mathbb{R}^d$, local loss functions f_i, initial momentum $m_i^0 = 0$, momentum coefficient $\beta = 0$, learning rate η_{op}, communication step size η, assumption on Byzantine local corruption b.
for $t = 0$ **to** T **do**
  for $i \in \mathcal{H}$ **in parallel do**
    Sample a noisy gradient: $\boldsymbol{g}_i^t = \nabla f_i(\boldsymbol{x}_i^t) + \boldsymbol{\xi}_i^t$.
    Update the momentum: $\boldsymbol{m}_i^t = \beta\boldsymbol{m}_i^{t-1} + (1-\beta)\boldsymbol{g}_i^t$.
    Optimization step: $\boldsymbol{x}_i^{t+1/2} = \boldsymbol{x}_i^t - \eta_{op}\boldsymbol{m}_i^t$.
    Communicate $\boldsymbol{x}_i^{t+1/2}$ with the neighbors $n(i)$.
    Update the model using the gossip scheme:
    $$\boldsymbol{x}_i^{t+1} = \text{F–RG}\left(\boldsymbol{x}_i^{t+1/2}; \{\boldsymbol{x}_j^{t+1/2};\ j \in n(i)\}\right).$$
  end for
end for

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

Assumption 4.6. Objective functions regularity.

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

Under these assumptions, we prove the following corollary.

Corollary 4.7. *Let F be a (b, ρ)–robust summation, let b and $\mu_{\min}$ be such that $2\rho b \leq \mu_{\min}$, and let $\mathcal{G} \in \Gamma_{\mu_{\min}, b}$. Under Assumption 4.6, 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 $\eta_{op} = \mathcal{O}(1/\sqrt{T})$ (depending also on problem parameters such as L, γ or δ), verify as T increases:*

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

If we perform $\tilde{\mathcal{O}}(\gamma^{-1}(1-\delta)^{-1})$ steps of F–RG *between each gradient computation, we obtain:*

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

As shown above, the guarantees improve when performing more aggregation steps between gradient computations, but this brings additional communication cost. This corollary is obtained by combining our Theorem 3.3 with Theorem 1 of Farhadkhani et al. (2023), which only requires that the robust aggregation satisfies an (α, λ)-*reduction property*. So to sum up, the (b, ρ)-robust summation property of an aggregation rule allows us to construct an r-robust communication algorithm, which verifies the (α, λ) reduction property, a parameter mixture property. This property is used to derive convergence guarantees when the algorithm is used in combination with an optimization scheme. Our Theorem 3.3 ensures that (F–RG) satisfies it with $\alpha = 1 - \gamma(1-\delta)$ and $\lambda = \gamma\delta$, while 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 F.

5. Attacking robust gossip algorithms

In this section, we design an attack that aims to disrupt robust gossip algorithms. To do this, we model communications as perturbations of a gossip scheme, and analyze their impact on the variance among nodes, which allows us to deduce what perturbation Byzantines nodes should enforce for effective attacks. Recall that $\boldsymbol{X}_{\mathcal{H}}^t = (\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_{|\mathcal{H}|}^t)^T \in \mathbb{R}^{|\mathcal{H}|\times d}$ denotes the matrix of honest parameters at communication round t. Each step of F–RG can be decomposed as a perturbed gossip update (cf. Lemma C.1),

$$\boldsymbol{X}_{\mathcal{H}}^{t+1} = (\boldsymbol{I}_{\mathcal{H}} - \eta\boldsymbol{W}_{\mathcal{H}})\boldsymbol{X}_{\mathcal{H}}^t + \eta\boldsymbol{E}^t. \tag{4}$$

Where $\boldsymbol{E}^t$ is the perturbation term due to Byzantine nodes. In the following, we assume that $[\boldsymbol{E}^t]_i = \zeta_i^t\boldsymbol{a}_i^t$ for any honest node i, where $\boldsymbol{a}_i^t$ is the direction of attack on node i, and ζ_i^t is a scaling factor, which is chosen to bypass the defenses. Typically, if ζ_i^t is small, a Byzantine node $j \in n_{\mathcal{B}}(i)$ can declare to node i the parameter $\boldsymbol{x}_j^t = \boldsymbol{x}_i^t + \zeta_i^t\boldsymbol{a}_i^t$.

Dissensus Attack. Byzantine nodes can disrupt decentralized communication by maximizing the variance of the

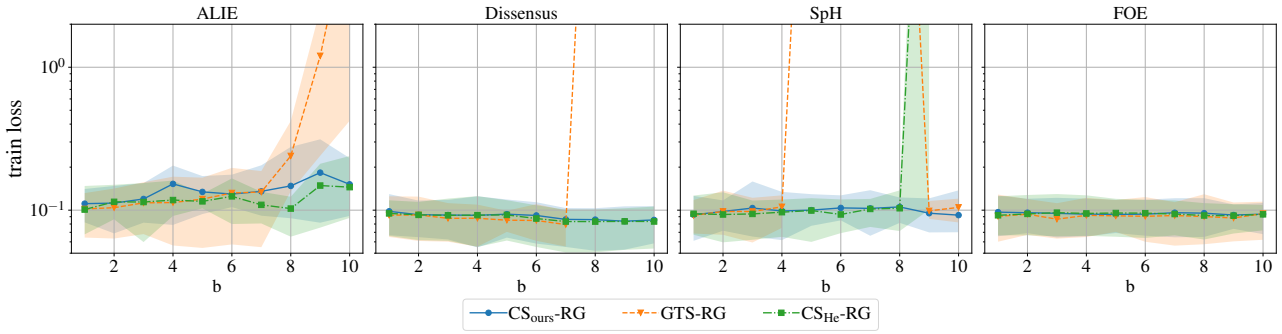

Figure 1. Training loss achieved by GTS–RG, CS_{ours}–RG and CS_{He}–RG on MNIST ($\alpha = 5$) against 4 attacks after 300 optimization and communication steps. The honest subgraph graph is $\mathcal{G}_{\mathcal{H}} := [\mathcal{G}_{m=13,k=8,c=1}]_{\mathcal{H}}$, as defined in Appendix D. Thus, $\mu_2(\mathcal{G}_{\mathcal{H}}) = 16$. The weight of Byzantines in the adjacency of honest nodes is equal to b, which varies.

honest parameters. A natural decentralized notion of variance is the *Laplacian heterogeneity* $\sum_{i,j\in\mathcal{H}} w_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$, which corresponds to $\|\boldsymbol{X}_{\mathcal{H}}\|^2_{\boldsymbol{W}_{\mathcal{H}}}$. Finding $\boldsymbol{a}_i^t$ such that this heterogeneity is maximized at $t+1$ writes

$$\begin{aligned}\underset{[\boldsymbol{E}^t]_i=\zeta_i^t\boldsymbol{a}_i^t}{\arg\max}\ &\|(\boldsymbol{I}_{\mathcal{H}} - \eta\boldsymbol{W}_{\mathcal{H}})\boldsymbol{X}_{\mathcal{H}}^t + \eta\boldsymbol{E}^t\|^2_{\boldsymbol{W}_{\mathcal{H}}} \\ &= \underset{[\boldsymbol{E}^t]_i=\zeta_i^t\boldsymbol{a}_i^t}{\arg\max}\ 2\eta\langle\boldsymbol{W}_{\mathcal{H}}\boldsymbol{X}_{\mathcal{H}}^t, \boldsymbol{E}^t\rangle + o(\eta^2).\end{aligned}$$

For small η, this suggests to take $\boldsymbol{a}_i^t = [\boldsymbol{W}_{\mathcal{H}}^t\boldsymbol{X}_{\mathcal{H}}^t]_i = \sum_{j\in n_{\mathcal{H}}(i)} w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t)$. This choice of $\boldsymbol{a}_i^t$ corresponds to the *Dissensus* attack proposed in He et al. (2023). However, as gossip communication is usually operated with multiple communication rounds, maximizing only the pairwise differences at the next step is a short-sighted approach.

Spectral Heterogeneity Attack. Byzantine nodes can take into account the fact that several rounds of communication will occur, and focus on increasing the heterogeneity over the long term. This leads, at any time t, to maximizing for any $s \geq 0$ the pairwise differences at time $t+s$, i.e, finding

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

Taking $s \to +\infty$ leads to approximating $\boldsymbol{W}_{\mathcal{H}}(\boldsymbol{I}_{\mathcal{H}} - \eta\boldsymbol{W}_{\mathcal{H}})^{2s}$ as a projection on its eigenspace associated with the largest eigenvalue of $\boldsymbol{W}_{\mathcal{H}}(\boldsymbol{I}_{\mathcal{H}} - \eta\boldsymbol{W}_{\mathcal{H}})^{2s}$. This eigenspace corresponds to the space spanned by the eigenvector of $\boldsymbol{W}_{\mathcal{H}}$ associated with the smallest non-zero eigenvalue of $\boldsymbol{W}_{\mathcal{H}}$, i.e $\mu_2(\mathcal{G}_{\mathcal{H}})$. This eigenvector (denoted $\boldsymbol{e}_{fied}$) 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 (least-connected) components. Our *Spectral Heterogeneity* attack consists in taking $\boldsymbol{a}_i^t = [\boldsymbol{e}_{fied}\boldsymbol{e}_{fied}^T\boldsymbol{X}_{\mathcal{H}}^t]_i$, which leads Byzantine nodes to cut the graph into two by pushing honest nodes in either plus or minus $\boldsymbol{e}_{fied}^T\boldsymbol{X}_{\mathcal{H}}^t$.

Experimental evaluation. We follow Farhadkhani et al. (2023) (on which our code is based), and present results for classification tasks on the MNIST dataset. We refer to Appendix B for experiments on the CIFAR-10 datasets and on an approximate averaging task. Similarly to Farhadkhani et al. (2023), heterogeneity is simulated by sampling data using a Dirichlet distribution of parameter α. We test the attacks Spectral Heterogeneity (SpH), Dissensus, A little is enough (ALIE) (Baruch et al., 2019), and Fall of Empire (FOE) (Xie et al., 2020). The main differences with Farhadkhani et al. (2023) are

(i) we consider sparse communication networks,
(ii) we implement GTS–RG instead of NNA, and ClippedGossip (denoted CS_{He}–RG) is implemented the adaptive rule of clipping of (He et al., 2023) instead of fixed thresholds,
(iii) we add Dissensus and Spectral Heterogeneity attacks,
(iv) we modify the generic design of attacks to adapt it really to the decentralized setting.

See Appendix A for further experimental details. In Figure 1 we note that the SpH attack breaks CS_{He}–RG and GTS–RG before (*i.e.*, while requiring less Byzantine nodes) Dissensus and ALIE. In this setting where heterogeneity is rather low ($\alpha = 5$), the connectivity of the graphs appears as the major limiting factor to the robustness of distributed algorithms, hence why Spectral Heterogeneity is very efficient. Furthermore, we note that CS_{ours}–RG and CS_{He}–RG have similar performances before $b = 9$.

6. Conclusion

This paper revisits robust averaging over sparse communication graphs. We introduce a general framework for robust decentralized averaging, which allows us to derive tight convergence guarantees for many robust summation rules. In particular, we show that some rules nearly match an upper bound on the breakdown point, *i.e.*, the maximum number of Byzantine nodes an algorithm can tolerate. Our experiments confirm that, contrary to ours, some existing methods

(such as NNA) indeed fail before the optimal breakdown point. We introduce a new *Spectral Heterogeneity* attack that exploits the graph topology for sparse graphs to obtain this result. An interesting future direction is the characterization of robustness when the constraint on the number of neighbors cannot be met globally, but convergence can be obtained within local neighborhoods. Conversely, this opens up questions on which nodes an attacker should corrupt to maximize its influence for a specific graph, in light of our results.

7. Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alistarh, D., Allen-Zhu, Z., and Li, J. Byzantine stochastic gradient descent. *Advances in neural information processing systems*, 31, 2018.

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

Allouah, Y., Guerraoui, R., Gupta, N., Jellouli, A., Rizk, G., and Stephan, J. The vital role of gradient clipping in byzantine-resilient distributed learning, 2024a. URL https://arxiv.org/abs/2405.14432.

Allouah, Y., Guerraoui, R., Gupta, N., Pinot, R., and Rizk, G. Robust distributed learning: tight error bounds and breakdown point under data heterogeneity. *Advances in Neural Information Processing Systems*, 36, 2024b.

Baruch, G., Baruch, M., and Goldberg, Y. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Dolev, D., Lynch, N. A., Pinter, S. S., Stark, E. W., and Weihl, W. E. Reaching approximate agreement in the presence of faults. *Journal of the ACM (JACM)*, 33(3): 499–516, 1986.

El-Mhamdi, E.-M., Guerraoui, R., Guirguis, A., Hoang, L. N., and Rouault, S. Genuinely distributed byzantine machine learning. In *Proceedings of the 39th Symposium on Principles of Distributed Computing*, pp. 355–364, 2020.

El-Mhamdi, E. M., Farhadkhani, S., Guerraoui, R., Guirguis, A., Hoang, L.-N., and Rouault, S. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in Neural Information Processing Systems*, 34:25044–25057, 2021.

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

Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pp. 6246–6283. PMLR, 2022.

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

He, L., Karimireddy, S. P., and Jaggi, M. Byzantine-robust decentralized learning via clippedgossip, 2023. URL https://arxiv.org/abs/2202.01545.

Karimireddy, S. P., He, L., and Jaggi, M. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pp. 5311–5319. PMLR, 2021.

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing, 2023. URL https://arxiv.org/abs/2006.09365.

Kovalev, D., Salim, A., and Richtárik, P. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33:18342–18352, 2020.

Lamport, L., Shostak, R., and Pease, M. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

LeBlanc, H. J., Zhang, H., Koutsoukos, X., and Sundaram, S. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31 (4):766–781, 2013.

Nedic, A. and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pp. 3027–3036. PMLR, 2017.

Vaidya, N. H. 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.

Vaidya, N. H., Tseng, L., and Liang, G. Iterative approximate byzantine consensus in arbitrary directed graphs. In *Proceedings of the 2012 ACM symposium on Principles of distributed computing*, pp. 365–374, 2012.

Wu, Z., Chen, T., and Ling, Q. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE transactions on signal processing*, 2023.

Xie, C., Koyejo, O., and Gupta, I. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pp. 261–270. PMLR, 2020.

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

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:

1. Each honest node receives different messages from Byzantine nodes: for an honest node $i \in \mathcal{H}$, Byzantines $j \in n_{\mathcal{B}}(i)$ send him at time t the vector $\boldsymbol{x}_j^t = \boldsymbol{x}_i^t + \zeta_i^t \boldsymbol{a}_i^t$. The reference point taken is the parameter of node i, instead of the average of all parameters $\overline{\boldsymbol{x}}_{\mathcal{H}}^t$, as performed in (Farhadkhani et al., 2022). Indeed $\overline{\boldsymbol{x}}_{\mathcal{H}}^t$ can be very far from the vectors in the honest neighborhood of node i since the network is not fully connected. Note that in opposition to (Farhadkhani et al., 2022), Byzantines declare *different* parameters to each of the honest nodes. Not only does it allow the use of attacks such as Dissensus and spectral heterogeneity (though the choice of $\boldsymbol{a}_i^t$), but it also allows to design ζ_i^t differently for each node. This is necessary for an optimal

2. Each scaling parameter ζ_i^t is designed separately through a linear search, such as to maximize for each honest node i

$$\left\| F\big((w_{ij}, \boldsymbol{x}_i - \boldsymbol{x}_j)_{j \in n(i)}\big) - \sum_{j \in n_{\mathcal{H}}(j)} w_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j) \right\|^2 .$$

3. The vector $\boldsymbol{a}_i^t$ is defined differently depending on the attack implemented: *Dissensus*, *Spectral Heterogeneity* (SpH), *Fall of Empire* (FOE) from Xie et al. (2020) or *A little is enough* (ALIE) from Baruch et al. (2019).

   - **Dissensus**. The Byzantines $j \in n_{\mathcal{H}}(i)$ takes as attack vector $\boldsymbol{a}_i^t = [\boldsymbol{W}_{\mathcal{H}}^t \boldsymbol{X}_{\mathcal{H}}^t]_i = \sum_{j \in n_{\mathcal{H}}(i)} w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t)$.
   - **Spectral Heterogeneity**. The Byzantines $j \in n_{\mathcal{H}}(i)$ takes as attack vector $\boldsymbol{a}_i^t = [\boldsymbol{e}_{fied} \boldsymbol{e}_{fied}^T \boldsymbol{X}_{\mathcal{H}}^t]_i$, where $\boldsymbol{e}_{fied}$ denotes an eigenvector of $\boldsymbol{W}_{\mathcal{H}}$ associated with $\mu_2(\boldsymbol{W}_{\mathcal{H}})$.
   - **ALIE**. The Byzantine nodes compute the mean of the honest parameters $\overline{\boldsymbol{x}}_{\mathcal{H}}^t$ and the coordinate-wise standard deviation $\boldsymbol{\sigma}^t$. Then they uses the attack vector $\boldsymbol{a}_i^t = \boldsymbol{\sigma}^t$.
   - **FOE**. The Byzantine nodes uses $\boldsymbol{a}_i^t = -\overline{\boldsymbol{x}}_{\mathcal{H}}^t$.

4. The aggregation is performed using a gossip update in the form of Equation (F–RG) with $\eta = \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$. The Laplacian used is the unitary-weighted Laplacian of the graph (i.e edges of the graph have unitary weights).

5. Instead of considering a constant clipping threshold for ClippedGossip of He et al. (2023), as done in the experiments of (Farhadkhani et al., 2023), we use the adaptive clipping rule suggested in He et al. (2023).

B. Experiments

In all experiments, we take as honest subgraph $\mathcal{G}_{\mathcal{H}} := [\mathcal{G}_{m=13,k=8,c=1}]_{\mathcal{H}}$, as defined in Appendix D. Thus, $\mu_2(\mathcal{G}_{\mathcal{H}}) = 16$.

B.1. MNIST dataset

In Figure 2 we provided the accuracy values of the experiments on the MNIST classification task of Figure 1 described in from Section 5. Note that the accuracy and training loss were measured after 300 optimization steps. Each configuration was run with 5 different seeds, we plotted the average loss and accuracy, as well as the minimal and maximal value over the different seeds.

B.2. CIFAR-10 dataset.

We provide in Figure 3 additional experiments on the CIFAR-10 dataset. Following (Farhadkhani et al., 2023), we use a CNN with four convolutional layers and two fully-connected layers. Furthermore, we set $\eta_{op} = 0.5$ and $T = 2000$ iterations. We consider the same communication network as for the MNIST topology, with $b = 5$ Byzantine nodes neighbor to each honest node, and $\mu_2(\mathcal{G}_{\mathcal{H}}) = 16$. We performed experiments using one seed only due to experimental running time.

We notice on the experiments that GTS–RG is non-robust in this setting to the Spectral Heterogeneity attack, while CS_{He}–RG and CS_{ours}–RGbreak under FOE attack.

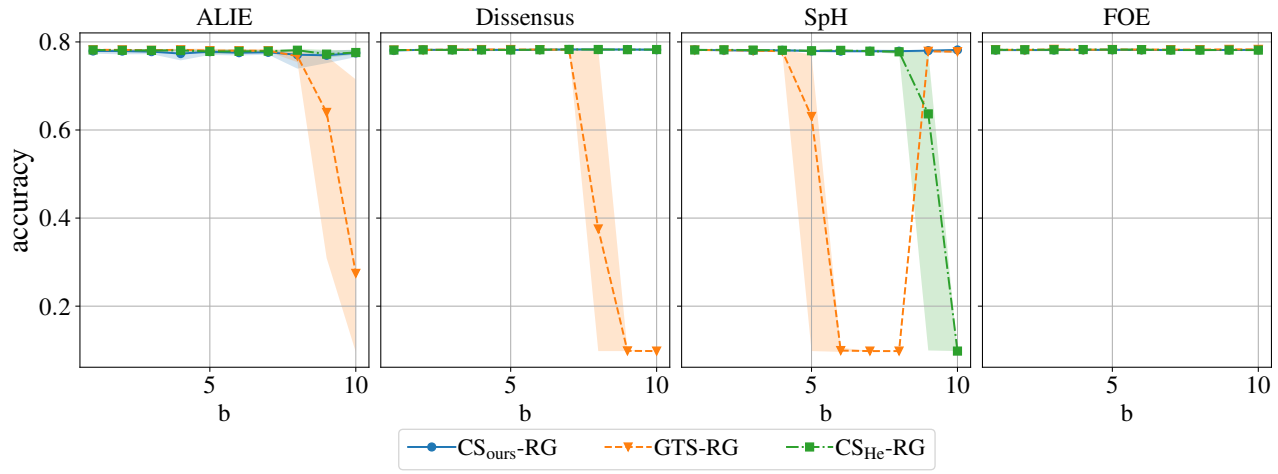

Figure 2. Accuracy achieved by GTS–RG, CS$_{ours}$–RG and CS$_{He}$–RG on MNIST against 4 attacks after 300 optimization and communication steps. The honest subgraph graph is $\mathcal{G}_{\mathcal{H}} := [\mathcal{G}_{m=13,k=8,c=1}]_{\mathcal{H}}$, as defined in Appendix D. Thus, $\mu_2(\mathcal{G}_{\mathcal{H}}) = 16$. The weight of Byzantines in the adjacency of honest nodes is equal to b, which varies.

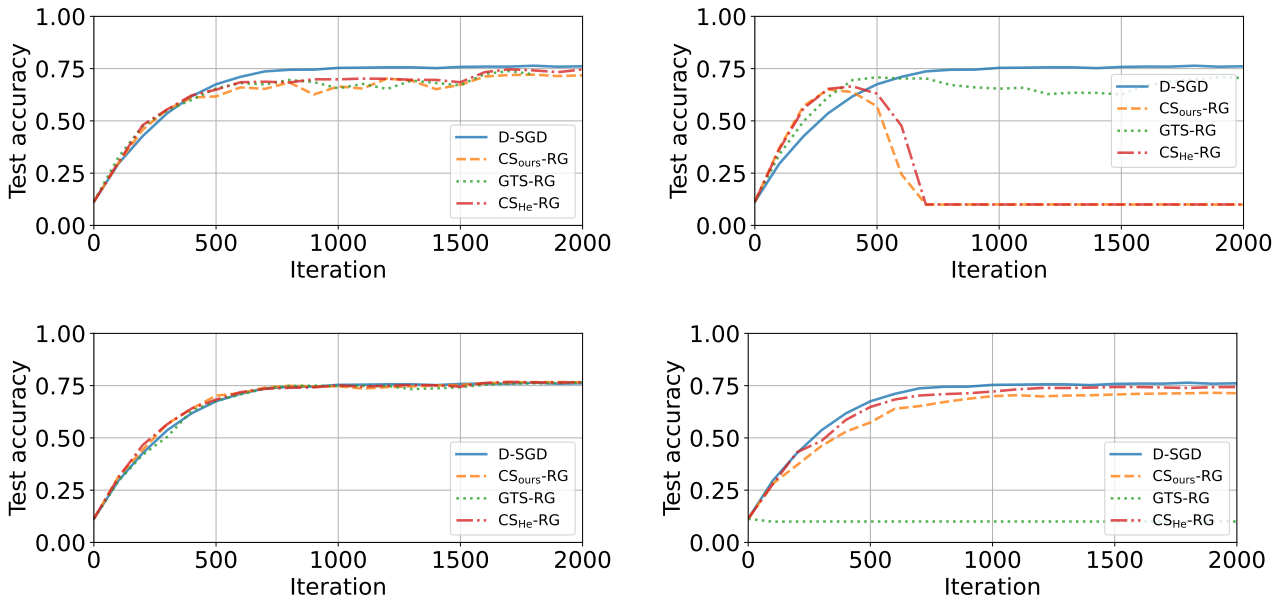

Figure 3. Test accuracies achieved by D-SGD, GTS–RG, CS$_{ours}$–RG and CS$_{He}$–RG on CIFAR-10 against 4 attacks, namely *ALIE* (row 1 left), *FOE* (row 1 right), *Dissensus* (row 2 left) and *Spectral Heterogeneity* (row 2 right). D-SGD is used as a reference and is thus not attacked by Byzantine nodes. There are $|\mathcal{H}| = 26$ honest workers, each is neighbor to $b = 5$ 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.

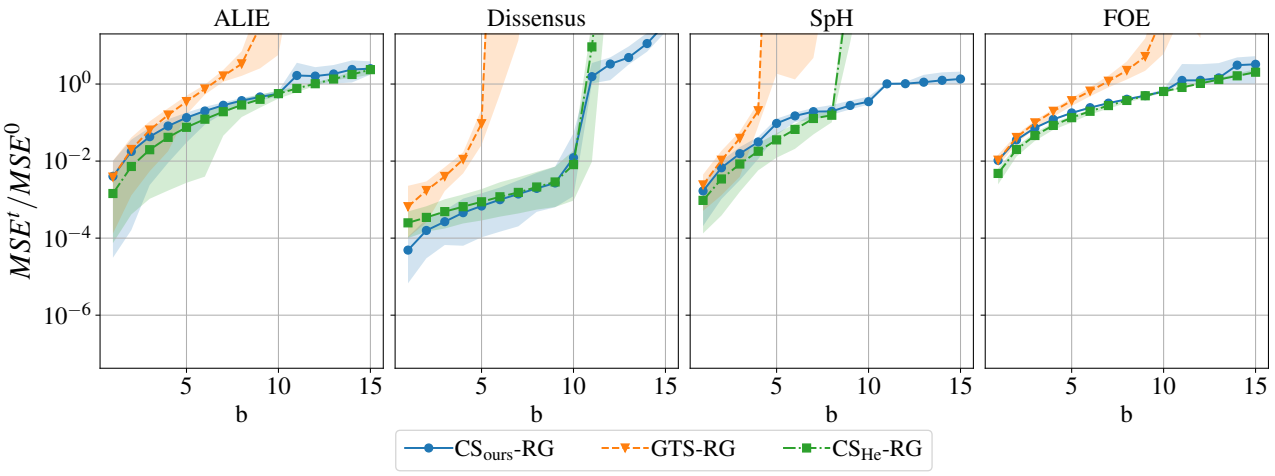

Figure 4. Relative mean square error of GTS–RG, CS_{ours}–RG and CS_{He}–RG on an averaging task. Here the optimal breakdown point is $b = 8$.

B.3. Averaging task

To finely compare the different communication schemes, we provide in Figure 4 further experiments on a simpler task of computing the average of the honest parameter values. We consider the previous graph based on two fully connected cliques of 13 honest nodes. Honest nodes parameters are initialized with a $\mathcal{N}(\boldsymbol{u}, I_d/d)$ distribution $(d = 5)$, where $\boldsymbol{u}$ is equal to $+(5, 0, \ldots, 0)^T$ for one of the two clique, and equal to $-(5, 0, \ldots, 0)^T$ for the other one. As previously, each honest node in one clique is connected to 8 honest nodes in the other clique. We perform only communication steps using GTS–RG, CS_{ours}–RG and CS_{He}–RG, and test these under the ALIE, FOE, Dissensus and Spectral Heterogeneity attacks. Eventually we plot the relative mean-square error $\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^t - \overline{\boldsymbol{x}}_{\mathcal{H}}^0\|^2 / \sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^0 - \overline{\boldsymbol{x}}_{\mathcal{H}}^0\|^2$ after $t = 100$ communication steps. For each data point, we sample 10 different initializations, and we plot the average result, as well as the minimal and maximal value.

C. Analysis of RG

C.1. Proof of Theorem 3.3

We now prove Theorem 3.3, and then use it to derive convergence for the Byzantine-robust decentralized stochastic gradient descent framework. Recall that nodes follow the update scheme below.

$$\begin{cases} \boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t - \eta F\left((w_{ij}; \boldsymbol{x}_j^t - \boldsymbol{x}_i^t)_{j\in n(i)}\right) & \text{if } i \in \mathcal{H} \\ \boldsymbol{x}_i^{t+1} = * \quad \text{if } i \in \mathcal{B}, \end{cases} \tag{5}$$

We recall the following notations

Before proving Theorem 3.3, we recall the following notations:

- $w_{ij} = -\boldsymbol{W}_{ij} \geq 0$ denote the weight associated with the edge $i \sim j$ on the graph.
- The matrix of honest parameters $\boldsymbol{X}_{\mathcal{H}}^t := \begin{pmatrix} (\boldsymbol{x}_1^t)^T \\ \vdots \\ (\boldsymbol{x}_{|\mathcal{H}|}^t)^T \end{pmatrix} \in \mathbb{R}^{|\mathcal{H}|\times d}$.
- The error due to robust aggregation and Byzantine corruption:

$$\forall i \in \mathcal{H}, \quad [\boldsymbol{E}^t]_i := \sum_{j\in n_{\mathcal{H}}(i)} w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t) - F\left((w_{ij}; \boldsymbol{x}_j^t - \boldsymbol{x}_i^t)_{j\in n(i)}\right)$$

Lemma C.1. *Equation (5) writes*

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

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

$$\begin{aligned}
\boldsymbol{x}_i^{t+1} &= \boldsymbol{x}_i^t - \eta F\left((w_{ij}; \boldsymbol{x}_j^t - \boldsymbol{x}_i^t)_{j\in n(i)}\right) \\
&= \boldsymbol{x}_i^t - \eta \sum_{j\in n_{\mathcal{H}}(i)} w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t) + \eta \left(\sum_{j\in n_{\mathcal{H}}(i)} w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t) - F\left((w_{ij}; \boldsymbol{x}_j^t - \boldsymbol{x}_i^t)_{j\in n(i)}\right) \right).
\end{aligned}$$

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

We begin by controlling the norm of the error term $\|\boldsymbol{E}^t\|_2^2$ in the case of CG$^+$.

Lemma C.2 (Control of the error)**.** *Assume F is a (b, ρ) robust summation. Then the error is controlled by the heterogeneity as measured by the Laplacian matrix:*

$$\|\boldsymbol{E}^t\|_2^2 \leq 2\rho b \|\boldsymbol{X}_{\mathcal{H}}^t\|_{\boldsymbol{W}_{\mathcal{H}}}^2 = \rho b \sum_{i\in\mathcal{H}, j\in n_{\mathcal{H}}(i)} w_{ij}\|\boldsymbol{x}_i^t - \boldsymbol{x}_j^t\|^2.$$

Proof. We recall that in this case,

$$\forall i \in \mathcal{H}, \quad [\boldsymbol{E}^t]_i := \sum_{j\in n_{\mathcal{H}}(i)} w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t) - F\left((w_{ij}; \boldsymbol{x}_j^t - \boldsymbol{x}_i^t)_{j\in n(i)}\right).$$

By assumption, for all honest node $i \in \mathcal{H}$, the weight of Byzantine in his neighborhood is smaller than b, i.e $\sum_{j\in n_{\mathcal{B}}(i)} w_{ij} \leq b$. Thus, applying the (b, ρ) robustness of F yields

$$\begin{aligned}
\|\boldsymbol{E}^t\|^2 &= \sum_{i\in\mathcal{H}} \left\| \sum_{j\in n_{\mathcal{H}}(i)} w_{ij}(\boldsymbol{x}_i^t - \boldsymbol{x}_j^t) - F\left((w_{ij}; \boldsymbol{x}_j^t - \boldsymbol{x}_i^t)_{j\in n(i)}\right) \right\|_2^2 \\
&\leq \sum_{i\in\mathcal{H}} \rho b \sum_{j\in n_{\mathcal{H}}(i)} w_{ij}\|\boldsymbol{x}_i^t - \boldsymbol{x}_j^t\|^2 \\
&= 2\rho b \|\boldsymbol{X}_{\mathcal{H}}^t\|_{\boldsymbol{W}_{\mathcal{H}}}^2,
\end{aligned}$$

Where the last equality follows by noting that $2\|\boldsymbol{X}_{\mathcal{H}}^t\|_{\boldsymbol{W}_{\mathcal{H}}}^2 = \sum_{i\in\mathcal{H}, j\in n_{\mathcal{H}}(i)} w_{ij}\|\boldsymbol{x}_i^t - \boldsymbol{x}_j^t\|^2$. 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:

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

□

Now that we control the error term, we can conclude the proof of Theorem 3.3 using standard optimization arguments. Before proving this theorem, we prove the following one, from which Corollary 3.4 is direct.

Theorem C.3. *Assume F is a (b,ρ) robust summation, and* RG *is the associated robust gossip algorithm. Let b and $\mu_{\min}$ be such that $2\rho b \leq \mu_{\min}$, and let $\mathcal{G} \in \Gamma_{\mu_{\min},b}$. Then, for any $\eta \leq \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$, the output $\boldsymbol{y} = \mathrm{RG}(\boldsymbol{x})$ (obtained by one step of* RG *on $\mathcal{G}$ from $\boldsymbol{x}$) verifies:*

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\boldsymbol{x}_i^{t+1}-\overline{\boldsymbol{x}}_{\mathcal{H}}^{t+1}\|^2 \leq (1-\eta\,(\mu_{\min}-2\rho b))\,\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\boldsymbol{x}_i^{t}-\overline{\boldsymbol{x}}_{\mathcal{H}}^{t}\|^2 \tag{6}$$

$$\|\overline{\boldsymbol{x}}_{\mathcal{H}}^{t+1}-\overline{\boldsymbol{x}}_{\mathcal{H}}^{t}\|^2 \leq \eta\frac{2\rho b}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\boldsymbol{x}_i^{t}-\overline{\boldsymbol{x}}_{\mathcal{H}}^{t}\|^2. \tag{7}$$

Proof. **Part I: Equation (7).**

Equation (7) is a direct consequence of Lemma C.2. Indeed applying $\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}} := \frac{1}{|\mathcal{H}|}\mathbf{1}_{\mathcal{H}}\mathbf{1}_{\mathcal{H}}^T$ - the orthogonal projection on the kernel of $\boldsymbol{W}_{\mathcal{H}}$ - on Lemma C.1 results in

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

Taking the norm yields

$$\|\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}}\boldsymbol{X}_{\mathcal{H}}^{t+1}-\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}}\boldsymbol{X}_{\mathcal{H}}^{t}\|^2 = \eta^2\|\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}}\boldsymbol{E}^t\|^2 \leq \eta^2\|\boldsymbol{E}^t\|^2. \tag{8}$$

We now apply Lemma C.2, and use that $\mu_{\max}(\mathcal{G}_{\mathcal{H}})$ is the largest eigenvalue of $\boldsymbol{W}_{\mathcal{H}}$. It gives

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

Finally, Equation (7) derives from $[\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}}\boldsymbol{X}_{\mathcal{H}}^{t}]_{i\in\mathcal{H}} = [\sum_{j\in\mathcal{H}}\boldsymbol{x}_j^t]_{i\in\mathcal{H}} = [\overline{\boldsymbol{x}}_{\mathcal{H}}^{t}]_{i\in\mathcal{H}}$ and $\eta\mu_{\max}(\mathcal{G}_{\mathcal{H}}) \leq 1$.

Part II: Equation (6).

To prove Equation (6), we consider the objective function $\|(\boldsymbol{I}_{\mathcal{H}}-\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}})\boldsymbol{X}^t\|^2$. We denote by $\boldsymbol{W}_{\mathcal{H}}^{\dagger}$ the Moore-Penrose pseudo inverse of $\boldsymbol{W}_{\mathcal{H}}$. We begin by applying Lemma C.1.

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

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

$$\begin{aligned}\|\boldsymbol{X}_{\mathcal{H}}^{t+1}\|^2_{(\boldsymbol{I}_{\mathcal{H}}-\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}})}-\|\boldsymbol{X}_{\mathcal{H}}^{t}\|^2_{(\boldsymbol{I}_{\mathcal{H}}-\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}})} &= -\eta\left\|\boldsymbol{X}_{\mathcal{H}}^{t}\right\|^2_{\boldsymbol{W}_{\mathcal{H}}}-\eta\left\|\boldsymbol{X}_{\mathcal{H}}^{t}-\boldsymbol{W}_{\mathcal{H}}^{\dagger}\boldsymbol{E}^t\right\|^2_{\boldsymbol{W}_{\mathcal{H}}}+\eta\left\|\boldsymbol{W}_{\mathcal{H}}^{\dagger}\boldsymbol{E}^t\right\|^2_{\boldsymbol{W}_{\mathcal{H}}}\\ &\quad+\eta^2\left\|\boldsymbol{X}_{\mathcal{H}}^{t}-\boldsymbol{W}_{\mathcal{H}}^{\dagger}\boldsymbol{E}^t\right\|_{\boldsymbol{W}_{\mathcal{H}}^2}\\ &= -\eta\left\|\boldsymbol{X}_{\mathcal{H}}^{t}\right\|^2_{\boldsymbol{W}_{\mathcal{H}}}+\eta\left\|\boldsymbol{E}^t\right\|^2_{\boldsymbol{W}_{\mathcal{H}}^{\dagger}}-\eta\left\|\boldsymbol{X}_{\mathcal{H}}^{t}-\boldsymbol{W}_{\mathcal{H}}^{\dagger}\boldsymbol{E}^t\right\|^2_{\boldsymbol{W}_{\mathcal{H}}}+\eta^2\left\|\boldsymbol{X}_{\mathcal{H}}^{t}-\boldsymbol{W}_{\mathcal{H}}^{\dagger}\boldsymbol{E}^t\right\|_{\boldsymbol{W}_{\mathcal{H}}^2}.\end{aligned} \tag{9}$$

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 $\boldsymbol{W}_{\mathcal{H}}$.

$$\|\boldsymbol{X}_{\mathcal{H}}^{t+1}\|^2_{(\boldsymbol{I}_{\mathcal{H}}-\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}})}-\|\boldsymbol{X}_{\mathcal{H}}^{t}\|^2_{(\boldsymbol{I}_{\mathcal{H}}-\boldsymbol{P}_{\mathbf{1}_{\mathcal{H}}})} \leq -\eta\left\|\boldsymbol{X}_{\mathcal{H}}^{t}\right\|^2_{\boldsymbol{W}_{\mathcal{H}}}+\eta\frac{1}{\mu_2(\mathcal{G}_{\mathcal{H}})}\left\|\boldsymbol{E}^t\right\|^2-\eta(1-\mu_{\max}(\mathcal{G}_{\mathcal{H}})\eta)\left\|\boldsymbol{X}_{\mathcal{H}}^{t}-\boldsymbol{W}_{\mathcal{H}}^{\dagger}\boldsymbol{E}^t\right\|^2_{\boldsymbol{W}_{\mathcal{H}}}.$$

Eventually Lemma C.2 with the assumption $\eta \leq 1/\mu_{\max}(\mathcal{G}_{\mathcal{H}})$ yields the result

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

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

□

To obtain Theorem C.3, we note that we can actually control the one-step variation of the MSE using $(1-\eta(\mu_{\min}-2\rho b))$ only, thus strengthening the first inequality. We rewrite the first part of Theorem 3.3 below for completeness.

Corollary C.4. *Assume F is a (b,ρ) robust summation, and* RG *is the associated robust gossip algorithm. Let b and $\mu_{\min}$ be such that $2\rho b \leq \mu_{\min}$, and let $\mathcal{G} \in \Gamma_{\mu_{\min},b}$. Then, for any $\eta \leq \mu_{\max}(\mathcal{G}_{\mathcal{H}})^{-1}$, the output $\boldsymbol{y} = \mathrm{RG}(\boldsymbol{x})$ (obtained by one step of* RG *on $\mathcal{G}$ from $\boldsymbol{x}$) verifies:*

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

Proof. We consider Equation (8) and Equation (9), which write

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

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

It follows from the bias - variance decomposition of the MSE

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

that

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

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

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

Which concludes the proof. □

C.2. Proof of Corollary 3.4

A direct consequence of the above results is Corollary 3.4, 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 $d_{t+1} = \|\boldsymbol{P}_{\mathbf{1}_\mathcal{H}} \boldsymbol{X}_\mathcal{H}^{t+1} - \boldsymbol{P}_{\mathbf{1}_\mathcal{H}} \boldsymbol{X}_\mathcal{H}^{t}\|$ and the variance at time t as $\sigma_t^2 = \|\boldsymbol{X}_\mathcal{H}^{t+1}\|^2_{(\boldsymbol{I}_\mathcal{H} - \boldsymbol{P}_{\mathbf{1}_\mathcal{H}})}$.

Corollary C.4 ensures that

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

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 C.3 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.

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

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

□

D. Proof of Theorem 3.5 - Upper bound on the breakdown point

The structure of the proof is as follows:

1. We introduce a family of communication networks $\{\mathcal{N}_{m,k}\}_{m \in \mathbb{N}, k \in [m]}$, such that $\mathcal{N}_{m,k}$ is composed of $2m$ honest nodes and m Byzantines nodes, in which each honest node is neighbor to k Byzantine nodes.
2. We show that no algorithm can be r–robust on any communication network within this family.
3. We show that uniform weighting of the graphs associated with these communication networks leads to a family of weighted graphs $\{\mathcal{G}_{m,k,c}\}_{m \in \mathbb{N}, k \in [m], c \in \mathbb{R}_+}$ such that
   (a) The weight of Byzantines in the neighborhood of any honest node is equal to $b = ck$.
   (b) The algebraic connectivity of the honest subgraph of any graph within this family always verifies:
   $$\mu_2([\mathcal{G}_{m,k,c}]_\mathcal{H}) = 2ck = 2b.$$
   (c) For any $h \geq b$, there exists a graph within $\{\mathcal{G}_{m,k,c}\}_{m \in \mathbb{N}, k \in [m], c \in \mathbb{R}_+}$ such that all honest nodes have a weight associated to honest neighbors $h(i)$ larger than h.

D.1. Definition of the family of graphs

Let $m \in \mathbb{N}^*$ r and $k \in [m]$.

We define the communication network $\mathcal{N}_{m,k}$ as composed of three cliques of m nodes: C_1, C_2, and C_3. Each node in C_i is additionally connected to exactly k nodes in $C_{i+1 \mod 3}$ and to k nodes in $C_{i-1 \mod 3}$. 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 $j, \ldots, j + k \mod m$ in $C_{i+1 \mod 3}$.

We assume that one of the cliques is composed of Byzantine nodes, each honest node having k Byzantine neighbors, and there are m Byzantine nodes among the $3m$ nodes.

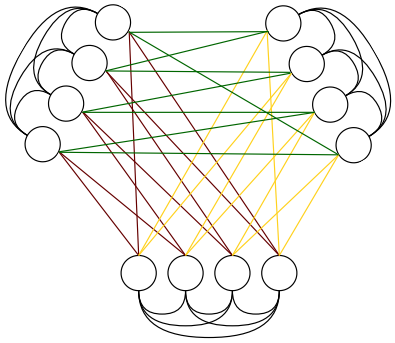

Figure 5. Topology of $\mathcal{N}_{m=4,k=2}$.

D.2. No algorithm can be α-robust on $\mathcal{N}_{m,k}$.

We now show that no algorithm can be robust on the communication network $\mathcal{N}_{m,k}$.

Lemma D.1. *Let one unknown clique within C_1, C_2, and C_3 be composed of Byzantine nodes, then no communication algorithm is α-robust on $\mathcal{N}_{m,k}$.*

Assume an r–robust algorithm exists on the communication network $\mathcal{G}_{H,k}$. Informally:

1. We first show that if all nodes within one clique hold a unique parameter $\boldsymbol{x}$, and receive this parameter from nodes of either of the two other cliques, then they cannot change their parameter.
2. We then consider a setting where the two honest cliques hold 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, $r < 1$ is impossible.

Proof.
Part I. Let Alg be an algorithm r–robust on $\mathcal{N}_{m,k}$ with $r < 1$. We denote $\boldsymbol{x}_i^+$ the output from node i after running Alg.

We know that one of the cliques say C_1, is composed of honest nodes. Let the honest nodes within C_1 hold the parameter $\boldsymbol{x}$. Nodes in another clique, say C_2, declare the parameter $\boldsymbol{x}$ as well, while nodes in C_3 declare another parameter, say $\boldsymbol{y} \neq \boldsymbol{x}$. We show that all nodes $i \in C_1$ must output the parameter $\hat{\boldsymbol{x}}_i = \boldsymbol{x}$.

From the point of view of nodes in C_1 considering that the Byzantine clique is unknown, it is impossible to distinguish between these situations:

1. Situation I: C_2 is honest, and C_3 is Byzantine,
2. Situation II: C_2 is Byzantine, and C_3 is honest,

thus the outputs given by nodes in C_1 after running Alg, denoted $(\boldsymbol{x}_i^+)_{i \in C_1}$, are the same in both situations.

In Situation I, nodes of C_2 are honest, and nodes in C_1 and C_2 have the same initial parameter; hence, the initial quadratic error is 0. The r criterion writes

$$\sum_{i \in \mathcal{H}} \|\boldsymbol{x}_i^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 \leq r \sum_{i \in \mathcal{H}} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 = 0.$$

It follows that for any node i in C_1, $\boldsymbol{x}_i^+ = \boldsymbol{x}$, *i.e.*, nodes in C_1 do not change their parameters (in both Situation I and Situation II).

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 $\boldsymbol{x}$, while nodes in C_2 hold the parameter $\boldsymbol{y} \neq \boldsymbol{x}$.

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 $\boldsymbol{x}$, and to nodes in C_2 that they hold the value $\boldsymbol{y}$. Following Part I, nodes in C_1 and in C_2 cannot update their parameter, (i.e. $\forall i \in \mathcal{H},\ \boldsymbol{x}_i^+ = \boldsymbol{x}_i$). Applying the r–robustness property brings:

$$\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 \leq r \sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 = r \sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2,$$

which implies that $r \geq 1$ since $\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 > 0$. It follows that the Algorithm Alg is not r–robust on $\mathcal{N}_{m,k}$ for an $r < 1$.

□

D.3. Spectral properties of the graph.

We define as $\mathcal{G}_{m,k,c}$ the graph associated with the network $\mathcal{N}_{m,k}$ where all edges have weighted as c.

We show here that $\mu_2([\mathcal{G}_{m,k,c}]_{\mathcal{H}}) = 2kc$. This concludes the proof as we have:

1. The weight of Byzantines in the neighborhood of any honest node is equal to $b = ck$;
2. Which is linked to the algebraic connectivity of the honest subgraph $\mu_2([\mathcal{G}_{m,k,c}]_{\mathcal{H}}) = c2k$;
3. While the weight of honest nodes in the neighborhood of honest nodes is equal to $h = c(m + k)$, thus growths to infinity with m;

Analysis. We denote $\boldsymbol{L}_{\mathcal{H}}$ the Laplacian of $[\mathcal{G}_{m,k,c=1}]_{\mathcal{H}}$, the honest subgraph of the network $\mathcal{N}_{m,k}$ in which we provided unitary weights to edges. We show that $\mu_2([\mathcal{G}_{m,k,c=1}]_{\mathcal{H}}) = 2k$, which brings that $\mu_2([\mathcal{G}_{m,k,c}]_{\mathcal{H}}) = 2kc$.

Lemma D.2. *The second smallest eigenvalue of the (unitary weighted) Laplacian matrix $\boldsymbol{L}_{\mathcal{H}}$ of $[\mathcal{G}_{m,k,1}]_{\mathcal{H}}$ is equal to $\mu_2(\boldsymbol{L}_{\mathcal{H}}) = 2k$.*

Proof. Recall that $m \triangleq m$. Let $\boldsymbol{M} \in \mathbb{R}^{m\times m}$ be a circulant matrix defined as $\boldsymbol{M} = \sum_{q=0}^{k-1} \boldsymbol{J}^q$, where $\boldsymbol{J}$ denotes the permutation

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

The Laplacian matrix of the honest subgraph of $\mathcal{G}_{H,k}$ can be written as:

$$\boldsymbol{L}_{\mathcal{H}} = \begin{pmatrix} m\boldsymbol{I}_m - \mathbf{1}_m\mathbf{1}_m^T & 0 \\ 0 & m\boldsymbol{I}_m - \mathbf{1}_m\mathbf{1}_m^T \end{pmatrix} + \begin{pmatrix} k\boldsymbol{I}_m & -\boldsymbol{M} \\ -\boldsymbol{M}^T & k\boldsymbol{I}_m \end{pmatrix}$$

Hence

$$\boldsymbol{L}_{\mathcal{H}} = (k+m)\boldsymbol{I}_{2m} - \begin{pmatrix} \mathbf{1}_m\mathbf{1}_m^T & 0 \\ 0 & \mathbf{1}_m\mathbf{1}_m^T \end{pmatrix} - \begin{pmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^T & 0 \end{pmatrix}. \tag{10}$$

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

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

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

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

Then the eigenvalues of $\boldsymbol{A}$ are $\{\pm|\mu_p|; p = 0, \dots, m-1\}$.

Proof of Lemma D.4. Let $\boldsymbol{M} = \boldsymbol{U}^* \boldsymbol{D} \boldsymbol{U}$ be the diagonalization of $\boldsymbol{M}$ where $\boldsymbol{D} = \mathrm{Diag}(\mu_0, \dots, \mu_{m-1})$, and $\boldsymbol{U}$ is a unitary matrix, i.e. $\boldsymbol{U}\boldsymbol{U}^* = \boldsymbol{I}$ where we denote as $\boldsymbol{U}^* = \overline{\boldsymbol{U}}^T$ the conjugate transpose of $\boldsymbol{U}$, where the (simple) conjugate matrix is denoted $\overline{\boldsymbol{U}}$.

Lemma D.4 follows from

$$\begin{pmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^T & 0 \end{pmatrix} = \begin{pmatrix} 0 & \boldsymbol{U}^* \boldsymbol{D} \boldsymbol{U} \\ \boldsymbol{U}^T \boldsymbol{D} \overline{\boldsymbol{U}} & 0 \end{pmatrix} \underset{\overline{\boldsymbol{A}} = \boldsymbol{A}}{=} \begin{pmatrix} 0 & \boldsymbol{U}^* \boldsymbol{D} \boldsymbol{U} \\ \boldsymbol{U}^* \overline{\boldsymbol{D}} \boldsymbol{U} & 0 \end{pmatrix}.$$

Hence

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

A simple calculus (using that $\boldsymbol{D}$ is diagonal) yields that all eigenvalues of $\begin{pmatrix} 0 & \boldsymbol{D} \\ \overline{\boldsymbol{D}} & 0 \end{pmatrix}$ are $\{\pm|\boldsymbol{D}_p|; p = 0, \dots, m-1\}$. □

Proof of Lemma D.3. We start from the decomposition of Equation (10) :

$$\boldsymbol{L}_{\mathcal{H}} = (k+m)\boldsymbol{I}_{2m} - \begin{pmatrix} \mathbf{1}_m \mathbf{1}_m^T & 0 \\ 0 & \mathbf{1}_m \mathbf{1}_m^T \end{pmatrix} - \begin{pmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^T & 0 \end{pmatrix}.$$

We first notice that the subspace spanned by $(\mathbf{1}_m^T, +\mathbf{1}_m^T)^T$ and $(\mathbf{1}_m^T, -\mathbf{1}_m^T)^T$ is an eigenspace of $\begin{pmatrix} \mathbf{1}_m \mathbf{1}_m^T & 0 \\ 0 & \mathbf{1}_m \mathbf{1}_m^T \end{pmatrix}$ associated the eigenvalue m, and the orthogonal subspace is associated with 0. Furthermore these are eigenvectors of $\begin{pmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^T & 0 \end{pmatrix}$ associated with k and $-k$. It follows that they are eigenvectors of $\boldsymbol{L}_{\mathcal{H}}$ with eigenvalues 0 and $2k$. We notice as well that the three matrices of Equation (10) can be diagonalized in the same orthogonal basis.

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

$$\boldsymbol{M} = \boldsymbol{U}^* \boldsymbol{D} \boldsymbol{U}.$$

Considering Lemma D.4, the eigenvalue of $\begin{pmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^T & 0 \end{pmatrix}$ are $\{\pm|\mu_p|; p = 0, \dots, m-1\}$, considering that $p = 0$ corresponds to the eigenvalues $+k$ and $-k$, 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 $\boldsymbol{L}_{\mathcal{H}}$ are $\{\pm|\mu_p|; p = 0 \dots m-1\}$. □

End of the proof of Lemma D.2.

To prove Lemma D.2, considering the decomposition of Equation (10), we only have to show that $m - k$ is always the second largest eigenvalue of the matrix

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

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

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

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

E. (b, ρ)-robust summation rules

We recall the definition of summation rules.

Definition E.1 ((b, ρ)–robust summation)**.** Let $b, \rho \geq 0$. An aggregation rule $F : (\mathbb{R}_+ \times \mathbb{R}^d)^n \to \mathbb{R}^d$ is a (b, ρ)–*robust summation*, when, for any vectors $(\boldsymbol{z}_i)_{i\in[n]} \in (\mathbb{R}^d)^n$, any weights $(\omega_i)_{i\in[n]} \in \mathbb{R}^n_+$ and any set $S \subset [n]$ such that $\sum_{i\in\overline{S}} \omega_i \leq b$, there is

$$\left\| F\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big) - \sum_{i\in S} \omega_i \boldsymbol{z}_i \right\|^2 \leq \rho b \sum_{i\in S} \omega_i \|\boldsymbol{z}_i\|^2.$$

where $\overline{S} := [n]\backslash S$.

E.1. Clipping

We first prove the robustness of the clipping-based summation.

Proposition E.2. *Let $b \geq 0$, then*

1. *(Practical)* CS_{ours} *is (b, ρ)–robust with $\rho = 2$.*
2. *(Oracle)* $\mathrm{CS}^{or.}_{ours}$ *is (b, ρ)–robust with $\rho = 1$.*
3. *(Oracle)* $\mathrm{CS}^{or.}_{He}$ *is (b, ρ)–robust with $\rho = 4$.*

Proof. Let $(\omega_i, \boldsymbol{z}_i)_{i\in[n]} \in (\mathbb{R}_+ \times \mathbb{R}^d)^n$, and $S \subset [n]$ such that $\sum_{i\in\overline{S}} \omega_i \leq b$,

We consider for $\tau \geq 0$

$$F\big((\omega_i, \boldsymbol{z}_i)_{i=1,\dots,n}\big) := \sum_{i=1}^n \omega_i \operatorname{Clip}(\boldsymbol{z}_i; \tau).$$

Then, by applying the triangle inequality we have

$$\begin{aligned}
\left\| F\big((\omega_i, \boldsymbol{z}_i)_{i=1,\dots,n}\big) - \sum_{i\in S} \omega_i \boldsymbol{z}_i \right\|^2 &= \left\| \sum_{i\in S} \omega_i \left(\operatorname{Clip}(\boldsymbol{z}_i; \tau) - \boldsymbol{z}_i\right) + \sum_{i\in\overline{S}} \omega_i \operatorname{Clip}(\boldsymbol{z}_i; \tau) \right\|^2 \\
&\leq \left(\sum_{i\in S} \omega_i \|\boldsymbol{z}_i - \operatorname{Clip}(\boldsymbol{z}_i; \tau)\| + \sum_{i\in\overline{S}} \omega_i \|\operatorname{Clip}(\boldsymbol{z}_i; \tau)\| \right)^2 \\
&\leq \left(\sum_{i\in S} \omega_i (\|\boldsymbol{z}_i\| - \tau)_+ + b\tau \right)^2. \qquad (11)
\end{aligned}$$

Where we used $\sum_{i\in\overline{S}} \omega_i \leq b$.

1. Case of our clipping threshold. We choose τ as

$$\tau = \max\left\{ \tau \geq 0 : \sum_{i=1}^n \omega_i \mathbf{1}_{\|\boldsymbol{z}_i\|\geq\tau} \geq 2b \right\} \triangleq \tau_{\text{ours}}\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big).$$

This corresponds to lowering the clipping threshold until the sum of the weights of clipped vectors is essentially equal to $2b$. This ensures that the total weight of the honest vectors that are clipped falls between b and $2b$. If there are ties at the clipping threshold, honest vectors can be arbitrarily denoted as *clipped* or *non-clipped*. Indeed, there is no clipping error incurred since the clipping threshold is the same as the actual value of the difference. Therefore, we do not have to accumulate error for the weight over $2b$, so the following equation always holds:

$$2b \geq \sum_{i\in S} \omega_i \mathbf{1}_{i \text{ clipped}} \geq b.$$

Which allows us to write:

$$\sum_{i\in S}\omega_i\left(\|\boldsymbol{z}_i\|-\tau\right)_+ + b\tau \leq \sum_{i\in S}\omega_i(\|\boldsymbol{z}_i\|-\tau)\mathbf{1}_{i \text{ clipped}} + b\tau$$
$$\leq \sum_{i\in S}\omega_i\|\boldsymbol{z}_i\|\mathbf{1}_{i \text{ clipped}}.$$

We conclude the proof using the Cauchy-Schwarz inequality:

$$\left\|F\big((\omega_i,\boldsymbol{z}_i)_{i=1,\dots,n}\big)-\sum_{i\in S}\omega_i\boldsymbol{z}_i\right\|^2 \leq \left(\sum_{i\in S}\sqrt{\omega_i}\|\boldsymbol{z}_i\|\cdot\sqrt{\omega_i}\mathbf{1}_{i \text{ clipped}}\right)^2$$
$$\leq \left(\sum_{i\in S}\omega_i\mathbf{1}_{i \text{ clipped}}\right)\left(\sum_{i\in S}\omega_i\|\boldsymbol{z}_i\|^2\right)$$
$$\leq 2b\sum_{i\in S}\omega_i\|\boldsymbol{z}_i\|^2.$$

Where we used $\sum_{j\in n_{\mathcal{H}}(i)}\omega_{ij}\mathbf{1}_{j \text{ clipped}} \leq 2b$. Note that, if somehow it is possible to choose the clipping threshold such that the weight of clipped vectors within S is equal exactly to b then this factor 2 disappears. Which correspond to our oracle clipping threshold $\tau^{\text{or.}}_{\text{ours}}$. The same result can be achieved when it is possible to identify a subset of $\{1,\dots,n\}$ of weight $2b$ which includes the set $\overline{S}$, and if the weight with $\overline{S}$ sums exactly to b. This is for instance the case of the communication graph used in Appendix D: nodes in (for instance) C_1 know that Byzantines nodes are among the nodes within C_2 and C_3, thus selecting all their neighbors that belong to these two other cliques leads to a subset of neighbors of weight $2b$ with exactly a weight b corresponding to Byzantine neighbors.

Clipping threshold of He et al. (2023) We plug in Equation (11) the following upper bound:

$$\left(\|\boldsymbol{z}_i\|-\tau\right)_+ = \tau\left(\frac{\|\boldsymbol{z}_i\|}{\tau}-1\right)_+ \leq \tau\left(\frac{\|\boldsymbol{z}_i\|^2}{\tau^2}-1\right)_+ \leq \frac{\|\boldsymbol{z}_i\|^2}{\tau}.$$

This yields

$$\left\|F\big((\omega_i,\boldsymbol{z}_i)_{i=1,\dots,n}\big)-\sum_{i\in S}\omega_i\boldsymbol{z}_i\right\|^2 \leq \left(\sum_{i\in S}\omega_i\frac{\|\boldsymbol{z}_i\|^2}{\tau}+b\tau\right)^2.$$

Then taking as clipping threshold the minimizer of the RHS, $\tau^* = \sqrt{\frac{1}{b}\sum_{i\in S}\omega_i\|\boldsymbol{z}_i\|^2} \triangleq \tau^{\text{or.}}_{\text{He}}\big((\omega_i,\boldsymbol{z}_i)_{i\in[n]}\big)$ leads to:

$$\left\|F\big((\omega_i,\boldsymbol{z}_i)_{i=1,\dots,n}\big)-\sum_{i\in S}\omega_i\boldsymbol{z}_i\right\|^2 \leq 4b\sum_{i\in S}\omega_i\|\boldsymbol{z}_i\|^2.$$

However, the clipping threshold here requires an exact knowledge of the set S. Furthermore, it is unclear to what extend making an approximate estimate of this clipping threshold allows us to derive robustness guarantees. □

Remark E.3. A key point here is that this oracle clipping threshold corresponds to the **unique minimizer** within each squared term of the sum. Hence, considering for instance the adaptive practical clipping rule of (He et al., 2023) leads to a **larger upper bound on the error**.

E.2. Geometric trimming a.k.a. NNA

Let $(\omega_i,\boldsymbol{z}_i)_{i\in[n]} \in (\mathbb{R}_+\times\mathbb{R}^d)^n$, and $S\subset[n]$ such that $\sum_{i\in\overline{S}}\omega_i \leq b$,

We recall the definition of GTS: Assume w.l.o.g. that $(\|\boldsymbol{z}_i\|)_{i\in[n]}$ are sorted, i.e. $\|\boldsymbol{z}_1\| \leq \ldots \leq \|\boldsymbol{z}_n\|$, and denote $k^*(b) := \max\{k \in [n]; \sum_{i\geq k} \omega_i \geq b\}$ the index of the largest vector which has at least a weight b of vector largest than him. (GTS) computes $\tilde{\omega}_{k^*(b)} := \sum_{i\geq k^*(b)} \omega_i - b$, and outputs

$$\mathrm{GTS}\big((\omega_i, \boldsymbol{z}_i)_{i\in[n]}\big) = \tilde{\omega}_{k^*(b)}\boldsymbol{z}_{k^*} + \sum_{i<k^*(b)} \omega_i \boldsymbol{z}_i.$$

Lemma E.4. *Geometric trimming is (b, ρ) -robust with $\rho = 4$.*

Proof. Let $(\omega_i, \boldsymbol{z}_i)_{i\in[n]} \in (\mathbb{R}_+ \times \mathbb{R}^d)^n$, and $S \subset [n]$ such that $\sum_{i\in\overline{S}} \omega_i \leq b$, Without loss of generality we assume that $\tilde{\omega}_{k^*(b)} = 0$[3].

Thus the aggregation rules write

$$F\big((\omega_i, \boldsymbol{z}_i)_{i=1,\ldots,n}\big) := \sum_{i<k^*(b)} \omega_i \boldsymbol{z}_i. = \sum_{i=1}^{n} \omega_i \boldsymbol{z}_i \mathbf{1}_{i \text{ not removed}}$$

Then, by applying the triangle inequality we have

$$\begin{aligned}\left\| F\big((\omega_i, \boldsymbol{z}_i)_{i=1,\ldots,n}\big) - \sum_{i\in S} \omega_i \boldsymbol{z}_i \right\|^2 &= \left\| \sum_{i\in S} \omega_i \boldsymbol{z}_i \mathbf{1}_{i \text{ removed}} + \sum_{i\in\overline{S}} \omega_i \boldsymbol{z}_i \mathbf{1}_{i \text{ not removed}} \right\|^2 \\ &\leq \left(\sum_{i\in S} \omega_i \|\boldsymbol{z}_i\| \mathbf{1}_{i \text{ removed}} + \sum_{i\in\overline{S}} \omega_i \|\boldsymbol{z}_i\| \mathbf{1}_{i \text{ not removed}} \right)^2.\end{aligned}$$

As $\sum_{i=1}^n \omega_i \mathbf{1}_{i \text{ removed}} = b$, it holds that

$$\begin{aligned}&\sum_{i\in\overline{S}} \omega_i \leq b = \sum_{i=1}^{n} \omega_i \mathbf{1}_{i \text{ removed}} = \sum_{i\in S} \omega_i \mathbf{1}_{i \text{ removed}} + \sum_{i\in\overline{S}} \omega_i \mathbf{1}_{i \text{ removed}} \\ \implies &\sum_{i\in\overline{S}} \omega_i \mathbf{1}_{i \text{ not removed}} \leq \sum_{i\in S} \omega_i \mathbf{1}_{i \text{ removed}}.\end{aligned}$$

Furthermore, if i is removed and j is not, then $\|\boldsymbol{z}_i\| \geq \|\boldsymbol{z}_j\|$. It follows that

$$\sum_{i\in S} \omega_i \|\boldsymbol{z}_i\| \mathbf{1}_{i \text{ removed}} \geq \sum_{i\in\overline{S}} \omega_i \|\boldsymbol{z}_i\| \mathbf{1}_{i \text{ not removed}}.$$

Consequently:

$$\begin{aligned}\left\| F\big((\omega_i, \boldsymbol{z}_i)_{i=1,\ldots,n}\big) - \sum_{i\in S} \omega_i \boldsymbol{z}_i \right\|^2 &\leq \left(2 \sum_{i\in S} \omega_i \|\boldsymbol{z}_i\| \mathbf{1}_{i \text{ removed}} \right)^2 \\ &= 4 \left(\sum_{i\in S} \sqrt{\omega_i} \|\boldsymbol{z}_i\| \sqrt{\omega_i} \mathbf{1}_{i \text{ removed}} \right)^2 \\ &\leq 4 \sum_{i\in S} \omega_i \mathbf{1}_{i \text{ removed}} \sum_{i\in S} \omega_i \|\boldsymbol{z}_i\|^2 \qquad \text{using Cauchy-Schwarz.}\end{aligned}$$

Thus:

$$\left\| F\big((\omega_i, \boldsymbol{z}_i)_{i=1,\ldots,n}\big) - \sum_{i\in S} \omega_i \boldsymbol{z}_i \right\|^2 \leq 4b \sum_{i\in S} \omega_i \|\boldsymbol{z}_i\|^2.$$

□

[3]Which can be ensured by adding artificially the entry $(\tilde{\omega}_{k^*(b)}, \boldsymbol{z}_{k^*(b)})$.

F. Proofs for $D-SGD$

Proof of Corollary 4.7. 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 (α, λ)-reduction, which they prove in their Lemma 2. Then, all subsequent results invoke this Lemma instead of the specific aggregation procedure. CG$^+$ also satisfies (α, λ)-reduction, as we prove in Theorem 3.3. 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:

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f_{\mathcal{H}}(\boldsymbol{x}_i^t)\right\|^2\right] = \mathcal{O}\left(\frac{L\sigma}{\sqrt{T}}(1+C) + \zeta^2 C\right), \tag{12}$$

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:

$$c_1 = \frac{(1-\gamma(1-\delta))(2-\gamma(1-\delta))}{\gamma^2(1-\delta)^2} = O\left(\frac{1}{\gamma^2(1-\delta)^2}\right). \tag{13}$$

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

Multi-step derivations. In the previous case, we see that C is dominated by the c_1 term since $c_1 >> \lambda$. In particular, the guarantees would increase if we were able to trade-off some α for some λ, which is possible by using multiple communications steps. This is what we do, and take enough steps that $c_1 << \lambda$ (i.e., $\alpha \approx 0$), so that $C \approx \lambda$. Following Corollary 3.4, this requires $\tilde{O}(\gamma^{-1}(1-\delta)^{-1})$ steps, where logarithmic factors are hidden in the $\tilde{O}$ notation. We then plug the multi-step λ value from Corollary 3.4 to obtain the result.

□