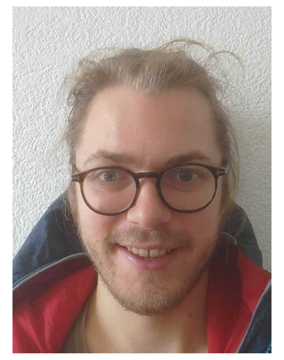
A Unified Breakdown Analysis for Byzantine Robust Gossip

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

[1]École polytechnique, Palaiseau [2]INRIA Grenoble





arxiv:2410.10418

Context

Many data providers (i.e. nodes) aim to train collaboratively a model using peer-to-peer synchronous communications. Some of them are omniscient adversaries called *Byzantine*.

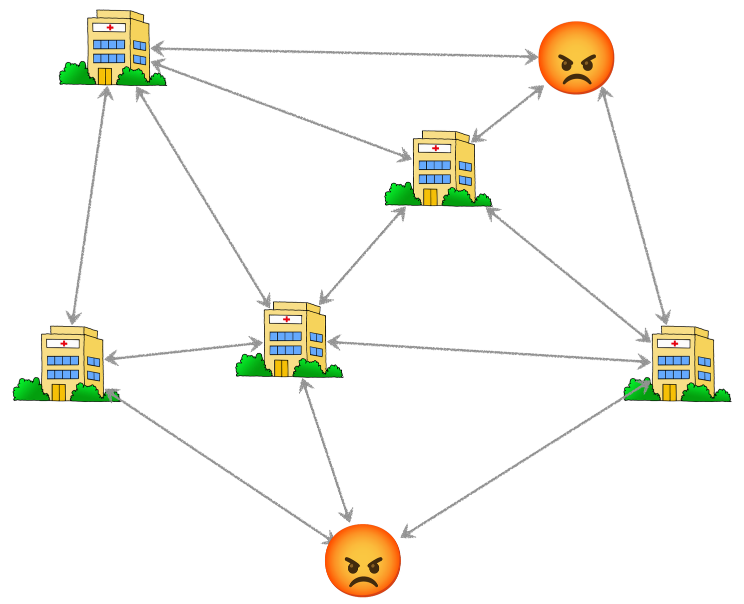
Takeaway

→ We combine 'any' robust average with gossip communication.

→ The second smallest eigenvalue of the graph's Laplacian & the number of adversarial neighbors measures the robustness of the resulting algorithm.

→ Our breakdown point is optimal up to a factor 2.

Setting



- Honest nodes $\mathcal{H}$ and Byzantine nodes $\mathcal{B}$ communicate in a graph $\mathcal{G} = (\mathcal{H} \cup \mathcal{B}, \mathcal{E})$.
- $\mu_{\max}$ and μ_2 are the largest and second smallest eigenvalue of the Laplacian matrix of the *honest subgraph*:

$$\mathrm{L} = \mathrm{Diagonal(degrees)} - \mathrm{Adjacency}.$$

Distributed optimization problem.

$$\text{Minimize} \quad f_{\mathcal{H}}(\boldsymbol{x}) := \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} f_i(\boldsymbol{x}).$$

Average consensus problem. Each node holds a parameter $\boldsymbol{x}_i \in \mathbb{R}^d$.

$$\text{Get close to} \quad \boldsymbol{x}_{\mathcal{H}} = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \boldsymbol{x}_i.$$

Assumption: Each honest node has at most b Byzantines neighbors.

Notation: $\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}) := \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2$, i.e. the variance of honest parameters.

r-Robust Communication

For $r<1$, the communication algorithm is r-robust on $\mathcal{G}$ if, for all $\boldsymbol{x}_i \in \mathbb{R}^d$, the outputs $(\boldsymbol{x}_i^+)_{i\in\mathcal{H}}$ satisfies

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^+ - \overline{\boldsymbol{x}}_{\mathcal{H}}\|^2 \le r\,\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}).$$

The Robust Gossip framework

Robust Aggregators

Let $b, \rho \ge 0$. An aggregation rule $F : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is a (b,ρ)–**robust summation** if, for any vectors $(\boldsymbol{z}_i)_{i\in[n]} \in (\mathbb{R}^d)^n$, any $S \subset [n]$ such that $|S| \ge n-b$,

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

↪ Weaker than (f,κ)-robustness[1]: it relies on a *second moment* instead of a variance.

Algorithm: F - Robust Gossip

Let F an aggregation rule, and $\eta \ge 0$ a communication step-size. At each iteration all honest nodes $i \in \mathcal{H}$ perform

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \eta F\left((\boldsymbol{x}_j^t - \boldsymbol{x}_i^t)_{j\in\mathrm{neighbors}(i)}\right). \tag{F-RG}$$

- The robust aggregation is performed on the *differences* of the parameters!
- If F is a simple sum, F-RG recovers the usual gossip update.

Instances of Robust Summation

Assume wlog that $\|\boldsymbol{z}_1\| \ge \ldots \ge \|\boldsymbol{z}_n\|$.

- **Clipped Sum$_+$** (CS$_+$). Denote $\mathrm{Clip}(\boldsymbol{z},\tau) = \min(\tau, \|\boldsymbol{z}\|)\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}$

$$\mathrm{CS}_+\big((\boldsymbol{z}_i)_{i\in[n]}\big) = \sum_{i\in[n]} \mathrm{Clip}(\boldsymbol{z}_i;\tau) \quad \text{with} \quad \tau = \|\boldsymbol{z}_{2b}\|.$$

- **Geometric Trimmed Sum** (GTS)

$$\mathrm{GTS}\big((\boldsymbol{z}_i)_{i\in[n]}\big) = \sum_{i\ge b+1} \boldsymbol{z}_i.$$

The following aggregator is called *oracle* since it requires knowing S.

- **Clipped Sum** [2] ($\mathrm{CS}_{\mathrm{He}}^{\mathrm{or}}$).

$$\mathrm{CS}_{\mathrm{He}}^{\mathrm{or}}\big((\boldsymbol{z}_i)_{i\in[n]}\big) = \sum_{i\in[n]} \mathrm{Clip}(\boldsymbol{z}_i;\tau) \quad \text{with} \quad \tau = \sqrt{\frac{1}{b}\sum_{i\in S}\|\boldsymbol{z}_i\|^2}.$$

> If $\mathcal{G}$ is fully connected, GTS-RG corresponds to NNA[1].
> $\mathrm{CS}_{\mathrm{He}}^{\mathrm{or}}$-RG corresponds to ClippedGossip[2].

Robustness Results

Theorem 1 - Convergence

If F is a (b,ρ) robust summand, and $\mu_2 \ge 2\rho b$, then for $\eta \le {}^1\!/_{\mu_{\max}}$, one step of F-RG verifies

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\boldsymbol{x}_i^1 - \overline{\boldsymbol{x}}_{\mathcal{H}}^0\|^2 \le (1 - \eta(\mu_2 - 2\rho b))\,\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^0).$$

Furthermore the additional bias is controlled

$$\|\overline{\boldsymbol{x}}_{\mathcal{H}}^1 - \overline{\boldsymbol{x}}_{\mathcal{H}}^0\|^2 \le 2\rho b\,\eta\,\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^0).$$

NB: In fully-connected graphs, $\mu_2 = |\mathcal{H}|$ and $\mu_2 \ge 2\rho b$ boils to

$$|\mathcal{B}|/|\mathcal{H}|+|\mathcal{B}| \le 1/2\rho+1.$$

Breakdown point assumption also written as $\delta := 2\rho b/\mu_2 < 1$.

Corollary

For t steps of F-RG, with $\eta = 1/\mu_{\max}$ and $\gamma = {}^{\mu_2}\!/_{\mu_{\max}}$:

$$\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^t) \le (1-\gamma(1-\delta))^t\,\mathrm{Var}_{\mathcal{H}}(\boldsymbol{x}^0) \xrightarrow[t\to\infty]{} 0,$$

Consensus is reached, and

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

Theorem 2 - Tightness

Let $b \in \mathbb{N}$. For any algorithm ALG and any $h \in \mathbb{N}$, there exists a graph $\mathcal{G}$, in which all honest nodes are neighbors to h other honest nodes, and for which $\mu_2 = 2b$, such that, for any $r<1$, ALG is not r-robust on $\mathcal{G}$.

↪ the breakdown assumption $\mu_2 \ge 2\rho b$ is tight for $\rho = 1$.

Theorem 3 - Robust Summation

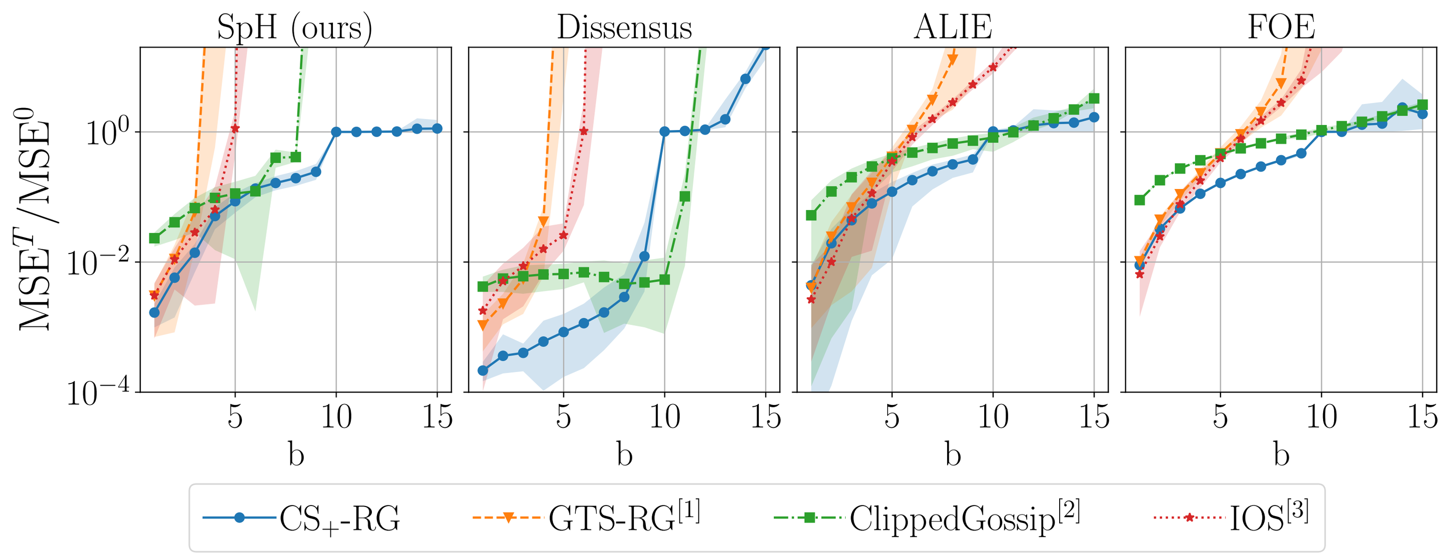
CS$_+$, GTS, $\mathrm{CS}_{\mathrm{He}}^{\mathrm{or}}$ and $\mathrm{CS}_{+}^{\mathrm{or}}$ are (b,ρ)-robust:

	CS$_+$	GTS	$\mathrm{CS}_{\mathrm{He}}^{\mathrm{or}}$	$\mathrm{CS}_{+}^{\mathrm{or}}$
ρ	2	4	4	1

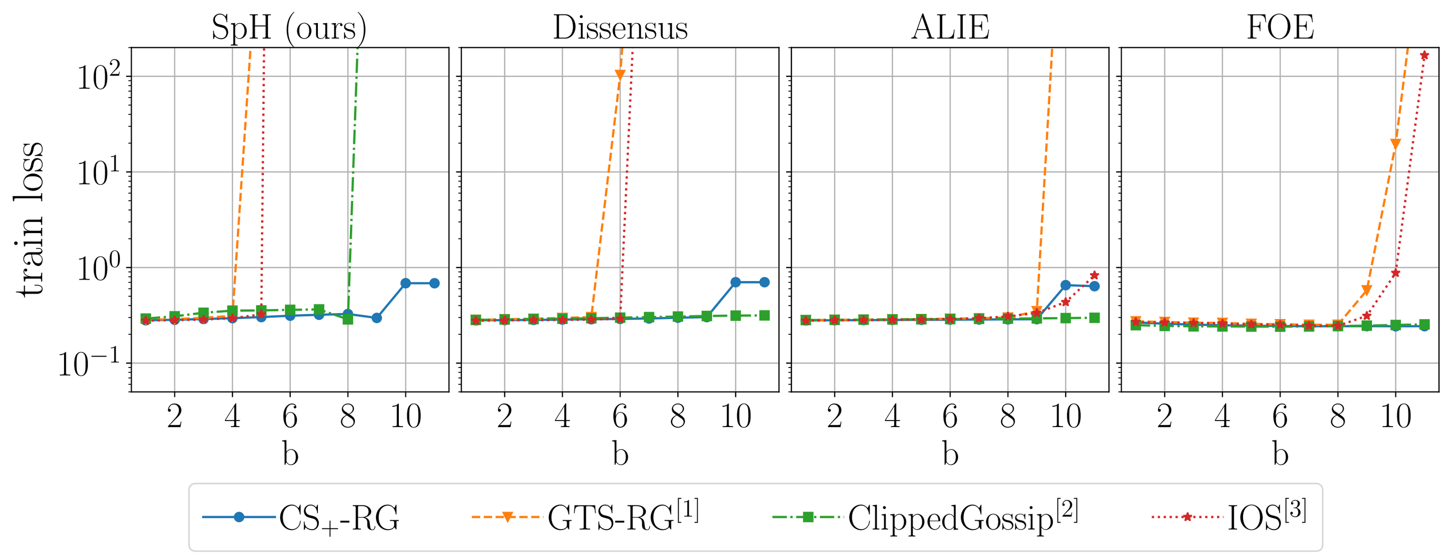
Experiments

Graph with two cliques of honest nodes *weakly* connected to each other, such that $\mu_2/2 = 8$ and $|\mathcal{H}| = 26$. Attacks tested are *Dissensus*[2], *ALIE*[4], *FOE*[5], and *Spectral Heterogeneity* (Ours).

- Average Consensus problem with gaussian initialization of the parameters.



- Optimization of a CNN on MNIST with local heterogeneity, using F-RG + momentum SGD.



More in the paper!

- Results stated with weighted graphs.
- Convergence results for D-SGD with F-RG communications.
- A new attack tailored to decentralized systems named Spectral Heterogeneity (SpH).

Bibliography.

[1] Robust collaborative learning with linear gradient overhead, Farhadkhani et al., ICML 2023
[2] Byzantine-Robust Decentralized Learning via ClippedGossip, He et. al. arxiv 2022
[3] Byzantine-resilient decentralized stochastic optimization with robust aggregation rules, Wu et. al. IEEE tsp 2023
[4] A little is enough: Circumventing defenses for distributed learning, Baruch et. al. NeurIPS 2019
[5] Fall of empires: Breaking byzantine tolerant SGD by inner product manipulation, Xie et. al., UAI, 2020