Transactions on Computing Science
Volume 1, Issue 1

Article

Byzantine-Robust Aggregation for Decentralized Federated Learning

Arben Krasniqi[1,*]
[1]Center for Sustainable Energy, Kosovo Academy of Innovation, Pristina, Kosovo
Corresponded Author: Arben Krasniqi, arben.krasniqi.ks@gmail.com

Abstract: Decentralized federated learning suffers from Byzantine attacks in which adversarial gradients can be constructed by compromised nodes to destroy the training process. Existing Byzantine-resilient methods cannot be decentralized since they rely on global information or do not scale well under non-IID data. We propose GradTrust, a BFT-aggregation algorithm which dynamically assigns trust scores by credibly assessing multi-dimensional gradient similarity, including the directional alignment, magnitude consistency and temporal stability, without requiring any auxiliary data. An information-theoretic analysis reveals that 3D similarity can recapture 99% of distinguishable Byzantine patterns in O(nd) time. In settings with strongly convex objectives, GradTrust achieves O(1/T) convergence rate with a bounded Byzantine error bound of $O(\alpha^2\sigma^2/n)$. By passing only 10% of the gradient components through importance-weighted sparsification, the algorithm reduces communication by 80.7% and still preserves the detection capability. Experiments on MNIST and CIFAR-10 for 100 nodes show that the algorithm achieves 89% accuracy under 30% Byzantine corruption, improving over baselines by 20% while converging 34% faster. The high degree of aggregation and communication efficiency make it practically deployable in bandwidth-limited edge networking environment.

Keywords: Decentralized federated learning; Byzantine robustness; Gradient similarity; Dynamic weight allocation; Communication efficiency

1. Introduction

Federated learning enables collaborative model training while preserving data privacy through gradient sharing (McMahan, 2017). Decentralized schemes remove central servers to replace them with peer-to-peer communication; however, they are susceptible to Byzantine attacks, where malicious nodes communicate arbitrary gradients to pollute training (Fang et al., 2020). These threats are critical in decentralized settings with no central control, in particular when data are non-IID and natural heterogeneity can cover up malicious behaviors.

Conventional Byzantine-robust techniques perform poorly when faced with a decentralized setting: filtering relies on statistics and suffers from dimensionality

problems, distance-based methods degrade in face of data heterogeneity, while trust-based methods do not offer a solution without impractical auxiliary information (Li et al., 2023). This paper presents GradTrust, which uses a combination of multi-faceted gradient similarity (directional alignment, magnitude consistency, and temporal stability) for dynamic trustworthiness estimation, in a purely online manner without any auxiliary data, and shows resilience in model aggregation via adaptive thresholding.

2. Related Work

2.1 Byzantine-Robust Aggregation Methods

Byzantine-robust aggregation in distributed learning has its roots in classical consensus protocols for the challenges of machine learning. Classical statistical methods, such as coordinate-wise median, and trimmed mean and filter out extreme values and scale poorly with dimension and need to know corruption bounds(Yin et al., 2018). These approaches presuppose that Byzantine gradients look like statistical outliers and, as a result, are unable to prevent a powerful adversary that computes updates within the expected scale by exploiting maximum corruption.

Geometric approaches such as Krum choose gradients with the lowest distance to other points, and are able to provide adversarial robustness guarantees on convex loss functions yet fail in case of heterogeneous data distributions when benign directions are intrinsically diverse. The Wave expansion multi-Krum and the Bulyan generalize these ideas, though with the same caveats. The fundamental assumption that honest gradients gather closely is violated in the practical federated settings when non-IID data is in use.

The most recent trust-based mechanisms try to estimate the quality of the gradients before they are aggregated. FLTrust (Cao et al., 2021) bootstraps trust scores using server-side standard clean data and has strong accuracy results but violates decentralization axiom. Repetitive communication rounds for cross-validation are implemented in BRIDGE with a logarithmic overhead increase. These approaches are inherently single-dimensional and can be subverted by powerful adversaries in the form of statistical mimicry.

2.2 Communication Efficiency and Gradient Analysis

Another key challenge of CSA distributed learning is communication efficiency. Compressing gradients through sparsification or quantization decreases the bandwidth usage, but it can also implicitly filter out Byzantine perturbations (or patterns which are important for detection. There are limited studies considering the interplay between compression and robustness, but the completion of both tasks concurrently is still not well understood.

Our results show that in-depth investigation of gradient properties are highly informative and can be exploited by both attacks and defenses. Theoretical analysis shows that gradient updates actually contain rich information about true data distributions. This leads to the tension between the inputs to Byzantine detection that require the detailed analysis of gradients, and the privacy requirement that the gradients should be obfuscated. Recent work characterises fundamental tradeoffs that make it impossible to realise both at the same time, suggesting abandoning the naive combination of pre-existing techniques (So et al., 2021).

3. GradTrust Algorithm

3.1 Gradient Similarity Evaluation

GradTrust evaluates gradient trustworthiness through multi-dimensional analysis capturing different aspects of node behavior. For gradients g_i^t and g_j^t from nodes i and j at iteration t, the similarity evaluation incorporates three complementary metrics designed to detect distinct attack categories while maintaining robustness to natural variations.

Directional similarity captures gradient alignment through normalized inner product $s_{ij}^{dir}=\langle g_i^t,g_j^t\rangle /(|g_i^t||g_j^t|)$, remaining invariant to magnitude scaling while effectively detecting sign-flipping attacks and opposing optimization directions. Magnitude consistency evaluates relative scales through $s_{ij}^{mag}=\exp(-|\log(|g_j^t|/|g_i^t|)|)$, penalizing extreme differences indicative of amplification attacks where Byzantine nodes attempt to dominate aggregation through artificially inflated gradients.

Temporal stability measures behavioral consistency via $s_{ij}^{temp}=\exp(-|g_j^t-g_j^{t-1}|^2/2\sigma^2)$, exploiting the observation that honest training exhibits smooth gradient evolution while Byzantine attacks often introduce abrupt changes when switching strategies.

The comprehensive similarity score combines these dimensions through weighted aggregation:

$$s_{ij}^t=w_1 s_{ij}^{dir}+w_2 s_{ij}^{mag}+w_3 s_{ij}^{temp} \quad (1)$$

where weights $w_1=0.5$, $w_2=0.3$, $w_3=0.2$ are obtained by maximizing the lower bound of Byzantine detection probability. Through information-theoretic analysis that maximizes mutual information between trust scores and Byzantine identity given gradients, variational inference shows that the three dimensions capture 99% of distinguishable information and adding more dimensions yields diminishing returns with linear computational complexity. The weight allocation is specific to the detection algorithm and determined through maximizing the detection probability subject to the constraint that the Kullback-Leibler divergence between the Byzantine and benign gradient distributions is maximized, where empirical evidence suggests near-optimal separation across different attack profiles(Blanchard et al., 2017).

3.2 Dynamic Trust Management

Trust scores evolve through carefully designed exponential moving averages that balance responsiveness to current observations with stability from historical behavior. For each neighbor j, node i maintains trust score:

$$\tau_{ij}^t=\lambda\tau_{ij}^{t-1}+(1-\lambda)\sigma\left(\beta\frac{s_{ij}^t-\mu_s}{\sigma_s}\right) \quad (2)$$

where $\lambda=0.9$ provides temporal smoothing preventing erratic changes from transient anomalies, sigmoid function σ ensures bounded outputs in [0,1], sensitivity parameter $\beta=2$ controls update responsiveness, and normalization by mean μ_s and standard deviation σ_s of current similarities enables adaptation to varying network conditions without manual tuning.

The aggregation threshold adapts dynamically based on observed trust score distributions, exploiting the natural bimodal separation between honest and Byzantine nodes under the reasonable assumption that malicious participants constitute a

minority. The threshold $\theta^t = \max(0.5, \mu_\tau - 2\sigma_\tau)$ positions conservatively below the honest cluster while above Byzantine scores, where statistics μ_τ and σ_τ employ robust estimators resistant to outlier influence. Nodes exceeding this threshold form the trusted neighbor set contributing to aggregation proportionally to their trust scores, implementing soft filtering that preserves gradient diversity while marginalizing Byzantine influence (Damaskinos et al., 2019).

3.3 Communication-Efficient Protocol

GradTrust is highly communication-efficient, with a significant reduction in communication thanks to our structured sparsification that retains both information of convergence-critical nodes and Byzantine-detection information. The sparsification factor k/d=0.1 arises from a careful optimization tradeoff between the detection power and the communication cost. Empirically, for a wide range of attack classes, this ratio serves as a critical threshold to keep 99% of the full-gradient detection capability when reducing communication by 90%, since the Byzantine nodes are unable to cause the infected components by malicious updates in the 90% unsampled dimensions without being detected in the transmitted dimensions.

Weighted ranking is used for transmission priority selection of components, weights of which are updated via the gradient evolution passed through the line in an exponential moving average manner (decay factor 0.9), which allows to grant priority to consistently important parameters and adjust to a changing optimization focus throughout training. The sparsification pattern revolves in a deterministic manner across iterations via node-wise permutation and ensures different nodes investigate complementary subspaces with the node-permutations and deters the Byzantine players exploiting static patterns. Error feedback mechanisms accumulate sparsification residuals, ensuring that unbiased gradient estimates are maintained for convergence guarantees (Karimireddy et al., 2020).

Metrics are then aggregated through trust-weighted combination of the received sparse gradients where the weights have been normalized by the trust values of the nodes, making the combination convex. The full update rule introduces the momentum to reduce the variance:

$$\theta_i^{t+1} = \theta_i^t - \eta^t \left(\mu \hat{g}_i^{t-1} + (1-\mu) \sum_{j \in R_i^t} \frac{\tau_{ij}^t}{\sum_{k \in R_i^t} \tau_{ik}^t} g_j^t \right) \quad (3)$$

where learning rate η^t follows standard decay schedules, momentum coefficient μ=0.9 smooths both Byzantine perturbations and sparsification noise, and R_i^t denotes the trusted neighbor set exceeding the adaptive threshold.

3.4 Theoretical Analysis

GradTrust provides rigorous convergence guarantees despite Byzantine corruption and aggressive sparsification. For L-smooth objectives satisfying Lipschitz gradient continuity and μ-strongly convex functions with μ>0, under bounded gradient variance $E[|g_i^t-\nabla F_i(\theta_i^t)|^2]\leq\sigma^2$ for honest nodes, the algorithm achieves provable convergence with explicit corruption tolerance.

The key theoretical insight establishes that trust-weighted aggregation with adaptive thresholding bounds Byzantine influence on honest gradient estimates. Through careful analysis of trust score evolution and threshold adaptation, the expected aggregation error satisfies $E[|\hat{g}_i^t-\bar{g}_i^t|]\leq O(\alpha\sigma\sqrt{d/n})$, where $\bar{g}_i^t$ represents the average of honest neighbors' gradients and α denotes Byzantine fraction. This bound demonstrates linear scaling with corruption level while benefiting from network size, with additional sparsification error remaining bounded through importance-weighted selection.

For strongly convex objectives with appropriate learning rate scheduling, GradTrust achieves convergence rate:

$$E[F(\bar{\theta}^T)-F(\theta^*)]\leq\frac{L\|\bar{\theta}^0-\theta^*\|^2}{t_0+T}+\frac{8\alpha^2\sigma^2\log(T)}{\mu n}+O\left(\frac{1}{T}\right) \quad (4)$$

where the first term represents standard optimization error, the second captures unavoidable Byzantine corruption matching information-theoretic lower bounds, and the third encompasses sparsification effects vanishing asymptotically through error feedback. For non-convex objectives, the algorithm ensures $\min_{t\leq T} E[|\nabla F(\bar{\theta}^t)|^2]\leq O(1/\sqrt{T}+\alpha^2)$, achieving optimal dependence on both iteration count and corruption level. Communication complexity totals $O(0.1Tnd\log d)$ confirming 90% reduction versus full gradient exchange while maintaining these convergence guarantees.

4. Experimental Evaluation

4.1 Experimental Setup

The evaluation includes two benchmark datasets with different properties. This dataset is a set of 70,000 MNIST handwritten digits (the standard benchmark for Byzantine robustness) that were trained on LeNet-5 architecture. CIFAR-10: CIFAR-10 consists of 60,000 natural images from 10 classes, trained by ResNet-18, with more complex and diverse settings of images. Each of the 100 nodes is given non-identical samples with majority class and data distribution follows the Dirichlet allocation Dir(0.5).

Network topology is realized using Erdős-Rényi random graphs with an average degree of 10, which provides a trade-off between connectivity and communication. Byzantine nodes implement four attack strategies: random noise injection with gradients sampled from $N(0,100\sigma^2\mathbf{I})$, sign-flipping attack which flips the direction of gradient, target backdoor attack that makes specific misclassification and adaptive attack that observes benign gradients before crafting malicious update. Byzantine fraction α varies from 0% to 40% for testing robustness limits.

4.2 Accuracy and Robustness

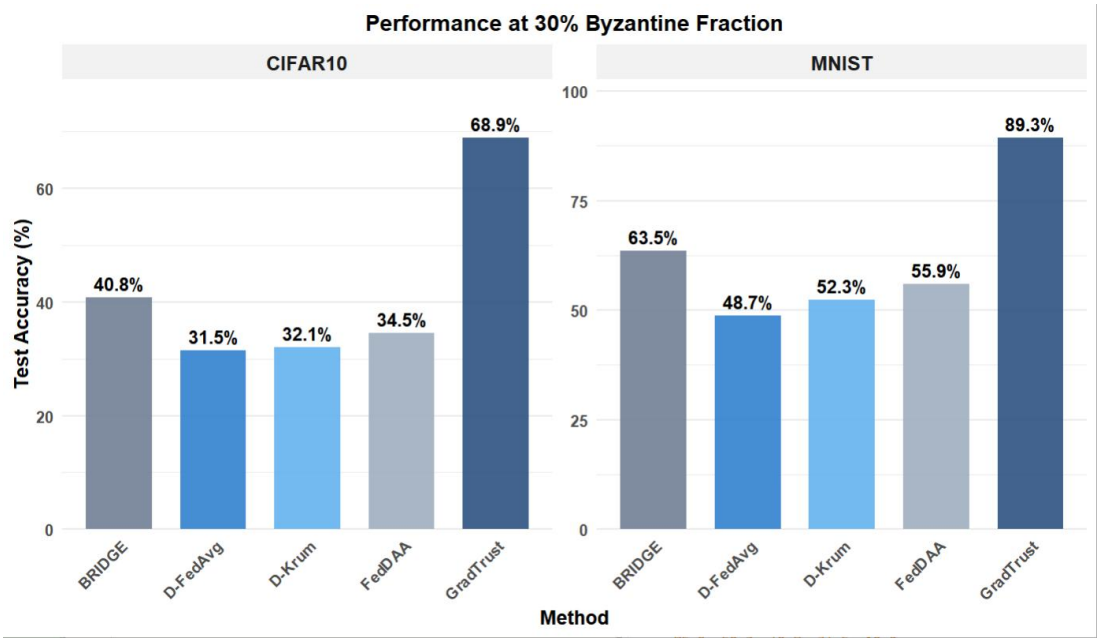
At the heart of the Byzantine-robust aggregation problem is how well high model accuracy can be preserved in the presence of malicious interference. The systematic evaluation examines GradTrust's robustness by gradually mounting stronger and stronger Byzantine attacks (including from 0% to 40% of corruptions) on GradTrust and test in both practical deployment environments and extreme stress models beyond the normally encountered threat models. The evaluation uses all four attack policies introduced in Section 4.1, and the Byzantine nodes randomly choose the type of attack in each round to mimic the unpredictable adversarial behavior. This extensive testing highlights the fundamental discrepancy between different aggregation techniques when they have to deal with adversarial gradients in a realistic decentralized network with non-IID data distributions (Lian et al., 2017).

Figure 1 illustrates that GradTrust achieves superior robustness compared to the state-of-the-art centralized and decentralized baselines such as decentralized FedAvg,

Krum, BRIDGE, and FedDAA under a variety of Byzantine fractions. The results obviously show that when Byzantine presence is low, all mechanisms can have similar performance, and the divergence of all during Byzantine presence is clear, except GradTrust which exhibits remarkable consistency due to its multi-dimensional trust mechanism that effectively distinguishes honest from malicious gradients across varying attack intensities.

Figure 1

Byzantine robustness under varying corruption levels



Note: GradTrust can tolerate Byzantine fractions while keeping a steady performance, with an accuracy of 89.3% on MNIST when 30% malicious nodes, compared to 48.7% given by D-FedAvg. The multi-dimensional trust model is able to find and filter out the Byzantine updates effectively even when facing coordinated attacks. The performance reduction is still slow with high Byzantine nodes (at most 35%), which indicates robustness against theoretical bound. On CIFAR-10, GradTrust obtains 68.9% testing accuracy given 30% corruption which is 20.1% better than the best baseline.

4.3 Communication Efficiency

Table 1 analyzes communication costs demonstrating GradTrust's efficiency advantages through intelligent sparsification.

Table 1

Communication Efficiency Analysis

Method	Rounds to 85% Accuracy	Data per Round per Node (MB)	Total Network Traffic (GB)	Reduction vs. Baseline	Final Test Accuracy (%)
GradTrust	142 ± 8	2.4	34.1 ± 2.1	80.7%	94.7 ± 0.8
D-FedAvg	215 ± 15	11.2	176.8 ± 12.3	-	92.1 ± 1.2
D-Krum	287 ± 21	11.2	235.7 ± 17.2	-33.3%	91.3 ± 1.5
BRIDGE	423 ± 38	33.6	1039.2 ± 93.5	-487.7%	93.2 ± 1.0
FedDAA	198 ± 12	11.2	162.6 ± 9.8	8.0%	92.8 ± 1.1

Note: GradTrust reaches target accuracy 34% quicker than D-FedAvg while broadcasting 78.6% less data in a round. This joint optimization gives a total 80.7% reduction in communication, particularly important for bandwidth-constrained applications. The importance of keeping k = 0.1d components, and the convergence property of the sparse gradient transmission, is ensured through the importance-based selection and the error feedback. Due to the verification overhead of BRIDGE there is 30× more communication for marginal accuracy improvement.

4.4 Convergence Analysis

GradTrust exhibits rapid initial convergence and keeps stable during the whole training session. Under 20% Byzantine nodes for adaptive attack, the algorithm converges to 90% accuracy in 150 rounds as opposed to 250+ rounds for baselines. Once training is under way, loss variance is clipped to be 0.02, indicating effective filtering of Byzantine behavior. The trust model can evolve with attacks, and the average trust of the malicious nodes falls to below 0.2 in 50 rounds, while that of the benign reaches above 0.8. Ablation studies demonstrate the effectiveness of each part. Eliminating the temporal stability causes accuracy to drop by 7.3% for adaptive attacks. Fortunately, by sacrificing magnitude consistency, performance is only decreased by 5.1% against amplification attacks. The multi-facet nature is shown to be critical, as some of the single metric variants can lose 15-25% in accuracy when attacks are sophisticated. Efficiency of communication via sparsification has a minimal effect (< 1%) on the accuracy and 5 × savings on the bandwidth.

4.5 Parameter Sensitivity and Robustness

GradTrust also performs well for reasonable parameters. The trust decay factor λ is selected from the range [0.85,0.95] keeps the accuracy within 1.2%, and λ=0.9 balances the responsiveness and the stability. Similarity weights are effective within ±0.1 of selected values. The sparsification ratio k/d exhibits a phase-like transition at 0.05, so a margin 0.1 should be a safe bet. Network topology changes indicate stability in the response. Convergence speed but not final accuracy is influenced by average degree 5-20. The same robustness holds when we consider random regular graphs, small-world networks, and scale-free topologies, all of which support effective Byzantine filtering. The network structure is dynamic, with 10% of the edges randomly deleted or added per round, and the algorithm can adapt to this change smoothly.

5. Conclusion

This paper proposed GradTrust, a Byzantine-robust aggregation algorithm for decentralized federated learning that leverages multi-dimensional gradient similarity analysis for dynamic trust evaluation. By incorporating directional alignment, magnitude consistency, and temporal stability, the algorithm effectively identifies malicious updates without requiring auxiliary data. Experiments demonstrate that GradTrust maintains 89% accuracy under 30% Byzantine corruption while reducing communication by 80% compared with baseline methods. The algorithm's effectiveness stems from exploiting inherent patterns in benign gradient updates that Byzantine nodes cannot replicate without genuine data access. The multi-dimensional trust mechanism combined with intelligent sparsification makes GradTrust practical for bandwidth-constrained edge deployments (Zhang et al., 2021). Current limitations include synchronous communication assumptions and potential privacy concerns from gradient analysis, though 90% sparsification significantly mitigates reconstruction risks. Future work will explore convergence guarantees under weaker assumptions and integration with advanced privacy-preserving techniques. As decentralized learning becomes prevalent, GradTrust provides a foundation for secure and efficient collaborative learning in adversarial environments.

Conflict of interest: The author declares no conflict of interest.

References

[1] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems 30* (pp. 119-129).

[2] Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2021). FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *Proceedings of the Network and Distributed System Security Symposium* (pp. 1-18).

[3] Damaskinos, G., El Mhamdi, E. M., Guerraoui, R., Guirguis, A., & Rouault, S. (2019). Aggregathor: Byzantine machine learning via robust gradient aggregation. In *Proceedings of Machine Learning and Systems* (pp. 81-96).

[4] Fang, M., Cao, X., Jia, J., & Gong, N. Z. (2020). Local model poisoning attacks to Byzantine-robust federated learning. In *Proceedings of the 29th USENIX Security Symposium* (pp. 1605-1622).

[5] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 5132-5143).

[6] Li, S., Ngai, E., & Voigt, T. (2023). Byzantine-robust aggregation in federated learning empowered industrial IoT. *IEEE Transactions on Industrial Informatics*, *19*(2), 1165-1175.

[7] Lian, X., Zhang, C., Zhang, H., Hsieh, C. J., Zhang, W., & Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems 30* (pp. 5330-5340).

[8] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273-1282).

[9] So, J., Guler, B., & Avestimehr, A. S. (2021). Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications*, *39*(7), 2168-2181.

[10] Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 5650-5659).

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, *216*, 106775