LASA: BYZANTINE-ROBUST DECENTRALIZED OPTIMIZATION VIA LOCAL-LOSS AWARE WEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

A key challenge in Byzantine-robust decentralized learning is to mitigate the adverse effects of Byzantine updates while preserving beneficial information exchange. To this end, we introduce aggregation budgets that provide loss-based control of Byzantine influence by bounding the weighted excess local loss relative to the self-candidate. We then propose a novel Local-Loss Aware Softmax Aggregator (LASA) with explicit aggregation budgets. Based on local-loss evaluations, LASA assigns positive weights to all candidates without prior Byzantine information, balancing robustness with opportunities for useful cooperation. For LASA and any other aggregator admitting such budgets, we establish global stationarity bounds for smooth convex objectives under a relative gradient heterogeneity condition without assuming an honest majority, and obtain sharper guarantees under an additional quadratic growth condition. We further provide sufficient conditions under which LASA admits a strictly tighter stationarity bound than self-only training. Experiments on MNIST and CIFAR-10 demonstrate the effectiveness of LASA across several communication graphs and Byzantine attacks.

1 INTRODUCTION

Modern machine learning increasingly relies on distributed data and computational resources to train a global model. Centralized distributed optimization depends on a server to aggregate local updates, while decentralized architectures use peer-to-peer communication over a graph to alleviate server-side communication and computation bottlenecks. However, these distributed systems are vulnerable to Byzantine nodes (Lamport et al., 1982) arising from malicious attacks, hardware malfunctions, or data poisoning. Such vulnerability can undermine the training process (Guerraoui et al., 2024) and thus necessitate Byzantine-robust algorithms.

1.1 MOTIVATION AND RELATED WORK

Much of the existing work considers server-based architectures. One line of research relies on aggregators inspired by robust statistics (Blanchard et al., 2017; Yin et al., 2018; Pillutla et al., 2022; Yi et al., 2024; Murata et al., 2024; Fang et al., 2025; Luan et al., 2025), while another exploits clipping or normalization to mitigate Byzantine effects (Karimireddy et al., 2021; Allouah et al., 2025; Zuo et al., 2025). Some unified robustness conditions are proposed to analyze these algorithms (Karimireddy et al., 2021; Farhadkhani et al., 2022; Allouah et al., 2023; 2024). Notably, a few approaches propose to use trusted data at the server to assess updates (Xie et al., 2019; Cao et al., 2021; Park et al., 2021; Molodtsov et al., 2026). However, all these methods rely on a central server to coordinate training. Methods using trusted data additionally require a clean evaluation dataset at the server, which may be inaccessible and raise privacy concerns (Fang et al., 2025).

In decentralized optimization, each honest node must aggregate potentially unreliable neighbor updates without central coordination while preserving useful information exchange. A prevalent approach is to identify and filter suspicious updates based on a prespecified bound on Byzantine participation, expressed as a number, fraction, or total weight (Fang et al., 2022; Wu et al., 2023; Farhadkhani et al., 2023; Guo et al., 2022; Kuwaranancharoen et al., 2024; Yang & Ghaderi, 2024; Kuwaranancharoen & Sundaram, 2025; Gaucher et al., 2025). These methods require sufficient neighborhood redundancy to accommodate this bound, limiting their applicability in sparse networks

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

or neighborhoods with a high Byzantine fraction. Moreover, overly conservative bounds may lead to excessive filtering and weaken effective mixing, whereas retaining more neighbors permits malicious influence. Moving beyond fixed bounds, Fang et al. (2024) adaptively exclude neighbors using a time-varying distance criterion, but their analysis relies on homogeneous data. Other approaches mitigate Byzantine influence through clipping (He et al., 2023; Han et al., 2025), but their theoretical guarantees depend on inaccessible quantities of the honest subnetwork.

The decentralized methods discussed above primarily assess received candidates through geometric relationships among model parameters, which may leave room for malicious yet geometrically plausible candidates. Local loss provides a complementary criterion by directly evaluating candidate performance on the receiving node's objective and can naturally be computed in decentralized learning using each honest node's own trusted data. Few methods exploit this information: Guo et al. (2022) use local losses as a secondary criterion after distance-based filtering, and Li et al. (2020) develop inverse-loss weighting for multitask learning with task-specific objectives, whereas we consider a shared global objective. Both of them exclude candidates with higher local loss than that of the receiver's own candidate. Under data heterogeneity, this exclusion can limit useful cooperation when optimizing a shared global objective, since a candidate with higher local loss may still help reduce the global objective value. All of these observations raise a central question:

Without requiring prior information about Byzantine nodes, can local-loss evaluations control Byzantine influence while retaining all candidates to preserve opportunities for useful cooperation?

1.2 Our Contributions

We address this question by introducing aggregation budgets and developing the Local-Loss Aware Softmax Aggregator (LASA) with explicit budget guarantees. Our main contributions are:

- We introduce deterministic aggregation budgets $\{\beta_{i,t}\}$ that bound the weighted excess local loss relative to the self-candidate: $\sum_{j\in\mathcal{S}_i} \omega_{ij}^t(\ell_{ij}^t - \ell_{ii}^t) \le \beta_{i,t}$. Here, the set $\mathcal{S}_i$ contains node i and its neighbors, while ℓ_{ij}^t and ω_{ij}^t denote the local loss evaluated by node i on candidate j and its aggregation weight, respectively. The method of Li et al. (2020) admits zero budgets as a special case. For any aggregator admitting such budgets, we derive convergence bounds under smoothness, convexity, and relative gradient heterogeneity, with explicit dependence on the aggregation budgets. We obtain sharper bounds under an additional quadratic growth assumption. Notably, these results do not assume an honest majority in each honest node's neighborhood.
- We propose LASA, which assigns positive weights $\omega_{ij}^t \propto \exp(-\tau_t \ell_{ij}^t)$ to all candidates without requiring prior Byzantine information or knowledge of the global network topology. This makes LASA readily applicable across a broader range of network configurations. An entropy-regularized characterization shows how the inverse temperature τ_t balances robustness and cooperation. We prove that LASA admits explicit aggregation budgets and further establish sufficient conditions under which it attains a strictly tighter global stationarity bound than self-only training.
- We evaluate LASA on MNIST and CIFAR-10 across multiple communication graphs and Byzantine attacks. Our results demonstrate that LASA consistently achieves excellent performance and attains the highest worst-node accuracy in most tested settings. Sensitivity studies on MNIST further show that its performance is stable across a range of inverse temperatures and with small batches for local-loss evaluation.

Paper organization. Section 2 presents the decentralized stochastic gradient descent framework and defines aggregation budgets. Section 3 derives convergence bounds for aggregators admitting such budgets. Section 4 introduces LASA and establishes its theoretical guarantees. Section 5 presents the empirical results. Section 6 discusses our findings and concludes the paper.

2 Aggregation Budgets for Decentralized SGD

In this section, we present the decentralized SGD (DSGD) framework under Byzantine attacks and define aggregation budgets. We consider n nodes connected by a fixed undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

The node set $\mathcal{V}$ is partitioned into honest and Byzantine sets $\mathcal{H}$ and $\mathcal{B}$, with $h := |\mathcal{H}| \geq 1$. Each honest node i holds a private dataset $\mathcal{D}_i$ defining a local loss $f_i : \mathbb{R}^d \to \mathbb{R}$. Its neighbor set including itself is $\mathcal{S}_i := \{i\} \cup \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$ with $m_i := |\mathcal{S}_i|$. Without prior information about Byzantine nodes, each honest node seeks to find a minimizer of the global objective

$$F(x) := \frac{1}{h} \sum_{i \in \mathcal{H}} f_i(x). \tag{2.1}$$

Most existing work follows the DSGD framework (Koloskova et al., 2020). To be specific, each honest node i starts from an initial model x_i^0. At round t, honest node i forms a local candidate $\tilde{x}_i^t = x_i^t - \gamma_t g_i^t$ and sends the candidate to its neighbors, where $\gamma_t > 0$ is the prescribed deterministic stepsize and g_i^t is a stochastic gradient of f_i evaluated at x_i^t. Let $\tilde{x}_{ij}^t$ denote the candidate received by node i from sender $j \in \mathcal{S}_i$. Honest senders satisfy $\tilde{x}_{ij}^t = \tilde{x}_j^t$, including $\tilde{x}_{ii}^t = \tilde{x}_i^t$, whereas Byzantine senders may transmit adaptive and receiver-dependent candidates. Having received the candidates, each honest node i then applies an aggregator to obtain its next iterate x_i^{t+1}. Appendix A summarizes the resulting DSGD framework.

We focus on weighted-average aggregators of the following form

$$x_i^{t+1} = \sum_{j \in \mathcal{S}_i} \omega_{ij}^t \tilde{x}_{ij}^t, \tag{2.2}$$

where ω_{ij}^t is the weight that node i assigns to the candidate from node j. In Byzantine-free DSGD, these weights are typically fixed in advance based on the communication graph. In the presence of Byzantine nodes, however, a fixed positive weight on a Byzantine message enables it to arbitrarily perturb the aggregation. To limit such influence, each honest node can evaluate the received candidates on its local objective. This motivates introducing aggregation budgets that bound the weighted excess local loss relative to the self-candidate.

Definition 1 (Aggregation budgets)**.** Let $\ell_{ij}^t := f_i(\tilde{x}_{ij}^t) \in \mathbb{R}$ denote the locally evaluated candidate loss. A weighted-average aggregator of the form (2.2) admits deterministic **aggregation budgets** $\{\beta_{i,t}\}_{i \in \mathcal{H}, t \geq 0}$ if all $\beta_{i,t}$ are finite and trajectory-independent, and the weights in (2.2) satisfy

$$\omega_{ij}^t \geq 0, \quad \sum_{j \in \mathcal{S}_i} \omega_{ij}^t = 1, \tag{2.3}$$

$$\sum_{j \in \mathcal{S}_i} \omega_{ij}^t (\ell_{ij}^t - \ell_{ii}^t) \leq \beta_{i,t}, \tag{2.4}$$

for every honest node $i \in \mathcal{H}$ and round $t \geq 0$ over every admissible trajectory.

Condition (2.3) keeps the update in the convex hull of the received candidates, including the honest self-candidate. Condition (2.4) bounds the weighted excess local loss relative to the self-candidate, thereby providing loss-based control of Byzantine influence. These conditions also yield a direct guarantee on the local loss after aggregation. In particular, if f_i is convex, then

$$f_i(x_i^{t+1}) \leq \sum_{j \in \mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t \leq f_i(x_i^t - \gamma_t g_i^t) + \beta_{i,t}.$$

Thus, aggregation preserves the local-loss guarantee of the gradient step up to an additional budget $\beta_{i,t}$. To obtain a zero budget, it suffices to assign nonnegative, normalized weights supported on $\mathcal{S}_i^{+,t} := \{j \in \mathcal{S}_i : \ell_{ij}^t \leq \ell_{ii}^t\}$. With positive losses, this includes the inverse-loss rule (Li et al., 2020):

$$\omega_{ij}^t = \begin{cases} \dfrac{1/\ell_{ij}^t}{\sum_{k \in \mathcal{S}_i^{+,t}} 1/\ell_{ik}^t}, & \text{if } j \in \mathcal{S}_i^{+,t}, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 1 also permits positive weights on higher-loss candidates to preserve opportunities for useful cooperation, as illustrated by LASA in Section 4. We next specify how these budgets should be controlled to obtain convergence guarantees in Section 3.

Notation. We write $\|\cdot\|$ for the Euclidean norm, $|\mathcal{S}|$ for the cardinality of a set $\mathcal{S}$, and $\mathbb{E}[\cdot]$ for expectation. $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ denotes conditional expectation given the history available at the beginning of iteration t. Define $F^\star = \inf_x F(x)$, $f_i^\star = \inf_x f_i(x)$, and the initial local objective gaps $e_{i,0} = f_i(x_i^0) - f_i^\star$ for $i \in \mathcal{H}$. For $m \geq 1$, let $\mathcal{P}_{m-1} = \{p \in \mathbb{R}^m : p \geq 0,\ \sum_{j=1}^m p_j = 1\}$ denote the probability simplex. The notation $\mathcal{O}(\cdot)$ hides constants independent of the iteration horizon.

3 Convergence Analysis with Aggregation Budgets

Throughout this section, we consider DSGD using a weighted-average aggregator admitting aggregation budgets $\{\beta_{i,t}\}$ as in Definition 1. We establish convergence guarantees under the following assumptions, none of which requires information about the honest network or an honest majority.

Assumption 1 (Stochastic gradients)**.** *For every honest node $i \in \mathcal{H}$ and every $t \geq 0$, the stochastic gradient g_i^t of f_i at x_i^t satisfies*

$$\mathbb{E}\left[g_i^t \mid \mathcal{F}_t\right] = \nabla f_i(x_i^t), \qquad \mathbb{E}\left[\left\|g_i^t - \nabla f_i(x_i^t)\right\|^2 \mid \mathcal{F}_t\right] \leq \sigma^2. \tag{3.1}$$

Assumption 2 (Smooth convex objectives)**.** *For each $i \in \mathcal{H}$, f_i is convex, differentiable, bounded below, and has an L-Lipschitz gradient.*

Assumption 3 (Relative gradient heterogeneity)**.** *There exist constants $\delta_1 \geq 0$ and $\delta_2 \in [0, 1)$ such that for all $i \in \mathcal{H}$ and $x \in \mathbb{R}^d$,*

$$\|\nabla f_i(x) - \nabla F(x)\|^2 \leq \delta_1^2 + \delta_2^2 \|\nabla F(x)\|^2. \tag{3.2}$$

The constants δ_1 and δ_2 capture absolute and relative gradient discrepancies, respectively. Setting $\delta_2 = 0$ recovers the bounded gradient heterogeneity condition of prior work (Wu et al., 2023; Yang & Ghaderi, 2024; Han et al., 2025).

Assumption 4 (Quadratic growth condition)**.** *For each $i \in \mathcal{H}$, the minimizer set $\mathcal{X}_i^\star = \arg\min_{x \in \mathbb{R}^d} f_i(x)$ is nonempty. There is a constant $\mu_{\mathrm{loc}} > 0$ such that for all $i \in \mathcal{H}$ and $x \in \mathbb{R}^d$,*

$$f_i(x) - f_i^\star \geq \frac{\mu_{\mathrm{loc}}}{2} \inf_{y \in \mathcal{X}_i^\star} \|x - y\|^2. \tag{3.3}$$

Moreover, F has a unique minimizer $x^\star$, and there exists a constant $\mu_F > 0$ such that for all $x \in \mathbb{R}^d$,

$$F(x) - F^\star \geq \frac{\mu_F}{2} \|x - x^\star\|^2. \tag{3.4}$$

Quadratic growth is strictly weaker than strong convexity (Necoara et al., 2019). Quadratic growth and uniqueness of the global minimizer allow sharper bounds on objective gaps and model disagreement.

We now establish convergence guarantees for the weighted-average aggregator admitting deterministic aggregation budgets $\{\beta_{i,t}\}$. The proofs are deferred to Appendix B. Fix

$$\rho > \frac{\delta_2^2}{1 - \delta_2^2}, \qquad \chi_\rho = \frac{1 - (1 + \rho^{-1})\delta_2^2}{1 + \rho} > 0, \qquad \Gamma_T = \sum_{t=0}^{T-1} \gamma_t. \tag{3.5}$$

Define the average of honest models, model disagreement, and disagreement among global gradients evaluated at honest models as follows:

$$\bar{x}^t = \frac{1}{h} \sum_{i \in \mathcal{H}} x_i^t, \quad \mathcal{C}_t = \frac{1}{h} \sum_{i \in \mathcal{H}} \left\|x_i^t - \bar{x}^t\right\|^2, \quad \mathcal{G}_t = \frac{1}{h} \sum_{i \in \mathcal{H}} \left\|\nabla F(x_i^t) - \frac{1}{h} \sum_{j \in \mathcal{H}} \nabla F(x_j^t)\right\|^2. \tag{3.6}$$

Theorem 3.1. *Suppose that Assumptions 1–3 hold. Fix an integer $T \geq 1$ and stepsizes γ_t satisfying $0 < \gamma_t \leq 1/L$. Let $R \in \{0, 1, \ldots, T-1\}$ be a random variable independent of the algorithmic randomness and distributed according to $\Pr(R = t) = \gamma_t / \Gamma_T$. Then, for every $i \in \mathcal{H}$, we have*

$$\mathbb{E}\left[\left\|\nabla F(x_i^R)\right\|^2\right] \leq \frac{2e_{i,0} + L\sigma^2 \sum_{t=0}^{T-1} \gamma_t^2 + 2\sum_{t=0}^{T-1} \beta_{i,t}}{\chi_\rho \Gamma_T} + \frac{\delta_1^2}{\rho \chi_\rho}, \tag{3.7}$$

$$\mathbb{E}\left[\mathcal{G}_R\right] \leq \frac{2\sum_{j \in \mathcal{H}} e_{j,0} + Lh\sigma^2 \sum_{t=0}^{T-1} \gamma_t^2 + 2\sum_{j \in \mathcal{H}} \sum_{t=0}^{T-1} \beta_{j,t}}{h \chi_\rho \Gamma_T} + \frac{\delta_1^2}{\rho \chi_\rho}. \tag{3.8}$$

In particular, choose $\rho = 1/(1 - \delta_2^2)$ and $\gamma_t = 1/(L\sqrt{T})$, and suppose that for each $i \in \mathcal{H}$, there exists a constant $q_i \geq 0$ such that $\beta_{i,t} \leq q_i / T$ for all $0 \leq t < T$. Then, for every $i \in \mathcal{H}$, we have

$$\max\left\{\mathbb{E}\left[\|\nabla F(x_i^R)\|^2\right], \mathbb{E}[\mathcal{G}_R]\right\} \leq \mathcal{O}\left(\frac{\max_{j \in \mathcal{H}}\{e_{j,0}\} + \max_{j \in \mathcal{H}}\{q_j\} + \sigma^2}{\sqrt{T}} + \frac{\delta_1^2}{(1 - \delta_2)^2}\right). \tag{3.9}$$

Equations (3.7) and (3.8) quantify how initial local objective gaps $e_{i,0}$, the stochastic gradient noise variance bound σ^2, the budgets $\{\beta_{i,t}\}$, and heterogeneity affect global stationarity and global gradient disagreement. The $\mathcal{O}(T^{-1/2})$ transient rate and $\mathcal{O}(\delta_1^2/(1-\delta_2)^2)$ heterogeneity floor in (3.9) match the corresponding orders in existing decentralized Byzantine-robust guarantees (Wu et al., 2023; Yang & Ghaderi, 2024; Han et al., 2025; Gaucher et al., 2025). This heterogeneity floor is generally unavoidable in Byzantine-robust optimization (Karimireddy et al., 2022; Murata et al., 2024). However, these global gradient disagreement guarantees do not by themselves control model disagreement. We next address this under the additional quadratic growth condition, with explicit statements and proofs given in Appendix B.3.

Theorem 3.2. *Suppose that Assumptions 1–4 hold and let $\mu_{\mathrm{PL}} = \min\{\mu_{\mathrm{loc}}/4, L\}$. The following results hold:*

(i) Constant stepsize. *Let $\gamma_t \equiv \gamma \in (0, 1/L]$. Assume that for each $i \in \mathcal{H}$, there exists a constant $\bar{\beta}_i \geq 0$ such that $\beta_{i,t} \leq \bar{\beta}_i$ for all $t \geq 0$. Then, for every $i \in \mathcal{H}$ and integer $T \geq 1$, we have*

$$\mathbb{E}[F(x_i^T) - F^\star] \leq \mathcal{O}\left(\frac{(1-\mu_{\mathrm{PL}}\gamma)^T e_{i,0}}{(1-\delta_2)^2} + \frac{\gamma\sigma^2 + \bar{\beta}_i/\gamma + \delta_1^2}{(1-\delta_2)^2}\right), \tag{3.10}$$

$$\mathbb{E}[\mathcal{C}_T] \leq \mathcal{O}\left(\frac{(1-\mu_{\mathrm{PL}}\gamma)^T \sum_{j\in\mathcal{H}} e_{j,0}}{(1-\delta_2)^2 h} + \frac{\gamma\sigma^2 + \frac{1}{h\gamma}\sum_{j\in\mathcal{H}} \bar{\beta}_j + \delta_1^2}{(1-\delta_2)^2}\right). \tag{3.11}$$

(ii) Diminishing stepsizes. *For all $t \geq 0$, choose the following stepsizes and suppose that the aggregation budgets satisfy*

$$\gamma_t = \frac{2}{\mu_{\mathrm{PL}}(t+t_0)}, \qquad \beta_{i,t} \leq \frac{q_i}{(t+t_0)^2}, \qquad t_0 \geq \max\left\{2, \frac{2L}{\mu_{\mathrm{PL}}}\right\}, \tag{3.12}$$

where $q_i \geq 0$ for each $i \in \mathcal{H}$. Then, for every $i \in \mathcal{H}$ and integer $T \geq 1$, we have

$$\mathbb{E}[F(x_i^T) - F^\star] \leq \mathcal{O}\left(\frac{e_{i,0} + q_i + \sigma^2}{(1-\delta_2)^2(T+t_0-1)} + \frac{\delta_1^2}{(1-\delta_2)^2}\right), \tag{3.13}$$

$$\mathbb{E}[\mathcal{C}_T] \leq \mathcal{O}\left(\frac{\sigma^2 + \frac{1}{h}\sum_{j\in\mathcal{H}}(e_{j,0} + q_j)}{(1-\delta_2)^2(T+t_0-1)} + \frac{\delta_1^2}{(1-\delta_2)^2}\right). \tag{3.14}$$

Under the quadratic growth assumption, constant stepsizes yield geometric decay to a neighborhood determined by the stochastic gradient noise variance bound σ^2, the budget bounds $\bar{\beta}_i$, and heterogeneity. With diminishing stepsizes and quadratically decaying budgets, both the global objective gap and model disagreement have an $\mathcal{O}(T^{-1})$ transient rate and an $\mathcal{O}(\delta_1^2/(1-\delta_2)^2)$ heterogeneity floor.

4 LASA: Local-Loss Aware Softmax Aggregator

In this section, we introduce LASA, which uses local-loss evaluations and requires no prior information about Byzantine nodes. We establish its entropy-regularized interpretation and derive explicit aggregation budgets. We further derive sufficient conditions for LASA to achieve a strictly tighter global stationarity bound than self-only training. Details and proofs are provided in Appendix C.

LASA assigns continuous aggregation weights based on differences in local model performance. At round t, honest node i evaluates the received candidates using its local objective and defines each candidate's local-loss improvement relative to its own candidate as $\Delta_{ij}^t = \ell_{ii}^t - \ell_{ij}^t$. LASA favors larger improvements while retaining all candidates through the following softmax weights:

$$\omega_{ij}^t = \frac{\exp(\tau_t \Delta_{ij}^t)}{\sum_{k\in\mathcal{S}_i}\exp(\tau_t \Delta_{ik}^t)} = \frac{\exp(-\tau_t \ell_{ij}^t)}{\sum_{k\in\mathcal{S}_i}\exp(-\tau_t \ell_{ik}^t)}, \tag{4.1}$$

where $\{\tau_t\}_{t\geq 0}$ is a prescribed deterministic sequence of positive inverse temperatures controlling sensitivity to loss improvements Δ_{ij}^t. Node i forms its next iterate x_i^{t+1} as the weighted average (2.2) using weights in (4.1). Each node can execute LASA using only locally evaluated losses without prior Byzantine information or knowledge of the global network topology, which supports deployment across diverse network configurations. Algorithm 2 in Appendix A summarizes LASA at an honest node i. The LASA weights in (4.1) admit the following entropy-regularized characterization.

Proposition 4.1. *For each $i \in \mathcal{H}$ and $t \geq 0$, the LASA weight vector $\omega_i^t = (\omega_{ij}^t)_{j \in \mathcal{S}_i}$ in* (4.1) *is the unique solution of*

$$\max_{p \in \mathcal{P}_{m_i - 1}} \left\{ \sum_{j \in \mathcal{S}_i} p_j \Delta_{ij}^t + \frac{1}{\tau_t} H_i(p) \right\}, \tag{4.2}$$

where the coordinates of p are reindexed by $\mathcal{S}_i$ and $H_i(p) = -\sum_{j \in \mathcal{S}_i} p_j \log p_j$ is the entropy of p, with the convention $0 \log 0 = 0$.

In (4.2), the first term rewards local-loss improvement, while entropy promotes weight dispersion. For fixed candidate losses, the weights approach the uniform distribution as $\tau_t \to 0$ and concentrate uniformly on the maximizers of Δ_{ij}^t (or minimizers of ℓ_{ij}^t) as $\tau_t \to \infty$. Thus, τ_t balances local-loss preference and weight dispersion. The next result provides explicit aggregation budgets for LASA.

Proposition 4.2. *For every $i \in \mathcal{H}$ and $t \geq 0$, LASA admits deterministic aggregation budgets*

$$\beta_{i,t} = \frac{\log m_i}{\tau_t}. \tag{4.3}$$

The guarantees in Section 3 therefore apply to LASA with suitable choices of τ_t. For any $\kappa > 0$ and fixed integer $T \geq 1$, choosing $\tau_t \geq T/\kappa$ satisfies the budget requirement of Theorem 3.1 with $q_i = \kappa \log m_i$. A constant $\tau_t \equiv \tau > 0$ gives $\bar{\beta}_i = (\log m_i)/\tau$ in Theorem 3.2(i), while $\tau_t \geq (t + t_0)^2/\kappa$ satisfies the budget requirement of Theorem 3.2(ii) with $q_i = \kappa \log m_i$. In particular, under the assumptions of Theorem 3.2(i), constant stepsizes $\gamma_t \equiv \gamma$ and a constant inverse temperature $\tau_t \equiv \tau$ yield geometric decay of the expected global objective gap at each honest node to a neighborhood with an additional aggregation term of order $\mathcal{O}(\log m_i/(\gamma\tau))$.

Existing convergence analyses of algorithms (Wu et al., 2023; Yang & Ghaderi, 2024; Gaucher et al., 2025; Han et al., 2025) do not explicitly compare their optimization bounds with those for self-only training. An error floor due to heterogeneity does not by itself establish a theoretical advantage, since self-only training admits a bound of the same order as shown in Section 3. We complement these guarantees by deriving sufficient conditions for LASA to achieve a strictly tighter global stationarity bound than self-only training. Define the LASA entropy and best honest-candidate gain by

$$H_{i,t} = -\sum_{j \in \mathcal{S}_i} \omega_{ij}^t \log \omega_{ij}^t, \qquad \Delta_{i,t}^{\mathrm{gain}} = \ell_{ii}^t - \min_{j \in \mathcal{S}_i \cap \mathcal{H}} \ell_{ij}^t \geq 0.$$

Also let $\mathcal{U}_{i,T}^{\mathrm{loc}}$ denote the bound in (3.7) with $\beta_{i,t} \equiv 0$, which corresponds to the bound of self-only training:

$$\mathcal{U}_{i,T}^{\mathrm{loc}} = \frac{2e_{i,0}}{\chi_\rho \Gamma_T} + \frac{\delta_1^2}{\rho \chi_\rho} + \frac{L\sigma^2 \sum_{t=0}^{T-1} \gamma_t^2}{\chi_\rho \Gamma_T}.$$

The following theorem provides a sharper convergence bound by accounting for local-loss improvements offered by honest candidates.

Theorem 4.3. *Suppose Assumptions 1–3 hold. Let $0 < \gamma_t \leq 1/L$ for $t \geq 0$ and let R be sampled as in Theorem 3.1. For every $i \in \mathcal{H}$, the iterates generated by LASA with the weights in* (4.1) *satisfy*

$$\mathbb{E}\left[\left\|\nabla F(x_i^R)\right\|^2\right] \leq \mathcal{U}_{i,T}^{\mathrm{loc}} + \frac{2}{\chi_\rho \Gamma_T} \sum_{t=0}^{T-1} \left(\frac{\mathbb{E}\left[H_{i,t}\right]}{\tau_t} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{gain}}\right] \right).$$

Consequently, the LASA bound is strictly smaller than the corresponding self-only bound whenever

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\Delta_{i,t}^{\mathrm{gain}}\right] > \sum_{t=0}^{T-1} \frac{\mathbb{E}\left[H_{i,t}\right]}{\tau_t}. \tag{4.4}$$

Theorem 4.3 provides a sufficient cooperation certificate when the expected cumulative gains from lower-loss honest candidates exceed the entropy penalties. Additional conditions and a similar bound comparison under the quadratic growth assumption are provided in Appendix C. Importantly, this certificate does not exhaust the benefits of aggregation, since candidates with higher losses may also improve the aggregated model through mixing. Appendix C.3 provides an example in which, for any fixed finite horizon, a suitable inverse temperature ensures that (4.4) holds, yielding a strict improvement over self-only trajectories. The example further illustrates an advantage over the inverse-loss rule arising from LASA's ability to retain higher-loss candidates.

Table 1: MNIST logistic regression on the regular graph (Figure 3a) with a Dirichlet partition ($\alpha = 0.5$). Entries show mean $\pm$ standard deviation over five seeds. Bold marks the best unrounded mean per setting and metric. All values are rounded for display, with CE means above 10^4 abbreviated.

	Benign	Gaussian	Sign Flip	Label Flip	FOE	Dissensus
Mean test accuracy (MTA, %) ↑						
DSGD	$\mathbf{92.07_{\pm 0.08}}$	$50.99_{\pm 7.73}$	$63.92_{\pm 1.81}$	$86.23_{\pm 0.15}$	$57.66_{\pm 2.49}$	$85.70_{\pm 0.17}$
IOS	$90.27_{\pm 0.29}$	$90.46_{\pm 0.17}$	$90.46_{\pm 0.17}$	$89.74_{\pm 0.18}$	$90.46_{\pm 0.17}$	$83.99_{\pm 0.57}$
NNA	$90.00_{\pm 0.35}$	$90.36_{\pm 0.32}$	$90.36_{\pm 0.32}$	$90.36_{\pm 0.33}$	$90.10_{\pm 0.48}$	$85.57_{\pm 0.51}$
CG	$91.98_{\pm 0.03}$	$91.71_{\pm 0.05}$	$86.73_{\pm 0.91}$	$88.34_{\pm 0.42}$	$85.50_{\pm 0.51}$	$89.98_{\pm 0.05}$
IL	$86.53_{\pm 1.19}$	$86.57_{\pm 2.22}$	$86.77_{\pm 1.98}$	$85.59_{\pm 1.62}$	$86.84_{\pm 2.19}$	$87.27_{\pm 1.81}$
SELF	$79.87_{\pm 0.09}$	$79.87_{\pm 0.09}$	$79.87_{\pm 0.09}$	$79.87_{\pm 0.09}$	$79.87_{\pm 0.09}$	$79.87_{\pm 0.09}$
LASA	$92.06_{\pm 0.07}$	$\mathbf{91.89_{\pm 0.06}}$	$\mathbf{91.90_{\pm 0.06}}$	$\mathbf{91.68_{\pm 0.14}}$	$\mathbf{91.90_{\pm 0.06}}$	$\mathbf{90.73_{\pm 0.49}}$
Worst-node test accuracy (WTA, %) ↑						
DSGD	$\mathbf{91.87_{\pm 0.15}}$	$24.05_{\pm 7.49}$	$31.51_{\pm 9.02}$	$83.60_{\pm 0.75}$	$13.99_{\pm 4.03}$	$76.25_{\pm 0.09}$
IOS	$89.68_{\pm 0.49}$	$89.94_{\pm 0.38}$	$89.94_{\pm 0.38}$	$87.58_{\pm 0.39}$	$89.94_{\pm 0.38}$	$32.84_{\pm 4.40}$
NNA	$88.99_{\pm 0.76}$	$89.66_{\pm 0.51}$	$89.66_{\pm 0.51}$	$89.67_{\pm 0.51}$	$89.40_{\pm 0.43}$	$49.37_{\pm 0.34}$
CG	$91.85_{\pm 0.02}$	$91.49_{\pm 0.06}$	$82.95_{\pm 5.54}$	$86.10_{\pm 2.66}$	$82.22_{\pm 2.74}$	$76.37_{\pm 0.20}$
IL	$79.11_{\pm 6.85}$	$82.34_{\pm 3.88}$	$82.86_{\pm 2.80}$	$79.18_{\pm 2.60}$	$82.69_{\pm 3.19}$	$81.86_{\pm 2.31}$
SELF	$73.05_{\pm 0.28}$	$73.05_{\pm 0.28}$	$73.05_{\pm 0.28}$	$73.05_{\pm 0.28}$	$73.05_{\pm 0.28}$	$73.05_{\pm 0.28}$
LASA	$91.85_{\pm 0.12}$	$\mathbf{91.59_{\pm 0.14}}$	$\mathbf{91.59_{\pm 0.13}}$	$\mathbf{91.17_{\pm 0.30}}$	$\mathbf{91.59_{\pm 0.13}}$	$\mathbf{89.25_{\pm 0.86}}$
Consensus error (CE) ↓						
DSGD	$0.002_{\pm 0.000}$	$187.2_{\pm 2.120}$	$0.044_{\pm 0.003}$	$\mathbf{0.001_{\pm 0.000}}$	$0.058_{\pm 0.002}$	$5.267_{\pm 0.010}$
IOS	$3.428_{\pm 0.053}$	$3.164_{\pm 0.061}$	$3.164_{\pm 0.061}$	$1.075_{\pm 0.212}$	$3.164_{\pm 0.061}$	$> 10^4$
NNA	$5.730_{\pm 0.245}$	$2.106_{\pm 0.701}$	$2.106_{\pm 0.701}$	$2.106_{\pm 0.701}$	$5.089_{\pm 1.163}$	$> 10^4$
CG	$\mathbf{0.001_{\pm 0.000}}$	$\mathbf{0.002_{\pm 0.000}}$	$0.026_{\pm 0.003}$	$0.020_{\pm 0.002}$	$0.012_{\pm 0.001}$	$0.904_{\pm 0.003}$
IL	$0.510_{\pm 0.204}$	$0.366_{\pm 0.142}$	$0.348_{\pm 0.141}$	$0.719_{\pm 0.195}$	$0.370_{\pm 0.152}$	$0.312_{\pm 0.066}$
SELF	$4.510_{\pm 0.014}$	$4.510_{\pm 0.014}$	$4.510_{\pm 0.014}$	$4.510_{\pm 0.014}$	$4.510_{\pm 0.014}$	$4.510_{\pm 0.014}$
LASA	$0.003_{\pm 0.000}$	$0.003_{\pm 0.000}$	$\mathbf{0.003_{\pm 0.000}}$	$0.009_{\pm 0.001}$	$\mathbf{0.003_{\pm 0.000}}$	$\mathbf{0.039_{\pm 0.004}}$

5 Empirical Evaluation

In this section, we evaluate LASA on MNIST logistic regression and CIFAR-10 ResNet-20 across several communication graphs and attack settings. We examine its robustness to Byzantine attacks and sensitivity to the inverse temperature and loss evaluation batch size. Table 3 summarizes experimental settings, while full details are provided in Appendix D.

5.1 MNIST Experiments

We evaluate LASA on MNIST logistic regression, reporting mean test accuracy (MTA), worst-node test accuracy (WTA), and consensus error (CE) over honest nodes. Unless otherwise stated, we use a 12-node regular graph (Figure 3a) of degree 4 with two Byzantine nodes and partition the training data across nodes using a Dirichlet distribution with $\alpha = 0.5$ (Hsu et al., 2019); LASA uses a constant inverse temperature $\tau = 3$. The baselines are standard DSGD, Iterative Outlier Scissor (IOS) (Wu et al., 2023), Nearest-Neighbor Averaging (NNA) (Farhadkhani et al., 2023), ClippedGossip (CG) (He et al., 2023; Han et al., 2025), inverse-loss weighting (IL) (Li et al., 2020) and self-only training (SELF).

Comparison with other aggregators. As shown in Table 1, LASA achieves the highest MTA and WTA under all five attacks and remains competitive in the benign setting. Under Dissensus, it raises WTA from the best baseline's 81.86% to 89.25% and achieves the lowest CE. LASA also outperforms SELF and IL in all metrics across all settings. Together, these results highlight LASA's ability to sustain accuracy at vulnerable honest nodes while retaining empirical benefits from cooperation.

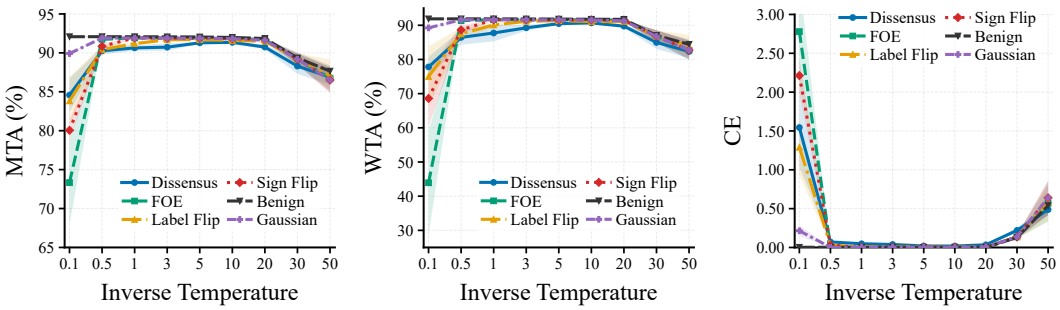


Figure 1: Sensitivity of LASA to the inverse temperature τ on MNIST under benign and adversarial settings. Lines show means over five seeds, and shaded regions indicate standard deviations.

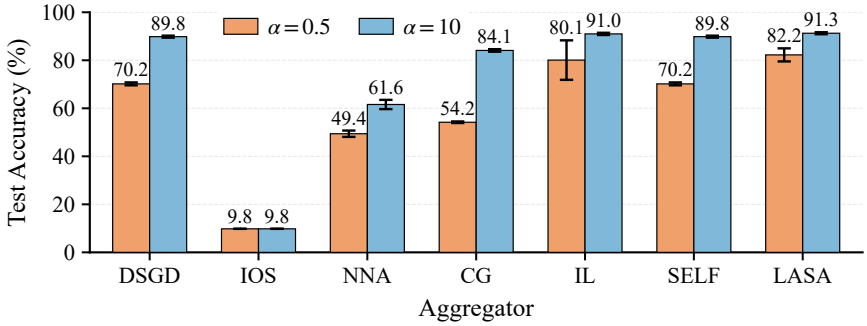


Figure 2: Final test accuracy of honest node 6 under Dissensus on MNIST, with three Byzantine neighbors and one honest neighbor (Figure 3c). Orange and blue bars correspond to $\alpha = 0.5$ and $\alpha = 10$, respectively. Bars show means over five seeds, and error bars indicate standard deviations.

Inverse temperature sensitivity. We next examine LASA's sensitivity to the inverse temperature τ. Figure 1 shows that the performance remains stable across a broad range of inverse temperatures $\tau \in [1, 20]$. However, $\tau = 0.1$ degrades performance in most tested settings, while setting τ to 30 or 50 lowers accuracy and increases CE in all tested settings. These results indicate stable performance over a broad range of inverse temperatures, while suggesting that excessively small or large values may lead to insufficient robustness or limited beneficial cooperation, respectively.

Loss evaluation batch size. To reduce the cost of local-loss evaluation, we evaluate candidate models using sampled minibatches. Figure 4 in Appendix D.6 shows minor changes in test accuracy and CE when using batches of 32 to 512 samples compared with full-data evaluation in the tested MNIST settings, suggesting that small batches can maintain good performance.

A different communication graph. We also evaluate logistic regression on an Erdős–Rényi graph with 20 nodes, edge probability 0.35, and 4 Byzantine nodes, as shown in Figure 3b in Appendix D. As reported in Table 4 in Appendix D.6, LASA achieves the highest MTA and WTA under Gaussian, Sign Flip, Label Flip, and FOE. LASA also achieves competitive performance in the benign setting and under Dissensus, providing evidence of its effectiveness across different communication topologies.

A Byzantine-dominance case. We further examine LASA's performance when Byzantine nodes form a majority of an honest node's neighbors, a setting permitted by our theoretical analysis in Section 3. Under Dissensus, we report this node's test accuracy for low ($\alpha = 10$) and high ($\alpha = 0.5$) heterogeneity. Figure 2 shows that LASA achieves the highest mean accuracy at both heterogeneity levels, while IOS remains near chance accuracy and NNA underperforms SELF. These results illustrate LASA's effectiveness at an honest node with predominantly Byzantine neighbors.

5.2 CIFAR-10 Experiments

We further evaluate LASA on CIFAR-10 with ResNet-20 (He et al., 2016). We use a Dirichlet data partition with $\alpha = 0.5$ and the same regular graph, attacks, and baselines as for MNIST. LASA

Table 2: CIFAR-10 ResNet-20 on the regular graph (Figure 3a) with a Dirichlet partition ($\alpha = 0.5$). Reporting conventions follow Table 1.

	Benign	Gaussian	Sign Flip	Label Flip	FOE	Dissensus
Mean test accuracy (MTA, %) ↑						
DSGD	$\mathbf{86.98_{\pm 0.23}}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$75.16_{\pm 0.98}$	$10.00_{\pm 0.00}$	$36.29_{\pm 1.40}$
IOS	$73.16_{\pm 0.43}$	$75.85_{\pm 0.74}$	$76.28_{\pm 0.80}$	$66.02_{\pm 3.78}$	$75.72_{\pm 0.83}$	$67.48_{\pm 0.61}$
NNA	$73.70_{\pm 0.70}$	$79.42_{\pm 0.80}$	$\mathbf{79.18_{\pm 0.97}}$	$70.30_{\pm 2.91}$	$\mathbf{79.18_{\pm 0.97}}$	$69.52_{\pm 0.99}$
CG	$83.05_{\pm 0.28}$	$81.63_{\pm 0.53}$	$10.00_{\pm 0.00}$	$73.69_{\pm 2.25}$	$10.00_{\pm 0.00}$	$69.65_{\pm 0.76}$
IL	$64.17_{\pm 1.66}$	$60.85_{\pm 1.75}$	$59.01_{\pm 2.22}$	$63.37_{\pm 1.61}$	$54.28_{\pm 2.69}$	$57.57_{\pm 1.16}$
SELF	$46.34_{\pm 0.70}$	$46.34_{\pm 0.70}$	$46.34_{\pm 0.70}$	$46.34_{\pm 0.70}$	$46.34_{\pm 0.70}$	$46.34_{\pm 0.70}$
LASA	$84.45_{\pm 0.26}$	$\mathbf{82.89_{\pm 0.47}}$	$79.05_{\pm 0.52}$	$\mathbf{76.68_{\pm 0.19}}$	$77.00_{\pm 0.60}$	$\mathbf{70.21_{\pm 1.04}}$
Worst-node test accuracy (WTA, %) ↑						
DSGD	$\mathbf{86.97_{\pm 0.23}}$	$10.00_{\pm 0.00}$	$10.00_{\pm 0.00}$	$75.14_{\pm 0.97}$	$10.00_{\pm 0.00}$	$25.24_{\pm 4.15}$
IOS	$68.64_{\pm 1.09}$	$75.84_{\pm 0.74}$	$76.27_{\pm 0.79}$	$40.69_{\pm 14.62}$	$75.72_{\pm 0.83}$	$10.00_{\pm 0.00}$
NNA	$68.78_{\pm 0.74}$	$79.02_{\pm 0.93}$	$78.90_{\pm 1.11}$	$44.10_{\pm 15.94}$	$\mathbf{78.90_{\pm 1.11}}$	$9.46_{\pm 0.89}$
CG	$83.04_{\pm 0.28}$	$81.60_{\pm 0.53}$	$10.00_{\pm 0.00}$	$73.65_{\pm 2.26}$	$10.00_{\pm 0.00}$	$41.31_{\pm 0.74}$
IL	$54.47_{\pm 2.49}$	$48.11_{\pm 8.18}$	$46.87_{\pm 8.16}$	$54.54_{\pm 1.48}$	$40.75_{\pm 6.77}$	$38.16_{\pm 6.76}$
SELF	$34.35_{\pm 1.38}$	$34.35_{\pm 1.38}$	$34.35_{\pm 1.38}$	$34.35_{\pm 1.38}$	$34.35_{\pm 1.38}$	$34.35_{\pm 1.38}$
LASA	$84.44_{\pm 0.26}$	$\mathbf{82.88_{\pm 0.46}}$	$\mathbf{79.04_{\pm 0.52}}$	$\mathbf{76.68_{\pm 0.19}}$	$76.99_{\pm 0.60}$	$\mathbf{63.85_{\pm 0.57}}$
Consensus error (CE) ↓						
DSGD	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$73.28_{\pm 2.653}$
IOS	$391.0_{\pm 54.66}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$336.0_{\pm 61.44}$	$\mathbf{0.000_{\pm 0.000}}$	$> 10^4$
NNA	$123.2_{\pm 25.30}$	$0.017_{\pm 0.011}$	$0.008_{\pm 0.004}$	$0.013_{\pm 0.004}$	$0.008_{\pm 0.004}$	$> 10^4$
CG	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$68.14_{\pm 2.005}$
IL	$38.45_{\pm 9.500}$	$82.88_{\pm 49.87}$	$124.4_{\pm 93.79}$	$54.73_{\pm 15.94}$	$161.2_{\pm 86.98}$	$119.6_{\pm 62.44}$
SELF	$895.2_{\pm 10.60}$	$895.2_{\pm 10.60}$	$895.2_{\pm 10.60}$	$895.2_{\pm 10.60}$	$895.2_{\pm 10.60}$	$895.2_{\pm 10.60}$
LASA	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{0.000_{\pm 0.000}}$	$\mathbf{5.483_{\pm 1.519}}$

uses a constant inverse temperature selected from $\tau \in \{5.0, 7.5, 10.0\}$ and 128 samples per loss evaluation. As shown in Table 2, LASA achieves the highest MTA under Gaussian, Label Flip, and Dissensus. Although NNA achieves slightly higher MTA under Sign Flip and higher MTA and WTA under FOE, LASA achieves the highest WTA under all other tested attacks. Specifically, LASA achieves substantially higher WTA (63.85% versus 9.46%) than NNA under Dissensus, highlighting its stronger performance at the most vulnerable honest nodes. These results demonstrate LASA's applicability in more challenging nonconvex deep learning tasks under diverse Byzantine attacks.

6 Discussion and Conclusion

In this paper, we have introduced deterministic aggregation budgets that bound the weighted excess local loss relative to the self-candidate. For aggregators admitting such budgets, we have established global stationarity bounds under smoothness, convexity, and relative gradient heterogeneity, with sharper guarantees under an additional quadratic growth condition. We have further proposed LASA, which admits explicit budgets while assigning positive weights to all candidates, preserving opportunities for useful cooperation. Moreover, LASA requires no prior knowledge of Byzantine participation or the global network topology and does not rely on an honest-majority assumption, allowing it to operate in sparse networks and in neighborhoods where Byzantine nodes may form a majority. We have further derived sufficient conditions under which the LASA global stationarity bound is strictly tighter than the corresponding bound for self-only training. Experiments on MNIST and CIFAR-10 demonstrate robust performance across diverse communication graphs and Byzantine attacks, particularly at vulnerable honest nodes. Future work includes extending the analysis to nonconvex objectives, establishing model-disagreement bounds under alternative structural assumptions, and developing adaptive inverse temperature schemes to further improve practical performance.

AI USE STATEMENT

In this work, we have used generative AI tools to assist with the writing of proofs, interpret results, assist with translation, create or edit software code, edit a research paper to improve readability, and create or modify scientific figures or images. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, or support qualitative and thematic data analysis. We have verified all AI-assisted proofs and tested LLM-generated code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on methodological and theoretical advances in Byzantine-robust decentralized optimization and does not involve human subjects or sensitive personal data. We are not aware of any specific ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of both our theoretical and empirical results. The complete DSGD and LASA procedures are given in Appendix A. All assumptions used in the theoretical analysis are stated explicitly in Sections 2–4, with complete proofs of the main results and additional theoretical details provided in Appendices B and C. Appendix D documents the experimental setup, including communication graphs, model architectures, data preprocessing, hyperparameter ranges, Byzantine attacks, baseline implementations, hyperparameter selection protocol, local-loss evaluation, and additional results.

REFERENCES

Opeyemi Ajibuwa, Olusola T. Odeyomi, Xiaohong Yuan, Kaushik Roy, and Danda Rawat. Data heterogeneity and Byzantine resiliency in decentralized federated learning. *IEEE Transactions on Cognitive Communications and Networking*, 12:8776–8794, 2026.

Zeyuan Allen-Zhu, Faeze Ebrahimianghazani, Jerry Li, and Dan Alistarh. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Fixing by mixing: A recipe for optimal Byzantine ML under heterogeneity. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 1232–1300, 2023.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, Geovani Rizk, and Sasha Voitovych. Byzantine-robust federated learning: Impact of client subsampling and local updates. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 1078–1114, 2024.

Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Ahmed Jellouli, Geovani Rizk, and John Stephan. Adaptive gradient clipping for robust federated learning. In *International Conference on Learning Representations*, pp. 84251–84295, 2025.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, pp. 118–128, 2017.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *Network and Distributed System Security Symposium*, 2021.

Xingrong Dong, Zhaoxian Wu, Qing Ling, and Zhi Tian. Byzantine-robust distributed online learning: Taming adversarial participants in an adversarial environment. *IEEE Transactions on Signal Processing*, 72:235–248, 2024.

Cheng Fang, Zhixiong Yang, and Waheed U. Bajwa. BRIDGE: Byzantine-resilient decentralized gradient descent. *IEEE Transactions on Signal and Information Processing over Networks*, 8: 610–626, 2022.

Minghong Fang, Zifan Zhang, Hairi, Prashant Khanduri, Jia Liu, Songtao Lu, Yuchen Liu, and Neil Gong. Byzantine-robust decentralized federated learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 2874–2888, 2024.

Minghong Fang, Seyedsina Nabavirazavi, Zhuqing Liu, Wei Sun, Sundararaja Sitharama Iyengar, and Haibo Yang. Do we really need to design new Byzantine-robust aggregation rules? In *Network and Distributed System Security Symposium*, 2025.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 6246–6283, 2022.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. Robust collaborative learning with linear gradient overhead. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 9761–9813, 2023.

Renaud Gaucher, Aymeric Dieuleveut, and Hadrien Hendrikx. Unified breakdown analysis for Byzantine robust gossip. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 18868–18896, 2025.

Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Byzantine machine learning: A primer. *ACM Computing Surveys*, 56(7):1–39, 2024.

Shangwei Guo, Tianwei Zhang, Han Yu, Xiaofei Xie, Lei Ma, Tao Xiang, and Yang Liu. Byzantine-resilient decentralized stochastic gradient descent. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):4096–4106, 2022.

Dongyu Han, Kun Liu, Yuanqing Xia, and Lihua Xie. Byzantine-robust distributed stochastic nonconvex optimization in adversarial environments over unbalanced networks. *IEEE Transactions on Automatic Control*, 70(12):8029–8043, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via ClippedGossip. *arXiv preprint arXiv:2202.01545*, 2023.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for Byzantine robust optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5311–5319, 2021.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5381–5393, 2020.

Kananart Kuwaranancharoen and Shreyas Sundaram. On the geometric convergence of Byzantine-resilient distributed optimization algorithms. *SIAM Journal on Optimization*, 35(1):210–239, 2025.

Kananart Kuwarananchaoen, Lei Xin, and Shreyas Sundaram. Scalable distributed optimization of multi-dimensional functions despite Byzantine adversaries. *IEEE Transactions on Signal and Information Processing over Networks*, 10:360–375, 2024.

Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

Jiani Li, Waseem Abbas, and Xenofon Koutsoukos. Byzantine resilient distributed multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18215–18225, 2020.

Zhirong Luan, Wenrui Li, Meiqin Liu, and Badong Chen. Robust federated learning: Maximum correntropy aggregation against Byzantine attacks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):62–75, 2025.

Gleb Molodtsov, Daniil Medyakov, Sergey Skorik, Nikolas Khachaturov, Shahane Tigranyan, Vladimir Aletov, Aram Avetisyan, Martin Takáč, and Aleksandr Beznosikov. Bant: Byzantine antidote via trial function and trust scores. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24431–24440, 2026.

Tomoya Murata, Kenta Niwa, Takumi Fukami, and Iifan Tyou. Simple minimax optimal Byzantine robust algorithm for nonconvex objectives with uniform gradient heterogeneity. In *International Conference on Learning Representations*, pp. 26449–26485, 2024.

Ion Necoara, Yurii Nesterov, and François Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

Jungwuk Park, Dong-Jun Han, Minseok Choi, and Jaekyun Moon. Sageflow: Robust federated learning against both stragglers and adversaries. In *Advances in Neural Information Processing Systems*, volume 34, pp. 840–851, 2021.

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

Murtaza Rangwala, Farag Azzedin, Richard O. Sinnott, and Rajkumar Buyya. SketchGuard: Scaling Byzantine-robust decentralized federated learning via sketch-based screening. *arXiv preprint arXiv:2510.07922*, 2026.

Yuxin Wu and Kaiming He. Group normalization. In *Computer Vision – ECCV 2018*, pp. 3–19, 2018.

Zhaoxian Wu, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE Transactions on Signal Processing*, 71:3179–3195, 2023.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6893–6901, 2019.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 261–270, 2020.

Caiyi Yang and Javad Ghaderi. Byzantine-robust decentralized learning via Remove-then-Clip aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21735–21743, 2024.

Yuhao Yi, Ronghui You, Hong Liu, Changxin Liu, Yuan Wang, and Jiancheng Lv. Near-optimal resilient aggregation rules for distributed learning using 1-center and 1-mean clustering with outliers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16469–16477, 2024.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 5650–5659, 2018.

Shiyuan Zuo, Xingrun Yan, Rongfei Fan, Li Shen, Puning Zhao, Jie Xu, and Han Hu. Efficient federated learning against Byzantine attacks and data heterogeneity via aggregating normalized gradients. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28998–29035, 2025.

A Pseudocode for DSGD and LASA

We summarize the DSGD framework and specify how LASA performs the aggregation step at each honest node. The DSGD framework allows a general aggregator $\mathcal{A}$ that combines received candidates using locally available information:

$$x_i^{t+1} = \mathcal{A}\big(i, x_i^t, \{\tilde{x}_{ij}^t\}_{j\in\mathcal{S}_i}\big)\,. \tag{A.1}$$

The weighted-average update (2.2) is a particular instance of this framework. Algorithm 1 summarizes the DSGD framework in Section 2. Each round consists of a local stochastic gradient step, communication with neighbors, and aggregation of the received candidates.

Algorithm 1 The Decentralized Stochastic Gradient Descent Framework

Input: Initial models $\{x_i^0\}_{i=1}^n$, stepsizes $\{\gamma_t\}_{t=0}^{T-1}$, aggregator $\mathcal{A}$
for $t = 0, 1, \ldots, T-1$ **do**
  for each honest node $i \in \mathcal{H}$ **do**
    Compute a stochastic gradient g_i^t from the local data
    Set $\tilde{x}_i^t = x_i^t - \gamma_t g_i^t$ and broadcast it to neighbors
  end for
  Byzantine nodes may send arbitrary, receiver-dependent candidates
  for each honest node $i \in \mathcal{H}$ **do**
    Receive $\{\tilde{x}_{ij}^t : j \in \mathcal{S}_i\}$, including $\tilde{x}_{ii}^t = \tilde{x}_i^t$
    Set $x_i^{t+1} = \mathcal{A}(i, x_i^t, \{\tilde{x}_{ij}^t\}_{j\in\mathcal{S}_i})$
  end for
end for
Output: Final honest models $\{x_i^T\}_{i\in\mathcal{H}}$

LASA discussed in Section 4 instantiates the aggregation step in Algorithm 1. As detailed in Algorithm 2, each honest node evaluates the candidates using its local objective and converts their loss improvements into softmax weights.

Algorithm 2 LASA at an Honest Node i at Round t

Input: x_i^t, candidates $\{\tilde{x}_{ij}^t\}_{j\in\mathcal{S}_i}$, local loss f_i, inverse temperature parameter τ_t
Evaluate $\ell_{ij}^t = f_i(\tilde{x}_{ij}^t)$ for every $j \in \mathcal{S}_i$
Compute $\Delta_{ij}^t = \ell_{ii}^t - \ell_{ij}^t$ for every $j \in \mathcal{S}_i$
Set $\omega_{ij}^t = \exp(\tau_t \Delta_{ij}^t) / \sum_{k\in\mathcal{S}_i} \exp(\tau_t \Delta_{ik}^t)$
Set $x_i^{t+1} = \sum_{j\in\mathcal{S}_i} \omega_{ij}^t \tilde{x}_{ij}^t$
Output: Aggregated model x_i^{t+1}

B Convergence Proofs for Aggregators with Budgets

In this section, we establish convergence guarantees for aggregators admitting aggregation budgets, covering smooth convex objectives and the sharper bounds under quadratic growth.

B.1 Proof of Theorem 3.1

We first establish a descent lemma for the honest local objectives.

Lemma B.1. *Suppose that Assumptions 1–2 hold, and let the stepsizes satisfy $0 < \gamma_t \le 1/L$. Then, for every $i \in \mathcal{H}$, the DSGD iterates $\{x_i^t\}$ generated with any weighted-average aggregator given by* (2.2) *admitting aggregation budgets $\{\beta_{i,t}\}$ satisfy*

$$\mathbb{E}\left[f_i(x_i^{t+1}) \mid \mathcal{F}_t\right] \le f_i(x_i^t) - \frac{\gamma_t}{2}\left\|\nabla f_i(x_i^t)\right\|^2 + \frac{L\gamma_t^2\sigma^2}{2} + \beta_{i,t}. \tag{B.1}$$

Proof. Convexity and the conditions (2.3) and (2.4) imply that

$$f_i(x_i^{t+1}) \le \sum_j \omega_{ij}^t f_i(\tilde{x}_{ij}^t) \le f_i(x_i^t - \gamma_t g_i^t) + \beta_{i,t}.$$

By smoothness of f_i at x_i^t, we obtain

$$f_i(x_i^{t+1}) \le f_i(x_i^t) - \gamma_t \left\langle \nabla f_i(x_i^t), g_i^t \right\rangle + \frac{L\gamma_t^2}{2} \left\| g_i^t \right\|^2 + \beta_{i,t}.$$

Taking conditional expectation on both sides gives

$$\begin{aligned} \mathbb{E}\left[f_i(x_i^{t+1}) \mid \mathcal{F}_t \right] &\le f_i(x_i^t) - \gamma_t \left\langle \nabla f_i(x_i^t), \mathbb{E}[g_i^t \mid \mathcal{F}_t] \right\rangle + \frac{L\gamma_t^2}{2} \mathbb{E}\left[\left\| g_i^t \right\|^2 \mid \mathcal{F}_t \right] + \beta_{i,t} \\ &= f_i(x_i^t) - \gamma_t \left\| \nabla f_i(x_i^t) \right\|^2 + \frac{L\gamma_t^2}{2} \mathbb{E}\left[\left\| g_i^t \right\|^2 \mid \mathcal{F}_t \right] + \beta_{i,t}. \end{aligned} \tag{B.2}$$

By Assumption 1 on stochastic gradients, we have

$$\mathbb{E}\left[\left\| g_i^t \right\|^2 \mid \mathcal{F}_t \right] = \left\| \nabla f_i(x_i^t) \right\|^2 + \mathbb{E}\left[\left\| g_i^t - \nabla f_i(x_i^t) \right\|^2 \mid \mathcal{F}_t \right] \le \left\| \nabla f_i(x_i^t) \right\|^2 + \sigma^2. \tag{B.3}$$

Substituting (B.3) into (B.2) and collecting the two gradient terms yields

$$\mathbb{E}\left[f_i(x_i^{t+1}) \mid \mathcal{F}_t \right] \le f_i(x_i^t) - \gamma_t \left(1 - \frac{L\gamma_t}{2} \right) \left\| \nabla f_i(x_i^t) \right\|^2 + \frac{L\gamma_t^2 \sigma^2}{2} + \beta_{i,t}.$$

Finally, using $0 < \gamma_t \le 1/L$ yields the desired inequality (B.1). □

The descent lemma controls the local gradient terms. The heterogeneity assumption then relates these terms to the global gradient, enabling the following proof of Theorem 3.1.

Proof of Theorem 3.1. We first prove the convergence result (3.7). From $\nabla F = \nabla f_i + (\nabla F - \nabla f_i)$ and Young's inequality, for any $\rho > 0$, we have

$$\left\| \nabla F(x) \right\|^2 \le (1+\rho) \left\| \nabla f_i(x) \right\|^2 + (1+\rho^{-1}) \left\| \nabla F(x) - \nabla f_i(x) \right\|^2.$$

Applying Assumption 3 to the last term yields

$$\left\| \nabla F(x) \right\|^2 \le (1+\rho) \left\| \nabla f_i(x) \right\|^2 + (1+\rho^{-1})\delta_1^2 + (1+\rho^{-1})\delta_2^2 \left\| \nabla F(x) \right\|^2.$$

Moving the last global gradient term to the left and dividing both sides by $1+\rho$ gives

$$\begin{aligned} \left\| \nabla f_i(x) \right\|^2 &\ge \frac{1 - (1+\rho^{-1})\delta_2^2}{1+\rho} \left\| \nabla F(x) \right\|^2 - \frac{1+\rho^{-1}}{1+\rho} \delta_1^2 \\ &= \chi_\rho \left\| \nabla F(x) \right\|^2 - \frac{\delta_1^2}{\rho}. \end{aligned} \tag{B.4}$$

Plugging (B.4) in the one-step descent inequality (B.1), we have

$$\mathbb{E}\left[f_i(x_i^{t+1}) \mid \mathcal{F}_t \right] \le f_i(x_i^t) - \frac{\gamma_t \chi_\rho}{2} \left\| \nabla F(x_i^t) \right\|^2 + \frac{\gamma_t \delta_1^2}{2\rho} + \frac{L\gamma_t^2 \sigma^2}{2} + \beta_{i,t}. \tag{B.5}$$

Taking total expectation and rearranging the gradient term to the left-hand side yields, for every t,

$$\frac{\gamma_t \chi_\rho}{2} \mathbb{E}\left[\left\| \nabla F(x_i^t) \right\|^2 \right] \le \mathbb{E}[f_i(x_i^t)] - \mathbb{E}[f_i(x_i^{t+1})] + \frac{\gamma_t \delta_1^2}{2\rho} + \frac{L\gamma_t^2 \sigma^2}{2} + \beta_{i,t}.$$

Telescoping the above inequality from $t = 0$ to $T-1$ and noting that $f_i(x_i^T) \ge f_i^\star$, we obtain

$$\frac{\chi_\rho}{2} \sum_{t=0}^{T-1} \gamma_t \mathbb{E}\left[\left\| \nabla F(x_i^t) \right\|^2 \right] \le f_i(x_i^0) - f_i^\star + \frac{\delta_1^2}{2\rho} \Gamma_T + \frac{L\sigma^2}{2} \sum_{t=0}^{T-1} \gamma_t^2 + \sum_{t=0}^{T-1} \beta_{i,t}.$$

By the sampling rule for R, we have

$$\mathbb{E}\left[\left\|\nabla F(x_i^R)\right\|^2\right] = \frac{1}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t \mathbb{E}\left[\left\|\nabla F(x_i^t)\right\|^2\right]. \tag{B.6}$$

Substitution of (B.6) into the preceding bound and division by $\chi_\rho \Gamma_T/2$ on both sides yield

$$\mathbb{E}\left[\left\|\nabla F(x_i^R)\right\|^2\right] \le \frac{2(f_i(x_i^0)-f_i^\star)}{\chi_\rho\Gamma_T} + \frac{\delta_1^2}{\rho\chi_\rho} + \frac{L\sigma^2\sum_{t=0}^{T-1}\gamma_t^2}{\chi_\rho\Gamma_T} + \frac{2\sum_{t=0}^{T-1}\beta_{i,t}}{\chi_\rho\Gamma_T}.$$

With the notation $e_{i,0} = f_i(x_i^0) - f_i^\star$, this establishes (3.7) in Theorem 3.1.

Now we prove the gradient disagreement bound (3.8). For arbitrary vectors $\{v_i\}$, their arithmetic mean minimizes $z \mapsto \sum_i \|v_i - z\|^2$. Applying this fact to $v_i = \nabla F(x_i^t)$ and comparing the minimizing value with the choice $z = 0$, we obtain

$$\mathcal{G}_t = \frac{1}{h}\sum_{i\in\mathcal{H}}\left\|\nabla F(x_i^t) - \frac{1}{h}\sum_{j\in\mathcal{H}}\nabla F(x_j^t)\right\|^2 = \min_{z\in\mathbb{R}^d}\frac{1}{h}\sum_{i\in\mathcal{H}}\left\|\nabla F(x_i^t) - z\right\|^2 \le \frac{1}{h}\sum_{i\in\mathcal{H}}\left\|\nabla F(x_i^t)\right\|^2.$$

Taking expectation in the above inequality at the common random index R yields

$$\mathbb{E}[\mathcal{G}_R] \le \frac{1}{h}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla F(x_i^R)\right\|^2\right].$$

Averaging (3.7) over $i \in \mathcal{H}$, we have

$$\frac{1}{h}\sum_{i\in\mathcal{H}}\mathbb{E}\left[\left\|\nabla F(x_i^R)\right\|^2\right] \le \frac{\frac{2}{h}\sum_{i\in\mathcal{H}}(f_i(x_i^0)-f_i^\star)}{\chi_\rho\Gamma_T} + \frac{\delta_1^2}{\rho\chi_\rho} + \frac{L\sigma^2\sum_{t=0}^{T-1}\gamma_t^2}{\chi_\rho\Gamma_T} + \frac{\frac{2}{h}\sum_{i\in\mathcal{H}}\sum_{t=0}^{T-1}\beta_{i,t}}{\chi_\rho\Gamma_T}.$$

Combining the two inequalities above yields (3.8).

Finally, the constant schedule gives

$$\Gamma_T = \frac{\sqrt{T}}{L}, \quad \sum_{t=0}^{T-1}\gamma_t^2 = \frac{1}{L^2}, \quad \sum_{t=0}^{T-1}\beta_{i,t} \le q_i.$$

Substituting these relations into (3.7) and (3.8), and bounding the averages of the initial gaps and budget constants by their respective maxima, we obtain

$$\max\left\{\mathbb{E}\left[\|\nabla F(x_i^R)\|^2\right], \mathbb{E}[\mathcal{G}_R]\right\} \le \frac{2L\max_{j\in\mathcal{H}}\{f_j(x_j^0)-f_j^\star\} + \sigma^2 + 2L\max_{j\in\mathcal{H}}\{q_j\}}{\chi_\rho\sqrt{T}} + \frac{\delta_1^2}{\rho\chi_\rho}.$$

For $\rho = 1/(1-\delta_2^2)$, we have

$$\frac{1}{\chi_\rho} = \frac{2-\delta_2^2}{(1-\delta_2^2)^3}, \qquad \frac{1}{\rho\chi_\rho} = \frac{2-\delta_2^2}{(1-\delta_2^2)^2} \le \frac{2}{(1-\delta_2)^2}.$$

For fixed $\delta_2 < 1$, absorbing χ_ρ^{-1} into the transient constant yields (3.9). □

B.2 Technical Lemmas for Quadratic Growth

We collect two consequences of quadratic growth that will be used to establish the last-iterate optimization and consensus error bounds.

Lemma B.2. *Suppose that Assumption 2 and* (3.3) *in Assumption 4 hold. Then the Polyak-Łojasiewicz condition holds for every $i \in \mathcal{H}$:*

$$\|\nabla f_i(x)\|^2 \ge \frac{\mu_{\mathrm{loc}}}{2}(f_i(x) - f_i^\star) \ge 2\mu_{\mathrm{PL}}(f_i(x) - f_i^\star). \tag{B.7}$$

Proof. Fix an honest node $i \in \mathcal{H}$. Since f_i is continuous and its nonempty minimizer set $\mathcal{X}_i^\star$ is closed, a Euclidean projection onto that set exists. Choose $x_i^\star(x) \in \mathcal{X}_i^\star$ which attains $\|x - x_i^\star(x)\| = \operatorname{dist}(x, \mathcal{X}_i^\star)$. By convexity and (3.3), we have

$$f_i(x) - f_i^\star \le \langle \nabla f_i(x), x - x_i^\star(x) \rangle \le \|\nabla f_i(x)\| \operatorname{dist}(x, \mathcal{X}_i^\star), \quad \operatorname{dist}^2(x, \mathcal{X}_i^\star) \le \frac{2}{\mu_{\mathrm{loc}}}(f_i(x) - f_i^\star).$$

Denote $d_i(x) = \operatorname{dist}(x, \mathcal{X}_i^\star)$ and $\Delta_i(x) = f_i(x) - f_i^\star$. If $\Delta_i(x) > 0$, the two inequalities above give

$$\Delta_i(x) \le \|\nabla f_i(x)\| \, d_i(x) \le \|\nabla f_i(x)\| \sqrt{\frac{2\Delta_i(x)}{\mu_{\mathrm{loc}}}}.$$

Dividing by $\sqrt{\Delta_i(x)}$ and squaring proves the first inequality in (B.7). If $\Delta_i(x) = 0$, x is a minimizer and $\nabla f_i(x) = 0$ by first-order optimality, so the same conclusion holds. Finally, $\mu_{\mathrm{PL}} \le \mu_{\mathrm{loc}}/4$ gives the second inequality in (B.7). □

Lemma B.3. *Suppose that Assumption 2 and* (3.4) *in Assumption 4 hold. Then*

$$\|x - x^\star\| \le \frac{2}{\mu_F} \|\nabla F(x)\|, \tag{B.8}$$

$$F(x) - F^\star \le \frac{2}{\mu_F} \|\nabla F(x)\|^2. \tag{B.9}$$

Proof. By uniqueness of $x^\star$, convexity, and (3.4), we have the following chain:

$$\frac{\mu_F}{2} \|x - x^\star\|^2 \le F(x) - F^\star \le \langle \nabla F(x), x - x^\star \rangle \le \|\nabla F(x)\| \, \|x - x^\star\|.$$

For $x \neq x^\star$, canceling one factor $\|x - x^\star\|$ between the first and last terms above gives $\|x - x^\star\| \le 2\|\nabla F(x)\| / \mu_F$; the same inequality is trivial at $x = x^\star$. Substituting this error bound back into the convexity upper bound $F(x) - F^\star \le \|\nabla F(x)\| \, \|x - x^\star\|$ gives (B.9). □

B.3 Details and Proof of Theorem 3.2

We give explicit last-iterate optimization and consensus error bounds under quadratic growth for constant and diminishing stepsizes, completing the statement and proof of Theorem 3.2.

Theorem B.4 (Quadratic growth with a constant stepsize)**.** *Suppose that Assumptions 1–4 hold. Consider DSGD using a weighted-average aggregator admitting aggregation budgets $\{\beta_{i,t}\}$ in the sense of Definition 1. Assume that $\beta_{i,t} \le \bar{\beta}_i$ and $\bar{\beta}_i \ge 0$ for all $i \in \mathcal{H}$ and $t \ge 0$. Let $0 < \gamma_t \equiv \gamma \le 1/L$, and define*

$$U_{i,t} = (1 - \mu_{\mathrm{PL}}\gamma)^t (f_i(x_i^0) - f_i^\star) + \frac{L\gamma\sigma^2}{2\mu_{\mathrm{PL}}} + \frac{\bar{\beta}_i}{\mu_{\mathrm{PL}}\gamma}.$$

Then for every integer $T \ge 1$ and $i \in \mathcal{H}$,

$$\mathbb{E}\left[F(x_i^T) - F^\star\right] \le \frac{8LU_{i,T} + 4\delta_1^2}{\mu_F(1 - \delta_2)^2}, \tag{B.10}$$

$$\mathbb{E}\left[\mathcal{C}_T\right] \le \frac{8}{\mu_F^2(1 - \delta_2)^2} \left(\frac{2L}{h} \sum_{j \in \mathcal{H}} U_{j,T} + \delta_1^2 \right). \tag{B.11}$$

Theorem B.5 (Quadratic growth with diminishing stepsizes)**.** *Suppose that Assumptions 1–4 hold. Consider DSGD using a weighted-average aggregator admitting aggregation budgets $\{\beta_{i,t}\}$ in the sense of Definition 1. Choose the stepsizes as below and suppose that the aggregation budgets satisfy*

$$\gamma_t = \frac{2}{\mu_{\mathrm{PL}}(t + t_0)}, \qquad t_0 \ge \max\left\{2, \frac{2L}{\mu_{\mathrm{PL}}}\right\}, \qquad \beta_{i,t} \le \frac{q_i}{(t + t_0)^2} \tag{B.12}$$

with $q_i \ge 0$. Define $A_i = \max\{(t_0 - 1)(f_i(x_i^0) - f_i^\star), 2L\sigma^2/\mu_{\mathrm{PL}}^2 + q_i\}$. Then for every integer $T \ge 1$ and $i \in \mathcal{H}$,

$$\mathbb{E}\left[F(x_i^T) - F^\star\right] \le \frac{8LA_i}{\mu_F(1 - \delta_2)^2(T + t_0 - 1)} + \frac{4\delta_1^2}{\mu_F(1 - \delta_2)^2}, \tag{B.13}$$

$$\mathbb{E}\left[\mathcal{C}_T\right] \le \frac{8}{\mu_F^2(1 - \delta_2)^2} \left(\frac{2L}{h(T + t_0 - 1)} \sum_{j \in \mathcal{H}} A_j + \delta_1^2 \right). \tag{B.14}$$

Denote

$$e_{i,t} = \mathbb{E}\left[f_i(x_i^t) - f_i^\star\right].$$

We first bound the local objective gaps using the PL inequality, then transfer these bounds to the global objective gap and model disagreement.

Lemma B.6. *Under the assumptions and schedules in Theorem B.4, for every honest node i and integer $T \geq 1$, we have*

$$e_{i,T} \leq (1 - \mu_{\mathrm{PL}}\gamma)^T e_{i,0} + \frac{L\gamma\sigma^2}{2\mu_{\mathrm{PL}}} + \frac{\bar{\beta}_i}{\mu_{\mathrm{PL}}\gamma} = U_{i,T}. \tag{B.15}$$

Proof. Taking total expectation in (B.1) and subtracting $f_i^\star$ from both sides, we obtain

$$e_{i,t+1} \leq e_{i,t} - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f_i(x_i^t)\right\|^2\right] + \frac{L\gamma^2\sigma^2}{2} + \beta_{i,t}.$$

By the local PL inequality (B.7), we have

$$\mathbb{E}\left[\left\|\nabla f_i(x_i^t)\right\|^2\right] \geq 2\mu_{\mathrm{PL}} e_{i,t}.$$

Combining these inequalities and using $\beta_{i,t} \leq \bar{\beta}_i$ yields

$$e_{i,t+1} \leq (1 - \mu_{\mathrm{PL}}\gamma)e_{i,t} + \frac{L\gamma^2\sigma^2}{2} + \bar{\beta}_i.$$

Set $q = 1 - \mu_{\mathrm{PL}}\gamma$. Since $0 < \mu_{\mathrm{PL}} \leq L$ and $0 < \gamma \leq 1/L$, we have $0 \leq q < 1$. Iterating the preceding recursion gives

$$\begin{aligned} e_{i,T} &\leq q^T e_{i,0} + \left(\frac{L\gamma^2\sigma^2}{2} + \bar{\beta}_i\right)\sum_{k=0}^{T-1} q^k \\ &= q^T e_{i,0} + \left(\frac{L\gamma\sigma^2}{2\mu_{\mathrm{PL}}} + \frac{\bar{\beta}_i}{\mu_{\mathrm{PL}}\gamma}\right)(1 - q^T) \\ &\leq (1 - \mu_{\mathrm{PL}}\gamma)^T e_{i,0} + \frac{L\gamma\sigma^2}{2\mu_{\mathrm{PL}}} + \frac{\bar{\beta}_i}{\mu_{\mathrm{PL}}\gamma}. \end{aligned}$$

This proves (B.15). □

Lemma B.7. *Under the assumptions and schedules in Theorem B.5, for any integer $T \geq 1$, we have*

$$e_{i,T} \leq \frac{A_i}{T + t_0 - 1}. \tag{B.16}$$

Proof. Taking total expectation in (B.1) gives

$$e_{i,t+1} \leq e_{i,t} - \frac{\gamma_t}{2}\mathbb{E}\left[\left\|\nabla f_i(x_i^t)\right\|^2\right] + \frac{L\gamma_t^2\sigma^2}{2} + \beta_{i,t}.$$

Applying the PL inequality (B.7) then gives

$$e_{i,t+1} \leq (1 - \mu_{\mathrm{PL}}\gamma_t)e_{i,t} + \frac{L\gamma_t^2\sigma^2}{2} + \beta_{i,t}. \tag{B.17}$$

Set $s(t) = t + t_0$ and $C_i = 2L\sigma^2/\mu_{\mathrm{PL}}^2 + q_i$. Note that $s(t) - 1 = s(t-1)$. The schedule in (B.12) gives $\mu_{\mathrm{PL}}\gamma_t = 2/s(t)$ and

$$\frac{L\gamma_t^2\sigma^2}{2} + \beta_{i,t} \leq \frac{2L\sigma^2/\mu_{\mathrm{PL}}^2 + q_i}{(s(t))^2} = \frac{C_i}{(s(t))^2}.$$

Substituting these relations in (B.17) gives

$$e_{i,t+1} \leq \left(1 - \frac{2}{s(t)}\right)e_{i,t} + \frac{C_i}{(s(t))^2}.$$

We prove $e_{i,t} \le A_i/s(t-1) = A_i/(t+t_0-1)$ by induction. At $t=0$, the definition $A_i \ge (t_0-1)e_{i,0}$ gives the base case. If the claim holds at t, then using $C_i \le A_i$ yields

$$e_{i,t+1} \le \left(1 - \frac{2}{s(t)}\right)\frac{A_i}{s(t-1)} + \frac{C_i}{(s(t))^2} \le \frac{A_i(s(t)-2)}{s(t)s(t-1)} + \frac{A_i}{(s(t))^2} \le \frac{A_i}{s(t)},$$

where the last inequality holds because

$$\begin{aligned}\frac{A_i(s(t)-2)}{s(t)s(t-1)} + \frac{A_i}{(s(t))^2} &= \frac{A_i}{s(t)}\left(\frac{s(t)-2}{s(t)-1} + \frac{1}{s(t)}\right) = \frac{A_i}{s(t)}\left(\frac{(s(t))^2 - s(t) - 1}{s(t)(s(t)-1)}\right) \\ &\le \frac{A_i}{s(t)}\left(\frac{(s(t))^2 - s(t)}{s(t)(s(t)-1)}\right) = \frac{A_i}{s(t)}.\end{aligned}$$

This completes the induction and proves (B.16). □

The following lemma relates the local objective gaps to the global objective gradients.

Lemma B.8. *Suppose that Assumptions 2 and 3 hold. Then for every $i \in \mathcal{H}$, $\varepsilon > 0$, and $x \in \mathbb{R}^d$,*

$$(1-\delta_2)^2 \left\|\nabla F(x)\right\|^2 \le 2L(1+\varepsilon)(f_i(x) - f_i^\star) + (1+\varepsilon^{-1})\delta_1^2. \tag{B.18}$$

Proof. By taking square roots in Assumption 3 and using $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$, we have

$$\left\|\nabla f_i(x) - \nabla F(x)\right\| \le \delta_1 + \delta_2 \left\|\nabla F(x)\right\|.$$

The above inequality and the triangle inequality then give

$$\left\|\nabla F(x)\right\| \le \left\|\nabla f_i(x)\right\| + \left\|\nabla f_i(x) - \nabla F(x)\right\| \le \left\|\nabla f_i(x)\right\| + \delta_1 + \delta_2 \left\|\nabla F(x)\right\|,$$

which further implies

$$(1-\delta_2)\left\|\nabla F(x)\right\| \le \left\|\nabla f_i(x)\right\| + \delta_1.$$

Squaring the inequality above and applying Young's inequality $(u+v)^2 \le (1+\varepsilon)u^2 + (1+\varepsilon^{-1})v^2$ gives, for every $\varepsilon > 0$,

$$(1-\delta_2)^2\left\|\nabla F(x)\right\|^2 \le (1+\varepsilon)\left\|\nabla f_i(x)\right\|^2 + (1+\varepsilon^{-1})\delta_1^2.$$

The smoothness of f_i implies $\left\|\nabla f_i(x)\right\|^2 \le 2L(f_i(x) - f_i^\star)$, which further yields the desired inequality (B.18). □

With all these lemmas, we are ready to prove the convergence guarantees under the additional quadratic growth condition.

Proof of Theorems B.4 and B.5. We first prove the convergence results (B.10) and (B.13). Taking expectation on both sides of (B.18) and using the definition of $e_{i,T}$, we obtain

$$(1-\delta_2)^2 \mathbb{E}\left[\left\|\nabla F(x_i^T)\right\|^2\right] \le 2L(1+\varepsilon)e_{i,T} + (1+\varepsilon^{-1})\delta_1^2.$$

Then, by taking expectation in (B.9) and inserting this gradient estimate, we have

$$\mathbb{E}[F(x_i^T) - F^\star] \le \frac{2}{\mu_F}\mathbb{E}\left[\left\|\nabla F(x_i^T)\right\|^2\right] \le \frac{2\left(2L(1+\varepsilon)e_{i,T} + (1+\varepsilon^{-1})\delta_1^2\right)}{\mu_F(1-\delta_2)^2}.$$

Finally, inserting (B.15) and (B.16) into the preceding bound and setting $\varepsilon = 1$, we obtain (B.10) and (B.13), respectively.

Now we prove the consensus error bounds (B.11) and (B.14). The arithmetic mean identity, followed by comparison with the unique minimizer $x^\star = \operatorname{argmin} F(x)$, gives

$$\sum_{i\in\mathcal{H}}\left\|x_i^T - \bar{x}^T\right\|^2 = \min_{z\in\mathbb{R}^d}\sum_{i\in\mathcal{H}}\left\|x_i^T - z\right\|^2 \le \sum_{i\in\mathcal{H}}\left\|x_i^T - x^\star\right\|^2.$$

The global minimizer is unique, so the pointwise error bound (B.8) applies to each x_i^T. By squaring that bound and summing, we obtain

$$\sum_{i\in\mathcal{H}}\left\|x_i^T - \bar{x}^T\right\|^2 \le \sum_{i\in\mathcal{H}}\left\|x_i^T - x^\star\right\|^2 \le \frac{4}{\mu_F^2}\sum_{i\in\mathcal{H}}\left\|\nabla F(x_i^T)\right\|^2.$$

Dividing both sides by h and using the definition of $\mathcal{C}_t$ in (3.6) gives

$$\mathcal{C}_T = \frac{1}{h}\sum_{i\in\mathcal{H}} \left\|x_i^T - \bar{x}^T\right\|^2 \le \frac{4}{h\mu_F^2}\sum_{i\in\mathcal{H}} \left\|\nabla F(x_i^T)\right\|^2. \tag{B.19}$$

Next, evaluate (B.18) at x_i^T, sum over honest nodes, and take expectations. Since $e_{i,T} = \mathbb{E}[f_i(x_i^T) - f_i^\star]$, we obtain

$$(1-\delta_2)^2 \sum_{i\in\mathcal{H}} \mathbb{E}\left[\left\|\nabla F(x_i^T)\right\|^2\right] \le 2L(1+\varepsilon)\sum_{i\in\mathcal{H}} e_{i,T} + h(1+\varepsilon^{-1})\delta_1^2. \tag{B.20}$$

By taking expectation in (B.19) and inserting (B.20), we obtain

$$\mathbb{E}[\mathcal{C}_T] \le \frac{4}{\mu_F^2(1-\delta_2)^2}\left(\frac{2L(1+\varepsilon)}{h}\sum_{i\in\mathcal{H}} e_{i,T} + (1+\varepsilon^{-1})\delta_1^2\right).$$

Finally, by applying the bounds on $e_{i,T}$ in (B.15) and (B.16), and setting $\varepsilon = 1$, we obtain (B.11) and (B.14), respectively. □

C LASA: Proofs and Additional Results

In this section, we prove the entropy-regularized characterization and establish the aggregation budgets of LASA, then examine when cooperation in LASA yields improvements over self-only training through theoretical comparisons and a concrete quadratic example.

C.1 Entropy-Regularized Characterization and Robustness

We establish LASA's entropy-regularized characterization in Proposition 4.1 and verify that it admits the budgets in Proposition 4.2.

Lemma C.1. *Fix any $c_1, c_2, \dots, c_m \in \mathbb{R}$. Let $m \ge 1$ and $\tau > 0$, and define*

$$Z_\tau(c) = \sum_{k=1}^m \exp\{-\tau c_k\}, \quad q_j = \frac{\exp\{-\tau c_j\}}{Z_\tau(c)}, \quad j = 1, 2, \dots, m. \tag{C.1}$$

For every $p \in \mathcal{P}_{m-1}$, define $H(p) = -\sum_{j=1}^m p_j \log p_j$. Then

$$\sum_{j=1}^m p_j c_j - \frac{1}{\tau}H(p) = -\frac{1}{\tau}\log Z_\tau(c) + \frac{1}{\tau}D_{\mathrm{KL}}(p\|q) \ge -\frac{1}{\tau}\log Z_\tau(c). \tag{C.2}$$

Consequently, $q = (q_1, q_2, \dots, q_m)$ is the unique minimizer of the entropy-regularized objective:

$$q = \operatorname{argmin}_{p\in\mathcal{P}_{m-1}} \left\{\sum_{j=1}^m p_j c_j - \frac{1}{\tau}H(p)\right\},$$

and it satisfies the following equality

$$\sum_{j=1}^m q_j c_j = -\frac{1}{\tau}\log Z_\tau(c) + \frac{1}{\tau}H(q). \tag{C.3}$$

Proof. Since $q_j > 0$, $p \in \mathcal{P}_{m-1}$, and $\log q_j = -\tau c_j - \log Z_\tau(c)$, we have

$$\begin{aligned} D_{\mathrm{KL}}(p\|q) &= \sum_{j=1}^m p_j \log\frac{p_j}{q_j} = \sum_{j=1}^m p_j \log p_j - \sum_{j=1}^m p_j\left(-\tau c_j - \log Z_\tau(c)\right) \\ &= -H(p) + \tau\sum_{j=1}^m p_j c_j + \log Z_\tau(c). \end{aligned}$$

Dividing both sides by τ and rearranging gives (C.2). Since $D_{\mathrm{KL}}(p\|q) \ge 0$, with equality if and only if $p = q$, the minimizer is unique. Setting $p = q$ in (C.2) and rearranging gives (C.3). □

We are now ready to prove the entropy-regularized characterization (Proposition 4.1) and provide the explicit budgets (Proposition 4.2) of LASA.

Proof of Proposition 4.1. Apply Lemma C.1 with $c_j = -\Delta_{ij}^t$ and $\tau = \tau_t$. The vector in (C.1) is exactly the LASA weight vector in (4.1). Moreover, minimizing $-\sum_j p_j \Delta_{ij}^t - H(p)/\tau_t$ is equivalent to maximizing $\sum_j p_j \Delta_{ij}^t + H(p)/\tau_t$. Thus, we obtain

$$\omega_i^t = \left(\frac{\exp\{\tau_t \Delta_{ij}^t\}}{\sum_{k \in \mathcal{S}_i} \exp\{\tau_t \Delta_{ik}^t\}} \right)_{j \in \mathcal{S}_i} = \operatorname*{arg\,min}_{p \in \mathcal{P}_{m_i - 1}} \left\{ -\sum_{j \in \mathcal{S}_i} p_j \Delta_{ij}^t - \frac{1}{\tau_t} H(p) \right\}$$
$$= \operatorname*{arg\,max}_{p \in \mathcal{P}_{m_i - 1}} \left\{ \sum_{j \in \mathcal{S}_i} p_j \Delta_{ij}^t + \frac{1}{\tau_t} H(p) \right\}.$$

□

Proof of Proposition 4.2. For LASA, the weights in (4.1) are nonnegative and sum to one, proving (2.3). Fix any $i \in \mathcal{H}$. Setting $c_j = \ell_{ij}^t$ in (C.3) yields

$$\sum_{j \in \mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t = -\frac{1}{\tau_t} \log \left(\sum_{j \in \mathcal{S}_i} \exp\{-\tau_t \ell_{ij}^t\} \right) - \frac{1}{\tau_t} \sum_{j \in \mathcal{S}_i} \omega_{ij}^t \log \omega_{ij}^t.$$

Since $i \in \mathcal{S}_i$, by using $\sum_{j \in \mathcal{S}_i} \exp\{-\tau_t \ell_{ij}^t\} \geq \exp\{-\tau_t \ell_{ii}^t\}$, we obtain

$$\sum_{j \in \mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t \leq \ell_{ii}^t - \frac{1}{\tau_t} \sum_{j \in \mathcal{S}_i} \omega_{ij}^t \log \omega_{ij}^t.$$

The tight upper bound $H(p) \leq \log m$ for all $p \in \mathcal{P}_{m-1}$ further yields $-\sum_{j \in \mathcal{S}_i} \omega_{ij}^t \log \omega_{ij}^t \leq \log m_i$. Thus, we obtain

$$\sum_{j \in \mathcal{S}_i} \omega_{ij}^t (\ell_{ij}^t - \ell_{ii}^t) = \sum_{j \in \mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t - \ell_{ii}^t \leq \frac{\log m_i}{\tau_t} = \beta_{i,t}.$$

This proves (2.4) with the budgets in (4.3). Since m_i is fixed and τ_t is prescribed deterministically, these budgets are independent of the realized optimization trajectory and Byzantine messages. Together with (2.3), this establishes that LASA satisfies Definition 1 with aggregation budgets $\{\log m_i / \tau_t\}$. □

C.2 Additional Cooperation Comparisons and Proofs

We compare LASA with self-only updates at a common state and derive conditions for tighter convergence bounds than those of self-only training. The results cover both smooth convex objectives and the additional quadratic growth setting. Denote

$$x_{i,\mathrm{loc}}^{t+1} = \tilde{x}_{ii}^t, \qquad f_i(x_{i,\mathrm{loc}}^{t+1}) = \ell_{ii}^t. \tag{C.4}$$

Here, $x_{i,\mathrm{loc}}^{t+1}$ denotes a one-step update computed from the current LASA state x_i^t, rather than an iterate of a separate self-only training trajectory. Define the signed aggregation gain, the entropy of the LASA weights, and the best honest-candidate gain by

$$\Delta_{i,t}^{\mathrm{agg}} = \ell_{ii}^t - \sum_{j \in \mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t,$$
$$H_{i,t} = -\sum_{j \in \mathcal{S}_i} \omega_{ij}^t \log \omega_{ij}^t,$$
$$\Delta_{i,t}^{\mathrm{gain}} = \ell_{ii}^t - \min_{j \in \mathcal{S}_i \cap \mathcal{H}} \ell_{ij}^t \geq 0.$$

Lemma C.2. *Suppose that Assumptions 1 and 2 hold. For every honest node i and every t, the iterates $\{x_i^t\}$ generated by LASA satisfy*

$$f_i(x_i^{t+1}) \le f_i(x_{i,\mathrm{loc}}^{t+1}) - \Delta_{i,t}^{\mathrm{agg}}, \tag{C.5}$$

$$\Delta_{i,t}^{\mathrm{agg}} \ge \Delta_{i,t}^{\mathrm{gain}} - \frac{H_{i,t}}{\tau_t} \ge \Delta_{i,t}^{\mathrm{gain}} - \frac{\log m_i}{\tau_t}. \tag{C.6}$$

Consequently, either $\Delta_{i,t}^{\mathrm{agg}} > 0$ or, more conservatively, $\Delta_{i,t}^{\mathrm{gain}} > H_{i,t}/\tau_t$ implies

$$f_i(x_i^{t+1}) < f_i(x_{i,\mathrm{loc}}^{t+1}).$$

Proof. Convexity and the definition of $\Delta_{i,t}^{\mathrm{agg}}$ give

$$f_i(x_i^{t+1}) = f_i\left(\sum_{j\in\mathcal{S}_i} \omega_{ij}^t \tilde{x}_{ij}^t\right) \le \sum_{j\in\mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t = \ell_{ii}^t - \Delta_{i,t}^{\mathrm{agg}}.$$

The right-hand side is $f_i(x_{i,\mathrm{loc}}^{t+1}) - \Delta_{i,t}^{\mathrm{agg}}$ by (C.4), which proves (C.5). Applying the equality (C.3) in Lemma C.1 with $c_j = \ell_{ij}^t$, $\tau = \tau_t$, and $q_j = \omega_{ij}^t$ gives

$$\sum_{j\in\mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t = -\frac{1}{\tau_t} \log Z_{i,t} + \frac{H_{i,t}}{\tau_t}, \tag{C.7}$$

where $Z_{i,t} = \sum_{j\in\mathcal{S}_i} \exp\{-\tau_t \ell_{ij}^t\}$. If $\ell_{\min}^t = \min_{j\in\mathcal{S}_i} \ell_{ij}^t$, then $Z_{i,t} \ge \exp\{-\tau_t \ell_{\min}^t\}$, and hence

$$-\frac{1}{\tau_t} \log Z_{i,t} \le \ell_{\min}^t \le \min_{j\in\mathcal{S}_i\cap\mathcal{H}} \ell_{ij}^t.$$

Substitute this estimate into (C.7) and subtract the result from ℓ_{ii}^t:

$$\begin{aligned}\Delta_{i,t}^{\mathrm{agg}} &= \ell_{ii}^t - \sum_{j\in\mathcal{S}_i} \omega_{ij}^t \ell_{ij}^t \\ &\ge \ell_{ii}^t - \min_{j\in\mathcal{S}_i\cap\mathcal{H}} \ell_{ij}^t - \frac{H_{i,t}}{\tau_t} = \Delta_{i,t}^{\mathrm{gain}} - \frac{H_{i,t}}{\tau_t}.\end{aligned}$$

If $\Delta_{i,t}^{\mathrm{agg}} > 0$ or $\Delta_{i,t}^{\mathrm{gain}} > H_{i,t}/\tau_t$, then (C.5) yields $f_i(x_i^{t+1}) < f_i(x_{i,\mathrm{loc}}^{t+1})$. □

Now, we give the full global stationarity bound underlying Theorem 4.3.

Theorem C.3 (Full version of Theorem 4.3)**.** *Under Assumptions 1–3, suppose $0 < \gamma_t \le 1/L$ and let R be distributed as in Theorem 3.1. Then the iterates generated by LASA satisfy*

$$\mathbb{E}\left[\left\|\nabla F(x_i^R)\right\|^2\right] \le \mathcal{U}_{i,T}^{\mathrm{loc}} - \frac{2}{\chi_\rho \Gamma_T} \sum_{t=0}^{T-1} \mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}}\right] \tag{C.8}$$

$$\le \mathcal{U}_{i,T}^{\mathrm{loc}} + \frac{2}{\chi_\rho \Gamma_T} \sum_{t=0}^{T-1} \left(\frac{\mathbb{E}\left[H_{i,t}\right]}{\tau_t} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{gain}}\right]\right) \tag{C.9}$$

$$\le \mathcal{U}_{i,T}^{\mathrm{loc}} + \frac{2}{\chi_\rho \Gamma_T} \sum_{t=0}^{T-1} \left(\frac{\log m_i}{\tau_t} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{gain}}\right]\right). \tag{C.10}$$

Proof. Smoothness at x_i^t and (C.5) imply

$$f_i(x_i^{t+1}) \le f_i(x_i^t) - \gamma_t \left\langle \nabla f_i(x_i^t), g_i^t \right\rangle + \frac{L\gamma_t^2}{2} \left\|g_i^t\right\|^2 - \Delta_{i,t}^{\mathrm{agg}}.$$

As in (B.2), by taking conditional expectation, using conditional unbiasedness, and then substituting (B.3), we obtain

$$\begin{aligned}\mathbb{E}\left[f_i(x_i^{t+1}) \mid \mathcal{F}_t\right] &\le f_i(x_i^t) - \gamma_t \left\|\nabla f_i(x_i^t)\right\|^2 + \frac{L\gamma_t^2}{2}\left(\left\|\nabla f_i(x_i^t)\right\|^2 + \sigma^2\right) - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}} \mid \mathcal{F}_t\right] \\ &= f_i(x_i^t) - \gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\left\|\nabla f_i(x_i^t)\right\|^2 + \frac{L\gamma_t^2\sigma^2}{2} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}} \mid \mathcal{F}_t\right].\end{aligned}$$

Since $\gamma_t \le 1/L$, we have

$$\mathbb{E}\left[f_i(x_i^{t+1}) \mid \mathcal{F}_t\right] \le f_i(x_i^t) - \frac{\gamma_t}{2}\left\|\nabla f_i(x_i^t)\right\|^2 + \frac{L\gamma_t^2\sigma^2}{2} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}} \mid \mathcal{F}_t\right]. \tag{C.11}$$

Substituting (B.4) into (C.11), we obtain

$$\mathbb{E}\left[f_i(x_i^{t+1}) \mid \mathcal{F}_t\right] \le f_i(x_i^t) - \frac{\gamma_t\chi_\rho}{2}\left\|\nabla F(x_i^t)\right\|^2 + \frac{\gamma_t\delta_1^2}{2\rho} + \frac{L\gamma_t^2\sigma^2}{2} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}} \mid \mathcal{F}_t\right].$$

Take total expectation, move the gradient term to the left, and sum from $t = 0$ to $T-1$. The objective terms telescope, and $f_i(x_i^T) \ge f_i^\star$, so

$$\frac{\chi_\rho}{2}\sum_{t=0}^{T-1}\gamma_t\mathbb{E}\left[\left\|\nabla F(x_i^t)\right\|^2\right] \le f_i(x_i^0) - f_i^\star + \frac{\delta_1^2}{2\rho}\Gamma_T + \frac{L\sigma^2}{2}\sum_{t=0}^{T-1}\gamma_t^2 - \sum_{t=0}^{T-1}\mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}}\right].$$

Use (B.6), divide by $\chi_\rho\Gamma_T/2$, and recognize the first three resulting terms as $\mathcal{U}_{i,T}^{\mathrm{loc}}$. This proves (C.8). Next, taking expectations in (C.6) gives

$$-\mathbb{E}[\Delta_{i,t}^{\mathrm{agg}}] \le \frac{\mathbb{E}[H_{i,t}]}{\tau_t} - \mathbb{E}[\Delta_{i,t}^{\mathrm{gain}}].$$

Substitution term by term proves (C.9). Using the inequality $\mathbb{E}[H_{i,t}] \le \log m_i$ further proves (C.10). □

Cooperation Comparison Under the Additional Quadratic Growth Condition

Under the additional quadratic growth condition, we identify when cooperation yields a tighter last-iterate local objective bound than self-only training. For the quadratic growth comparison, set

$$P_{r,T} = \prod_{j=r}^{T-1}(1 - \mu_{\mathrm{PL}}\gamma_j)$$

with the convention that an empty product equals 1, so $P_{T,T} = 1$. Define

$$\mathcal{U}_{i,T}^{\mathrm{loc,qg}} = P_{0,T}e_{i,0} + \sum_{j=0}^{T-1}P_{j+1,T}\frac{L\gamma_j^2\sigma^2}{2},$$

which is the self-only bound obtained by iterating (B.17) with $\beta_{i,t} = 0$.

Theorem C.4 (Quadratic growth comparison with self-only training)**.** *Under Assumptions 1, 2, and 4, suppose $0 < \gamma_t \le 1/L$. Then the iterates generated by LASA satisfy*

$$e_{i,T} \le \mathcal{U}_{i,T}^{\mathrm{loc,qg}} - \sum_{t=0}^{T-1}P_{t+1,T}\mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}}\right] \tag{C.12}$$

$$\le \mathcal{U}_{i,T}^{\mathrm{loc,qg}} + \sum_{t=0}^{T-1}P_{t+1,T}\left(\frac{\mathbb{E}\left[H_{i,t}\right]}{\tau_t} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{gain}}\right]\right). \tag{C.13}$$

LASA has a strictly smaller bound whenever the contraction-weighted sum of $\mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}}\right]$ is positive. A sufficient condition is

$$\sum_{t=0}^{T-1}P_{t+1,T}\mathbb{E}\left[\Delta_{i,t}^{\mathrm{gain}}\right] > \sum_{t=0}^{T-1}P_{t+1,T}\frac{\mathbb{E}\left[H_{i,t}\right]}{\tau_t}.$$

Proof. Take total expectation in (C.11). Since (B.7) implies $\mathbb{E}\left\|\nabla f_i(x_i^t)\right\|^2 \ge 2\mu_{\mathrm{PL}}e_{i,t}$, collecting the two gap terms gives

$$e_{i,t+1} \le (1 - \mu_{\mathrm{PL}}\gamma_t)e_{i,t} + \frac{L\gamma_t^2\sigma^2}{2} - \mathbb{E}\left[\Delta_{i,t}^{\mathrm{agg}}\right].$$

Write $q_t = 1 - \mu_{\mathrm{PL}}\gamma_t$. By repeated substitution in this recursion, we obtain

$$e_{i,T} \le \left(\prod_{u=0}^{T-1} q_u\right) e_{i,0} + \sum_{j=0}^{T-1} \left(\prod_{u=j+1}^{T-1} q_u\right) \left(\frac{L\gamma_j^2\sigma^2}{2} - \mathbb{E}[\Delta_{i,j}^{\mathrm{agg}}]\right)$$
$$= P_{0,T}e_{i,0} + \sum_{j=0}^{T-1} P_{j+1,T}\frac{L\gamma_j^2\sigma^2}{2} - \sum_{j=0}^{T-1} P_{j+1,T}\mathbb{E}[\Delta_{i,j}^{\mathrm{agg}}].$$

The first two terms are exactly $\mathcal{U}_{i,T}^{\mathrm{loc,qg}}$, proving (C.12). Moreover, $\mu_{\mathrm{PL}} \le L$ and $\gamma_t \le 1/L$ imply $0 \le q_t \le 1$, hence $P_{t+1,T} \ge 0$. We can therefore multiply $-\mathbb{E}[\Delta_{i,t}^{\mathrm{agg}}] \le \mathbb{E}[H_{i,t}]/\tau_t - \mathbb{E}[\Delta_{i,t}^{\mathrm{gain}}]$ by each $P_{t+1,T}$ and sum, which proves (C.13). □

C.3 A Quadratic Example of Cooperation

We construct a heterogeneous quadratic example with a specified Byzantine strategy to illustrate the cooperation condition (4.4) and a strict improvement over self-only training. The example also shows how retaining an honest candidate with higher local loss can improve the global objective relative to inverse-loss weighting.

Construction. Consider a complete graph with honest nodes $\mathcal{H} = \{1, 2\}$ and one Byzantine node $\mathcal{B} = \{3\}$. The honest local objectives and their global average are

$$f_1(x) = \frac{1}{2}(x-1)^2 + 1, \qquad f_2(x) = \frac{1}{2}(x+1)^2 + 1, \qquad F(x) = \frac{1}{2}x^2 + \frac{3}{2}.$$

The assumptions of Theorem 4.3 hold with $L = 1$, $\sigma = 0$, $\delta_1 = 1$, and $\delta_2 = 0$, since

$$\nabla f_1(x) - \nabla F(x) = -1, \qquad \nabla f_2(x) - \nabla F(x) = 1.$$

Each honest node uses exact gradients and the constant stepsize $\gamma_t = 1/2$. Initialize the honest models at

$$(x_1^0, x_2^0) = (7, -1),$$

and use a constant inverse temperature $\tau_t = \tau > 0$. At every round, the Byzantine node sends receiver-dependent candidates

$$\tilde{x}_{13}^t = 1, \qquad \tilde{x}_{23}^t = -1.$$

This is an oracle attack in which these messages are the receivers' respective local minimizers. All comparisons below use this fixed Byzantine strategy.

The honest candidates generated in round t are

$$\tilde{x}_1^t = \frac{x_1^t + 1}{2}, \qquad \tilde{x}_2^t = \frac{x_2^t - 1}{2}.$$

In particular, the first-round honest candidates are $(4, -1)$.

The condition (4.4)**.** At node 1 in the first round, the best honest-candidate gain is

$$\Delta_{1,0}^{\mathrm{gain}} = f_1(4) - f_1(-1) = \frac{11}{2} - 3 = \frac{5}{2}.$$

At every subsequent round t, we have $\Delta_{1,t}^{\mathrm{gain}} \ge 0$ and $H_{1,t} \le \log 3$ since each node has three candidates. For any fixed integer $T \ge 1$, choose

$$\tau > \frac{2T\log 3}{5}. \tag{C.14}$$

The trajectories are deterministic, and hence

$$\sum_{t=0}^{T-1} \Delta_{1,t}^{\mathrm{gain}} \ge \frac{5}{2} > \frac{T\log 3}{\tau} \ge \sum_{t=0}^{T-1} \frac{H_{1,t}}{\tau}.$$

Thus, condition (4.4) holds at node 1.

Strict improvement over self-only trajectories. Under (C.14), the first LASA iterates are

$$x_1^1 = \frac{4e^{-9\tau/2} - e^{-2\tau} + 1}{e^{-9\tau/2} + e^{-2\tau} + 1}, \qquad x_2^1 = -1 + \frac{5}{1 + 2e^{25\tau/2}}.$$

By (C.14), we have $e^{5\tau/2} > 3^T \geq 3$, and thus $3e^{-9\tau/2} < 2e^{-2\tau}$. Hence,

$$-1 < x_1^1 < 1, \qquad -1 < x_2^1 < 1.$$

All subsequent LASA iterates ($t > 1$) lie in the interval $(-1, 1)$. Indeed, if both honest models lie in this interval, then

$$\tilde{x}_1^t \in (0, 1), \qquad \tilde{x}_2^t \in (-1, 0).$$

The Byzantine candidates are the endpoints 1 and -1. Since LASA assigns strictly positive weights to all candidates, each aggregate remains strictly inside $(-1, 1)$. Induction therefore gives

$$|x_i^t| < 1, \qquad i = 1, 2, \quad t \geq 1.$$

For comparison, let u_i^t denote self-only iterates with the same initialization and stepsize. Their trajectories are

$$u_1^t = 1 + 6 \cdot 2^{-t}, \qquad u_2^t = -1, \qquad t \geq 0.$$

Consequently,

$$F(x_i^t) < 2 = \min_{j=1,2} F(u_j^t), \qquad i = 1, 2, \quad t \geq 1.$$

Thus, after choosing τ for the fixed horizon T, the strict trajectory comparison holds at every positive iteration, including those beyond T. Since $\nabla F(x) = x$, the same strict comparison holds for the squared global gradients.

Comparison with inverse-loss weighting. Initialize IL at the same honest models and use the same exact gradients, stepsize, and Byzantine messages. At node 2 in the first round, both the self-candidate and the Byzantine candidate equal -1, with local loss $f_2(-1) = 1$. The honest neighbor's candidate is 4, with local loss $f_2(4) = 27/2$. IL therefore excludes the honest neighbor and averages the two identical remaining candidates, yielding

$$x_{2,\mathrm{IL}}^1 = -1.$$

Under (C.14), LASA retains the higher-loss honest candidate with positive weight and satisfies

$$F(x_2^1) < 2 = F(x_{2,\mathrm{IL}}^1).$$

In fact, the IL iterate at node 2 remains at -1 at every round: its self-candidate is the unique minimizer of f_2, so any retained candidate must also equal -1. Hence, for all $t \geq 1$,

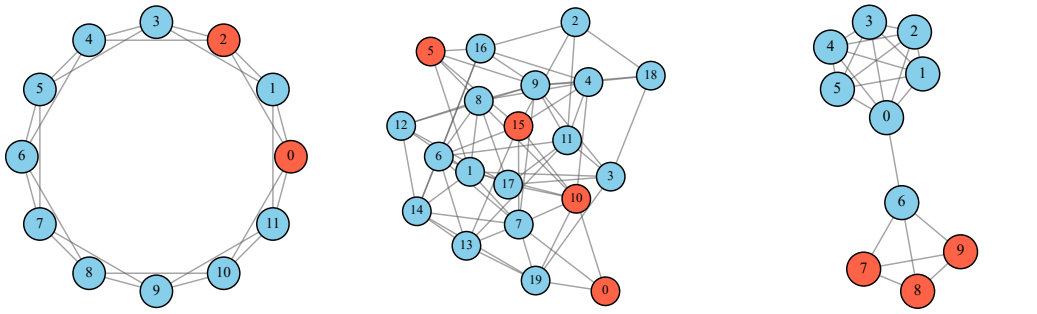
$$F(x_2^t) < 2 = F(x_{2,\mathrm{IL}}^t).$$

D Experimental Details and Additional Results

We provide the experimental setup and evaluation procedures for reproducing the MNIST and CIFAR-10 experiments, together with additional results supplementing the main text.

D.1 Communication Graphs

We describe the three communication graphs used in Section 5, including a setting with a local Byzantine majority. The regular graph in Figure 3a has 12 nodes, each connected to four neighbors, with two Byzantine nodes. The Erdős–Rényi graph in Figure 3b has 20 nodes, with each undirected edge included independently with probability 0.35, and four Byzantine nodes. The graph in Figure 3c has seven honest nodes and three Byzantine nodes; in particular, honest node 6 is connected to one honest neighbor and all three Byzantine nodes, creating a local Byzantine majority. Blue and red nodes indicate honest and Byzantine nodes, respectively.



(a) Regular graph (b) Erdős–Rényi graph (c) Byzantine majority graph

Figure 3: Graphs used in our experiments.

Table 3: Model architectures and experimental settings. The stepsize grid specifies constant stepsizes for MNIST and initial stepsizes for CIFAR-10. LASA settings refer to the main comparisons; the sensitivity studies vary the inverse temperature or the loss-evaluation batch size separately.

Dataset	MNIST			CIFAR-10
Model	Multiclass logistic regression			ResNet-20
Architecture details	Linear ($784 \to 10$)			(He et al., 2016)
Loss function	Cross-entropy			Cross-entropy
Local optimizer	SGD			SGD
Training batch size	32			64
Momentum	0.0			0.0
Weight decay	5×10^{-4}			5×10^{-4}
Stepsize grid	$\{0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1.0\}$			
Stepsize schedule	Constant			Cosine annealing
Communication rounds	$T = 2000$			$T = 20000$
Inverse temperature τ	3.0			$\{5, 7.5, 10\}$
Loss batch size	full			128
Communication graph	Regular	Erdős–Rényi	Byzantine majority	Regular
Total nodes	12	20	10	12
Honest nodes	10	16	7	10
Byzantine nodes	2	4	3	2
Maximum number of Byzantine neighbors b'	2	3	3	2
Graph configuration	Degree 4	$p = 0.35$	Figure 3c	Degree 4
Dirichlet parameter	$\alpha = 0.5$	$\alpha = 0.5$	$\alpha \in \{0.5, 10\}$	$\alpha = 0.5$
Number of seeds	5	5	5	5

D.2 Model Architecture and Hyperparameters

We specify the data preprocessing, model architectures, and training hyperparameters for MNIST and CIFAR-10. Table 3 summarizes the main settings. Unless otherwise stated, we partition the training data across nodes using a Dirichlet distribution with parameter $\alpha = 0.5$.

For MNIST, we use multiclass logistic regression without data augmentation. Inputs are scaled to $[0, 1]$ and normalized with mean 0.1307 and standard deviation 0.3081. For CIFAR-10, we use the

ResNet-20 model (He et al., 2016), replacing batch normalization with group normalization using eight groups (Wu & He, 2018). Training augmentation consists of random 32×32 crops with four-pixel padding and random horizontal flips. Inputs are scaled to $[0, 1]$ and normalized channel-wise with means $(0.4914, 0.4822, 0.4465)$ and standard deviations $(0.2470, 0.2435, 0.2616)$.

For the MNIST experiments, each run consists of $T = 2000$ communication rounds. We tune the constant stepsize over $\gamma \in \{0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1.0\}$ and use a constant inverse temperature $\tau = 3.0$ for LASA. Each node uses the full local dataset to evaluate local losses, except in the ablation study for loss-evaluation batch sizes.

For the CIFAR-10 experiments, we apply the cosine annealing learning rate schedule, and tune the initial learning rate over $\gamma \in \{0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1.0\}$. We run each algorithm for 20000 rounds. For the inverse temperature of LASA, we search over $\tau \in \{5.0, 7.5, 10.0\}$ and use a loss-evaluation batch size of 128.

D.3 Adversarial Settings

Let z_j^t denote node j's local candidate model before message manipulation. We specify the benign setting and five attacks used in our experiments as follows.

- Benign: all Byzantine nodes follow the DSGD protocol in Algorithm 1.
- Gaussian (Dong et al., 2024): each Byzantine node replaces its outgoing candidate with a random vector whose coordinates are sampled independently from the standard Gaussian distribution $\mathcal{N}(0, 1)$.
- Label Flip (Allen-Zhu et al., 2021; Allouah et al., 2023): each training label y at a Byzantine node is replaced with $9 - y$ for the ten-class tasks considered here. Byzantine nodes then train using the corrupted labels.
- Sign Flip (Allen-Zhu et al., 2021): each Byzantine node b replaces its outgoing candidate with $-z_b^t$, reversing the entire model vector.
- Fall-of-Empire (FOE) (Xie et al., 2020): each Byzantine node sends $-\bar{z}_{\mathcal{H}}^t$ to its neighbors, where $\bar{z}_{\mathcal{H}}^t = h^{-1} \sum_{j \in \mathcal{H}} z_j^t$ is the mean of all honest candidates.
- Dissensus (He et al., 2023): each Byzantine node constructs a receiver-specific message using the receiving node's candidate, its honest neighbors' candidates, and the mixing weights (chosen according to the Metropolis weights). Let $\bar{\omega}_{ij}$ denote the fixed mixing weights used to construct the attack. For each honest receiver i with positive Byzantine-neighbor weight $B_i = \sum_{b \in \mathcal{S}_i \cap \mathcal{B}} \bar{\omega}_{ib}$, every Byzantine neighbor sends

$$\tilde{x}_{ib}^t = z_i^t - \frac{1}{B_i} \sum_{j \in \mathcal{S}_i \cap \mathcal{H}} \bar{\omega}_{ij} (z_j^t - z_i^t).$$

  This message opposes the weighted displacement induced by honest neighbors. Byzantine messages are identical for a given receiver but may differ across receivers.

Sign Flip and FOE act on model parameters in our implementation. FOE and Dissensus assume oracle access to honest candidates; Dissensus additionally uses the mixing weights and may send different models to different neighbors.

D.4 Aggregators

We compare LASA with the following baselines, each implemented through the aggregation step (A.1).

- DSGD: aggregates candidates using fixed mixing weights in (2.2).
- Iterative outlier scissor (IOS) (Wu et al., 2023) is a filtering-based aggregator that requires an estimate b' of the local Byzantine count. At each honest node i, IOS initializes the retained set as $\mathcal{S}_i$, including its own candidate. It then repeatedly computes the normalized weighted average of the retained candidates and removes the neighboring candidate farthest from this average in Euclidean distance, while always retaining the local candidate $\tilde{x}_{ii}^t$. The

average is recomputed after each removal. After discarding b' neighboring candidates, IOS returns the weighted average of the remaining candidates, with their original mixing weights renormalized to sum to one.

- Nearest-neighbor averaging (NNA) (Farhadkhani et al., 2023) is a filtering-based aggregator that also requires an estimate b' of the local Byzantine count. At each honest node i, NNA retains the $|\mathcal{S}_i| - b'$ candidates with the smallest distances to the self-candidate and returns their unweighted arithmetic mean.
- ClippedGossip (CG) (He et al., 2023; Han et al., 2025): At each honest node i, CG computes $d_{ij}^t = \tilde{x}_{ij}^t - \tilde{x}_{ii}^t$ and replaces each candidate with $\tilde{x}_{ii}^t + \min\{1, c/\|d_{ij}^t\|_2\} d_{ij}^t$ if $\|d_{ij}^t\| > 0$, where $c > 0$ is the clipping threshold. If $d_{ij}^t = 0$, the candidate is left unchanged. The resulting candidates are aggregated using the original mixing weights. We tune c over $\{0.1, 0.5, 1.0\}$ and keep the selected threshold fixed across nodes and communication rounds, as suggested by experiments in He et al. (2023); Han et al. (2025).
- IL (Li et al., 2020): at each honest node i, the inverse-loss aggregator assigns weights to candidates in (2.2) according to

$$\omega_{ij}^t = \begin{cases} \dfrac{1/\ell_{ij}^t}{\sum_{k \in \mathcal{S}_i^{+,t}} 1/\ell_{ik}^t}, & \text{if } j \in \mathcal{S}_i^{+,t}, \\ 0, & \text{otherwise,} \end{cases}$$

  where $\mathcal{S}_i^{+,t} = \{j \in \mathcal{S}_i : \ell_{ij}^t \le \ell_{ii}^t\}$.
- Self-only training (SELF): each honest node runs SGD on its local data without communication, corresponding to

$$\omega_{ij}^t = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise} \end{cases}$$

  in the weighted-average update (2.2).

For LASA and IL, the local loss used for candidate evaluation includes the same weight decay regularization term as the training objective, and therefore corresponds to f_i in our theoretical formulation. For methods that require a mixing matrix, we use weights constructed by the Metropolis rule (Xiao & Boyd, 2004). For IOS and NNA, we use a common Byzantine-count estimate at all honest nodes and keep it fixed throughout training:

$$b' = \max_{i \in \mathcal{H}} \{|\mathcal{B} \cap \mathcal{S}_i|\}.$$

This choice is suggested by the theoretical analysis (Wu et al., 2023; Gaucher et al., 2025), ensuring that the prescribed filtering budget is at least the actual number of Byzantine neighbors at every honest node. This gives these filtering-based baselines more prior information than LASA. Specifically, we set $b' = 2$ for the regular graph (Figure 3a) and $b' = 3$ for both the Erdős–Rényi graph (Figure 3b) and the Byzantine majority graph (Figure 3c).

On the selection of baselines. The baselines cover filtering-based methods (IOS and NNA), a clipping method (CG), and loss-based weighting (IL). IOS and NNA represent filtering-based approaches that continue to receive attention in recent work (Gaucher et al., 2025; Ajibuwa et al., 2026; Rangwala et al., 2026). CG adapts the centered-clipping method from centralized Byzantine-robust optimization (Karimireddy et al., 2021) to decentralized learning and has been further modified in the decentralized literature (Yang & Ghaderi, 2024; Han et al., 2025). IL provides a direct comparison with LASA because both use local losses to determine aggregation weights. DSGD and SELF provide reference points for standard communication and training without communication, respectively.

D.5 Experimental Protocol

Hyperparameter selection. For both MNIST and CIFAR-10, we hold out 10% of the original training set for validation and tune hyperparameters separately for each aggregator and setting. After hyperparameter selection, we fix the chosen configuration and retrain each method on the full original training set using five random seeds $\{1, 2, 3, 4, 5\}$. In the inverse temperature study, we retune the learning rate for each inverse temperature. In the loss-evaluation batch size study, we keep the learning rate fixed across batch sizes.

Evaluation protocol. All honest nodes are evaluated on the same test set. For each seed, MTA and WTA are respectively the mean and minimum test accuracy over honest nodes, and CE is the model disagreement $\mathcal{C}_T$ defined in (3.6). MTA, WTA, and CE are all computed at the final communication round separately for each seed. The communication graph, Byzantine-node assignment, and data partition are fixed across the five seeds. We report the means and standard deviations of MTA, WTA, and CE across five seeds. In the benign setting, the designated Byzantine nodes follow the training protocol honestly, while evaluation remains restricted to the predetermined honest-node set $\mathcal{H}$ to keep the evaluation population consistent across settings.

Local-loss evaluation. LASA and IL use the same local-loss evaluation protocol. Each node evaluates all candidate models on the same local data. MNIST experiments use the full local dataset unless otherwise stated. For CIFAR-10 and the sampled-loss experiments on MNIST, each node resamples an evaluation minibatch at every communication round and uses that batch for all candidates.

D.6 Additional Results

Table 4 presents the MNIST experiments on the Erdős–Rényi graph.

Table 4: MNIST logistic regression on the Erdős–Rényi graph (Figure 3b) with a Dirichlet partition ($\alpha = 0.5$). MTA: mean test accuracy; CE: consensus error; WTA: worst-node test accuracy. Entries show mean $\pm$ standard deviation over five seeds. Bold marks the best unrounded mean per attack and metric. All values are rounded for display, with CE means above 10^4 abbreviated.

	Benign	Gaussian	Sign Flip	Label Flip	FOE	Dissensus
Mean test accuracy (MTA, %) ↑						
DSGD	$\mathbf{91.95_{\pm 0.10}}$	$53.68_{\pm 1.74}$	$63.68_{\pm 1.95}$	$86.19_{\pm 0.20}$	$61.12_{\pm 2.99}$	$77.97_{\pm 0.36}$
IOS	$90.27_{\pm 0.13}$	$90.05_{\pm 0.13}$	$90.05_{\pm 0.13}$	$90.05_{\pm 0.14}$	$90.05_{\pm 0.13}$	$89.98_{\pm 0.19}$
NNA	$90.54_{\pm 0.36}$	$90.28_{\pm 0.46}$	$90.32_{\pm 0.44}$	$90.29_{\pm 0.46}$	$90.28_{\pm 0.46}$	$90.20_{\pm 0.50}$
CG	$91.83_{\pm 0.06}$	$91.53_{\pm 0.06}$	$84.74_{\pm 0.42}$	$88.23_{\pm 0.10}$	$84.34_{\pm 0.26}$	$\mathbf{91.12_{\pm 0.04}}$
IL	$86.68_{\pm 0.70}$	$85.37_{\pm 1.67}$	$84.95_{\pm 1.11}$	$85.62_{\pm 0.68}$	$86.36_{\pm 1.58}$	$86.22_{\pm 0.75}$
SELF	$73.02_{\pm 0.18}$	$73.02_{\pm 0.18}$	$73.02_{\pm 0.18}$	$73.02_{\pm 0.18}$	$73.02_{\pm 0.18}$	$73.02_{\pm 0.18}$
LASA	$91.94_{\pm 0.12}$	$\mathbf{91.77_{\pm 0.05}}$	$\mathbf{91.77_{\pm 0.04}}$	$\mathbf{91.15_{\pm 0.12}}$	$\mathbf{91.77_{\pm 0.04}}$	$90.48_{\pm 0.36}$
Worst-node test accuracy (WTA, %) ↑						
DSGD	$91.31_{\pm 0.19}$	$34.58_{\pm 4.48}$	$48.81_{\pm 4.63}$	$84.26_{\pm 0.35}$	$40.25_{\pm 6.23}$	$60.33_{\pm 0.17}$
IOS	$81.17_{\pm 0.53}$	$81.17_{\pm 0.53}$	$81.17_{\pm 0.53}$	$81.17_{\pm 0.53}$	$81.17_{\pm 0.53}$	$81.17_{\pm 0.53}$
NNA	$86.78_{\pm 2.26}$	$86.50_{\pm 2.34}$	$86.82_{\pm 2.33}$	$86.61_{\pm 2.31}$	$86.50_{\pm 2.34}$	$86.65_{\pm 2.09}$
CG	$\mathbf{91.66_{\pm 0.05}}$	$91.08_{\pm 0.14}$	$76.40_{\pm 4.68}$	$85.48_{\pm 0.91}$	$76.27_{\pm 3.32}$	$\mathbf{89.32_{\pm 0.46}}$
IL	$76.55_{\pm 4.84}$	$73.19_{\pm 6.77}$	$73.31_{\pm 2.06}$	$73.94_{\pm 3.55}$	$76.39_{\pm 5.00}$	$73.77_{\pm 1.43}$
SELF	$60.39_{\pm 0.18}$	$60.39_{\pm 0.18}$	$60.39_{\pm 0.18}$	$60.39_{\pm 0.18}$	$60.39_{\pm 0.18}$	$60.39_{\pm 0.18}$
LASA	$91.37_{\pm 0.28}$	$\mathbf{91.42_{\pm 0.10}}$	$\mathbf{91.42_{\pm 0.10}}$	$\mathbf{89.23_{\pm 1.14}}$	$\mathbf{91.43_{\pm 0.09}}$	$86.15_{\pm 1.61}$
Consensus error (CE) ↓						
DSGD	$0.005_{\pm 0.000}$	$95.19_{\pm 1.201}$	$0.017_{\pm 0.001}$	$\mathbf{0.001_{\pm 0.000}}$	$0.019_{\pm 0.001}$	$29.99_{\pm 0.181}$
IOS	$1.460_{\pm 0.001}$	$1.492_{\pm 0.004}$	$1.492_{\pm 0.004}$	$1.492_{\pm 0.004}$	$1.492_{\pm 0.004}$	$1.484_{\pm 0.004}$
NNA	$0.027_{\pm 0.006}$	$0.029_{\pm 0.006}$	$0.028_{\pm 0.006}$	$0.028_{\pm 0.006}$	$0.029_{\pm 0.006}$	$0.030_{\pm 0.006}$
CG	$\mathbf{0.001_{\pm 0.000}}$	$\mathbf{0.001_{\pm 0.000}}$	$0.019_{\pm 0.001}$	$0.003_{\pm 0.000}$	$0.006_{\pm 0.000}$	$\mathbf{0.003_{\pm 0.000}}$
IL	$0.272_{\pm 0.044}$	$0.419_{\pm 0.150}$	$0.427_{\pm 0.086}$	$0.358_{\pm 0.046}$	$0.348_{\pm 0.057}$	$0.362_{\pm 0.055}$
SELF	$20.12_{\pm 0.04}$	$20.12_{\pm 0.04}$	$20.12_{\pm 0.04}$	$20.12_{\pm 0.04}$	$20.12_{\pm 0.04}$	$20.12_{\pm 0.04}$
LASA	$0.004_{\pm 0.000}$	$0.001_{\pm 0.000}$	$\mathbf{0.001_{\pm 0.000}}$	$0.015_{\pm 0.002}$	$\mathbf{0.001_{\pm 0.000}}$	$0.024_{\pm 0.003}$

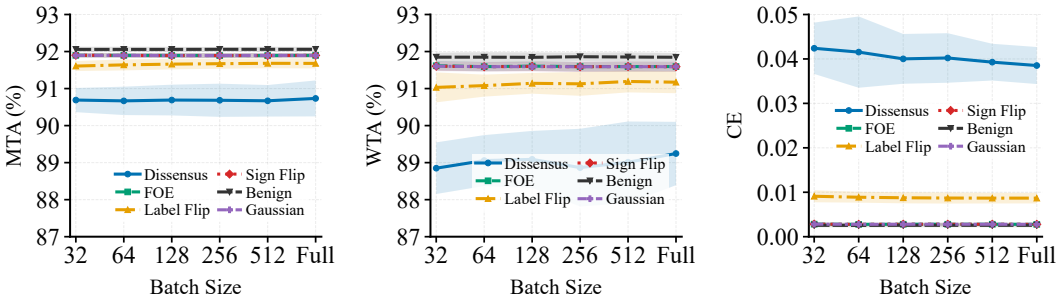


Figure 4: Sensitivity of LASA to the batch size used for local-loss evaluation on MNIST over the regular graph (Figure 3a) under benign and adversarial settings. From left to right: mean test accuracy, worst-node test accuracy, and consensus error.

Figure 4 provides the batch size sensitivity results discussed in Section 5.