ADAPTIVE DISCOUNTING OF COMPRESSION ERROR FEEDBACK FOR DECENTRALIZED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized learning enables collaborative model training across multiple nodes without relying on a central server. Each node performs local updates using its own data and exchanges compressed updates with neighboring nodes. Although numerous methods have been proposed for decentralized non-convex optimization under communication constraints, correcting compression error while mitigating the effect of network topology, particularly in sparse networks, remains challenging. We propose A-DEFEAT-GT (adaptively discounted error feedback algorithm for decentralized learning with gradient tracking), which combines error-feedback with an adaptively selected discount coefficient and gradient tracking, without a momentum term. We also propose A-DEFEAT, a one-message variant that removes gradient tracking for less complexity and communication cost. The analysis gives explicit dependence on the spectral gap and compression level under stated stability conditions. The experiments show strong performance on sparse networks. A-DEFEAT-GT attains the highest test accuracy in the principal fixed-aggregation comparisons, while A-DEFEAT outperforms all existing baselines in several settings despite sending half as many vectors per aggregation as the two-message tracking baselines, and it attains the highest test accuracy in multiple fixed-communication comparisons.

1 INTRODUCTION

In recent years, decentralized learning has emerged as an important paradigm in machine learning due to its advantages in computational scalability for parallel computing Nedic & Ozdaglar (2009); Lian et al. (2017), communication efficiency Tang et al. (2019; 2018); Koloskova et al. (2019); Pu & Nedić (2021), and preservation of data privacy (Wangni et al., 2018; Reisizadeh et al., 2019). Unlike centralized learning, training is distributed across nodes connected by a decentralized communication network, which collaboratively train a shared model without a central server. Each node performs local updates on its own data, whose distribution may differ across nodes, and communicates only with its neighbors in the network topology.

A fundamental challenge in decentralized learning is communication efficiency. To reduce communication costs, recent works focus on designing gossip-based algorithms with *compressed communication* (Wangni et al., 2018; Koloskova et al., 2019; Zhao et al., 2022; Huang & Pu, 2024; Yau & Wai, 2023; Islamov et al., 2024). However, compression inevitably introduces information loss that accumulates over iterations. To mitigate this effect, *error-feedback* mechanisms have been proposed, which accumulate and compensate for the information lost due to compression. Recent studies incorporate error-feedback in centralized optimization (Stich et al., 2018; Wu et al., 2018; Karimireddy et al., 2019). It is non-trivial to incorporate error-feedback in a decentralized setting. One reason is that error-feedback propagates past compression residuals among nodes through multi-hop communication, which can persist along poorly mixed graphs and eventually dominate the diminishing gradients. Discounted error-feedback is one way to limit the impact of stale residuals, ensuring that error correction remains proportional to the current gradient values and maintaining stability over sparse networks. Discounted error-feedback has been introduced in centralized settings Gao et al. (2023). Nevertheless, extending it to decentralized networks is challenging since compression residual errors propagate across nodes and interact with slow consensus processes. Figure 1 illustrates how using a fixed discount coefficient on the error-feedback can lead to *under-correction*

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.

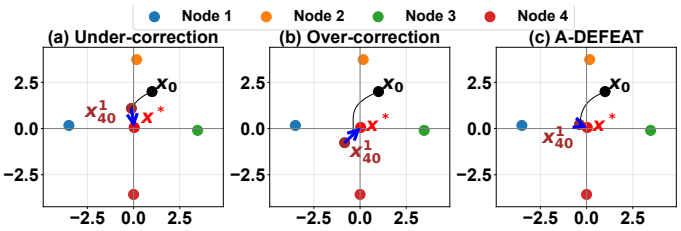


Figure 1. Motivating example with different choices of the coefficient applied to the residual correction. The trajectory from $x_0^1 = x_0$ to x_{40}^1 illustrates how node 1's model parameters evolve from round $t = 0$ to round $t = 40$ under different coefficient choices. The red point denotes the global minimizer x^*, the black point represents the common initial model $x_0^1 = x_0$, and the four colored points represent the nodes' local data means. The brown point indicates node 1's model after 40 updates. The network is a four-node ring. The under-correction and over-correction panels use the fixed values $\gamma = 0.1$ and $\gamma = 1.0$, respectively. The blue arrow shows the distance between x_{40}^1 and x^*. Further details can be found in Appendix A.

or *over-correction*, especially during long training runs. If the coefficient is too small, the residual may fail to correct the local update sufficiently, as in Figure 1(a). If it is too large, the residual may dominate the combined update and misdirect the trajectory, as in Figure 1(b). A single fixed coefficient can therefore be difficult to tune across different stages of training.

Existing decentralized learning methods with compressed communication, such as (Zhao et al., 2022; Yau & Wai, 2023; Islamov et al., 2024), still face two major limitations. First, their performance can deteriorate over sparse network topologies because their convergence bounds depend strongly on the network spectral gap. Second, they rely on gradient tracking, momentum propagation, or auxiliary consensus variables, requiring the transmission of additional gradient-related information at every communication round and thereby increasing communication overhead. These limitations can restrict scalability and efficiency in bandwidth-constrained decentralized networks. Motivated by these challenges, we ask the following question.

Can we design a lightweight decentralized learning algorithm with compressed communication that automatically controls the amount of error-feedback, achieving competitive convergence over sparse networks?

To answer this question, we propose an adaptively discounted error-feedback algorithm for decentralized learning with gradient tracking, or A-DEFEAT-GT, and its variant A-DEFEAT without gradient tracking to save communication cost. These algorithms use discounted error-feedback with a coefficient selected from the current update and residual norms. The coefficient scales past residual errors before they are reintegrated into the local update. Figure 1(c) illustrates how this adaptive coefficient avoids the under-correction and over-correction seen in the fixed-coefficient examples. Our main contributions are summarized as follows.

- We propose A-DEFEAT-GT, a decentralized learning algorithm that combines gradient tracking with an adaptively discounted error-feedback correction. The discount coefficient is computed at every iteration from the current tracker and residual norms, so it adapts the feedback magnitude without a searched discount hyperparameter. We also develop A-DEFEAT, which removes the gradient tracker and communicates one compressed vector per aggregation instead of two.
- We analyze both algorithms for non-convex objectives under explicit stability margins. For A-DEFEAT-GT, the network-dependent powers improve to $\rho^{-2/3}$ and ρ^{-1}, compared with $\rho^{-5/3}$ and ρ^{-3} in MoTEF and the ρ^{-3} transient power in BEER. For A-DEFEAT, under bounded gradient divergence, the network-dependent stochastic and heterogeneity contributions scale as $\rho^{-1/3}$ and $\rho^{-2/3}$, respectively.
- To validate our approach, we perform extensive experiments on standard benchmarks (FMNIST, KMNIST, CIFAR-10, and CIFAR-100) using deep neural networks, on networks of up to 200 nodes, with different data heterogeneity levels, network topologies, and top-k compression ratios. In the fixed-aggregation comparisons, A-DEFEAT-GT attains the highest accuracy, including gains of 6.1% on FMNIST with 50 nodes and 6.3% and 5.5% on CIFAR-

10 and CIFAR-100. A-DEFEAT also outperforms all five existing baselines on three of the four datasets in the 10-node comparison while using half the vector transmissions of BEER and MoTEF, and it attains the highest mean test accuracy in several fixed-communication settings. The appendix presents all additional comparisons, including cases in which another method is competitive.

2 Related Works

For decentralized learning, gossip algorithms are one of the most popular approaches (Kempe et al., 2003; Johansson & Johansson, 2008; Boyd et al., 2006; Lian et al., 2017). In a gossip algorithm, each node first computes the local update vector using SGD and sends it to its neighboring nodes. Given specific network topologies and standard assumptions, Lian et al. (2017) demonstrated an $\mathcal{O}(1/\sqrt{nT})$ convergence rate for gossip algorithms, where n is the number of nodes and T is the number of iterations. Notably, this convergence rate matches that of centralized SGD, where a central server aggregates the SGDs and updates the model parameters in each iteration.

Compressed communication is a common technique for reducing communication overhead between neighboring nodes. Typical methods include biased compressors such as sparsification. These techniques are widely used in the federated learning literature (Basu et al., 2019; Haddadpour et al., 2021; Alistarh et al., 2017; Bernstein et al., 2018; Shlezinger et al., 2020; Reisizadeh et al., 2020; Wangni et al., 2018; Sattler et al., 2019; Albasyoni et al., 2020; Gorbunov et al., 2021; Alistarh et al., 2018; Stich & Karimireddy, 2020; Karimireddy et al., 2019), as well as the decentralized learning literature (Wangni et al., 2018; Reisizadeh et al., 2019; Tang et al., 2019; 2018). However, compression causes information loss. Applying it directly can degrade the convergence of a gossip algorithm and can even cause divergence at high compression levels.

In centralized settings, Stich et al. (2018) and Wu et al. (2018) introduced error-compensated stochastic gradient descent with compressed communication and analyzed convex or strongly convex objectives. Karimireddy et al. (2019) analyzed error-feedback for sign-SGD and other compressors for weakly convex and non-convex objectives. Instead of adding the full residual, Gao et al. (2023) introduced a coefficient that discounts the error-feedback term. These centralized algorithms rely on global aggregation to maintain a shared model and residual, whereas decentralized learning has multiple local models and residuals that propagate through multi-hop communication.

Koloskova et al. (2019) introduced a consensus step size multiplying the compressed gossip correction. BEER (Zhao et al., 2022) and MoTEF (Islamov et al., 2024) likewise multiply the model and tracking consensus corrections by a mixing gain, while also incorporating gradient or momentum tracking. Such damping is a standard way to stabilize the feedback between compressed surrogates and network mixing. It makes the effective mixing matrix more self-weighted, at the cost of a smaller effective spectral gap. These tracking-based methods also transmit an additional gradient-related vector at every iteration, which increases communication cost. Several recent works have also explored adaptive mechanisms in decentralized learning. For instance, Aketi et al. (2024) introduces an adaptive consensus step based on momentum tracking. However, this adaptive strategy can reduce the amount of information exchanged among neighboring nodes, potentially slowing consensus and degrading performance over sparse networks. Similarly, Liu et al. (2025) introduces an additional coefficient into the local update using a momentum-based adaptive method, but this coefficient simultaneously scales the local gradient term and can undesirably reduce the gradient step. We are not aware of a prior decentralized method that selects the residual discount coefficient from the current update and residual magnitudes. This motivates the adaptive rule in A-DEFEAT-GT and A-DEFEAT.

3 Problem Formulation and Assumptions

We consider a decentralized optimization problem as follows.

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} f(\boldsymbol{x}), \qquad f(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^N$ represents the model parameters and $f_i(\boldsymbol{x}) \coloneqq \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{D}_i}[F_i(\boldsymbol{x};\boldsymbol{\zeta})]$ for $i \in \{1, 2, \ldots, n\}$. Here $\mathcal{D}_i$ is node i's local data distribution, $\boldsymbol{\zeta}$ is a sample from that distribution, and $F_i(\boldsymbol{x};\boldsymbol{\zeta})$ is the

corresponding sample loss. At each gradient evaluation, node i draws $\zeta_t^i \sim \mathcal{D}_i$ independently of all randomness revealed before that evaluation and independently across nodes.

In decentralized learning, such as decentralized parallel stochastic gradient descent (D-PSGD) (Lian et al., 2017), each node communicates only with its neighbors. These communications are encoded by a mixing matrix $W \in \mathbb{R}^{n\times n}$. Each entry $W_{ij} \in [0,1]$ is the weight that node i assigns to node j; a nonzero off-diagonal entry represents a communication edge. We make the following standard assumption on the W, as used in the literature (Tang et al., 2018; Koloskova et al., 2019; Tang et al., 2019; Zhao et al., 2022; Huang & Pu, 2024; Yau & Wai, 2023; Islamov et al., 2024).

Assumption 3.1 (Connectivity matrix)**.** For $n \geq 2$, the matrix W is entrywise nonnegative, symmetric, and doubly stochastic. Thus $W = W^T$ and $W\mathbf{1}_n = \mathbf{1}_n$, where $\mathbf{1}_n$ is the all-ones vector. Let $J := \mathbf{1}_n\mathbf{1}_n^\top/n$ be the averaging matrix and assume in addition that $\|W - J\|_2 < 1$. Index the non-consensus eigenvalues by decreasing absolute value, so that $1 = \lambda_1(W) > |\lambda_2(W)| \geq \cdots \geq |\lambda_n(W)|$. We denote the spectral gap as

$$\rho := 1 - \|W - J\|_2 = 1 - |\lambda_2(W)| \in (0, 1]. \tag{2}$$

Also define the extremal spectral quantities

$$\lambda_- := \min_i \lambda_i(W), \quad \mu := \max_{i\geq 2} |1 - \lambda_i(W)| = \|I - W\|_2 = 1 - \lambda_-, \quad \nu := 1 + \lambda_- = 2 - \mu > 0. \tag{3}$$

The spectral gap is commonly used as a metric for the network topology (Zhao et al., 2022; Islamov et al., 2024; Koloskova et al., 2019; Huang & Pu, 2024; Yau & Wai, 2023).

In this paper, we consider decentralized learning with compressed communication. In particular, we apply the contractive compression operators commonly used in the literature (Islamov et al., 2024; Gao et al., 2023; Zhao et al., 2022; Tang et al., 2019) and defined as follows.

Definition 3.2 (Contractive compressor)**.** For a fixed $0 \leq \beta < 1$, the possibly randomized compression operator satisfies, for every input b,

$$\mathbb{E}_{\mathcal{C}}\|\mathcal{C}(b) - b\|^2 \leq \beta\|b\|^2, \tag{4}$$

where the expectation is over the compressor randomness. The same parameter β is used for the model and tracker compressors.

We make the following additional standard assumptions.

Assumption 3.3 (Smoothness)**.** Assume that there is an $L > 0$ such that every local objective function $f_i(\cdot)$ satisfies

$$\|\nabla f_i(\boldsymbol{y}) - \nabla f_i(\boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N. \tag{5}$$

Assume that the global objective is bounded below by a finite f^*, and define $\Delta := f(x_0) - f^* < \infty$.

Assumption 3.4 (Unbiased stochastic gradients)**.** For every node i and every $x \in \mathbb{R}^N$, assume

$$\mathbb{E}_{\zeta\sim\mathcal{D}_i}[\nabla F_i(x;\zeta)] = \nabla f_i(x), \qquad \mathbb{E}_{\zeta\sim\mathcal{D}_i}\left[\|\nabla F_i(x;\zeta) - \nabla f_i(x)\|^2\right] \leq \sigma^2. \tag{6}$$

Assumption 3.5 (Bounded gradient divergence)**.** There is an $\epsilon \geq 0$ such that, for every $x \in \mathbb{R}^N$,

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \epsilon^2. \tag{7}$$

4 Proposed Algorithms and Convergence Analysis

The stepsizes η and τ are selected before a run and held fixed across its iterations. Their selected values may depend on the horizon and on the problem, network, and compression parameters.

4.1 Adaptive clipping-range selection

Let $d_t^i := u_t^i$ for A-DEFEAT-GT and $d_t^i := g_t^i$ for A-DEFEAT . The discount coefficient compares the current update magnitude with the residual magnitude.

$$\gamma_t^i := \begin{cases} \mathrm{clip}_{[g_c, 2-g_c]}\left(\eta\|d_t^i\|/\|e_t^i\|\right), & \|e_t^i\| > 0, \\ 1, & \|e_t^i\| = 0, \end{cases} \qquad g_c := 1 - 1/\sqrt{2}. \tag{8}$$

Here $\mathrm{clip}_{[a,b]}(r) := \min\{b, \max\{a, r\}\}$. g_c denotes the clipping endpoint, while d_t^i denotes the update direction used by the selected algorithm.

4.2 A-DEFEAT-GT

Algorithm 1 A-DEFEAT-GT

1 **Input.** Fixed symmetric stochastic matrix W, fixed initial point x_0, horizon T, model stepsize $\eta > 0$ and tracker stepsize $\tau > 0$ used at every iteration, and consensus gain $0 < \omega \leq 1$.
2 At every node initialize $x_0^i = x_0$, $e_0^i = 0$, $\hat{u}_0^i = 0$; initialize neighbor caches $x_0^{i,j} = x_0$, $\hat{u}_0^{i,j} = 0$ for $W_{ij} > 0$.
3 Draw independent initial samples, store $g_0^i = \nabla F_i(x_0; \zeta_0^i)$, and set $u_0^i = g_0^i$.
4 **for** $t = 0, \ldots, T-1$ **do**
5 **for** each node i, in parallel **do**
6 Set $\gamma_t^i = \mathrm{clip}_{[g_c, 2-g_c]}(\eta\|u_t^i\|/\|e_t^i\|)$, where $g_c \coloneqq 1 - 1/\sqrt{2}$; set $\gamma_t^i = 1$ if $e_t^i = 0$.
7 $b_t^i = \omega\big(\sum_{j|W_{ij}>0} W_{ij} x_t^{i,j} - x_t^i\big) - \eta u_t^i + \gamma_t^i e_t^i$.
8 $v_t^i = \mathcal{C}_{t,i}^x(b_t^i)$; set $x_{t+1}^i = x_t^i + v_t^i$ and $e_{t+1}^i = b_t^i - v_t^i + (1-\gamma_t^i)e_t^i$.
9 After forming x_{t+1}^i, draw a fresh sample ζ_{t+1}^i and set $g_{t+1}^i = \nabla F_i(x_{t+1}^i; \zeta_{t+1}^i)$.
10 $u_{t+1}^i = u_t^i + \tau\omega\big(\sum_{j|W_{ij}>0} W_{ij}\hat{u}_t^{i,j} - \hat{u}_t^i\big) + g_{t+1}^i - g_t^i$; retain g_{t+1}^i for the next iteration.
11 $z_t^i = \mathcal{C}_{t,i}^u(u_{t+1}^i - \hat{u}_t^i)$; set $\hat{u}_{t+1}^i = \hat{u}_t^i + z_t^i$.
12 Send v_t^i, z_t^i to neighbors; receive v_t^j, z_t^j from each neighbor j.
13 Set $x_{t+1}^{i,j} = x_t^{i,j} + v_t^j$ and $\hat{u}_{t+1}^{i,j} = \hat{u}_t^{i,j} + z_t^j$ for $W_{ij} > 0$.
14 **end for**
15 **end for**

Algorithm 1 introduces a node wise and adaptive choice of the error-feedback coefficient γ_t^i. Each node selects it from the relative magnitudes of its tracked update and accumulated compression residual, adjusting residual reinjection (lines 6–8). The algorithm combines this adaptive rule with compressed gradient tracking.

Initialization and caches. Lines 2–3 initialize every model at the common point x_0, set the residuals and communicated tracker surrogates to zero, draw g_0^i, and set $u_0^i = g_0^i$. The neighbor caches use the same initial values. In lines 12–13, node j and every neighbor caching its state add the same increments v_t^j and z_t^j. Consequently, $x_t^{i,j} = x_t^j$ and $\hat{u}_t^{i,j} = \hat{u}_t^j$ whenever $W_{ij} > 0$.

Model update. Line 6 computes γ_t^i from the relative magnitudes of ηu_t^i and e_t^i and clips it to $[g_c, 2 - g_c]$. The convention $\gamma_t^i = 1$ when $e_t^i = 0$ avoids an undefined ratio without changing the update. Line 7 forms b_t^i from three parts. The first is the weighted neighbor consensus correction, the second is the tracked descent direction $-\eta u_t^i$, and the third reinjects the discounted residual $\gamma_t^i e_t^i$.

Residual update. Line 8 compresses b_t^i into v_t^i and applies it to the model. The new residual combines the current compression error $b_t^i - v_t^i$ with the carryover $(1-\gamma_t^i)e_t^i$. Substituting this residual update into the model update shows that the total residual contribution is $e_t^i - e_{t+1}^i$. It therefore telescopes across iterations even though γ_t^i is selected adaptively.

Gradient tracking. Line 9 evaluates g_{t+1}^i after forming the new model. Line 10 updates the tracker using the consensus correction and $g_{t+1}^i - g_t^i$, then retains g_{t+1}^i for the next iteration.

Tracker compression and communication. Line 11 compresses the innovation $u_{t+1}^i - \hat{u}_t^i$ into z_t^i and adds it to the local surrogate. Lines 12–13 send v_t^i and z_t^i and update the corresponding neighbor caches before the next iteration. Thus each iteration sends one compressed model increment and one compressed tracker increment to every neighbor. Each randomized compression call uses a new seed independent of its input and previously revealed randomness.

4.3 A-DEFEAT AS A ONE-MESSAGE VARIANT

A-DEFEAT is obtained from Algorithm 1 by removing the tracker u_t^i, its compressed surrogate $\hat{u}_t^i$, and the tracker message z_t^i. At iteration t, node i computes a local stochastic gradient $g_t^i = \nabla F_i(x_t^i; \zeta_t^i)$, replaces u_t^i by g_t^i in the clipping rule, and forms $b_t^i = \omega\left(\sum_j W_{ij} x_t^{i,j} - x_t^i\right) - \eta g_t^i + \gamma_t^i e_t^i$. It then compresses b_t^i, updates the model, residual, and neighbor caches exactly as in the model part of A-DEFEAT-GT, and communicates only the model increment v_t^i. Appendix E gives the complete

algorithm and proof. Removing gradient tracking reduces each aggregation from two compressed vectors to one and eliminates the tracker state and tracker compression. The residual-corrected average is still driven by the mean local stochastic gradient. What changes is the disagreement bound, which now uses the bounded-gradient-divergence assumption (Assumption 3.5) to control persistent differences among local objectives. A-DEFEAT-GT remains preferable when arbitrary or extreme data heterogeneity must be covered without a uniform divergence bound.

4.4 Convergence Analysis

Write $X_t \coloneqq [x_t^1, \dots, x_t^n]$, $F_t \coloneqq [\nabla f_1(x_t^1), \dots, \nabla f_n(x_t^n)]$, and $G_t \coloneqq [g_t^1, \dots, g_t^n]$. Also write $\bar{x}_t \coloneqq X_t \mathbf{1}_n/n$, $\Pi \coloneqq I - \mathbf{1}_n\mathbf{1}_n^\top/n$, and $\mathcal{R}_T \coloneqq T^{-1}\sum_{t<T}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2$. Define

$$\begin{aligned} d_{\mathrm{sp}} &\coloneqq \frac{2\mu}{2-\mu}, & \vartheta &\coloneqq g_{\mathrm{c}} - \sqrt{\beta}(d_{\mathrm{sp}} + g_{\mathrm{c}}) > 0, \\ m &\coloneqq 1 - \sqrt{\beta}(1+2\tau\mu) > 0, & g_{\mathrm{c}} &\coloneqq 1 - 1/\sqrt{2}. \end{aligned} \tag{9}$$

Theorem 4.1 (A-DEFEAT-GT with explicit network dependence)**.** *When Assumptions 3.1, 3.3, and 3.4 hold and the compressor satisfies Definition 3.2, use Algorithm 1 with consensus gain $\omega = 1$ and $0 < \tau \le \min\{1, 1/\mu\}$. If $\min\{\nu, \vartheta, m, \tau\} = \Omega(1)$, let $D \ge \Delta$ be any positive number and let η be specified in equation 34 of Appendix D.1. Then*

$$\mathcal{R}_T = O\left(\frac{\sigma\sqrt{LD}}{\sqrt{nT}} + \frac{(LD\sigma)^{2/3}}{\rho^{2/3}T^{2/3}} + \frac{LD + \mathcal{J}_0}{\rho T}\right), \qquad \mathcal{J}_0 \coloneqq \frac{\mathbb{E}\|G_0\|_F^2}{n}. \tag{10}$$

Main proof idea. BEER and MoTEF control compression, consensus, and tracking errors with separate one-step estimates (Zhao et al., 2022; Islamov et al., 2024). Here the model update contains successive compression residuals as the single difference $e_t^i - e_{t+1}^i$. Propagating this difference through the disagreement recursion gives gain $2/\nu$; bounding the two residuals separately would instead introduce an inverse spectral-gap factor. The tracker similarly uses the gradient difference $g_{t+1}^i - g_t^i$. Keeping its stochastic-noise difference intact gives a tracker noise-filter gain independent of ρ, rather than amplifying noise by an inverse gap. The compression bound and the positive margins in equation 9 then control the coupled residual and tracker errors. Finally, set $y_t \coloneqq \bar{x}_t + \bar{e}_t$, where $\bar{e}_t \coloneqq n^{-1}\sum_i e_t^i$. Since the tracker consensus corrections sum to zero and $u_0^i = g_0^i$, the mean tracker equals the mean stored gradient. Averaging the model update and adding the new mean residual cancels the residual terms and gives $y_{t+1} = y_t - \eta n^{-1}\sum_i g_t^i$. Smoothness, the disagreement and residual bounds, and the gradient-noise assumption give descent for $f(y_t)$. Appendix D.1 gives the full derivation and constants. Theorem F.2 gives the corresponding result for every $0 < \omega \le 1$ that satisfies its transformed stability margins. Corollary F.3 then gives a sufficient choice of ω for every fixed $\beta < 1$ and states the resulting explicit dependence on β and ρ.

Theorem 4.2 (A-DEFEAT with separate noise and heterogeneity terms)**.** *When Assumptions 3.1, 3.3, 3.4, and 3.5 hold and the compressor satisfies Definition 3.2, use Algorithm 2 with consensus gain $\omega = 1$. If $\min\{\nu, \vartheta\} = \Omega(1)$, let $D \ge \Delta$ be any positive number and let η be specified in equation 64 of Appendix E.2. Then*

$$\mathcal{R}_T = O\left(\frac{\sigma\sqrt{LD}}{\sqrt{nT}} + \frac{(LD)^{2/3}}{T^{2/3}}\left(\frac{\sigma^2}{\rho} + \frac{\epsilon^2}{\rho^2}\right)^{1/3} + \frac{LD}{\rho T}\right). \tag{11}$$

The complete proof appears in Appendices E.2 and F. Theorem F.4 gives the corresponding result for every $0 < \omega \le 1$ that satisfies its transformed stability margin. Corollary F.5 then gives a sufficient choice of ω for every fixed $\beta < 1$ and states the resulting explicit dependence on β, ρ, and ϵ.

Improved dependence on the spectral gap. At a fixed compression level and under the stated positive stability margins, A-DEFEAT-GT has a network-dependent stochastic term of order $\rho^{-2/3}T^{-2/3}$ and a transient and initialization term of order $\rho^{-1}T^{-1}$. These improve on the corresponding powers $\rho^{-4/3}T^{-2/3}$ for CHOCO-SGD (Koloskova et al., 2020) and $\rho^{-5/3}T^{-2/3}$ for MoTEF (Islamov et al., 2024), while the T^{-1} network power improves from ρ^{-3} in MoTEF and BEER (Zhao et al., 2022) to ρ^{-1}. A-DEFEAT has the sharper network-dependent stochastic term, of order $\rho^{-1/3}T^{-2/3}$ when $\epsilon = 0$, whereas persistent gradient divergence contributes $\rho^{-2/3}T^{-2/3}$. Thus A-DEFEAT can have a

better ρ dependence than A-DEFEAT-GT when the local gradients agree, while A-DEFEAT-GT does not require bounded gradient divergence. Both methods have a $\rho^{-1}T^{-1}$ term. These comparisons concern the powers of ρ in the proven convergence bounds, under each method's stated assumptions. Appendix G gives the compression factors, initialization terms, and assumptions for every method.

5 EXPERIMENTS

In this section, we provide the experimental results including details on the model structures, datasets, compression schemes, baseline algorithms and performance comparison. The code for the experiments is available at `https://anonymous.4open.science/r/A-DEFEAT-38D8/`.

5.1 EXPERIMENTAL SETUP

Datasets and Models. We evaluate A-DEFEAT, A-DEFEAT-GT, and the baselines on FashionMNIST (FMNIST), KMNIST, CIFAR-10, and CIFAR-100 under different network topologies. FMNIST, KMNIST, and CIFAR-10 contain 10 image classes, while CIFAR-100 contains 100. For FMNIST, we use a two-layer fully connected network with a hidden width of 50 (0.04M parameters). For KMNIST, we use a convolutional network with four 3×3 layers of widths 32, 32, 64, and 128, two max-pooling layers, and two fully connected layers with a 256-unit hidden layer (0.24M parameters). For CIFAR-10 and CIFAR-100, we use a VGG-style network with five 3×3 convolutional layers of widths 64, 64, 128, 128, and 256, three max-pooling layers, and two fully connected layers with a 256-unit hidden layer (1.6M parameters); only the output dimension differs (10 or 100 classes). The convolutional models use group normalization and no dropout. Group normalization has no running statistics, allowing the node models to remain independent while their per-node gradients are computed in one vectorized pass, and it remains stable with small local mini-batches. Every node draws a mini-batch of 64 samples per iteration. Each experimental entry is the mean and standard deviation over three final runs with different random seeds.

Baselines and Parameter Selection. We compare the proposed algorithms with DCD (Tang et al., 2018), CHOCO (Koloskova et al., 2019), DeepSqueeze Tang et al. (2019), BEER (Zhao et al., 2022), and MoTEF (Islamov et al., 2024) on different network topologies. We use grid search (Canonaco et al., 2021) to select the hyperparameters for each algorithm. Appendix B gives the details.

Data Distribution. In our experiments, we use the Dirichlet distribution with parameter $\alpha > 0$ to control how different the local data sets are across the nodes. The Dirichlet partition is described in detail in (Tzu-Ming Harry et al., 2019). The data distribution becomes more heterogeneous when α decreases. We evaluate $\alpha = 0.05$, 0.2 and 5.0, together with a separately constructed partition labeled $\alpha = 0$; this label does not denote a Dirichlet distribution with a zero parameter. For FMNIST we use all four values, for KMNIST we use $\alpha = 0$, 0.05 and 0.2, and for CIFAR-10 and CIFAR-100 we use $\alpha = 0$ and 0.05.

Compression Operator. In our experiments, all algorithms use the same commonly used biased top-k compression operator, which keeps the k fraction of the entries with the largest magnitude and sets the remaining entries to zero. The operator is applied at each node to the whole flattened vector that the node sends, so a message contains $\lfloor kN \rfloor$ of the N entries, and the algorithms that send two messages per round compress each of them in the same way. The main experiments use $k = 0.05$. The additional results with different topologies and compressors can be found in Appendix I.

Communication accounting. One communication unit is one compressed vector transmitted over a node–neighbor link. A-DEFEAT, DCD, CHOCO, and DeepSqueeze send one such vector per aggregation, whereas A-DEFEAT-GT, BEER, and MoTEF send two. Therefore, under a fixed budget of communication units, a two-message method performs half as many aggregation rounds as a one-message method. The fixed-aggregation and fixed-communication tables present these two comparisons separately.

Implementation details. The experiments implement Algorithm 1 in Section 4.2 and Algorithm 2 in Appendix E.1. In A-DEFEAT-GT, each node retains g_t^i, forms the tracker increment from the gradient difference $g_{t+1}^i - g_t^i$, and uses the tracker direction u_t^i in the clipping rule. Both methods use the Euclidean norm and the interval $[g_c, 2 - g_c]$ in equation 8. Every experiment uses the unit consensus gain $\omega = 1$ for both proposed methods.

5.2 Experimental Results

We evaluate all algorithms using the globally averaged model $\bar{x}_t$ at each aggregation. In every figure, the x-axis is the number of aggregation rounds and the y-axis is the test accuracy or global loss, averaged over the last 50 rounds. Results at fixed communication costs appear in Appendix I.

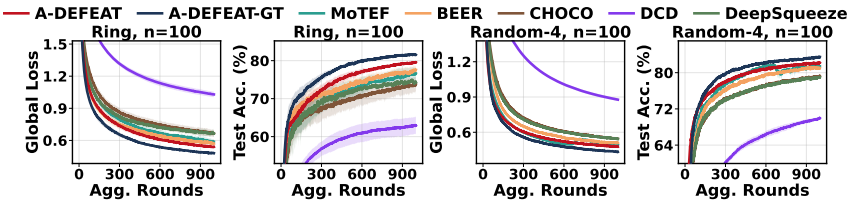


Figure 2. Global training loss and test accuracy on FashionMNIST with $\alpha = 0.05$ and 100 nodes, on a ring and on a random-4 graph (random-4 means each node randomly selects 4 neighbors). All methods use top-k compression with $k = 0.05$. Results are averaged over 3 seeds.

Table 1. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 aggregation rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	65.8 ± 0.2	65.8 ± 0.2	65.8 ± 0.2	65.7 ± 0.2	65.7 ± 0.2
CHOCO	65.1 ± 1.1	65.5 ± 1.4	65.8 ± 0.8	65.8 ± 0.8	65.7 ± 1.2
DeepSqueeze	66.1 ± 0.4	67.8 ± 1.5	66.7 ± 2.0	68.7 ± 1.7	70.5 ± 1.0
BEER	78.2 ± 0.1	78.2 ± 0.2	78.2 ± 0.2	78.2 ± 0.2	78.2 ± 0.2
MoTEF	79.6 ± 0.4	79.6 ± 0.1	79.3 ± 0.3	79.8 ± 0.3	80.1 ± 0.3
A-DEFEAT	75.2 ± 0.8	74.2 ± 0.0	74.2 ± 0.1	74.2 ± 0.1	74.3 ± 0.1
A-DEFEAT-GT	$\mathbf{80.3 \pm 0.3}$	$\mathbf{80.2 \pm 0.7}$	$\mathbf{79.6 \pm 0.4}$	$\mathbf{80.5 \pm 0.6}$	$\mathbf{80.2 \pm 0.1}$

Table 2. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 aggregation rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.1 ± 0.1	65.7 ± 1.2	62.1 ± 0.3	62.8 ± 2.2	65.3 ± 0.7
CHOCO	74.0 ± 0.4	75.0 ± 0.1	68.1 ± 0.4	73.5 ± 2.0	76.3 ± 0.4
DeepSqueeze	71.8 ± 0.6	75.5 ± 0.1	68.7 ± 0.5	74.5 ± 0.5	76.5 ± 0.5
BEER	79.2 ± 0.6	79.8 ± 0.2	73.8 ± 0.3	77.2 ± 1.3	76.8 ± 0.6
MoTEF	80.9 ± 0.1	80.0 ± 0.6	74.9 ± 0.9	76.5 ± 1.7	75.5 ± 0.2
A-DEFEAT	78.0 ± 0.5	79.9 ± 0.3	76.2 ± 0.3	79.4 ± 0.2	78.8 ± 0.1
A-DEFEAT-GT	$\mathbf{81.4 \pm 0.2}$	$\mathbf{81.7 \pm 0.1}$	$\mathbf{81.0 \pm 0.4}$	$\mathbf{81.6 \pm 0.4}$	$\mathbf{80.8 \pm 0.4}$

Table 3. Test accuracy % comparison. The experiments use a 50-node ring under different levels of data heterogeneity. This table summarizes the final FashionMNIST test accuracy of each algorithm after 1000 aggregation rounds.

Algorithms	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.2$	$\alpha = 5$
DCD	65.8 ± 0.2	62.1 ± 0.3	70.2 ± 0.9	79.3 ± 0.3
CHOCO	65.8 ± 0.8	68.1 ± 0.4	81.1 ± 0.1	83.6 ± 0.1
DeepSqueeze	66.7 ± 2.0	68.7 ± 0.5	80.1 ± 0.2	83.2 ± 0.1
BEER	78.2 ± 0.2	73.8 ± 0.3	82.0 ± 0.4	83.8 ± 0.2
MoTEF	79.3 ± 0.3	74.9 ± 0.9	82.4 ± 0.1	83.6 ± 0.1
A-DEFEAT	74.2 ± 0.1	76.2 ± 0.3	82.6 ± 0.1	84.0 ± 0.1
A-DEFEAT-GT	$\mathbf{79.6 \pm 0.4}$	$\mathbf{81.0 \pm 0.4}$	$\mathbf{83.4 \pm 0.2}$	$\mathbf{84.1 \pm 0.2}$

Effects of Network Topology. We use ring networks with 10, 20, 50, 100, and 200 nodes and top-k sparsification with $k = 0.05$. In these experiments, every algorithm uses the same number of aggregation rounds. Table 2 presents the FashionMNIST results for $\alpha = 0.05$, and Table 1 presents the extreme partition labeled $\alpha = 0$, in which each node holds a single class. A-DEFEAT-GT achieves the highest accuracy at every network size. For $\alpha = 0.05$, its accuracy remains between 80.8% and 81.7% from 10 to 200 nodes. In contrast, MoTEF and DCD drop by 6.0 and 7.0 percentage points,

respectively, from 10 to 50 nodes. The margin of A-DEFEAT-GT over the best baseline grows from 0.5 percentage points at 10 nodes to 6.1 at 50 nodes and remains at least 4 percentage points at 100 and 200 nodes. For $\alpha = 0$, the three methods with gradient tracking reach 78%–81%, while DCD, CHOCO, and DeepSqueeze remain between 65% and 71%. Figure 2 shows the loss and accuracy curves for 100 nodes on the ring and random-4 graphs. A-DEFEAT-GT reaches the lowest training loss and highest test accuracy on both, with 0.483 and 81.6% on the ring, and 0.436 and 83.4% on the random-4 graph. Appendix I presents the fixed-communication comparisons.

Effects of Data Heterogeneity. Table 3 compares A-DEFEAT-GT with the baselines on Fashion-MNIST using a 50-node ring and a fixed number of aggregation rounds. Smaller positive values of the Dirichlet parameter α produce more heterogeneous local datasets. A-DEFEAT-GT attains the highest accuracy at every tested heterogeneity level, with its largest lead, 6.1 percentage points, at $\alpha = 0.05$. From $\alpha = 5$ to $\alpha = 0.05$, DCD loses 17.2 percentage points, CHOCO 15.5, DeepSqueeze 14.5, BEER 10.0, and MoTEF 8.7, while A-DEFEAT-GT loses 3.1. These results indicate greater empirical robustness to non-IID data in the tested settings.

Table 4. Test accuracy % comparison. The experiments use a ring network with 10 nodes across different datasets with the same data heterogeneity, controlled by the Dirichlet parameter $\alpha = 0.05$. All results are compared using 1000 aggregation rounds for FashionMNIST, 5000 aggregation rounds for KMNIST and 10000 aggregation rounds for CIFAR-10 and CIFAR-100. † marks a run whose curve stops before the reading round because a seed diverged; the table uses its last available value.

Algorithms	FashionMNIST	KMNIST	CIFAR-10	CIFAR-100
DCD	69.1 ± 0.1	72.6 ± 0.7	57.9 ± 2.1	25.0 ± 0.4
CHOCO	74.0 ± 0.4	65.5 ± 1.3	61.1 ± 1.3	24.1 ± 16.3 †
DeepSqueeze	71.8 ± 0.6	67.4 ± 0.5	62.1 ± 0.9	35.5 ± 0.6
BEER	79.2 ± 0.6	75.6 ± 0.8	64.6 ± 1.4	36.5 ± 0.8
MoTEF	80.9 ± 0.1	59.1 ± 6.2	56.2 ± 2.2	24.2 ± 8.8
A-DEFEAT	78.0 ± 0.5	82.4 ± 2.0	67.0 ± 1.2	39.2 ± 1.3
A-DEFEAT-GT	$\mathbf{81.4 \pm 0.2}$	$\mathbf{87.6 \pm 2.9}$	$\mathbf{70.9 \pm 0.5}$	$\mathbf{42.0 \pm 1.5}$

Results across Datasets. We evaluate both proposed algorithms on FMNIST, KMNIST, CIFAR-10, and CIFAR-100 using the models described in Section 5.1. Table 4 compares all methods after the same number of aggregation rounds on a 10-node ring with $\alpha = 0.05$. A-DEFEAT-GT reaches the highest accuracy on all four datasets. Its gain over the best existing baseline is 0.5 percentage points on FMNIST, 6.3 on CIFAR-10, and 5.5 on CIFAR-100. On KMNIST, the gain is 12.0 percentage points; several baselines are unstable under high heterogeneity, and MoTEF, for example, reaches 59.1% with a standard deviation of 6.2%. The one-message A-DEFEAT variant is also stronger than every existing baseline on KMNIST (82.4%), CIFAR-10 (67.0%), and CIFAR-100 (39.2%), although it sends only one compressed vector per aggregation rather than the two sent by BEER and MoTEF. With the same communication budget, A-DEFEAT is tied for or attains the highest mean test accuracy at every tested network size on the FashionMNIST ring for $\alpha = 0.2$ and $\alpha = 5$, as shown in Tables 10 and 11 of Appendix I. These comparisons identify settings in which the simpler variant is the most communication-efficient choice; they do not assert uniform superiority over all data distributions. Appendix I contains the remaining experimental settings.

6 Conclusion

We proposed A-DEFEAT-GT and A-DEFEAT, two decentralized methods with adaptively discounted compression residuals. A-DEFEAT-GT uses compressed gradient tracking without bounded gradient divergence, while A-DEFEAT removes tracking and sends one compressed vector per aggregation under a bounded-divergence assumption. By retaining differences of successive residuals and stochastic gradients, our non-convex analysis improves the proved dependence on network connectivity under the stated stability conditions. In four-dataset equal-aggregation experiments, A-DEFEAT-GT achieves the highest test accuracy throughout. A-DEFEAT exceeds existing baselines on three datasets while using half the compressed vectors of BEER and MoTEF, and leads or ties in several equal-communication comparisons.

AI USE STATEMENT

In this work, we used generative AI tools for mathematical check and code implementation assist. We have not used generative AI tools for the algorithm design and theoretical development, the problem formulation, the convergence analysis or its proofs, and synthetic data generation is not applicable to this work. Additionally, we used generative AI tools for figure and table generation, code editing, and readability editing of the manuscript. We have reviewed all AI-assisted work. The implementation was verified against an independent reference implementation of the same algorithms, every reported number is produced by the released code from stored per-seed records, and all AI-assisted text was read and revised by the authors. 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 studies optimization algorithms for decentralized training and uses only publicly available benchmark datasets (FashionMNIST, KMNIST, CIFAR-10 and CIFAR-100), which contain no personally identifiable information and no human subjects. We see no direct ethical concerns arising from the method itself.

REPRODUCIBILITY STATEMENT

The assumptions and the complete proofs of all theoretical claims are given in Appendix G. Algorithms 1 and 2 specify both methods. Appendices D and E give complete proofs. The experimental setup, including datasets, model architectures, data partition, compression operator and the full hyper-parameter search protocol, is described in Section 5.1, and the complete search ranges together with the selection procedure are stated in Appendix B. All reported numbers are averages over three random seeds that are not used anywhere in the hyper-parameter search, and the per-setting results for every dataset, topology and compression ratio are given in Appendix I. The source code that produces all experiments, tables and figures is available at `https://anonymous.4open.science/r/A-DEFEAT-38D8/`.

REFERENCES

Sai Aparna Aketi, Abolfazl Hashemi, and Kaushik Roy. AdaGossip—adaptive consensus step-size for decentralized deep learning with communication compression, 2024.

Alyazeed Albasyoni, Mher Safaryan, Laurent Condat, and Peter Richtárik. Optimal gradient compression for distributed and federated learning. *arXiv preprint 2010.03246*, 2020.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD—communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The convergence of sparsified gradient methods. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd—distributed sgd with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems*, 32, 2019.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD—compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Giuseppe Canonaco, Alex Bergamasco, Alessio Mongelluzzo, and Manuel Roveri. Adaptive federated learning in presence of concept drift. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2021. doi: 10.1109/IJCNN52387.2021.9533710.

Yuan Gao, Rustem Islamov, and Sebastian Stich. EControl—fast distributed optimization with compression and error control, 2023.

Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. MARINA—faster nonconvex distributed learning with compression. In *International Conference on Machine Learning*, pp. 3788–3798. PMLR, 2021.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression—unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021.

Kun Huang and Shi Pu. CEDAS—a compressed decentralized stochastic gradient method with improved convergence, 2024.

Rustem Islamov, Yuan Gao, and Sebastian U. Stich. Towards faster decentralized stochastic optimization with communication compression, 2024.

Björn Johansson and Mikael Johansson. Faster linear iterations for distributed averaging. *IFAC Proceedings Volumes*, 41(2):2861–2866, 2008.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, 2019.

David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 482–491. IEEE, 2003.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pp. 3478–3487, 2019.

Anastasia Koloskova, Tao Lin, Sebastian U. Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*, 2020. arxiv.org/abs/1907.09356.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *NeurIPS*, pp. 5330–5340, 2017.

Wei Liu, Anweshit Panda, Ujwal Pandey, Christopher Brissette, Yikang Shen, George M. Slota, Naigang Wang, Jie Chen, and Yangyang Xu. Compressed decentralized momentum stochastic gradient methods for nonconvex optimization, 2025.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009. doi: 10.1109/TAC.2008.2009515.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. An exact quantized decentralized gradient descent algorithm. *IEEE Transactions on Signal Processing*, 67(19):4934–4947, 2019.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq—a communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031. PMLR, 2020.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.

Nir Shlezinger, Mingzhe Chen, Yonina C Eldar, H Vincent Poor, and Shuguang Cui. UVeQFed—universal vector quantization for federated learning. *IEEE Transactions on Signal Processing*, 69: 500–514, 2020.

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework—better rates for sgd with delayed gradients and compressed updates. *The Journal of Machine Learning Research*, 21(1):9613–9648, 2020.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory, 2018.

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Hanlin Tang, Xiangru Lian, Shuang Qiu, Lei Yuan, Ce Zhang, Tong Zhang, and Ji Liu. Deepsqueeze—decentralization meets error-compensated compression. *arXiv preprint 1907.07346*, 2019.

Hsu Tzu-Ming Harry, Qi Hang, and Brown Matthew. Measuring the effects of non-identical data distribution for federated visual classification. *Neurips Workshop on Federated Learning*, 2019.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *NeurIPS*, 2018.

Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized sgd and its applications to large-scale distributed optimization, 2018.

Chung-Yiu Yau and Hoi-To Wai. DoCoM—compressed decentralized optimization with near-optimal sample complexity, 2023.

Haoyu Zhao, Boyue Li, Zhize Li, Peter Richtarik, and Yuejie Chi. BEER—fast o(1/t) rate for decentralized nonconvex optimization with communication compression. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31653–31667. Curran Associates, Inc., 2022.

APPENDIX CONTENTS

A Details of Motivating Example **14**

B Grid-search procedure **14**

C Clipping bounds **15**

D Proof of A-DEFEAT-GT **15**

D.1 Finite-horizon guarantee for A-DEFEAT-GT 15

D.2 Sparse graph families and the role of the consensus gain 20

E A-DEFEAT **21**

E.1 A-DEFEAT algorithm 21

E.2 Finite-horizon guarantee for A-DEFEAT 21

F A constructive consensus gain for every contractivity level **26**

F.1 A-DEFEAT-GT 27

F.2 A-DEFEAT 29

G Convergence and algorithm comparisons **30**

H Limitations of network connectivity and compression **31**

I Additional experiments **32**

I.1 FashionMNIST 32

I.2 KMNIST 44

I.3 CIFAR-10 45

I.4 CIFAR-100 46

A Details of Motivating Example

The motivating example uses four nodes on a ring, with three points $z_i^\ell \in \mathbb{R}^2$ per node. Define the sample loss and local objective by

$$F_i(x;\ell) \coloneqq \tfrac{1}{2}\|x - z_i^\ell\|^2, \qquad f_i(x) \coloneqq \tfrac{1}{3}\sum_{\ell=1}^{3} F_i(x;\ell). \tag{12}$$

The global minimizer is

$$x^* \coloneqq \frac{1}{12}\sum_{i=1}^{4}\sum_{\ell=1}^{3} z_i^\ell = (1/30, 7/120). \tag{13}$$

The point sets are

$$\begin{aligned} z_1 &\coloneqq [(-3.5, 0.2), (-3.0, -0.3), (-4.0, 0.6)], \\ z_2 &\coloneqq [(0.3, 3.8), (-0.4, 4.2), (0.6, 3.2)], \\ z_3 &\coloneqq [(3.4, -0.2), (3.9, 0.5), (3.1, -0.6)], \\ z_4 &\coloneqq [(-0.5, -3.6), (0.4, -3.9), (0.1, -3.2)]. \end{aligned}$$

Every node starts at $x_0 = (1, 2)$ and cycles through its three points. This is a deterministic illustration, not an experiment satisfying the independent-sampling assumption. Figure 1 labels the trajectory of node 1 through x_{40}^1; the average model would instead be denoted by $\bar{x}_t$.

B Grid-search procedure

To select competitive settings for the proposed and baseline algorithms, we search hyperparameters such as the learning rate η and the consensus parameters used by the baselines. We perform a grid search for every combination of dataset, compression level, network topology, number of nodes, and Dirichlet parameter. The same learning-rate range is applied across all methods. It is $\eta \in \{0.00316, 0.00562, 0.01, 0.0178, 0.0316, 0.0562, 0.1\}$ for FMNIST and KMNIST, extended by 0.001 and 0.00178 for CIFAR-10 and CIFAR-100, where the error-feedback baselines require smaller steps over the longer horizon. A single common range is shared by the consensus-type parameters of the baselines. The step γ of BEER and MoTEF, the consensus step of CHOCO, and the averaging step of DeepSqueeze are selected from $\{0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$. For MoTEF, the momentum parameter is additionally selected from $\lambda \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$. A-DEFEAT and A-DEFEAT-GT use $\omega = 1$ in every experiment, so their consensus gain is not searched. Their discount coefficient is also not searched. Accordingly to Lemma C.1, we apply two general clipping ranges for discount coefficient. Algorithms 2 and 1 compute it using equation 8, with $\|g_t^i\|$ for A-DEFEAT and $\|u_t^i\|$ for A-DEFEAT-GT and with the stated zero-residual convention. The main theorems analyze this unit-gain implementation under their stated stability conditions.

To achieve reasonable performance for each algorithm, we conduct extensive simulations. For the FMNIST and KMNIST datasets, we run 1000 aggregation rounds with a single random seed for each parameter pair under different network topologies and levels of data heterogeneity. For the CIFAR-10 and CIFAR-100 datasets, we run 2000 aggregation rounds for each parameter pair in each experimental setting. For KMNIST, CIFAR-10 and CIFAR-100, the grid search uses fewer rounds than the final experiments. It uses 1000 rounds instead of 5000 for KMNIST, and 2000 instead of 10000 for CIFAR-10 and CIFAR-100. For these datasets we take the best five to eight settings from the search, run them again for the full number of rounds, and then keep the best one. In the final experiments, settings are compared by the average test accuracy over the last 10% of the rounds, so the choice does not depend on a single evaluation. The chosen setting is then run again with three new random seeds that are not used in the search, and the tables give the mean and standard deviation over these three runs. Because test accuracy was used for selection, the test comparison is subject to selection bias; independent final seeds do not remove that limitation. A validation split should be used for selection in a rerun.

We emphasize that, although the grid search is done before the formal experiments, we cannot guarantee that the chosen parameters are the best possible ones, especially for the algorithms with a

fixed error-feedback step. One reason is that the search uses only one random seed, and for KMNIST and the CIFAR datasets it uses fewer rounds than the final experiments, so a setting can look best in the search and then do worse when it is run for the full number of rounds. We reduce this risk by rerunning the best settings for the full horizon and replacing a divergent setting with another searched candidate when one is available. Some baseline entries remain unstable, as marked in Appendix I. The adaptive coefficient is not an additional searched hyperparameter in these experiments. The resulting tables describe the selected runs, while the use of test accuracy for selection remains a limitation of the comparison.

C CLIPPING BOUNDS

Lemma C.1 (General clipping bounds)**.** *For an interval $[a, b]$ containing 1, set $\delta := \max\{1-a, b-1\}$. Every $\gamma \in [a, b]$ satisfies $\|(1-\gamma)e\| \le \delta\|e\|$, and $b - a \le 2\delta$, with equality for $[a, b] = [1-\delta, 1+\delta]$. In particular, the adaptive rule in equation 8 satisfies, for each node and iteration,*

$$|1 - \gamma_t^i|\,\|e_t^i\| \le (1 - g_c)\|e_t^i\|, \qquad \gamma_t^i\|e_t^i\| \le \eta\|d_t^i\| + g_c\|e_t^i\|.$$

Proof. The first claims follow from the distances of the interval's endpoints from 1. For the second clipping inequality, an unclipped or upper-clipped ratio gives $\gamma_t^i\|e_t^i\| \le \eta\|d_t^i\|$. A lower-clipped ratio gives $\gamma_t^i\|e_t^i\| = g_c\|e_t^i\|$. Both imply the stated bound, and the zero-residual case is immediate. We stick on $\delta = \frac{1}{\sqrt{2}}$ in the main theorem and its proof. If $\delta \to 1$, the more general range is $(0, 2)$, which consists with $|1 - \gamma_t^i| < 1$. □

D PROOF OF A-DEFEAT-GT

The proof below establishes Theorem 4.1 by finite-horizon bounds with explicit constants. All node vectors are stored as matrix columns. We assume $n \ge 2$; this is the decentralized setting under consideration.

D.1 FINITE-HORIZON GUARANTEE FOR A-DEFEAT-GT

We use column-stacked matrices and the Frobenius norm throughout. Write

$$X_t := [x_t^1, \dots, x_t^n], \qquad E_t := [e_t^1, \dots, e_t^n], \qquad U_t := [u_t^1, \dots, u_t^n], \qquad \widehat{U}_t := [\widehat{u}_t^1, \dots, \widehat{u}_t^n].$$

Also write

$$J := \frac{\mathbf{1}\mathbf{1}^\top}{n}, \qquad \Pi := I - J, \qquad \mathcal{L} := I - W,$$
$$\rho := 1 - \|W - J\|_2 > 0, \qquad \mu := \|\mathcal{L}\|_2 = 1 - \lambda_{\min}(W) < 2.$$

Write

$$F_t := [\nabla f_i(x_t^i)]_{i=1}^n, \qquad G_t := [g_t^1, \dots, g_t^n].$$

Use the initialization $X_0 = x_0\mathbf{1}^\top$, $E_0 = 0$, $U_0 = G_0$, and $\widehat{U}_0 = 0$, where the deterministic point x_0 is arbitrary.

Define the following explicit constants.

$$g_c := 1 - 1/\sqrt{2}, \qquad s := \sqrt{\beta}, \qquad h := \frac{2}{2-\mu}, \tag{14}$$

$$d_{sp} := \frac{2\mu}{2-\mu}, \qquad \kappa := \max\left\{1, \frac{\mu}{2-\mu}\right\},$$

$$\vartheta := g_c - s(d_{sp} + g_c), \qquad C := \frac{s(\kappa+2)}{\vartheta}, \qquad v := h(1+2C). \tag{15}$$

We require $\vartheta > 0$. Choose

$$0 < \tau \le \min\{1, 1/\mu\}, \qquad m := 1 - s(1+2\tau\mu) > 0, \qquad k_{tr} := \frac{2}{2-\tau\mu}.$$

Thus $\tau = 1/2$ is admissible whenever the two displayed stability margins are positive. In particular, τ need not decrease with ρ. Let

$$A_{\mathrm{GT}} \coloneqq \frac{1}{\tau\rho} + \frac{2s}{m}, \tag{16}$$

$$B_{\mathrm{GT}} \coloneqq k_{\mathrm{tr}} + \frac{s(2+\tau\mu k_{\mathrm{tr}})}{m},$$

$$I_0 \coloneqq \frac{\|F_0\Pi\|_F}{\sqrt{\tau\rho}} + \frac{\sqrt{\mathbb{E}\|G_0\|_F^2} + s\sqrt{\frac{\tau\mu}{2-\tau\mu}}\|F_0\Pi\|_F}{m}, \tag{17}$$

$$H_{\mathrm{GT}} \coloneqq \frac{1}{2}\left(\frac{1}{\rho} + hC\right)^2 + C^2,$$

$$\eta_* \coloneqq \min\left\{\frac{1}{4L}, \frac{1}{2A_{\mathrm{GT}}Lv}, \frac{1}{\sqrt{192}\,L\sqrt{H_{\mathrm{GT}}}}\right\}. \tag{18}$$

For a nontrivial connected network $\mu > 0$; the one-node problem can instead be treated directly as centralized error-feedback SGD.

Theorem D.1 (Finite-horizon guarantee for A-DEFEAT-GT)**.** *When Assumptions 3.1, 3.3, and 3.4 and the positivity conditions $\vartheta > 0$ and $m > 0$ above hold and the compressor satisfies Definition 3.2, use Algorithm 1 with consensus gain $\omega = 1$, the stated value of τ, and clipping endpoints g_{c} and $2 - g_{\mathrm{c}}$, defining $\gamma_t^i = 1$ when $e_t^i = 0$. Let $T \geq 1$ be an integer and use the same stepsize $0 < \eta \leq \eta_*$ at every iteration. The selected value may depend on T and on the problem, network, and compression parameters. With $\bar{x}_t \coloneqq X_t\mathbf{1}/n$ and $\Delta \coloneqq f(x_0) - f^*$, the finite-horizon guarantee is*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 \leq \frac{2\Delta}{\eta T} + \frac{\eta L\sigma^2}{n} + 24L^2\eta^2 H_{\mathrm{GT}}\left((B_{\mathrm{GT}}^2+1)\sigma^2 + \frac{I_0^2}{nT}\right). \tag{19}$$

Proof. Let $\mathcal{F}_t$ contain the algorithmic history before the iteration-t samples are drawn, and write

$$\Xi_t \coloneqq G_t - F_t = [\xi_t^1, \ldots, \xi_t^n], \qquad \xi_t^i \coloneqq g_t^i - \nabla f_i(x_t^i).$$

The sampling rule below equation 1 and Assumption 3.4 give

$$\mathbb{E}[\Xi_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\Xi_t\|_F^2 \mid \mathcal{F}_t] \leq n\sigma^2.$$

The draws at different nodes are conditionally independent, so $\mathbb{E}\|\Xi_t\mathbf{1}/n\|^2 \leq \sigma^2/n$. Index all model and tracker compression calls chronologically by k. Let Z_k be the input to call k and let $\mathcal{G}_k$ be the sigma-algebra generated by Z_k and all random variables revealed before that call. The compressor seed is independent of $\mathcal{G}_k$. Since Z_k is $\mathcal{G}_k$-measurable, Definition 3.2 gives

$$\mathbb{E}[\|\mathcal{C}_k(Z_k) - Z_k\|_F^2 \mid \mathcal{G}_k] \leq \beta\|Z_k\|_F^2, \qquad 0 \leq \beta < 1.$$

No unbiasedness of compression is required. The tracker uses the stored gradient difference $G_{t+1} - G_t$.

All finite-horizon expectations below are well defined. Conditional Minkowski's inequality and the compression bound above give

$$\left(\mathbb{E}[\|\mathcal{C}_k(Z_k)\|_F^2 \mid \mathcal{G}_k]\right)^{1/2} \leq (1+\sqrt{\beta})\|Z_k\|_F.$$

Smoothness gives $\|F_t\|_F \leq \|F_0\|_F + L\|X_t - X_0\|_F$, and the conditional noise bound gives $\mathbb{E}\|\Xi_t\|_F^2 \leq n\sigma^2$. Starting from the deterministic initialization, induction through the algorithm shows that every model, tracker, residual, cache, compressed message, and stochastic gradient has a finite second moment at each finite iteration. For the residual-adjusted average used below,

$$f(y_t) \leq f(x_0) + \langle\nabla f(x_0), y_t - x_0\rangle + \frac{L}{2}\|y_t - x_0\|^2.$$

Together with $f(y_t) \geq f^*$, this shows that $\mathbb{E}f(y_t)$ is finite and justifies the finite-horizon telescoping argument.

Finite-horizon norms and elementary filter bounds. For a sequence of random matrices, define $\|Z\|_{0\text{-}T} \coloneqq (\sum_{t=0}^{T} \mathbb{E}\|Z_t\|_F^2)^{1/2}$. These are Hilbert-space norms; in particular, Minkowski's inequality applies without independence assumptions between terms. For a stable scalar recursion with eigenvalue $\lambda \in (-1, 1)$, Parseval's identity and zero extension of its finite input sequence give the finite-horizon ℓ_2 operator bounds

$$\left\| \frac{1}{1-\lambda z} \right\|_\infty = \frac{1}{1-|\lambda|}, \qquad \left\| \frac{1-z}{1-\lambda z} \right\|_\infty = \frac{2}{1+\lambda}, \qquad |z| = 1. \tag{20}$$

Here z is the scalar transform variable and is unrelated to the tracker message z_t^i in Algorithm 1. The second identity follows by maximizing $2(1-\cos\varphi)/(1+\lambda^2-2\lambda\cos\varphi)$ over $\cos\varphi \in [-1, 1]$. Diagonalizing the symmetric matrix W applies these scalar bounds to matrix recursions. Multiplication by a time delay has unit gain. Restricting the resulting output to a finite horizon can only decrease its norm. These deterministic operator bounds continue to hold after taking expectations, including for adapted inputs. The residual-to-increment map has a one-step advance, so its bound uses the terminal residual E_T included in the norm below.

Step 1. Model residuals and increments. The adaptive coefficient in line 6 of Algorithm 1 compares the magnitude of the tracker step with that of the stored model residual. Equating $\gamma_t^i\|e_t^i\|$ and $\eta\|u_t^i\|$ gives the ratio $\eta\|u_t^i\|/\|e_t^i\|$. The algorithm clips that ratio to $[g_c, 2-g_c]$, where $g_c = 1-\delta$ and $\delta = 1/\sqrt{2}$. This symmetric interval ensures $|1-\gamma_t^i| \le \delta$ and hence contracts the carried fraction of the residual. The choice $\delta^2 = 1/2$ is a convenient fixed margin, not an optimality claim. When $e_t^i = 0$, the convention $\gamma_t^i = 1$ leaves every residual product zero.

The coefficient appears twice in the model update. In matrix notation, $B_t = -X_t\mathcal{L} - \eta U_t + \Gamma_t \odot E_t$ and $E_{t+1} = B_t - V_t + (1-\Gamma_t)\odot E_t$, where $V_t = X_{t+1} - X_t$. Thus $V_t = B_t + (1-\Gamma_t)\odot E_t - E_{t+1}$, and the two parts of the old residual add to E_t. The exact model identity is

$$X_{t+1} = X_t W - \eta U_t + E_t - E_{t+1}. \tag{21}$$

For this proof set

$$u \coloneqq \|U\|_{0\text{-}T-1}, \qquad e \coloneqq \|E\|_{0\text{-}T}, \qquad p \coloneqq \|X\Pi\|_{0\text{-}T-1},$$
$$\ell \coloneqq \|X\mathcal{L}\|_{0\text{-}T-1}, \qquad w \coloneqq \|(X_{t+1}-X_t)_{t=0}^{T-1}\|_{0\text{-}T-1}.$$

Applying equation 20 to equation 21, using $X_0\Pi = E_0 = 0$, yields

$$p \le \frac{\eta}{\rho}u + he, \qquad \ell \le \eta\kappa u + d_{\mathrm{sp}}e, \qquad w \le \eta h u + 2he. \tag{22}$$

For completeness, the input-to-output transfer factors for $X\Pi$ are $-\eta z/(1-\lambda z)$ from $U\Pi$ and $-(1-z)/(1-\lambda z)$ from $E\Pi$. Multiplying these by $1-\lambda$ gives the factors for $X\mathcal{L}$. The increment factors have gains bounded by ηh and $2h$, respectively. For the consensus mode, the apparent pole cancels and these increment bounds remain valid.

The clipping rule gives, pointwise,

$$\|\Gamma_t \odot E_t\|_F \le \eta\|U_t\|_F + g_c\|E_t\|_F, \qquad \|(1-\Gamma_t)\odot E_t\|_F \le (1-g_c)\|E_t\|_F.$$

For the first inequality, let $r_t^i \coloneqq \eta\|u_t^i\|/\|e_t^i\|$ when $e_t^i \neq 0$. If $r_t^i \ge g_c$, the clipped value obeys $\gamma_t^i \le r_t^i$, so $\gamma_t^i\|e_t^i\| \le \eta\|u_t^i\|$. If $r_t^i < g_c$, the lower clip gives $\gamma_t^i\|e_t^i\| = g_c\|e_t^i\|$. Consequently, $(\gamma_t^i)^2\|e_t^i\|^2 \le \eta^2\|u_t^i\|^2 + g_c^2\|e_t^i\|^2$. The second inequality follows from $|1-\gamma_t^i| \le \delta = 1-g_c$. These two bounds are where the adaptive coefficient enters the energy argument. The first controls its contribution to the compressed input B_t, while the second controls residual carryover in E_{t+1}. Since $B_t = -X_t\mathcal{L} - \eta U_t + \Gamma_t \odot E_t$, conditional contractivity and Minkowski's inequality give

$$e \le s(\ell + 2\eta u + g_c e) + (1-g_c)e \le s(\kappa+2)\eta u + [1 - g_c + s(d_{\mathrm{sp}} + g_c)]e.$$

The coefficient of e is below one precisely when $g_c - s(d_{\mathrm{sp}} + g_c) = \vartheta > 0$. Moving that term to the left yields $e \le [s(\kappa+2)/\vartheta]\eta u = C\eta u$. Thus clipping supplies the positive margin that closes the model residual estimate; the exact difference $E_t - E_{t+1}$ supplies the favorable disagreement filter in equation 21. Combining this residual bound with the model gains above gives

$$e \le C\eta u, \qquad p \le \eta(\rho^{-1} + hC)u, \qquad w \le v\eta u. \tag{23}$$

Both endpoints E_0 and E_T are included in e, so the summation of the residual recursion has no omitted terminal term.

Step 2. Tracker noise must retain its difference structure. Let $R_t^{\mathrm{tr}} \coloneqq U_t - \widehat{U}_t$ and $W_\tau \coloneqq I - \tau\mathcal{L}$. The exact tracker and compression identities are

$$U_{t+1} = U_t W_\tau + \tau R_t^{\mathrm{tr}}\mathcal{L} + (F_{t+1} - F_t) + (\Xi_{t+1} - \Xi_t), \tag{24}$$

$$R_{t+1}^{\mathrm{tr}} = (I - \mathcal{C}_t^u)\big(R_t^{\mathrm{tr}} - \tau U_t\mathcal{L} + \tau R_t^{\mathrm{tr}}\mathcal{L} + (F_{t+1} - F_t) + (\Xi_{t+1} - \Xi_t)\big). \tag{25}$$

Here $(I - \mathcal{C})(Z)$ denotes $Z - \mathcal{C}(Z)$, not a linear operator. Multiplying the tracker recursion by J gives $U_{t+1}J = U_tJ + (G_{t+1} - G_t)J$, since $W_\tau J = J$ and $\mathcal{L}J = 0$. Because $U_0 = G_0$, induction yields $U_tJ = G_tJ$ for every t.

Define $Q_t \coloneqq U_t\Pi$, and let

$$q \coloneqq \|Q\|_{0\text{-}T-1}, \qquad r \coloneqq \|R^{\mathrm{tr}}\|_{0\text{-}T-1}, \qquad \xi \coloneqq \|\Xi\|_{0\text{-}T-1}, \qquad a_F \coloneqq \|(F_{t+1} - F_t)_{t=0}^{T-2}\|_{0\text{-}T-2}.$$

An empty sequence has norm zero when $T = 1$. The eigenvalues of W_τ are nonnegative, and its disagreement eigenvalues are at most $1 - \tau\rho$. Unrolling equation 24 gives

$$Q_t = F_0\Pi W_\tau^t + \sum_{j=0}^{t-1}(F_{j+1} - F_j)\Pi W_\tau^{t-1-j} + \Xi_t\Pi - \tau\sum_{j=0}^{t-1}\Xi_j\Pi\mathcal{L}W_\tau^{t-1-j} + \tau\sum_{j=0}^{t-1}R_j^{\mathrm{tr}}\mathcal{L}W_\tau^{t-1-j}. \tag{26}$$

The four terms in this decomposition have gains bounded, respectively, by the initial-condition norm, $(\tau\rho)^{-1}$, $k_{\mathrm{tr}} = 2/(2-\tau\mu)$, and 1. In particular, the noise term has transfer function $(1-z)/(1-(1-\tau\lambda_{\mathcal{L}})z)$ in a Laplacian mode with eigenvalue $\lambda_{\mathcal{L}} > 0$. It therefore has a bounded gain independent of the positive spectral gap. Treating $\Xi_{t+1} - \Xi_t$ as arbitrary undifferenced noise would lose this cancellation. Thus

$$q \le \frac{\|F_0\Pi\|_F}{\sqrt{\tau\rho}} + \frac{a_F}{\tau\rho} + k_{\mathrm{tr}}\xi + r. \tag{27}$$

Multiplying equation 26 by $\mathcal{L}$ gives

$$\|U\mathcal{L}\|_{0\text{-}T-1} \le I_L + \frac{a_F}{\tau} + \mu k_{\mathrm{tr}}\xi + \mu r, \qquad I_L \coloneqq \sqrt{\frac{\mu}{\tau(2-\tau\mu)}}\,\|F_0\Pi\|_F. \tag{28}$$

Indeed, the initial-condition gain squared in a Laplacian mode is

$$\sum_{t=0}^{\infty}\lambda_{\mathcal{L}}^2(1-\tau\lambda_{\mathcal{L}})^{2t} = \frac{\lambda_{\mathcal{L}}}{\tau(2-\tau\lambda_{\mathcal{L}})} \le \frac{\mu}{\tau(2-\tau\mu)}.$$

The gains from $F_{t+1} - F_t$ and R_t^{tr} to $U_t\mathcal{L}$ are $1/\tau$ and μ, respectively.

Because $R_0^{\mathrm{tr}} = G_0$, the compression recursion equation 25 and $\|\Xi_{t+1} - \Xi_t\|_{0\text{-}T-2} \le 2\xi$ give

$$\begin{aligned} r &\le \sqrt{\mathbb{E}\|G_0\|_F^2} + s\big((1+\tau\mu)r + \tau\|U\mathcal{L}\|_{0\text{-}T-1} + a_F + 2\xi\big) \\ &\le \sqrt{\mathbb{E}\|G_0\|_F^2} + s\big((1+2\tau\mu)r + \tau I_L + 2a_F + (2+\tau\mu k_{\mathrm{tr}})\xi\big). \end{aligned} \tag{29}$$

Using the margin $m > 0$ to absorb r, then substituting into equation 27, proves

$$q \le I_0 + A_{\mathrm{GT}}a_F + B_{\mathrm{GT}}\xi. \tag{30}$$

Smoothness and equation 23 imply $a_F \le Lw \le Lv\eta u$.

Write

$$M \coloneqq \sum_{t=0}^{T-1}\mathbb{E}\,\|F_t\mathbf{1}/n\|^2.$$

Conditional unbiasedness and node independence yield

$$\|UJ\|_{0\text{-}T-1}^2 = \|GJ\|_{0\text{-}T-1}^2 \le nM + T\sigma^2, \qquad \xi^2 \le nT\sigma^2.$$

Since $u \le q + \|UJ\|_{0\text{-}T-1}$, the bound in equation 30 and $\eta \le (2A_{\mathrm{GT}}Lv)^{-1}$ imply

$$u \le 2\left(I_0 + B_{\mathrm{GT}}\xi + \sqrt{nM + T\sigma^2}\right).$$

Squaring this inequality and using $n \ge 1$ proves

$$u^2 \le 12I_0^2 + 12(B_{\mathrm{GT}}^2 + 1)nT\sigma^2 + 12nM. \tag{31}$$

Step 3. Objective descent. Set $\bar{e}_t := E_t\mathbf{1}/n$ and $y_t := \bar{x}_t + \bar{e}_t$. Averaging the model identity equation 21 from Step 1 cancels the consensus correction because $W\mathbf{1} = \mathbf{1}$. The identity $U_tJ = G_tJ$ established in Step 2 then gives the exact update

$$y_{t+1} = y_t - \eta G_t\mathbf{1}/n.$$

The variables y_t, X_t, E_t are measurable before drawing G_t. By Assumption 3.3 and $f = n^{-1}\sum_i f_i$, the global gradient is also L-Lipschitz, since

$$\|\nabla f(y) - \nabla f(x)\| \le \frac{1}{n}\sum_i \|\nabla f_i(y) - \nabla f_i(x)\| \le L\|y - x\|.$$

Taking $y = y_t$ and $x = \bar{x}_t$, and using $y_t - \bar{x}_t = \bar{e}_t$, gives $\|\nabla f(y_t) - \nabla f(\bar{x}_t)\| \le L\|\bar{e}_t\|$. Write $\bar{F}_t := F_t\mathbf{1}/n$. Conditional smoothness applied to $y_{t+1} = y_t - \eta G_t\mathbf{1}/n$, together with conditional unbiasedness and $\mathbb{E}[\|\Xi_t\mathbf{1}/n\|^2 \mid \mathcal{F}_t] \le \sigma^2/n$, gives

$$\mathbb{E}[f(y_{t+1}) \mid \mathcal{F}_t] \le f(y_t) - \eta\langle\nabla f(y_t), \bar{F}_t\rangle + \frac{L\eta^2}{2}\left(\|\bar{F}_t\|^2 + \frac{\sigma^2}{n}\right).$$

We split the inner product at $\bar{x}_t$. The polarization identity gives

$$-\langle\nabla f(\bar{x}_t), \bar{F}_t\rangle = -\tfrac{1}{2}\|\nabla f(\bar{x}_t)\|^2 - \tfrac{1}{2}\|\bar{F}_t\|^2 + \tfrac{1}{2}\|\nabla f(\bar{x}_t) - \bar{F}_t\|^2,$$

and Jensen's inequality and local smoothness bound the last term by

$$\|\nabla f(\bar{x}_t) - \bar{F}_t\|^2 \le \frac{L^2}{n}\sum_i \|x_t^i - \bar{x}_t\|^2 = \frac{L^2}{n}\|X_t\Pi\|_F^2.$$

For the remaining part, Young's inequality and $\|\nabla f(y_t) - \nabla f(\bar{x}_t)\| \le L\|\bar{e}_t\|$ give

$$-\langle\nabla f(y_t) - \nabla f(\bar{x}_t), \bar{F}_t\rangle \le L^2\|\bar{e}_t\|^2 + \tfrac{1}{4}\|\bar{F}_t\|^2.$$

Since $\|\bar{e}_t\|^2 \le \|E_t\|_F^2/n$, combining these three bounds shows that the coefficient of $\|\bar{F}_t\|^2$ is $-\eta/2 + \eta/4 + L\eta^2/2 = -\eta/4 + L\eta^2/2$. Therefore

$$\begin{aligned}\mathbb{E}f(y_{t+1}) \le \mathbb{E}f(y_t) - \frac{\eta}{2}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 - \left(\frac{\eta}{4} - \frac{L\eta^2}{2}\right)\mathbb{E}\|F_t\mathbf{1}/n\|^2 \\ + \frac{\eta L^2}{2n}\mathbb{E}\|X_t\Pi\|_F^2 + \frac{\eta L^2}{n}\mathbb{E}\|E_t\|_F^2 + \frac{L\eta^2\sigma^2}{2n}.\end{aligned} \tag{32}$$

We retain the negative $\|\bar{F}_t\|^2$ term because Step 2's tracker-energy bound contains $M = \sum_{t=0}^{T-1}\mathbb{E}\|\bar{F}_t\|^2$; this term allows us to absorb that contribution after summing equation 32. The inequality uses the exact identities from Steps 1 and 2; their quantitative energy bounds enter when we sum it. Indeed, the disagreement and residual terms sum to $\eta L^2(\frac{1}{2}p^2 + e^2)/n$. By equation 23 from Step 1,

$$\tfrac{1}{2}p^2 + e^2 \le \eta^2\left[\tfrac{1}{2}(\rho^{-1} + hC)^2 + C^2\right]u^2 = \eta^2 H_{\mathrm{GT}}u^2.$$

Step 2 supplies the bound on u^2 in equation 31; it is substituted after the summed descent inequality below. For $\eta \le 1/(4L)$, summing equation 32 and using $y_0 = x_0$ yields

$$\begin{aligned}\frac{1}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 + \frac{1}{8}M &\le \frac{\Delta}{\eta} + \frac{\eta L\sigma^2 T}{2n} + \frac{L^2}{n}\left(\tfrac{1}{2}p^2 + e^2\right) \\ &\le \frac{\Delta}{\eta} + \frac{\eta L\sigma^2 T}{2n} + \frac{L^2\eta^2 H_{\mathrm{GT}}}{n}u^2.\end{aligned} \tag{33}$$

After substituting equation 31, the coefficient of M on the right is at most $1/16$, because $\eta \le (\sqrt{192}L\sqrt{H_{\mathrm{GT}}})^{-1}$. Absorb this term and discard the remaining nonnegative multiple of M. Multiplying by $2/T$ proves equation 19. □

Explicit step-size optimization and interpretation. Choose any $D \geq \Delta$ with $D > 0$ (in particular, set $D = \Delta$ when $\Delta > 0$), and put

$$a_{\mathrm{GT}} \coloneqq \frac{L\sigma^2}{n}, \qquad b_{\mathrm{GT}} \coloneqq 24L^2 H_{\mathrm{GT}}(B_{\mathrm{GT}}^2 + 1)\sigma^2, \qquad c_{\mathrm{GT}} \coloneqq \frac{24L^2 H_{\mathrm{GT}} I_0^2}{n},$$

$$\eta \coloneqq \min\left\{\eta_*, \sqrt{\frac{2D}{a_{\mathrm{GT}}T}}, \left(\frac{2D}{b_{\mathrm{GT}}T}\right)^{1/3}\right\}. \tag{34}$$

where a term with zero denominator is omitted. The finite-horizon guarantee implies

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 \leq \frac{2D}{\eta_* T} + 2\sqrt{\frac{2Da_{\mathrm{GT}}}{T}} + \frac{2(2D)^{2/3}b_{\mathrm{GT}}^{1/3}}{T^{2/3}} + \frac{c_{\mathrm{GT}}\eta_*^2}{T}. \tag{35}$$

If $\sigma = 0$, simply choose $\eta = \eta_*$ in the finite-horizon bound. This convention also defines a positive step size when $\Delta = 0$; the finite-horizon theorem itself uses the actual value of Δ.

Consider a family of networks with $\lambda_{\min}(W) \geq -1 + \nu_0$ for a fixed $\nu_0 > 0$, fixed $\tau > 0$, and fixed positive lower bounds on ϑ and m. All constants C, v, B_{GT} are then bounded independently of ρ, while

$$H_{\mathrm{GT}} = O(\rho^{-2}), \qquad A_{\mathrm{GT}} = O(\rho^{-1}), \quad \eta_*^{-1} = O(L/\rho),$$
$$I_0^2 = O\left(\rho^{-1}\|F_0\Pi\|_F^2 + \mathbb{E}\|G_0\|_F^2\right).$$

Accordingly, for $\Delta > 0$ and $D = \Delta$, equation 35 gives

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 = O\left(\sqrt{\frac{L\Delta\sigma^2}{nT}} + \frac{(L\Delta\sigma)^{2/3}}{\rho^{2/3}T^{2/3}} + \frac{L\Delta + n^{-1}\|F_0\Pi\|_F^2 + n^{-1}\mathbb{E}\|G_0\|_F^2}{\rho T}\right). \tag{36}$$

At the deterministic common initialization, unbiasedness gives

$$\mathbb{E}\|G_0\|_F^2 = \|F_0\|_F^2 + \mathbb{E}\|G_0 - F_0\|_F^2 \geq \|F_0\|_F^2 \geq \|F_0\Pi\|_F^2.$$

Consequently, with $\mathcal{J}_0 \coloneqq \mathbb{E}\|G_0\|_F^2/n$, the single term $\mathcal{J}_0/(\rho T)$ bounds both initialization contributions in equation 36 up to a numerical factor. This gives the more compact form in equation 10. The constants hidden here depend on the stated compression and negative spectrum margins, but not on the positive spectral gap. This is a rigorous spectral-gap rate for the stated regime. It does not establish the same guarantee for arbitrary $\beta < 1$, including aggressive Top-k compression that violates these margins, nor does it imply uniform superiority over other methods under different assumptions.

D.2 Sparse graph families and the role of the consensus gain

Why the consensus gain is needed. This gain is a sufficient stabilization device rather than the source of the improved network exponent. Write $s \coloneqq \sqrt{\beta}$ and let $\widetilde{\mu} \coloneqq \|I - W_\omega\|_2 = \omega\mu$. Compression leaves a factor s of its input error, while the disagreement operator feeds part of that error into the next compressed input. For the effective matrix, the two small-gain margins in equation 9 take the form $\vartheta = g_{\mathrm{c}}(1-s) - s\frac{2\widetilde{\mu}}{2-\widetilde{\mu}}$ and $m = (1-s) - 2\tau s\widetilde{\mu}$. When s is close to one, the remaining compression contraction $1 - s$ can be too small to dominate this network feedback at unit gain. The choice in equation 67 reduces $\widetilde{\mu}$ enough to make both margins positive. If they are already positive for the prescribed W, one may use $\omega = 1$; the displayed choice is sufficient, not necessary. This use of a scalar mixing gain is analogous to the consensus stepsizes in CHOCO, BEER, and MoTEF. The adaptive residual coefficient γ_t^i has a different role. It controls how much accumulated residual is reinjected at iteration t, whereas ω controls how strongly the network propagates the resulting perturbation. The improved powers of ρ in Theorems 4.1 and 4.2 instead come from the difference cancellations in the model recursion and, for A-DEFEAT-GT, the tracker recursion. The proofs use only the clipping interval and the bounds in Lemma C.1. They therefore certify the proposed adaptive rule but do not establish a convergence advantage over every fixed residual coefficient that satisfies the same inequalities.

E A-DEFEAT

E.1 A-DEFEAT ALGORITHM

Algorithm 2 A-DEFEAT

Input. Fixed symmetric stochastic matrix W, fixed initial point x_0, horizon T, stepsize $\eta > 0$ used at every iteration, and consensus gain $0 < \omega \leq 1$.
Initialize $x_0^i = x_0$, $e_0^i = 0$, and $x_0^{i,j} = x_0$ for every actual neighbor, including self-caches when $W_{ii} > 0$.
for $t = 0, \ldots, T-1$ **do**
  for each node i, in parallel **do**
    Draw a fresh sample and compute $g_t^i = \nabla F_i(x_t^i; \zeta_t^i)$.
    Set $\gamma_t^i = \mathrm{clip}_{[g_c, 2-g_c]}(\eta\|g_t^i\|/\|e_t^i\|)$, $g_c \coloneqq 1 - 1/\sqrt{2}$; use $\gamma_t^i = 1$ if $e_t^i = 0$.
    $b_t^i = \omega\big(\sum_{j|W_{ij}>0} W_{ij} x_t^{i,j} - x_t^i\big) - \eta g_t^i + \gamma_t^i e_t^i$; $v_t^i = \mathcal{C}_{t,i}^x(b_t^i)$.
    $x_{t+1}^i = x_t^i + v_t^i$; $e_{t+1}^i = b_t^i - v_t^i + (1-\gamma_t^i)e_t^i$.
    Send v_t^i to neighbors and receive v_t^j; set $x_{t+1}^{i,j} = x_t^{i,j} + v_t^j$ for $W_{ij} > 0$.
  end for
end for

The input matrix W is fixed. The scalar ω is selected once before training and multiplies the model-consensus correction at every iteration. For analysis only, this correction is equivalent to mixing with $W_\omega \coloneqq (1-\omega)I + \omega W$; the implementation need not form or store W_ω. Each randomized compression call uses a new seed independent of its input and previously revealed randomness.

E.2 FINITE-HORIZON GUARANTEE FOR A-DEFEAT

Throughout this section, matrices store node vectors as columns. Write

$$X_t \coloneqq [x_t^1, \ldots, x_t^n], \qquad E_t \coloneqq [e_t^1, \ldots, e_t^n].$$

Let

$$J \coloneqq \frac{\mathbf{1}\mathbf{1}^\top}{n}, \qquad \Pi \coloneqq I - J, \qquad P_t \coloneqq X_t\Pi, \qquad \bar{x}_t \coloneqq \frac{X_t\mathbf{1}}{n}.$$

The symmetric doubly stochastic matrix W satisfies $WJ = JW = J$ and $\|W - J\|_2 < 1$. Define

$$\rho \coloneqq 1 - \|W - J\|_2 > 0, \qquad \lambda_- \coloneqq \min_{2\leq j\leq n} \lambda_j(W) > -1, \qquad \mu \coloneqq 1 - \lambda_-, \tag{37}$$

$$h \coloneqq \frac{2}{1+\lambda_-}, \qquad d_{\mathrm{sp}} \coloneqq \frac{2(1-\lambda_-)}{1+\lambda_-}, \qquad \kappa \coloneqq \max\left\{1, \frac{1-\lambda_-}{1+\lambda_-}\right\}. \tag{38}$$

Here the nontrivial eigenvalues need not be ordered by absolute value. In particular, $\mu = \|I - W\|_2$; its ratio to ρ is not presumed bounded. The conclusions below hold with the displayed values of $h, d_{\mathrm{sp}}, \kappa$ for every such W. Statements that hide these constants uniformly over networks additionally require $\lambda_- \geq -1 + \nu_0$ for some fixed $\nu_0 > 0$.

Finite-horizon theorem. Set

$$g_c \coloneqq 1 - 1/\sqrt{2}, \qquad s_\rho \coloneqq \frac{1}{\rho(2-\rho)}.$$

Define the model-compression margin and residual amplification by

$$\vartheta \coloneqq g_c - \sqrt{\beta}(d_{\mathrm{sp}} + g_c) > 0, \qquad C \coloneqq \frac{\sqrt{\beta}(\kappa+2)}{\vartheta}. \tag{39}$$

The remaining constants are

$$\begin{aligned} A_{\mathrm{A}} &\coloneqq 6h^2C^2 + \frac{6}{\rho^2}, \quad & B_\sigma &\coloneqq 3h^2C^2 + 3s_\rho, \\ B_\epsilon &\coloneqq A_{\mathrm{A}}, & B_Q &\coloneqq 6h^2C^2. \end{aligned} \tag{40}$$

$$D_\sigma \coloneqq (12h^2+2)C^2 + 12s_\rho,$$
$$D_\epsilon \coloneqq (24h^2+4)C^2 + \frac{24}{\rho^2}, \qquad D_Q \coloneqq (24h^2+4)C^2. \tag{41}$$

$$\eta_{\max} \coloneqq \min\left\{\frac{1}{2L}, \frac{1}{L\sqrt{2A_{\mathrm{A}}}}, \frac{1}{L\sqrt{2D_Q}}\right\}. \tag{42}$$

An entry containing $1/\sqrt{D_Q}$ is interpreted as $+\infty$ if $D_Q = 0$.

Theorem E.1 (Finite-horizon guarantee for A-DEFEAT)**.** *When Assumptions 3.1, 3.3, 3.4, and 3.5 and the positivity condition in equation 39 hold and the compressor satisfies Definition 3.2, use Algorithm 2 with consensus gain $\omega = 1$ and the clipping interval $[g_{\mathrm{c}}, 2-g_{\mathrm{c}}]$, setting $\gamma_t^i = 1$ when $e_t^i = 0$. Let $T \geq 1$ be an integer and use the same stepsize $0 < \eta \leq \eta_{\max}$ at every iteration. The selected value may depend on T and on the problem, network, and compression parameters. With $\bar{x}_t \coloneqq X_t\mathbf{1}/n$ and $\Delta \coloneqq f(x_0) - f^*$, the finite-horizon guarantee is*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 \leq \frac{4\Delta}{\eta T} + \frac{2\eta L\sigma^2}{n} + 2\eta^2L^2(D_\sigma\sigma^2 + D_\epsilon\epsilon^2). \tag{43}$$

The stochastic noise and systematic gradient divergence remain separate in equation 43.

Proof. We first record the stochastic and compression relations used below.

Probability assumptions and initialization. Each f_i has an L-Lipschitz gradient, $f \coloneqq n^{-1}\sum_i f_i$ is bounded below by f^*, and

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \epsilon^2 \quad \text{for every } x. \tag{44}$$

Let $\mathcal{F}_t$ contain the complete history before drawing the iteration-t gradients; in particular, X_t, E_t are $\mathcal{F}_t$-measurable. Write

$$F_t \coloneqq [\nabla f_1(x_t^1), \dots, \nabla f_n(x_t^n)], \quad G_t \coloneqq [g_t^1, \dots, g_t^n] = F_t + \Xi_t,$$
$$\Xi_t \coloneqq [\xi_t^1, \dots, \xi_t^n].$$

where $\xi_t^i \coloneqq g_t^i - \nabla f_i(x_t^i)$. The sampling rule below equation 1 and Assumption 3.4 imply, for each column ξ_t^i,

$$\mathbb{E}[\Xi_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi_t^i\|^2 \mid \mathcal{F}_t] \leq \sigma^2, \qquad \mathbb{E}[\langle \xi_t^i, \xi_t^j\rangle \mid \mathcal{F}_t] = 0 \quad (i \neq j). \tag{45}$$

Conditional independence across nodes is sufficient for the last condition. The histories are nested and include all gradients and compression outputs from earlier iterations. Initialize $X_0 = x_0\mathbf{1}^\top$ and $E_0 = 0$.

The coefficient γ_t^i determines how much of the stored residual e_t^i is put into the current compressed input. To choose it, compare the desired reinjection magnitude $\gamma_t^i\|e_t^i\|$ with the stochastic-gradient step magnitude $\eta\|g_t^i\|$. Equating these magnitudes gives the ratio below, which is clipped to keep both the reinjection and the residual carryover uniformly bounded. Set $\delta \coloneqq 1/\sqrt{2}$ and $g_{\mathrm{c}} \coloneqq 1-\delta$. The symmetric interval $[1-\delta, 1+\delta] = [g_{\mathrm{c}}, 2-g_{\mathrm{c}}]$ gives $|1-\gamma_t^i| \leq \delta$, and the particular choice $\delta^2 = 1/2$ provides a fixed contraction margin. This numerical choice is sufficient for the proof, not claimed to be optimal. The algorithm uses

$$\gamma_t^i \coloneqq \mathrm{clip}_{[g_{\mathrm{c}}, 2-g_{\mathrm{c}}]}\left(\frac{\eta\|g_t^i\|}{\|e_t^i\|}\right). \tag{46}$$

When $e_t^i = 0$, the algorithm sets $\gamma_t^i = 1$; all products involving e_t^i then vanish. Let $\Gamma_t \odot E_t$ denote the matrix with columns $\gamma_t^i e_t^i$. In matrix form the algorithm is

$$B_t = X_t(W-I) - \eta G_t + \Gamma_t \odot E_t, \qquad V_t = \mathcal{C}_t(B_t), \tag{47}$$

$$R_t^{\mathrm{mod}} = B_t - V_t, \qquad E_{t+1} = R_t^{\mathrm{mod}} + (\mathbf{1} - \Gamma_t)\odot E_t, \qquad X_{t+1} = X_t + V_t. \tag{48}$$

Let $\mathcal{G}_t^x$ be the sigma-algebra generated by B_t and all random variables revealed before the model-compression call at iteration t. The compressor seed is independent of $\mathcal{G}_t^x$. Since B_t is $\mathcal{G}_t^x$-measurable, Definition 3.2 gives

$$\mathbb{E}[\|R_t^{\mathrm{mod}}\|_F^2 \mid \mathcal{G}_t^x] \le \beta \|B_t\|_F^2, \qquad 0 \le \beta < 1. \tag{49}$$

No unbiasedness of the compressor or independence of its columns is required. All arguments below concern expectations, and no pathwise contraction inequality is used.

The expectations in the finite-horizon argument are well defined. Conditional Minkowski's inequality and the preceding bound give

$$\left(\mathbb{E}[\|\mathcal{C}_t(B_t)\|_F^2 \mid \mathcal{G}_t^x]\right)^{1/2} \le (1+\sqrt{\beta})\|B_t\|_F.$$

Moreover, smoothness gives $\|F_t\|_F \le \|F_0\|_F + L\|X_t - X_0\|_F$, while equation 45 gives $\mathbb{E}\|\Xi_t\|_F^2 \le n\sigma^2$. Starting from the deterministic initialization, induction through the algorithm shows that every model, residual, cache, compressed message, and stochastic gradient has a finite second moment at each finite iteration. For the residual-adjusted average used below, smoothness gives

$$f(y_t) \le f(x_0) + \langle \nabla f(x_0), y_t - x_0\rangle + \frac{L}{2}\|y_t - x_0\|^2.$$

Together with $f(y_t) \ge f^*$, this shows that $\mathbb{E}f(y_t)$ is finite and justifies the finite-horizon telescoping argument.

Exact transfer bounds on a finite horizon. For a random matrix sequence A, write

$$\|A\|_{T,2} \coloneqq \left(\sum_{t=0}^{T-1} \mathbb{E}\|A_t\|_F^2\right)^{1/2}.$$

The same γ_t^i appears in B_t and in the carried residual. Indeed, equation 48 gives $V_t = B_t + (\mathbf{1} - \Gamma_t) \odot E_t - E_{t+1}$. Substituting equation 47 makes $\Gamma_t \odot E_t + (\mathbf{1} - \Gamma_t) \odot E_t = E_t$. Thus γ_t^i controls the residual energy without changing the exact model recursion. Eliminating $B_t, V_t, R_t^{\mathrm{mod}}$ gives

$$X_{t+1} = X_t W - \eta G_t + E_t - E_{t+1}, \tag{50}$$

$$P_t = -E_t\Pi + \sum_{k=1}^{t-1} E_k(I-W)W^{t-1-k}\Pi - \eta\sum_{k=0}^{t-1} G_k W^{t-1-k}\Pi. \tag{51}$$

For each nontrivial eigenvalue $\lambda \in (-1, 1)$, the transfer function from its residual coordinate to its disagreement coordinate is

$$K_\lambda(z) \coloneqq -\frac{1-z^{-1}}{1-\lambda z^{-1}}.$$

Here z is the scalar transform variable and is unrelated to the tracker-message notation used by A-DEFEAT-GT. The transfer from a gradient coordinate is $-\eta z^{-1}/(1-\lambda z^{-1})$. On the unit circle,

$$\sup_{|z|=1} |K_\lambda(z)| = \frac{2}{1+\lambda}, \qquad \sup_{|z|=1} \frac{1}{|1-\lambda z^{-1}|} = \frac{1}{1-|\lambda|}, \tag{52}$$

$$\sup_{|z|=1} |(1-\lambda)K_\lambda(z)| = \frac{2(1-\lambda)}{1+\lambda}, \qquad \sup_{|z|=1} \frac{1-\lambda}{|1-\lambda z^{-1}|} = \frac{1-\lambda}{1-|\lambda|}.$$

For completeness, the square of the first modulus is $2(1-c)/(1+\lambda^2-2\lambda c)$, where $c \coloneqq \cos\varphi$. Its derivative in c is $-2(1-\lambda)^2/(1+\lambda^2-2\lambda c)^2$, so its maximum is attained at $c=-1$. The other formulas follow by minimizing the denominator over $c \in [-1,1]$.

Diagonalize W orthogonally and apply Parseval's identity to each scalar coordinate. To apply these bounds on the finite horizon $0,\dots,T-1$, extend the inputs $E_0,\dots,E_{T-1}$ and $G_0,\dots,G_{T-1}$ by zero at all other times. The causal output agrees with equation 51 through time $T-1$. Its energy on this horizon is no larger than its full output energy, which is bounded by the corresponding transfer gain times the input energy. The bounds first hold for each realization and then in the Hilbert space of square-integrable random sequences. Thus, with $Y_t \coloneqq P_t(I-W)$,

$$\|P\|_{T,2} \le h\|E\|_{T,2} + \frac{\eta}{\rho}\|G\Pi\|_{T,2}, \tag{53}$$

$$\|Y\|_{T,2} \le d_{\mathrm{sp}}\|E\|_{T,2} + \eta\kappa\|G\Pi\|_{T,2} \le d_{\mathrm{sp}}\|E\|_{T,2} + \eta\kappa\|G\|_{T,2}. \tag{54}$$

In particular, E_T is not included in the inputs of these two inequalities. The cancellation between the direct residual term and its filtered difference is kept intact. Treating E_t and E_{t+1} as unrelated inputs to the resolvent would instead give the larger gain $2/\rho$.

Residual energy with an explicit stability margin. The clipping rule gives, including when the denominator is zero,

$$\|\Gamma_t \odot E_t\|_F \le \eta\|G_t\|_F + g_c\|E_t\|_F, \qquad \|(\mathbf{1}-\Gamma_t)\odot E_t\|_F \le \delta\|E_t\|_F. \tag{55}$$

To see the first bound, put $r_t^i \coloneqq \eta\|g_t^i\|/\|e_t^i\|$ when $e_t^i \neq 0$. If $r_t^i \ge g_c$, whether or not the upper clip is active, $\gamma_t^i \le r_t^i$ and hence $\gamma_t^i\|e_t^i\| \le \eta\|g_t^i\|$. If $r_t^i < g_c$, the lower clip gives $\gamma_t^i\|e_t^i\| = g_c\|e_t^i\|$. The two cases imply the sharper columnwise bound $(\gamma_t^i)^2\|e_t^i\|^2 \le \eta^2\|g_t^i\|^2 + g_c^2\|e_t^i\|^2$. The second bound follows from the symmetric clipping interval, $|1-\gamma_t^i| \le \delta$; when $e_t^i = 0$, both products are zero. Thus the first bound controls the compressed input B_t and the second controls the carryover in E_{t+1}. Minkowski's inequality now implies

$$\|B\|_{T,2} \le \|Y\|_{T,2} + 2\eta\|G\|_{T,2} + g_c\|E\|_{T,2}.$$

Since $E_0 = 0$,

$$\|E\|_{T,2}^2 \le \sum_{t=0}^{T-1}\mathbb{E}\|E_{t+1}\|_F^2.$$

Consequently, equation 48, equation 49, and Minkowski's inequality give

$$\begin{aligned}\|E\|_{T,2} &\le \sqrt{\beta}\|B\|_{T,2} + \delta\|E\|_{T,2}\\ &\le \{\sqrt{\beta}(d_{sp}+g_c)+\delta\}\|E\|_{T,2} + \sqrt{\beta}(\kappa+2)\eta\|G\|_{T,2}.\end{aligned}$$

The coefficient multiplying $\|E\|_{T,2}$ is strictly below one exactly when $1-\delta-\sqrt{\beta}(d_{sp}+g_c) = g_c - \sqrt{\beta}(d_{sp}+g_c) = \vartheta > 0$. Moving this term to the left yields $\|E\|_{T,2} \le C\eta\|G\|_{T,2}$, where $C = \sqrt{\beta}(\kappa+2)/\vartheta$. This is where the clipped choice of γ_t^i closes the residual energy estimate. The positive margin in equation 39 is the explicit sufficient compressor restriction $\beta < g_c^2/(d_{sp}+g_c)^2$. It is independent of the positive spectral gap when λ_- is bounded away from -1, but it does not certify arbitrarily aggressive compression. Its margin must be retained in every constant. We have proved, on every finite horizon,

$$S_E \coloneqq \sum_{t=0}^{T-1}\mathbb{E}\|E_t\|_F^2 \le C^2\eta^2 S_G, \qquad S_G \coloneqq \sum_{t=0}^{T-1}\mathbb{E}\|G_t\|_F^2. \tag{56}$$

When $\beta = 0$, $C = 0$ and the conclusion correctly gives $E_t = 0$.

Separating martingale noise from systematic disagreement. Define the filtered noise

$$N_t \coloneqq \sum_{k=0}^{t-1}\Xi_k W^{t-1-k}\Pi.$$

For $r < k$, Ξ_r is $\mathcal{F}_k$-measurable, so conditional unbiasedness makes all time cross terms vanish. Thus

$$\begin{aligned}\mathbb{E}\|N_t\|_F^2 &= \sum_{k=0}^{t-1}\mathbb{E}\|\Xi_k W^{t-1-k}\Pi\|_F^2 \\ &\le n\sigma^2\sum_{k=0}^{t-1}(1-\rho)^{2(t-1-k)} \le \frac{n\sigma^2}{\rho(2-\rho)}.\end{aligned} \tag{57}$$

This calculation does not require that future iterates or residuals be independent of Ξ_k. In particular, no cancellation with the residual component of P_t is assumed. Set

$$Q \coloneqq \sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2, \qquad S_P \coloneqq \sum_{t=0}^{T-1}\mathbb{E}\|P_t\|_F^2.$$

Apply equation 51 separately to F and Ξ, then use $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ to obtain

$$S_P \le 3h^2 S_E + \frac{3\eta^2}{\rho^2} \sum_{t=0}^{T-1} \mathbb{E}\|F_t \Pi\|_F^2 + 3\eta^2 n \sigma^2 T s_\rho. \tag{58}$$

For $F_t^\circ \coloneqq [\nabla f_1(\bar{x}_t), \dots, \nabla f_n(\bar{x}_t)]$, smoothness and equation 44 give

$$\|F_t \Pi\|_F^2 \le 2\|(F_t - F_t^\circ)\Pi\|_F^2 + 2\|F_t^\circ \Pi\|_F^2 \le 2L^2 \|P_t\|_F^2 + 2n\epsilon^2, \tag{59}$$

$$\|F_t\|_F^2 \le 2\|F_t - F_t^\circ\|_F^2 + 2\|F_t^\circ\|_F^2 \le 2L^2 \|P_t\|_F^2 + 2n\epsilon^2 + 2n\|\nabla f(\bar{x}_t)\|^2. \tag{60}$$

The last step uses the exact orthogonal decomposition of F_t° into its column mean and its mean-zero part. Conditional unbiasedness also removes the cross term between F_t and Ξ_t. It follows that

$$S_G \le 2L^2 S_P + nT(\sigma^2 + 2\epsilon^2) + 2nQ. \tag{61}$$

Combining equation 56, equation 58, and equation 61 with the constants in equation 40 gives

$$(1 - \eta^2 L^2 A_{\mathrm{A}}) S_P \le n\eta^2 (B_\sigma \sigma^2 T + B_\epsilon \epsilon^2 T + B_Q Q). \tag{62}$$

Descent of the residual-adjusted average. Let $\bar{e}_t \coloneqq E_t \mathbf{1}/n$, $y_t \coloneqq \bar{x}_t + \bar{e}_t$, and $\bar{F}_t \coloneqq F_t \mathbf{1}/n$. Taking the column average in equation 50 gives the exact recursion

$$y_{t+1} = y_t - \eta G_t \mathbf{1}/n.$$

Conditional smoothness and equation 45 imply

$$\mathbb{E}[f(y_{t+1}) \mid \mathcal{F}_t] \le f(y_t) - \eta \langle \nabla f(y_t), \bar{F}_t \rangle + \frac{L\eta^2}{2} \left(\|\bar{F}_t\|^2 + \frac{\sigma^2}{n} \right).$$

Furthermore,

$$\begin{aligned} \langle \nabla f(y_t), \bar{F}_t \rangle &\ge \frac{1}{2}\|\nabla f(\bar{x}_t)\|^2 + \frac{1}{2}\|\bar{F}_t\|^2 - \frac{1}{2}\|\nabla f(\bar{x}_t) - \bar{F}_t\|^2 \\ &\quad - \|\nabla f(y_t) - \nabla f(\bar{x}_t)\|^2 - \frac{1}{4}\|\bar{F}_t\|^2 \\ &\ge \frac{1}{2}\|\nabla f(\bar{x}_t)\|^2 + \frac{1}{4}\|\bar{F}_t\|^2 - \frac{L^2}{2n}\|P_t\|_F^2 - \frac{L^2}{n}\|E_t\|_F^2. \end{aligned}$$

Here Jensen's inequality bounds the two squared gradient differences, and $\|\bar{e}_t\|^2 \le \|E_t\|_F^2/n$. If $\eta L \le 1/2$, the remaining coefficient of $\|\bar{F}_t\|^2$ in the descent inequality is nonpositive. Sum the inequality and use $y_0 = x_0$ and $f(y_T) \ge f^*$ to conclude, with $\Delta \coloneqq f(x_0) - f^*$,

$$Q \le \frac{2\Delta}{\eta} + \frac{\eta L \sigma^2 T}{n} + \frac{L^2}{n}(S_P + 2S_E). \tag{63}$$

Final absorption. To verify the remaining absorption explicitly, the stepsize bound gives $\eta^2 L^2 A_{\mathrm{A}} \le 1/2$ and therefore

$$S_P \le 2n\eta^2 (B_\sigma \sigma^2 T + B_\epsilon \epsilon^2 T + B_Q Q).$$

Since $h \ge 1$ and $A_{\mathrm{A}} \ge 6C^2$, it also gives $4C^2\eta^2 L^2 \le 1/3$. Hence

$$\begin{aligned} S_P + 2S_E &\le (1 + 4C^2\eta^2 L^2) S_P + 2C^2 n\eta^2 (\sigma^2 + 2\epsilon^2) T + 4C^2 n\eta^2 Q \\ &\le 2S_P + 2C^2 n\eta^2 (\sigma^2 + 2\epsilon^2) T + 4C^2 n\eta^2 Q \\ &\le n\eta^2 (D_\sigma \sigma^2 T + D_\epsilon \epsilon^2 T + D_Q Q). \end{aligned}$$

Substitute this into equation 63 and use $\eta^2 L^2 D_Q \le 1/2$ to obtain equation 43. □

Explicit step-size optimization and interpretation. Choose any positive number $D \ge \Delta$; thus the following choice remains well-defined even when $\Delta = 0$. Put

$$a_T \coloneqq \frac{4D}{T}, \qquad b \coloneqq \frac{2L\sigma^2}{n}, \qquad c \coloneqq 2L^2 (D_\sigma \sigma^2 + D_\epsilon \epsilon^2),$$

and choose

$$\eta \coloneqq \min\left\{\eta_{\max}, \sqrt{\frac{a_T}{b}}, \left(\frac{a_T}{c}\right)^{1/3}\right\}. \tag{64}$$

A balancing entry is omitted, or assigned $+\infty$, when its denominator is zero. This always yields a positive step size. Because the reciprocal of a minimum is the maximum of the reciprocals and is no greater than their sum,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 \le \frac{a_T}{\eta_{\max}} + 2\sqrt{a_T b} + 2a_T^{2/3}c^{1/3}. \tag{65}$$

If $\sigma = \epsilon = 0$, both balancing entries are omitted and $\eta = \eta_{\max}$ gives the stronger bound $4\Delta/(\eta_{\max}T)$. If also $\Delta = 0$, the finite-step inequality proves that all the displayed expected gradient norms are zero. When $\Delta = 0$ with nonzero noise or heterogeneity, using any positive D avoids the invalid choice $\eta = 0$.

What the network dependence does and does not prove. For a family of networks and compressors satisfying fixed bounds $1 + \lambda_- \ge \nu_0 > 0$ and $\vartheta \ge \vartheta_0 > 0$, the quantities $h, d_{\mathrm{sp}}, \kappa, C$ are bounded independently of ρ. Then

$$D_\sigma = O(\rho^{-1}), \qquad D_\epsilon = O(\rho^{-2}), \qquad \eta_{\max}^{-1} = O(L/\rho).$$

Consequently, equation 65 yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 = O\left(\frac{\sigma\sqrt{LD}}{\sqrt{nT}} + \frac{(LD)^{2/3}}{T^{2/3}}\left(\frac{\sigma^2}{\rho} + \frac{\epsilon^2}{\rho^2}\right)^{1/3} + \frac{LD}{\rho T}\right). \tag{66}$$

The hidden constants may depend on ν_0 and ϑ_0, and this dependence cannot be suppressed if either margin deteriorates with the network. When $\epsilon = 0$, the network-dependent stochastic term has $\rho^{-1/3}$ dependence, improving the $\rho^{-2/3}$ dependence obtained by treating all stochastic gradients as arbitrary inputs. For heterogeneous objectives, the displayed systematic-disagreement term retains $\rho^{-2/3}$. This theorem does not establish a uniform advantage over every competing algorithm, nor does it imply that the stability restriction covers the experimental top-k ratios. Both comparisons require their own assumptions and constants.

F A CONSTRUCTIVE CONSENSUS GAIN FOR EVERY CONTRACTIVITY LEVEL

The preceding theorems have explicit stability margins. These margins can be enforced for any fixed $\beta < 1$ while keeping the given mixing matrix W fixed. The algorithms multiply their consensus corrections by a scalar gain ω. This changes the numerical updates and the effective mixing matrix seen by the recursion, while preserving the communication graph and the number of transmitted messages.

Proposition F.1 (An admissible consensus gain for arbitrary $\beta < 1$)**.** *Let $n \ge 2$ and let the fixed matrix W be symmetric, entrywise nonnegative, and doubly stochastic, with $\|W - J\|_2 < 1$. Define*

$$\rho \coloneqq 1 - \|W - J\|_2, \qquad \mu \coloneqq \|I - W\|_2 \in (0, 2), \qquad \rho_{\mathrm{alg}} \coloneqq 1 - \max_{\substack{v \perp \mathbf{1} \\ \|v\|=1}} v^\top W v.$$

Thus ρ_{alg} is the algebraic spectral gap and uses the largest algebraic *nontrivial eigenvalue; it is not defined by the largest absolute eigenvalue. Put*

$$\begin{aligned} g_{\mathrm{c}} &\coloneqq 1 - 1/\sqrt{2}, \qquad s \coloneqq \sqrt{\beta}, \qquad \chi \coloneqq 1 - s > 0, \\ \omega &\coloneqq \min\left\{1, \frac{1}{\mu}, \frac{g_{\mathrm{c}}\chi}{4s\mu}\right\}, \qquad W_\omega \coloneqq (1-\omega)I + \omega W. \end{aligned} \tag{67}$$

The last entry in the minimum is $+\infty$ when $s = 0$. Use the fixed W and consensus gain ω in both algorithms, and set $\tau = 1/2$ in A-DEFEAT-GT. Then the auxiliary effective matrix W_ω is symmetric, nonnegative, doubly stochastic and positive semidefinite, and

$$\widetilde{\mu} \coloneqq \|I - W_\omega\|_2 = \omega\mu \le 1, \qquad \widetilde{\rho} \coloneqq 1 - \|W_\omega - J\|_2 = \omega\rho_{\mathrm{alg}} \ge \omega\rho, \tag{68}$$

$$\vartheta = g_{\mathrm{c}} - s(d_{\mathrm{sp}} + g_{\mathrm{c}}) \ge \frac{g_{\mathrm{c}}\chi}{2} > 0, \qquad m = 1 - s(1 + 2\tau\widetilde{\mu}) \ge \chi\left(1 - \frac{g_{\mathrm{c}}}{4}\right) > 0, \tag{69}$$

where $d_{\mathrm{sp}} \coloneqq 2\widetilde{\mu}/(2-\widetilde{\mu})$. *Moreover,*

$$\omega \geq \min\left\{\frac{1}{2}, \frac{g_{\mathrm{c}}\chi}{8s}\right\} \geq \frac{g_{\mathrm{c}}\chi}{8}, \qquad \widetilde{\rho} \geq \frac{g_{\mathrm{c}}\chi\rho}{8}. \tag{70}$$

The entry involving s^{-1} *is again interpreted as* $+\infty$ *at* $s = 0$. *Consequently, for every fixed* $\beta < 1$, *evaluating the spectral quantities at* W_ω *makes the model-compression margin* ϑ *and tracker-compression margin* m *required by Theorem D.1 for A-DEFEAT-GT strictly positive, with* $\tau = 1/2$ *satisfying its tracker-stepsize restriction. It also makes the model-compression margin* ϑ *required by Theorem E.1 for A-DEFEAT strictly positive.*

Proof. All eigenvalues of W_ω are of the form $1-\omega(1-\lambda)$, where λ is an eigenvalue of W. The choice $\omega\mu \leq 1$ places these eigenvalues in $[0,1]$, and the nontrivial eigenvalues are strictly smaller than 1. This proves positive semidefiniteness and the exact formulas in equation 68. The elementary comparison

$$\max_{v\perp\mathbf{1},\,\|v\|=1} v^\top W v \leq \|W - J\|_2$$

gives $\rho_{\mathrm{alg}} \geq \rho$. Convex combination with I preserves entrywise nonnegativity, symmetry and double stochasticity.

When $s > 0$, the third entry in the minimum defining ω gives $s\widetilde{\mu} \leq g_{\mathrm{c}}\chi/4$; the same inequality holds trivially when $s = 0$. Since $\widetilde{\mu} \leq 1$,

$$h \coloneqq \frac{2}{2-\widetilde{\mu}} \leq 2, \qquad \kappa \coloneqq \max\left\{1, \frac{\widetilde{\mu}}{2-\widetilde{\mu}}\right\} = 1, \qquad d_{\mathrm{sp}} = \frac{2\widetilde{\mu}}{2-\widetilde{\mu}} \leq 2\widetilde{\mu}.$$

It follows that

$$\vartheta = g_{\mathrm{c}}(1-s) - s d_{\mathrm{sp}} \geq g_{\mathrm{c}}\chi - 2s\widetilde{\mu} \geq g_{\mathrm{c}}\chi/2, \qquad m = 1 - s - s\widetilde{\mu} \geq \chi - g_{\mathrm{c}}\chi/4.$$

The choice $\tau = 1/2$ also satisfies $0 < \tau \leq \min\{1, 1/\widetilde{\mu}\}$.

Finally, $\mu < 2$ implies $1/\mu > 1/2$ and, for $s > 0$, $g_{\mathrm{c}}\chi/(4s\mu) > g_{\mathrm{c}}\chi/(8s)$. Taking minima proves the first lower bound on ω. Both $1/2$ and $g_{\mathrm{c}}\chi/(8s)$ are at least $g_{\mathrm{c}}\chi/8$, since $0 < \chi \leq 1$ and $0 < s < 1$; at $s = 0$ the claim follows directly. The remaining gap bound follows from equation 68. □

Explicit compression costs. The exact constants in the finite-horizon convergence theorems remain the authoritative bounds. The following estimates expose the cost of enforcing their margins through this construction. In the estimates below, all asymptotic constants are numerical constants, since g_{c} is fixed. When applying the fixed-margin proofs in this subsection, their spectral symbols ρ and μ are evaluated at the auxiliary matrix W_ω and therefore equal $\widetilde{\rho}$ and $\widetilde{\mu}$, respectively. The unadorned ρ below remains the spectral gap of the fixed input matrix W. Write $q_0 \coloneqq 1 - g_{\mathrm{c}}/4$ and

$$C \coloneqq \frac{s(\kappa+2)}{\vartheta}.$$

By the proposition, $\kappa = 1$ and

$$C \leq \frac{6s}{g_{\mathrm{c}}\chi}, \qquad C \leq \frac{3}{2\widetilde{\rho}}. \tag{71}$$

For $s > 0$, the second inequality follows from $\widetilde{\rho} \leq \widetilde{\mu} \leq g_{\mathrm{c}}\chi/(4s)$ and the first; at $s = 0$, $C = 0$.

F.1 A-DEFEAT-GT

Theorem F.2 (A-DEFEAT-GT with general consensus gain and compression level)**.** *Fix* $0 \leq \beta < 1$ *and choose any* $0 < \omega \leq 1$ *that satisfies the conditions below. The value in equation 67 is an explicit sufficient choice and is used in Corollary F.3. Write* $\rho \coloneqq 1 - \|W - J\|_2$ *and* $\chi \coloneqq 1 - \sqrt{\beta}$. *Define* $W_\omega \coloneqq (1-\omega)I + \omega W$, $\widetilde{\rho} \coloneqq 1 - \|W_\omega - J\|_2$, $\widetilde{\mu} \coloneqq \|I - W_\omega\|_2$, $\widetilde{\nu} \coloneqq 2 - \widetilde{\mu}$, *and* $\widetilde{d}_{\mathrm{sp}} \coloneqq 2\widetilde{\mu}/(2-\widetilde{\mu})$. *Choose* $0 < \tau \leq \min\{1, 1/\widetilde{\mu}\}$ *and define*

$$\widetilde{\vartheta} \coloneqq g_{\mathrm{c}} - \sqrt{\beta}(\widetilde{d}_{\mathrm{sp}} + g_{\mathrm{c}}), \qquad \widetilde{m} \coloneqq 1 - \sqrt{\beta}(1 + 2\tau\widetilde{\mu}).$$

When Assumptions 3.1, 3.3, and 3.4 hold and the compressor satisfies Definition 3.2, use Algorithm 1 with the fixed matrix W and consensus gain ω. If $\min\{\widetilde{\nu}, \widetilde{\vartheta}, \widetilde{m}, \tau\} = \Omega(1)$, let $D \geq \Delta$ be any positive number and let η be specified in equation 34 of Appendix D.1, with all spectral and stability quantities evaluated at W_ω. Then

$$\mathcal{R}_T = O\left(\frac{\sigma\sqrt{LD}}{\sqrt{nT}} + \frac{(LD\sigma)^{2/3}}{\widetilde{\rho}^{2/3}T^{2/3}} + \frac{LD + \mathcal{J}_0}{\widetilde{\rho}T}\right),$$

where $\mathcal{J}_0 \coloneqq \mathbb{E}\|G_0\|_F^2/n$.

Proof. The two consensus corrections in Algorithm 1 are algebraically identical to the unit-gain corrections formed with W_ω. Applying Theorem D.1 to this auxiliary matrix proves the result. □

For A-DEFEAT-GT, define numerical constants and compression-dependent upper bounds

$$a_0 \coloneqq 2 + \frac{g_c}{2q_0}, \qquad V_\beta \coloneqq 2 + \frac{24s}{g_c\chi}, \qquad B_\beta \coloneqq \frac{4}{3} + \frac{8s}{3q_0\chi}, \tag{72}$$

$$K_\beta \coloneqq \max\{4,\, 2a_0V_\beta,\, \sqrt{1968}\}, \qquad \bar{\eta}_{\mathrm{GT}} \coloneqq \frac{\widetilde{\rho}}{LK_\beta}. \tag{73}$$

The constants in equation 18 obey

$$H_{\mathrm{GT}} \leq \frac{41}{4\widetilde{\rho}^2}, \qquad A_{\mathrm{GT}} \leq \frac{a_0}{\widetilde{\rho}}, \qquad B_{\mathrm{GT}} \leq B_\beta, \qquad v \leq V_\beta, \qquad \bar{\eta}_{\mathrm{GT}} \leq \eta_*. \tag{74}$$

Indeed, $h \leq 2$ and equation 71 give

$$H_{\mathrm{GT}} = \tfrac{1}{2}(\widetilde{\rho}^{-1} + hC)^2 + C^2 \leq \frac{8}{\widetilde{\rho}^2} + \frac{9}{4\widetilde{\rho}^2} = \frac{41}{4\widetilde{\rho}^2}.$$

Also $m \geq q_0\chi$, $s\widetilde{\mu} \leq g_c\chi/4$, and $\widetilde{\rho} \leq \widetilde{\mu}$ imply

$$A_{\mathrm{GT}} = \frac{2}{\widetilde{\rho}} + \frac{2s}{m} \leq \frac{2}{\widetilde{\rho}} + \frac{g_c}{2q_0\widetilde{\rho}}.$$

Since $\tau = 1/2$ and $\widetilde{\mu} \leq 1$, one has $k_{\mathrm{tr}} \leq 4/3$ and $\tau\widetilde{\mu}k_{\mathrm{tr}} \leq 2/3$, proving $B_{\mathrm{GT}} \leq B_\beta$. The bound on v is immediate from $v = h(1 + 2C)$, and the three terms defining K_β verify each of the three step-size restrictions separately.

Corollary F.3 (Explicit dependence on β for A-DEFEAT-GT)**.** *When Assumptions 3.1, 3.3, and 3.4 hold and the compressor satisfies Definition 3.2 for a fixed $0 \leq \beta < 1$, use Algorithm 1 with the fixed matrix W, consensus gain ω specified in equation 67, and $\tau = 1/2$. Let $D \geq \Delta$ be any positive number, and let η be specified in equation 34 of Appendix D.1, using the cap $\bar{\eta}_{\mathrm{GT}}$ in equation 73. Write $F_{\Pi,0} \coloneqq \|F_0\Pi\|_F$ and $\mathcal{G}_0^2 \coloneqq \mathbb{E}\|G_0\|_F^2$. Then*

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla f(\bar{x}_t)\|^2 = O\Bigg(\frac{\sigma\sqrt{LD}}{\sqrt{nT}} + \frac{(LD\sigma)^{2/3}}{(1-\beta)^{4/3}\rho^{2/3}T^{2/3}} + \frac{LD}{(1-\beta)^2\rho T} + \frac{F_{\Pi,0}^2}{n\rho T} + \frac{\mathcal{G}_0^2}{nT}\Bigg). \tag{75}$$

The hidden constants in the last display are numerical and independent of n, T, ρ, and β.

Proof. Proposition F.1 verifies the margins for the prescribed gain, and Theorem F.2 applies with the auxiliary matrix W_ω. The estimates above give the stated cap and all noninitial terms in equation 75. It remains to check that no compression penalty was hidden in the two initialization terms. The formula for I_0 and $\widetilde{\mu}/(4 - \widetilde{\mu}) \leq 1/3$ give

$$I_0^2 \leq \frac{4F_{\Pi,0}^2}{\widetilde{\rho}} + \frac{16\mathcal{G}_0^2}{3q_0^2\chi^2},$$

where $F_{\Pi,0}^2 \le \mathcal{G}_0^2$ follows from the deterministic initialization and unbiasedness. The optimized initialization term is at most

$$\frac{24L^2 H_{\mathrm{GT}} \bar{\eta}_{\mathrm{GT}}^2 I_0^2}{nT} \le \frac{246 I_0^2}{K_\beta^2 nT} \le \frac{492}{a_0^2 g_{\mathrm{c}}} \frac{F_{\Pi,0}^2}{n\rho T} + \frac{82}{a_0^2 q_0^2} \frac{\mathcal{G}_0^2}{nT}.$$

For the last inequality, note that $K_\beta \chi \ge 2a_0 V_\beta \chi \ge 4a_0$ and $\widetilde{\rho} \ge g_{\mathrm{c}} \chi \rho / 8$. The remaining terms in equation 75 follow from $B_\beta = O(\chi^{-1})$, $K_\beta = O(\chi^{-1})$, $H_{\mathrm{GT}} = O(\chi^{-2}\rho^{-2})$, and the same step-size balancing calculation as in the GT proof. Since

$$\chi = 1 - \sqrt{\beta} = \frac{1-\beta}{1+\sqrt{\beta}},$$

the displayed powers of χ^{-1} are equivalent up to numerical constants to the corresponding powers of $(1-\beta)^{-1}$. □

F.2 A-DEFEAT

Theorem F.4 (A-DEFEAT with general consensus gain and compression level)**.** *Fix $0 \le \beta < 1$ and choose any $0 < \omega \le 1$. Define*

$$W_\omega \coloneqq (1-\omega)I + \omega W, \qquad \widetilde{\rho} \coloneqq 1 - \|W_\omega - J\|_2, \qquad \widetilde{\mu} \coloneqq \|I - W_\omega\|_2,$$

$$\widetilde{\nu} \coloneqq 2 - \widetilde{\mu}, \qquad \widetilde{d}_{\mathrm{sp}} \coloneqq \frac{2\widetilde{\mu}}{2-\widetilde{\mu}}, \qquad \widetilde{\vartheta} \coloneqq g_{\mathrm{c}} - \sqrt{\beta}(\widetilde{d}_{\mathrm{sp}} + g_{\mathrm{c}}).$$

The value in equation 67 is an explicit sufficient choice and is used in Corollary F.5. When Assumptions 3.1, 3.3, 3.4, and 3.5 hold and the compressor satisfies Definition 3.2, use Algorithm 2 with the fixed matrix W and consensus gain ω. If $\min\{\widetilde{\nu}, \widetilde{\vartheta}\} = \Omega(1)$, let $D \ge \Delta$ be any positive number and let η be specified in equation 64 of Appendix E.2, with all spectral and stability quantities evaluated at W_ω. Then

$$\mathcal{R}_T = O\left(\frac{\sigma\sqrt{LD}}{\sqrt{nT}} + \frac{(LD)^{2/3}}{T^{2/3}} \left(\frac{\sigma^2}{\widetilde{\rho}} + \frac{\epsilon^2}{\widetilde{\rho}^2} \right)^{1/3} + \frac{LD}{\widetilde{\rho} T} \right).$$

Proof. The model consensus correction in Algorithm 2 is algebraically identical to the unit-gain correction formed with W_ω. Applying Theorem E.1 to this auxiliary matrix and using the balancing argument in equation 65 proves the result. □

For A-DEFEAT, use the constants in equation 40 and equation 41. They satisfy

$$A_{\mathrm{A}} = 6h^2C^2 + 6/\widetilde{\rho}^2 \le 60/\widetilde{\rho}^2, \qquad D_Q \le 225/\widetilde{\rho}^2, \tag{76}$$

$$D_\sigma \le \frac{1800 s^2}{g_{\mathrm{c}}^2 \chi^2} + \frac{12}{\widetilde{\rho}}, \qquad D_\epsilon \le \frac{3600 s^2}{g_{\mathrm{c}}^2 \chi^2} + \frac{24}{\widetilde{\rho}^2}. \tag{77}$$

Thus $\bar{\eta}_A \coloneqq \widetilde{\rho}/(\sqrt{450}L)$ is an admissible upper step-size cap. In particular, the displayed constants imply

$$D_\sigma = O\left(\frac{1}{\chi^2 \rho}\right), \qquad D_\epsilon = O\left(\frac{1}{\chi^2 \rho^2}\right), \qquad \bar{\eta}_A^{-1} = O\left(\frac{L}{\chi\rho}\right).$$

Corollary F.5 (Explicit dependence on β for A-DEFEAT)**.** *When Assumptions 3.1, 3.3, 3.4, and 3.5 hold and the compressor satisfies Definition 3.2 for a fixed $0 \le \beta < 1$, use Algorithm 2 with the fixed matrix W and consensus gain ω specified in equation 67. Let $D \ge \Delta$ be any positive number, and let η be specified in equation 64 of Appendix E.2, using $\bar{\eta}_A \coloneqq \widetilde{\rho}/(\sqrt{450}L)$ as its cap. Then*

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\|\nabla f(\bar{x}_t)\|^2 = O\left(\frac{\sigma\sqrt{LD}}{\sqrt{nT}} + \frac{(LD)^{2/3}}{(1-\beta)^{2/3}T^{2/3}} \left(\frac{\sigma^2}{\rho} + \frac{\epsilon^2}{\rho^2} \right)^{1/3} + \frac{LD}{(1-\beta)\rho T} \right). \tag{78}$$

The hidden constants are numerical and independent of n, T, ρ, and β.

Proof. The gain in equation 67 satisfies the margin required by Theorem F.4. To check the remaining constants, use $h \le 2$ and $C \le 3/(2\widetilde{\rho})$ in equation 76, and use $C \le 6s/(g_c\chi)$ with $[\widetilde{\rho}(2-\widetilde{\rho})]^{-1} \le \widetilde{\rho}^{-1}$ in equation 77. The admissible cap then follows directly from equation 42. Substitution of $\widetilde{\rho} \ge g_c\chi\rho/8$ and $\rho, \chi \le 1$ gives the stated orders; the balancing inequality proves equation 78. Since

$$\chi = 1 - \sqrt{\beta} = \frac{1-\beta}{1+\sqrt{\beta}},$$

the powers $\chi^{-2/3}$ and χ^{-1} are equivalent up to numerical constants to $(1-\beta)^{-2/3}$ and $(1-\beta)^{-1}$, respectively. □

Interpretation and experimental scope. For each fixed compression level, the construction preserves the proved powers of the fixed matrix's spectral gap ρ. Increasing compression strength reduces the consensus gain and can increase the iteration complexity through the explicit penalties above. The matrix W and every communication edge remain fixed; only the consensus corrections are multiplied by ω. The auxiliary matrix W_ω has gap $\widetilde{\rho} = \omega\rho_{\mathrm{alg}}$ and is used only in the proof. The experiments use the same fixed-W update and set ω equal to one. Theorems 4.1 and 4.2 apply to this unit-gain implementation when their stated margins are positive. Theorem F.2 extends the analysis to general admissible gains. Proposition F.1 and Corollaries F.5 and F.3 give a sufficient theoretical choice and explicit rates for every fixed $\beta < 1$.

G Convergence and algorithm comparisons

We compare expected squared-gradient criteria. Our rows use $\mathcal{R}_T$ at the averaged model, while the DoCoM row retains the source paper's average over local iterates. The baseline rows follow DCD (Tang et al., 2018), CHOCO-SGD (Koloskova et al., 2020), DeepSqueeze (Tang et al., 2019), BEER (Zhao et al., 2022), CEDAS (Huang & Pu, 2024), DoCoM (Yau & Wai, 2023), and MoTEF (Islamov et al., 2024). Table 5 presents the leading stochastic term together with the topology-dependent terms. It suppresses smoothness constants and positive initial objective gaps, but displays gradient bounds and initial tracker energy where they remain. Only spectral-gap quantities are colored; compression factors remain black. For A-DEFEAT and A-DEFEAT-GT, $\rho := 1 - \|W - J\|_2$ is the gap of the fixed input matrix W, and the table uses the consensus gain in Proposition F.1. The derivation uses $\chi := 1 - \sqrt{\beta}$; because $\chi = (1-\beta)/(1+\sqrt{\beta})$, Corollaries F.5 and F.3 state the equivalent powers of $(1-\beta)^{-1}$ directly.

Model communication and CHOCO-style methods. CHOCO-SGD (Koloskova et al., 2019; 2020) introduced compressed-surrogate gossip for the model. BEER (Zhao et al., 2022) adds compressed gradient tracking to that scheme, and MoTEF (Islamov et al., 2024) combines the surrogate gossip with momentum tracking. In BEER (Algorithm 1, lines 3–5) and MoTEF (Algorithm 1, lines 4–6), the actual model X_t is updated separately from a communicated surrogate H_t. Both use $Q^h_{t+1} = \mathcal{C}(X_{t+1} - H_t)$ and $H_{t+1} = H_t + Q^h_{t+1}$, while model consensus uses $H_t(W - I)$. Compression can leave $X_t - H_t \neq 0$, so consensus can use approximate model states. By contrast, in A-DEFEAT-GT (Algorithm 1, lines 7–8 and 12–13) and A-DEFEAT (Algorithm 2, lines 7–9), node i applies the compressed increment v^i_t to its model and sends the same increment to its neighbors. Since the copies start equal and receive identical updates, $x^{j,i}_t = x^i_t$ whenever $W_{ji} > 0$. Their model-consensus corrections thus use actual neighbor models. A-DEFEAT-GT still communicates a compressed tracker surrogate.

Adaptive error feedback. The second difference is how the model compression error is corrected. Both proposed algorithms store it in e^i_t and choose $\gamma^i_t = \mathrm{clip}_{[g_c, 2-g_c]}(\eta\|d^i_t\|/\|e^i_t\|)$ at each node and iteration, with $d^i_t = u^i_t$ for A-DEFEAT-GT and $d^i_t = g^i_t$ for A-DEFEAT and the stated zero-residual convention. They inject $\gamma^i_t e^i_t$ into the current compressed input and carry $(1-\gamma^i_t)e^i_t$ to the next iteration. The fixed consensus gains in CHOCO-SGD, BEER, and MoTEF control mixing, rather than selecting this residual fraction from the current update and residual magnitudes. DeepSqueeze (Tang et al., 2019) uses error compensation with the full stored residual, but not this clipped, node-specific rule. The comparison does not place every baseline in the CHOCO family. DCD (Tang et al., 2018) predates CHOCO, while CEDAS (Huang & Pu, 2024) and DoCoM (Yau & Wai, 2023) use different compressed-diffusion and momentum-tracking constructions.

Table 5. Convergence-rate comparison for the expected squared-gradient criterion. Assumptions and initialization differ between rows and are summarized in the final column.

Algorithm	Convergence Rate	Assumption
DCD	$\mathcal{O}_{D_1,D_2}\left(\frac{\sigma}{\sqrt{nT}}+\frac{\epsilon^{2/3}}{T^{2/3}}+\frac{1}{T}\right)$	Unbiased compression; $\rho^2-4\mu^2\beta>0$; D_1, D_2 retain the reciprocal of this margin
CHOCO-SGD	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}}+\frac{G^{2/3}}{(1-\beta)^{2/3}\rho^{4/3}T^{2/3}}+\frac{1}{T}\right)$	Bounded stochastic gradients $\mathbb{E}\|g_i(x)\|^2\le G^2$
DeepSqueeze	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}}+\frac{q_\beta\epsilon^{2/3}}{\rho_{\text{alg}}^2T^{2/3}}+\frac{1}{T}+\frac{q_\beta\sigma}{\rho_{\text{alg}}^2T}\right)$, $q_\beta\coloneqq 1+\frac{\beta}{\beta^{-1/3}-1}$	Bounded variance and gradient divergence; $\rho_{\text{alg}}\coloneqq 1-\lambda_2(W)$; q_0 is defined by continuity
BEER	$\mathcal{O}\left(\frac{\Phi_0-\Phi_T}{(1-\beta)\rho^3T}+\frac{\sigma^2}{b(1-\beta)}\right)$	Source Lyapunov initialization; mini-batch size b
CEDAS	$\mathcal{O}\left(\frac{1}{\sqrt{nT}}+\frac{n}{\rho_{\text{CEDAS}}T}+\frac{\|x_{-1}\|^2+\|H_0\|^2}{n\rho_{\text{CEDAS}}T}+\frac{\|\nabla F(x_{-1})\|^2}{\rho_{\text{CEDAS}}^2T^2}\right)$	Unbiased compression; source initialization; $\widetilde{W}\coloneqq I-\gamma(I-W)/2$, $\rho_{\text{CEDAS}}\coloneqq 1-\lambda_2(\widetilde{W})>0$
DoCoM	$\mathcal{O}\left(\frac{1+\sigma^2}{(nT)^{2/3}}+\frac{n\overline{G}_0}{(1-\beta)^2\rho^4T}+\frac{\sigma^2n^{5/3}}{(1-\beta)^3\rho^6T^{4/3}}\right)$	Average over local iterates; mean-square smoothness; source initial-gradient energy $\overline{G}_0$
MoTEF	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}}+\frac{\sigma^{2/3}}{(1-\beta)^{2/3}\rho^{5/3}T^{2/3}}+\frac{1}{(1-\beta)\rho^3T}\right)$	Source-theorem initialization and initial-batch condition
A-DEFEAT	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}}+\frac{\left(\sigma^2/\rho+\epsilon^2/\rho^2\right)^{1/3}}{(1-\beta)^{2/3}T^{2/3}}+\frac{1}{(1-\beta)\rho T}\right)$	Bounded gradient divergence; fixed W; consensus gain ω chosen by equation 67
A-DEFEAT-GT	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}}+\frac{\sigma^{2/3}}{(1-\beta)^{4/3}\rho^{2/3}T^{2/3}}+\frac{1}{(1-\beta)^2\rho T}+\frac{\mathcal{J}_0}{\rho T}\right)$	Fixed W; consensus gain ω chosen by equation 67; $\mathcal{J}_0\coloneqq\mathbb{E}\|G_0\|_F^2/n$; no bounded gradient or heterogeneity assumption

H LIMITATIONS OF NETWORK CONNECTIVITY AND COMPRESSION

This section clarifies how network connectivity and compression affect the guarantees for A-DEFEAT-GT and A-DEFEAT. The discussion concerns the sufficient upper bounds proved in this paper. It does not establish lower bounds, and therefore the powers below should not be interpreted as showing that the corresponding deterioration is unavoidable for either algorithm.

Impact of the spectral gap ρ. The quantity $\rho = 1 - \|W - J\|_2$ measures the worst-case contraction of disagreement for the prescribed mixing matrix. A small ρ corresponds to slow information propagation and enlarges the finite-horizon network terms in both methods. At a fixed compression level and under uniform positive stability margins, A-DEFEAT-GT has a network-dependent stochastic term of order $\rho^{-2/3}T^{-2/3}$ and a transient term of order $\rho^{-1}T^{-1}$; see equation 10. For A-DEFEAT, the network-dependent stochastic term is $\rho^{-1/3}T^{-2/3}$ when the gradient-divergence parameter is $\epsilon = 0$, whereas persistent heterogeneity contributes $\rho^{-2/3}T^{-2/3}$; see equation 11. Thus the network-independent term of order $T^{-1/2}$ can be masked at practical horizons by slow mixing. For example, the equal-weight ring considered in Appendix D.2 has $\rho = \Theta(n^{-2})$, so enlarging the network can make these lower-order terms substantial. The algorithms improve the proved powers of ρ relative to the comparison methods, but they do not remove this connectivity bottleneck.

The analysis also assumes a fixed, connected, undirected network with a symmetric, nonnegative, doubly stochastic mixing matrix. Its exact constants retain the lower-spectrum margin $\nu = 1 + \lambda_-$. Consequently, a statement that is uniform over a graph family requires ν as well as the relevant stability margins to remain bounded away from zero. The present guarantees do not directly cover directed, time-varying, or asynchronous communication.

Impact of the compression parameter β. Definition 3.2 permits any fixed $0 \le \beta < 1$; larger β gives a weaker contraction guarantee and represents more aggressive compression. When β is close to one, the unit-gain stability margins in equation 9 can fail. The constructive choice in Proposition F.1

restores the margins by reducing the consensus gain ω. This stabilization makes the effective spectral gap proportional, in the worst case, to $(1-\sqrt{\beta})\rho$ and therefore slows mixing.

The resulting explicit rates quantify this trade-off. For A-DEFEAT, the $T^{-2/3}$ network term is multiplied by $(1-\beta)^{-2/3}$ and the transient term by $(1-\beta)^{-1}$ in equation 78. For A-DEFEAT-GT, the corresponding factors are $(1-\beta)^{-4/3}$ and $(1-\beta)^{-2}$ in equation 75. The stronger worst-case dependence for A-DEFEAT-GT reflects the need to stabilize compressed model and tracker recursions simultaneously. Hence poor connectivity and aggressive compression compound one another: a small ρ and a β near one can make the provable constants and admissible stepsizes conservative even though convergence holds for every fixed $\beta < 1$ under the stated construction. These factors arise from a sufficient gain and proof argument; they are not claimed to be optimal compression exponents or exact empirical thresholds.

Algorithm-specific trade-offs. A-DEFEAT communicates one compressed vector per aggregation, but without a tracker it requires Assumption 3.5. Its guarantee can therefore be loose or inapplicable when local objectives have unbounded or extreme gradient divergence, and the associated heterogeneity penalty is amplified on poorly connected graphs. A-DEFEAT-GT avoids this assumption, but it stores and communicates an additional tracker, sends two compressed vectors per aggregation, and has the stronger explicit dependence on β above. Neither variant is uniformly preferable: their relative benefit depends on data heterogeneity, the graph, the compression level, and whether cost is measured by aggregation rounds or transmitted vectors.

Finally, the experiments use $\omega = 1$, while Proposition F.1 supplies a conservative sufficient value for arbitrary fixed $\beta < 1$. Empirical success with unit gain does not by itself verify the theorem's stability margins. Moreover, the proofs use the clipping interval implied by the adaptive residual rule; they certify the proposed rule but do not prove that it dominates every fixed residual coefficient satisfying the same bounds. Natural extensions include topology-aware mixing or multiple gossip steps for small ρ, less conservative compression-aware gain selection for large β, and guarantees for directed, time-varying, or asynchronous networks.

I Additional experiments

Evaluation details. Each entry is the mean and standard deviation across the three final seeds after first averaging each run over its last 50 rounds. The A-DEFEAT-GT runs retain g_t^i and use $g_{t+1}^i - g_t^i$ in the tracker update, together with the tracker direction u_t^i in the clipping rule. The clipping coefficient uses the Euclidean norm and the interval $[g_c, 2 - g_c]$. Both proposed methods use $\omega = 1$ in every experiment. These choices implement Algorithms 1 and 2.

I.1 FashionMNIST

I.1.1 Ring with fixed aggregations

Table 6. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	72.2 ± 0.8	73.5 ± 0.5	70.2 ± 0.9	71.9 ± 0.1	72.9 ± 0.3
CHOCO	78.3 ± 0.2	80.8 ± 0.2	81.1 ± 0.1	80.5 ± 0.1	80.1 ± 0.4
DeepSqueeze	76.9 ± 0.3	80.5 ± 0.1	80.1 ± 0.2	80.0 ± 0.1	79.6 ± 0.4
BEER	79.9 ± 0.1	80.8 ± 0.2	82.0 ± 0.4	81.4 ± 0.1	81.5 ± 0.6
MoTEF	81.6 ± 0.2	82.3 ± 0.1	82.4 ± 0.1	81.8 ± 0.6	79.8 ± 0.3
A-DEFEAT	80.7 ± 0.1	82.5 ± 0.1	82.6 ± 0.1	82.2 ± 0.1	81.9 ± 0.1
A-DEFEAT-GT	$\mathbf{82.5 \pm 0.2}$	$\mathbf{83.2 \pm 0.2}$	$\mathbf{83.4 \pm 0.2}$	$\mathbf{83.1 \pm 0.0}$	$\mathbf{83.0 \pm 0.1}$

Table 7. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.4 ± 0.1	79.6 ± 0.2	79.3 ± 0.3	79.3 ± 0.2	79.0 ± 0.2
CHOCO	83.7 ± 0.2	83.8 ± 0.1	83.6 ± 0.1	83.7 ± 0.1	83.5 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.6 ± 0.1	83.2 ± 0.1	83.3 ± 0.0	83.0 ± 0.1
BEER	83.8 ± 0.2	83.9 ± 0.2	83.8 ± 0.2	83.8 ± 0.2	83.7 ± 0.3
MoTEF	83.7 ± 0.2	83.8 ± 0.1	83.6 ± 0.1	83.8 ± 0.2	83.7 ± 0.2
A-DEFEAT	84.0 ± 0.2	84.1 ± 0.1	84.0 ± 0.1	84.0 ± 0.2	83.9 ± 0.2
A-DEFEAT-GT	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.2 \pm 0.1}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.0 \pm 0.1}$

I.1.2 Ring with fixed communications

Table 8. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	65.8 ± 0.2	65.8 ± 0.2	65.8 ± 0.2	65.7 ± 0.2	65.7 ± 0.2
CHOCO	65.1 ± 1.1	65.5 ± 1.4	65.8 ± 0.8	65.8 ± 0.8	65.7 ± 1.2
DeepSqueeze	66.1 ± 0.4	67.8 ± 1.5	66.7 ± 2.0	68.7 ± 1.7	70.5 ± 1.0
BEER	72.9 ± 0.3	73.4 ± 0.3	73.4 ± 0.2	73.3 ± 0.3	73.4 ± 0.3
MoTEF	75.7 ± 0.2	75.6 ± 0.2	75.4 ± 0.2	76.4 ± 0.4	76.2 ± 0.2
A-DEFEAT	75.2 ± 0.8	74.2 ± 0.0	74.2 ± 0.1	74.2 ± 0.1	74.3 ± 0.1
A-DEFEAT-GT	$\mathbf{77.1 \pm 0.2}$	$\mathbf{77.3 \pm 0.5}$	$\mathbf{77.9 \pm 0.5}$	$\mathbf{77.7 \pm 0.7}$	$\mathbf{77.0 \pm 0.1}$

Table 9. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.1 ± 0.1	65.7 ± 1.2	62.1 ± 0.3	62.8 ± 2.2	65.3 ± 0.7
CHOCO	74.0 ± 0.4	75.0 ± 0.1	68.1 ± 0.4	73.5 ± 2.0	76.3 ± 0.4
DeepSqueeze	71.8 ± 0.6	75.5 ± 0.1	68.7 ± 0.5	74.5 ± 0.5	76.5 ± 0.5
BEER	76.4 ± 0.9	76.1 ± 0.0	66.2 ± 0.8	73.1 ± 1.5	74.3 ± 1.4
MoTEF	77.8 ± 0.3	78.1 ± 0.5	69.1 ± 0.2	72.2 ± 2.3	72.3 ± 1.4
A-DEFEAT	78.0 ± 0.5	$\mathbf{79.9 \pm 0.3}$	76.2 ± 0.3	$\mathbf{79.4 \pm 0.2}$	$\mathbf{78.8 \pm 0.1}$
A-DEFEAT-GT	$\mathbf{79.7 \pm 0.3}$	79.7 ± 0.2	$\mathbf{77.2 \pm 0.6}$	78.7 ± 0.4	77.9 ± 0.4

Table 10. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	72.2 ± 0.8	73.5 ± 0.5	70.2 ± 0.9	71.9 ± 0.1	72.9 ± 0.3
CHOCO	78.3 ± 0.2	80.8 ± 0.2	81.1 ± 0.1	80.5 ± 0.1	80.1 ± 0.4
DeepSqueeze	76.9 ± 0.3	80.5 ± 0.1	80.1 ± 0.2	80.0 ± 0.1	79.6 ± 0.4
BEER	76.4 ± 0.2	79.6 ± 0.2	79.0 ± 0.2	78.8 ± 0.2	79.2 ± 0.5
MoTEF	78.8 ± 0.2	79.2 ± 0.2	79.1 ± 0.1	79.2 ± 1.0	77.0 ± 0.8
A-DEFEAT	$\mathbf{80.7 \pm 0.1}$	$\mathbf{82.5 \pm 0.1}$	$\mathbf{82.6 \pm 0.1}$	$\mathbf{82.2 \pm 0.1}$	$\mathbf{81.9 \pm 0.1}$
A-DEFEAT-GT	80.7 ± 0.2	81.2 ± 0.2	81.4 ± 0.2	80.9 ± 0.1	80.9 ± 0.1

Table 11. Test accuracy $\%$ comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.4 ± 0.1	79.6 ± 0.2	79.3 ± 0.3	79.3 ± 0.2	79.0 ± 0.2
CHOCO	83.7 ± 0.2	83.8 ± 0.1	83.6 ± 0.1	83.7 ± 0.1	83.5 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.6 ± 0.1	83.2 ± 0.1	83.3 ± 0.0	83.0 ± 0.1
BEER	81.8 ± 0.1	81.9 ± 0.1	81.8 ± 0.1	81.9 ± 0.2	81.7 ± 0.1
MoTEF	82.2 ± 0.1	82.3 ± 0.0	81.9 ± 0.0	82.2 ± 0.2	82.2 ± 0.1
A-DEFEAT	$\mathbf{84.0 \pm 0.2}$	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.0 \pm 0.1}$	$\mathbf{84.0 \pm 0.2}$	$\mathbf{83.9 \pm 0.2}$
A-DEFEAT-GT	82.4 ± 0.1	82.4 ± 0.1	82.4 ± 0.1	82.4 ± 0.2	82.3 ± 0.2

I.1.3 CYCLIC-4 WITH FIXED AGGREGATIONS

Table 12. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.4 ± 0.4	69.5 ± 0.5	69.4 ± 0.4	69.5 ± 0.4	69.5 ± 0.4
CHOCO	69.7 ± 0.1	68.4 ± 0.7	69.7 ± 0.1	68.5 ± 0.7	68.6 ± 0.6
DeepSqueeze	71.0 ± 0.2	72.0 ± 0.9	73.4 ± 0.4	73.1 ± 0.2	73.7 ± 0.6
BEER	78.5 ± 0.1	78.9 ± 0.7	79.0 ± 0.5	78.5 ± 0.2	78.5 ± 0.2
MoTEF	81.1 ± 0.2	80.5 ± 0.5	80.5 ± 0.6	80.8 ± 0.5	80.9 ± 0.2
A-DEFEAT	77.2 ± 0.8	78.7 ± 0.6	79.4 ± 0.5	79.5 ± 0.7	79.7 ± 0.2
A-DEFEAT-GT	$\mathbf{81.8 \pm 0.3}$	$\mathbf{81.9 \pm 0.0}$	$\mathbf{82.1 \pm 0.2}$	$\mathbf{81.9 \pm 0.2}$	$\mathbf{81.7 \pm 0.2}$

Table 13. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	70.9 ± 0.3	69.9 ± 0.3	65.7 ± 0.8	68.6 ± 0.4	68.4 ± 0.4
CHOCO	74.8 ± 0.6	76.5 ± 0.1	72.0 ± 0.9	76.2 ± 1.1	76.3 ± 0.3
DeepSqueeze	73.7 ± 1.0	77.1 ± 0.3	72.0 ± 0.5	76.9 ± 0.7	77.7 ± 0.3
BEER	80.1 ± 0.1	80.5 ± 0.6	78.0 ± 0.5	80.0 ± 0.5	79.6 ± 0.5
MoTEF	81.4 ± 0.1	80.7 ± 1.6	78.7 ± 1.2	79.9 ± 0.4	78.7 ± 0.9
A-DEFEAT	79.6 ± 0.3	80.6 ± 0.2	79.1 ± 0.2	80.9 ± 0.1	81.1 ± 0.2
A-DEFEAT-GT	$\mathbf{82.2 \pm 0.2}$	$\mathbf{82.8 \pm 0.0}$	$\mathbf{82.1 \pm 0.2}$	$\mathbf{82.6 \pm 0.1}$	$\mathbf{82.6 \pm 0.1}$

Table 14. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	71.1 ± 0.7	70.7 ± 0.1	71.4 ± 0.2	71.5 ± 0.2	72.1 ± 0.5
CHOCO	79.5 ± 0.1	81.4 ± 0.1	81.5 ± 0.1	80.9 ± 0.3	80.6 ± 0.4
DeepSqueeze	78.4 ± 0.3	80.9 ± 0.1	80.9 ± 0.0	80.5 ± 0.2	80.2 ± 0.4
BEER	81.3 ± 0.2	82.1 ± 0.6	82.4 ± 0.3	82.3 ± 0.2	82.0 ± 0.7
MoTEF	82.6 ± 0.5	81.7 ± 0.2	82.1 ± 0.1	81.7 ± 0.7	82.1 ± 0.5
A-DEFEAT	81.7 ± 0.1	82.8 ± 0.1	82.9 ± 0.1	82.8 ± 0.0	82.5 ± 0.1
A-DEFEAT-GT	$\mathbf{83.2 \pm 0.5}$	$\mathbf{83.6 \pm 0.2}$	$\mathbf{83.7 \pm 0.1}$	$\mathbf{83.5 \pm 0.1}$	$\mathbf{83.3 \pm 0.2}$

Table 15. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.5 ± 0.2	79.6 ± 0.2	79.3 ± 0.1	79.4 ± 0.2	79.2 ± 0.1
CHOCO	83.7 ± 0.2	83.9 ± 0.1	83.7 ± 0.1	83.8 ± 0.1	83.7 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.6 ± 0.1	83.4 ± 0.1	83.5 ± 0.1	83.3 ± 0.1
BEER	83.9 ± 0.1	84.0 ± 0.2	83.9 ± 0.2	83.9 ± 0.2	83.9 ± 0.2
MoTEF	83.8 ± 0.1	83.9 ± 0.1	83.8 ± 0.2	83.9 ± 0.2	83.8 ± 0.3
A-DEFEAT	84.0 ± 0.1	84.1 ± 0.2	84.0 ± 0.2	84.1 ± 0.2	84.0 ± 0.1
A-DEFEAT-GT	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.2 \pm 0.2}$	$\mathbf{84.2 \pm 0.1}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.1 \pm 0.1}$

I.1.4 CYCLIC-4 WITH FIXED COMMUNICATIONS

Table 16. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.4 ± 0.4	69.5 ± 0.5	69.4 ± 0.4	69.5 ± 0.4	69.5 ± 0.4
CHOCO	69.7 ± 0.1	68.4 ± 0.7	69.7 ± 0.1	68.5 ± 0.7	68.6 ± 0.6
DeepSqueeze	71.0 ± 0.2	72.0 ± 0.9	73.4 ± 0.4	73.1 ± 0.2	73.7 ± 0.6
BEER	75.1 ± 0.6	75.2 ± 0.3	75.2 ± 0.1	73.9 ± 0.1	74.0 ± 0.2
MoTEF	78.2 ± 0.4	78.1 ± 0.2	77.9 ± 0.3	78.0 ± 0.3	78.4 ± 0.2
A-DEFEAT	77.2 ± 0.8	78.7 ± 0.6	79.4 ± 0.5	79.5 ± 0.7	$\mathbf{79.7 \pm 0.2}$
A-DEFEAT-GT	$\mathbf{79.6 \pm 0.2}$	$\mathbf{79.5 \pm 0.2}$	$\mathbf{79.9 \pm 0.2}$	$\mathbf{80.0 \pm 0.2}$	79.7 ± 0.4

Table 17. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	70.9 ± 0.3	69.9 ± 0.3	65.7 ± 0.8	68.6 ± 0.4	68.4 ± 0.4
CHOCO	74.8 ± 0.6	76.5 ± 0.1	72.0 ± 0.9	76.2 ± 1.1	76.3 ± 0.3
DeepSqueeze	73.7 ± 1.0	77.1 ± 0.3	72.0 ± 0.5	76.9 ± 0.7	77.7 ± 0.3
BEER	77.3 ± 0.3	78.2 ± 0.3	72.4 ± 0.3	76.2 ± 1.0	76.5 ± 0.5
MoTEF	78.6 ± 0.2	79.3 ± 0.3	75.0 ± 1.2	76.0 ± 0.1	73.7 ± 1.4
A-DEFEAT	79.6 ± 0.3	80.6 ± 0.2	79.1 ± 0.2	$\mathbf{80.9 \pm 0.1}$	$\mathbf{81.1 \pm 0.2}$
A-DEFEAT-GT	$\mathbf{80.2 \pm 0.1}$	$\mathbf{80.9 \pm 0.2}$	$\mathbf{80.2 \pm 0.1}$	80.5 ± 0.1	80.3 ± 0.3

Table 18. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	71.1 ± 0.7	70.7 ± 0.1	71.4 ± 0.2	71.5 ± 0.2	72.1 ± 0.5
CHOCO	79.5 ± 0.1	81.4 ± 0.1	81.5 ± 0.1	80.9 ± 0.3	80.6 ± 0.4
DeepSqueeze	78.4 ± 0.3	80.9 ± 0.1	80.9 ± 0.0	80.5 ± 0.2	80.2 ± 0.4
BEER	79.4 ± 0.1	79.8 ± 0.3	80.0 ± 0.2	79.6 ± 0.1	80.0 ± 0.5
MoTEF	80.3 ± 0.3	81.0 ± 0.2	79.7 ± 0.1	80.4 ± 0.3	80.0 ± 0.6
A-DEFEAT	$\mathbf{81.7 \pm 0.1}$	$\mathbf{82.8 \pm 0.1}$	$\mathbf{82.9 \pm 0.1}$	$\mathbf{82.8 \pm 0.0}$	$\mathbf{82.5 \pm 0.1}$
A-DEFEAT-GT	81.5 ± 0.2	81.5 ± 0.1	81.8 ± 0.0	81.5 ± 0.0	81.6 ± 0.0

Table 19. Test accuracy % comparison. The experiments use a cyclic-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.5 ± 0.2	79.6 ± 0.2	79.3 ± 0.1	79.4 ± 0.2	79.2 ± 0.1
CHOCO	83.7 ± 0.2	83.9 ± 0.1	83.7 ± 0.1	83.8 ± 0.1	83.7 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.6 ± 0.1	83.4 ± 0.1	83.5 ± 0.1	83.3 ± 0.1
BEER	82.0 ± 0.2	82.0 ± 0.1	82.0 ± 0.1	82.1 ± 0.1	81.9 ± 0.1
MoTEF	82.2 ± 0.2	82.2 ± 0.1	82.2 ± 0.1	82.3 ± 0.1	82.2 ± 0.1
A-DEFEAT	$\mathbf{84.0 \pm 0.1}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.0 \pm 0.2}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.0 \pm 0.1}$
A-DEFEAT-GT	82.3 ± 0.1	82.5 ± 0.1	82.4 ± 0.1	82.5 ± 0.2	82.4 ± 0.1

I.1.5 RANDOM-4 WITH FIXED AGGREGATIONS

Table 20. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.4 ± 1.0	69.6 ± 0.6	69.1 ± 0.6	68.7 ± 0.6	68.7 ± 0.7
CHOCO	71.4 ± 0.6	69.6 ± 0.8	71.6 ± 0.5	70.9 ± 0.2	71.6 ± 0.2
DeepSqueeze	72.0 ± 0.6	68.1 ± 1.9	73.0 ± 0.5	73.4 ± 0.3	74.2 ± 0.4
BEER	79.0 ± 0.5	79.5 ± 0.1	80.2 ± 0.1	80.1 ± 0.1	80.2 ± 0.1
MoTEF	$\mathbf{81.2 \pm 0.2}$	81.1 ± 0.2	80.7 ± 0.2	80.9 ± 0.1	81.0 ± 0.2
A-DEFEAT	79.2 ± 0.3	78.3 ± 1.2	78.9 ± 0.3	78.5 ± 0.4	78.9 ± 0.1
A-DEFEAT-GT	81.1 ± 0.7	$\mathbf{81.7 \pm 0.0}$	$\mathbf{81.6 \pm 0.0}$	$\mathbf{82.1 \pm 0.1}$	$\mathbf{82.1 \pm 0.2}$

Table 21. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	71.4 ± 0.2	72.3 ± 0.5	70.6 ± 0.2	69.7 ± 0.4	70.2 ± 0.1
CHOCO	75.4 ± 0.6	78.4 ± 0.2	78.4 ± 0.2	79.1 ± 0.1	79.6 ± 0.1
DeepSqueeze	73.1 ± 1.2	78.2 ± 0.1	78.2 ± 0.1	78.9 ± 0.2	79.5 ± 0.1
BEER	79.1 ± 0.1	80.9 ± 0.1	81.0 ± 0.3	81.0 ± 0.2	81.4 ± 0.4
MoTEF	81.7 ± 0.1	81.4 ± 0.5	81.5 ± 0.1	81.4 ± 0.4	82.6 ± 0.2
A-DEFEAT	80.1 ± 0.5	81.2 ± 0.2	81.5 ± 0.2	82.1 ± 0.0	82.4 ± 0.1
A-DEFEAT-GT	$\mathbf{82.4 \pm 0.0}$	$\mathbf{82.6 \pm 0.1}$	$\mathbf{83.2 \pm 0.0}$	$\mathbf{83.4 \pm 0.1}$	$\mathbf{83.6 \pm 0.0}$

Table 22. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	72.5 ± 0.7	71.2 ± 0.2	71.5 ± 0.1	71.2 ± 0.2	71.3 ± 0.1
CHOCO	79.4 ± 0.1	81.4 ± 0.2	81.9 ± 0.0	81.9 ± 0.2	81.9 ± 0.2
DeepSqueeze	78.6 ± 0.4	81.1 ± 0.2	81.5 ± 0.1	81.4 ± 0.1	81.3 ± 0.2
BEER	81.6 ± 0.1	81.3 ± 0.2	82.7 ± 0.3	82.9 ± 0.3	83.0 ± 0.3
MoTEF	82.9 ± 0.2	82.4 ± 0.3	83.2 ± 0.3	83.1 ± 0.2	83.1 ± 0.2
A-DEFEAT	81.8 ± 0.2	82.8 ± 0.1	83.1 ± 0.1	83.2 ± 0.1	83.2 ± 0.1
A-DEFEAT-GT	$\mathbf{83.3 \pm 0.2}$	$\mathbf{83.5 \pm 0.1}$	$\mathbf{84.0 \pm 0.1}$	$\mathbf{84.0 \pm 0.1}$	$\mathbf{84.0 \pm 0.1}$

Table 23. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.2 ± 0.3	79.0 ± 0.3	78.6 ± 0.2	78.2 ± 0.1	77.8 ± 0.1
CHOCO	83.7 ± 0.2	83.8 ± 0.1	83.8 ± 0.1	83.8 ± 0.1	83.9 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.7 ± 0.1	83.6 ± 0.1	83.7 ± 0.1	83.5 ± 0.1
BEER	83.8 ± 0.2	83.9 ± 0.1	84.0 ± 0.2	84.0 ± 0.2	84.0 ± 0.2
MoTEF	83.8 ± 0.1	84.0 ± 0.1	83.9 ± 0.3	84.0 ± 0.2	84.0 ± 0.2
A-DEFEAT	84.0 ± 0.1	84.1 ± 0.2	84.1 ± 0.2	84.2 ± 0.2	84.2 ± 0.1
A-DEFEAT-GT	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.2 \pm 0.1}$	$\mathbf{84.3 \pm 0.1}$	$\mathbf{84.2 \pm 0.2}$	$\mathbf{84.3 \pm 0.2}$

I.1.6 RANDOM-4 WITH FIXED COMMUNICATIONS

Table 24. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.4 ± 1.0	69.6 ± 0.6	69.1 ± 0.6	68.7 ± 0.6	68.7 ± 0.7
CHOCO	71.4 ± 0.6	69.6 ± 0.8	71.6 ± 0.5	70.9 ± 0.2	71.6 ± 0.2
DeepSqueeze	72.0 ± 0.6	68.1 ± 1.9	73.0 ± 0.5	73.4 ± 0.3	74.2 ± 0.4
BEER	76.8 ± 1.0	76.2 ± 0.3	77.3 ± 0.2	77.9 ± 0.1	77.9 ± 0.1
MoTEF	78.3 ± 0.2	78.2 ± 0.4	77.7 ± 0.2	77.9 ± 0.4	78.5 ± 0.2
A-DEFEAT	79.2 ± 0.3	78.3 ± 1.2	78.9 ± 0.3	78.5 ± 0.4	78.9 ± 0.1
A-DEFEAT-GT	$\mathbf{79.4 \pm 0.2}$	$\mathbf{80.3 \pm 0.0}$	$\mathbf{80.2 \pm 0.1}$	$\mathbf{79.4 \pm 0.5}$	$\mathbf{80.6 \pm 0.0}$

Table 25. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	71.4 ± 0.2	72.3 ± 0.5	70.6 ± 0.2	69.7 ± 0.4	70.2 ± 0.1
CHOCO	75.4 ± 0.6	78.4 ± 0.2	78.4 ± 0.2	79.1 ± 0.1	79.6 ± 0.1
DeepSqueeze	73.1 ± 1.2	78.2 ± 0.1	78.2 ± 0.1	78.9 ± 0.2	79.5 ± 0.1
BEER	77.3 ± 0.1	78.9 ± 0.2	78.9 ± 0.3	78.4 ± 0.1	79.2 ± 0.1
MoTEF	79.3 ± 0.2	79.1 ± 0.3	79.9 ± 0.1	80.1 ± 0.2	80.1 ± 0.1
A-DEFEAT	80.1 ± 0.5	$\mathbf{81.2 \pm 0.2}$	$\mathbf{81.5 \pm 0.2}$	$\mathbf{82.1 \pm 0.0}$	$\mathbf{82.4 \pm 0.1}$
A-DEFEAT-GT	$\mathbf{80.6 \pm 0.4}$	81.0 ± 0.3	81.2 ± 0.1	81.5 ± 0.1	81.8 ± 0.1

Table 26. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	72.5 ± 0.7	71.2 ± 0.2	71.5 ± 0.1	71.2 ± 0.2	71.3 ± 0.1
CHOCO	79.4 ± 0.1	81.4 ± 0.2	81.9 ± 0.0	81.9 ± 0.2	81.9 ± 0.2
DeepSqueeze	78.6 ± 0.4	81.1 ± 0.2	81.5 ± 0.1	81.4 ± 0.1	81.3 ± 0.2
BEER	79.2 ± 0.2	79.0 ± 0.1	80.5 ± 0.2	80.7 ± 0.3	80.6 ± 0.3
MoTEF	80.6 ± 0.2	80.5 ± 0.2	81.1 ± 0.3	81.0 ± 0.2	81.1 ± 0.2
A-DEFEAT	$\mathbf{81.8 \pm 0.2}$	$\mathbf{82.8 \pm 0.1}$	$\mathbf{83.1 \pm 0.1}$	$\mathbf{83.2 \pm 0.1}$	$\mathbf{83.2 \pm 0.1}$
A-DEFEAT-GT	81.5 ± 0.1	81.7 ± 0.0	82.2 ± 0.1	82.1 ± 0.2	82.3 ± 0.1

Table 27. Test accuracy % comparison. The experiments use a random-4 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.2 ± 0.3	79.0 ± 0.3	78.6 ± 0.2	78.2 ± 0.1	77.8 ± 0.1
CHOCO	83.7 ± 0.2	83.8 ± 0.1	83.8 ± 0.1	83.8 ± 0.1	83.9 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.7 ± 0.1	83.6 ± 0.1	83.7 ± 0.1	83.5 ± 0.1
BEER	81.9 ± 0.2	82.0 ± 0.1	82.0 ± 0.2	82.1 ± 0.1	82.1 ± 0.1
MoTEF	82.3 ± 0.1	82.3 ± 0.1	82.3 ± 0.1	82.4 ± 0.1	82.3 ± 0.1
A-DEFEAT	$\mathbf{84.0 \pm 0.1}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.2 \pm 0.2}$	$\mathbf{84.2 \pm 0.1}$
A-DEFEAT-GT	82.4 ± 0.1	82.3 ± 0.2	82.5 ± 0.2	82.6 ± 0.1	82.6 ± 0.1

I.1.7 RANDOM-9 WITH FIXED AGGREGATIONS

Table 28. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.8 ± 0.4	69.8 ± 0.6	69.8 ± 0.5	69.4 ± 0.5	69.5 ± 0.4
CHOCO	73.1 ± 0.1	71.7 ± 0.1	72.4 ± 0.2	71.7 ± 0.1	72.2 ± 0.3
DeepSqueeze	71.1 ± 0.8	72.9 ± 0.1	73.3 ± 0.5	73.6 ± 0.3	73.6 ± 0.4
BEER	79.0 ± 0.5	79.5 ± 0.4	80.0 ± 0.1	80.5 ± 0.2	79.9 ± 0.1
MoTEF	81.6 ± 0.2	81.5 ± 0.2	81.5 ± 0.2	81.4 ± 0.2	81.0 ± 0.4
A-DEFEAT	78.6 ± 0.1	78.1 ± 0.7	77.9 ± 0.6	78.5 ± 0.5	78.8 ± 0.1
A-DEFEAT-GT	$\mathbf{82.1 \pm 0.2}$	$\mathbf{82.1 \pm 0.1}$	$\mathbf{82.0 \pm 0.3}$	$\mathbf{82.0 \pm 0.2}$	$\mathbf{81.9 \pm 0.2}$

Table 29. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	71.2 ± 0.4	72.3 ± 0.4	70.6 ± 0.3	70.4 ± 0.3	70.7 ± 0.3
CHOCO	75.8 ± 0.5	78.5 ± 0.0	79.0 ± 0.3	79.6 ± 0.2	80.1 ± 0.0
DeepSqueeze	75.9 ± 0.5	78.9 ± 0.1	78.9 ± 0.2	79.3 ± 0.2	79.7 ± 0.0
BEER	80.0 ± 0.2	81.2 ± 0.1	80.8 ± 0.1	81.1 ± 0.1	81.8 ± 0.2
MoTEF	81.7 ± 0.1	81.7 ± 0.1	82.2 ± 0.3	82.2 ± 0.2	$\mathbf{82.6 \pm 0.4}$
A-DEFEAT	80.9 ± 0.2	81.8 ± 0.2	81.8 ± 0.2	81.9 ± 0.1	82.3 ± 0.2
A-DEFEAT-GT	$\mathbf{82.1 \pm 0.3}$	$\mathbf{81.9 \pm 1.3}$	$\mathbf{82.7 \pm 0.1}$	$\mathbf{82.4 \pm 0.1}$	82.6 ± 0.1

Table 30. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	73.4 ± 0.3	71.4 ± 0.3	71.3 ± 0.2	71.4 ± 0.1	71.4 ± 0.1
CHOCO	79.6 ± 0.1	81.6 ± 0.2	82.0 ± 0.1	82.0 ± 0.1	82.0 ± 0.2
DeepSqueeze	78.5 ± 0.1	81.3 ± 0.1	81.6 ± 0.1	81.5 ± 0.2	81.5 ± 0.1
BEER	81.0 ± 0.2	82.4 ± 0.4	82.6 ± 0.1	82.9 ± 0.3	83.2 ± 0.2
MoTEF	83.0 ± 0.2	83.1 ± 0.2	83.3 ± 0.3	$\mathbf{83.3 \pm 0.3}$	$\mathbf{83.3 \pm 0.3}$
A-DEFEAT	82.2 ± 0.1	82.9 ± 0.1	83.1 ± 0.0	83.1 ± 0.0	83.0 ± 0.1
A-DEFEAT-GT	$\mathbf{83.6 \pm 0.1}$	$\mathbf{83.5 \pm 0.1}$	$\mathbf{83.7 \pm 0.2}$	83.1 ± 0.4	83.1 ± 0.3

Table 31. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.5 ± 0.2	78.9 ± 0.2	78.5 ± 0.2	78.0 ± 0.2	77.3 ± 0.3
CHOCO	83.6 ± 0.2	83.9 ± 0.1	83.8 ± 0.1	83.9 ± 0.2	83.9 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.7 ± 0.1	83.7 ± 0.1	83.7 ± 0.1	83.7 ± 0.1
BEER	83.9 ± 0.2	84.0 ± 0.2	83.9 ± 0.2	84.0 ± 0.2	84.1 ± 0.2
MoTEF	83.8 ± 0.1	83.9 ± 0.1	83.9 ± 0.2	84.0 ± 0.2	83.9 ± 0.2
A-DEFEAT	84.0 ± 0.1	84.1 ± 0.1	84.1 ± 0.1	84.1 ± 0.1	84.1 ± 0.1
A-DEFEAT-GT	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.2 \pm 0.2}$	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.2 \pm 0.1}$

I.1.8 RANDOM-9 WITH FIXED COMMUNICATIONS

Table 32. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.8 ± 0.4	69.8 ± 0.6	69.8 ± 0.5	69.4 ± 0.5	69.5 ± 0.4
CHOCO	73.1 ± 0.1	71.7 ± 0.1	72.4 ± 0.2	71.7 ± 0.1	72.2 ± 0.3
DeepSqueeze	71.1 ± 0.8	72.9 ± 0.1	73.3 ± 0.5	73.6 ± 0.3	73.6 ± 0.4
BEER	75.2 ± 0.4	76.6 ± 0.5	77.8 ± 0.1	77.6 ± 0.4	77.5 ± 0.0
MoTEF	79.1 ± 0.2	78.5 ± 0.3	78.9 ± 0.3	78.6 ± 0.2	78.5 ± 0.2
A-DEFEAT	78.6 ± 0.1	78.1 ± 0.7	77.9 ± 0.6	78.5 ± 0.5	78.8 ± 0.1
A-DEFEAT-GT	$\mathbf{79.7 \pm 0.1}$	$\mathbf{79.7 \pm 0.2}$	$\mathbf{79.6 \pm 0.0}$	$\mathbf{78.8 \pm 0.4}$	$\mathbf{79.7 \pm 0.1}$

Table 33. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	71.2 ± 0.4	72.3 ± 0.4	70.6 ± 0.3	70.4 ± 0.3	70.7 ± 0.3
CHOCO	75.8 ± 0.5	78.5 ± 0.0	79.0 ± 0.3	79.6 ± 0.2	80.1 ± 0.0
DeepSqueeze	75.9 ± 0.5	78.9 ± 0.1	78.9 ± 0.2	79.3 ± 0.2	79.7 ± 0.0
BEER	77.1 ± 0.1	79.3 ± 0.2	78.4 ± 0.1	79.1 ± 0.1	79.9 ± 0.2
MoTEF	79.4 ± 0.1	79.4 ± 0.2	79.7 ± 0.3	79.9 ± 0.1	79.7 ± 0.5
A-DEFEAT	$\mathbf{80.9 \pm 0.2}$	$\mathbf{81.8 \pm 0.2}$	$\mathbf{81.8 \pm 0.2}$	$\mathbf{81.9 \pm 0.1}$	$\mathbf{82.3 \pm 0.2}$
A-DEFEAT-GT	80.1 ± 0.2	79.3 ± 1.8	80.3 ± 0.1	79.8 ± 0.1	80.2 ± 0.1

Table 34. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	73.4 ± 0.3	71.4 ± 0.3	71.3 ± 0.2	71.4 ± 0.1	71.4 ± 0.1
CHOCO	79.6 ± 0.1	81.6 ± 0.2	82.0 ± 0.1	82.0 ± 0.1	82.0 ± 0.2
DeepSqueeze	78.5 ± 0.1	81.3 ± 0.1	81.6 ± 0.1	81.5 ± 0.2	81.5 ± 0.1
BEER	79.6 ± 0.2	80.1 ± 0.2	80.4 ± 0.2	80.9 ± 0.3	81.0 ± 0.3
MoTEF	80.6 ± 0.3	80.7 ± 0.3	81.1 ± 0.4	81.1 ± 0.3	81.2 ± 0.2
A-DEFEAT	$\mathbf{82.2 \pm 0.1}$	$\mathbf{82.9 \pm 0.1}$	$\mathbf{83.1 \pm 0.0}$	$\mathbf{83.1 \pm 0.0}$	$\mathbf{83.0 \pm 0.1}$
A-DEFEAT-GT	81.5 ± 0.1	81.1 ± 0.4	81.0 ± 0.2	80.9 ± 0.3	80.7 ± 0.4

Table 35. Test accuracy % comparison. The experiments use a random-9 network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	79.5 ± 0.2	78.9 ± 0.2	78.5 ± 0.2	78.0 ± 0.2	77.3 ± 0.3
CHOCO	83.6 ± 0.2	83.9 ± 0.1	83.8 ± 0.1	83.9 ± 0.2	83.9 ± 0.1
DeepSqueeze	83.5 ± 0.1	83.7 ± 0.1	83.7 ± 0.1	83.7 ± 0.1	83.7 ± 0.1
BEER	82.0 ± 0.2	82.0 ± 0.1	82.0 ± 0.2	82.2 ± 0.1	82.2 ± 0.1
MoTEF	82.3 ± 0.1	82.3 ± 0.1	82.3 ± 0.1	82.4 ± 0.1	82.3 ± 0.1
A-DEFEAT	$\mathbf{84.0 \pm 0.1}$	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.1 \pm 0.1}$	$\mathbf{84.1 \pm 0.1}$
A-DEFEAT-GT	82.3 ± 0.1	82.3 ± 0.2	82.2 ± 0.3	82.4 ± 0.3	82.4 ± 0.3

I.1.9 RING ($k = 0.01$) WITH FIXED AGGREGATIONS

Table 36. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.01$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	57.4 ± 2.0	57.4 ± 2.0	57.4 ± 1.9	57.4 ± 1.9	57.4 ± 1.9
CHOCO	52.8 ± 4.1	52.8 ± 4.1	52.8 ± 4.2	52.8 ± 4.2	52.8 ± 4.2
DeepSqueeze	47.1 ± 0.3	48.2 ± 3.3	48.2 ± 2.4	49.2 ± 4.7	49.2 ± 4.4
BEER	69.8 ± 1.1	69.7 ± 1.1	69.3 ± 1.0	70.3 ± 0.6	70.2 ± 0.6
MoTEF	67.5 ± 0.4	67.4 ± 0.9	66.5 ± 1.7	68.4 ± 0.7	67.0 ± 1.4
A-DEFEAT	69.3 ± 0.9	69.8 ± 0.8	69.2 ± 1.3	68.8 ± 0.3	68.7 ± 1.0
A-DEFEAT-GT	$\mathbf{77.4 \pm 0.6}$	$\mathbf{77.3 \pm 0.4}$	$\mathbf{77.2 \pm 0.7}$	$\mathbf{77.0 \pm 0.6}$	$\mathbf{77.3 \pm 0.5}$

Table 37. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.01$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	59.3 ± 1.1	57.1 ± 1.3	56.7 ± 0.4	55.1 ± 1.8	55.5 ± 2.8
CHOCO	70.6 ± 0.5	69.9 ± 3.1	64.0 ± 0.7	68.9 ± 2.6	72.9 ± 0.2
DeepSqueeze	62.1 ± 1.8	65.9 ± 4.0	62.6 ± 0.4	69.3 ± 3.6	72.8 ± 0.3
BEER	73.4 ± 1.0	72.0 ± 0.5	64.5 ± 0.4	68.3 ± 1.7	73.5 ± 1.4
MoTEF	68.3 ± 3.5	73.4 ± 2.5	62.3 ± 0.9	65.1 ± 3.0	67.3 ± 2.8
A-DEFEAT	75.5 ± 0.3	75.8 ± 0.4	67.6 ± 0.9	70.2 ± 1.6	73.2 ± 0.8
A-DEFEAT-GT	$\mathbf{77.4 \pm 0.3}$	$\mathbf{78.9 \pm 0.2}$	$\mathbf{73.1 \pm 1.0}$	$\mathbf{75.3 \pm 1.6}$	$\mathbf{75.5 \pm 0.5}$

Table 38. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.01$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	64.7 ± 0.8	65.5 ± 0.6	58.9 ± 1.4	62.0 ± 0.8	62.8 ± 1.3
CHOCO	74.3 ± 0.2	79.3 ± 0.3	79.3 ± 0.2	79.5 ± 0.6	78.8 ± 0.5
DeepSqueeze	74.0 ± 0.0	78.2 ± 0.4	78.2 ± 0.0	78.6 ± 0.2	78.2 ± 0.5
BEER	75.5 ± 0.5	79.2 ± 0.3	78.9 ± 0.3	79.8 ± 0.6	79.4 ± 0.5
MoTEF	75.9 ± 0.2	78.8 ± 0.3	72.4 ± 4.7	69.7 ± 3.6	74.9 ± 2.6
A-DEFEAT	78.6 ± 0.4	79.9 ± 0.2	79.5 ± 0.1	78.8 ± 0.5	77.7 ± 0.3
A-DEFEAT-GT	$\mathbf{79.6 \pm 0.5}$	$\mathbf{80.9 \pm 0.1}$	$\mathbf{80.4 \pm 0.2}$	$\mathbf{80.2 \pm 0.1}$	$\mathbf{79.6 \pm 0.1}$

I.1.10 RING ($k = 0.01$) WITH FIXED COMMUNICATIONS

Table 39. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.01$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	57.4 ± 2.0	57.4 ± 2.0	57.4 ± 1.9	57.4 ± 1.9	57.4 ± 1.9
CHOCO	52.8 ± 4.1	52.8 ± 4.1	52.8 ± 4.2	52.8 ± 4.2	52.8 ± 4.2
DeepSqueeze	47.1 ± 0.3	48.2 ± 3.3	48.2 ± 2.4	49.2 ± 4.7	49.2 ± 4.4
BEER	62.0 ± 1.7	62.1 ± 1.5	62.0 ± 1.4	62.6 ± 1.6	62.4 ± 1.5
MoTEF	58.2 ± 3.2	58.3 ± 3.4	58.2 ± 3.5	59.1 ± 2.9	58.5 ± 4.3
A-DEFEAT	69.3 ± 0.9	69.8 ± 0.8	69.2 ± 1.3	68.8 ± 0.3	68.7 ± 1.0
A-DEFEAT-GT	$\mathbf{70.4 \pm 0.6}$	$\mathbf{70.5 \pm 0.7}$	$\mathbf{70.6 \pm 0.7}$	$\mathbf{70.0 \pm 1.0}$	$\mathbf{70.4 \pm 0.8}$

Table 40. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.01$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	59.3 ± 1.1	57.1 ± 1.3	56.7 ± 0.4	55.1 ± 1.8	55.5 ± 2.8
CHOCO	70.6 ± 0.5	69.9 ± 3.1	64.0 ± 0.7	68.9 ± 2.6	72.9 ± 0.2
DeepSqueeze	62.1 ± 1.8	65.9 ± 4.0	62.6 ± 0.4	69.3 ± 3.6	72.8 ± 0.3
BEER	66.8 ± 0.4	61.3 ± 2.0	62.3 ± 0.3	67.0 ± 2.2	71.6 ± 0.8
MoTEF	64.5 ± 1.8	48.9 ± 4.8	57.4 ± 2.3	61.3 ± 2.2	59.3 ± 5.1
A-DEFEAT	$\mathbf{75.5 \pm 0.3}$	$\mathbf{75.8 \pm 0.4}$	$\mathbf{67.6 \pm 0.9}$	$\mathbf{70.2 \pm 1.6}$	$\mathbf{73.2 \pm 0.8}$
A-DEFEAT-GT	71.6 ± 0.7	73.4 ± 0.6	65.5 ± 1.7	67.8 ± 2.2	70.9 ± 1.2

I.1.11 RING ($k = 0.03$) WITH FIXED AGGREGATIONS

Table 41. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	63.3 ± 0.8	63.3 ± 0.7	63.3 ± 0.8	63.3 ± 0.8	63.3 ± 0.8
CHOCO	61.6 ± 1.7	61.6 ± 1.6	55.6 ± 6.3	58.4 ± 3.5	57.6 ± 4.9
DeepSqueeze	63.3 ± 3.0	63.2 ± 3.6	65.1 ± 3.3	66.0 ± 1.9	66.5 ± 1.6
BEER	77.2 ± 0.2	77.1 ± 0.1	77.3 ± 0.1	77.2 ± 0.1	77.2 ± 0.0
MoTEF	77.8 ± 0.0	75.2 ± 3.2	77.4 ± 0.1	77.6 ± 0.2	77.2 ± 0.6
A-DEFEAT	75.5 ± 1.0	74.4 ± 0.7	73.8 ± 0.4	73.9 ± 0.3	75.2 ± 0.4
A-DEFEAT-GT	$\mathbf{80.0 \pm 0.3}$	$\mathbf{80.3 \pm 0.4}$	$\mathbf{80.3 \pm 0.2}$	$\mathbf{79.8 \pm 0.4}$	$\mathbf{79.9 \pm 0.4}$

Table 42. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	67.2 ± 0.1	64.3 ± 0.7	60.4 ± 0.4	61.0 ± 2.3	63.4 ± 1.3
CHOCO	72.6 ± 0.4	74.7 ± 0.5	67.1 ± 0.6	71.3 ± 2.3	74.3 ± 0.4
DeepSqueeze	69.1 ± 1.0	75.2 ± 0.4	65.6 ± 0.7	72.5 ± 2.0	75.4 ± 0.3
BEER	78.7 ± 0.1	78.6 ± 0.4	71.2 ± 0.3	74.1 ± 1.3	75.1 ± 1.3
MoTEF	79.6 ± 0.2	78.6 ± 0.7	70.3 ± 0.5	72.0 ± 0.7	71.6 ± 1.5
A-DEFEAT	77.5 ± 0.5	79.7 ± 0.2	75.4 ± 0.5	78.8 ± 0.3	78.3 ± 0.1
A-DEFEAT-GT	$\mathbf{81.8 \pm 0.1}$	$\mathbf{81.6 \pm 0.1}$	$\mathbf{79.9 \pm 0.3}$	$\mathbf{81.0 \pm 0.4}$	$\mathbf{80.1 \pm 0.3}$

Table 43. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.9 ± 0.8	70.8 ± 0.3	67.3 ± 1.6	69.8 ± 0.4	70.2 ± 0.7
CHOCO	77.6 ± 0.3	80.7 ± 0.3	80.6 ± 0.1	80.3 ± 0.2	79.7 ± 0.4
DeepSqueeze	75.8 ± 0.3	79.9 ± 0.3	79.5 ± 0.0	79.5 ± 0.1	79.2 ± 0.4
BEER	78.8 ± 0.2	80.8 ± 0.4	81.4 ± 0.3	80.7 ± 0.4	80.9 ± 0.6
MoTEF	76.5 ± 4.9	81.9 ± 0.2	80.5 ± 1.2	79.8 ± 1.3	79.2 ± 1.3
A-DEFEAT	80.6 ± 0.1	82.5 ± 0.1	82.5 ± 0.1	82.1 ± 0.2	81.9 ± 0.1
A-DEFEAT-GT	$\mathbf{82.2 \pm 0.4}$	$\mathbf{83.2 \pm 0.1}$	$\mathbf{83.1 \pm 0.2}$	$\mathbf{82.8 \pm 0.2}$	$\mathbf{82.6 \pm 0.1}$

Table 44. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 aggregation rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	77.7 ± 0.7	78.2 ± 0.1	77.9 ± 0.2	77.9 ± 0.1	77.6 ± 0.1
CHOCO	83.7 ± 0.2	83.8 ± 0.1	83.5 ± 0.1	83.6 ± 0.1	83.4 ± 0.1
DeepSqueeze	83.4 ± 0.2	83.5 ± 0.1	83.1 ± 0.1	83.3 ± 0.1	82.8 ± 0.1
BEER	83.7 ± 0.2	83.9 ± 0.1	83.8 ± 0.2	83.7 ± 0.2	83.5 ± 0.2
MoTEF	83.7 ± 0.2	83.7 ± 0.1	83.5 ± 0.2	83.5 ± 0.1	83.3 ± 0.1
A-DEFEAT	83.8 ± 0.2	84.0 ± 0.2	83.9 ± 0.1	84.0 ± 0.2	83.8 ± 0.2
A-DEFEAT-GT	$\mathbf{84.0 \pm 0.1}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.1 \pm 0.2}$	$\mathbf{84.1 \pm 0.1}$	$\mathbf{83.9 \pm 0.2}$

I.1.12 RING ($k = 0.03$) WITH FIXED COMMUNICATION

Table 45. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	63.3 ± 0.8	63.3 ± 0.7	63.3 ± 0.8	63.3 ± 0.8	63.3 ± 0.8
CHOCO	61.6 ± 1.7	61.6 ± 1.6	55.6 ± 6.3	58.4 ± 3.5	57.6 ± 4.9
DeepSqueeze	63.3 ± 3.0	63.2 ± 3.6	65.1 ± 3.3	66.0 ± 1.9	66.5 ± 1.6
BEER	71.8 ± 0.4	72.1 ± 0.7	72.0 ± 0.5	71.9 ± 0.5	71.8 ± 0.5
MoTEF	72.3 ± 0.5	71.0 ± 0.9	72.2 ± 0.5	72.5 ± 0.4	72.2 ± 0.8
A-DEFEAT	$\mathbf{75.5 \pm 1.0}$	74.4 ± 0.7	73.8 ± 0.4	73.9 ± 0.3	75.2 ± 0.4
A-DEFEAT-GT	75.1 ± 0.3	$\mathbf{76.5 \pm 0.5}$	$\mathbf{75.5 \pm 0.9}$	$\mathbf{76.0 \pm 0.2}$	$\mathbf{77.2 \pm 0.3}$

Table 46. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	67.2 ± 0.1	64.3 ± 0.7	60.4 ± 0.4	61.0 ± 2.3	63.4 ± 1.3
CHOCO	72.6 ± 0.4	74.7 ± 0.5	67.1 ± 0.6	71.3 ± 2.3	74.3 ± 0.4
DeepSqueeze	69.1 ± 1.0	75.2 ± 0.4	65.6 ± 0.7	72.5 ± 2.0	75.4 ± 0.3
BEER	74.9 ± 0.3	73.1 ± 0.6	66.1 ± 0.4	70.4 ± 2.9	73.0 ± 1.6
MoTEF	75.1 ± 0.4	71.9 ± 1.6	64.3 ± 0.7	67.4 ± 2.8	68.3 ± 1.6
A-DEFEAT	77.5 ± 0.5	$\mathbf{79.7 \pm 0.2}$	$\mathbf{75.4 \pm 0.5}$	$\mathbf{78.8 \pm 0.3}$	$\mathbf{78.3 \pm 0.1}$
A-DEFEAT-GT	$\mathbf{78.8 \pm 0.5}$	79.0 ± 0.1	73.6 ± 0.9	77.7 ± 0.8	76.5 ± 1.0

Table 47. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	69.9 ± 0.8	70.8 ± 0.3	67.3 ± 1.6	69.8 ± 0.4	70.2 ± 0.7
CHOCO	77.6 ± 0.3	80.7 ± 0.3	80.6 ± 0.1	80.3 ± 0.2	79.7 ± 0.4
DeepSqueeze	75.8 ± 0.3	79.9 ± 0.3	79.5 ± 0.0	79.5 ± 0.1	79.2 ± 0.4
BEER	74.9 ± 0.6	78.2 ± 0.5	78.5 ± 0.2	78.4 ± 0.2	78.5 ± 0.5
MoTEF	76.8 ± 0.8	78.2 ± 0.4	74.2 ± 2.5	77.4 ± 2.1	77.5 ± 0.5
A-DEFEAT	$\mathbf{80.6 \pm 0.1}$	$\mathbf{82.5 \pm 0.1}$	$\mathbf{82.5 \pm 0.1}$	$\mathbf{82.1 \pm 0.2}$	$\mathbf{81.9 \pm 0.1}$
A-DEFEAT-GT	80.1 ± 0.2	80.9 ± 0.1	81.2 ± 0.2	80.8 ± 0.1	79.8 ± 0.3

Table 48. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 5$ and FashionMNIST. All results are compared using 1000 communication rounds, each value averaged over the last 50 rounds. The compressor is top-k with $k = 0.03$.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	77.7 ± 0.7	78.2 ± 0.1	77.9 ± 0.2	77.9 ± 0.1	77.6 ± 0.1
CHOCO	83.7 ± 0.2	83.8 ± 0.1	83.5 ± 0.1	83.6 ± 0.1	83.4 ± 0.1
DeepSqueeze	83.4 ± 0.2	83.5 ± 0.1	83.1 ± 0.1	83.3 ± 0.1	82.8 ± 0.1
BEER	81.9 ± 0.2	81.8 ± 0.1	81.7 ± 0.1	81.8 ± 0.1	81.7 ± 0.1
MoTEF	82.1 ± 0.1	82.0 ± 0.1	81.7 ± 0.2	81.7 ± 0.5	81.1 ± 0.3
A-DEFEAT	$\mathbf{83.8 \pm 0.2}$	$\mathbf{84.0 \pm 0.2}$	$\mathbf{83.9 \pm 0.1}$	$\mathbf{84.0 \pm 0.2}$	$\mathbf{83.8 \pm 0.2}$
A-DEFEAT-GT	82.4 ± 0.1	82.4 ± 0.0	82.4 ± 0.1	82.4 ± 0.1	82.2 ± 0.1

I.2 KMNIST

I.2.1 Ring with fixed aggregations

Table 49. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and KMNIST. All results are compared using 5000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	44.1 ± 1.7	47.5 ± 1.2	49.0 ± 1.8	47.7 ± 2.2	47.7 ± 2.2
CHOCO	33.4 ± 4.3	32.8 ± 4.7	12.1 ± 1.4	33.3 ± 4.4	33.5 ± 4.2
DeepSqueeze	25.2 ± 1.3	33.3 ± 1.1	34.5 ± 1.1	35.2 ± 0.8	35.6 ± 1.4
BEER	38.0 ± 4.7	53.8 ± 12.2	59.9 ± 3.6	55.1 ± 2.0	60.3 ± 0.5
MoTEF	26.2 ± 3.7	11.3 ± 0.5	16.4 ± 4.5	15.3 ± 5.3	16.0 ± 2.5
A-DEFEAT	62.3 ± 5.4	65.3 ± 3.7	64.7 ± 4.6	67.1 ± 1.8	70.7 ± 0.3
A-DEFEAT-GT	$\mathbf{82.2 \pm 5.7}$	$\mathbf{80.1 \pm 3.1}$	$\mathbf{80.2 \pm 7.0}$	$\mathbf{87.2 \pm 0.4}$	$\mathbf{86.7 \pm 0.6}$

Table 50. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and KMNIST. All results are compared using 5000 aggregation rounds, each value averaged over the last 50 rounds. † marks a run whose curve stops before the reading round because a seed diverged; the table uses its last available value.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	72.6 ± 0.7	84.1 ± 0.7	80.2 ± 1.5	80.7 ± 0.5	77.2 ± 3.9
CHOCO	65.5 ± 1.3	87.6 ± 0.5	84.0 ± 0.9	83.0 ± 0.2	80.3 ± 0.5
DeepSqueeze	67.4 ± 0.5	82.9 ± 0.7	76.6 ± 0.9	76.6 ± 0.4	74.5 ± 0.3
BEER	75.6 ± 0.8	92.1 ± 0.3	72.3 ± 26.2 †	90.2 ± 0.2	89.6 ± 0.2
MoTEF	59.1 ± 6.2	88.0 ± 1.1	91.5 ± 0.3	90.5 ± 0.4	89.8 ± 0.8
A-DEFEAT	82.4 ± 2.0	92.6 ± 0.0	91.9 ± 0.1	91.6 ± 0.1	90.8 ± 0.3
A-DEFEAT-GT	$\mathbf{87.6 \pm 2.9}$	$\mathbf{94.6 \pm 0.1}$	$\mathbf{94.7 \pm 0.1}$	$\mathbf{94.6 \pm 0.1}$	$\mathbf{93.7 \pm 0.1}$

Table 51. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and KMNIST. All results are compared using 5000 aggregation rounds, each value averaged over the last 50 rounds. † marks a run whose curve stops before the reading round because a seed diverged; the table uses its last available value.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	90.1 ± 0.4	90.6 ± 0.3	90.1 ± 0.1	89.0 ± 0.3	86.0 ± 0.8
CHOCO	92.9 ± 0.4	93.9 ± 0.1	92.9 ± 0.2	91.2 ± 0.6	88.1 ± 0.6
DeepSqueeze	92.0 ± 0.1	92.8 ± 0.1	91.1 ± 0.1	88.2 ± 0.4	85.7 ± 0.4
BEER	94.7 ± 0.3	95.3 ± 0.1	95.0 ± 0.2	85.1 ± 0.5 †	92.6 ± 0.4
MoTEF	10.0 ± 0.0 †	95.4 ± 0.3	95.5 ± 0.1	94.7 ± 0.1	92.3 ± 0.2
A-DEFEAT	94.7 ± 0.1	94.9 ± 0.0	94.7 ± 0.1	94.2 ± 0.1	92.9 ± 0.1
A-DEFEAT-GT	$\mathbf{95.6 \pm 0.2}$	$\mathbf{95.9 \pm 0.0}$	$\mathbf{95.7 \pm 0.1}$	$\mathbf{95.4 \pm 0.1}$	$\mathbf{94.5 \pm 0.1}$

I.2.2 RING WITH FIXED COMMUNICATIONS

Table 52. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and KMNIST. All results are compared using 5000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	44.1 ± 1.7	47.5 ± 1.2	49.0 ± 1.8	47.7 ± 2.2	47.7 ± 2.2
CHOCO	33.4 ± 4.3	32.8 ± 4.7	12.1 ± 1.4	33.3 ± 4.4	33.5 ± 4.2
DeepSqueeze	25.2 ± 1.3	33.3 ± 1.1	34.5 ± 1.1	35.2 ± 0.8	35.6 ± 1.4
BEER	34.8 ± 3.7	38.7 ± 9.4	46.1 ± 10.9	56.0 ± 1.7	53.2 ± 1.3
MoTEF	22.2 ± 1.3	12.7 ± 1.8	29.7 ± 1.0	23.0 ± 3.9	20.9 ± 3.6
A-DEFEAT	62.3 ± 5.4	65.3 ± 3.7	64.7 ± 4.6	67.1 ± 1.8	70.7 ± 0.3
A-DEFEAT-GT	**77.1 ± 2.6**	**71.1 ± 5.5**	**75.0 ± 3.2**	**79.4 ± 0.3**	**81.4 ± 0.6**

Table 53. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and KMNIST. All results are compared using 5000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	72.6 ± 0.7	84.1 ± 0.7	80.2 ± 1.5	80.7 ± 0.5	77.2 ± 3.9
CHOCO	65.5 ± 1.3	87.6 ± 0.5	84.0 ± 0.9	83.0 ± 0.2	80.3 ± 0.5
DeepSqueeze	67.4 ± 0.5	82.9 ± 0.7	76.6 ± 0.9	76.6 ± 0.4	74.5 ± 0.3
BEER	67.3 ± 0.6	85.7 ± 0.4	83.4 ± 0.9	82.8 ± 0.2	82.4 ± 0.3
MoTEF	48.1 ± 3.0	78.6 ± 3.5	85.4 ± 0.5	83.2 ± 1.9	81.8 ± 1.4
A-DEFEAT	82.4 ± 2.0	**92.6 ± 0.0**	91.9 ± 0.1	**91.6 ± 0.1**	**90.8 ± 0.3**
A-DEFEAT-GT	**85.7 ± 0.9**	92.1 ± 0.2	**91.9 ± 0.3**	91.3 ± 0.1	90.7 ± 0.4

Table 54. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.2$ and KMNIST. All results are compared using 5000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes	100 nodes	200 nodes
DCD	90.1 ± 0.4	90.6 ± 0.3	90.1 ± 0.1	89.0 ± 0.3	86.0 ± 0.8
CHOCO	92.9 ± 0.4	93.9 ± 0.1	92.9 ± 0.2	91.2 ± 0.6	88.1 ± 0.6
DeepSqueeze	92.0 ± 0.1	92.8 ± 0.1	91.1 ± 0.1	88.2 ± 0.4	85.7 ± 0.4
BEER	92.0 ± 0.2	92.6 ± 0.0	92.3 ± 0.2	90.9 ± 0.3	89.3 ± 0.7
MoTEF	65.5 ± 39.3	93.2 ± 0.4	93.2 ± 0.1	92.1 ± 0.2	88.2 ± 0.2
A-DEFEAT	**94.7 ± 0.1**	**94.9 ± 0.0**	**94.7 ± 0.1**	**94.2 ± 0.1**	**92.9 ± 0.1**
A-DEFEAT-GT	94.0 ± 0.2	94.3 ± 0.1	94.0 ± 0.1	93.6 ± 0.2	92.6 ± 0.1

I.3 CIFAR-10

I.3.1 RING WITH FIXED AGGREGATIONS

Table 55. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and CIFAR-10. All results are compared using 10000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	34.4 ± 2.5	39.0 ± 0.2	39.1 ± 0.9
CHOCO	24.5 ± 1.2	21.4 ± 3.6	25.0 ± 1.4
DeepSqueeze	18.5 ± 3.1	26.3 ± 2.4	26.6 ± 1.8
BEER	39.1 ± 1.9	39.5 ± 1.8	42.9 ± 0.5
MoTEF	14.7 ± 4.7	11.3 ± 0.7	12.3 ± 2.3
A-DEFEAT	43.7 ± 1.0	46.3 ± 1.8	46.5 ± 1.4
A-DEFEAT-GT	**58.0 ± 2.5**	**55.3 ± 3.5**	**59.9 ± 1.0**

Table 56. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and CIFAR-10. All results are compared using 10000 aggregation rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	57.9 ± 2.1	64.4 ± 0.6	56.1 ± 1.5
CHOCO	61.1 ± 1.3	61.7 ± 0.3	59.0 ± 0.2
DeepSqueeze	62.1 ± 0.9	61.5 ± 0.5	58.9 ± 0.1
BEER	64.6 ± 1.4	67.2 ± 0.6	59.7 ± 0.5
MoTEF	56.2 ± 2.2	63.9 ± 1.5	63.3 ± 1.4
A-DEFEAT	67.0 ± 1.2	68.2 ± 1.7	62.7 ± 0.7
A-DEFEAT-GT	$\mathbf{70.9 \pm 0.5}$	$\mathbf{71.2 \pm 0.2}$	$\mathbf{69.9 \pm 0.5}$

I.3.2 Ring with fixed communications

Table 57. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and CIFAR-10. All results are compared using 10000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	34.4 ± 2.5	39.0 ± 0.2	39.1 ± 0.9
CHOCO	24.5 ± 1.2	21.4 ± 3.6	25.0 ± 1.4
DeepSqueeze	18.5 ± 3.1	26.3 ± 2.4	26.6 ± 1.8
BEER	38.1 ± 0.9	37.3 ± 0.4	40.0 ± 0.6
MoTEF	17.6 ± 2.4	12.1 ± 1.1	15.3 ± 3.1
A-DEFEAT	43.7 ± 1.0	46.3 ± 1.8	46.5 ± 1.4
A-DEFEAT-GT	$\mathbf{50.8 \pm 5.0}$	$\mathbf{54.4 \pm 0.4}$	$\mathbf{57.1 \pm 0.7}$

Table 58. Test accuracy % comparison. The experiments use a ring network with different numbers of nodes. This table summarizes the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and CIFAR-10. All results are compared using 10000 communication rounds, each value averaged over the last 50 rounds.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	57.9 ± 2.1	64.4 ± 0.6	56.1 ± 1.5
CHOCO	61.1 ± 1.3	61.7 ± 0.3	59.0 ± 0.2
DeepSqueeze	62.1 ± 0.9	61.5 ± 0.5	58.9 ± 0.1
BEER	59.7 ± 0.6	57.1 ± 1.4	58.2 ± 0.6
MoTEF	48.7 ± 2.9	60.3 ± 3.6	59.0 ± 1.3
A-DEFEAT	$\mathbf{67.0 \pm 1.2}$	$\mathbf{68.2 \pm 1.7}$	62.7 ± 0.7
A-DEFEAT-GT	65.2 ± 0.7	66.9 ± 0.7	$\mathbf{67.7 \pm 0.3}$

I.4 CIFAR-100

I.4.1 Ring with fixed aggregations

Table 59. Test accuracy % comparison. Results are reported for a ring network with different number of nodes. This table summarize the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and CIFAR-100. All results are compared using 10000 aggregation rounds, each value averaged over the last 50 rounds. $-$ represents that algorithm diverges for all seeds.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	23.2 ± 0.3	–	5.9 ± 0.4
CHOCO	30.9 ± 0.4	–	8.9 ± 0.1
DeepSqueeze	30.8 ± 0.5	–	8.3 ± 0.7
BEER	28.2 ± 2.5	26.0 ± 0.6	15.2 ± 0.7
MoTEF	23.9 ± 2.8	15.5 ± 0.6	15.3 ± 1.3
A-DEFEAT	35.8 ± 0.4	–	17.3 ± 0.3
A-DEFEAT-GT	$\mathbf{38.4 \pm 0.8}$	$\mathbf{35.6 \pm 0.7}$	$\mathbf{25.3 \pm 0.5}$

Table 60. Test accuracy % comparison. Results are reported for a ring network with different number of nodes. This table summarize the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and CIFAR-100. All results are compared using 10000 aggregation rounds, each value averaged over the last 50 rounds. † marks a run whose curve stops before the reading round because a seed diverged; its last available value is reported.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	25.0 ± 0.4	30.5 ± 0.5	23.7 ± 0.6
CHOCO	24.1 ± 16.3 †	33.8 ± 0.4	27.2 ± 0.3
DeepSqueeze	35.5 ± 0.6	34.2 ± 0.3	27.9 ± 0.4
BEER	36.5 ± 0.8	35.2 ± 0.4	27.5 ± 0.7
MoTEF	24.2 ± 8.8	30.0 ± 1.0	36.0 ± 0.5
A-DEFEAT	39.2 ± 1.3	38.1 ± 0.6	32.6 ± 0.3
A-DEFEAT-GT	$\mathbf{42.0 \pm 1.5}$	$\mathbf{41.3 \pm 0.4}$	$\mathbf{38.9 \pm 0.2}$

I.4.2 Ring with fixed communications

Table 61. Test accuracy % comparison. Results are reported for a ring network with different number of nodes. This table summarize the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0$ and CIFAR-100. All results are compared using 10000 communication rounds, each value averaged over the last 50 rounds. − represents that algorithm diverges for all seeds.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	23.2 ± 0.3	–	5.9 ± 0.4
CHOCO	30.9 ± 0.4	–	8.9 ± 0.1
DeepSqueeze	30.8 ± 0.5	–	8.3 ± 0.7
BEER	26.3 ± 0.4	22.2 ± 0.3	10.9 ± 0.5
MoTEF	24.9 ± 1.8	16.9 ± 0.1	12.5 ± 0.8
A-DEFEAT	$\mathbf{35.8 \pm 0.4}$	–	17.3 ± 0.3
A-DEFEAT-GT	32.7 ± 0.9	$\mathbf{31.8 \pm 0.2}$	$\mathbf{21.8 \pm 0.5}$

Table 62. Test accuracy % comparison. Results are reported for a ring network with different number of nodes. This table summarize the final test accuracy achieved by each algorithm for heterogeneity parameter $\alpha = 0.05$ and CIFAR-100. All results are compared using 10000 communication rounds, each value averaged over the last 50 rounds. † marks a run whose curve stops before the reading round because a seed diverged; its last available value is reported.

Algorithms	10 nodes	20 nodes	50 nodes
DCD	25.0 ± 0.4	30.5 ± 0.5	23.7 ± 0.6
CHOCO	24.1 ± 16.3 †	33.8 ± 0.4	27.2 ± 0.3
DeepSqueeze	35.5 ± 0.6	34.2 ± 0.3	27.9 ± 0.4
BEER	29.4 ± 0.9	30.8 ± 0.5	28.6 ± 0.5
MoTEF	25.8 ± 3.0	34.9 ± 1.3	32.6 ± 0.9
A-DEFEAT	39.2 ± 1.3	38.1 ± 0.6	32.6 ± 0.3
A-DEFEAT-GT	$\mathbf{39.7 \pm 1.3}$	$\mathbf{38.7 \pm 0.6}$	$\mathbf{37.7 \pm 0.3}$