EQUIVALENT ON ACCURACY, ORDERED IN FIDELITY: AGGREGATION UNDER PARTIAL RECEPTION IN DECENTRALIZED FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In decentralized federated learning over unreliable links without retransmission, a receiver obtains only a subset of each neighbor's model coordinates with reception probability q, which are unobservable in advance and *channel-imposed* rather than designed or selected. We show that final accuracy, the standard comparison metric, fails to distinguish aggregation rules under standard annealing schedules. We propose a direct measure of *aggregation fidelity*, the distance between the aggregate formed under coordinate loss and its lossless counterpart, and use it to give a two-part characterization. First, fidelity separates aggregators into two families. Convex per-coordinate weights cannot amplify disagreement in any realization, whereas naive loss compensation incurs variance $\Theta((1-q)/q)$ that is coupled to parameter magnitude rather than to the step size. This property is path-wise, so learning-rate schedules cannot change family membership, as confirmed under constant and cosine schedules. Second, within this convex family, the difference between rules persists, but their visibility depends on the schedule. Final accuracy separates aggregation rules at a constant rate, but converges under cosine annealing. Fidelity, consensus, and convergence cost nevertheless preserve their ordering, with the fidelity gap widening monotonically from $1.56\times$ to $7.62\times$ against the strongest baseline on CIFAR-10 as availability falls. Benchmarking on final accuracy under a single schedule therefore identifies a rule's family but never the ordering within it.

1 INTRODUCTION

Decentralized federated learning (DFL) enables privacy-preserving machine learning through peer-to-peer gossip, in which nodes train locally and exchange model updates with their neighbors to reach consensus (agreement), avoiding the bottlenecks of centralized federated learning (Yuan et al., 2024; McMahan et al., 2017). Most analyses assume a neighbor's update arrives intact, but two deployment conditions break that assumption. Over wireless links, ad-hoc networks, and commodity IP fabrics without retransmission (Imteaj et al., 2022; Park et al., 2021; Abad et al., 2020), a single checksum failure discards a packet of the model outright. More generally, synchronous rounds with deadlines must accept partial updates, since a retransmitted packet may still miss the deadline. Partial reception is therefore a property of latency-bounded round-based training as much as lossy links. In both cases, the receiver holds an arbitrary subset of each neighbor's coordinates, and the subset changes at every communication round.

Partial reception is easily conflated with compression, but under sparsification or quantization (Alistarh et al., 2017; Vargaftik et al., 2022; Tang et al., 2023) the sender *chooses* what to transmit and both endpoints know the retained structure. Here the partiality mask is channel-imposed, receiver-specific, and unobservable in advance. A compressed payload can itself be lost in transit, so the two compose rather than compete, and this work is not a contribution to communication efficiency.

Existing solutions impute surrogates for missing coordinates (Vargaftik et al., 2022; Li et al., 2024; Ye et al., 2022; Movahedian et al., 2023), biasing the aggregate, and then correct the bias by reweighting with an assumed, known, or estimated reception probability, through channel-aware weighting, inverse-probability weighting (Horvitz & Thompson, 1952), its normalized vari-

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.

ants (Kaluannakkage & Buyya, 2026), or rotation-based compressors that rescale after loss (Vargaftik et al., 2022). All of these attempt to reconstruct the update a perfect channel would have delivered, from information that does not contain it. For decentralized consensus, however, unbiasedness is not essential. What matters most is that nodes remain in agreement while continuing to make progress toward optimization. This motivates *faithful averaging* over what actually arrived, inferring nothing. The rule coincides with the Hájek estimator (Datta & Polson, 2026) under homogeneous reception, and its convex weights prevent the operator from amplifying disagreement in any realization. Unnormalized rules instead couple error to parameter magnitude, so it survives step-size decay.

Establishing which aggregators are stable is only half the question. The difference between convex rules appears under every schedule we test, but *which measurement reveals it* depends on the schedule. At a constant learning rate, it appears in the final accuracy. Under cosine annealing (Loshchilov & Hutter, 2017), the vanishing late-stage step size suppresses each convex rule's residual aggregation error, and their final accuracies overlap, yet the same runs remain clearly separated in aggregation fidelity, consensus, and convergence cost. The unnormalized family fails under both schedules. The schedule therefore never changes family membership or removes the within-family difference, only *where that difference is visible*, so a benchmark that evaluates final accuracy under a single schedule identifies family membership while systematically under-measuring within-family robustness. We use Mask-Faithful-Gossip (MFG), a realized-coordinate rule, as the control condition. It is the convex-family member that adds nothing beyond count normalization, and thus isolates faithful normalization from the channel estimates, weights, and coordination introduced by the other convex rules. That this minimal rule is also the most faithful across all setups is a result about the regime rather than a claim about MFG. The machinery the others add does not pay for itself under channel-imposed loss.

Our contributions are: **(i)** ***The imputation trap.*** No coordination-free aggregator is unbiased for the loss-free neighbor mean at every loss profile (Thm. 4.1), so imputation carries an irreducible bias floor (Lem. 4.2), while inverse-probability correction inflates variance to $\Theta((1-q)/q)$ and couples the error to parameter magnitude rather than to the step size (Props. 4.4–4.5). **(ii)** ***A schedule-robust dichotomy.*** Aggregators separate by whether their error couples only to the step size or has additional coupling to the parameter magnitude (Def. 5.1). The convex family cannot amplify disagreement in any realization (Prop. 5.3), and the unnormalized family can, so no schedule alters family membership. **(iii)** ***The communication-optimization interaction.*** Within the convex family, which measurement reveals the difference depends on the schedule. The displacement ratio $R_t = \|\Delta_{\text{cons}}\|/\|\Delta_{\text{SGD}}\|$, the share of each round's movement due to aggregation, shows the realized-coordinate rule consensus-dominated under both schedules, while the others are SGD-dominated until annealing lifts them above unity. **(iv)** ***A persistent within-family fidelity ordering.*** Measured against the loss-free neighbor mean, aggregation fidelity keeps the convex rules ordered throughout training, with the gap against the strongest baseline widening monotonically from $1.56\times$ to $7.62\times$ (CIFAR-10) and $2.20\times$ to $6.06\times$ (EMNIST) even where final accuracies coincide. The rules also differ in rounds-to-target, which is proportional to total communication. **(v)** ***An evaluation confound.*** Single-schedule, final accuracy benchmarking measures neither within-family fidelity nor convergence cost.

2 Related Work

Decentralized optimization and compression. Gossip protocols (Kempe et al., 2003; Boyd et al., 2006) established decentralized averaging, and D-PSGD (Lian et al., 2017) matched the convergence rates of centralized Stochastic Gradient Descent (SGD). Later work extends this to compression (Koloskova et al., 2019), directed and time-varying graphs (Assran et al., 2019; Nedić & Olshevsky, 2015), general mixing (Koloskova et al., 2020), and sporadic updates (Zehtabi et al., 2025) and these ideas carry into DFL systems (Roy et al., 2019; Wang et al., 2023; Hua et al., 2024). Quantization, sparsification (Alistarh et al., 2017) and rotation-based unbiased compressors (Vargaftik et al., 2022) reduce communication, but the sender chooses what to transmit. This entire line assumes the payload arrives intact, and without that assumption exact consensus is unattainable.

Aggregation under partial reception. Existing works impute a surrogate for each missing coordinate and compensate with channel-aware weighting. Imputing zero drags the aggregation towards

the origin (Li et al., 2024), whereas imputing local models' coordinates avoids this but requires a correction (Ye et al., 2022; Movahedian et al., 2023), which has been extended to directed and asynchronous settings (Kaluannakkage & Buyya, 2026). Others recover via low-rank approximation with system network coding (Kwon & Park, 2026) at 172% packet overhead, or by exchanging sender-chosen sparse coordinates to mitigate reconstruction risks (Sha et al., 2026). The count-normalized operator we study has been independently proposed for parameter dilution under deliberate sparsity (Bai et al., 2025; Dai et al., 2022; Bibikar et al., 2022; Ma et al., 2026), but in those settings the mask is algorithmically chosen and known to participants. Here it is stochastic, receiver-specific, and unobservable, making coordination-free unbiasedness impossible and leaving operator stability as the central question. In survey sampling terms, Horvitz-Thompson (Horvitz & Thompson, 1952) is unbiased but unstable at low inclusion probabilities, while its self-normalization counterpart, the Hájek estimator (Datta & Polson, 2026), trades unbiasedness for convex weights.

Learning rate schedule and consensus dynamics. Standard analysis combines a consensus recursion with an optimization recursion (Koloskova et al., 2020), giving a consensus floor that scales with η^2, so the step size directly controls residual disagreement. (Kong et al., 2021) identify a critical consensus distance that makes this coupling explicit, and (Vogels et al., 2022) show that sparse averaging permits larger step sizes and that overly tight consensus can hurt optimization. Scheduling has been applied to the mixing rate itself (Aketi et al., 2024). All of this assumes reliable delivery. The interaction between the schedule and aggregation under channel-imposed partial reception, where the effective mixing strength is itself a random function of what arrived, has not been studied.

3 Problem Formulation

N nodes form a connected undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = N$, where node i has degree $d_i = |\mathcal{N}_i|$ and model $\mathbf{x}_i \in \mathbb{R}^P$ with P parameters ($p \in P$). Nodes minimize $F(\mathbf{x}) = \frac{1}{N}\sum_i f_i(\mathbf{x}), \quad f_i(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i}[\ell(\mathbf{x};\xi)]$, over non-IID private dataset $\mathcal{D}_i$. In each synchronous round, node i performs k local SGD steps to produce $\tilde{\mathbf{x}}_i$ and broadcasts it in chunks. Each chunk from j reaches i independently with probability q_{ji}, so all coordinates in a chunk share one indicator $m_{ji}^{(p,t)} \sim \text{Bernoulli}(q_{ji})$, with $m_{ii}^{(p,t)} \equiv 1$ and independence across links and rounds. Our results are per-coordinate and require independence only across links, so this within-link correlation does not affect them. The delivering set is $\mathcal{S}_i^{(p,t)} = \{j \in \mathcal{N}_i \mid m_{ji}^{(p,t)} = 1\}$ with realized count $K_i^{(p,t)} = |\mathcal{S}_i^{(p,t)}|$. Because $K_i^{(p,t)}$ varies across coordinates within a round, even a fixed nominal mixing matrix induces a random, coordinate-dependent effective mixing operation.

Agreement is measured by the consensus distance $D^{(t)} = \frac{1}{N}\sum_i \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2$, with $\bar{\mathbf{x}}^t = \frac{1}{N}\sum_i \mathbf{x}_i^t$, and an algorithm achieves bounded consensus if $\limsup_{t\to\infty} \mathbb{E}[D^{(t)}] < \infty \; \forall q \in (0, 1]$. We do not establish this bound (Rem. 3.1) and instead report measured consensus distance in §7.1. We study linear, mask-measurable per-coordinate rules:

$$A_i^{(p)} = a_{ii}(\mathcal{S}_i^{(p)})x_i^{(p)} + \sum_{j\in\mathcal{N}_i} a_{ij}(\mathcal{S}_i^{(p)})m_{ji}^{(p)}x_j^{(p)}. \tag{1}$$

whose coefficients may depend on the realized arrival pattern but not on the unobserved missing values, a class that includes ZeroFill, local-fill, inverse-probability weighting, and the realized-coordinate rule. Collecting these coefficients gives the per-coordinate mixing matrix $W^{(p)} \in \mathbb{R}^{N\times N}$, where $[W^{(p)}]_{ij}$ is the weight node i places on node j's value, so that $x_i^{(p),+} = \sum_j [W^{(p)}]_{ij}x_j^{(p)}$. Because its entries depend on what arrives, $W^{(p)}$ is random and differs across coordinates even when the nominal mixing matrix W of the communication graph is fixed. W_{ii} denotes node i's self-weight under lossless delivery. We also track the coordinate range $\text{diam}^{(p)}(x) = \max_{i\in\mathcal{V}} x_i^{(p)} - \min_{i\in\mathcal{V}} x_i^{(p)}$

Remark 3.1 (Scope). *Prop. 5.3 bounds the aggregation step in isolation, not* $\limsup_t \mathbb{E}[D^{(t)}]$ *over the full trajectory, since local SGD injects fresh disagreement between aggregation steps and* $W^{(p)}$ *is row-stochastic but not column-stochastic, so the network average is not preserved. The iterates therefore need not converge to the minimizer of the unweighted objective* F*, and our claims concern agreement and aggregation fidelity rather than optimality with respect to* F*.*

Definition 3.2 (Coordination-free aggregation)**.** *A rule is coordination-free if its coefficients depend only on the values received and the arrival mask of parameters, thereby requiring no knowledge of or estimates of q_{ji}, no shared randomness, and no globally coordinated mixing weights.*

4 The Imputation Trap

Theorem 4.1 (Unbiasedness requires the reception probabilities)**.** *No coordination-free aggregator in the class of Eq. 1 is unbiased for the loss-free neighbor mean $T = \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i\cup\{i\}} \mathbf{x}_j$ at every reception profile $\{q_{ji}\} \in (0,1]^{d_i}$. Unbiasedness at known profile requires the Horvitz-Thompson (HT) weights $a_{ij} = \frac{1}{q_{ji}(d_i+1)}$, which depend on q_{ji} explicitly.*

Lemma 4.2 (Imputation bias floor)**.** *A fixed-weight rule that substitutes a mask-independent surrogate u for each missing coordinate has bias $\mathbb{E}_S[A^{imp,(p)} - T] = \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i}(1-q_{ji})(u - x_j^{(p)})$, which is non-zero whenever $q < 1$ and cannot be removed by tuning the weights.*

Corollary 4.3 (Zero-fill is a contraction, not a bias)**.** *For $u = 0$ the missing weight mass is multiplied by zero rather than redistributed, so each round applies expected total weight $\rho(q) = W_{ii} + q(1 - W_{ii}) < 1$ to real values. The iterate then contracts geometrically, $\|x^{(t)}\| \approx \rho(q)^t\|x^{(0)}\| \to 0$, after which the classifier emits constant logits and accuracy falls to chance. An error term proportional to the state is a decay rate, not an offset.*

Proposition 4.4 (Inverse-probability trap)**.** *The Horvitz–Thompson estimator $\hat{A}_i = \frac{1}{d_i+1}\big[x_i^{(p)} + \sum_j \frac{m_{ji}^{(p)}}{q_{ji}} x_j^{(p)}\big]$ is unbiased for T, since each term satisfies $\mathbb{E}\big[\frac{m_{ji}^{(p)}}{q_{ji}} x_j^{(p)}\big] = x_j^{(p)}$. However, its per-neighbor variance is $\mathrm{Var}_{\mathcal{S}}\big(\frac{m_{ji}^{(p)}}{q_{ji}} x_j^{(p)}\big) = \frac{1-q_{ji}}{q_{ji}}\|x_j^{(p)}\|^2 = \Theta\big((1-q)/q\big) \to \infty$ as $q \to 0$. The variance is proportional to the parameter magnitude and contains no factor of η, so step-size decay cannot suppress it.*

Proposition 4.5 (Correction destabilizes the operator)**.** *Unnormalized inverse-probability correction assigns realized weight $1/q_{ji} > 1$ to each arrived neighbor. It's realized row sum is $\frac{1}{d_i+1}\big(1 + K_i^{(p)}/q\big)$, which exceeds one with positive probability for every $q \in (0,1)$. The update is therefore not a convex combination of existing values, and the per-round aggregation error has variance $\Theta((1-q)/q)\|x\|^2$ (Prop. 4.4).*

We do not derive a closed-form divergence rate for $\mathbb{E}[D^{(t)}]$. The trajectory-level consequence is established empirically, where the unnormalized family's consensus distance reaches unpresentable range on CIFAR-10 and 10^{24} on EMNIST (Tab. 1). In-depth theoretical analysis is in App. A.

5 The Two-Family Dichotomy and MFG

Definition 5.1 (Convex vs. unnormalized)**.** *An aggregator is convex if its per-coordinate weights are non-negative and sum to one for every arrival pattern, so that $W^{(p)}$ is row-stochastic with $\|W^{(p)}\|_\infty = 1$ path-wise. It is unnormalized otherwise.*

The main dividing line is convexity, not correction vs. faithfulness. Soft-DSGD (Ye et al., 2022) reaches it by construction through fixed weights; DFL-AA (Kaluannakkage & Buyya, 2026) by self-normalizing over the realized weight sum despite its reception and staleness correction; and MFG by normalizing over the realized count. Zero-fill applies no correction at all yet is unnormalized, since the weight on each missing coordinate is discarded rather than redistributed. The distinction matters because the two families' errors depend on different quantities. A convex rule's disagreement arises only from the local steps and therefore scales with the step-size. An unnormalized rule adds an error that is not proportional to parameter magnitude when weights exceed one (Prop. 4.4) and contracts the iterate when they fall below one (Cor. 4.3), and step-size decay cannot suppress it.

5.1 The Convex Family and its Minimal Member: Mask-Faithful-Gossip

The minimal operator is the count-normalization rule, which we call *Mask-Faithful Gossip (MFG)*. It coincides with the Hájek estimator (Datta & Polson, 2026) when reception is homogeneous, since

the inverse weights then cancel, and under heterogeneous q_{ji} it takes the unweighted mean over the realized arrival set rather than weighting by inclusion probabilities, which Def. 3.2 excludes. It is faithful rather than imputing, and therefore belongs to a convex family. For each coordinate p, node i averages its own value together with the values received from neighbors whose corresponding chunks arrived successfully, so the MFG update is:

$$x_i^{(p),+} = \frac{x_i^{(p)} + \sum_{j \in S_i^{(p)}} x_j^{(p)}}{1 + |S_i^{(p)}|} = \underbrace{\frac{1}{K_i^{(p)} + 1}}_{\text{self-weight}} x_i^{(p)} + \underbrace{\frac{K_i^{(p)}}{K_i^{(p)} + 1}}_{\alpha_i^{(p)}} \cdot \underbrace{\frac{1}{K_i^{(p)}} \sum_{j \in S_i^{(p)}} x_j^{(p)}}_{\text{realized-coordinate mean}}. \tag{2}$$

We use this equivalent form throughout the analysis, with $\alpha_i^{(p)} = 0$ when $K_i^{(p)} = 0$, so the node keeps its own value. Two weights sum to 1, which makes the operator row-stochastic (Lem. 5.2). Note that α decreases as fewer neighbor coordinates are received, resulting in less mixing under scarce reception and tending towards self-retention. Two structural properties follow directly from Eq. 2,

Lemma 5.2 (Row-stochastic weights)**.** *For every arrival realization* $[W^{(p)}]_{ij} \geq 0$ *and* $\sum_j [W^{(p)}]_{ij} = 1$*, so* $\|W^{(p)}\|_\infty = 1$ *path-wise.*

Proposition 5.3 (Path-wise diameter non-expansiveness)**.** *For every arrival realization, the realized-coordinate update satisfies*

$$\operatorname{diam}^{(p)}\big(x^+\big) \;\leq\; \operatorname{diam}^{(p)}(x). \tag{3}$$

The proof (App. A.6) requires only non-negative weights summing to one. The unnormalized family violates this in either direction. Inverse-probability correction exceeds one with positive probability (Prop. 4.5), and ZeroFill falls below it (Cor. 4.3), so in both cases the coordinate range is no longer controlled. Since Eq. 3 holds path-wise, the result is independent of the learning-rate schedule.

The effective mixing strength also has a closed-form expression, which allows us to quantify the self-retention noted above.

Proposition 5.4 (Expected effective mixing)**.** *With* d_i *potential contributors each delivering parameter* p *independently with homogeneous reception probability* $q_{ji} = q$*, so* $K_i^{(p)} \sim$ *Binomial*(d_i, q)

$$\mathbb{E}_{\mathcal{S}}[\alpha_i^{(p)}] = \mathbb{E}_{\mathcal{S}}\Big[\frac{K_i^{(p)}}{K_i^{(p)} + 1}\Big] = 1 - \frac{1 - (1-q)^{d_i+1}}{(d_i+1)q}. \tag{4}$$

Tab. 2 in Appendix A.7 provides empirical evidence for both self and neighbor weights for this case where the self-weight rises monotonically from 0.200 at full reception to 0.819 at 90% loss.

Remark 5.5 (Coverage: degree, not loss rate)**.** *A coordinate contributes extra disagreement only when no neighbor delivers it, and we describe it as coverage failure, which occurs with probability* $(1-q)^{\bar{d}}$ *under* homogeneous reception *for an idealized* $\bar{d}$*-regular proxy graph, where every node is assigned the expected degree* $\bar{d} = \mathbb{E}[d_i]$*. These failures are drawn each round independently, so over* T *rounds a coordinate is still mixed* $T \cdot (1 - (1-q)^{\bar{d}})$ *times (Tab. 3 in Appendix A.8). Coverage failure therefore delays mixing rather than preventing it. This is why faithful aggregation approaches the lossless references even at 90% loss, and why degree, rather than loss rate, governs fidelity.*

The aggregation cost is $O(P\, d_i)$ per round for a model of P parameters, with no per-coordinate loop in vectorized implementation.

6 The Communication-Optimization Interaction

In each round, local SGD and aggregation together update a node's model. Local SGD moves it toward a minimum of the node's own objective. However, under non-IID data, this is not the ideal objective of decentralized learning. Aggregation moves it back toward what its neighbors hold. An aggregation rule therefore cannot be understood from the aggregation step alone, because what

matters is the size of the correction relative to the local step it corrects. Node i begins at $\mathbf{x}_i^t$, takes k local steps to reach $\tilde{\mathbf{x}}_i^t$, and aggregates to reach $\mathbf{x}_i^{t+1}$. The total movement decomposes as local displacement(Δ_{SGD}) and consensus displacement(Δ_{cons}):

$$\underbrace{\mathbf{x}_i^{t+1} - \mathbf{x}_i^t}_{\text{total}} = \underbrace{(\mathbf{x}_i^{t+1} - \tilde{\mathbf{x}}_i^t)}_{\Delta_{\mathrm{cons}}} + \underbrace{(\tilde{\mathbf{x}}_i^t - \mathbf{x}_i^t)}_{\Delta_{\mathrm{SGD}}}, \quad \Delta_{\mathrm{SGD}} = -\eta_t \sum_{s=1}^{k} \mathbf{g}_i^{t,s}. \tag{5}$$

For each coordinate p, the consensus term adopts the mixing form described in Eq. 2, where $\Delta_{\mathrm{cons}}^{(p)} = \alpha_i^{(p)} \left(\bar{x}_{\mathcal{S},i}^{(p)} - \tilde{x}_i^{(p)}\right)$ and $\bar{x}_{\mathcal{S},i}^{(p)}$ represents the mean of coordinate p over the neighbors that provided this coordinate, as defined in Eq. 2.

The two terms respond differently to the schedule. Under a constant learning rate, both displacements decline despite the fixed η_t. The local displacement decreases as the gradient norm shrinks, while the consensus displacement decreases as disagreement diminishes. Because these quantities decay at different rates, a gap emerges between them. During annealing, η_t decreases, driving both displacements down. Smaller steps also limit model drift, causing the two terms to decay together and the gap to close. Their magnitudes therefore do not distinguish the rules, but their ratio $R_t = \frac{\frac{1}{N}\sum_i \|\Delta_{\mathrm{cons},i}^t\|}{\frac{1}{N}\sum_i \|\Delta_{\mathrm{SGD},i}^t\| + \epsilon}$ with $\epsilon = 10^{-12}$ does, which measures the relative contribution of aggregation to each round's movement, compared with local descent (§ 7.3).

Why the schedule moves the difference without removing it. A rule that aggregates faithfully introduces less error in each round. With a constant step size, these per-round differences accumulate and become visible in the final accuracy. During annealing, the late-stage decay of the step size suppresses error accumulation and allows less faithful rules to settle at comparable final accuracies. However, the per-round difference arises from the aggregation operator rather than the schedule. It remains visible in fidelity (§7.2), the displacement ratio (§7.3), and rounds-to-target (§7.4). Annealing equalizes the final outcomes, not the cost of reaching them.

7 Performance Evaluation

Setup. 20 nodes on an Erdős–Rényi graph of average degree 4, with CIFAR-10 (Krizhevsky, 2009)/ResNet-20 and EMNIST (Cohen et al., 2017)/MLP (GroupNorm throughout, so no normalization statistics are averaged under non-IID data), Dirichlet(0.1) non-IID training data, and a shared test set. Each coordinate is received independently with probability $q \in \{1, 0.9, 0.8, 0.7, 0.5, 0.3, 0.1\}$, where 6 loss levels and $q = 1$ lossless reference. We use SGD with momentum 0.9 and weight decay 5×10^{-4}. Learning rates are tuned per schedule: $\eta = 0.03$ for constant, and $\eta_0 = 0.1$ for cosine peak. Results are mean $\pm$ std over three seeds, with identical data partitions, initializations, and reception masks across methods within each seed. **Baselines**. Our baselines span both families of Def. 5.1. The *unnormalized* family comprises ZeroFill (Li et al., 2024), which imputes zeros; the loss-handling component of EDEN (Vargaftik et al., 2022), which rescales received coordinates by inverse completeness; and HT-Gossip, which applies inverse-probability weighting (Horvitz & Thompson, 1952) with a fixed divisor and no renormalization. The *convex* family comprises Soft-DSGD (Ye et al., 2022), which imputes the receiver's own coordinates under a loss-aware mixing matrix; DFL-AA (Kaluannakkage & Buyya, 2026), which weights by staleness and estimated reception per link and then renormalizes; and MFG. D-PSGD under lossless communication (Lian et al., 2017) is the oracle reference. We establish the main phenomena on both datasets, while the mechanism and ablation studies use CIFAR-10 and CIFAR-100.

7.1 The two-family split is schedule-robust

Tab. 1 shows that the convex family degrades gracefully under both schedules, whereas the unnormalized family falls to chance as availability drops. Under a constant learning rate, final accuracy alone already separates the convex rules, since a fixed step size allows each round's aggregation error to compound throughout training. MFG injects the least error by remaining faithful to the received values rather than imputing and correcting missing messages (Props. 4.4 and 4.5). Consensus distance further clarifies this contrast. The convex family remains between 4.3×10^{-3} and 4.9×10^{-1}

on CIFAR-10, and between 5.7×10^{-2} and 2.9 on EMNIST, whereas EDEN reaches unpresentable range and 3.9×10^{24}, respectively.

Cosine annealing preserves the separation between the two families, as Prop. 5.3 requires, because non-amplification holds for every arrival realization and cannot depend on the step size. What changes is the final accuracy. The late reduction in the step size suppresses accumulated error, allowing rules with moderate aggregation error to settle close to the best-performing rule. At 20% and 50% loss, DFL-AA and Soft-DSGD match or exceed MFG in accuracy, while MFG retains the lowest consensus distance in both cells. It also preserves the ordering within the convex family, which spans 1.2×10^{-6} to 4.7×10^{-4} on CIFAR-10 and 1.8×10^{-5} to 8.9×10^{-3} on EMNIST. Thus, the trajectories remain separated even when their endpoints do not.

Why the three failures look different. All three unnormalized rules eventually reach chance-level accuracy, but their consensus distances converge to very different values. ZeroFill contracts because a missing coordinate is multiplied by zero and discarded, rather than reassigned (Cor. 4.3). As a result, the parameters decay toward zero, and all nodes converge to the same empty model. HT-Gossip, in contrast, inflates the received coordinates by a factor of $1/q$ to compensate for missing messages (Props. 4.4, 4.5). Its fixed denominator (eg., $d_i + 1$) then dampens this effect, keeping the resulting values within the representable range. Once accuracy reaches chance level, the gradients carry little to no useful signal, no substantial new disagreement is introduced, and repeated averaging reduces the consensus distance. EDEN also inflates the parameters, but without the same damping effect because individual coordinates cannot be recovered in the rotated domain. The entire received vector is rescaled by the inverse of the realized completeness. This inflation compounds over iterations until the parameter values exceed the representable range.

7.2 Measured aggregation faithfulness

Accuracy and consensus describe the model, not the aggregation. We therefore measure how far the aggregate a node forms sits from the aggregate it would have formed under lossless delivery as $Q^{(t)} = \frac{1}{N}\sum_i \frac{1}{P}\left\|A_i^{(t)} - \bar{x}^{(t)}_{\mathcal{N}_i \cup \{i\}}\right\|_2^2$. The predicted onset is shown in Fig. 1. At 20% loss, the unnormalized rules are elevated but flat. From 50% loss onward, they grow, as implied by Prop. 4.4, since $(1-q)/q$ increases from 0.25 at $q = 0.8$ to 9 at $q = 0.1$. Fidelity identifies where the transition begins, whereas accuracy indicates that these rules fail. Within the convex family, the ordering MFG $<$ DFL-AA $<$ Soft-DSGD holds at every loss level and for essentially every round on both datasets. Against the strongest baseline, the gap widens monotonically from $1.56\times$ to $7.62\times$ on CIFAR-10 and from $2.20\times$ to $6.06\times$ on EMNIST. To separate loss-induced disagreement from client drift, experiments on near IID data partitioning, reported in App. D.2, repeat the CIFAR-10 experiment with 70% loss and near IID data ($\mathrm{Dir}(\alpha) = 1.0$), and observe the same ordering.

The two measurements also differ in resolution. Under cosine annealing at 50% loss on CIFAR-10, MFG and Soft-DSGD are statistically indistinguishable in accuracy (74.0 ± 2.6 versus 75.5 ± 4.0), yet in the same runs, MFG's consensus distance is approximately $4.8\times$ lower and its aggregation error approximately $6.2\times$ lower. Thus, accuracy cannot distinguish the two rules, whereas both consensus distance and aggregation error can. The slightly higher point estimate for Soft-DSGD may reflect a mild regularizing effect of the looser consensus induced by local-parameter imputation under non-IID data, which final accuracy cannot distinguish from faithful aggregation.

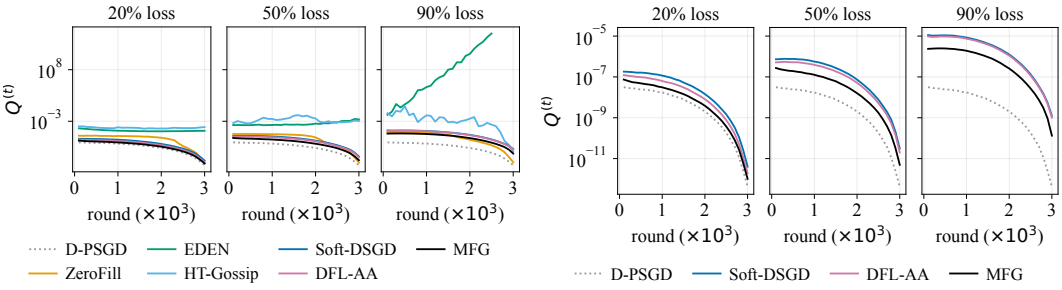
7.3 Mechanism: the displacement ratio

Fig. 2 measures both displacements at 50% loss. Under a constant rate, both decline even though η_t is fixed. The local displacement decreases with a shrinking gradient norm, while the consensus displacement decreases with a shrinking disagreement. Because these quantities decrease at different rates, a gap opens between them. During annealing, η_t decays toward zero, driving both displacements down, since smaller steps also limit the divergence between models. The two displacements therefore decay together, and the gap closes. Their magnitudes alone do not distinguish the rules, but their ratio does.

MFG maintains $R_t \approx 1.07$ under both schedules and is consensus-dominated in every round, while DFL-AA and Soft-DSGD remain near 0.91 and 0.87, respectively. Under a constant rate, this or-

Table 1: Final test accuracy (%) and consensus distance $D^{(T)}$, mean $\pm$ std over 3 seeds, for 3000 (CIFAR-10) and 1000 (EMNIST) rounds. 20 nodes, Dirichlet(0.1), average degree 4. Dashed rule and shaded rows separate the unnormalized from the convex family. Lossless D-PSGD: $59.0 \pm 5.0/81.0 \pm 1.1$ (CIFAR-10/EMNIST) at constant step size, $74.3 \pm 2.7/84.8 \pm 0.1$ under cosine. Consensus std shares the mean exponent, so $4.3\,\mathrm{e}{-3}$ with std $1.0\,\mathrm{e}{-3}$ prints $4.3 \pm 1.0\,\mathrm{e}{-3}$. †Accuracy at chance: consensus reflects a collapsed model (Para. 7.1). Per-column best in bold and full loss levels in Appendix D.1.

			10%		**20%**		**50%**		**90%**	
		Method	Acc.	Cons.	Acc.	Cons.	Acc.	Cons.	Acc.	Cons.
Constant Step-Size	CIFAR-10	ZeroFill	10.0 ± 0.0	$9.4 \pm 2.2\,\mathrm{e}{-2}^{\dagger}$	10.0 ± 0.0	$1.1 \pm 0.3\,\mathrm{e}{-1}^{\dagger}$	10.0 ± 0.0	$8.8 \pm 1.4\,\mathrm{e}{-2}^{\dagger}$	10.0 ± 0.0	$3.6 \pm 2.3\,\mathrm{e}{-2}^{\dagger}$
		EDEN	58.5 ± 5.1	1.0 ± 0.2	53.5 ± 5.4	3.8 ± 0.5	22.3 ± 0.3	$8.3 \pm 2.9\,\mathrm{e}2$	10.0 ± 0.0	—
		HT-Gossip	60.0 ± 2.2	6.0 ± 2.0	59.2 ± 4.7	$1.9 \pm 1.0\,\mathrm{e}1$	47.7 ± 3.7	$2.5 \pm 2.0\,\mathrm{e}2$	10.0 ± 0.0	$5.0 \pm 8.2\,\mathrm{e}2^{\dagger}$
		Soft-DSGD	58.9 ± 4.6	$7.4 \pm 1.7\,\mathrm{e}{-3}$	57.5 ± 3.7	$1.1 \pm 0.3\,\mathrm{e}{-2}$	54.7 ± 5.3	$3.8 \pm 0.6\,\mathrm{e}{-2}$	42.0 ± 4.6	$4.9 \pm 0.6\,\mathrm{e}{-1}$
		DFL-AA	57.5 ± 4.1	$4.9 \pm 1.1\,\mathrm{e}{-3}$	58.0 ± 4.5	$7.4 \pm 1.1\,\mathrm{e}{-3}$	55.2 ± 4.5	$2.6 \pm 0.3\,\mathrm{e}{-2}$	43.3 ± 3.7	$4.2 \pm 0.5\,\mathrm{e}{-1}$
		MFG	$\mathbf{60.1} \pm 2.4$	$\mathbf{4.3} \pm 1.0\,\mathrm{e}{-3}$	$\mathbf{60.1} \pm 3.2$	$\mathbf{5.1} \pm 1.2\,\mathrm{e}{-3}$	$\mathbf{60.0} \pm 4.6$	$\mathbf{9.7} \pm 1.9\,\mathrm{e}{-3}$	$\mathbf{51.4} \pm 5.3$	$\mathbf{1.1} \pm 0.2\,\mathrm{e}{-1}$
	EMNIST	ZeroFill	32.4 ± 1.7	$3.9 \pm 0.9\,\mathrm{e}{-1}$	13.7 ± 1.6	$4.0 \pm 1.1\,\mathrm{e}{-1}$	2.2 ± 0.0	$1.9 \pm 0.5\,\mathrm{e}{-1}^{\dagger}$	2.1 ± 0.0	$9.6 \pm 3.3\,\mathrm{e}{-2}^{\dagger}$
		EDEN	80.6 ± 1.1	3.2 ± 0.3	80.3 ± 1.2	9.0 ± 1.2	71.7 ± 4.9	$2.1 \pm 1.0\,\mathrm{e}2$	2.2 ± 0.1	$3.9 \pm 6.6\,\mathrm{e}24^{\dagger}$
		HT-Gossip	79.0 ± 1.7	6.1 ± 1.2	77.0 ± 1.6	$1.7 \pm 0.3\,\mathrm{e}1$	62.8 ± 1.9	$9.4 \pm 2.2\,\mathrm{e}2$	2.4 ± 0.3	$6.0 \pm 7.7\,\mathrm{e}3^{\dagger}$
		Soft-DSGD	80.0 ± 0.9	$8.2 \pm 1.8\,\mathrm{e}{-2}$	79.0 ± 1.1	$1.2 \pm 0.2\,\mathrm{e}{-1}$	74.7 ± 0.9	$3.2 \pm 0.3\,\mathrm{e}{-1}$	62.1 ± 1.4	2.9 ± 0.3
		DFL-AA	81.1 ± 0.9	$6.4 \pm 1.5\,\mathrm{e}{-2}$	80.2 ± 1.0	$8.7 \pm 1.7\,\mathrm{e}{-2}$	76.2 ± 1.1	$2.4 \pm 0.3\,\mathrm{e}{-1}$	64.0 ± 1.1	2.7 ± 0.3
		MFG	$\mathbf{81.5} \pm 0.7$	$\mathbf{5.7} \pm 1.3\,\mathrm{e}{-2}$	$\mathbf{81.3} \pm 0.7$	$\mathbf{6.9} \pm 1.4\,\mathrm{e}{-2}$	$\mathbf{80.3} \pm 0.7$	$\mathbf{1.3} \pm 0.2\,\mathrm{e}{-1}$	$\mathbf{72.8} \pm 1.7$	$\mathbf{8.7} \pm 1.6\,\mathrm{e}{-1}$
Cosine Annealing	CIFAR-10	ZeroFill	10.0 ± 0.0	$8.9 \pm 1.0\,\mathrm{e}{-7}^{\dagger}$	10.0 ± 0.0	$7.5 \pm 1.1\,\mathrm{e}{-7}^{\dagger}$	10.0 ± 0.0	$5.6 \pm 1.2\,\mathrm{e}{-7}^{\dagger}$	10.0 ± 0.0	$3.6 \pm 1.2\,\mathrm{e}{-7}^{\dagger}$
		EDEN	67.8 ± 3.1	$7.0 \pm 1.4\,\mathrm{e}{-1}$	57.1 ± 2.6	2.6 ± 0.4	10.4 ± 0.3	$6.8 \pm 6.8\,\mathrm{e}2^{\dagger}$	10.0 ± 0.0	—
		HT-Gossip	64.2 ± 3.9	5.4 ± 4.2	57.5 ± 4.8	$1.5 \pm 0.1\,\mathrm{e}1$	22.2 ± 1.9	$3.4 \pm 3.4\,\mathrm{e}2$	10.0 ± 0.0	$9.7 \pm 12\,\mathrm{e}{-5}^{\dagger}$
		Soft-DSGD	73.3 ± 3.5	$1.8 \pm 0.3\,\mathrm{e}{-6}$	$\mathbf{74.6} \pm 3.0$	$2.9 \pm 0.5\,\mathrm{e}{-6}$	$\mathbf{75.5} \pm 4.0$	$1.5 \pm 0.2\,\mathrm{e}{-5}$	73.5 ± 4.3	$4.7 \pm 0.1\,\mathrm{e}{-4}$
		DFL-AA	73.3 ± 1.6	$1.3 \pm 0.4\,\mathrm{e}{-6}$	73.3 ± 2.9	$1.9 \pm 0.6\,\mathrm{e}{-6}$	$\mathbf{75.5} \pm 3.4$	$9.4 \pm 1.4\,\mathrm{e}{-6}$	74.0 ± 3.0	$3.6 \pm 0.2\,\mathrm{e}{-4}$
		MFG	$\mathbf{73.9} \pm 1.2$	$\mathbf{1.2} \pm 0.4\,\mathrm{e}{-6}$	73.2 ± 1.3	$\mathbf{1.5} \pm 0.5\,\mathrm{e}{-6}$	74.0 ± 2.6	$\mathbf{3.1} \pm 1.1\,\mathrm{e}{-6}$	$\mathbf{76.1} \pm 2.7$	$\mathbf{4.9} \pm 0.9\,\mathrm{e}{-5}$
	EMNIST	ZeroFill	2.1 ± 0.0	$7.6 \pm 3.2\,\mathrm{e}{-7}^{\dagger}$	2.1 ± 0.0	$6.1 \pm 2.5\,\mathrm{e}{-7}^{\dagger}$	2.1 ± 0.0	$4.2 \pm 1.5\,\mathrm{e}{-7}^{\dagger}$	2.1 ± 0.0	$2.9 \pm 1.3\,\mathrm{e}{-7}^{\dagger}$
		EDEN	83.8 ± 0.3	3.6 ± 0.3	82.4 ± 0.5	$1.1 \pm 0.2\,\mathrm{e}1$	48.0 ± 13.9	$3.1 \pm 1.5\,\mathrm{e}2$	2.3 ± 0.1	$3.0 \pm 5.2\,\mathrm{e}24^{\dagger}$
		HT-Gossip	78.4 ± 1.6	7.0 ± 1.4	71.3 ± 1.9	$2.0 \pm 0.4\,\mathrm{e}1$	37.4 ± 1.5	$1.4 \pm 0.1\,\mathrm{e}3$	2.1 ± 0.1	$2.1 \pm 3.3^{\dagger}$
		Soft-DSGD	84.5 ± 0.1	$5.9 \pm 1.1\,\mathrm{e}{-5}$	84.3 ± 0.1	$9.9 \pm 1.5\,\mathrm{e}{-5}$	83.6 ± 0.1	$4.1 \pm 0.3\,\mathrm{e}{-4}$	81.9 ± 0.3	$8.9 \pm 1.5\,\mathrm{e}{-3}$
		DFL-AA	84.7 ± 0.2	$2.8 \pm 0.7\,\mathrm{e}{-5}$	84.4 ± 0.1	$5.5 \pm 1.1\,\mathrm{e}{-5}$	83.9 ± 0.0	$2.7 \pm 0.2\,\mathrm{e}{-4}$	82.4 ± 0.1	$7.1 \pm 1.0\,\mathrm{e}{-3}$
		MFG	$\mathbf{84.8} \pm 0.0$	$\mathbf{1.8} \pm 0.7\,\mathrm{e}{-5}$	$\mathbf{84.7} \pm 0.3$	$\mathbf{2.4} \pm 0.8\,\mathrm{e}{-5}$	$\mathbf{84.5} \pm 0.1$	$\mathbf{5.6} \pm 1.6\,\mathrm{e}{-5}$	$\mathbf{83.9} \pm 0.1$	$\mathbf{1.1} \pm 0.2\,\mathrm{e}{-3}$



(a) Aggregation fidelity $Q^{(t)}_{(\log-\mathrm{scale})}$: all methods

(b) Aggregation fidelity $Q^{(t)}_{(\log-\mathrm{scale})}$: convex family

Figure 1: Aggregation fidelity $Q^{(t)}_{(\log-\mathrm{scale})}$ for CIFAR-10 under cosine annealing, averaged over three seeds. Convergence across the convex family, with $Q^{(t)}$ remaining ordered throughout training.

dering is stable across all 3000 rounds. Under annealing, it persists for most of training, and the other rules cross unity only in the final fifth, where their accuracies converge. This is consistent with Prop. 5.4 rather than in tension with it. A falling α compares MFG with itself under full reception, whereas, across methods at the same loss rate, a fixed-weight rule assigns part of its weight to coordinates that have not arrived. The substituted value then contributes no displacement, allowing MFG to move further toward its neighbors despite its nominally smaller coefficient.

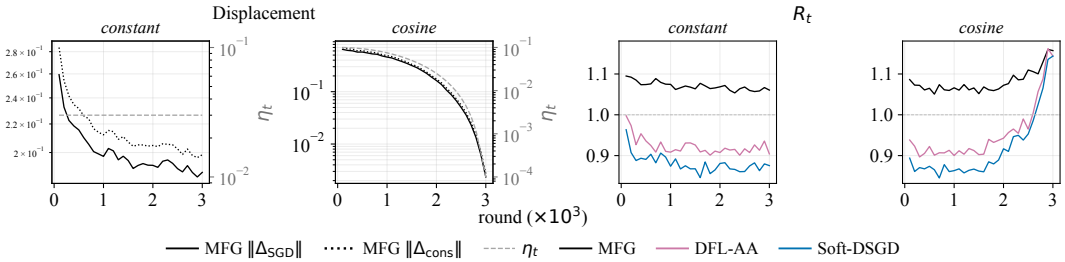


Figure 2: Per-round displacements and their ratio R_t at 50% loss on CIFAR-10, averaged over three seeds. Left pair: local and consensus displacements for MFG, with step size η_t on the right axis. Right pair: displacement ratio R_t across convex aggregation rules, with a dashed line at unity.

7.4 Convergence cost and Controls

At a 90% loss under a constant step size, the seed-averaged accuracy of MFG reaches 40% at round 1122, compared with 2265 and 2458 for DFL-AA and Soft-DSGD, within the 100-round evaluation interval. Under cosine annealing, MFG reaches the same threshold in 745 rounds, compared with 1377 and 1452, to within the 100-round evaluation interval. All three rules cross this threshold within the first 1452 rounds under annealing. At this point, the step size remains above 45% of its peak, so the observed ordering reflects the aggregation rule rather than the schedule's decay. Since each round corresponds to one full round of peer-to-peer communication, rounds-to-target is proportional to the communication required to reach a usable model. Thus, the communication cost is ordered by aggregation faithfulness under both schedules, including annealing, where the same runs achieve comparable final accuracies. Further results are in the appendix, including full results at every loss level (App. D.1), ablations over the learning rate, mixing rate, and topology (App. D.2). CIFAR-100 with ResNet-18, a harder task on a roughly $40\times$ larger model is also presented in App. D.2.

8 Conclusions

Our results suggest that partial reception is better treated as a measurement problem than as a communication problem. Aggregators fall into two families: one that cannot amplify disagreement in any realization, and one that can; no schedule alters this split. Within the convex family, the differences in fidelity and convergence cost persist even when final accuracies coincide or invert. That the minimal rule is also the most faithful throughout is a result about the regime rather than a claim about any method, since the estimates and weights the other convex rules add do not pay for themselves here. Our scope is bounded in four ways. Nodes share a single architecture per experiment. We cover three architectures across three datasets, without a transformer model. Reception probability is homogeneous within each run. We study memory-less aggregation that acts only on the current round's arrivals, because our analysis concerns the per-round operator. Stateful concealment, such as sample-and-hold or error feedback, and drift-correction methods, such as gradient tracking or exact diffusion, which operate on historical information, can be layered on any aggregation rule, so we leave them to future work. Two claims are bounded more sharply. Under annealing, the convex rules are statistically indistinguishable in accuracy. We claim no superiority there, and the constant-rate advantage is degree-dependent. We establish path-wise non-amplification (Prop. 5.3) rather than a quantitative bound on $\limsup_t \mathbb{E}[D^{(t)}]$, which is left to future work. Reporting both schedules and fidelity alongside accuracy would make these differences visible at negligible cost.

Use of large language models

We used large language models to assist with the manuscript's grammar correction, polish writing, code development, and the development of Prop. 5.3 and its proof. The research direction, problem formulation, experimental design, implementation, and analysis are the authors' own work. They also conducted and validated all experiments and independently checked all analytical results. The authors take full responsibility for the manuscript content, including any errors.

REFERENCES

M. S. H. Abad, E. Ozfatura, D. GUndUz, and O. Ercetin. Hierarchical federated learning across heterogeneous cellular networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8866–8870, 2020. doi: 10.1109/ICASSP40776.2020.9054634.

Sai Aparna Aketi, Sakshi Choudhary, and Kaushik Roy. Averaging rate scheduler for decentralized learning on heterogeneous data. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 344–353. PMLR, 09–15 Jun 2019.

Guangji Bai, Yijiang Li, Zilinghan Li, Liang Zhao, and Kibaek Kim. FedSpaLLM: Federated Pruning of Large Language Models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8361–8373, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.424.

Sameer Bibikar, Haris Vikalo, Zhangyang Wang, and Xiaohan Chen. Federated dynamic sparse training: Computing less, communicating less, yet learning better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6080–6088, 2022. doi: 10.1609/aaai.v36i6.20555.

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. DisPFL: Towards Communication-Efficient Personalized Federated Learning via Decentralized Sparse Training. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4587–4604. PMLR, July 2022.

Jyotishka Datta and Nicholas Polson. Inverse probability weighting: From survey sampling to evidence estimation. *The New England Journal of Statistics in Data Science*, pp. 1–13, 2026. ISSN 2693-7166. doi: 10.51387/26-NEJSDS100.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Yifan Hua, Jinlong Pang, Xiaoxue Zhang, Yi Liu, Xiaofeng Shi, Bao Wang, Yang Liu, and Chen Qian. Towards practical overlay networks for decentralized federated learning. In *2024 IEEE 32nd International Conference on Network Protocols (ICNP)*, pp. 1–11, 2024. doi: 10.1109/ICNP61940.2024.10858583.

Ahmed Imteaj, Urmish Thakker, Shiqiang Wang, Jian Li, and M. Hadi Amini. A survey on federated learning for resource-constrained iot devices. *IEEE Internet of Things Journal*, 9(1):1–24, 2022. doi: 10.1109/JIOT.2021.3095077.

Chanuka A. S. Hewa Kaluannakkage and Rajkumar Buyya. Asynchronous decentralized federated learning over lossy wireless links via reception- and age-aware aggregation. *arXiv preprint arXiv:2606.10774*, 2026.

D. Kempe, A. Dobra, and J. Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 482–491, 2003. doi: 10.1109/SFCS.2003.1238221.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3478–3487. PMLR, 09–15 Jun 2019.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 13–18 Jul 2020.

Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, and Sebastian Stich. Consensus control for decentralized deep learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5686–5696. PMLR, 18–24 Jul 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Jungmin Kwon and Hyunggon Park. Efficient and resilient packet recovery for federated learning via approximation. *IEEE Transactions on Mobile Computing*, 25(5):6413–6428, 2026. doi: 10.1109/TMC.2025.3636594.

Weicai Li, Tiejun Lv, Wei Ni, Jingbo Zhao, Ekram Hossain, and H. Vincent Poor. Decentralized federated learning over imperfect communication channels. *IEEE Transactions on Communications*, 72(11):6973–6991, 2024. doi: 10.1109/TCOMM.2024.3407208.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5336–5346, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Chengjie Ma, Seungeun Oh, Jihong Park, and Seong-Lyun Kim. Breaking the capacity bottleneck in model-heterogeneous federated learning via gradual model restoration. In *Forty-third International Conference on Machine Learning*, 2026.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

Mahtab Movahedian, Mahdi Dolati, and Majid Ghaderi. Adaptive model aggregation for decentralized federated learning in vehicular networks. In *2023 19th International Conference on Network and Service Management (CNSM)*, pp. 1–9, 2023. doi: 10.23919/CNSM59352.2023.10327827.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015. doi: 10.1109/TAC.2014.2364096.

Jihong Park, Sumudu Samarakoon, Anis Elgabli, Joongheon Kim, Mehdi Bennis, Seong-Lyun Kim, and Mérouane Debbah. Communication-efficient and distributed learning over wireless networks: Principles and applications. *Proceedings of the IEEE*, 109(5):796–819, 2021. doi: 10.1109/JPROC.2021.3055679.

Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. Braintorrent: A peer-to-peer environment for decentralized federated learning. *arXiv preprint arXiv:1905.06731*, 2019.

Shan Sha, Shenglong Zhou, Xin Wang, Lingchen Kong, and Geoffrey Ye Li. Decentralized federated learning by partial message exchange. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2026. doi: 10.1109/TPAMI.2026.3722165.

Zhenheng Tang, Shaohuai Shi, Bo Li, and Xiaowen Chu. Gossipfl: A decentralized federated learning framework with sparsified and adaptive communication. *IEEE Transactions on Parallel and Distributed Systems*, 34(3):909–922, 2023. doi: 10.1109/TPDS.2022.3230938.

Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, and Michael Mitzenmacher. EDEN: Communication-Efficient and Robust Distributed Mean Estimation for Federated Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 21984–22014. PMLR, June 2022.

Thijs Vogels, Hadrien Hendrikx, and Martin Jaggi. Beyond spectral gap: the role of the topology in decentralized learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15039–15050. Curran Associates, Inc., 2022. doi: 10.52202/068431-1094.

Lun Wang, Yang Xu, Hongli Xu, Min Chen, and Liusheng Huang. Accelerating decentralized federated learning in heterogeneous edge computing. *IEEE Transactions on Mobile Computing*, 22(9):5001–5016, 2023. doi: 10.1109/TMC.2022.3178378.

Hao Ye, Le Liang, and Geoffrey Ye Li. Decentralized federated learning with unreliable communications. *IEEE Journal of Selected Topics in Signal Processing*, 16(3):487–500, 2022. doi: 10.1109/JSTSP.2022.3152445.

Liangqi Yuan, Ziran Wang, Lichao Sun, Philip S. Yu, and Christopher G. Brinton. Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*, 11(21):34617–34638, 2024. doi: 10.1109/JIOT.2024.3407584.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher Brinton. Decentralized Sporadic Federated Learning: A Unified Algorithmic Framework with Convergence Guarantees. *International Conference on Learning Representations*, 2025:59567–59636, May 2025.

SUPPLEMENTARY MATERIALS

A PROOFS

A.1 PROOF OF THEOREM 4.1

Fix receiver i, coordinate p, round t. Let $d_i = |\mathcal{N}_i|$, $m_{ji} \sim \mathrm{Bern}(q_{ji})$ independent with $m_{ii} \equiv 1$, $\mathcal{S} = \{j \in \mathcal{N}_i : m_{ji} = 1\}$, $n = 1 + |\mathcal{S}|$. The $d_i + 1$ scalar values entering the aggregator form the tuple $v = (\mathbf{x}_i, (\mathbf{x}_j)_{j\in\mathcal{N}_i}) \in \mathbb{R}^{d_i+1}$, while the link probabilities range only over the d_i neighbors, $q = (q_{ji})_{j\in\mathcal{N}_i} \in (0,1)^{d_i}$. A rule (Eq. 1) is *coordination-free* if its coefficients do not use $\{q_{ji}\}$.

Proof. Taking the expectation of Eq. 1 over the mask and using that the values v are fixed while the indicators m_{ji} are independent of v, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[A_i] &= \mathbb{E}_{\mathcal{S}}[a_{ii}(\mathcal{S}_i)]\mathbf{x}_i + \sum_{j\in\mathcal{N}_i} \mathbb{E}_{\mathcal{S}}[a_{ij}(\mathcal{S}_i)\, m_{ji}]\, \mathbf{x}_j \\ &= a_{ii}(\mathcal{S}_i)\mathbf{x}_i + \sum_{j\in\mathcal{N}_i} \mathbb{E}_{\mathcal{S}}[a_{ij}(\mathcal{S}_i)\, m_{ji}]\, \mathbf{x}_j \\ &= a_{ii}\mathbf{x}_i + \sum_{j\in\mathcal{N}_i} \mathbb{E}_{\mathcal{S}}[a_{ij}(\mathcal{S}_i)\, m_{ji}]\, \mathbf{x}_j.\end{aligned} \tag{6}$$

Unbiasedness requires Eq. 6 to equal $T = \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i\cup\{i\}} \mathbf{x}_j$ for every $v \in \mathbb{R}^{d_i+1}$ and therefore,

$$a_{ii} = \frac{1}{d_i+1}, \qquad \mathbb{E}_{\mathcal{S}}\big[a_{ij}(\mathcal{S})\, m_{ji}\big] = \frac{1}{d_i+1} \quad \forall\, j \in \mathcal{N}_i. \tag{7}$$

Because $a_{ij}(\mathcal{S})$ multiplies with $m_{ji} = \mathbf{1}\{j \in \mathcal{S}\}$, which is an indicator of arrival, conditional expectation can be formed. We define $\bar{a}_{ij} = \mathbb{E}_{\mathcal{S}}[a_{ij}(\mathcal{S}) \mid j \in \mathcal{S}]$,

$$\mathbb{E}_{\mathcal{S}}\big[a_{ij}(\mathcal{S})\, m_{ji}\big] = \Pr[j \in \mathcal{S}]\, \mathbb{E}_{\mathcal{S}}\big[a_{ij}(\mathcal{S}) \mid j \in \mathcal{S}\big] = q_{ji}\, \bar{a}_{ij}. \tag{8}$$

Combining Eq. 7–Eq. 8 yields the *arrival-conditional identity*

$$\bar{a}_{ij} = \frac{1}{q_{ji}\,(d_i+1)}, \qquad j \in \mathcal{N}_i. \tag{9}$$

If the rule is coordination-free, each coefficient must be chosen without reference to q_{ji}, i.e. $\bar{a}_{ij}$ is constant in q_{ji}. This contradicts Eq. 9, which requires $\bar{a}_{ij} = \frac{1}{q_{ji}(d_i+1)}$, $\forall q_{ji} \in (0,1)$. Hence no coordination-free rule is unbiased for T.

The impossibility of unbiasedness also holds for weighted averages mean for $T = \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i\cup\{i\}} \mathrm{w}\cdot\mathbf{x}_j^{(p)}$, where w denotes the fixed weight for each neighbor and the own model itself, forming a non-random variable across the domain.

□

Remark A.1 (Remark (MFG))**.** MFG uses the coefficient $1/n$, which ignores q_{ji}. By the argument above, it is therefore biased for the loss-free mean T. This is by design, and MFG does not target T but the *realized-set mean* $\frac{1}{n}\sum_{j\in\mathcal{S}\cup\{i\}} \mathbf{x}_j$, the average over the sources that actually arrive, for which it is exactly unbiased and which requires no knowledge of q_{ji}.

A.2 PROOF OF LEMMA 4.2

Fix receiver i, coordinate p, and round t, and retain the notation of App. A.1. Let u be a surrogate value that does not depend on the mask (the indicator), and consider the fixed-weight rule following Eq. 1 that uses u to impute for missing coordinate,

$$A_i^{\mathrm{imp},(p)} = \frac{1}{d_i+1}\Big[x_i^{(p)} + \sum_{j\in\mathcal{N}_i}\Big(m_{ji}^{(p)}\, x_j^{(p)} + (1 - m_{ji}^{(p)})\, u\Big)\Big]. \tag{10}$$

Proof. Taking the expectation of Eq. 10 over the mask and using $\mathbb{E}_{\mathcal{S}}[m_{ji}^{(p)}] = q_{ji}$ (Bernoulli conditions), together with independence of the indicators from the values, gives

$$\mathbb{E}_{\mathcal{S}}\big[A_i^{\mathrm{imp},(p)}\big] = \frac{1}{d_i+1}\Big[x_i^{(p)} + \sum_{j\in\mathcal{N}_i}\Big(q_{ji}\,x_j^{(p)} + (1-q_{ji})\,u\Big)\Big]. \tag{11}$$

Subtracting the target, which is the loss-free neighbor mean for a single coordinate, which also follows T in Thm.4.1.

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\big[A_i^{\mathrm{imp},(p)} - T\big] &= \frac{1}{d_i+1}\Big[x_i^{(p)} + \sum_{j\in\mathcal{N}_i}\Big(q_{ji}\,x_j^{(p)} + (1-q_{ji})\,u\Big)\Big] \\
&\quad - \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i\cup\{i\}} x_j^{(p)} \\
&= \frac{1}{d_i+1}x_i^{(p)} + \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i} q_{ji}\,x_j^{(p)} + \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i}(1-q_{ji})\,u \\
&\quad - \left[\frac{1}{d_i+1}x_i^{(p)} + \frac{1}{d_i+1}\sum_{j\in\mathcal{N}_i} x_j^{(p)}\right] \\
&= \frac{1}{d_i+1}\left[\sum_{j\in\mathcal{N}_i} q_{ji}\,x_j^{(p)} + \sum_{j\in\mathcal{N}_i}(1-q_{ji})\,u - \sum_{j\in\mathcal{N}_i} x_j^{(p)}\right].
\end{aligned} \tag{12}$$

Considering the per-coordinate bias for a single neighbor j, the bias term forms

$$\begin{aligned}
&= \frac{1}{d_i+1}\Big(q_{ji}\,x_j^{(p)} + (1-q_{ji})\,u - x_j^{(p)}\Big) \\
&= \frac{1}{d_i+1}\Big((q_{ji}-1)\,x_j^{(p)} + (1-q_{ji})\,u\Big) \\
&= \frac{1-q_{ji}}{d_i+1}\big(u - x_j^{(p)}\big).
\end{aligned} \tag{13}$$

Summing over $j \in \mathcal{N}_i$ gives the stated bias. It vanishes only if $q_{ji} = 1$ or $u = x_j^{(p)}$, and $x_j^{(p)}$ is exactly the quantity that failed to arrive, so no choice of u made in advance can remove it. □

A.3 Proof of Corollary 4.3

ZeroFill considers a fixed weight matrix $[W]_{ij}$ along with imputed $u = 0$ for the Eq. 10, an aggregation operator therefore can be derived:

$$\mathbf{x}_i^{+} = W_{ii}\,\mathbf{x}_i + \sum_{j\in\mathcal{N}_i} W_{ij}\,m_{ji}\,\mathbf{x}_j\,, \tag{14}$$

where the weight matrix can take any fixed values and satisfies the row-stochastic property. For App. A.1 and App. A.2, consider $1/(d_i+1)$. It can also serve as the Metropolis weight. Even if the weights are fixed, zero imputation for missing coordinates *discards* the weights for those coordinates.

Proof. Taking the expectation of Eq.14 under homogeneous reception, where $\mathbb{E}_{\mathcal{S}}[m_{ji}] = q_{ji}$ gives

$$\mathbb{E}_{\mathcal{S}}\big[\mathbf{x}_i^{+}\big] = W_{ii}\,\mathbf{x}_i + q\sum_{j\in\mathcal{N}_i} W_{ij}\,\mathbf{x}_j. \tag{15}$$

The total weight applied to values that actually exist is $W_{ii} + q\sum_j W_{ij}$. Since the initial matrix preserves row-stochasticity,

$$\sum_{j\in\mathcal{N}_i\cup\{i\}} W_{ij} = W_{ii} + \sum_{j\in\mathcal{N}_i} W_{ij} = 1, \quad W_{ii} = 1 - \sum_{j\in\mathcal{N}_i} W_{ij} \tag{16}$$

and define $\rho(q) = W_{ii} + q\sum_{j\in\mathcal{N}_i} W_{ij}$ as the total weight component of the aggregation step formed by ZeroFill. By assuming the initial conditions derived via Eq. 16,

$$\rho(q) = W_{ii} + q\,(1 - W_{ii}), \qquad \rho(q) < 1 \text{ whenever } q < 1 \text{ and } W_{ii} < 1. \tag{17}$$

So, with lossy aggregation, the total weight component for a single aggregation step is less than 1, and, when accumulated over communication rounds, this results in a contraction. For t rounds, Eq. 15 reduces to use total weight component and its expectation $\mathbb{E}_{\mathcal{S}}[\mathbf{x}^{(t)}] \approx \rho(q)^{(t)}\mathbf{x}^{(0)}$, where $\mathbf{x}^{(0)}$ is the initially defined model.

Iterating over t rounds gives $\|\mathbf{x}^{(t)}\| \approx \rho(q)^t\|\mathbf{x}^{(0)}\| \to 0$. Once the parameters vanish, the network emits constant logits, and accuracy falls to chance.

The result is also Lemma 4.2 with $u = 0$. The per-neighbor bias for a coordinate p in Eq. 13 becomes $-\frac{1-q_{ji}}{d_i+1}x_j^{(p)}$, which is proportional to the state rather than being an additive offset. An error proportional to the state therefore acts as a decay rate. □

Remark A.2 (Numerical consequence)**.** At average degree 4, the Metropolis–Hastings self-weight is $W_{ii} = 1/5$, so $\rho(q) = 0.2 + 0.8q$. The number of rounds needed to reach 10^{-6} of the initial scale is $\log(10^{-6})/\log\rho(q)$, which is approximately 166 at 10% loss and 11 at 90% loss. This predicts the step-function behavior of ZeroFill in Tab. 1, namely reference accuracy at $q = 1$ and chance accuracy at every $q < 1$, with no intermediate regime.

A.4 Proof of Proposition 4.4

By App. A.1 following the notations, unbiasedness for lossless neighbor mean T forces the Horvitz–Thompson (HT) coefficients of the aggregator rule $\bar{a} = \frac{1}{q_{ji}\,(d_i+1)}$, producing the estimator for Inverse-Probability-Weighting by reception probability q_{ji} following HT estimator:

$$A_i^{HT} = \frac{1}{d_i+1}\left[\mathbf{x}_i + \sum_{j\in\mathcal{N}_i}\frac{m_{ji}}{q_{ji}}\mathbf{x}_j\right], \qquad \text{let } Z_j = \frac{m_{ji}}{q_{ji}}\mathbf{x}_j. \tag{18}$$

Proof. Following the property $\mathbb{E}_{\mathcal{S}}[m_{ji}] = q_{ji}$, the estimator satisfies $\mathbb{E}_{\mathcal{S}}[A_i^{HT}] = T$. Thus, unbiasedness appears as:

$$\mathbb{E}_{\mathcal{S}}[Z_j] = \frac{\mathbb{E}_{\mathcal{S}}[m_{ji}]}{q_{ji}}\mathbf{x}_j \;=\; \mathbf{x}_j. \tag{19}$$

For the second moment of the estimator, we consider its squared Euclidean norm. Since $m_{ji} \in \{0,1\}$ is a binary indicator, $m_{ji}^2 = m_{ji}$, and hence

$$\begin{aligned}
\mathrm{Var}_{\mathcal{S}}(Z_j) &= \mathbb{E}_{\mathcal{S}}\left[\|Z_j\|_2^2\right] \;-\; \|\mathbb{E}_{\mathcal{S}}[Z_j]\|_2^2 \\
&= \frac{\mathbb{E}_{\mathcal{S}}[m_{ji}^2]\,\|\mathbf{x}_j\|^2}{q_{ji}^2} \;-\; \|\mathbf{x}_j\|^2 \\
&= \frac{\|\mathbf{x}_j\|^2}{q_{ji}} \;-\; \|\mathbf{x}_j\|^2 \\
&= \frac{(1-q_{ji})}{q_{ji}}\,\|\mathbf{x}_j\|^2,
\end{aligned} \tag{20}$$

Its dependence on the reception probability is $\Theta((1-q_{ji})/q_{ji})$ as $q \to 0$ at fixed parameter magnitude. It is coupled to that magnitude rather than to the step size η. Decreasing the step size, therefore, cannot suppress the variance, since $\eta \to 0$ shrinks the local contribution while leaving the aggregation noise untouched. This coupling is what Def. 5.1 uses to separate the two families, and it explains why no learning-rate schedule rescues the unnormalized family. □

A.5 Proof of Proposition 4.5

Proof. Retaining the notation of App. A.1, fix a node i and a coordinate p, and consider the homogeneous reception setting $q_{ji} = q$.

By App. A.4, unbiasedness for T forces the use of Horvitz–Thompson coefficients, which assign a weight of $1/\big(q(d_i+1)\big)$ to each received neighbor and $1/(d_i+1)$ to the receiver's own value. Thus, each received neighbor carries a relative weight of $1/q > 1$ compared with the receiver's own value.

Summing the coefficients actually applied in a given round gives the realized row sum

$$\begin{aligned}\Sigma_i^{(p)} &= \frac{1}{d_i+1}\left(1+\sum_{j\in\mathcal{N}_i}\frac{m_{ji}^{(p)}}{q}\right)\\ &= \frac{1}{d_i+1}\left(1+\frac{K_i^{(p)}}{q}\right), \qquad K_i^{(p)} = \sum_{j\in\mathcal{N}_i} m_{ji}^{(p)}.\end{aligned} \tag{21}$$

Only received neighbors contribute to the sum, since $m_{ji}^{(p)} = 0$ for missing messages.

Because the arrival indicators are independent Bernoulli(q) random variables, we have $K_i^{(p)} \sim \mathrm{Binomial}(d_i, q)$ and $\mathbb{E}_{\mathcal{S}}[K_i^{(p)}] = d_i q$. Therefore,

$$\mathbb{E}_{\mathcal{S}}\left[\Sigma_i^{(p)}\right] = \frac{1}{d_i+1}\left(1+\frac{d_i q}{q}\right) = \frac{1+d_i}{d_i+1} = 1. \tag{22}$$

Hence, the operator is row-stochastic in expectation, which is what ensures unbiasedness.

However, Prop. 5.3 requires the weights to sum to one for *every* arrival pattern, rather than only in expectation. From Eq. 21, we have $\Sigma_i^{(p)} > 1$ precisely when $K_i^{(p)} > q d_i$. This event has positive probability, since

$$\Pr\left[K_i^{(p)} > q d_i\right] \geq \Pr\left[K_i^{(p)} = d_i\right] = q^{\,d_i} > 0, \qquad \text{for every } q \in (0,1), \tag{23}$$

where the inequality follows from $d_i > q d_i$ whenever $q < 1$.

On any realization for which $\Sigma_i^{(p)} > 1$, the update is not a convex combination of the existing values. Consequently, the bound Eq. 26 used in Prop. 5.3 no longer holds, and the coordinate range is not controlled path wise. The variance statement follows from Prop. 4.4, whose per-neighbor variance $\frac{1-q}{q}\|x_j^{(p)}\|^2$ is inherited by the full estimator when the contributions from independent neighbors are summed. □

A.6 Proof of Lemma 5.2 and Proposition 5.3

Proof of Lemma 5.2. Retaining the notation of App. A.1, fix a node i and a coordinate p. By Eq. 2, node i averages $K_i^{(p)} + 1$ values. They are its own value and those of the $K_i^{(p)}$ neighbors that delivered coordinate p, each with equal weight $1/(K_i^{(p)}+1)$. All other entries in row i are zero. The weights are therefore non-negative, and there are exactly $K_i^{(p)} + 1$ non-zero weights. Hence, their sum is

$$(K_i^{(p)}+1)\frac{1}{K_i^{(p)}+1} = 1. \tag{24}$$

Therefore, this holds for every node and every arrival realization, and $\|W^{(p)}\|_\infty = 1$ path-wise. □

Proof of Proposition 5.3. Define M and m globally over the entire network graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ for any $j \in \mathcal{V} = \mathcal{N}_i \cup \{i\}$ for any node i:

$$M := \max_{j \in \mathcal{V}} x_j^{(p)}, \qquad m := \min_{j \in \mathcal{V}} x_j^{(p)}. \tag{25}$$

By Lem. 5.2 for a fix a node i, the updated value is a convex combination of existing values, so

$$\begin{aligned} x_i^{(p),+} &= \sum_{j \in \mathcal{N}_i \cup \{i\}} \left[W^{(p)}\right]_{ij} x_j^{(p)} \\ &\leq \sum_{j \in \mathcal{N}_i \cup \{i\}} \left[W^{(p)}\right]_{ij} M \\ &= M \sum_{j \in \mathcal{N}_i \cup \{i\}} \left[W^{(p)}\right]_{ij} \\ &= M, \end{aligned} \tag{26}$$

where the inequality uses non-negativity of the weights and the final equality uses row stochasticity and similarly,

$$\begin{aligned} x_i^{(p),+} &= \sum_{j \in \mathcal{N}_i \cup \{i\}} \left[W^{(p)}\right]_{ij} x_j^{(p)} \\ &\geq \sum_{j \in \mathcal{N}_i \cup \{i\}} \left[W^{(p)}\right]_{ij} m \\ &= m \sum_{j \in \mathcal{N}_i \cup \{i\}} \left[W^{(p)}\right]_{ij} \\ &= m. \end{aligned} \tag{27}$$

Therefore, for every $i \in \mathcal{V}$,

$$x_i^{(p),+} \leq M, \qquad x_i^{(p),+} \geq m, \tag{28}$$

and hence

$$\max_{i \in \mathcal{V}} x_i^{(p),+} \leq M, \qquad \min_{i \in \mathcal{V}} x_i^{(p),+} \geq m. \tag{29}$$

□

Remark A.3 (Structural counterpart of the unnormalized family)**.** The proof uses only two facts: the weights are non-negative, and they sum to one. The unnormalized family violates the second condition in either direction. Inverse-probability correction exceeds one with positive probability Eq. 23, while zero-fill falls below one Cor. 4.3. Thus, in both cases, the bound Eq. 26 fails, and the coordinate range is no longer controlled. This is the precise sense in which one family cannot amplify disagreement while the other can. Since Eq. 3 holds for every realization, it also holds under every learning-rate schedule.

A.7 Proof of Proposition 5.4

Proof. Retaining the notation of App. A.1, fix a node i and a coordinate p, and consider the homogeneous reception setting $q_{ji} = q$. Since $\alpha_i^{(p)} = \frac{K_i^{(p)}}{K_i^{(p)}+1} = 1 - \frac{1}{K_i^{(p)}+1}$, let us take the expectation $\mathbb{E}_{\mathcal{S}}\left[\frac{1}{K_i^{(p)}+1}\right]$ for $K_i^{(p)} \sim \text{Binomial}\,(d_i, q)$.

$$f(k, d_i, q) = \Pr\left[K_i^{(p)} = k\right] = \binom{d_i}{k} q^k (1-q)^{d_i - k}, \qquad \text{for } k = 0, 1, 2, ..., d_i. \tag{30}$$

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\frac{1}{K_i^{(p)}+1}\right] &= \sum_{k=0}^{d_i} \frac{1}{k+1} \Pr\left[K_i^{(p)}=k\right] \\
&= \sum_{k=0}^{d_i} \frac{1}{k+1}\binom{d_i}{k} q^k (1-q)^{d_i-k} \\
&= \sum_{k=0}^{d_i} \frac{1}{k+1}\cdot\frac{d_i+1}{d_i+1}\cdot\frac{d_i!}{k!\,(d_i-k)!}\, q^k (1-q)^{d_i-k} \\
&= \sum_{k=0}^{d_i} \frac{1}{d_i+1}\cdot\frac{(d_i+1)!}{(k+1)!\,(d_i-k)!}\, q^k (1-q)^{d_i-k} \\
&= \sum_{k=0}^{d_i} \frac{1}{d_i+1}\binom{d_i+1}{k+1} q^k (1-q)^{d_i-k} \\
&= \sum_{k=0}^{d_i} \frac{1}{d_i+1}\binom{d_i+1}{k+1} \frac{q^{k+1}}{q} (1-q)^{(d_i+1)-(k+1)} \\
&= \frac{1}{q(d_i+1)} \sum_{k=0}^{d_i} \binom{d_i+1}{k+1} q^{k+1} (1-q)^{(d_i+1)-(k+1)}
\end{aligned}
\tag{31}
$$

Re-indexing Eq. 31 with $j = k+1$, therefore $J_i^{(p)} \sim \text{Binomial}\,(d_i+1, q)$ for $1 \le j \le d_i+1$

$$
\mathbb{E}_{\mathcal{S}}\left[\frac{1}{K_i^{(p)}+1}\right] = \frac{1}{q(d_i+1)} \sum_{j=1}^{d_i+1} \binom{d_i+1}{j} q^j (1-q)^{d_i+1-j} \tag{32}
$$

Using the property of $\sum_{j=0}^{d_i+1} \Pr\left[J_i^{(p)}=j\right] = 1$ and subtracting the omitted the zeroth term $j=0$: $\binom{d_i+1}{0} q^0 (1-q)^{d_i+1}$

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\frac{1}{K_i^{(p)}+1}\right] &= \frac{1}{q(d_i+1)}\left[1 - \binom{d_i+1}{0} q^0 (1-q)^{d_i+1}\right] \\
&= \frac{1}{q(d_i+1)}\left[1-(1-q)^{d_i+1}\right] \\
&= \frac{1-(1-q)^{d_i+1}}{q(d_i+1)}, \qquad \mathbb{E}_{\mathcal{S}}[\alpha_i^{(p)}] = 1 - \frac{1-(1-q)^{d_i+1}}{q(d_i+1)}.
\end{aligned}
\tag{33}
$$

□

Table 2: Effective mixing under realized-coordinate normalization at average degree $d = 4$, computed from Eq. 4. As reception becomes less reliable, more weight is retained locally, reaching 0.819 at 90% loss.

Loss	0%	10%	20%	30%	50%	70%	90%
Reception probability q	1.0	0.9	0.8	0.7	0.5	0.3	0.1
Neighbor weight $\mathbb{E}[\alpha]$	0.800	0.778	0.750	0.715	0.613	0.445	0.181
Self-weight $1-\mathbb{E}[\alpha]$	**0.200**	**0.222**	**0.250**	**0.285**	**0.387**	**0.555**	**0.819**

A.8 Proof of the coverage law (Remark 5.5)

Proof. Retaining the notation of App. A.1, fix a node i and a coordinate p, the coverage failure at coordinate p is the event $\{m_{ji}^{(p)} = 0 \text{ for all } j \in \mathcal{N}_i\}$ for an idealized $\bar{d}$-regular proxy graph, where every node is assigned the expected degree $\bar{d} = \mathbb{E}[d_i]$. The indicators are independent across links, so

$$\Pr\left[K_i^{(p)} = 0\right] = \prod_{j \in \mathcal{N}_i} (1 - q) = (1 - q)^{d_i}, \tag{34}$$

and mixing occurs on the complement with probability $1 - (1 - q)^{\bar{d}}$. Arrivals are drawn independently in each round, so the mixing events across T rounds are i.i.d. Bernoulli, and their sum is $\text{Binomial}\big(T, 1 - (1 - q)^{\bar{d}}\big)$ with mean $T \cdot \big(1 - (1 - q)^{\bar{d}}\big)$. □

Table 3: Coverage failure at average degree $d = 4$ of the Erdős–Rényi graph over $T = 3000$ rounds. Even at 90% loss, where two thirds of coordinates receive nothing in a given round, each coordinate is still mixed roughly a thousand times over training.

Loss	10%	20%	30%	50%	70%	90%
Per-round coverage failure $(1 - q)^4$	0.0001	0.0016	0.0081	0.0625	0.2401	**0.6561**
Expected mixes over 3000 rounds	3000	2995	2976	2813	2280	**1032**

B Bacground

As discussed in the main paper, partial reception in decentralized learning, where there is no control authority, can be categorized as either deliberate or stochastic partiality. There is a considerable body of research addressing the same issue and offering solutions across various domains. Fig. 3 illustrates the partial-reception phenomenon in decentralized learning for a model with P parameters. It also shares the visual representation of effective mixing in MFG, as discussed in Eq. 2. However, the related work studied in this paper (Tab. 4) highlights the importance of considering faithful aggregation in the problem regime. Later, we discuss the paper's baselines, their scope, and relevance to the problem we are discussing. As mentioned in the paper, although we do not claim superiority in minimal faithful operator (MFG) itself, we highlight the regime in which it is beneficial at no extra cost or knowledge compared to baselines.

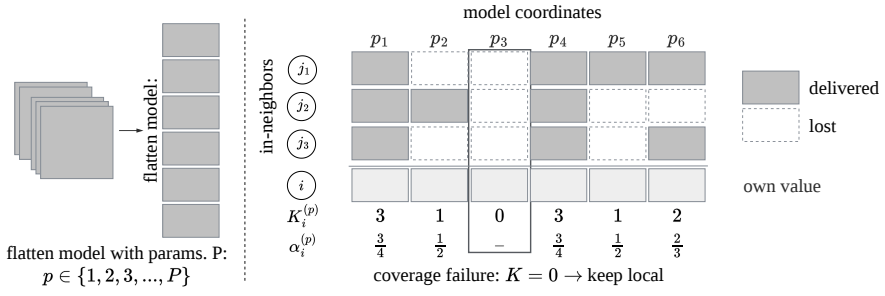


Figure 3: **Partial reception is a per-coordinate event.** One communication round at receiver i with three in-neighbors. Each neighbor model appears as coordinates, and the received subset of coordinates is unobservable in advance. The realized contributor count $K_i^{(p)}$ therefore varies across coordinates within a single round, and with it the effective mixing strength $\alpha_i^{(p)} = K_i^{(p)}/(K_i^{(p)} + 1)$ induced by realized-coordinate normalization.

D-PSGD (Lian et al., 2017) is a foundational paper in decentralized learning that compares decentralized algorithms to centralized algorithms and demonstrates that decentralized methods can

Table 4: Faithful averaging (MFG) is compared with related works in the literature across six factors. The methods are categorized according to the source of partiality.

Method	Partial reception	Decentralized Architecture	Coordination -free	Imputation -free	Non-amplify	Path-wise
Reliable communication						
D-PSGD (Lian et al., 2017)	✗	✓	–	–	✓	✓
CHOCO-SGD (Koloskova et al., 2019)	✗	✓	–	–	✓	✓
Deliberate sparsity (masks chosen and known)						
FedGMR (Ma et al., 2026)	✓	✗	–	✓	✗	✗
FedSpaLLM (ℓ_0) (Bai et al., 2025)	✓	✗	–	✓	✗	✗
DisPFL (Dai et al., 2022)	✓	✓	–	✓	✗	✗
PaME (Sha et al., 2026)[a]	✓	✓	–	✓	✗	✗
Stochastic partial delivery (masks imposed by channel)						
ZeroFill (Li et al., 2024)	✓	✓	✓	✗	✗	✗
Soft-DSGD (Ye et al., 2022)[b]	✓	✓	✓	✗	✓	✗
EDEN (Vargaftik et al., 2022)[c]	∼	✗	✗	✗	✗	✗
HT / IPW gossip[d]	✓	✓	✗	✗	✗	✗
DFL-AA (Kaluannakkage & Buyya, 2026)	✓	✓	✗	✗	✓	✗
MFG (Faithful averaging)	✓	✓	✓	✓	✓	✓

[a] Sender-selected sparse coordinate exchange is implemented using deliberate, known masks. This approach offers a privacy and reconstruction-risk analysis and demonstrates linear convergence to a stationary point. However, it does not guarantee consensus. This method is orthogonal to MFG as PaME determines the transmitted information for communication and privacy, while MFG addresses constraints imposed by the communication channel.

[b] The uniform-weight variant is evaluated in this context. The optimal-weight variant requires access to the reception-probability matrix and is therefore not coordination-free. This method is convex by construction and thus non-amplifying, but no guarantee is established for individual realizations.

[c] EDEN operates in a rotated (Hadamard) domain, so individual coordinates cannot be recovered individually. It rescales the entire received vector by completeness rather than handling per-coordinate reception directly. This needs estimation therefore not a coordination-free baseline.

[d] The canonical unnormalized Horvitz-Thompson reweighting is evaluated here as the reference estimator within the correction-family.

✓ = yes / addressed. ✗ = no or not addressed. ∼ = partial or conditional. – = not applicable. **Coordination-free** = requires no knowledge of or estimates for the reception probabilities (Def. 3.2). **Imputation-free** = missing coordinates are not filled with substituted values. **Non-amplify** = the aggregation operator cannot amplify disagreement in any realization, as its per-coordinate weights form a convex combination. **Path-wise** = this non-amplification property is established for every arrival realization (Prop. 5.3) and therefore holds under any learning-rate schedule. The trajectory-level consensus-distance bound remains an open question for all methods in this setting (Rem. 3.1).

achieve equivalent convergence rates while maintaining similar or lower computational complexity. This work serves as a lossless oracle upper bound for mechanisms in which an aggregation rule is implemented in a decentralized setting. Subsequent research, inspired by this work, has focused on optimizing communication in decentralized training, as communication can become a significant bottleneck in peer-to-peer systems. However, these studies have also identified limitations in existing compression and quantization techniques, noting that such methods accumulate errors during training that do not diminish asymptotically. To address this, CHOCO-SGD and CHOCO-GOSSIP (Koloskova et al., 2019) were introduced, maintaining a public vector for each neighbor and updating it using received or locally owned quantized parameters. Both of these approaches assume reliable, full transmission between peers, which is distinct from the focus of the present work. Nevertheless, they remain among the most prominent foundational contributions to decentralized learning. As outlined in the main paper, this study does not primarily address communication efficiency. Therefore, the baselines are limited to partial-model studies.

FedGMR (Ma et al., 2026) addresses challenges in heterogeneous federated learning, particularly under bandwidth constraints. The study considers partial model exchange as a deliberate strategy,

with both client and server aware of the masking process. This approach enables optimal training for heterogeneous clients. The authors introduce a federated learning framework based on Gradual Model Restoring, in which each client model is a submodel of the full model with density p. In each round, a proportion p of parameters is activated, and this density is gradually restored using learning metrics to maximize client utilization while optimizing communication efficiency. The authors also note that classical methods such as FedAvg tend to shrink masked model parameters toward zero. Mask selection is determined by the importance of each parameter, thereby orthogonal to the origin of the mask we are highlighting here. FedSpaLLM (Bai et al., 2025) applies pruning to its local model in the large language model (LLM) setting to generate a binary mask. This mask is determined by the client's pruning criterion, such as weight magnitude, and is communicated to the server. The approach also employs the same count normalization as MFG, so only clients contributing a nonzero value participate in the averaging process. DisPFL (Dai et al., 2022) proposes a decentralized sparse training strategy for personalized federated learning, in which each node maintains a fixed binary mask over the shared dense model, enabling sparse training on the retained parameters. Gradients are also masked, making the backward pass sparse. Their aggregation method is similar to MFG's count normalization, and the mask is applied to the aggregated model. In this method, a new mask is generated for each communication round based on parameter magnitude, consistently retaining highly important parameters. This contrasts with channel-imposed masks, which are stochastic and independent, and do not account for parameter importance. PaME (Sha et al., 2026) addresses communication efficiency issues, accuracy issues, and privacy concerns arising from existing methods that rely heavily on theoretical assumptions. The primary concern is reconstruction risk under partial exchange involving sensitive model parameter communication. Although coordinate selection is random, it is a deliberate and mutually known choice. All these methods utilize partial model parameter exchange, but the mask selection is a deliberate decision between parties, unlike the channel-imposed, unobservable setting emphasized in the main paper. These approaches use only available coordinates rather than imputing missing ones, as the underlying pattern of availability is known. Consequently, these methods are imputation-free, and there is no clear evidence in the literature regarding amplified disagreement studies or their demonstration in a path-wise manner.

A separate line of research examines the channel-imposed loss regime in learning, including the primary baselines discussed in this paper. ZeroFill (Li et al., 2024) serves as a basic baseline, substituting zeros for missing coordinates and employing classical aggregation over reconstructed parameters at the receiver. Because the missing coordinates are unknown, zero imputation alone results in contraction, as established in Cor. 4.3. EDEN (Vargaftik et al., 2022) is a rotation-based distributed mean-estimation method. It applies a Randomized Hadamard Transform to distribute the vector's energy uniformly across coordinates, quantizes the transformed values, and rescales them to produce an unbiased estimate. In the presence of packet loss, the rotation ensures that losses have an approximately uniform effect, allowing EDEN to compensate through inverse-probability rescaling. Consequently, EDEN achieves robustness by uniformly distributing information across coordinates and then correcting for the fraction lost. In this work, communication efficiency is not considered. Therefore, only the transformation and rescaling components of EDEN are adopted to address lossiness. Here, EDEN refers specifically to the transformation and its inverse, with rescaling to handle losses, rather than to the full method as originally proposed. Soft-DSGD (Ye et al., 2022) primarily addresses the problem domain of aggregation under loss in decentralized learning. In this approach, local parameters are used to reconstruct missing values resulting from communication loss. The method employs two strategies for determining mixing weights. The first uses uniform weights across nodes, while the second seeks optimal weights based on knowledge of channel conditions. Consequently, the optimal weighting strategy is not coordination-free. The communication model assumes full broadcast, which limits the evaluation of robustness to changes in network topology. Additionally, the method does not consider disagreement amplification metrics or address the bias introduced by imputation, a limitation shared with other methods in this category. HT/IPW Gossip utilizes canonical unnormalized Horvitz-Thompson weighting to illustrate the behavior of unnormalized rules and to isolate the effect of correction destabilization under IPW weighting, as discussed in Props. 4.4-4.5. DFL-AA (Kaluannakkage & Buyya, 2026) evaluates the bias terms introduced by imputation baselines and considers directed and asynchronous communication models. This method employs a self-normalized aggregator followed by local-fill imputation with staleness handling to mitigate selection bias and staleness effects on aggregators. Although DFL-AA demonstrates strong performance in the evaluated settings, it still requires estimating channel conditions q,

which is obtained via EWMA, and is primarily assessed under single-scheduling. In contrast to these lines of work, which either assume the mask is known (deliberate sparsity) or estimate the reception probabilities to correct for loss (the correction family), we study aggregation under channel-imposed masks that are unobservable in advance, and characterize when the choice of rule matters and which measurement reveals it.

C Experimental Details

C.1 Datasets and partitioning

Evaluation is conducted on CIFAR-10, CIFAR-100, and EMNIST-Balanced (47 classes). Each dataset is partitioned across N = 20 nodes using a symmetric Dirichlet distribution over classes, with a single draw per seed. For each class, mixing proportions $(\mathrm{Dir}(\alpha))$ are sampled, and class indices are split accordingly, ensuring that every node receives at least one sample. The parameter α is set to 0.1 for CIFAR-10 and EMNIST (indicating high heterogeneity) and 0.5 for CIFAR-100. The test set follows the standard split, is shared and identical across all nodes, and reported accuracy is measured on the same held-out data for every method. CIFAR images are augmented with random 32 × 32 cropping, 4-pixel padding, horizontal flipping, and per-channel normalization. EMNIST images are normalized only.

C.2 Models

All models are buffer-free. Each BatchNorm layer is replaced with GroupNorm, eliminating the use of running statistics. As a result, no statistics are gossiped or averaged, and train/test statistic mismatch is avoided under non-IID label skew. This approach isolates the aggregation comparison from potential normalization confounds. CIFAR-10 employs a GroupNorm ResNet-20 architecture (approximately 0.27 million parameters). CIFAR-100 utilizes a CIFAR-adapted GroupNorm ResNet-18 (approximately 11.2 million parameters, 3×3 stride-1 stem, with max-pooling removed). EMNIST is modeled using a two-hidden-layer MLP with 200 units and approximately 0.2 million parameters (Tab. 5). Group counts are selected as the largest divisor of the channel count that is no greater than 32.

Table 5: Models used in the paper - all utilize GroupNorm (buffer-free).

Dataset	Architecture	Params
CIFAR-10	ResNet-20 (GN)	0.27M
EMNIST	MLP (2×200)	0.20M
CIFAR-100	ResNet-18 (GN)	11.2M

C.3 Communication model

The communication graph follows an Erdős–Rényi random structure with an average degree of 4, generated once per seed. Each model is serialized into 1400-byte packets, with each coordinate occupying 4 bytes, which is within the safe UDP payload size. Packets are delivered independently with probability q according to a homogeneous IID Bernoulli process. If a packet is lost, all coordinates within that packet are lost simultaneously, resulting in chunk-level rather than per-coordinate loss. The reported loss rate is $1 - q$. Only learnable parameters (those requiring gradients) are serialized and exchanged. Missing packets are addressed according to the aggregation rule: ZeroFill, local-fill using the receiver's own value, or exclusion from the realized-coordinate average. For ablation studies, experiments are conducted across both ring and fully connected topologies to provide evidence for Prop. 5.4 in App. 10.

C.4 Optimization

Each node executes $k = 3$ local stochastic gradient descent (SGD) steps per round, employing momentum of 0.9, weight decay of 5×10^{-4}, a batch size of 64, and gradient clipping at an ℓ_2 norm of 5.0. Learning rates are optimized separately for each schedule: $\eta = 0.03$ for the constant

schedule and $\eta_0 = 0.1$ for the cosine schedule, with $\eta_{\min} = 10^{-3} \times \eta_0$ and no warmup phase. Cosine annealing sets $T_{\max}$ equal to the training horizon, ensuring both schedules operate for the same number of rounds per dataset: 3000 for CIFAR-10 and CIFAR-100, and 1000 for EMNIST. Cross-entropy loss is consistently applied, with label smoothing set to 0.0.

C.5 Baseline implementations

All methods utilize identical node configurations, optimizers, learning-rate schedules, data partitions, communication graphs, and reception masks within each seed. The sole distinction lies in how the aggregation step processes received packets. **D-PSGD** (Lian et al., 2017) employs Metropolis–Hastings mixing under lossless delivery and serves as the oracle reference. **ZeroFill** (Li et al., 2024) maintains nominal weights and replaces missing coordinates with zeros. **Soft-DSGD** (Ye et al., 2022) substitutes the receiver's own coordinate values for missing entries (local-fill) under the uniform-weight variant. **EDEN** (Vargaftik et al., 2022) rescales the received vector by the inverse realized completeness after applying a randomized Hadamard rotation. **HT-Gossip** applies the Horvitz–Thompson estimator, scaling each received coordinate by $1/\hat{q}$ using a fixed divisor $d_i + 1$ and omitting self-normalization, where $\hat{q}$ is an exponential moving average of observed completeness, clipped to $[0.05, 0.999]$. **DFL-AA** Kaluannakkage & Buyya (2026) weights each arrival by its estimated reception probability and an age-of-information factor, then normalizes by the realized sum of weights. **MFG** averages each coordinate across the receiver and neighbors that successfully delivered it.

Reception masks. For each seed, packet-loss masks are generated from a shared random number generator stream and applied identically across all methods. Consequently, methods evaluated at the same loss level experience identical loss patterns. The seeds used are $42, 52, 62$, and results are reported as the mean $\pm$ standard deviation across the three seeds.

D Additional Results

D.1 Main results for all six loss levels

The main text, Tab. 1, reports four loss levels spanning the transition from mild to severe reception loss. A natural concern is whether this selection flattens the method by omitting levels at which the ordering breaks down. It does not. Across all six levels (Tabs. 6–7) for CIFAR-10 and across 4 loss levels (Tabs. 8–9) for EMNIST, the two-family split is absolute, where every convex rule degrades gracefully, while every unnormalized rule falls to chance as availability drops, on both datasets and under both schedules. The within-family consensus ordering MFG$<$DFL-AA$<$Soft-DSGD likewise holds at every intermediate level, without a single inversion. The intermediate levels also make the onset visible, as the unnormalized family's consensus distance is elevated but stable at a 10–20% loss. It grows by orders of magnitude from 50% onward, tracking the $(1 - q)/q$ variance factor in Prop. 4.4 rather than degrading smoothly.

The accuracy trajectories (Figs. 4 and 6) show that this endpoint behavior is not an artifact of where training happens to stop. The convex rules approach the lossless reference throughout training and maintain their relative order from the early rounds, so the final-round numbers in Tab. 1 reflect a separation that is present throughout the run. The fidelity trajectories (Figs. 5 and 7) make the same point about $Q^{(t)}$ across two datasets in a convex family, with the ordering established within the first few hundred rounds. It never crosses, which justifies reading it from the final round in the main text.

D.2 Ablation study

The four alternative explanations underscore the importance of faithful aggregation in diverse contexts and demonstrate its robustness. These explanations consider factors such as a shared learning rate, the cosine peak value, specific network topology, limited model architecture, and non-IID data-client drift. Each explanation is systematically excluded, thereby outlining the scope of the claim.

The advantage is not a tuning artifact. If MFG's performance advantage were solely due to a favorable learning rate, a per-method sweep would eliminate this difference. However, Fig. 8

Table 6: Final test accuracy (%) on CIFAR-10, reported as mean ± std over three random seeds at each loss level. Shaded rows distinguish the unnormalized family (top) from the convex family (bottom). The highest value in each column is indicated in bold, with ties also bolded.

	Loss rate					
Method	10%	20%	30%	50%	70%	90%
Constant step size						
D-PSGD (lossless)	58.6 ±5.0					
ZeroFill	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0
EDEN	58.5 ±5.1	53.5 ±5.4	48.4 ±7.4	22.3 ±0.3	10.0 ±0.2	10.0 ±0.0
HT-Gossip	60.0 ±2.2	59.2 ±4.7	57.1 ±6.2	47.7 ±3.7	19.9 ±1.9	10.0 ±0.0
Soft-DSGD	58.9 ±4.6	57.5 ±3.7	57.3 ±5.0	54.7 ±5.3	49.4 ±4.4	42.0 ±4.6
DFL-AA	57.5 ±4.1	58.0 ±4.5	57.7 ±3.6	55.2 ±4.5	51.1 ±4.3	43.3 ±3.7
MFG	**60.1 ±2.4**	**60.1 ±3.2**	**60.4 ±3.5**	**60.2 ±4.6**	**59.7 ±6.2**	**51.4 ±5.3**
Cosine annealing						
D-PSGD (lossless)	74.3 ±2.7					
ZeroFill	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0	10.0 ±0.0
EDEN	67.8 ±3.1	57.1 ±2.6	38.5 ±4.0	10.4 ±0.3	10.0 ±0.0	10.0 ±0.0
HT-Gossip	64.2 ±3.9	57.5 ±4.8	40.7 ±11.0	22.2 ±1.9	10.1 ±0.1	10.0 ±0.0
Soft-DSGD	73.3 ±3.5	**74.6 ±3.0**	**75.4 ±3.2**	**75.5 ±4.0**	76.0 ±3.4	73.5 ±4.3
DFL-AA	73.3 ±1.6	73.3 ±2.9	75.0 ±2.2	**75.5 ±3.4**	**76.2 ±2.5**	74.0 ±3.0
MFG	**73.9 ±1.2**	73.2 ±1.3	74.1 ±2.8	74.0 ±2.6	75.6 ±3.0	**76.1 ±2.7**

Table 7: Final consensus distance $D^{(T)}$ on CIFAR-10, reported as mean±std over three seeds, with both values sharing a common exponent. For example, $9.4 \pm 2.2\,\mathrm{e}{-2}$ represents $(9.4 \pm 2.2) \times 10^{-2}$. Bold font indicates the lowest value within the convex family. †Accuracy is at chance level, indicating that the consensus distance reflects a collapsed model rather than faithful aggregation (Para. 7.1). ‡Parameters exceeded the floating-point range.

	Loss rate					
Method	10%	20%	30%	50%	70%	90%
Constant step size						
D-PSGD (lossless)	$5.0 \pm 1.4\,\mathrm{e}{-3}$					
ZeroFill	$9.4 \pm 2.2\,\mathrm{e}{-2}^{\dagger}$	$1.1 \pm 0.3\,\mathrm{e}{-1}^{\dagger}$	$1.0 \pm 0.2\,\mathrm{e}{-1}^{\dagger}$	$8.8 \pm 1.4\,\mathrm{e}{-2}^{\dagger}$	$5.1 \pm 2.3\,\mathrm{e}{-2}^{\dagger}$	$3.6 \pm 2.3\,\mathrm{e}{-2}^{\dagger}$
EDEN	1.0 ± 0.2	3.8 ± 0.5	$1.2 \pm 0.4\,\mathrm{e}1$	$8.3 \pm 2.9\,\mathrm{e}2$	$6.4 \pm 3.2\,\mathrm{e}4^{\dagger}$	—‡
HT-Gossip	6.0 ± 2.0	$1.9 \pm 1.0\,\mathrm{e}1$	$8.4 \pm 5.2\,\mathrm{e}1$	$2.5 \pm 2.0\,\mathrm{e}2$	$1.3 \pm 0.9\,\mathrm{e}3$	$5.0 \pm 8.2\,\mathrm{e}2^{\dagger}$
Soft-DSGD	$7.4 \pm 1.7\,\mathrm{e}{-3}$	$1.1 \pm 0.3\,\mathrm{e}{-2}$	$1.6 \pm 0.3\,\mathrm{e}{-2}$	$3.8 \pm 0.6\,\mathrm{e}{-2}$	$1.0 \pm 0.1\,\mathrm{e}{-1}$	$4.9 \pm 0.6\,\mathrm{e}{-1}$
DFL-AA	$4.9 \pm 1.1\,\mathrm{e}{-3}$	$7.4 \pm 1.1\,\mathrm{e}{-3}$	$1.1 \pm 0.2\,\mathrm{e}{-2}$	$2.6 \pm 0.3\,\mathrm{e}{-2}$	$7.3 \pm 0.8\,\mathrm{e}{-2}$	$4.2 \pm 0.5\,\mathrm{e}{-1}$
MFG	$\mathbf{4.3 \pm 1.0\,e{-3}}$	$\mathbf{5.1 \pm 1.2\,e{-3}}$	$\mathbf{6.2 \pm 1.3\,e{-3}}$	$\mathbf{9.7 \pm 1.9\,e{-3}}$	$\mathbf{2.1 \pm 0.4\,e{-2}}$	$\mathbf{1.1 \pm 0.2\,e{-1}}$
Cosine annealing						
D-PSGD (lossless)	$1.4 \pm 0.4\,\mathrm{e}{-6}$					
ZeroFill	$8.9 \pm 1.0\,\mathrm{e}{-7}^{\dagger}$	$7.5 \pm 1.1\,\mathrm{e}{-7}^{\dagger}$	$6.8 \pm 1.4\,\mathrm{e}{-7}^{\dagger}$	$5.6 \pm 1.2\,\mathrm{e}{-7}^{\dagger}$	$4.3 \pm 1.1\,\mathrm{e}{-7}^{\dagger}$	$3.6 \pm 1.2\,\mathrm{e}{-7}^{\dagger}$
EDEN	$7.0 \pm 1.4\,\mathrm{e}{-1}$	2.6 ± 0.4	8.6 ± 3.5	$6.8 \pm 6.8\,\mathrm{e}2^{\dagger}$	$2.2 \pm 0.8\,\mathrm{e}4^{\dagger}$	—‡
HT-Gossip	5.4 ± 4.2	$1.5 \pm 0.1\,\mathrm{e}1$	$8.0 \pm 4.6\,\mathrm{e}1$	$3.4 \pm 3.4\,\mathrm{e}2$	$4.0 \pm 3.2\,\mathrm{e}2^{\dagger}$	$9.7 \pm 12\,\mathrm{e}{-5}^{\dagger}$
Soft-DSGD	$1.8 \pm 0.3\,\mathrm{e}{-6}$	$2.9 \pm 0.5\,\mathrm{e}{-6}$	$5.1 \pm 0.5\,\mathrm{e}{-6}$	$1.5 \pm 0.2\,\mathrm{e}{-5}$	$5.7 \pm 0.2\,\mathrm{e}{-5}$	$4.7 \pm 0.1\,\mathrm{e}{-4}$
DFL-AA	$1.3 \pm 0.4\,\mathrm{e}{-6}$	$1.9 \pm 0.6\,\mathrm{e}{-6}$	$3.4 \pm 1.0\,\mathrm{e}{-6}$	$9.4 \pm 1.4\,\mathrm{e}{-6}$	$3.6 \pm 0.4\,\mathrm{e}{-5}$	$3.6 \pm 0.2\,\mathrm{e}{-4}$
MFG	$\mathbf{1.2 \pm 0.4\,e{-6}}$	$\mathbf{1.5 \pm 0.5\,e{-6}}$	$\mathbf{1.7 \pm 0.6\,e{-6}}$	$\mathbf{3.1 \pm 1.1\,e{-6}}$	$\mathbf{6.4 \pm 1.2\,e{-6}}$	$\mathbf{4.9 \pm 0.9\,e{-5}}$

Table 8: Final test accuracy (%) on EMNIST, reported as mean ± std over three random seeds at each loss level. Shaded rows distinguish the unnormalized family (top) from the convex family (bottom). The highest value in each column is shown in bold, with ties also bolded.

	Loss rate			
Method	10%	20%	50%	90%
Constant step size				
D-PSGD (lossless)	80.7 ± 1.1			
ZeroFill	32.4 ± 1.7	13.7 ± 1.6	2.1 ± 0.0	2.1 ± 0.0
EDEN	80.6 ± 1.1	80.3 ± 1.2	71.7 ± 4.9	2.2 ± 0.1
HT-Gossip	79.0 ± 1.7	77.0 ± 1.6	62.8 ± 1.9	2.4 ± 0.3
Soft-DSGD	80.0 ± 0.9	79.0 ± 1.1	74.7 ± 0.9	62.1 ± 1.4
DFL-AA	81.1 ± 0.9	80.2 ± 1.0	76.2 ± 1.1	64.0 ± 1.1
MFG	$\mathbf{81.5 \pm 0.7}$	$\mathbf{81.3 \pm 0.7}$	$\mathbf{80.3 \pm 0.7}$	$\mathbf{72.8 \pm 1.7}$
Cosine annealing				
D-PSGD (lossless)	84.8 ± 0.1			
ZeroFill	2.1 ± 0.0	2.1 ± 0.0	2.1 ± 0.0	2.1 ± 0.0
EDEN	83.8 ± 0.3	82.4 ± 0.5	48.0 ± 14.0	2.3 ± 0.1
HT-Gossip	78.4 ± 1.6	71.3 ± 1.9	37.7 ± 1.5	2.1 ± 0.1
Soft-DSGD	84.5 ± 0.1	84.3 ± 0.1	83.6 ± 0.1	81.9 ± 0.3
DFL-AA	84.7 ± 0.2	84.4 ± 0.1	83.9 ± 0.0	82.4 ± 0.1
MFG	$\mathbf{84.8 \pm 0.0}$	$\mathbf{84.7 \pm 0.3}$	$\mathbf{84.5 \pm 0.1}$	$\mathbf{83.9 \pm 0.1}$

Table 9: Final consensus distance $D^{(T)}$ on EMNIST, reported as mean±std over three seeds, with both values sharing a common exponent. For example, $9.4 \pm 2.2\,\mathrm{e}{-2}$ represents $(9.4 \pm 2.2) \times 10^{-2}$. Bold font indicates the lowest value within the convex family. †Accuracy is at chance level, indicating that the consensus distance reflects a collapsed model rather than faithful aggregation (Para. 7.1). ‡Parameters exceeded the floating-point range.

	Loss rate			
Method	10%	20%	50%	90%
Constant step size				
D-PSGD (lossless)	$5.9 \pm 1.9\,\mathrm{e}{-2}$			
ZeroFill	$3.9 \pm 0.9\,\mathrm{e}{-1}$	$4.0 \pm 1.1\,\mathrm{e}{-1}$	$1.9 \pm 0.5\,\mathrm{e}{-1}^{\dagger}$	$9.6 \pm 3.3\,\mathrm{e}{-2}^{\dagger}$
EDEN	3.2 ± 0.3	9.0 ± 1.2	$2.1 \pm 1.0\,\mathrm{e}2$	$3.9 \pm 6.6\,\mathrm{e}24^{\dagger}$
HT-Gossip	6.1 ± 1.2	$1.7 \pm 0.3\,\mathrm{e}1$	$9.4 \pm 2.2\,\mathrm{e}2$	$6.0 \pm 7.7\,\mathrm{e}3^{\dagger}$
Soft-DSGD	$8.2 \pm 1.8\,\mathrm{e}{-2}$	$1.2 \pm 0.2\,\mathrm{e}{-1}$	$3.2 \pm 0.3\,\mathrm{e}{-1}$	2.9 ± 0.3
DFL-AA	$6.4 \pm 1.5\,\mathrm{e}{-2}$	$8.7 \pm 1.7\,\mathrm{e}{-2}$	$2.4 \pm 0.3\,\mathrm{e}{-1}$	2.7 ± 0.3
MFG	$\mathbf{5.7 \pm 1.3\,e{-2}}$	$\mathbf{6.9 \pm 1.4\,e{-2}}$	$\mathbf{1.3 \pm 0.2\,e{-1}}$	$\mathbf{8.7 \pm 1.6\,e{-1}}$
Cosine annealing				
D-PSGD (lossless)	$3.4 \pm 0.8\,\mathrm{e}{-5}$			
ZeroFill	$7.6 \pm 3.2\,\mathrm{e}{-7}^{\dagger}$	$6.1 \pm 2.5\,\mathrm{e}{-7}^{\dagger}$	$4.2 \pm 1.5\,\mathrm{e}{-7}^{\dagger}$	$2.9 \pm 1.3\,\mathrm{e}{-7}^{\dagger}$
EDEN	3.6 ± 0.3	$1.1 \pm 0.2\,\mathrm{e}1$	$3.1 \pm 1.5\,\mathrm{e}2$	$3.0 \pm 5.2\,\mathrm{e}24^{\dagger}$
HT-Gossip	7.0 ± 1.4	$2.0 \pm 0.4\,\mathrm{e}1$	$1.4 \pm 0.1\,\mathrm{e}3$	$2.1 \pm 3.3^{\dagger}$
Soft-DSGD	$5.9 \pm 1.1\,\mathrm{e}{-5}$	$9.9 \pm 1.5\,\mathrm{e}{-5}$	$4.1 \pm 0.3\,\mathrm{e}{-4}$	$8.9 \pm 1.5\,\mathrm{e}{-3}$
DFL-AA	$2.8 \pm 0.7\,\mathrm{e}{-5}$	$5.5 \pm 1.1\,\mathrm{e}{-5}$	$2.7 \pm 0.2\,\mathrm{e}{-4}$	$7.1 \pm 1.0\,\mathrm{e}{-3}$
MFG	$\mathbf{1.8 \pm 0.7\,e{-5}}$	$\mathbf{2.4 \pm 0.8\,e{-5}}$	$\mathbf{5.6 \pm 1.6\,e{-5}}$	$\mathbf{1.1 \pm 0.2\,e{-3}}$

demonstrates that the performance ordering remains consistent across all fixed rates. Consequently, faithful aggregation yields the greatest benefit with a constant step size, independent of the specific rate. This advantage cannot be eliminated by parameter tuning in the presence of partial reception.

The confound at moderate loss is annealing, not the peak value. Under cosine annealing, the convex rules converge in final accuracy, and one might ask whether this reflects the specific peak $\eta_0 = 0.1$ used in the main results. Fig. 9 repeats the comparison for $\eta_0 \in 0.1, 0.03$ and finds the same pattern. Accuracy converges while fidelity remains ordered. The observed convergence is therefore a property of the schedule's late-stage decay, rather than of any particular peak rate.

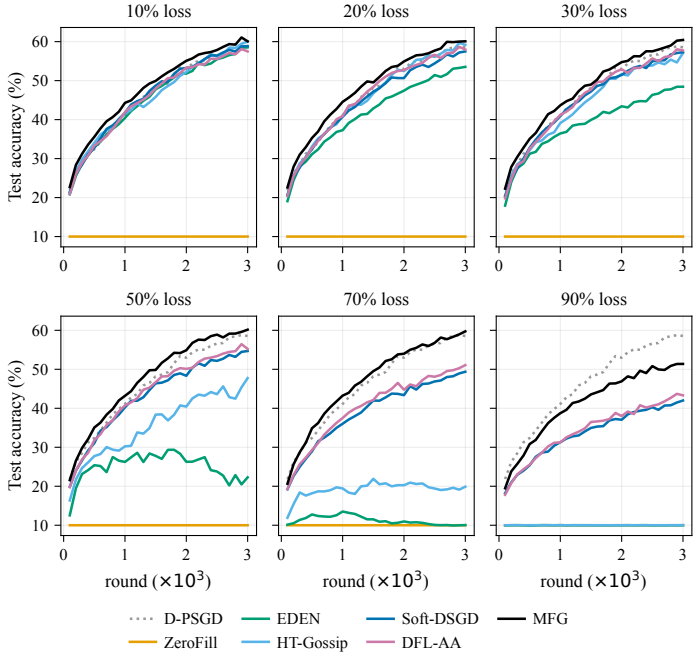
Degree, not loss rate, governs the advantage. The strongest test is topology because it separates the effect of loss from that of connectivity. Fig. 10 shows how the graph changes from a ring to a fully connected graph. On the ring, the convex rules are indistinguishable from one another and from the oracle. With only two neighbors, the topology, rather than the loss, limits mixing, and faithful normalization has little opportunity to concentrate weight over arriving neighbors. The advantage appears at average degree 4 and is largest when the graph is fully connected. This is the mechanism of Rem. 5.5 made empirical. What matters is the expected number of arriving neighbors, $d\,q$, not the loss rate alone, and this bounds the claim. Under a constant rate and sufficient degree, faithful aggregation leads and at degree 2, the advantage vanishes.

The characterization is not specific to the main architectures. Finally, Fig. 11 reruns the constant-rate comparison on CIFAR-100 with ResNet-18 under Dirichlet(0.5) under 70% loss, a harder task on a model roughly $40\times$ larger than the ResNet-20 used for CIFAR-10. The two-family split and the convex ordering carry over unchanged, suggesting that the phenomena we characterize are properties of the aggregation operators rather than of a particular dataset or network.

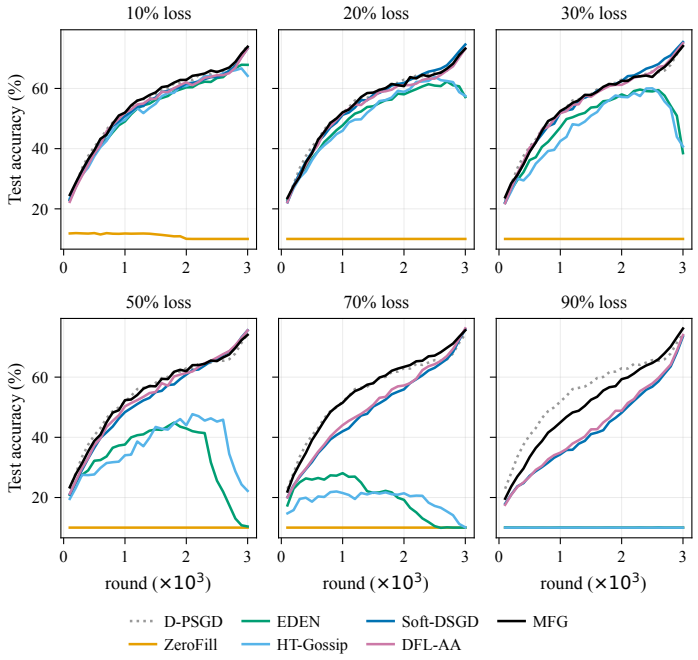
Separation of loss-induced disagreement from client drift Channel-imposed loss of coordinates may also be affected by client drift introduced by the extreme non-IID setup ($\mathrm{Dir}(\alpha) = 0.1$). Figure 12 shows the behavior of the aggregation rules in the near IID setting with $\mathrm{Dir}(\alpha) = 1.0$. The two-family dichotomy persists in this setting, indicating that lossy reception from data distribution statistics does not resolve this issue. The ordering of the convex family rules remains consistent throughout learning, similar to the extreme non-IID case. Final accuracies are higher in the near IID case due to reduced learning difficulty compared to the extreme scenario examined in the main experiments. These results confirm that the findings are robust and not artifacts of client drift. The choice of aggregation rule is significant in the context of partial reception, as well as the metric used to assess robustness.

D.3 Reproducibility statement

Every result in this paper is produced by the DFedNexus event-driven simulator under fixed seeds $42, 52, 62$. The design deliberately removes all sources of variation except the aggregation rule. Within a seed, the data partition, model initialization, communication graph, and per-round reception masks are drawn once and shared across all methods, so two methods at the same loss level see byte-identical loss patterns. Reported figures are the mean and standard deviation over the three seeds. Full dataset, partitioning, model, and optimizer settings are given in App. C, and each baseline's exact update rule, including the estimator clipping and imputation choices that distinguish HT-Gossip, Soft-DSGD, and DFL-AA, is specified in App. C.5. To facilitate reproducibility, the source code, experimental configurations, and supplementary materials are publicly available at `https://zenodo.org/records/22959994`.

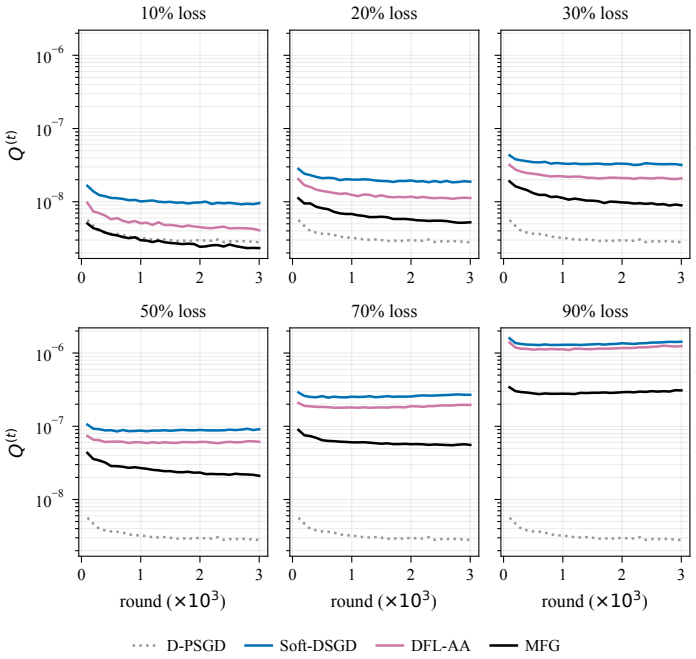


(a) Test accuracy under constant step size

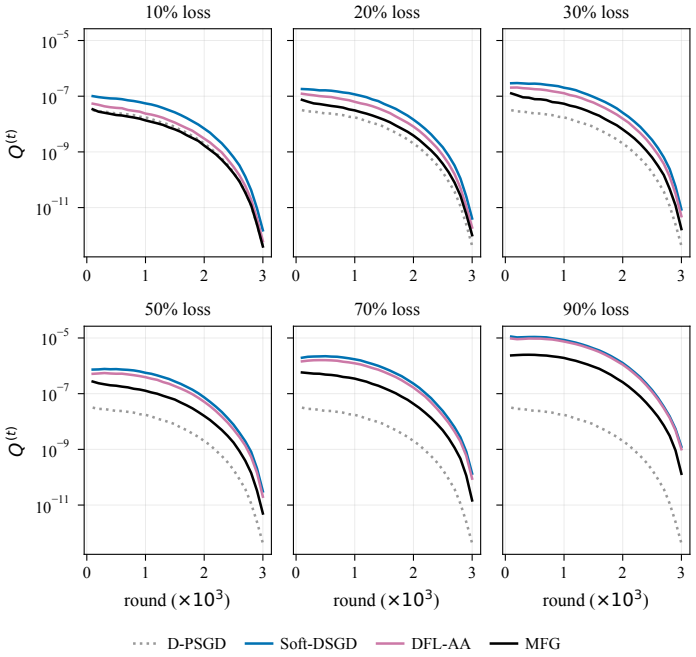


(b) Test accuracy under cosine annealing

Figure 4: Test accuracy over training on CIFAR-10 under both schedules, averaged over three seeds, at every 6 loss levels. (a) All methods are compared with the oracle lossless method, D-PSGD, under a constant step size. (b) All methods under cosine annealing across all 6 loss levels. Curves are evaluated at 100 rounds. Unnormalized methods start to collapse under loss as described in Props. 4.4-4.5 and Cor. 4.3. Fidelity is evident in the trajectories of both schedules, as described in the paper, despite the final accuracies converging, as shown in Fig. 4b.

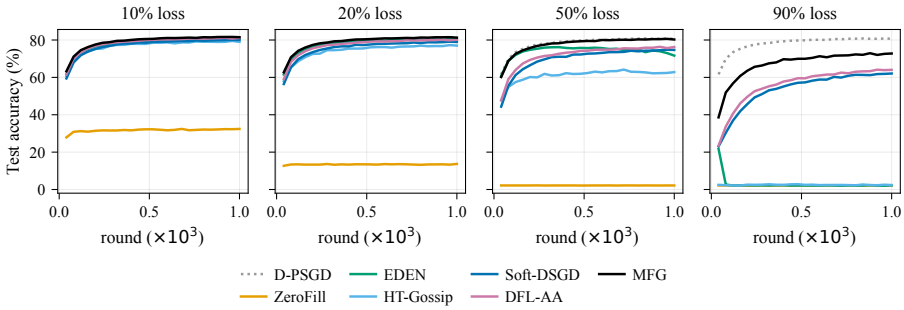


(a) Aggregation fidelity $Q^{(t)}_{(\log-\text{scale})}$ under constant step-size for convex-family

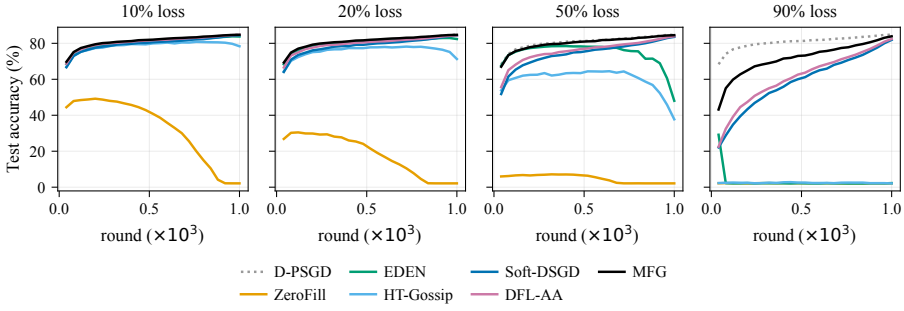


(b) Aggregation fidelity $Q^{(t)}_{(\log-\text{scale})}$ under cosine annealing for convex-family

Figure 5: Aggregation fidelity $Q^{(t)}$ on a log scale over training on CIFAR-10 under both schedules, averaged over three seeds, at every 6 loss levels. (a) All convex-family methods are compared with the oracle lossless method, D-PSGD, under a constant step size. (b) Convex-family members under cosine annealing across all 6 loss levels. Curves are evaluated at 100 rounds. Convergence across the convex family is visible, with $Q^{(t)}$ remaining ordered throughout training.

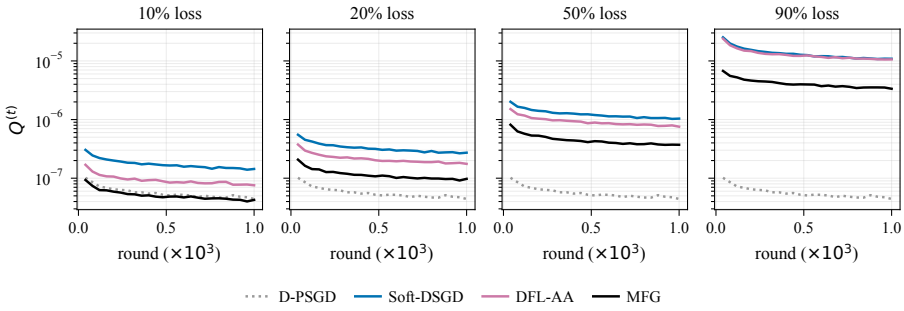


(a) Test accuracy under constant step size

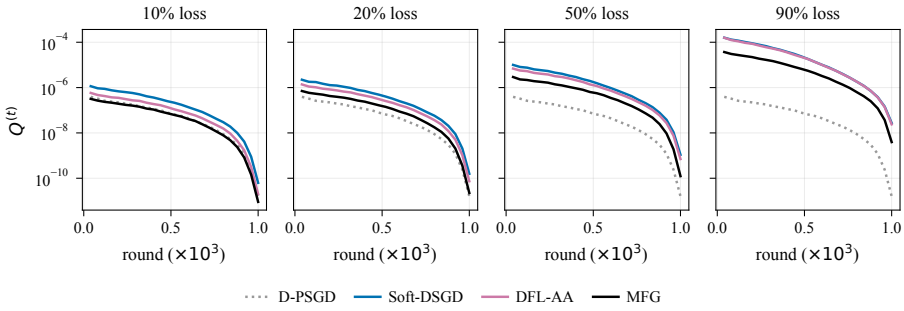


(b) Test accuracy under cosine annealing

Figure 6: Test accuracy during training on EMNIST under both schedules, averaged over three seeds, at four loss levels (10%, 20%, 50%, and 90%). (a) All methods are compared with the oracle lossless method, D-PSGD, under a constant step size. (b) All methods are compared under cosine annealing across 4 loss levels. The runtime is evaluated over 1000 communication rounds. Unnormalized methods start to collapse under loss as described in Props. 4.4-4.5 and Cor. 4.3. A clear separation between convex family members remains visible in both cases, even when final accuracies converge under cosine annealing (Fig. 6b). This maintains a consistent gap between methods at a constant step size as the loss increases (Fig. 6a), despite different datasets and model architectures, thereby proving the claims in the paper.

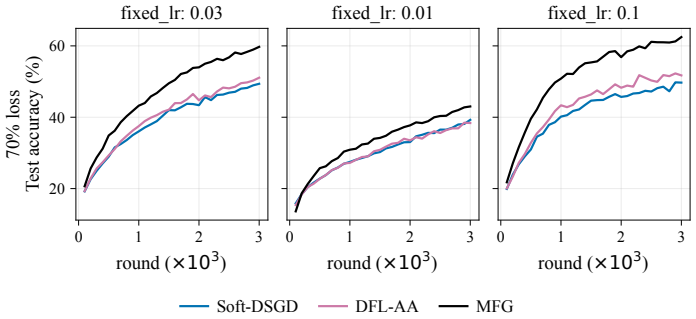


(a) Aggregation fidelity $Q^{(t)}_{(\log-\text{scale})}$ under constant step size for convex family

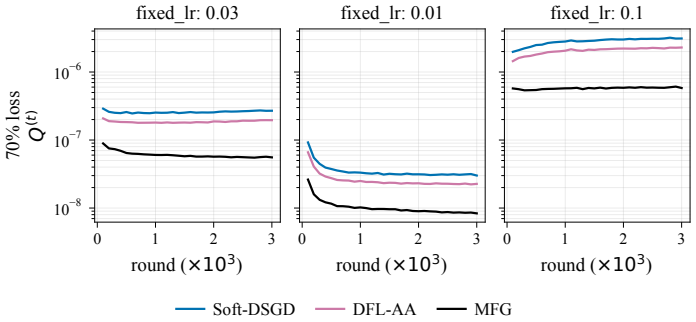


(b) Aggregation fidelity $Q^{(t)}_{(\log-\text{scale})}$ under cosine annealing for convex family

Figure 7: Aggregation fidelity $Q^{(t)}$ on a log scale during training on EMNIST under both schedules, averaged over three seeds and across four loss levels (10%, 20%, 50%, and 90%). (a) All convex-family methods are compared with the oracle lossless method, D-PSGD, using a constant step size. (b) Convex family members are evaluated using cosine annealing across 4 loss levels. The runtime is evaluated over 1000 communication rounds. The advantage of the faithful rule over other family members is confirmed by the observed fidelity within the convex family for both schedules.

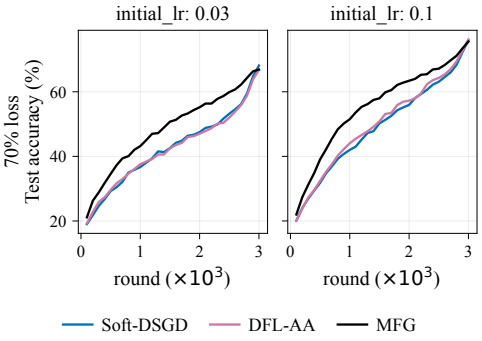


(a) Test accuracy under different fixed step sizes for convex family

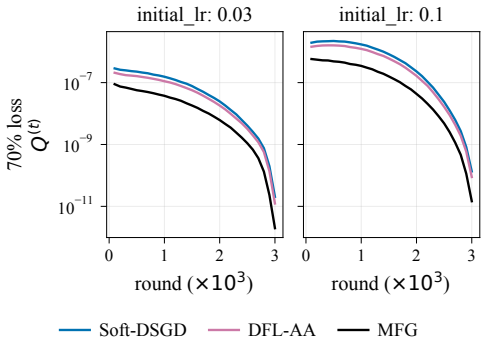


(b) Aggregation fidelity $Q^{(t)}_{(\log-\text{scale})}$ under different fixed step sizes for convex family

Figure 8: Learning-rate control at 70% loss on CIFAR-10 with a constant step size, averaged over three seeds. (a) Test accuracy and (b) aggregation fidelity $Q^{(t)}$ on a log scale at fixed rates $\eta \in 0.01, 0.03, 0.1$. MFG's advantage over the other convex rules persists at every rate, confirming that the constant-rate result is not an artifact of a single shared learning rate.



(a) Test accuracy for convex family under different peak η_0 under cosine annealing



(b) Aggregation fidelity $Q^{(t)}_{(\log-\text{scale})}$ for convex family under different peak η_0 under cosine annealing

Figure 9: Cosine schedule at peak rates $\eta_0 \in \{0.1, 0.03\}$, with a 70% loss on CIFAR-10, averaged over three seeds. (a) Test accuracy and (b) aggregation fidelity $Q^{(t)}$ on a log scale. The convex rules converge to similar final accuracies while remaining ordered in fidelity throughout training, showing that the accuracy–fidelity separation is not specific to the $\eta_0 = 0.1$ peak rate used in the main results.

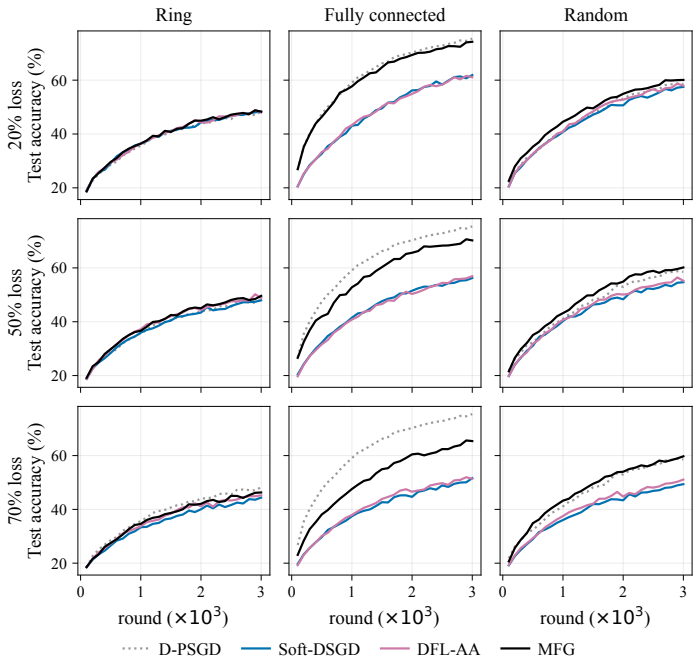
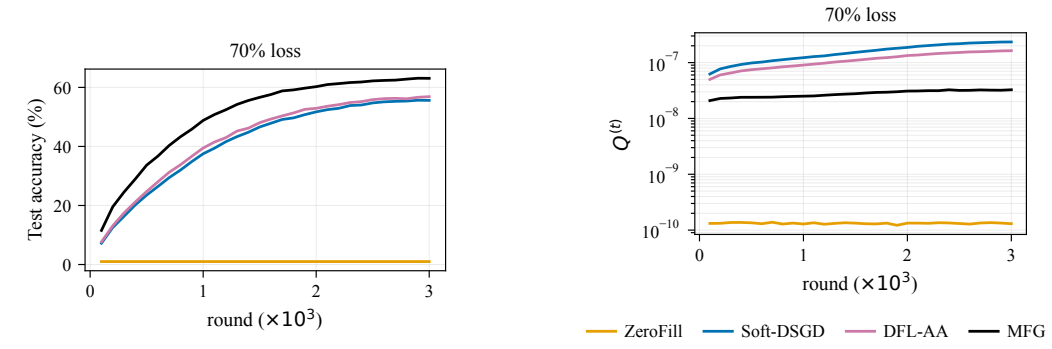


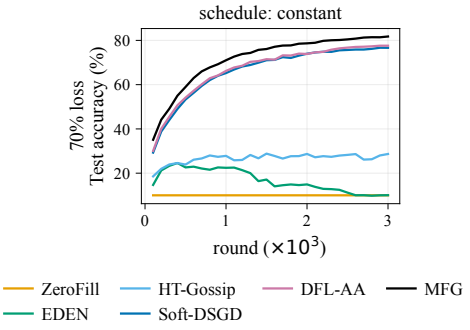
Figure 10: Effect of topology on CIFAR-10 under a constant step size, averaged over three seeds. Test accuracy over training on a ring (degree 2), a random Erdős–R'enyi graph (average degree 4), and a fully connected graph (degree 19), at 20%, 50%, and 70% loss, for the convex rules compared with the lossless D-PSGD reference. As the degree increases, MFG separates from the other convex rules and remains closest to D-PSGD, with the gap widening on the fully connected graph. This matches Rem. 5.5: the advantage of faithful normalization is governed by the expected number of arriving neighbors ($d\,q$), rather than by the loss rate alone.



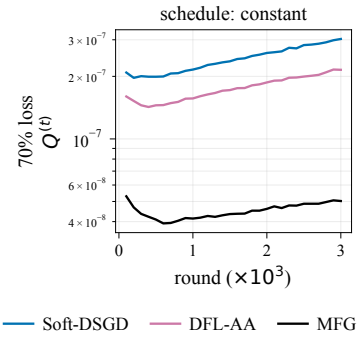
(a) Test accuracy under constant step size

(b) Aggregation fidelity $Q^{(t)}_{(\log-\mathrm{scale})}$ under constant step size

Figure 11: CIFAR-100 at 70% loss with a constant step size, averaged over three seeds. (a) Test accuracy: ZeroFill remains at chance level, while the convex rules learn, with MFG achieving the highest accuracy. (b) Aggregation fidelity $Q^{(t)}$ (log scale): the convex rules remain ordered as MFG $<$ DFL-AA $<$ Soft-DSGD throughout training. At the same time, ZeroFill's low value reflects a collapsed model rather than faithful aggregation.



(a) Test accuracy under near IID data distribution: $\mathrm{Dir}(\alpha) = 1.0$



(b) Aggregation fidelity $Q^{(t)}_{(\log-\mathrm{scale})}$ for convex family under near IID data distribution: $\mathrm{Dir}(\alpha) = 1.0$

Figure 12: Evaluation is conducted under a near IID data distribution ($\mathrm{Dir}(\alpha) = 1.0$) across constant step size to isolate the effect of loss-induced disagreement from client drifts introduced by an extreme non-IID data distribution under 70% loss on CIFAR-10, averaged over three seeds. (a) Test accuracy and (b) aggregation fidelity $Q^{(t)}$ are reported on a log scale. The separation between the two method families is evident even under near IID conditions, and the ordering of convex family members observed under non-IID settings is also present.