EFFICIENT PRIVATE FEDERATED LEARNING WITH ERROR FEEDBACK UNDER RANDOM ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient sparsification is widely used in differentially private federated learning to reduce communication cost. However, sparsification drops coordinates that may carry useful gradient information. Error feedback can recover this information by storing the dropped coordinates and reusing them later, but its effectiveness is limited by how often a client participates: under partial client participation, each client is selected infrequently, so the dropped coordinates become stale and rarely reintroduced. Under differential privacy two further problems arise: adding the stored coordinates back inflates the update norm before clipping, and storing them after the noise carries that noise into later rounds. In this paper, we propose Clean Error Residual Feedback (CERF), where each client stores the coordinates dropped by sparsification from its clipped model update before noise is added, overwrites them at each participation, and feeds them back only at the coordinates the current mask keeps. To tackle the problem of infrequent participation, we use random allocation for client selection, which lets each client participate more regularly while still benefiting from the privacy amplification of subsampling. We derive the privacy and convergence analysis of CERF, showing the trade-off between compression, staleness, and noise terms. Empirical experiments show that at a fixed privacy level, there exists a sparsification rate and allocation cycle length where CERF matches or improves the accuracy of DP-FedAvg at a fraction of the communication and energy cost per client.

1 INTRODUCTION

Federated learning (FL) trains a shared model across many clients through a central server, and the clients never share their local data (McMahan et al., 2017). The model updates sent to the server can still be exploited to infer sensitive information about the clients' data (Zhu et al., 2019), so differential privacy (DP) is adopted as the standard defense (McMahan et al., 2018; Hu et al., 2024). The price of this is the model's accuracy, since the noise it adds degrades the learned model. FL also carries a communication cost, since a model update is as large as the model itself, often millions of parameters, and each selected client uploads its update over a wireless link. In cross-device deployments this cost is paid in battery energy, and it grows with the update size and with how often a client participates (Yang et al., 2021; Li et al., 2024).

Gradient sparsification is used to reduce the communication cost by having each client transmit only a subset of its model update coordinates (Hu et al., 2024; Stich et al., 2018; Ozfatura et al., 2021). Under DP it reduces both costs, since only the transmitted coordinates carry noise, so fewer of them means less noise reaching the global model. By dropping coordinates, gradient sparsification introduces a compression error that grows as the update is sparsified more aggressively. Error feedback is introduced to tackle this problem by storing the dropped coordinates and adding them back later (Seide et al., 2014).

However, error feedback faces several obstacles when applied to DP-FL. The first problem is infrequent participation. Under partial client participation, only a fraction of clients is selected in each round, so a client can go many rounds without participating and has no guarantee of when it will be selected again. Its residual was computed at a model the server has since moved far away from, so adding it back can push the model in the wrong direction. This is the staleness problem, and stale residuals are known to degrade the convergence of error-feedback methods (Li & Li, 2023; Greidi

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.

& Cohen, 2024). The DP pipeline itself creates two more. Adding the stored residual to the update inflates its norm, so the clip scales every coordinate down and the round loses signal at the coordinates it actually transmits. Accumulating the residual makes this worse, since it keeps growing each time the client participates and the clipping then binds harder. Storing the residual after the noise is added carries that noise into the next participation, where it accumulates across rounds.

In this paper, we propose Clean Error Residual Feedback (CERF), which recovers part of the gradient information discarded by sparsification while keeping the same privacy guarantee and communication savings. CERF stores the residual from the clipped update before noise is added, so no DP noise is fed back, and overwrites it at each participation instead of accumulating, so the residual does not grow across rounds and inflate the norm before clipping. The residual is reintroduced only at the coordinates the current mask keeps, so the clip eats less into the transmitted signal. CERF selects clients by random allocation, where every client participates once in each cycle of K rounds, which bounds how long a residual can age while keeping the privacy amplification of subsampling. A shorter cycle is not free, since a client that participates more often carries a higher communication cost and spends more energy. Our contributions are:

- We propose CERF, an error feedback mechanism for sparsified DP-FL under client-level DP, combined with random allocation.
- We derive the privacy guarantee of CERF and a convergence bound whose compression-dependent part splits into a staleness term, a discarded-residual term, and a DP noise term.
- We evaluate CERF across sparsification rates and cycle lengths at fixed privacy. The rate that maximizes accuracy falls as the cycle grows. The gap to DP-FedAvg is largest at the longest cycle, so CERF matches it at short cycles and exceeds it at long ones, at a fraction of the communication and energy cost per client.

2 Related Work

Error feedback (EF) is a method that stores the part of the gradient discarded by the compressor and adds it back in a later iteration (Seide et al., 2014). This recovers the convergence rate of uncompressed training even for biased compressors (Stich et al., 2018; Karimireddy et al., 2019; Richtárik et al., 2021). Sparsification is one of these compressors. The residual holds the dropped coordinates and the client accumulates it across iterations, then adds it back to the next update (Basu et al., 2020). Partial participation in FL creates a staleness problem, since the model keeps moving during the rounds when a client is absent, and this slows convergence (Li & Li, 2023; Yang et al., 2025). In (Greidi & Cohen, 2024), a penalty term is added to the local objective that steers each local gradient step toward the global model plus the stored residual, rather than only transmitting the residual later. In (Li & Li, 2023), zeroing residuals that have aged past a threshold is shown to speed up norm convergence empirically. None of these methods operate under DP, where clipping and noise limit how the residual can be reused.

How well error feedback works depends on how often clients participate. In (Cho et al., 2023), a cyclic client participation scheme splits clients into groups that take turns in a fixed order, and in (Malinovsky et al., 2023), a regularized client participation scheme lets each client participate once within a fixed window. Both bound how long a client waits between its turns, which is needed to deal with staleness. In (Feldman & Shenfeld, 2025), random allocation assigns each client to a round of the cycle at random, so the waiting time stays bounded while the scheme still benefits from privacy amplification.

Error feedback has been studied for compression error without DP (Stich et al., 2018; Basu et al., 2020; Li & Li, 2023), and under DP for clipping bias (Zhang et al., 2024; Islamov et al., 2025). In (Zhang et al., 2024; Islamov et al., 2025), error feedback corrects clipping bias with every client participating every round, while CERF recovers the compression error from sparsification under partial participation and client-level DP. In (Zhang et al., 2024), the residual is clipped to a second threshold and added after clipping the gradient, which raises the total amount of noise the algorithm has to inject. CERF adds the residual before the clip, so the sensitivity stays at C and the noise does not change. In exchange the residual shares the clip budget with the gradient, and adding all of it would spend that budget on coordinates the mask drops, so CERF adds the residual only where the current mask keeps it. In (Basu et al., 2020; Stich et al., 2018), the residual accumulates, which

under DP inflates the norm before clipping and costs gradient signal at the transmitted coordinates, so CERF overwrites it instead. Table 1 summarizes these differences.

Table 1: CERF against prior error-feedback methods.

	Error source	Participation	Privacy	Residual update	Feedback
(Stich et al., 2018; Basu et al., 2020)	compression	full	none	accumulate	full
(Li & Li, 2023)	compression	partial	none	accumulate	full
(Islamov et al., 2025; Zhang et al., 2024)	clipping bias	full	record-level	accumulate	full
CERF (ours)	compression	partial	client-level	overwrite	selective

3 System Model and Problem Formulation

3.1 System Model

We consider a federated learning system with n clients and a central server, as shown in Fig. 1. Each client $m \in [n]$ holds a private local dataset $\mathcal{D}_m$. A client is a battery-powered mobile device connected to the server over a wireless channel. The objective is to minimize the global loss:

$$\min_{\theta \in \mathbb{R}^d} f(\theta) := \frac{1}{n} \sum_{m=1}^{n} f_m(\theta) \quad (1)$$

where $f_m(\theta)$ is the local loss function of client m computed on $\mathcal{D}_m$, and θ is the global model.

At each global round t, the server selects a set $\mathcal{W}^t \subset [n]$ of clients by random allocation, r on average, generates a shared random sparsification mask $m^t \in \{0,1\}^d$, and broadcasts θ^t and m^t to all clients in $\mathcal{W}^t$. Each selected client $m \in \mathcal{W}^t$ then performs the following steps.

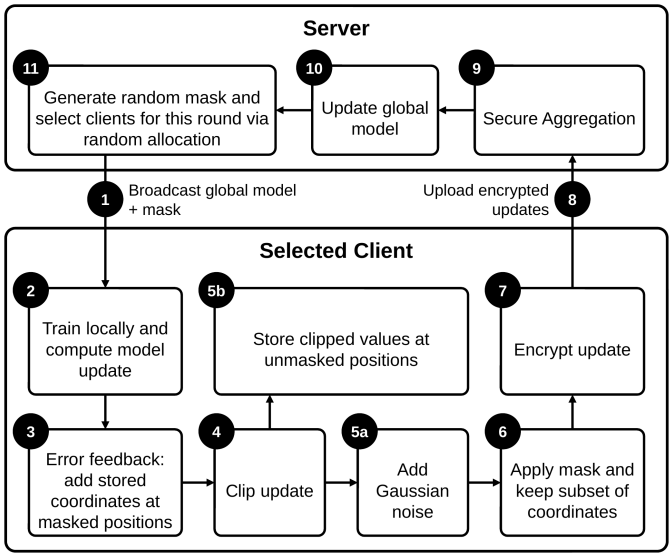


Figure 1: System architecture of CERF.

First, the client updates its local model using gradient descent:

$$\theta_m^{t,s} = \theta_m^{t,s-1} - \eta \nabla f_m(\theta_m^{t,s-1}), \quad s = 1, \ldots, \tau \quad (2)$$

and after τ local epochs computes the model update $g_m^t = \theta_m^{t,\tau} - \theta^t$. The client injects its stored residual from its last participation at the selected coordinates:

$$\hat{g}_m^t = g_m^t + e_m \odot m^t. \quad (3)$$

The update is clipped to norm C and perturbed with Gaussian noise:

$$\bar{g}_m^t = \hat{g}_m^t \cdot \min\left(1, \frac{C}{\|\hat{g}_m^t\|}\right), \qquad \tilde{g}_m^t = \bar{g}_m^t + \mathcal{N}\left(0, \frac{C^2\sigma^2}{r} I_d\right). \quad (4)$$

The shared mask is then applied to retain k out of d coordinates, $s_m^t = \tilde{g}_m^t \odot m^t$, and the client stores the clean residual from the dropped coordinates locally:

$$e_m \leftarrow \bar{g}_m^t \odot (1 - m^t). \quad (5)$$

This residual is never transmitted and is selectively fed back at the client's next participation. Finally, the client encrypts s_m^t and uploads it via secure aggregation (Bonawitz et al., 2017).

The server decrypts only the aggregate and updates the global model:

$$\theta^{t+1} = \theta^t + \frac{1}{r} \sum_{m \in \mathcal{W}^t} s_m^t. \quad (6)$$

3.2 THREAT MODEL

We consider the server to be honest but curious: it follows the protocol but may attempt to infer a client's local data from the updates it receives, for example by gradient inversion (Zhu et al., 2019). A passive eavesdropper on the client-server link could mount the same attack on an intercepted update.

4 PROPOSED SOLUTION

4.1 CERF ALGORITHM

Algorithm 1 CERF Algorithm

Require: n: clients, K: cycle length, T: rounds, τ: local epochs, η: learning rate, C: clipping threshold, σ: noise multiplier, k: coordinates to keep

1: Initialize θ^0; $e_m \leftarrow \mathbf{0}, \forall m \in [n]$
2: **for** $t = 0$ **to** $T - 1$ **do**
  ▷ **Server**
3: **if** $t \bmod K = 0$ **then**
4: $(\mathcal{W}^t, \ldots, \mathcal{W}^{t+K-1}) \leftarrow$ Random allocation
5: **end if**
6: $m^t \leftarrow \texttt{Random-k}(d, k)$
7: Broadcast (θ^t, m^t) to all $m \in \mathcal{W}^t$
  ▷ **Client** $m \in \mathcal{W}^t$
8: $\theta_m^{t,0} \leftarrow \theta^t$
9: **for** $s = 1$ **to** τ **do**
10: $\theta_m^{t,s} \leftarrow \theta_m^{t,s-1} - \eta \nabla f_m(\theta_m^{t,s-1})$
11: **end for**
12: $g_m^t \leftarrow \theta_m^{t,\tau} - \theta^t$
13: $\hat{g}_m^t \leftarrow g_m^t + e_m \odot m^t$
14: $\bar{g}_m^t \leftarrow \hat{g}_m^t \cdot \min(1, C/\|\hat{g}_m^t\|)$
15: $\tilde{g}_m^t \leftarrow \bar{g}_m^t + \mathcal{N}(0, \frac{C^2\sigma^2}{r} I_d)$
16: $s_m^t \leftarrow \tilde{g}_m^t \odot m^t$
17: $e_m \leftarrow \bar{g}_m^t \odot (1 - m^t)$
18: Encrypt s_m^t and upload via secure aggregation
  ▷ **Server aggregation**
19: $\theta^{t+1} \leftarrow \theta^t + \frac{1}{r} \sum_{m \in \mathcal{W}^t} s_m^t$
20: **end for**
21: **return** θ^T

Definition 1 (Random-k Sparsifier)**.** *For $1 \le k \le d$, the random-k sparsifier draws a binary mask $m \in \{0,1\}^d$ with exactly k entries equal to one, chosen uniformly at random from the d coordinates, and keeps only those coordinates:*

$$[\mathrm{rand}_k(x, m)]_j = [x]_j \, m[j] = \begin{cases} [x]_j, & m[j] = 1 \\ 0, & m[j] = 0. \end{cases} \tag{7}$$

Each coordinate is kept with probability $\mathbb{E}[m[j]] = k/d = p$.

We made three specific design choices in CERF. The first is overwriting the residual in line 17 rather than accumulating it. Since the residual is injected before the clip, it shares the clip budget with the gradient, so letting it grow across rounds inflates the norm before clipping, which costs gradient signal at the coordinates sent. Overwriting keeps it bounded, but it also removes the mechanism that top-k relies on, where a small coordinate is eventually sent once its residual has accumulated enough to enter the top-k. This is why we use the random-k mask of Definition 1 in line 6, which changes every round, so any coordinate in the residual has a chance to be sent regardless of its magnitude, and which needs no public dataset, unlike a shared top-k mask (Hu et al., 2024).

The second is selective feedback in line 13, where CERF adds the residual only at the coordinates the current mask keeps, instead of all of it as standard error feedback does. The residual can be as large as C, so adding all of it inflates the norm and the clip in line 14 shrinks every coordinate

harder, costing signal at the coordinates we send. Adding it only where the mask overlaps inflates the norm less, so we recover dropped coordinates while paying less of that clipping penalty.

The third is keeping the residual clean in line 17, where we store it from the clipped update $\bar{g}_m^t$ before the noise, not from the noisy $\tilde{g}_m^t$. Stored after the noise in line 15, the residual would carry that noise on top of the fresh noise at the coordinates it is fed back into.

4.2 Random Allocation

To achieve regular client participation while keeping the privacy amplification of subsampling, CERF uses random allocation for client selection.

Definition 2 (Random Allocation, Feldman & Shenfeld, 2025)**.** *In each cycle of K rounds, every client participates in exactly κ of the K rounds, chosen uniformly at random and independently of all other clients. This is the κ-out-of-K random allocation scheme.*

We assume $\kappa = 1$ throughout, so each client is selected once in every cycle of K rounds, and on average n/K clients participate in each round, giving $r = n/K$. What helps error feedback is the regularity this imposes. Under uniform subsampling a client may not be selected for many rounds, and its residual becomes stale. Random allocation bounds the gap between a client's participations at $2K - 1$ rounds, and a smaller K shortens it further, so residuals are fed back more often and stay fresher.

5 Theoretical Analysis

5.1 Privacy Guarantee

In this subsection we provide the privacy guarantee of CERF under client-level DP. The relevant definitions and lemmas are given in Appendix C.

Theorem 1 (Privacy Guarantee of CERF)**.** *Under $\kappa = 1$ random allocation with cycle length $K \geq 5$, $r = n/K$ clients per round, and $\sigma \geq 4$, for any $\delta \in (0, 1)$ and $\epsilon > 0$, CERF satisfies (ϵ, δ)-DP after T communication rounds if*

$$\sigma^2 \geq \frac{4T(\epsilon + 2\log(1/\delta))}{K^2\epsilon^2}. \tag{8}$$

The proof is given in Appendix C.

Remark 1. *Theorem 1 is not tight, since it uses a closed-form bound on Poisson subsampling at rate $1/K$. The noise multipliers used in our experiments come from the numerical accountant of (Feldman & Shenfeld, 2025) instead.*

5.2 Signal Preservation Under Selective Feedback

Standard error feedback adds the full residual to the gradient before compression, $\hat{g}_m^t = g_m^t + e_m$. Without clipping this is optimal, since it feeds back the most information. Under DP clipping, a problem arises: adding back residual inflates the input norm and shrinks the useful gradient signal at transmitted coordinates.

Lemma 1 (Clipping signal loss)**.** *Let $g \in \mathbb{R}^d$ with $\|g\| > C$ and $e \in \mathbb{R}^d$ with $\|e\| < \|g\|$. For any coordinate j, the signal loss at that coordinate is*

$$\left| \frac{g[j]C}{\|g\|} - \frac{g[j]C}{\|g + e\|} \right| \leq |g[j]| \cdot \frac{C\|e\|}{\|g\|(\|g\| - \|e\|)}. \tag{9}$$

The proof is given in Appendix D.

Remark 2. *We sum Lemma 1 over the pd transmitted coordinates, where the expectation is taken over the random mask m^t. Standard feedback adds the full residual e_m, with $\|e_m\| \leq C$, so the signal loss is*

$$\Delta_{std} = \sum_{j \in m^t} |g[j]| \cdot \frac{C\|e_m\|}{\|g\|(\|g\| - \|e_m\|)}. \tag{10}$$

Selective feedback adds only $e_m \odot m^t$. Each coordinate has probability $\mathbb{E}[m^t[j]] = p$, so $\mathbb{E}\|e_m \odot m^t\|^2 = p\|e_m\|^2$. The square root is concave on $[0, \infty)$, so Jensen's inequality gives $\mathbb{E}\|e_m \odot m^t\| \leq \sqrt{p}\,\|e_m\|$. Substituting this norm into Lemma 1 and summing over the transmitted coordinates, we get

$$\mathbb{E}[\Delta_{sel}] \leq \sum_{j \in m^t} |g[j]| \cdot \frac{C\sqrt{p}\,\|e_m\|}{\|g\|(\|g\| - \sqrt{p}\,\|e_m\|)}. \tag{11}$$

Since $\sqrt{p}\,\|e_m\| \leq \|e_m\|$ and $\|g\| - \sqrt{p}\,\|e_m\| \geq \|g\| - \|e_m\|$, this yields

$$\mathbb{E}[\Delta_{sel}] \leq \Delta_{std}. \tag{12}$$

Thus, selective feedback costs less signal loss compared to standard feedback.

5.3 Convergence Analysis

Assumption 1 (Smoothness)**.** *Each local objective f_m is L-smooth: $\|\nabla f_m(\theta) - \nabla f_m(\theta')\| \leq L\|\theta - \theta'\|$ for all θ, θ'.*

Assumption 2 (Unbiased Gradient and Bounded Variance)**.** *The stochastic gradient at each client is an unbiased estimator of the local gradient, $\mathbb{E}[g_m(\theta)] = \nabla f_m(\theta)$, with bounded variance $\mathbb{E}\|g_m(\theta) - \nabla f_m(\theta)\|^2 \leq \zeta_m^2$, and we write $\bar{\zeta}^2 := \frac{1}{n}\sum_{m=1}^n \zeta_m^2$.*

Assumption 3 (Bounded Dissimilarity)**.** *There exist constants $\beta^2 \geq 1$, $\varkappa^2 \geq 0$ such that $\frac{1}{n}\sum_{m=1}^n \|\nabla f_m(\theta)\|^2 \leq \beta^2\|\nabla f(\theta)\|^2 + \varkappa^2$.*

Lemma 2 (Bounded Residual)**.** *At each participation, client m computes $\hat{g}_m^t = g_m^t + e_m^{(t)} \odot m^t$, clips it to $\bar{g}_m^t = \hat{g}_m^t \cdot \min(1, C/\|\hat{g}_m^t\|)$, and stores $e_m^{(t+1)} \leftarrow \bar{g}_m^t \odot (1 - m^t)$. Then for every client m and every round,*

$$\|e_m^{(t+1)}\| \leq \|\bar{g}_m^t\| \leq C, \tag{13}$$

and, in expectation over the rand-k mask of Definition 1,

$$\mathbb{E}\|e_m^{(t+1)}\|^2 = (1-p)\,\|\bar{g}_m^t\|^2 \leq (1-p)\,C^2. \tag{14}$$

The proof is given in Appendix E.

Theorem 2 (Convergence of CERF)**.** *Under Assumptions 1–3, suppose $\|\hat{g}_m^t\| \leq C$ so that clipping is inactive, each round selects $r = n/K$ clients uniformly at random from all n clients, and the learning rate satisfies $\eta \leq \min\{\frac{1}{144\tau L\beta^2}, \frac{1}{48\tau\beta^2}, \frac{1}{12\tau L}\}$. Then the sequence of outputs θ^t generated by Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\theta^t)\|^2 \leq \frac{4\big(f(\theta^0) - f^*\big)}{\eta\tau\,T} + 2\eta\tau\big(32\eta\tau L^2 + 6L + 3\big)\varkappa^2 + 4\eta\tau\big(2\eta L^2 + 2L + 1\big)\bar{\zeta}^2$$
$$+ \Big(\frac{2}{\eta\tau} + \frac{1}{2}\Big)L^2K^2(1-p)\,C^2 + \Big(\frac{16}{\eta^2\tau^2} + \frac{4\,L}{\eta\tau}\Big)(1-p)^2\,C^2 + \frac{2\,L\,p\,d\,C^2\sigma^2K^2}{\eta\tau\,n^2}. \tag{15}$$

where f^* is the optimal objective value.

The proof is given in Appendix E.

Remark 3. *Theorem 2 characterizes the compression, staleness, and noise trade-off when clipping is not active, thus it does not model the clipping bias that selective feedback reduces, which Lemma 1 and Remark 2 treat separately. The first three terms are the standard FedAvg errors, with the optimization error vanishing as T grows. The fourth term is the staleness error, which scales with K^2 because a client has to wait a full cycle of K rounds before it participates again. The fifth term is the error from the coordinates that are dropped and never fed back, and it shrinks as p grows. The last term is the DP noise error, linear in p, since fewer transmitted coordinates admit less noise into the model.*

6 Accuracy and Energy Trade-off

We formulate the trade-off between the convergence error and the energy each client spends over the training run, at a fixed privacy level, as problem (16) over the sparsification rate, the cycle length, and the noise multiplier.

$$\mathbf{P}: \quad \min_{p,\,K,\,\sigma} \left\{ \Big(\frac{2}{\eta\tau}+\frac{1}{2}\Big) L^2 K^2 (1-p)\, C^2 + \Big(\frac{16}{\eta^2\tau^2}+\frac{4L}{\eta\tau}\Big)(1-p)^2 C^2 + \frac{2LpdC^2\sigma^2K^2}{\eta\tau\, n^2} \right. \\ \left. + \lambda \max_{m\in[n]} \frac{T}{K}\left[\frac{\chi\,\tau\,|\mathcal{D}_m|\,\Phi}{2}\big(f^{\mathrm{cpu}}\big)^2 + \frac{32\,p\,d\,P^{\mathrm{tx}}}{B\log_2\big(1+P^{\mathrm{tx}}h/(N_0B)\big)} \right] \right\} \tag{16}$$

subject to:

$$\sigma^2 \geq \frac{4T(\epsilon + 2\log(1/\delta))}{K^2\epsilon^2}, \tag{16a}$$

$$\frac{T}{K}\left[\frac{\chi\,\tau\,|\mathcal{D}_m|\,\Phi}{2}\big(f^{\mathrm{cpu}}\big)^2 + \frac{32\,p\,d\,P^{\mathrm{tx}}}{B\log_2\big(1+P^{\mathrm{tx}}h/(N_0B)\big)} \right] \leq E^{\max}, \quad \forall m \in [n], \tag{16b}$$

$$p \in (0,1], \tag{16c}$$

$$K \in \{5, 6, \ldots, n\}. \tag{16d}$$

The first part of the objective is the compression-dependent error of Theorem 2. Constraint (16a) is the noise required by Theorem 1.

The second part of the objective is the energy a client spends over the training. The energy of one participation is the sum of a computation cost and a transmission cost. We assume the clients share the same transmit power and channel gain, and we neglect the co-channel interference between them. Under $\kappa = 1$ random allocation a client participates T/K times over T rounds. Clients can hold local datasets of different sizes, so the objective takes the largest energy over the clients. Constraint (16b) caps this energy at the budget $E^{\max}$ of the client. We do not solve $\mathbf{P}$ directly, but use it to guide a grid search in Section 7.2.2, where we set $\lambda = 0$ and pick the tested (K, p) with the highest accuracy within the budget.

7 Experiments

7.1 Dataset and Experiment Setup

We evaluate CERF on Fashion-MNIST and FEMNIST. For Fashion-MNIST, we follow the setup of (Hu et al., 2024), which uses a CNN with two convolutional layers (32 and 64 channels, 5×5 kernels) and a fully connected layer of 512 units, $n = 6000$ clients, 10 local epochs, batch size 10, learning rate 0.125 with 0.99 decay, momentum 0.5, and clipping threshold $C = 1.0$. We shuffle the training set and split it equally across the n clients, so the partition is IID.

For FEMNIST, a cross-device benchmark used for client-level DP (Chen et al., 2022), we keep the $n = 3400$ writers with at least 30 samples and reserve 10% of each writer's data for testing. The rest of the protocol is unchanged, except that the output layer has 62 units and clients run 1 local epoch, since a FEMNIST writer has 161 samples at the median against 10 on Fashion-MNIST. Each client is one writer, so clients differ in handwriting and in which characters they wrote, which makes the partition non-IID.

We set $\delta = n^{-1.1}$ and calibrate σ per K so that every run reaches the same ϵ after $T = 180$ rounds, using the accountant of (Feldman & Shenfeld, 2025). Fashion-MNIST is run at $\epsilon = 0.7442$ and FEMNIST at $\epsilon = 3.0$.

We measure communication as the bits one client sends on the uplink, where each of the $k = pd$ coordinates kept by the mask is a 32-bit float. The mask is generated at the server and shared, so no coordinate indices are sent, and we ignore the finite precision and modular arithmetic of secure aggregation. We price these bits in the energy model of (16). We set the CPU cycles per sample $\Phi = 2\times10^4$, the CPU frequency $f^{\mathrm{cpu}} = 2$ GHz, the capacitance coefficient $\chi = 10^{-28}$, the transmit power $P^{\mathrm{tx}} = 23$ dBm, the uplink bandwidth $B = 400$ kHz, the channel gain $h = 1.6 \times 10^{-13}$, and the noise power spectral density $N_0 = -174$ dBm/Hz.

7.2 Experimental Results

7.2.1 Accuracy Across Sparsification Rates and Cycle Lengths

Table 2: Best test accuracy (%), mean ± standard deviation over 5 seeds. MB is the cost of one upload.

	Fashion-MNIST				FEMNIST			
p	K=20	K=45	K=90	MB	K=20	K=45	K=90	MB
σ_K	3.226	1.635	1.168		1.132	0.807	0.664	
CERF (with error feedback)								
0.001	28.35 ±6.47	27.31 ±4.67	20.23 ±3.09	0.007	5.12 ±0.38	5.20 ±0.50	5.61 ±0.56	0.007
0.01	65.48 ±2.23	64.05 ±2.03	61.06 ±2.17	0.067	27.46 ±2.04	22.56 ±3.11	19.62 ±5.24	0.068
0.1	78.05 ±0.62	76.98 ±0.23	74.58 ±0.65	0.665	68.86 ±0.34	66.87 ±0.40	58.30 ±0.16	0.676
0.2	79.92 ±0.39	78.79 ±0.57	**75.13 ±0.42**	1.331	75.05 ±0.19	71.98 ±0.27	**61.61 ±0.44**	1.352
0.4	**81.18 ±0.46**	**79.50 ±0.36**	73.47 ±0.61	2.661	78.22 ±0.13	73.86 ±0.24	59.69 ±0.42	2.704
0.8	80.62 ±0.41	79.18 ±0.17	72.35 ±0.61	5.323	**79.77 ±0.17**	**73.99 ±0.23**	50.09 ±2.31	5.408
Without error feedback								
0.001	20.14 ±3.64	19.18 ±4.20	17.14 ±2.81	0.007	4.27 ±0.79	5.04 ±0.54	5.59 ±0.83	0.007
0.01	62.42 ±2.16	61.05 ±1.68	56.51 ±2.11	0.067	16.53 ±4.92	17.85 ±3.49	18.55 ±3.32	0.068
0.1	75.69 ±0.52	75.35 ±0.29	73.46 ±0.69	0.665	64.16 ±0.56	63.70 ±0.60	57.62 ±0.34	0.676
0.2	77.98 ±0.61	77.23 ±0.79	**74.38 ±0.54**	1.331	72.93 ±0.26	70.72 ±0.29	**60.57 ±0.50**	1.352
0.4	79.97 ±0.50	78.62 ±0.43	73.42 ±0.69	2.661	77.35 ±0.04	73.10 ±0.28	58.81 ±0.46	2.704
0.8	**80.40 ±0.43**	**79.09 ±0.25**	72.47 ±0.85	5.323	**79.65 ±0.15**	**73.86 ±0.25**	48.47 ±4.96	5.408
Accumulate, full feedback								
0.001	21.33 ±4.10	20.02 ±4.12	17.63 ±2.76	0.007	5.16 ±0.42	5.17 ±0.59	5.61 ±0.76	0.007
0.01	62.06 ±2.26	60.44 ±1.57	56.64 ±1.83	0.067	20.18 ±4.30	15.61 ±3.77	17.22 ±3.92	0.068
0.1	76.77 ±0.64	75.64 ±0.29	73.53 ±0.73	0.665	64.03 ±0.52	62.99 ±0.61	56.11 ±0.24	0.676
0.2	79.15 ±0.32	77.99 ±0.59	**74.48 ±0.55**	1.331	72.71 ±0.24	70.01 ±0.18	**60.48 ±0.39**	1.352
0.4	**80.83 ±0.54**	79.07 ±0.26	73.51 ±0.65	2.661	77.27 ±0.08	73.05 ±0.30	56.90 ±4.42	2.704
0.8	80.59 ±0.40	**79.11 ±0.19**	72.30 ±1.21	5.323	**79.65 ±0.16**	**73.97 ±0.24**	48.71 ±3.04	5.408
Noisy residual storage								
0.001	15.32 ±2.48	15.73 ±3.10	16.13 ±2.88	0.007	3.32 ±1.91	4.82 ±0.61	4.72 ±1.31	0.007
0.01	22.18 ±1.55	33.51 ±4.45	47.47 ±3.62	0.067	5.54 ±0.47	7.89 ±1.99	12.82 ±4.24	0.068
0.1	57.85 ±4.06	68.08 ±1.15	71.30 ±0.49	0.665	20.74 ±2.55	35.77 ±1.20	43.77 ±1.23	0.676
0.2	66.13 ±1.86	72.29 ±1.05	**73.91 ±0.36**	1.331	34.15 ±0.83	44.25 ±1.06	54.23 ±0.87	1.352
0.4	71.12 ±0.75	75.61 ±0.44	73.48 ±0.55	2.661	40.85 ±1.14	57.87 ±0.38	**57.03 ±0.54**	2.704
0.8	**74.54 ±0.35**	**78.06 ±0.29**	72.75 ±0.71	5.323	**55.36 ±0.90**	**66.96 ±0.60**	52.02 ±3.13	5.408
Baselines (no sparsification)								
DP-FedAvg	80.60 ±0.33	79.01 ±0.23	71.28 ±0.87	6.653	80.01 ±0.15	72.74 ±2.17	48.10 ±2.51	6.760
FedAvg	85.84 ±0.22	85.89 ±0.23	85.77 ±0.17	6.653	84.52 ±0.02	84.62 ±0.11	84.48 ±0.10	6.760

Table 2 reports accuracy across sparsification rates and cycle lengths on both datasets. Within a fixed K, accuracy rises with p up to an optimal point. On Fashion-MNIST the peak is at $p = 0.4$ for $K = 20$ and 45 and at $p = 0.2$ for $K = 90$; on FEMNIST it is at $p = 0.8$ for $K = 20$ and 45 and at $p = 0.2$ for $K = 90$. The peak moves to a smaller p at $K = 90$, since the noise term of Theorem 2 grows with $p\,\sigma_K^2 K^2$ and $\sigma_K^2 K^2$ is largest at $K = 90$ on both datasets, so a smaller p is needed to keep the noise error low. The cycle length also matters on its own. At $K = 45$ and $p = 0.2$ a client sends the same number of coordinates over the run as at $K = 90$ and $p = 0.4$, but reaches 78.79% against 73.47% on Fashion-MNIST and 71.98% against 59.69% on FEMNIST, since the staleness term, which grows with $(1-p)K^2$, and the noise term, which grows with $p\,\sigma_K^2 K^2$, are both larger at $K = 90$ and $p = 0.4$.

DP-FedAvg falls as the cycle grows, since fewer clients take part in each round and the average is noisier. At its best p, CERF beats DP-FedAvg by 0.58, 0.49 and 3.85 points on Fashion-MNIST for $K = 20$, 45 and 90, and on FEMNIST it trails by 0.24 at $K = 20$ and leads by 1.25 and 13.51 at $K = 45$ and 90. DP-FedAvg is CERF at $p = 1$, so it pays the noise error in full, while CERF pays a fraction p of the same noise error and takes on the compression errors instead. This trade barely pays at $K = 20$, where the noise error is small, and pays most at $K = 90$, where it is larger.

The remaining blocks isolate the design choices. Without error feedback, accuracy falls behind CERF at every K and every $p < 0.8$, by 3.06 points on Fashion-MNIST and 10.93 on FEMNIST at $K = 20$ and $p = 0.01$. Accumulating and injecting the full residual loses 3.42 and 7.28 points at the same setting, which is the signal loss of Remark 2. Both gaps close at $p = 0.8$, where few coordinates are dropped. Storing the residual from the noisy update loses 43.30 and 21.92 points at the same setting, and at $p = 0.1$ it is the only arm whose accuracy rises with K, since a larger K has a smaller σ_K and stores less noise.

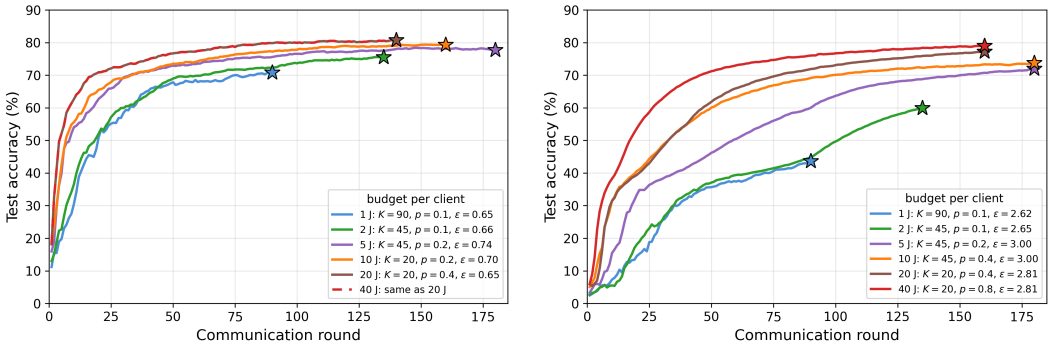
7.2.2 Communication and Energy Cost

Table 2 reports the uplink cost of one upload. Every client sends the same pd coordinates, so the cost is exactly a fraction p of the 6.653 MB that DP-FedAvg sends on Fashion-MNIST and the 6.760 MB on FEMNIST. At the best p for each K, CERF cuts the uplink cost by $2.5\times$ at $K = 20$ and 45 and $5\times$ at $K = 90$ on Fashion-MNIST, and by $1.25\times$ at $K = 20$ and 45 and $5\times$ at $K = 90$ on FEMNIST.

Table 3 shows the most accurate tested configuration under each budget, where a run stops at the end of a cycle since energy and privacy are charged per participation. As the budget grows, the best configuration moves to shorter cycles and larger sparsification rates. At 20 J, CERF reaches 80.72% on Fashion-MNIST for 16.9 J and $\epsilon = 0.65$ with $K = 20$ and $p = 0.4$, and at 40 J it reaches 79.06% on FEMNIST for 39.4 J and $\epsilon = 2.81$ with $K = 20$ and $p = 0.8$. A budget is needed since clients run on batteries, and a weaker device or one with low charge drops out if training drains it. For example, the phones in (Arouj & Abdelmoniem, 2022) hold 3000 to 4000 mAh, about 40 to 53 kJ at 3.7 V, so our budgets of 1 to 40 J use at most 0.1% of the weakest device. A larger model costs more, since transmission energy grows linearly with d, and uploading a model with 1B parameters at $K = 20$ and $p = 0.4$ would alone take 33% of the same battery.

Table 3: Best configuration under each per-client energy budget. ϵ and J are the privacy and energy spent at the stop.

	Fashion-MNIST				FEMNIST			
Budget	(K, p)	acc	ϵ	J	(K, p)	acc	ϵ	J
1	$(90, 0.1)$	71.05	0.65	0.6	$(90, 0.1)$	43.58	2.62	0.6
2	$(45, 0.1)$	75.61	0.66	1.8	$(45, 0.1)$	59.89	2.65	1.9
5	$(45, 0.2)$	78.41	0.74	4.8	$(45, 0.2)$	71.92	3.00	4.9
10	$(20, 0.2)$	79.38	0.70	9.7	$(45, 0.4)$	73.76	3.00	9.8
20	$(20, 0.4)$	80.72	0.65	16.9	$(20, 0.4)$	77.23	2.81	19.7
40	$(20, 0.4)$	80.72	0.65	16.9	$(20, 0.8)$	79.06	2.81	39.4



(a) Fashion-MNIST (b) FEMNIST

Figure 2: Best configuration under each per-client energy budget.

8 Limitations

The allocation schedule assumes every allocated client participates on its assigned round, but in cross-device deployments a client may be offline, slow, or drop mid-round. Its residual then ages by another cycle for every missed turn, and reallocating it within the cycle would break the once-per-cycle property the privacy analysis depends on. The convergence bound also assumes that clipping does not bind, so it does not capture the clipping bias that selective feedback reduces.

9 Conclusion

This paper has proposed CERF, an error feedback mechanism for sparsified federated learning under client-level differential privacy with random allocation, and derived its privacy guarantee and convergence bound. Experiments on Fashion-MNIST and FEMNIST show that CERF matches DP-FedAvg at short cycles and exceeds it at long ones, at a fraction of the communication and energy cost per client. In future work, we will consider asynchronous participation and heterogeneous client-level privacy.

AI USE STATEMENT

In the process of developing this work, we used LLMs as a rewriting tool for the paper and for reviewing and debugging experimental code. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amna Arouj and Ahmed M. Abdelmoniem. Towards energy-aware federated learning on battery-powered clients. In *Proceedings of the 1st ACM Workshop on Data Privacy and Federated Learning Technologies for Mobile Edge Network (FedEdge)*, pp. 7–12, 2022. doi: 10.1145/3556557.3557952.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas N. Diggavi. Qsparse-Local-SGD: Distributed SGD with quantization, sparsification, and local computations. *IEEE Journal on Selected Areas in Information Theory*, 1(1):217–226, 2020.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Wei-Ning Chen, Christopher A. Choquette-Choo, Peter Kairouz, and Ananda Theertha Suresh. The fundamental price of secure aggregation in differentially private federated learning. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3056–3089, 2022.

Yae Jee Cho, Pranay Sharma, Gauri Joshi, Zheng Xu, Satyen Kale, and Tong Zhang. On the convergence of federated averaging with cyclic client participation. In *Proc. 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

Vitaly Feldman and Moshe Shenfeld. Privacy amplification by random allocation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ran Greidi and Kobi Cohen. Sparse training for federated learning with regularized error correction. *IEEE Journal of Selected Topics in Signal Processing*, 18(6):1085–1099, 2024.

Rui Hu, Yuanxiong Guo, and Yanmin Gong. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy. *IEEE Transactions on Mobile Computing*, 23(8):8242–8255, 2024.

Rustem Islamov, Samuel Horváth, Aurelien Lucchi, Peter Richtárik, and Eduard Gorbunov. Double momentum and error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*, 2025.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proc. 36th International Conference on Machine Learning (ICML)*, 2019.

Liang Li, Chenpei Huang, Dian Shi, Hao Wang, Xiangwei Zhou, Minglei Shu, and Miao Pan. Energy and spectrum efficient federated learning via high-precision over-the-air computation. *IEEE Transactions on Wireless Communications*, 23(2):1228–1242, 2024. doi: 10.1109/TWC.2023.3287549.

Xiaoyun Li and Ping Li. Analysis of error feedback in federated non-convex optimization with biased compression: Fast convergence and partial participation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19638–19688. PMLR, 2023.

Grigory Malinovsky, Samuel Horváth, Konstantin Burlachenko, and Peter Richtárik. Federated learning with regularized client participation. *arXiv preprint arXiv:2302.03662*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Emre Ozfatura, Kerem Ozfatura, and Deniz Gündüz. Time-correlated sparsification for communication-efficient federated learning. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 461–466, Melbourne, Australia, 2021.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Diying Yang, Yingwei Hou, and Weigang Wu. Analysis of asynchronous federated learning: Unraveling the interactions between gradient compression, delay, and data heterogeneity. *arXiv preprint arXiv:2504.19903*, 2025.

Zhaohui Yang, Mingzhe Chen, Walid Saad, Choong Seon Hong, and Mohammad Shikh-Bahaei. Energy efficient federated learning over wireless communication networks. *IEEE Transactions on Wireless Communications*, 20(3):1935–1949, 2021. doi: 10.1109/TWC.2020.3037554.

Xinwei Zhang, Zhiqi Bu, Zhiwei Steven Wu, and Mingyi Hong. Differentially private SGD without clipping bias: An error-feedback approach. In *International Conference on Learning Representations*, 2024.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A NOTATION

Table 4: Summary of key notation.

Symbol	Description
n	Total number of clients
r	Number of clients selected per round
K	Random allocation cycle length
κ	Rounds per cycle each client participates in
T	Total number of communication rounds
τ	Number of local epochs per round
d	Model dimension (number of parameters)
k	Number of coordinates kept after sparsification
$p = k/d$	Sparsification rate
$q = r/n$	Client sampling probability per round
θ^t	Global model at round t
$\mathcal{W}^t$	Set of selected clients at round t, $\lvert\mathcal{W}^t\rvert = r$
$\mathcal{D}_m$	Local dataset of client m
$m^t \in \{0,1\}^d$	Shared random sparsification mask at round t
g_m^t	Model update of client m at round t
e_m	Local residual, stored before noise, persists across rounds
$\hat{g}_m^t$	Update after selective residual injection
$\bar{g}_m^t$	Clipped update, $\lVert\bar{g}_m^t\rVert \le C$
b_m^t	DP noise added by client m at round t
$\tilde{g}_m^t$	Noisy clipped update
s_m^t	Sparsified upload
C, σ	Clipping threshold; noise multiplier
ϵ, δ	DP privacy parameters
η	Local learning rate
Φ	CPU cycles to process one data sample
f^{cpu}	CPU frequency of the client
χ	Effective capacitance coefficient of the computing chipset
P^{tx}	Transmit power of the client
B	Uplink bandwidth
h	Uplink channel gain
N_0	Noise power spectral density
$E^{\max}$	Energy budget of a client over the training run
λ	Weight on the energy term in (16)

B RANDOM ALLOCATION

Algorithm 2 gives the $\kappa = 1$ random allocation of Definition 2, which the server runs at the start of every cycle of K rounds.

Algorithm 2 Random Allocation ($\kappa = 1$)

Require: n: clients, K: cycle length
1: $\mathcal{W}^0, \ldots, \mathcal{W}^{K-1} \leftarrow \emptyset$
2: **for** each client $m \in [n]$ **do**
3: Client m draws a round $j \in \{0, \ldots, K-1\}$ uniformly at random
4: $\mathcal{W}^j \leftarrow \mathcal{W}^j \cup \{m\}$
5: **end for**
6: **return** $(\mathcal{W}^0, \ldots, \mathcal{W}^{K-1})$

C Privacy Analysis

C.1 Definitions and Lemmas

Definition 3 ((ϵ, δ)-DP, Dwork et al., 2006)**.** *A randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if for any two adjacent datasets D, D' and any output set $O \subseteq \mathrm{range}(\mathcal{M})$,*

$$\Pr[\mathcal{M}(D) \in O] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in O] + \delta. \tag{17}$$

Two datasets are adjacent if they differ by the entire local data of one client, giving client-level DP. When $\delta = 0$, $\mathcal{M}$ satisfies pure ϵ-DP.

Definition 4 ((α, ρ)-RDP, Mironov, 2017)**.** *For a real number $\alpha > 1$ and $\rho \geq 0$, a mechanism $\mathcal{M}$ satisfies (α, ρ)-RDP if for any two adjacent datasets D, D', the Rényi α-divergence between $\mathcal{M}(D)$ and $\mathcal{M}(D')$ satisfies*

$$D_\alpha[\mathcal{M}(D) \| \mathcal{M}(D')] := \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{\mathcal{M}(D)}{\mathcal{M}(D')}\right)^{\alpha}\right] \leq \rho. \tag{18}$$

Definition 5 (ℓ_2-sensitivity, Dwork et al., 2006)**.** *The ℓ_2-sensitivity of a function $h : \mathcal{D} \to \mathbb{R}^d$ is*

$$\psi(h) = \sup_{D \sim D'} \|h(D) - h(D')\|_2, \tag{19}$$

where $D \sim D'$ denotes that D and D' are adjacent datasets.

Lemma 3 (Gaussian Mechanism, Mironov, 2017)**.** *Let $h : \mathcal{D} \to \mathbb{R}^d$ be a function with ℓ_2-sensitivity $\psi(h)$. The Gaussian mechanism $\mathcal{M}(D) = h(D) + \mathcal{N}(0, \sigma^2 \psi^2(h) I_d)$ satisfies $(\alpha, \alpha/2\sigma^2)$-RDP.*

Lemma 4 (RDP of the Sampled Gaussian Mechanism, Mironov et al., 2019)**.** *Let $\mathcal{M}$ be the Sampled Gaussian Mechanism: each element is sampled independently with probability q (Poisson subsampling), then Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$ is added to a function of ℓ_2-sensitivity 1. If $q \leq \frac{1}{5}$, $\sigma \geq 4$, and α satisfies*

$$1 < \alpha \leq \tfrac{1}{2}\sigma^2 L - 2 \ln \sigma, \tag{20}$$

$$\alpha \leq \frac{\frac{1}{2}\sigma^2 L^2 - \ln 5 - 2 \ln \sigma}{L + \ln(q\alpha) + 1/(2\sigma^2)}, \tag{21}$$

where $L = \ln\big(1 + \frac{1}{q(\alpha-1)}\big)$, then $\mathcal{M}$ satisfies $(\alpha,\, 2q^2\alpha/\sigma^2)$-RDP.

Lemma 5 (Composition, Mironov, 2017)**.** *If a mechanism $\mathcal{M}$ consists of a sequence of T mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_T$, where each $\mathcal{M}_t$ satisfies (α, ρ_t)-RDP, then $\mathcal{M}$ satisfies $(\alpha, \sum_{t=1}^{T} \rho_t)$-RDP. In particular, if each round satisfies (α, ρ)-RDP, then T rounds satisfy $(\alpha, T\rho)$-RDP.*

Lemma 6 (RDP to DP Conversion, Mironov, 2017)**.** *If a mechanism $\mathcal{M}$ satisfies $(\alpha, \rho(\alpha))$-RDP, then for any $\delta \in (0, 1)$ it satisfies (ϵ, δ)-DP with*

$$\epsilon = \rho(\alpha) + \frac{\log(1/\delta)}{\alpha - 1}. \tag{22}$$

Theorem 3 (Amplification by Random Allocation, Feldman & Shenfeld, 2025)**.** *Given $\epsilon_0 > 0$ and $\delta_0 \in [0, 1]$, let R be an (ϵ_0, δ_0)-DP randomizer and let $A_K(R)$ denote the $\kappa = 1$ random allocation scheme over K rounds. For any $\epsilon, \delta > 0$,*

$$\delta_{A_K(R)}(\epsilon) \leq \delta_{P_{K,\eta}(R)}(\epsilon) + K\delta_0 + \delta, \tag{23}$$

where $P_{K,\eta}$ denotes Poisson subsampling at rate $\eta = \frac{1}{K(1-\gamma)}$ and $\gamma = \min\Big\{\cosh(\epsilon_0)\sqrt{\frac{2}{K}\ln\frac{1}{\delta}},\, 1 - \frac{1}{K}\Big\}$. Since η corresponds to a sampling probability of $1/K$ up to a lower-order term, the privacy of random allocation is asymptotically upper-bounded by Poisson subsampling at rate $1/K$.

C.2 Proof of Theorem 1

Proof. Under secure aggregation the server sees only the aggregate $\sum_{m \in \mathcal{W}^t} s_m^t$. For adjacent datasets differing in one client c, by Definition 5 the ℓ_2-sensitivity of the sum is

$$\psi = \sup_{D, D'} \left\| \sum_{m \in \mathcal{W}^t} \bar{g}_m^t \odot m^t - \sum_{m \in \mathcal{W}^t \setminus \{c\}} \bar{g}_m^t \odot m^t \right\|_2 = \sup_{D} \|\bar{g}_c^t \odot m^t\|_2. \tag{24}$$

Since the noise-free part of s_c^t is $\bar{g}_c^t \odot m^t$ with $\bar{g}_c^t = \mathrm{clip}_C(g_c^t + e_c \odot m^t)$, clipping gives $\|\bar{g}_c^t\|_2 \leq C$ even when the residual is added. Hence $\psi \leq C$, and with $|\mathcal{W}^t| = r$ clients the aggregate noise has per-coordinate variance $C^2\sigma^2$. We assume $\kappa = 1$ random allocation, where each client participates in exactly one round per cycle of K rounds. While random allocation has an exact RDP form (Feldman & Shenfeld, 2025), it is a partition sum that is impractical to compose over T rounds. Instead, by Theorem 3 random allocation is upper-bounded, up to a lower-order term, by Poisson subsampling at rate $q = 1/K$, to which the subsampled Gaussian mechanism bound of Lemma 4 applies, since $q = 1/K \leq 1/5$, $\sigma \geq 4$, and we assume the remaining conditions of Lemma 4 on α hold. Combined with Lemma 3, each round satisfies $(\alpha, \rho_t(\alpha))$-RDP with

$$\rho_t(\alpha) = \frac{2q^2\alpha}{\sigma^2}. \tag{25}$$

Composing T rounds by Lemma 5 gives $(\alpha, T\rho_t(\alpha))$-RDP. Converting to (ϵ, δ)-DP by Lemma 6 requires

$$\frac{2q^2T\alpha}{\sigma^2} + \frac{\log(1/\delta)}{\alpha - 1} \leq \epsilon. \tag{26}$$

We choose $\alpha = 1 + 2\log(1/\delta)/\epsilon$, so the second term equals $\epsilon/2$, and the first term is at most $\epsilon/2$ when

$$\sigma^2 \geq \frac{4q^2T(\epsilon + 2\log(1/\delta))}{\epsilon^2}, \qquad q = \tfrac{1}{K}. \tag{27}$$

Finally, the mask is data-independent and applied after the noise, so it does not affect the privacy guarantee (Dwork & Roth, 2014), and the residual is stored locally and never transmitted. □

C.3 Theorem 1 Versus the Numerical Accountant

Table 5 compares the noise multipliers for which Theorem 1 holds with the one from the accountant of (Feldman & Shenfeld, 2025) on Fashion-MNIST. The accountant needs $4\times$ to $5.4\times$ less noise at every K, since the proof of Theorem 1 uses the closed-form bound of Lemma 4 at a single order α, which only holds under its conditions on α, while the accountant computes the privacy of random allocation numerically over many orders.

Table 5: Smallest noise multiplier for which all conditions of Theorem 1 hold, and the noise multiplier from the accountant of (Feldman & Shenfeld, 2025), on Fashion-MNIST with $\epsilon = 0.7442$, $\delta = n^{-1.1}$ and $T = 180$.

K	15	20	30	45	60	90
Theorem 1	18.25	13.46	8.97	7.81	7.14	6.36
Accountant	4.26	3.23	2.22	1.63	1.39	1.17

D Signal Preservation Under Selective Feedback

Proof. The difference in clipping factors is $\frac{C}{\|g\|} - \frac{C}{\|g+e\|} = C\frac{\|g+e\| - \|g\|}{\|g\|\|g+e\|}$.
For the numerator, the triangle inequality gives $\|g + e\| - \|g\| \leq \|e\|$.
For the denominator, applying the triangle inequality to $g = (g+e)+(-e)$ gives $\|g\| \leq \|g+e\|+\|e\|$, hence $\|g + e\| \geq \|g\| - \|e\|$ and multiplying by $\|g\|$ gives $\|g\|\|g + e\| \geq \|g\|(\|g\| - \|e\|)$.
Replacing the denominator and the numerator, the difference is at most $\frac{C\|e\|}{\|g\|(\|g\|-\|e\|)}$. Multiplying by $|g[j]|$ gives the bound. □

E Convergence Analysis

E.1 Supporting Lemmas

Lemma 7. *For any $x, y \in \mathbb{R}^d$, $\langle x, y\rangle = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x - y\|^2$.*

Lemma 8 (Fenchel–Young inequality)**.** *For any $x, y \in \mathbb{R}^d$ and $\lambda > 0$, $\langle x, y\rangle \leq \frac{\lambda}{2}\|x\|^2 + \frac{1}{2\lambda}\|y\|^2$.*

Lemma 9. *For any $x_1, \dots, x_s \in \mathbb{R}^d$, $\left\| \sum_{j=1}^s x_j \right\|^2 \le s \sum_{j=1}^s \|x_j\|^2$. In particular, $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$ and $\|x+y+z\|^2 \le 3\|x\|^2 + 3\|y\|^2 + 3\|z\|^2$.*

Lemma 10 (Jensen's inequality)**.** *For any $x_1, \dots, x_n \in \mathbb{R}^d$, $\left\| \frac{1}{n} \sum_{m=1}^n x_m \right\|^2 \le \frac{1}{n} \sum_{m=1}^n \|x_m\|^2$.*

Lemma 11 (Bounded Local Divergence, Hu et al., 2024)**.** *Under Assumptions 1–3 and $\eta \le \frac{1}{12\tau L}$, the local model difference at round t is bounded as*

$$\frac{1}{n} \sum_{m=1}^n \mathbb{E}\|\theta^t - \theta_m^{t,s}\|^2 \le 32\eta^2\tau^2\beta^2\|\nabla f(\theta^t)\|^2 + 32\eta^2\tau^2\varkappa^2 + 4\tau\eta^2\bar{\zeta}^2. \tag{28}$$

Lemma 12 (Bounded Local Model Update, Hu et al., 2024)**.** *Let $d_m^t := \frac{1}{\tau} \sum_{s=0}^{\tau-1} g_m^{t,s}$ denote client m's average stochastic gradient over local steps. Under Assumptions 1–3,*

$$\frac{1}{n} \sum_{m=1}^n \mathbb{E}\|d_m^t\|^2 \le \frac{2L^2}{n\tau} \sum_{m=1}^n \sum_{s=0}^{\tau-1} \mathbb{E}\|\theta^t - \theta_m^{t,s}\|^2 + 2\big(\beta^2\|\nabla f(\theta^t)\|^2 + \varkappa^2\big) + \bar{\zeta}^2. \tag{29}$$

E.2 Proof of Lemma 2

Proof. Multiplying by $(1-m^t)$ only zeros entries of $\bar{g}_m^t$, so $\|e_m^{(t+1)}\| \le \|\bar{g}_m^t\|$. The clip scales $\hat{g}_m^t$ by $\min(1, C/\|\hat{g}_m^t\|)$, which leaves the update unchanged when $\|\hat{g}_m^t\| \le C$ and scales it to norm exactly C otherwise, so $\|\bar{g}_m^t\| \le C$ in both cases. For the expectation, by Definition 1 each coordinate is dropped with probability $1-p$, so

$$\mathbb{E}\|\bar{g}_m^t \odot (1-m^t)\|^2 = \sum_j \bar{g}_m^t[j]^2\, \mathbb{E}[1-m^t[j]] = (1-p)\|\bar{g}_m^t\|^2. \tag{30}$$

□

E.3 Proof of the Convergence Theorem

Proof. Throughout the proof, the expectation is taken over the mini-batch sampling, the rand-k masks, the DP noise, and the random allocation scheme. Let $p = k/d$ be the probability for rand-k. We assume each round selects $r = n/K$ clients uniformly at random from all n clients. Let $e_m^{(t)}$ denote the residual of client m at round t, with $e_m^{(0)} = \mathbf{0}$. We assume that $\|\hat{g}_m^t\| \le C$. We define the upload and residual as

$$s_m^t = g_m^t \odot m^t + e_m^{(t)} \odot m^t + b_m^t \odot m^t, \qquad e_m^{(t+1)} = \bar{g}_m^t \odot (1-m^t) = g_m^t \odot (1-m^t), \tag{31}$$

with $b_m^t \sim \mathcal{N}(0, \frac{C^2\sigma^2}{r} I_d)$, while $e_m^{(t+1)} = e_m^{(t)}$ for $m \notin \mathcal{W}^t$. We define

$$\mathcal{E}^t := \frac{1}{r} \sum_{m=1}^n e_m^{(t)}, \qquad \tilde{\theta}^t := \theta^t + \mathcal{E}^t, \tag{32}$$

so that $\tilde{\theta}^0 = \theta^0$. Adding the two equations in (31) and subtracting $e_m^{(t)}$ gives $s_m^t + e_m^{(t+1)} - e_m^{(t)} = g_m^t + b_m^t \odot m^t - e_m^{(t)} \odot (1-m^t)$ for $m \in \mathcal{W}^t$, and summing over $\mathcal{W}^t$ with weight $1/r$ yields

$$\tilde{\theta}^{t+1} = \tilde{\theta}^t + u^t + \bar{b}^t - \ell^t; \qquad u^t := \frac{1}{r} \sum_{m \in \mathcal{W}^t} g_m^t, \quad \bar{b}^t := \frac{1}{r} \sum_{m \in \mathcal{W}^t} b_m^t \odot m^t, \quad \ell^t := \frac{1}{r} \sum_{m \in \mathcal{W}^t} e_m^{(t)} \odot (1-m^t), \tag{33}$$

where ℓ^t is the discarded residual, the residual from the previous round sitting on coordinates that the current mask drops, which the overwrite rule erases instead of carrying forward. By Lemma 2, $\|e_m^{(t)}\| \le C$ and $\mathbb{E}\|e_m^{(t)}\|^2 \le (1-p)C^2$. Since $n/r = K$, Lemma 10 gives

$$\mathbb{E}\|\mathcal{E}^t\|^2 = \frac{n^2}{r^2}\, \mathbb{E}\Big\| \frac{1}{n} \sum_{m=1}^n e_m^{(t)} \Big\|^2 \le K^2 \cdot \frac{1}{n} \sum_{m=1}^n \mathbb{E}\|e_m^{(t)}\|^2 \le K^2(1-p)\, C^2, \tag{34}$$

and, since the mask is independent of the residuals so that $\mathbb{E}\|e_m^{(t)} \odot (1-m^t)\|^2 = (1-p)\, \mathbb{E}\|e_m^{(t)}\|^2$, Lemma 10 gives

$$\mathbb{E}\|\ell^t\|^2 \le \frac{1}{r} \sum_{m \in \mathcal{W}^t} (1-p)\, \mathbb{E}\|e_m^{(t)}\|^2 \le (1-p)^2\, C^2. \tag{35}$$

By Assumption 1, $f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$. Taking $x = \tilde{\theta}^t$ and $y = \tilde{\theta}^{t+1}$ in (33), adding and subtracting $\nabla f(\theta^t)$ in the first inner product, and using that $\bar{b}^t$ has zero mean and is independent of $\tilde{\theta}^t$, we get

$$\mathbb{E}[f(\tilde{\theta}^{t+1}) - f(\tilde{\theta}^t)] \le \mathbb{E}\langle \nabla f(\theta^t), u^t\rangle + \mathbb{E}\langle \nabla f(\tilde{\theta}^t) - \nabla f(\theta^t), u^t\rangle + \mathbb{E}\langle \nabla f(\tilde{\theta}^t), -\ell^t\rangle + \frac{L}{2}\mathbb{E}\|u^t + \bar{b}^t - \ell^t\|^2. \tag{36}$$

We bound the first term of (36). By Assumption 2, $\mathbb{E}[u^t] = -\eta \sum_{s=1}^{\tau} \frac{1}{n}\sum_{m=1}^{n} \nabla f_m(\theta_m^{t,s-1})$,

$$\begin{aligned}\mathbb{E}\langle \nabla f(\theta^t), u^t\rangle &= -\eta \sum_{s=1}^{\tau} \Big(\tfrac{1}{2}\|\nabla f(\theta^t)\|^2 + \tfrac{1}{2}\mathbb{E}\big\|\tfrac{1}{n}\textstyle\sum_m \nabla f_m(\theta_m^{t,s-1})\big\|^2 \\ &\qquad - \tfrac{1}{2}\mathbb{E}\big\|\nabla f(\theta^t) - \tfrac{1}{n}\textstyle\sum_m \nabla f_m(\theta_m^{t,s-1})\big\|^2 \Big) \\ &\le -\frac{\eta\tau}{2}\|\nabla f(\theta^t)\|^2 + \frac{\eta L^2}{2}\sum_{s=1}^{\tau}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}\|\theta^t - \theta_m^{t,s-1}\|^2, \end{aligned} \tag{37}$$

where the equality comes from Lemma 7, and we drop the second negative term from the inequality. Since $\nabla f(\theta^t) = \frac{1}{n}\sum_m \nabla f_m(\theta^t)$, we apply Lemma 10 then Assumption 1. Substituting Lemma 11 and summing over the τ steps, we get

$$\mathbb{E}\langle \nabla f(\theta^t), u^t\rangle \le \Big(-\frac{\eta\tau}{2} + 16\eta^3\tau^3 L^2\beta^2\Big)\|\nabla f(\theta^t)\|^2 + 16\eta^3\tau^3 L^2\varkappa^2 + 2\eta^3\tau^2 L^2\bar{\zeta}^2. \tag{38}$$

For the second term of (36), we use Lemma 8 with $\lambda = 1$, Assumption 1, and (34),

$$\mathbb{E}\langle \nabla f(\tilde{\theta}^t) - \nabla f(\theta^t), u^t\rangle \le \frac{L^2}{2}\,\mathbb{E}\|\mathcal{E}^t\|^2 + \frac{1}{2}\,\mathbb{E}\|u^t\|^2 \le \frac{L^2K^2(1-p)\,C^2}{2} + \frac{1}{2}\,\mathbb{E}\|u^t\|^2. \tag{39}$$

For the third term, we use Lemma 8 choosing $\lambda = \frac{\eta\tau}{8}$, then $\|\nabla f(\tilde{\theta}^t)\|^2 \le 2\|\nabla f(\theta^t)\|^2 + 2L^2\|\mathcal{E}^t\|^2$ by Lemma 9 and Assumption 1, together with (34) and (35), give

$$\mathbb{E}\langle \nabla f(\tilde{\theta}^t), -\ell^t\rangle \le \frac{\eta\tau}{16}\,\mathbb{E}\|\nabla f(\tilde{\theta}^t)\|^2 + \frac{4}{\eta\tau}\,\mathbb{E}\|\ell^t\|^2 \le \frac{\eta\tau}{8}\|\nabla f(\theta^t)\|^2 + \frac{\eta\tau L^2K^2(1-p)\,C^2}{8} + \frac{4(1-p)^2\,C^2}{\eta\tau}. \tag{40}$$

We now bound the last term of (36). Using Lemma 9 and (35) gives

$$\mathbb{E}\|u^t + \bar{b}^t - \ell^t\|^2 = \mathbb{E}\|u^t - \ell^t\|^2 + \mathbb{E}\|\bar{b}^t\|^2 \le 2\,\mathbb{E}\|u^t\|^2 + 2(1-p)^2C^2 + \mathbb{E}\|\bar{b}^t\|^2. \tag{41}$$

Since every client receives the same mask, each of the k kept coordinates carries averaged noise of variance $\frac{1}{r^2}\cdot r\cdot\frac{C^2\sigma^2}{r} = \frac{C^2\sigma^2}{r^2}$, giving $\mathbb{E}\|\bar{b}^t\|^2 = \frac{kC^2\sigma^2}{r^2}$.
Let $g_m^t = -\eta\tau\, d_m^t$ where $d_m^t = \frac{1}{\tau}\sum_{s=0}^{\tau-1} g_m^{t,s}$ is the average local gradient. Since $\mathcal{W}^t$ is chosen uniformly at random and using Lemma 10 we get $\mathbb{E}\|u^t\|^2 \le \eta^2\tau^2\cdot\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}\|d_m^t\|^2$. Applying Lemma 12 and then Lemma 11, if we choose $\eta \le \frac{1}{12\tau L}$,

$$\mathbb{E}\|u^t\|^2 \le 3\eta^2\tau^2\beta^2\,\|\nabla f(\theta^t)\|^2 + 3\eta^2\tau^2\varkappa^2 + 2\eta^2\tau^2\bar{\zeta}^2. \tag{42}$$

Substituting (38), (39), (40), (41), and (42) into (36),

$$\begin{aligned}\mathbb{E}[f(\tilde{\theta}^{t+1}) - f(\tilde{\theta}^t)] \le{}& \Big(-\frac{3\eta\tau}{8} + 16\eta^3\tau^3L^2\beta^2 + \frac{3}{2}\eta^2\tau^2\beta^2 + 3L\eta^2\tau^2\beta^2\Big)\|\nabla f(\theta^t)\|^2 \\ &+ 16\eta^3\tau^3L^2\varkappa^2 + \frac{3}{2}\eta^2\tau^2\varkappa^2 + 3L\eta^2\tau^2\varkappa^2 + 2\eta^3\tau^2L^2\bar{\zeta}^2 + \eta^2\tau^2\bar{\zeta}^2 + 2L\eta^2\tau^2\bar{\zeta}^2 \\ &+ \Big(\frac{1}{2} + \frac{\eta\tau}{8}\Big)L^2K^2(1-p)\,C^2 + \Big(\frac{4}{\eta\tau} + L\Big)(1-p)^2\,C^2 + \frac{L\,k\,C^2\sigma^2K^2}{2\,n^2}. \end{aligned} \tag{43}$$

If

$$\eta \le \min\left\{\frac{1}{144\,\tau L\beta^2},\ \frac{1}{48\,\tau\beta^2},\ \frac{1}{12\,\tau L}\right\}, \tag{44}$$

then $16\eta^2\tau^2L^2\beta^2 + \frac{3}{2}\eta\tau\beta^2 + 3L\eta\tau\beta^2 \leq \frac{1}{8}$, so the $\|\nabla f(\theta^t)\|^2$ coefficient is at most $-\frac{\eta\tau}{4}$ and

$$\begin{aligned}\mathbb{E}[f(\tilde{\theta}^{t+1}) - f(\tilde{\theta}^t)] \leq &-\frac{\eta\tau}{4}\|\nabla f(\theta^t)\|^2 + 16\eta^3\tau^3L^2\varkappa^2 + \frac{3}{2}\eta^2\tau^2\varkappa^2 + 3L\eta^2\tau^2\varkappa^2 \\ &+ 2\eta^3\tau^2L^2\bar{\zeta}^2 + \eta^2\tau^2\bar{\zeta}^2 + 2L\eta^2\tau^2\bar{\zeta}^2 \\ &+ \Big(\frac{1}{2} + \frac{\eta\tau}{8}\Big)L^2K^2(1-p)\,C^2 + \Big(\frac{4}{\eta\tau} + L\Big)(1-p)^2\,C^2 + \frac{L\,k\,C^2\sigma^2K^2}{2\,n^2}.\end{aligned} \tag{45}$$

Summing over $t = 0, \ldots, T-1$, with $\tilde{\theta}^0 = \theta^0$ and $\mathbb{E}[f(\tilde{\theta}^T)] \geq f^*$, and dividing by $\frac{\eta\tau}{4}T$, we get

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\theta^t)\|^2 \leq &\frac{4\big(f(\theta^0) - f^*\big)}{\eta\tau\,T} + \big(64\,\eta^2\tau^2L^2 + 6\,\eta\tau + 12\,L\eta\tau\big)\varkappa^2 + \big(8\,\eta^2\tau L^2 + 4\,\eta\tau + 8\,L\eta\tau\big)\bar{\zeta}^2 \\ &+ \Big(\frac{2}{\eta\tau} + \frac{1}{2}\Big)L^2K^2(1-p)\,C^2 + \Big(\frac{16}{\eta^2\tau^2} + \frac{4\,L}{\eta\tau}\Big)(1-p)^2\,C^2 + \frac{2\,L\,k\,C^2\sigma^2K^2}{\eta\tau\,n^2}.\end{aligned} \tag{46}$$

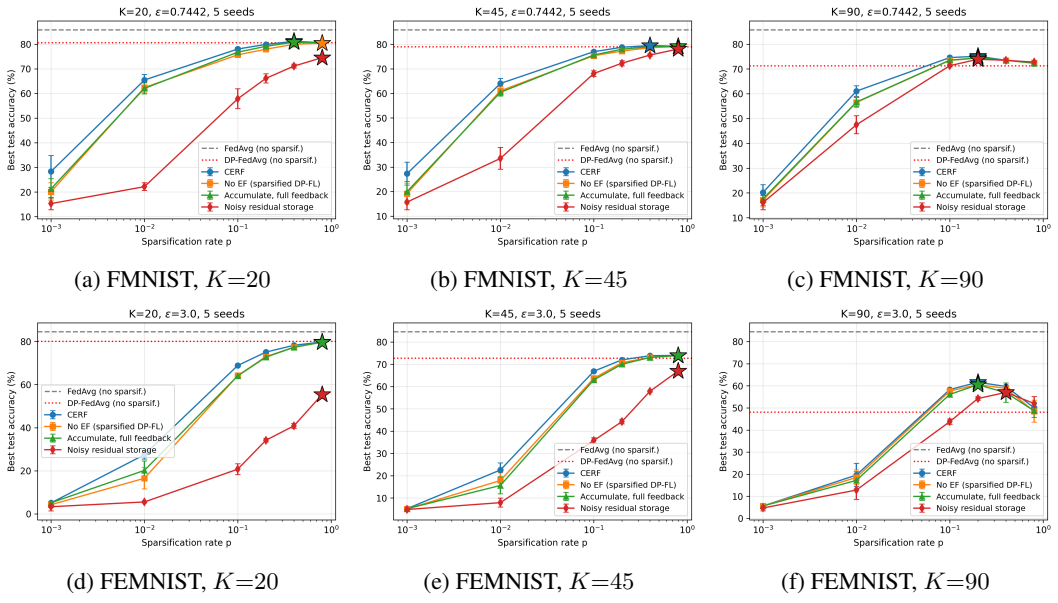
Grouping terms and using $k = pd$, this is simplified to

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\theta^t)\|^2 \leq &\frac{4\big(f(\theta^0) - f^*\big)}{\eta\tau\,T} + 2\eta\tau\big(32\eta\tau L^2 + 6L + 3\big)\varkappa^2 + 4\eta\tau\big(2\eta L^2 + 2L + 1\big)\bar{\zeta}^2 \\ &+ \Big(\frac{2}{\eta\tau} + \frac{1}{2}\Big)L^2K^2(1-p)\,C^2 + \Big(\frac{16}{\eta^2\tau^2} + \frac{4\,L}{\eta\tau}\Big)(1-p)^2\,C^2 + \frac{2\,L\,p\,d\,C^2\sigma^2K^2}{\eta\tau\,n^2}.\end{aligned} \tag{47}$$

□

F Additional Results

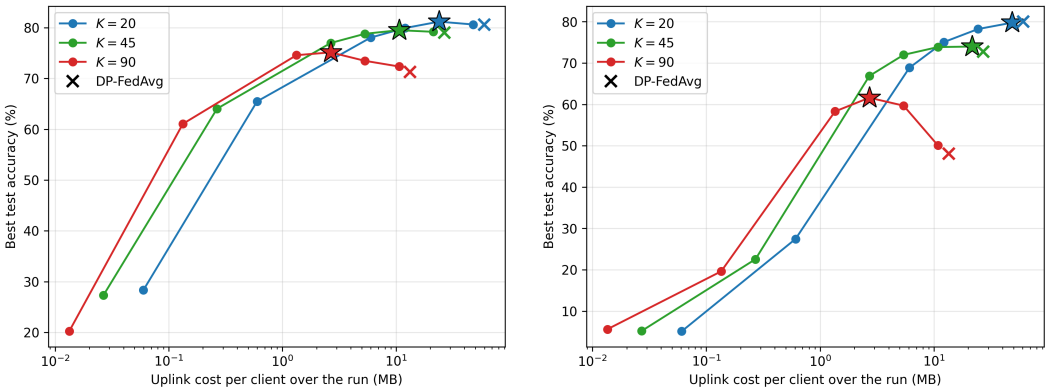
Fig. 3 shows the best test accuracy versus the sparsification rate p for each K, and Fig. 4 shows the best test accuracy versus the uplink cost per client over the run.



(a) FMNIST, K=20 (b) FMNIST, K=45 (c) FMNIST, K=90

(d) FEMNIST, K=20 (e) FEMNIST, K=45 (f) FEMNIST, K=90

Figure 3: Best test accuracy versus sparsification rate p for each K.



(a) Fashion-MNIST (b) FEMNIST

Figure 4: Best test accuracy versus the uplink cost per client over the run. Each point is one sparsification rate p, the star marks the best p for each K, and the cross marks DP-FedAvg.