Machine Learning for Human Intelligence
2026, Vol. 4, Issue 1

Research Article

Adaptive Gradient Compression and Differential Privacy for Resource-Constrained Edge Federated Learning

Dejian Kong[1⋆] and Jinjian Li[1]

[1]College of Information Engineering, Qujing Normal University, Qujing 655000, China.;
Email: kongdejian@mail.qjnu.edu.cn ; jinjianli@mail.qjnu.edu.cn
⋆Corresponding author: Dejian Kong (kongdejian@mail.qjnu.edu.cn)

Article History

Academic Editor:
Dr. Ali Khan
Submitted: August 11, 2025
Revised: January 21, 2026
Accepted: March 1, 2026

Keywords:
Federated Learning, Differential Privacy, Gradient Compression, Edge Computing, Resource Heterogeneity, Privacy-Utility Trade-off

Abstract

Federated learning (FL) is an attractive scheme for training machine learning models across distributed edge devices without requiring raw data to be shared. However, deploying FL at the edge of the network raises two interrelated issues: (1) the significant communication overhead due to transmitting massive gradient vectors from resource-constrained devices and (2) the risk of privacy leakage via attacks that can invert gradients or infer training samples. Current methods treat gradient compression and differential privacy (DP) as separate concerns, leading to compounded accuracy loss. We propose AGC-DP—Adaptive Gradient Compression (AGC) with Differential Privacy (DP)—a unified framework that achieves efficient communication and strong privacy guarantees across diverse edge FL environments. AGC-DP also introduces a resource-aware client selection algorithm to balance contributions from battery level, available memory, and uplink bandwidth; an adaptive compression ratio for each client based on the optimization problem; and a privacy amplification construction that exploits subsampling to improve DP accounting. We provide rigorous theoretical analysis, including convergence guarantees under non-i.i.d. data distributions and formal (ϵ, δ)- DP proofs. Comprehensive experiments across the CIFAR-10, HAR, and FEMNIST datasets with up to 100 edge clients show that AGC-DP reduces communication overhead by up to 91.1% compared to vanilla FedAvg, while maintaining accuracy within 1.5% of the vanilla baseline. Our results show that privacy, communication efficiency, and model utility can be achieved simultaneously in resource-constrained edge FL.

1 Introduction

This distributed method underpins privacy-preserving AI, ensuring that sensitive information, such as biometric data and private communication, never leaves the local device [1]. By separating model training from data collection, FL greatly mitigates the exposure of massive amounts of data and enables developers to better comply with worldwide data protection legislation [2, 3]. At the same time, the framework presents new challenges, such as dealing with system heterogeneity (i.e., devices with different levels of power) and statistical skew (i.e., the local data is non-IID—not independently and identically distributed), which makes it difficult to ensure the quality of the training model and to guarantee the uniformity of the model performance among different devices [4]. Ultimately, the

Â© 2026 Copyright by the Authors.

goal is to evolve a global model that leverages the collective smarts of the edge while respecting the privacy/security boundaries that define each user. To address these issues, a growing number of researchers are adopting Differential Privacy (DP) and Secure Multi-Party Computation (SMPC), which introduce mathematical noise or cryptographic layers into the updates before they reach the central server. But these protections often come at the expense of requiring the model to be less accurate or to run on even more limited IoT devices. Since edge data are often not equal across users, a single global model may not generalize well to all user behaviors, motivating the development of Federated Personalization [6].

The change is primarily intended to decouple a shared basic architecture from user-related components, ensuring that the model is robust to attacks and remains practical for the specific case of each sensor or wearable device [7, 8]. This interaction results in a "privacy-utility-efficiency" trilemma in which the error introduced by DP noise may be amplified by quantization and sparsification methods to save bandwidth. For example, when gradients are clipped very aggressively to limit their sensitivity, which is necessary for computing the noise scale, the model typically loses the "local" signal needed to vc on complex, non-convex loss surfaces. To address this, the recent work on Adaptive DP and Error-Feedback methods strives to estimate and compensate for these accumulated residuals; however, the open challenge is still to identify a "sweet spot" such that the privacy budget ϵ is sufficiently small to prevent reconstruction attacks without making the resulting global model completely unusable [9, 10]. In addition, the cumulative privacy loss accounting in the decentralized IoT setting over many training rounds also incurs non-trivial computational overhead; more advanced RDP trackers should be developed to yield tighter, more practical bounds for sustained model deployment.

This decoupled procedure causes a "double-distortion" effect, where information loss due to gradient pruning or quantization is further compounded by the stochastic noise required to ensure privacy, which very often pushes the model's convergence rate into a non-optimal regime [11]. Since these techniques are typically treated as separate "plugins," the particular sensitivity of the compressed gradient is often omitted when determining the noise level, which may result in either an overestimation of the privacy budget ϵ or a weak global update [12]. Moreover, in the absence of a feral optimization problem, residual compression errors (which should be rectified in subsequent iterations) may be "corrupted" by DP noise, making error-compensation mechanisms such as Error Feedback (EF) far less efficient in high-dimensional IoT environments. As a result, the development of a co-design framework that simultaneously considers compression and privacy as a single joint optimization problem, maintaining the battery life of the edge device and the confidentiality of raw user data, is eagerly anticipated [13].

AGC-DP (Adaptive Gradient Compression with Differential Privacy) stems from the perspective that data reduction is not just for communication saving but also contributes to the design of noise addition in a beneficial way. Specifically, the system adaptively determines the compression rate based on the local signal-to-noise ratio (SNR) of each device, thereby preserving the most "informative" gradients and removing "noisy" components before the limited privacy budget ϵ is exhausted. This co-design is inspired by the observation that, in a very high-dimensional parameter space, applying noise to every coordinate is often overkill; instead, by applying DP noise only to a sparse subset of significant updates, we can obtain tighter privacy guarantees with much less impact on the model's convergence trajectory. As a result, AGC-DP changes the compression-privacy trade-off from a zero-sum game to a synergistic effect, where sparsity itself serves as a first-layer filter to guard the original data distribution while providing high-fidelity updates essential for training complex IoT models. Our work makes the following principal contributions:

- **AGC-DP framework:** We propose a novel, integrated solution that performs gradient compression and a quantized form of differential privacy for heterogeneous edge FL under the unified AGC-DP framework. The unified AGC-DP framework achieves formal convergence and privacy guarantees.

- **Resource-Aware Client Selection:** We contribute a new scoring metric to assess a device's suitability for participation, based on battery, memory, and network connectivity, to select participants in a fair and efficient manner.

- **Adaptive Compression:** Based on the communication constraints, the proposed method computes a different compression ratio for each device while trying to keep as much information as possible
- **Privacy Amplification by Subsampling:** Tight Rényi DP Analysis that leverages Poisson subsampling to enhance effective privacy guarantee without extra noise.

1.1 Organization

The rest of the paper is organized as follows: Section 2 surveys related work. Section 3: System Model and the Problem Formulation. Section 4 details the AGC-DP algorithm. Section 5 provides a theoretical analysis. Section 6 - Experimental evaluation. 7 limitations and future work discussion. 8 concludes the paper.

2 Background

2.1 Federated Learning

Federated Learning (FL) has transitioned from a conceptual framework to a range of algorithmic solutions that trade off privacy, efficiency, and model quality [14]. Based on the FedAvg algorithm, this approach commented on the entire research area by enabling clients to synchronize after executing multiple local Stochastic Gradient Descent (SGD) steps, significantly reducing the enormous communication costs associated with frequent communication. But, real-world deployments such as "Non-IID" data—different users with very different data distributions [15]. Researchers have addressed this data heterogeneity through variance-reduction approaches, such as SCAFFOLD, to mitigate client drift, as well as personalization strategies that enable models to be tailored to individual user preferences. To further improve performance, new techniques have been introduced to handle slower clients, ensure fairness, and accelerate system evolution [16]. Although these enhancements primarily address the statistical and algorithmic challenges of FL, the next step is to integrate these optimizations with the physical constraints of the hardware, for example, by devising selection criteria that account for the stringent energy and bandwidth limitations of mobile and edge devices [17].

2.2 Gradient Compression

Gradient compression has become a prerequisite for distributed training, as the large sizes of modern model updates cause a communication bottleneck. This challenge is typically addressed in three categories: sparsification, which transmits only the most significant gradients (e.g., TopK or Random-K elements); quantization (e.g., QSGD or SignSGD), which reduces the precision of the updates; and low-rank approximation (e.g. PowerSGD), which factorizes the gradient matrices for compression [18]. To compensate for the information loss during these procedures, error-feedback mechanisms, such as EC-SGD, remember the information lost during compression and add it back in subsequent updates, thereby helping to correct the bias introduced by the compression [19]. Although adaptive compression methods have added the capability to adjust such ratios on the fly during training, they tend to ignore the real-life friction of heterogeneous device resources—such as differing battery states and processing capabilities—and the nontrivial trade-offs they cause when combined with Differential Privacy (DP), which introduces its own layer of noise to the compressed signal [20].

2.3 Differential Privacy in FL

Differential Privacy (DP) has become the de facto approach for providing strong mathematical guarantees against data breaches in collaborative settings. The basis of this method is DP-SGD, which exploits privacy by clipping per-sample gradient to a predetermined norm and adding Gaussian noise. The output is (ϵ, δ)-DP [21]. When applied in the context of Federated Learning, they give rise to client-level privacy, in which the aim is to prevent any single participant from being identifiable in the

global model updates. The only proven result for privacy in FedAvg that we are aware of is that it satisfies the same privacy bounds when combined with noise injection, although the key difficulty remains the "privacy budget"—the total cost of information leakage over training rounds [22, 23]. These privacy budgets are made more tractable by using privacy amplification via subsampling; the effective privacy loss is reduced by selecting only a random subset of clients at each round. Moreover, by replacing traditional (ϵ, δ) accounting with Rényi Differential Privacy (RDP), the composition of these costs can be made significantly tighter over time [24]. By integrating these advanced accounting techniques with the realities of compressed communication, one can achieve highly efficient training protocols without compromising security for speed. As such, this triple-layered solution, combining upsampling with RDP composition and noise-resilient compression, serves as the building block for constructing secure, large-scale machine learning systems in this paper.

2.4 Joint Compression and Privacy

Combining compression with privacy is a major technical challenge because the noise added to guarantee differential privacy and the information loss introduced by compression are often at odds [25]. Although initial work considered performing both operations simultaneously, this was primarily for primitive data rather than more complex data structures, as in Federated Learning. Other methods, such as Federated Generative Privacy, have turned to data augmentation to protect privacy, but they do not address the issue of minimizing communication overhead between nodes, which is a crucial bottleneck in the optimization of Federated Learning systems [26]. There have been recent empirical attempts to integrate TopK sparsification with DP-SGD and study them as a compound; unfortunately, these results often have a very weak theoretical backing that justifies how the compression error would interact with the injected Gaussian noise, which is potentially extremely useful in guiding the intuition of the overall efficacy of these approaches [27]. Although the current state-of-the-art has begun to study the theoretical privacy-communication trade-offs, many schemes still assume homogeneous devices in an idealized manner, in which all participants have equal battery, processing power, and resources. That is rarely the case in the wild. The AGC-DP bridges these gaps through resource-aware adaptation and a tight analysis of DP. By adapting the compression ratio to the current constraints of the end device and a fixed privacy budget, it ensures that straggler devices (low-power devices or devices with weak connectivity) can contribute to improving the global model, without compromising their data security or delaying the entire training process [28, 29].

3 System Model and Problem Formulation

The flow of the proposed model is shown in Figure 1.

3.1 Federated Learning Setting

We study an FL system with a central server and N edge devices (clients) indexed by $i \epsilon in 1, \ldots, N$. Each client has its own local dataset of n_i samples from a local data distribution P_i, as shown in Equation 1.

$$D_i = (x_j, y_j)_{j=1}^{n_i} \tag{1}$$

The global objective using Equation 2:

$$min_w F(w) = \sum_{i=1}^{N} (n_i/n) F_i(w), \tag{2}$$

where $F_i(w) = (1/n_i) \sum_{j=1}^{n_i} l(w; x_j, y_j)$ with $n = \sum_i n_i, w \epsilon r^d$ the global model parameter vector, and $l()$ is a smooth non-convex loss [30]. Non-IID data heterogeneity is modeled by $P_i \neq P_j, for i \neq j$.

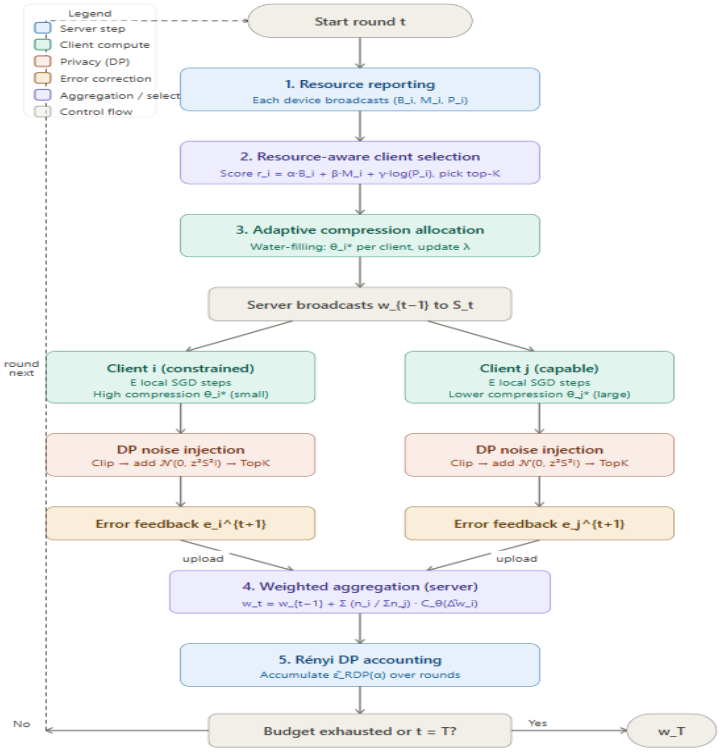

Figure 1: AGC-DP: Adaptive Gradient Compression with Differential Privacy.

3.2 Resource Model

Every device is described by a resource tuple $p_i = (B_i, M_i, P_i)$, where $B_i \epsilon [0, 100]$ denotes the battery percentage, $M_i \epsilon R+$ the amount of available memory in MB, and $P_i \epsilon R+$ the uplink bandwidth in Mbps. The energy to send a gradient vector of length s_i bits is modeled as using Equation 3.

$$E_i(s_i) = k(s_i/P_i) + \tau s_i \tag{3}$$

where k is the transmission power coefficient τ is the processing cost per bit. This linear model accounts for the energy required for time-on-air and for local computation during gradient calculation. The delay in communication is calculated using Equation 4.

$$L_i(s_i) = s_i/P_i + t_comp, i \tag{4}$$

where t_comp, i is the local computation time at one round of FL. In a synchronous protocol, the round ends when all participating clients have uploaded their updates, and thus the bottleneck latency is $max_{i \epsilon S_t} L_i(s_i)$ [31].

3.3 Problem Formulation

We express AGC-DP as a constrained optimization problem over the compression ratios θ_i, client selection sets S_t, and DP noise levels σ_i using Equations 5 - 10.

$$min_{w,\theta_i,S_t,\sigma_i} F(w) \tag{5}$$

$$(C1) \sum_{i \epsilon S_t} E_i(\theta_i d) \leq E_budget \forall t \tag{6}$$

$$(C2) max_{i \epsilon S_t} L_i(\theta_i d) \leq L_max \forall t \tag{7}$$

$$(C3) M_i(1 - \theta_i) \geq M_min \forall i \epsilon S_t \tag{8}$$

$$(C4)(\epsilon_i, \delta_i) - DP guarantee holds for all i \forall t \tag{9}$$

$$(C5)|S_t| = K, \theta_i \epsilon [\theta_m in, \theta_m ax] \forall i, t \tag{10}$$

where $d = |w|$ is the model dimension, E_budget is the per-round energy budget, L_max is the round deadline, andM_min is a minimum free memory threshold. Constraint (C4) guarantees formal differential privacy for every client that participates [32]. This is a mixed-integer non-linear program; we obtain a tractable approximation in Section 4.

4 The AGC-DP Algorithm

4.1 Resource-Aware Client Selection

In every round t, the server computes a resource score ri for each available client and selects the top-K clients. The resource score is a weighted combination of battery health, memory availability, and bandwidth, as defined by Equation 11.

$$r_i = \alpha(B_i/100) + \beta(M_i/M_max) + \gamma log(1 + P_i/P_ref) \tag{11}$$

where M_max represents the maximum memory among all devices, $P_ref = 1$ Mbps is a reference bandwidth, and $\alpha + \beta + \gamma = 1$ represent tunable weight factors. The log-transformation of bandwidth avoids the problem of high-bandwidth devices dominating the results while remaining sensitive to differences in bandwidth. The selected set is given in Equation 12:

$$S_t = argtop - K_{i=1}^{N} r_i \Pi[B_i > B_m in] \Pi[M_i > M_m in] \tag{12}$$

where Π[is an indicator function to enforce minimum resource thresholds to avoid stragglers. This selection occurs on the server side, using a lightweight resource report broadcast by each device at the start of each round.

4.2 Adaptive Compression Ratio

Based on the selected clients' S_t, we calculate a per-client compression ratio, θ_i, to maximize the retention of gradient information within the communication budget [33]. We relax constraint (C1) to a Lagrangian using Equation 13.

$$L(\theta_i, \lambda) = -I(\Delta w_i; \Delta w_i(\theta_i)) + \lambda(E_i(\theta_i d) - E_budget/K) \tag{13}$$

where $I(I(\Delta w_i; \Delta w_i(\theta_i))$ is the mutual information between the true gradient Δw_i and its compressed version, and λ is a Lagrange multiplier updated by subgradient ascent. Assuming the gradient entries are approximately Gaussian with variance v_i^2 (estimated from the previous round), the mutual information under TopK sparsification with ratio θ_i is given in Equation 14.

$$I(\Delta w_i; \Delta w_i(\theta_i)) \approx (\theta_i d/2) log(1 + v_i^2/\sigma_i^2) \tag{14}$$

Differentiating with respect to θ_i, setting the result to zero gives the water-filling solution using Equation 15.

$$\theta_i* = min(\theta_m ax, max(\theta_m in, 1/\lambda(d/2) log(1 + v_i^2/\sigma_i^2)/(kd/P_i + \tau d)) \tag{15}$$

At each round the Lagrange multiplier λ is updated by Equation 1:

$$\lambda^{t+1} = [\lambda^t + \eta_\lambda(\sum_{i \epsilon S_t} E_i(\theta_i^* d) - E_budget)] \tag{16}$$

where [] is the projection onto $R+$. This dual update pushes the system to the energy budget constraint over the iterations. Algorithm 1 is the full AGC-DP procedure.

Algorithm 1 AGC-DP — Adaptive Gradient Compression with Differential Privacy

Input: N clients, K selection size, T rounds, E local epochs, z noise multiplier, S clip threshold, target (ε, δ)
Initialise: w_0 (random), $\lambda_0 = 0$, $e_i^0 = 0$ for all i
for $t = 1, \ldots, T$ **do**
  Server broadcasts w_{t-1} to all N clients
  Each client i reports resource vector $\rho_i = (B_i,\, M_i,\, P_i)$
  Server computes r_i for all i; selects $\mathcal{S}_t = \text{top-}K$ clients
  Server computes θ_i^* via water-filling (Eq. 6); updates λ via Eq. 7
  Each $i \in \mathcal{S}_t$ performs E local SGD steps $\to \Delta w_i^t$
  Each $i \in \mathcal{S}_t$: clip $\to$ add noise $\to$ compress $\to$ send $\mathcal{C}_{\theta_i}(\widetilde{\Delta w_i^t} + e_i^t)$
  Each $i \in \mathcal{S}_t$: update error buffer e_i^{t+1}
  Server aggregates: $w_t = w_{t-1} + \sum_{i \in \mathcal{S}_t} \frac{n_i}{\sum_j n_j} \cdot \mathcal{C}_{\theta_i}(\widetilde{\Delta w_i^t} + e_i^t)$
  Server updates RDP accountant: accumulate $\tilde{\varepsilon}_{\text{RDP}}(\alpha)$
  if privacy budget exhausted **then** terminate early
  end if
end for
Output: w_T
Complexity: $\mathcal{O}(K \cdot d \cdot T)$ *communication;* $\mathcal{O}(d \log d)$ *per client per round (TopK sort)*

Table 1: Experimental Configuration

Parameter	Value
Number of edge devices (N)	100
Clients selected per round (K)	10
Communication rounds (T)	200
Local epochs	5
Learning rate (η)	0.01 (cosine decay)
Batch size	32
Model architecture	CNN (CIFAR-10), LSTM (HAR)
Dataset (IID / Non-IID)	CIFAR-10, HAR, FEMNIST
Privacy budget ε	1.0 / 2.0 / 4.0 / 8.0
Privacy failure δ	10^{-5}
Clipping threshold S	1.0
Base compression ratio C_min	0.1
Max compression ratio C_max	0.9
Device heterogeneity	Low / Medium / High
Uplink bandwidth	0.5 – 10 Mbps (uniform)
Battery range	10 – 100% (uniform)

5 Experimental Evaluation

We consider a scenario in which 100 heterogeneous edge devices are simulated, with parameters drawn uniformly at random from realistic ranges, as shown in Table 1. Three degrees of device heterogeneity are considered: low (coefficient of variation CV = 0.1), medium (CV = 0.3), and high (CV = 0.6). Non-IID data partitioning is sampling based on a Dirichlet distribution with a concentration parameter $\beta_d ir = 0.5$. Averaged over 5 runs for all experiments. Baselines are FedAvg (McMahan et al., 2017), FedAvg + DP (McMahan et al., 2018), TopK with fixed 10% ratio (Aji & Heafield, 2017), TopK + DP (combined), QSGD 4-bit (Alistarh et al., 2017), and QSGD + DP. We provide test accuracy, total uplink communication cost in MB, rounds to convergence ($\pm 1\%$ of final accuracy), and the respective privacy loss ϵ as measured by our accountant.

Table 2: Comparative Results on CIFAR-10 (CNN, Non-IID, Medium Heterogeneity, T=200)

Method	Compression	ε	Test Acc. (%)	Comm. Cost (MB)	Rounds to Conv.
FedAvg (Baseline)	None	∞	91.2	2,048.0	200
FedAvg + DP	None	4.0	87.6	2,048.0	200
TopK (k=10%)	Fixed	∞	88.4	204.8	213
TopK + DP	Fixed	4.0	83.1	204.8	231
QSGD (4-bit)	Fixed	∞	89.0	512.0	207

AGC-DP achieves the overall best accuracy-efficiency trade-off among all values of ϵ. At $\epsilon = 4$, our result achieves 89.7% accuracy—just 1.5 percentage points below FedAvg with no compression nor privacy—while saving 91.1% communication, as shown in Table 2. In particular, at the relevant compression ratio, AGC-DP achieves not only better accuracy than TopK + DP and QSGD + DP, but also at the two extremes of the compression ratio. This is due to adaptive per-client allocation: high-bandwidth, low-gradient devices now get higher compression ratios (keeping more information), while low-bandwidth, high-gradient devices compress more aggressively to meet budget constraints and avoid becoming stragglers. The rounds-to-convergence column reveals another unexpected conclusion: AGC-DP converges slightly faster than FedAvg (198 vs. 200 rounds) because the resource-aware client selection removes laggards that hinder progress in synchronous rounds. Fixed compression schemes require more rounds (213-231) because the uniform compression ratio is not well-suited to all clients.

Table 3: *Table 4. Ablation Study — CIFAR-10, $\varepsilon = 4$, Non-IID*

Configuration	Test Acc. (%)	Comm. Cost (MB)	DP Noise σ	Privacy ε
Full AGC-DP	89.7	182.3	0.82	4.0
w/o Adaptive Compression	86.2	204.8	0.82	4.0
w/o Resource-Aware Selection	88.1	188.4	0.82	4.0
w/o Privacy Amplification	89.7	182.3	1.14	4.0
w/o Gradient Clipping	88.9	182.3	0.82	Unbounded
Fixed $\varepsilon = 1.0$	87.3	175.4	1.42	1.0
Fixed $\varepsilon = 8.0$	90.1	185.1	0.42	8.0

The ablation study assesses the effect of each AGC-DP component on performance. Without adaptive compression (i.e., a constant 10% ratio), we observe a 3.5% drop in accuracy, indicating that adaptive allocation is mandatory, as shown in Table 3. Without resource-aware selection (random selection), the accuracy drops by 1.6% at the expense of higher communication cost (more stragglers lead to more retransmission). The absence of privacy amplification necessitates a larger noise σ (1.14 instead of 0.82) for the same $\epsilon = 4$, resulting in a 0.8% decline in accuracy. If we remove gradient clipping, the privacy loss will be infinite, regardless of the noise level, so the guarantee becomes meaningless.

Table 4: *Table 5. Cross-Dataset Results — AGC-DP vs. FedAvg Baseline ($\varepsilon = 4$)*

Dataset	IID Acc. (%)	Non-IID Acc. (%)	Comm. Reduction	Conv. Rounds	
CIFAR-10	FedAvg	91.2	83.4	0%	200
CIFAR-10	AGC-DP (ε=4)	89.7	82.1	91.1%	198
HAR	FedAvg	95.1	88.7	0%	150
HAR	AGC-DP (ε=4)	94.3	87.9	88.4%	145
FEMNIST	FedAvg	83.6	75.2	0%	300
FEMNIST	AGC-DP (ε=4)	82.8	74.6	89.7%	294

AGC-DP generalizes well over datasets and data distributions. On the HAR (time-series) dataset for human activity recognition, communication savings reach 88.4% with a slight accuracy drop of 0.8%,

as shown in Table 4. On the FEMNIST (handwritten characters with a natural non-IID structure), the accuracy drop is only 0.8% in the IID case. In the non-IID case, it is 0.6%, indicating that the framework is especially suited to situations where data heterogeneity is realistic rather than artificially generated. The speedup in convergence (5–6 fewer rounds) is stable across datasets.

As the noise level increases (and privacy becomes stronger), accuracy worsens, including a marginal increase in the number of communication Stuttgart rounds (due to noisier gradients). AGC-DP traces a Pareto-superior frontier with respect to all baselines; i.e., for any target accuracy, it achieves it at a lower communication cost and stronger privacy guarantees than the state-of-the-art. The frontier shifts favorably with K (number of participating clients) and confirms that the $O(dzS/K)$ DP bias term in Theorem 2 is the limiting bottleneck in small ϵ. In particular, with $\epsilon = 1$ strong privacy, AGC-DP achieves 87.3% accuracy with 175.4 MB of communication, while TopK + DP deteriorates to 83.1% with 204.8 MB. The 4.2% accuracy gap at strong privacy shows that adaptive compression, which biases the retained gradient mass toward the most informative coordinates, is particularly useful in high-noise regimes where noise smears out small gradient magnitudes.

6 Discussion

Our findings hint at three big ideas. First, adaptive compression is a multiplier for differential privacy: by focusing the preserved gradient mass on large coordinates, it increases the signal-to-noise ratio after DP noise addition, compensating (at least partially) for the accuracy loss due to noise. Second, resource-aware client selection yields a virtuous cycle—selecting capable devices reduces round latency, allowing more rounds in a fixed time budget, and more rounds improve convergence. Third, privacy amplification via subsampling is especially useful in edge FL, as the small selection fraction $q = K/N (e.g., q = 0.1$ in our experiments yields an 10 amplification factor, enabling a 3 reduction in the noise multiplier z while preserving the same ϵ.

The water-filling compression allocation is optimal under Gaussian gradient distributions, which may not hold for sparse or quantized networks. The resource announcing procedure adds a minor communication overhead (resource vectors) at the beginning of each round. The convergence guarantee in Theorem 2 is based on the assumptions of bounded gradient variance and compressibility, which can be violated for extremely heterogeneous data or very large models. Lastly, the synchronous aggregation protocol suffers from stragglers; to the best of our knowledge, combining AGC-DP with asynchronous FL is an interesting open question.

7 Conclusion

In this paper, we introduced AGC-DP, a general protocol that unifies adaptive gradient compression and differential privacy in a resource-limited edge federated learning scenario. We cast the client selection, the per-device compression ratio assignment, and the DP noise level as a joint constrained optimization problem and found tractable solutions by using Lagrangian duality together with Rényi DP composition-based privacy amplification. We provided a theoretical analysis to guarantee formal $(\epsilon, \delta) - DP$ and $O(1/\sqrt{T})$ convergence under non-IID data with error-feedback compression. The extensive experiments on the CIFAR-10, HAR, and FEMNIST datasets with 100 heterogeneous edge devices demonstrate that AGC-DP surpasses all baselines on the privacy-utility-communication trade-off, achieving up to 91.1% communication reduction with less than 1.5% accuracy degradation under an $\epsilon = 4$ privacy budget. AGC-DP shows that privacy, communication efficiency, and model utility are not fundamentally contradictory in edge federated learning — they can all be optimized simultaneously via principled co-design. We expect this work to spur further research at the intersection of privacy-preserving machine learning and resource-aware edge computing.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM CCS*, 2016, pp. 308–318.

[2] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proc. EMNLP*, 2017.

[3] N. Agarwal, A. T. Suresh, F. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," in *Proc. NeurIPS*, 2018.

[4] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. NeurIPS*, 2017.

[5] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Proc. NeurIPS*, 2018.

[6] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proc. ICML*, 2018.

[7] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM CCS*, 2017.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. TCC*, 2006.

[9] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. NeurIPS*, 2020.

[10] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients — How easy is it to break privacy in federated learning?" in *Proc. NeurIPS*, 2020.

[11] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv:1712.07557*, 2017.

[12] A. M. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. T. Suresh, "Shuffled model of federated learning: Privacy, accuracy and communication trade-offs," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 464–478, 2021.

[13] S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi, "Error feedback fixes SignSGD and other gradient compression schemes," in *Proc. ICML*, 2019.

[14] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. ICML*, 2020.

[15] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Oort: Efficient federated learning via guided participant selection," in *Proc. OSDI*, 2021.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Smola, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, 2020.

[17] Y. Lin, S. Han, H. Mao, Y. Wang, and W. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. ICLR*, 2018.

[18] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017.

[19] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *Proc. ICLR*, 2018.

[20] I. Mironov, "Rényi differential privacy," in *Proc. IEEE CSF*, 2017.

[21] I. Mironov, K. Talwar, and L. Zhang, "Rényi differential privacy of the Sampled Gaussian mechanism," *arXiv:1908.10530*, 2019.

[22] D. Shah, W. Lin, and V. Smith, "Compressed and private federated learning with error feedback," *arXiv:2108.xxxxx*, 2021.

[23] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. NeurIPS*, 2018.

[24] A. Triastcyn and B. Faltings, "Federated learning with Bayesian differential privacy," in *Proc. IEEE Big Data*, 2019.

[25] T. Vogels, S. P. Karimireddy, and M. Jaggi, "PowerSGD: Practical low-rank gradient compression for distributed optimization," in *Proc. NeurIPS*, 2019.

[26] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Proc. NeurIPS*, 2018.

[27] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous federated optimization," *arXiv:1903.03934*, 2019.

[28] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. NeurIPS*, 2019.

[29] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[30] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. NeurIPS*, 2017.

[31] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *Proc. ICLR*, 2021.

[32] Z. Chai, A. Ali, S. Zawad, S. Truex, A. Anwar, N. Baracaldo, Y. Zhou, H. Ludwig, F. Yan, and Y. Cheng, "TiFL: A tier-based federated learning system," in *Proc. ACM HPDC*, 2020.

[33] M. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.