

The Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS 2026)

FedAdaVR: Adaptive Variance Reduction for Robust Federated Learning under Limited Client Participation

S M Ruhul Kabir Howlader


Xiao Chen


Yifei Xie


Lu Liu


01 / CROSS-DEVICE FL BOTTLENECK

Extreme Partial Participation

< 1% – 2%

Active client cohort per round in real-world cross-device deployments

1. Stochastic Gradient Error

Local mini-batch sampling noise.

2. Client Drift Error

Non-IID local objective divergence.

3. Participation Error

Dominant unvisited client bias.

Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. FedVARP: Tackling the variance due to partial client participation in federated learning. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 906–916. PMLR, 01–05 Aug 2022.

Coordinate Imbalance Under Severe Label Skew

Frequent-Class Coordinates

Update Frequency: Sampled in nearly every round.

Signal Profile: Dense low-variance gradient trajectories.

Large Fixed Step η_s: Amplifies oscillation near optima.

Rare-Class Coordinates (LQ-1)

Update Frequency: Sporadic participation across rounds.

Signal Profile: Sparse high-variance gradient spikes.

Small Fixed Step η_s: Freezes learning on stale weights.

Key Insight: Variance reduction requires coordinate-wise adaptive step sizes to balance heterogeneous update frequencies.

Stage 1: SAGA-Style Server Variance Reduction

HISTORICAL CLIENT TABLE UPDATE (SERVER MEMORY)

$$y_j^{(t+1)} = g_j^{(t)} \quad \text{if } j \in S^{(t)} \quad \text{else } y_j^{(t)}$$

VARIANCE-REDUCED UPDATE DIRECTION (EQ. 6)

$$r^{(t)} = \frac{1}{M} \sum_{i \in S^{(t)}} (g_i^{(t)} - y_i^{(t)}) + \frac{1}{N} \sum_{j=1}^{N} y_j^{(t)}$$

Emulated Presence: Retains latest update for all N clients on server.

Unbiased Estimator: Active cohort corrects stale historical average.

Fast O(Md) Update: Maintained via running sum of active clients.

Zero Client Burden: Identical device computation to standard FedAvg.

Stage 2: Decoupled Adaptive Preconditioning

STEP 01

Unscaled Signal

Form pseudo-gradient $G(t) = \eta_c\, r(t)$.

STEP 02 • HIGHLIGHT

Clean Moments

Avoids distorting numerical stabiliser.

STEP 03 • HIGHLIGHT

Exact Debiasing

Restores early-round moment scaling.

STEP 04

Global Step

Supports Adagrad, Adam, Adabelief, Yogi, Lamb.

FedAdaVR-Quant: Quantised State Maintenance

$$r^{(t)} = \frac{1}{M}\sum_{i \in S^{(t)}} \left(g_i^{(t)} - \mathrm{DEQUANT}(y_i^{(t)})\right) + \frac{1}{N}\sum_{j=1}^{N} \mathrm{DEQUANT}(y_j^{(t)})$$

-50%

FP16 Storage

-75%

Int8

-87.5%

Int4

Full-precision FP32 client communication preserved; quantisation noise acts as a mild regulariser.

Non-Convex Convergence Guarantee

KEY THEORETICAL BREAKTHROUGH

Zero First-Order Participation Bias

Eliminates global data heterogeneity from dominant convergence terms.

Key Properties (Server-Side Adagrad)

Canonical Speed: Matches standard federated Adagrad under non-convex loss.

Higher-Order Drift Only: Confines non-IID dissimilarity to fast-decaying local drift.

Monotonic Cohort Scaling: Sampling variance drops steadily for larger active cohorts.

Full-Participation Limit: All partial-participation overheads vanish at full participation.

Extreme Low-Participation Benchmarks

MNIST & FMNIST

1% active client participation.

CIFAR-10 Vision

2% active ResNet-18 cohort.

Shakespeare NLP

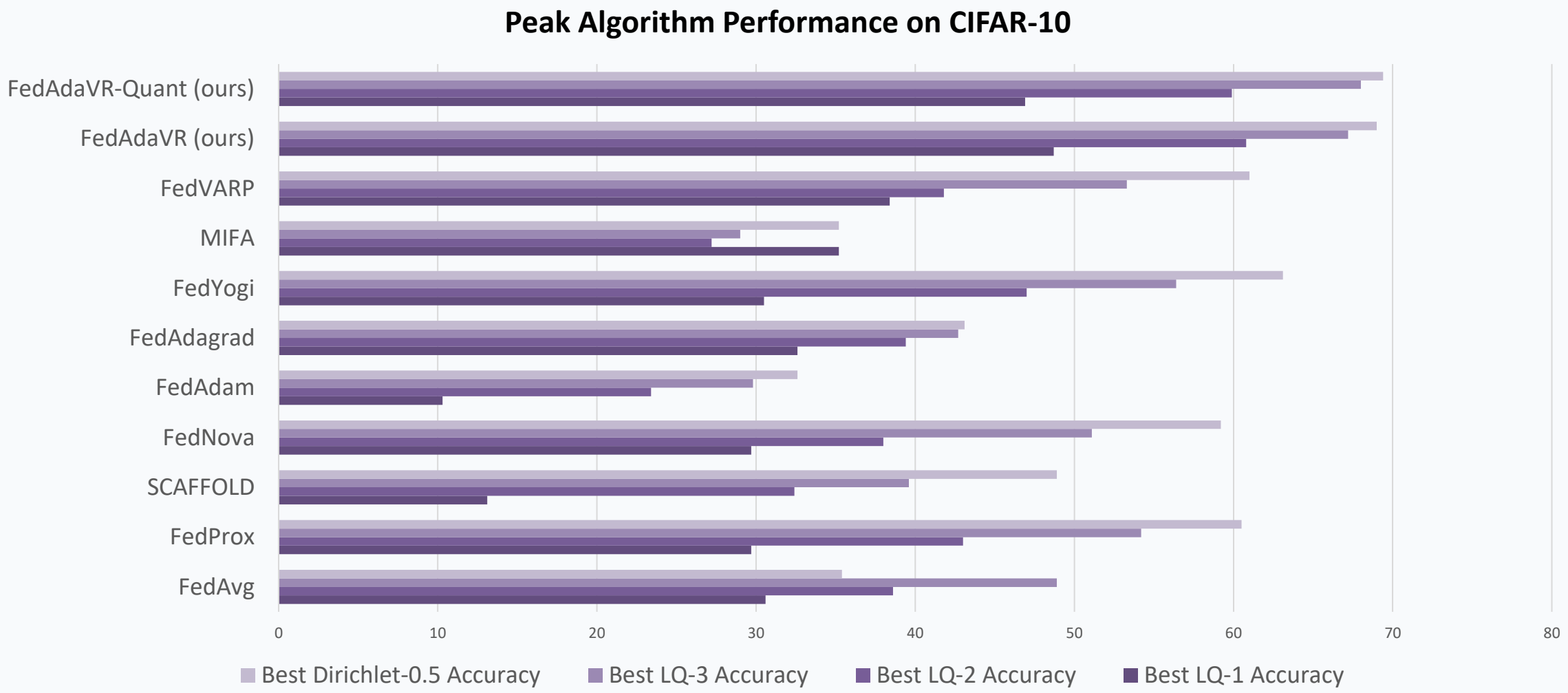
1% natural role participation.

7 Non-IID Regimes

Single-class LQ-1 extreme skew.

Evaluated against 9 state-of-the-art baselines
(FedAvg, FedProx, SCAFFOLD, FedNova, FedAdagrad, FedAdam, FedYogi, MIFA, FedVARP)

Outperforming State-of-the-Art



09 / ABLATION STUDY • CIFAR-10 (LQ-1 PARTITION)

Both Components Are Essential

FedAdaVR-NoVR

Severe late-stage loss spikes.

~39.0% Acc

FedAdaVR-NoOPT

Stagnant rare-class learning progress.

~43.5% Acc

Full FedAdaVR

Fast stable convergence trajectory.

~49.2% Peak

Variance reduction suppresses partial-participation oscillations; adaptive preconditioning accelerates rare-class updates.

Faster End-to-End Training to Target Accuracy

Algorithm	Client Compute	Server Memory	Rounds to 45%	Time to 45% Accuracy (Seconds)
FedAvg	O(K)	O(d)	1,396 (1.5×)	232,336
FedNova	O(K) + Norm	O(d)	1,491 (1.6×)	251,457
FedVARP	O(K)	O(Nd)	1,385 (1.5×)	242,749
FedAdaVR-Quant (Ours)	**O(K)**	**O(Nd_{quant})**	**1,118 (1.2×)**	**199,351**
FedAdaVR (Ours)	**O(K)**	**O(Nd)**	**918 (1.0×)**	**163,689**



SUMMARY & TAKEAWAYS

Thank You for Watching!

FedAdaVR unifies SAGA-like variance reduction with server-side adaptivity to eliminate first-order partial client participation bias, while its variant, FedAdaVR-Quant, achieves up to 87.5% server memory compression.

Authors Contacts

S M Ruhul Kabir Howlader: smrkh1@leicester.ac.uk
Xiao Chen: xiao.chen@ncu.edu.cn
Yifei Xie: xieyfzju@outlook.com
Lu Liu: l.liu3@exeter.ac.uk