COSAGE: FEDERATED LEARNING WITH GRADIENT SUMMARIES FOR CENTRALIZED CLIENT SELECTION

Houman Asgari[*]*, Stefano Rini*[†]*, and Andrea Munari*[‡]

[*] Technical University of Munich (TUM), Germany
[†] College of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan
[‡] Inst. of Communication and Navigation, German Aerospace Center (DLR), Germany

ABSTRACT

Heterogeneous data and partial participation hinder the effectiveness of federated learning (FL). To compare client selection policies under a common yardstick, we adopt the *Federated Learning with Gradient Summaries for Centralized Client Selection (FL-GSCCS)* model, where each client transmits a lightweight *gradient summary* for selection and only chosen clients perform full local training with sparsified updates. Within this framework, we propose COSAGE, a hybrid centralized policy that combines *Age of Information* (AoI) with gradient dissimilarity computed from a proxy update via the $\cos_4$ metric. Simulation results show that COSAGE consistently outperforms AoI-only and dissimilarity-only baselines in non-IID settings, and approaches the performance of clustering-based upper bounds without requiring client-to-client coordination or server access to client statistics.

***Index Terms*—** Federated learning, Client selection, Heterogeneous data distribution, Gradient similarity measures.

I. INTRODUCTION

Federated learning (FL) typically has to operate under pronounced data heterogeneity: clients collect data in disparate contexts, yielding non-IID, imbalanced, and often limited local datasets [1], [2]. In such settings, naïve aggregation (e.g., FedAvg [3]) can converge slowly, bias the global model, and generalize poorly. The challenge is amplified at scale by *partial participation*: per-round uplink budgets force the server to train with only a subset of clients. When this subset is chosen without regard to statistical diversity, the aggregated update may under-represent the population, further degrading convergence. This motivates investing a small additional uplink cost from all clients to expose selection-relevant information.

To tackle this problem, we formalize *Federated Learning with Gradient Summaries for Centralized Client Selection (FL-GSCCS)*: in every round, each client transmits a small gradient summary (typically one real value) that the server uses for participant selection; only the selected clients then perform full local training and upload sparsified updates. This framework provides a clean yardstick to compare selection policies under the same *uplink* budget which is the critical bottleneck for classic FL.

Relevant literature. Client selection has evolved from availability/throughput heuristics [3], [4] to content-aware criteria based on losses, norms, cosine similarity, or projections [5]–[12]. Clustering methods (CFL, FedGroup) [13], [14] promote diversity but assume stable metadata or server visibility into client statistics-assumptions that are often unrealistic and that add maintenance overhead. Joint selection–communication schemes (e.g., FedCG [15]) can be effective but are centralized and heavy. In contrast, FL-GSCCS is *cluster-free* and *lightweight*: a tiny per-client summary enables diversity- and fairness-aware selection without client-to-client coordination or server access to private features.

Contributions. Building on the *FL-GSCCS* system model [16], we develop and evaluate COSAGE, a mixed-policy client selection algorithm for FL. Specifically:

- We instantiate a hybrid centralized policy that combines (i) participation time-*Age of Information (AoI)* [17]-to ensure fairness, with (ii) gradient dissimilarity computed from a one-epoch proxy update via the $\cos_4$ metric to promote diversity under non-IID data.
- We provide explicit per-round communication accounting that distinguishes downlink broadcast, per-client summary overhead, and sparse uplink from selected clients.
- Through experiments on CIFAR-10 with VGG16/ResNet18 (feature extractors frozen) and top-k sparsification, we show that COSAGE consistently outperforms baselines that rely only on AoI or $\cos_4$ across heterogeneity levels and uplink budgets, and approaches the performance of a clustering-based upper bound-without requiring client-to-client coordination or server access to client statistics.

Notations: Calligraphic letters denote sets. The set $\{0, \dots, n-1\}$ is denoted by $[n]$. Bold lowercase letters represent vectors when the dimension is clear. Random variables

The work of S.R. is partially funded by the NSTC grant number 111-2221-E-A49 -068 -MY3. H.A. acknowledges the financial support of the Munich Aerospace scholarship within the group "Multi-access and Security Coding for Massive IoT Satellite Systems".

are denoted by uppercase letters, and their realizations by lowercase.

II. SYSTEM MODEL AND PROBLEM FORMULATION

II-A. System Model

Let us begin by detailing the FL-GSCCS setting. We consider U clients, each with private dataset $\mathcal{D}_u$, collaboratively minimizing a global loss $\mathcal{L}(\mathbf{w})$ over model weights $\mathbf{w} \in \mathbb{R}^m$:

$$\mathcal{L}(\mathbf{w}) \triangleq \frac{1}{|\mathcal{D}|} \sum_{u\in[U]} \mathcal{L}_u(\mathcal{D}_u, \mathbf{w}), \qquad \mathcal{D} \triangleq \biguplus_{u\in[U]} \mathcal{D}_u. \quad (1)$$

Each local dataset is composed of S shards. We divide the training process in four phases, repeated iteratively, described in the following and summarized in Fig. 1.

Phase 1– Global model update and local training: The parameter server (PS) distributes a global model update $\mathbf{g}(t)$ to all the local clients. Each local client performs SGD of the global loss function over the local dataset to obtain a local model update:

$$\mathbf{g}_u(t) \approx \nabla\mathcal{L}(\mathcal{D}_u, \mathbf{w}(t)). \quad (2)$$

Phase 2– Gradient summary transmission: Each local client transmits a *gradient summary* to the PS, following the idea of lightweight feedback in [16]. Specifically, each client reports a scalar

$$q_u(t) = f_q(\mathbf{g}(t-1), \mathbf{g}_u(t)), \quad (3)$$

where $f_q(\cdot)$ is a summary function that the current local gradient update and the previous global update to a single real-valued score. This summary is orders of magnitude smaller than $\mathbf{g}_u(t)$, so its overhead is negligible. We denote by $\mathbf{q}(t) = [q_0(t) \dots q_{U-1}(t)]$ the vector of gradient summaries transmitted by all the users to the PS.

Phase 3– Client selection: The PS selects a subset $\mathcal{S}(t)$ of clients, following policy

$$\mathcal{S}(t) = \pi(\mathbf{q}(t)) \quad (4)$$

with $\|\mathcal{S}(t)\| = S$ and broadcasts the selected client IDs in the downlink.

Phase 4– Gradient transmission and global model update: Clients in the set $\mathcal{S}(t)$ transmit their full gradient to the PS. The PS then produces a global model update as

$$\mathbf{g}(t+1) = \sum_{u\in\mathcal{S}(t)} \mathbf{g}_u(t). \quad (5)$$

II-B. Problem Formulation

Given a per-round *client budget* S, e.g., driven by the amount of available uplink resources, we consider the problem of choosing the gradient summarization function f_q and the client selection policy π that maximizes the learning performance. We focus specifically on the practically relevant case of f_q and π *not* being time dependent. Conceptually, one may seek to minimize the deviation of the global model from the ideal trajectory:

$$\min_{f_q,\pi} \sum_t \Delta_t, \qquad \Delta_t = L_k - \mathbb{E}[\mathcal{L}(t, \mathbf{w}_t)]\,. \quad (6)$$

Rather than attempting to solve (6) explicitly, in the remainder of this work we focus on a specific, practically motivated instantiation of f_q and π, and evaluate its impact on the learning trajectory empirically.

Note that a more complete model as in [16] would consider a given size of the gradient summary, say R and consider the objective function in (6) as a function of R. For simplicity, we consider here only the case of scalar summaries, i.e. $\mathsf{R} = 1$.

III. PROPOSED SOLUTION: COSAGE

We now introduce COSAGE, a hybrid centralized policy for client selection in the FL-GSCCS framework. The key idea is to balance two complementary criteria: temporal fairness and update diversity.

Age of Information. Each client u maintains an *age counter* $a_u(t)$, which is increased by 1 every round in which the client is not selected by PS. The counter is akin to AoI [18], [19]. To prevent repeatedly sampling recently active clients, we implement a cooldown via a *silent ratio* $\rho \in [0, 1]$: the youngest ρU clients are excluded from the candidate pool at round t.

Gradient dissimilarity. For each eligible client, the server requests a proxy update obtained from a short local training phase, yielding a proxy gradient $\nabla\tilde{w}_u(t)$. This proxy is summarized into a scalar quantity

$$f_q\big(\nabla w(t-1), \nabla\tilde{w}_u(t)\big) \triangleq \cos_4\big(\nabla w(t-1), \nabla\tilde{w}_u(t)\big), \quad (7)$$

where $\nabla w(t-1)$ denotes the previous global gradient. The $\cos_4$ similarity is defined as

$$\cos_4(\mathbf{g}_u, \mathbf{g}_v) = \frac{\langle \mathbf{g}_u, \mathbf{g}_v\rangle_4}{\|\mathbf{g}_u\|_4 \, \|\mathbf{g}_v\|_4}, \quad (8)$$

$$\langle \mathbf{u}, \mathbf{v}\rangle_4 = \tfrac{1}{4}(\|\mathbf{u}+\mathbf{v}\|_4 - \|\mathbf{u}-\mathbf{v}\|_4)\,, \quad (9)$$

with $\|x\|_4 \triangleq \left(\sum_i |x_i|^4\right)^{1/4}$. Unlike standard cosine similarity, $\cos_4$ emphasizes dominant gradient components, making it effective for identifying informative and dissimilar updates under non-IID data [16].

Hybrid selection. At each round, the server ranks clients by their dissimilarity scores, partitions them into S bins, and from each bin selects the client with the largest AoI. This ensures that the selected set combines both fresh participants and diverse updates.

Training and aggregation. Selected clients perform E epochs of local SGD, sparsify their gradients via top-k selec-

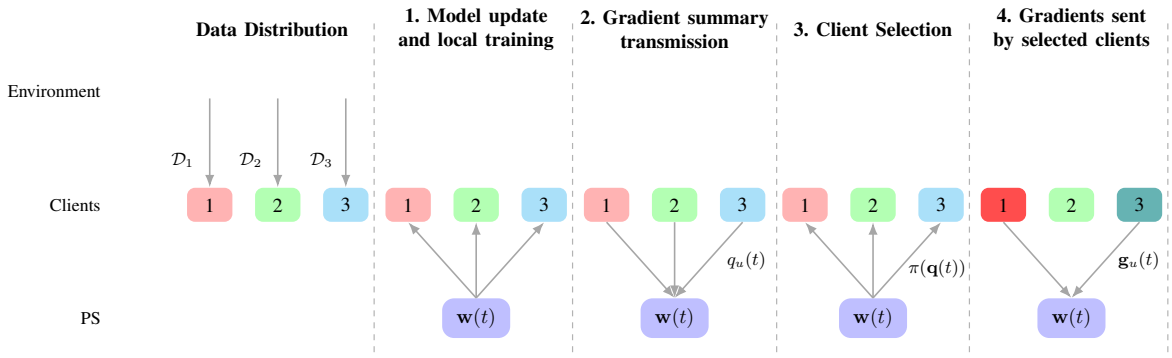

Fig. 1. Workflow phases for *Federated Learning with Gradient Summaries for Centralized Client Selection (FL-GSCCS).*

tion, and transmit them to the server. The server aggregates the received gradients as

$$w^{(t)} \leftarrow w^{(t-1)} + \frac{1}{S} \sum_{u \in \mathcal{S}(t)} \mathsf{Sparse}\big(\nabla \tilde{w}_u(t)\big). \tag{10}$$

For a gradient vector $g \in \mathbb{R}^m$, the operator $\mathsf{Sparse}(g)$ retains the k entries of g with the largest absolute values and sets the remaining $m-k$ entries to zero:

$$\mathsf{Sparse}(g)_i = \begin{cases} g_i, & i \in \text{Top-}k(|g|), \\ 0, & \text{otherwise.} \end{cases}$$

Conceptually, the $\cos_4$ metric accentuates dominant gradient directions, making it well-suited for capturing subtle but significant changes in client updates. Also, clients that pass the cool-down filter and exhibit high gradient dissimilarity are prioritized. This hybrid strategy balances the need to incorporate both fresh clients and diverse information.

IV. SIMULATION RESULTS AND DISCUSSIONS

We evaluate the proposed solution in two FL scenarios (targeting smaller and larger populations, respectively), on the CIFAR-10 dataset using a parameter-server architecture. All models are initialized with ImageNet-pretrained weights [20], and only the classifier heads are updated during training; the convolutional backbones remain frozen.

In the small-scale setting, we consider $U = 10$ clients with 5 shards each, using ResNet50. Two clients are selected per round. In the large-scale setting, $U = 100$ clients have 3 shards each, and VGG16 is used as the backbone, with 10 clients selected per round. In both cases, updates are aggregated via gradient averaging. As part of Phase 2 in the FL-GSCCS workflow, each client simulates a short local update over four mini-batches to generate a lightweight proxy gradient. This proxy is used to compute a scalar summary (e.g., gradient dissimilarity) transmitted to the server for centralized client selection in Phase 3. If selected, the client proceeds to Phase 4, performs one full local training epoch, and uploads a sparsified gradient

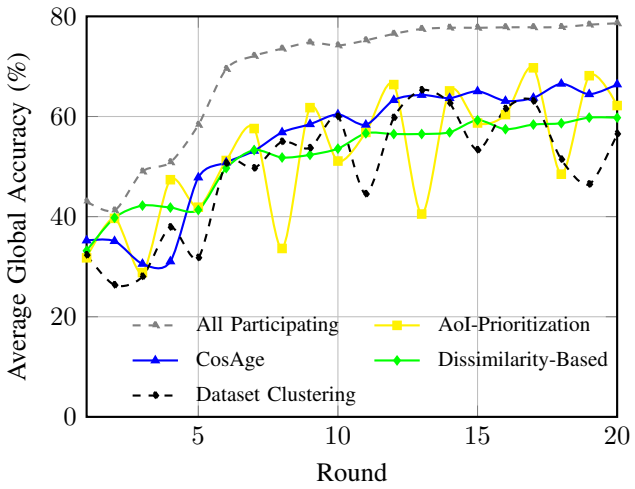

Fig. 2. Global accuracy comparison using ResNet50 and 10 clients with 5 shards each.

to contribute to the next global model update. Training is conducted using SGD with momentum 0.9, a learning rate of 0.01, and batch size 64. We adopt the standard cross-entropy loss for classification. To reduce communication overhead, only the top 10% of classifier-head gradients are transmitted. Results are averaged across 25 random seeds. We compare the following client selection strategies:

- **CosAge (proposed):** Clients are ranked by gradient dissimilarity with respect to the previous global update and partitioned into bins. From each bin, the client with the highest AoI is selected.
- **AoI-Based Prioritization:** Clients with the highest AoI are selected without considering gradient information, promoting fairness in participation.
- **Dissimilarity-Based Prioritization:** Clients with the most divergent updates (in terms of $\cos_4$ distance) are selected, promoting update diversity but disregarding freshness.

As additional references, we also report the performance of an all-participating upper bound (gray curves in the

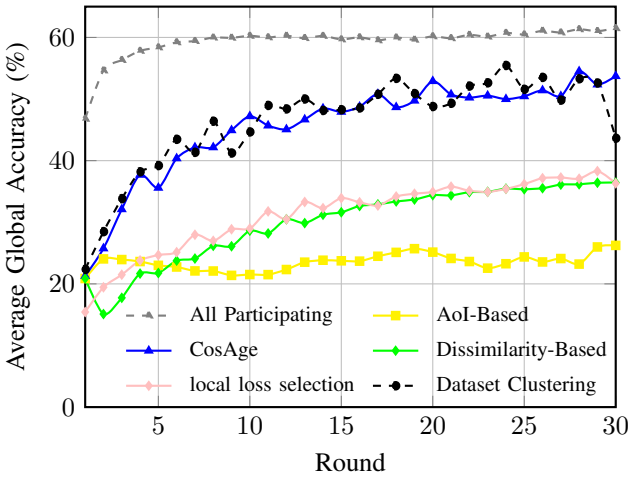

Fig. 3. Global accuracy comparison for different client selection strategies using VGG16 and 100 clients.

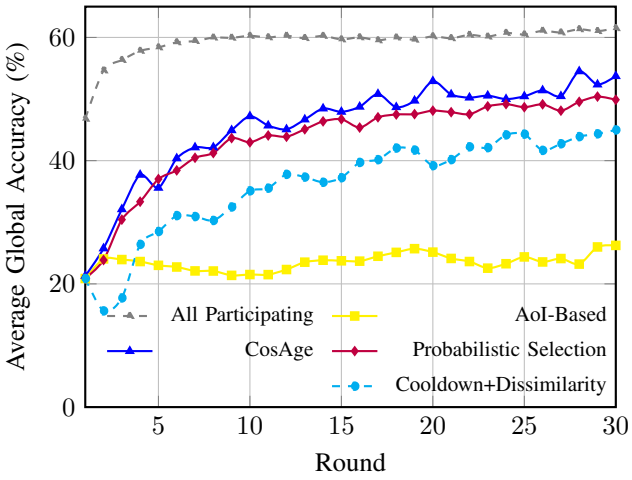

Fig. 4. Performance comparison of CosAge to other client selections mechanisms.

figures). Furthermore, we include a cluster-based client selection baseline (dashed black lines). In this setting, clients are first partitioned into clusters based on their local data distributions, obtained by applying k-means to label histograms. The number of clusters is set equal to the number of clients selected per round, and one client is selected from each cluster in a round-robin manner, providing a data-aware benchmark.

In the small-scale scenario (Fig. 2), all strategies perform relatively well, including AoI-based selection. This is likely due to the small population size and the relatively rich local datasets per client, which reduce both statistical heterogeneity and the impact of client selection. Nevertheless, our method shows smoother convergence and lower variance, suggesting more stable training under constrained participation. However, as the network scales up, the advantages of our method become more apparent. In the large-scale setting with $U = 100$ clients the training becomes significantly more sensitive to client selection (Fig. 3). AoI-only selection (yellow) quickly loses effectiveness and fails to sustain learning, while dissimilarity-only selection (green) suffers from slow convergence due to its lack of participation diversity. The local loss selection mechanism [7] (pink) also yields slower progress and lags behind CosAge policy. In contrast, our hybrid strategy, by combining temporal freshness (AoI) with gradient diversity ($\cos_4$ dissimilarity), maintains steady improvement and achieves performance close to the cluster-based oracle (black), despite not using any data distribution information.

These results demonstrate that our method scales well with system size and statistical heterogeneity, making it especially interesting for practical FL deployments. It consistently outperforms AoI-only and dissimilarity-only baselines and adapts to challenging conditions without relying on access to client data statistics.

Finally, we examine the performance of different client selection policies that incorporate similarity-based scoring. Fig. 4 reports the results using $\cos_4$ dissimilarities. The cyan curve corresponds to a variant where the cooldown mechanism is enforced: only the top 50% of clients in terms of AoI are considered, and dissimilarity is evaluated within this restricted set to discourage repeated participation of recently active clients. The purple curve illustrates a probabilistic selection rule, where each candidate is assigned a weight proportional to its dissimilarity score. Formally, letting s_i denote the similarity-based score of client i, we construct shifted values $\tilde{s}_i \geq 0$ and define the selection probabilities as

$$p_i = \frac{\tilde{s}_i}{\sum_j \tilde{s}_j}. \tag{11}$$

A fixed number of representatives is then sampled without replacement according to $\{p_i\}$. This randomized mapping favors clients with more diverse gradient summaries while still allowing occasional exploration of lower-ranked candidates. As a benchmark, we also include an AoI-only policy (yellow) and an all-participating upper bound (grey). Our findings suggest that CosAge offers a more sensitive measure of update diversity in high-dimensional settings.

V. CONCLUSION

We studied client selection in the FL-GSCCS framework using lightweight gradient summaries. By combining the Age of Information with update diversity via gradient dissimilarity, the proposed policy achieves stable convergence, scales effectively under heterogeneous data and approaches clustering-based upper bounds without requiring access to client statistics.

VI. REFERENCES

[1] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[2] J. Zhang, Z. Li, B. Li, J. Xu, S. Wu, S. Ding, and C. Wu, "Federated learning with label distribution skew via logits calibration," in *in Proc. Int. Conference on Machine Learning*. PMLR, 2022, pp. 26 311–26 329.

[3] B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," *Google Research Blog*, vol. 3, 2017.

[4] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE ICC Int. Cnfererence on Communications*. IEEE, 2019.

[5] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-iid data with reinforcement learning," in *IEEE INFOCOM*. IEEE, 2020.

[6] M. Ribero and H. Vikalo, "Communication-efficient federated learning via optimal client sampling," *arXiv preprint arXiv:2007.15197*, 2020.

[7] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," *arXiv preprint arXiv:2010.01243*, 2020.

[8] O. Marnissi, H. E. Hammouti, and E. H. Bergou, "Client selection in federated learning based on gradients importance," in *AIP Conference Proceedings*, vol. 3034, no. 1. AIP Publishing, 2024.

[9] R. Liu, Y. Cao, M. Yoshikawa, and H. Chen, "Fedsel: Federated SGD under local differential privacy with top-k dimension selection," in *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I 25*. Springer, 2020, pp. 485–501.

[10] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "Fedcorr: Multi-stage federated learning for label noise correction," in *Proc. IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 10 184–10 193.

[11] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, J. Cao, and H. Guan, "Gpfl: Simultaneously learning global and personalized feature information for personalized federated learning," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023.

[12] Y. Zeng, L. Liu, L. Liu, L. Shen, S. Liu, and B. Wu, "Global balanced experts for federated long-tailed learning," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023.

[13] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints," *arXiv preprint arXiv:1910.01991*, 2019.

[14] M. Duan, D. Liu, X. Ji, R. Liu, L. Liang, X. Chen, and Y. Tan, "Fedgroup: Efficient federated learning via decomposed similarity-based clustering," in *Proc. IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking*, 2021.

[15] Y. Xu, Z. Jiang, H. Xu, Z. Wang, C. Qian, and C. Qiao, "Federated learning with client selection and gradient compression in heterogeneous edge systems," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 5446–5461, 2024.

[16] L. Li, Y. Liu, Y. Ning, S. Rini, and J. Chen, "PNCS: Power-Norm cosine similarity for diverse client selection in federated learning," *arXiv preprint arXiv:2506.15923*, 2025.

[17] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Select. Areas Commun.*, vol. 39, no. 5, 2021.

[18] S. Kaul, R. Yates, and M. Gruteser, "On piggybacking in vehicular networks," in *Proc. IEEE GLOBECOM*, Dec 2011.

[19] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *Proc. IEEE SECON*, June 2011.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.