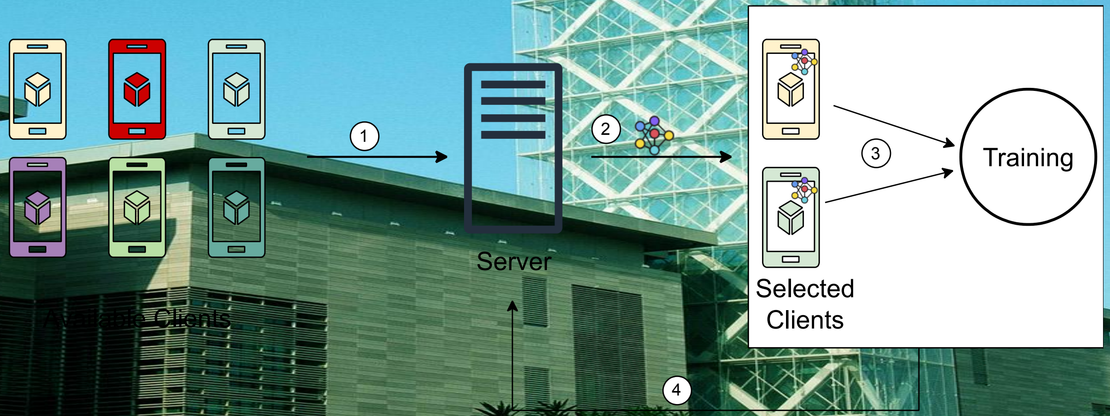
FilFL: Client **Fil**tering for Optimized Client Participation in **F**ederated **L**earning

Salma Kharrat, Fares Fourati, Vaneet Aggarwal, M-Slim Alouini, Marco Canini

KAUST

Federated Learning

In every communicaiton round:

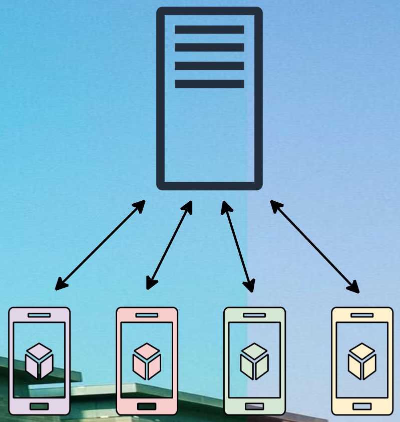


- Available clients check in with the server.
- Server selects subset of the available clients and broadcasts the latest version of the global model
- [illegible] local data starting from the received [illegible]
- [illegible]

Problem: Can we optimize client selection in FL

Challenge 1: Large number of clients

- ➢ partial client participation

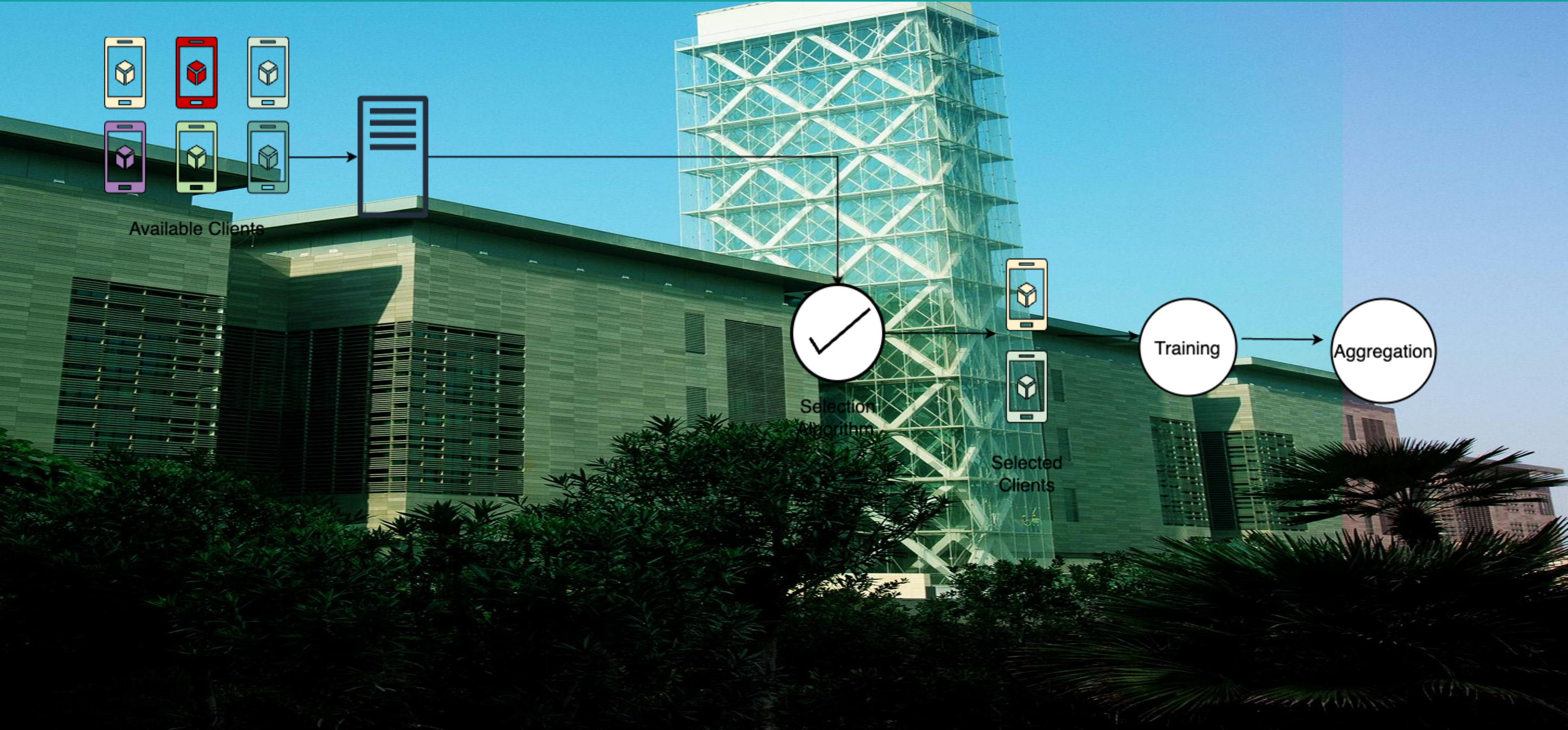


Challenge 2: Heterogeneity (e.g., data, device, behavior)

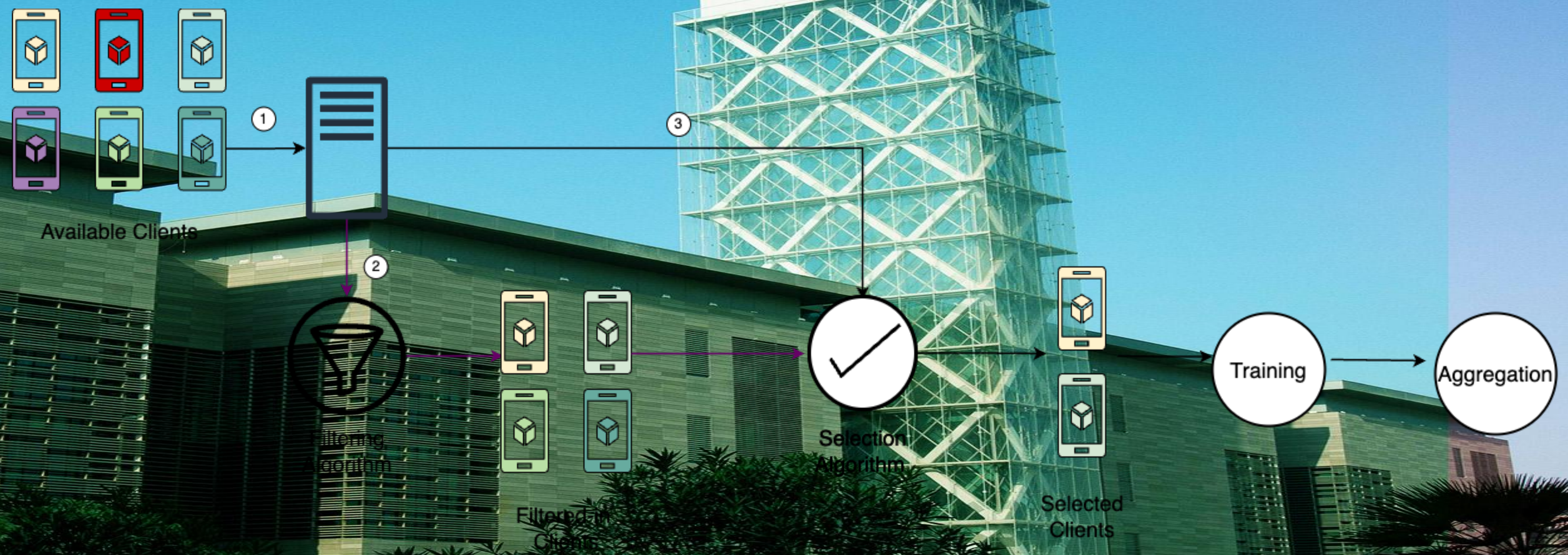
- ➢ optimize client selection

Problem: Previous approaches rely on selecting participants from the entire available pool without considering whether they are all appropriate for collaboration at the current stage of the training process.

Solution: Introduce client filtering in FL

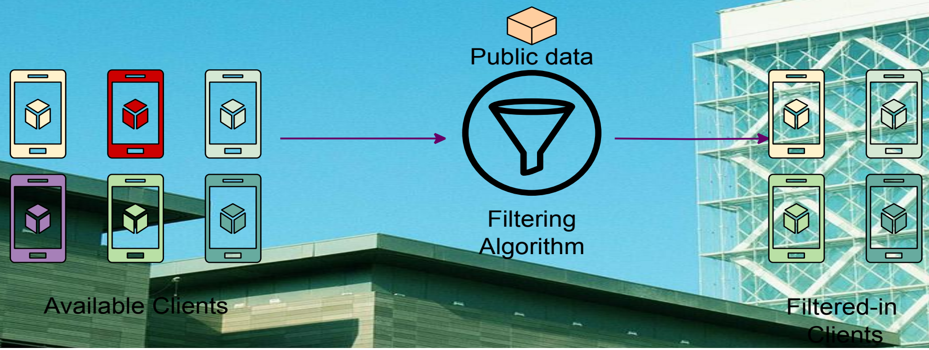


Solution: Introduce client filtering in FL



Filtering Algorithm … clients to be considered …

Solution: Introduce client filtering in FL

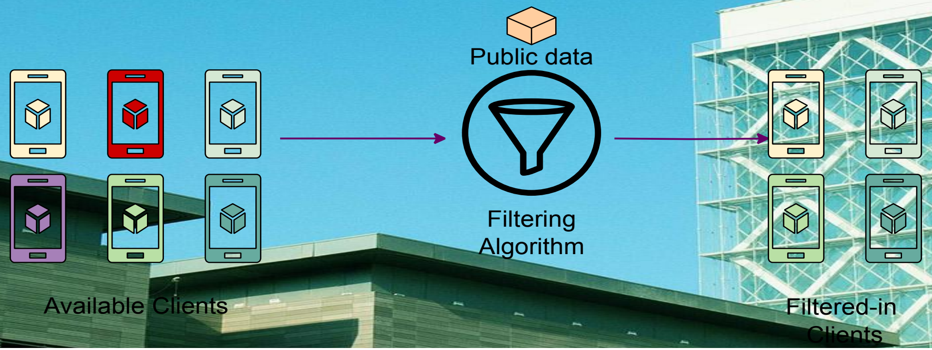


Filtering Objective. Our filtering objective is to find a subset of clients $\mathcal{S}_t^f$ that approximates a solution to the following combinatorial maximization problem:

$$\max_{\mathcal{S} \in \mathcal{S}_t} \left\{ \mathcal{R}(\mathcal{S}) \triangleq \mathcal{C} - F^{\mathcal{P}} \left(\frac{\mathbf{1}}{|\mathcal{S}|} \sum_{\mathbf{k} \in \mathcal{S}} \mathbf{w}_{\mathbf{t}}^{\mathbf{k}} \right) \right\} \tag{2}$$

where $\mathcal{C}$ is a sufficiently large constant, such that $\mathcal{R}(\mathcal{S})$ is positive, $\mathbf{w}_t^k$ is the weight of the k^{th} client in round t, and $F^{\mathcal{P}}(\mathbf{w}) \triangleq \frac{1}{m} \sum_{j=1}^{m} \ell(\mathbf{w}; x_j)$ as the loss on the server-held public dataset $\mathcal{P}$, which has m training data: $x_1, x_2, \cdots, x_m$.

Solution: Introduce client filtering in FL



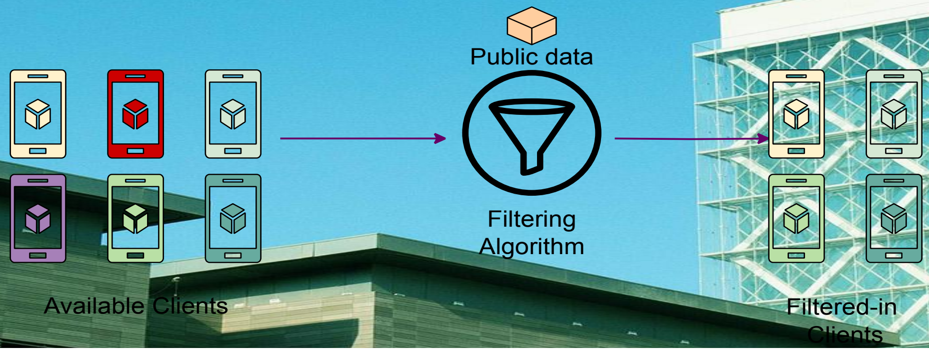
Solving this problem exactly requires exponential queries to the objective function!

Filtering Objective. Our filtering objective is to find a subset of clients $\mathcal{S}_t^f$ that approximates a solution to the following combinatorial maximization problem:

$$\max_{\mathcal{S} \in \mathcal{S}_t} \left\{ \mathcal{R}(\mathcal{S}) \triangleq \mathcal{C} - F^{\mathcal{P}} \left(\frac{\mathbf{1}}{|\mathcal{S}|} \sum_{\mathbf{k} \in \mathcal{S}} \mathbf{w}_{\mathbf{t}}^{\mathbf{k}} \right) \right\} \tag{2}$$

where $\mathcal{C}$ is a sufficiently large constant, such that $\mathcal{R}(\mathcal{S})$ is positive, $\mathbf{w}_t^k$ is the weight of the k^{th} client in round t, and $F^{\mathcal{P}}(\mathbf{w}) \triangleq \frac{1}{m} \sum_{j=1}^{m} \ell(\mathbf{w}; x_j)$ as the loss on the server-held public dataset $\mathcal{P}$, which has m training data: $x_1, x_2, \cdots, x_m$.

Solution: Introduce client filtering in FL



We use a greedy algorithm instead for non-monotone combinatorial maximization, which approximates the solution in linear time!

Filtering Objective. Our filtering objective is to find a subset of clients $\mathcal{S}_t^f$ that approximates a solution to the following combinatorial maximization problem:

$$\max_{\mathcal{S} \in \mathcal{S}_t} \left\{ \mathcal{R}(\mathcal{S}) \triangleq \mathcal{C} - F^{\mathcal{P}} \left(\frac{\mathbf{1}}{|\mathcal{S}|} \sum_{\mathbf{k} \in \mathcal{S}} \mathbf{w_t^k} \right) \right\} \tag{2}$$

where $\mathcal{C}$ is a sufficiently large constant, such that $\mathcal{R}(\mathcal{S})$ is positive, $\mathbf{w}_t^k$ is the weight of the k^{th} client in round t, and $F^{\mathcal{P}}(\mathbf{w}) \triangleq \frac{1}{m} \sum_{j=1}^{m} \ell(\mathbf{w}; x_j)$ as the loss on the server-held public dataset $\mathcal{P}$, which has m training data: $x_1, x_2, \cdots, x_m$.

Theoretical guarantees

Theorem 1. Under some assumptions (L-smoothness, μ-strong convexity, bounded variance of stochastic gradients, gradient norms and heterogeneity) we have:

$$\mathbb{E}\left[\| \bar{w}_{t+1} - w^* \|^2 \right] \leq \mathcal{O}\left(\frac{1}{t}\right) + \mathcal{O}(\varphi)$$

The above result guarantees the convergence rate of $\mathcal{O}\left(\frac{1}{t}\right)$ of FilFL up to a certain neighborhood $\mathcal{O}(\varphi)$, which depends on the quality of filtering. The φ-term encodes the approximation error of the filtering algorithm.

$\bar{\mathbf{w}}_t \triangleq \sum_{k\in[N]} p_k \mathbf{w}_t^k$

$\mathbf{w}^* \in \arg\min_{\mathbf{w}} F^{\mathcal{D}}(\mathbf{w})$

$F^{\mathcal{D}}(\mathbf{w}) \triangleq \sum_{k=1}^{N} p_k F_k(\mathbf{w})$

Client Filtering enhances FL algorithms

Best test Accuracy over Rounds			
	CIFAR-10	FEMNIST	Shakespeare
FedAvg	68%	70%	45%
FilFL	**75%**	**78%**	**55%**

Table: Best achieved test accuracy for FedAvg vs FilFL both using PoC [illegible] selection method.

FilFL sensitivity to Hyperparameters

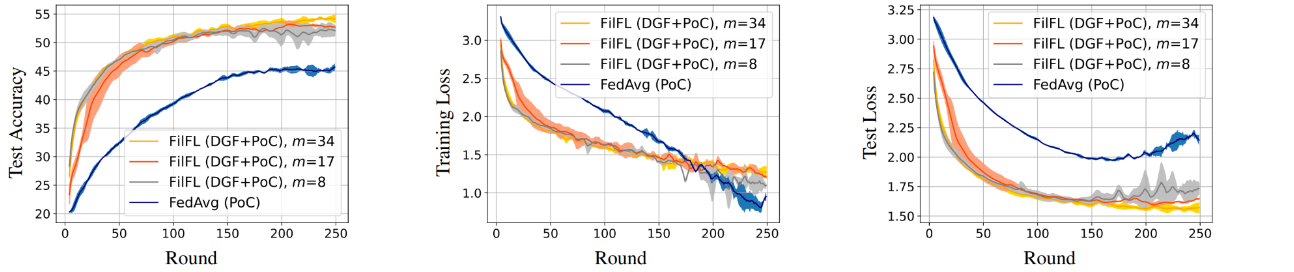


Figure 4: FilFL (FedAvg with DGF) sensitivity to public dataset size m on Shakespeare dataset with PoC for client selection, $N = 143$, $n = 100$, $K = 10$, and $h = 5$.

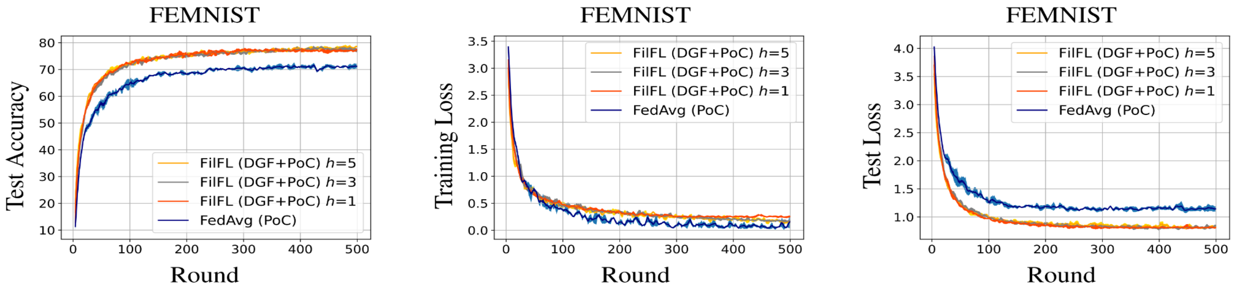


Figure 13: FilFL (FedAvg + χGF + PoC) sensitivity to periodicity h

Filtering Behavior

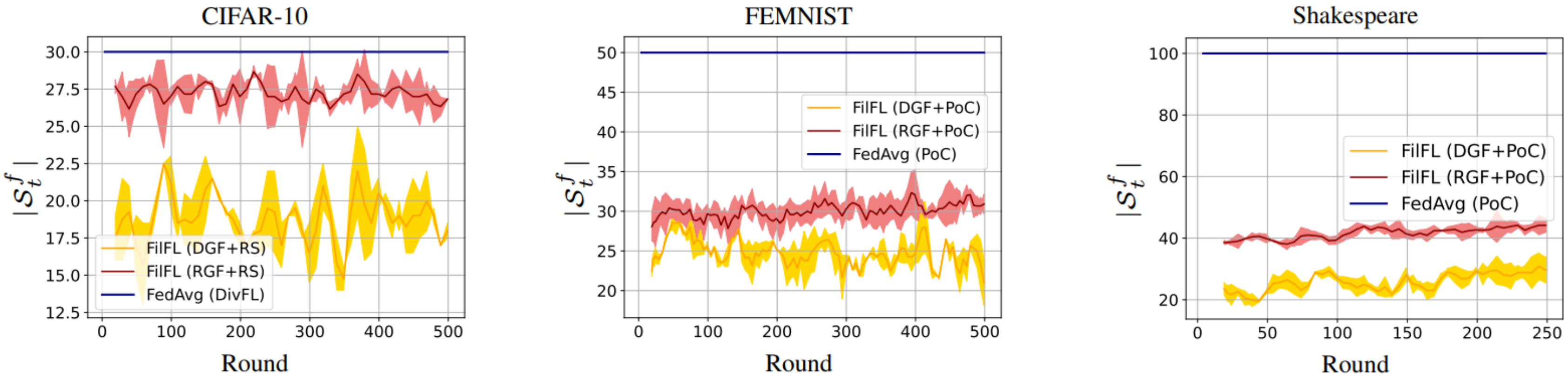


Figure 5: The number of filtered-in clients, denoted as $|\mathcal{S}_t^f|$, for FilFL (FedAvg with χGF), over the rounds in different settings of CIFAR-10, FEMNIST, and Shakespeare datasets, with available clients n being 30, 50, and 100, respectively. For FedAvg without filtering, we consider $\mathcal{S}_t^f$ to be equal to $\mathcal{S}_t$.

Approximation Ratio

Approximation Ratio. Fig. 6 shows the approximation ratios of both χGF versions compared to the optimal filtering (OPT) on CIFAR-10 with $N = 200$ and $n = 10$, which we find by evaluating $2^n - 1$ combinations. We find that both χGF versions achieve approximation ratios higher than 0.96, meaning that $\mathcal{R}(\mathcal{S}_t^f) \geq 0.96\mathcal{R}(OPT)$ over the multiple rounds. This indicates that greedy filtering identifies near-optimal combinations of clients.

Filtering Performance. The filtering performance can be measured by the improved FL performance and the higher approximation ratios. Since both versions of χGF show similarly high ratios and improved FL performance, both can be considered effective for filtering.

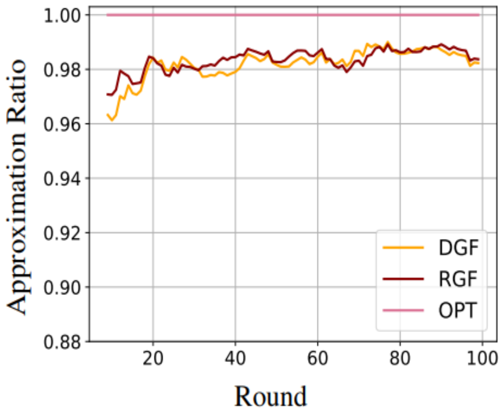


Figure 6: Approximation ratios of filtering objective solution on CIFAR-10 dataset.

Conclusion

We proposed client filtering as a promising technique to optimize client participation and training in FL.

Our proposed FL algorithm, FilFL, which incorporates our greedy filtering algorithm, has:

- Theoretical **convergence guarantees**.
- **Better learning efficiency**.
- **Accelerated convergence**.
- **Higher test accuracy** across different vision and language tasks.
- Potentially more robust selecti [illegible]

References

1. Buchbinder, N., Feldman, M., Seffi, J., & Schwartz, R. (2015). A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, *44*(5), 1384-1402.

1. Cho, Y. J., Wang, J., & Joshi, G. (2020). Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*.

1. Fourati, F., Aggarwal, V., Quinn, C., & Alouini, M. S. (2023, April). Randomized greedy learning for non-monotone stochastic submodular maximization under full-bandit feedback. In *International Conference on Artificial Intelligence and Statistics* (pp. 7455-7471). PMLR.

1. Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

1. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, *2*, 429-450.

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282).

Test Accuracies

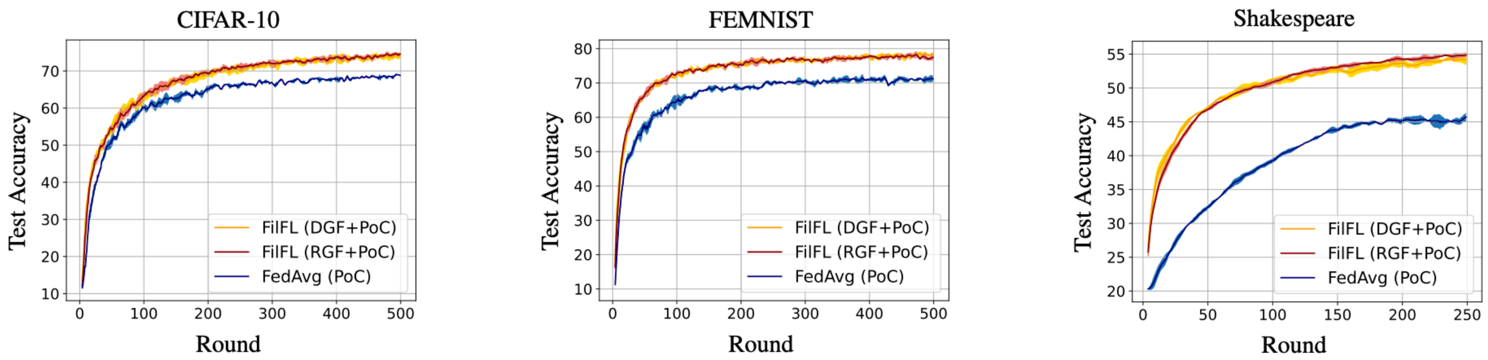


Figure 7: FilFL vs FedAvg test accuracies both using PoC as a client selection method.

Assumptions

Assumption 1. *$F_1, \cdots, F_N$ are all L-smooth*[7].

Assumption 2. *$F_1, \cdots, F_N$ are all μ-strongly convex*[8].

Assumption 3. *Let ψ_t^k be sampled from the k-th client's local data uniformly at random. The variance of stochastic gradients in each client is bounded*[9] *by σ_k^2.*

Assumption 4. *The norms of the stochastic gradients are uniformly bounded by G*[10].

Assumption 5. *Statistical heterogeneity defined as $F^* - \sum_{k \in [N]} p_k F_k^*$ is bounded, where $F^* := \min_{\mathbf{w}} F(\mathbf{w})$ and $F_k^* := \min_{\mathbf{v}} F_k(\mathbf{v})$.*

Assumption 6. *Assume $\mathcal{A}_t$ contains a subset of K indices randomly selected with replacement according to the sampling probabilities $p_1, \cdots, p_N$, with simple averaging for aggregation* [11].