Communication-Efficient Federated Learning via Optimal Client Sampling

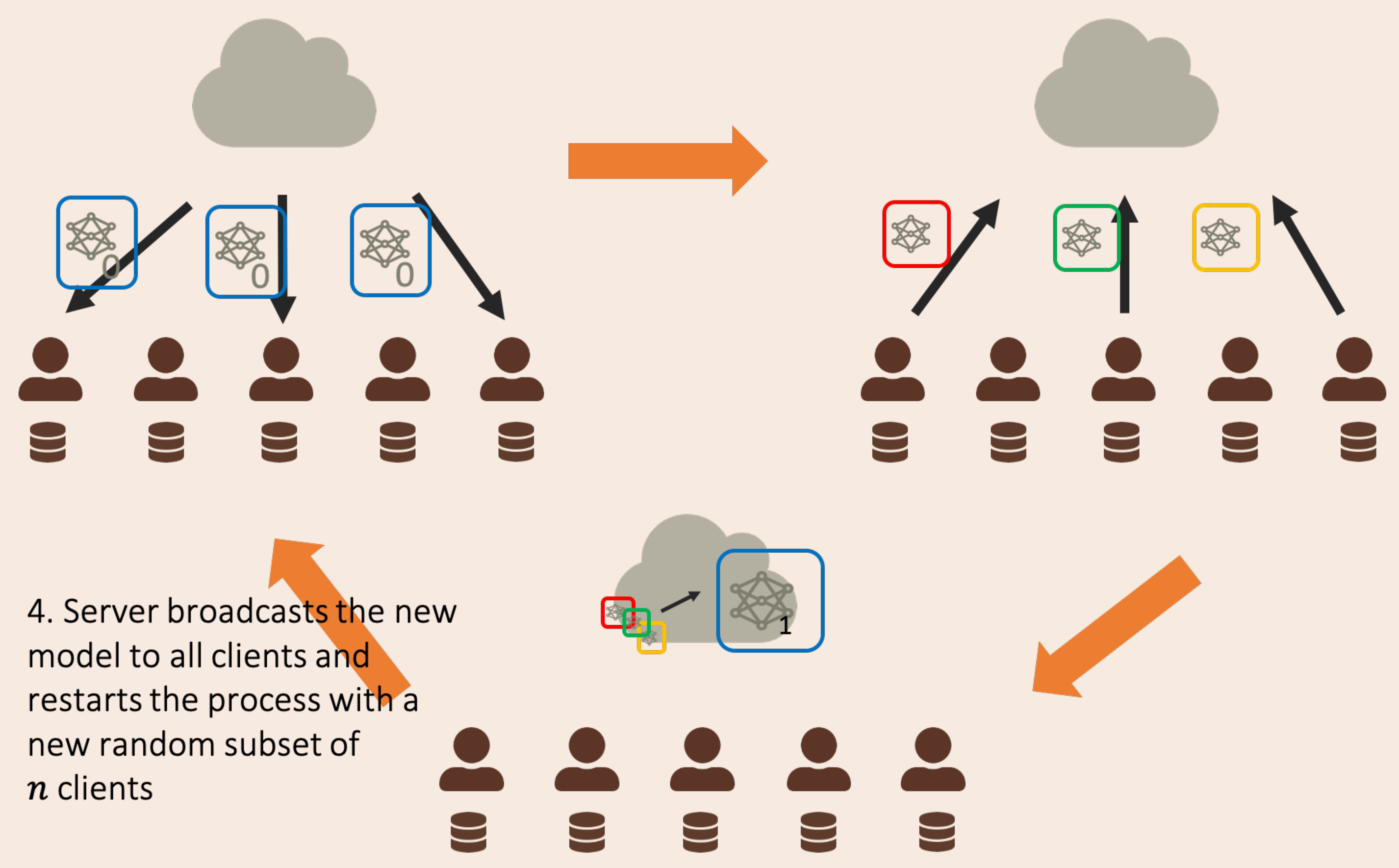
Mónica Ribero[1], Haris Vikalo[1]

[1]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas. United States.

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Federated Learning

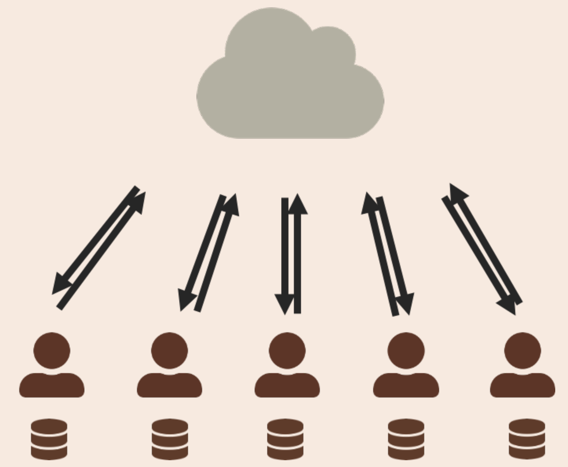
Federated learning provides a private and efficient way of learning machine learning models when the data is distributed across many clients[1]



Communication challenge

Federated Learning requires massive communication
- Millions of clients
- Thousands of iterations
- Models ~ 1-20 MB

Most of current communication reduction strategies focus on compressing the model [2-7] using a variety of strategies:
- Reducing network architecture
- Quantization
- Sparsification
- Low rank matrices



Research Goal

Reduce communication by reducing the amount of clients communicating with the server. Only the clients whose updates are deemed informative communicate their updates.
- Clients train locally
- Clients assess how informative their update is and decide to communicate or not.

Methods

We model the progression of each user's vector of weights during SGD as a stochastic process.

OU process definition: An Ornstein-Uhlenbeck processes is a stationary Gauss-Markov processes that, over time, tends to drift towards a mean value. Formally, let W_t be a standard Wiener process, then an OU process is defined by

$$d\theta_t = \lambda(\mu - \theta_t) + \sigma W_t$$

SGD as an OU process: Consider the loss function $\mathcal{L}(\theta; X) = \sum_{i=1}^{N} \ell_i(\theta)$ where X is a dataset with N samples. In SGD, $\mathcal{L}$ is minimized by evaluating an approximation of the gradient on a mini-batch $\mathcal{S} \subseteq X$,

$$\theta_{t+1} \leftarrow \theta_t - \eta \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} g_i(\theta)$$

Under some mild assumptions,

$$\Delta\theta = \theta_{t+1} - \theta_t \approx \eta g(\theta) - \sqrt{\frac{\eta}{N}} B\, \mathcal{N}(0, \eta I)$$

where BB^T approximates the covariance of gradients. Essentially, this is a discretization of an OU process.

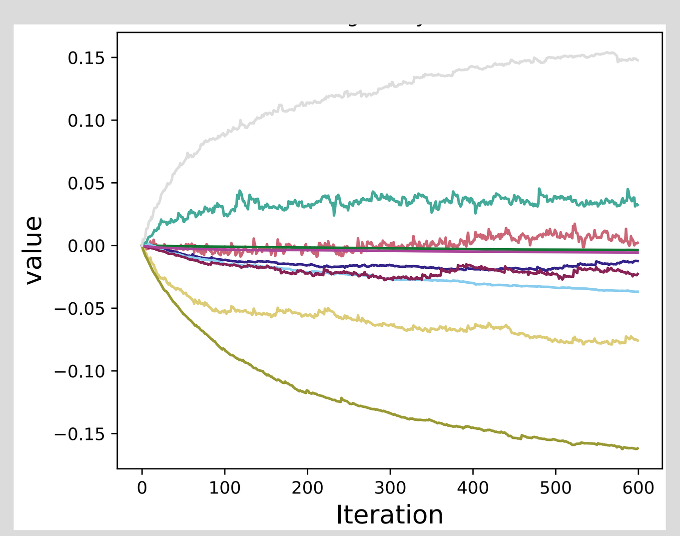


Figure 1. Weight updates can be interpreted as a realization of an OU process.

Optimal sampling: The following thresholding strategy outperforms deterministic sampling; the threshold is derived from a frequency constraint [8],

$$|\theta_t - E[\theta_t|\theta_0]| > \gamma$$

Two different strategies for selecting the threshold:

- **Fixed Threshold (FT):** Fix a threshold γ for the entire duration of the iterative process.
- **Adaptive Threshold (AT):** Compute threshold γ_t based on the previous updates' statistics.

Benchmark the proposed strategies vs. two baselines:

- Full communication
- Randomly dropping clients to match the communication rate of the best approach (FT or AT).

Experiments

We test our methods in two settings:

1. A classification task on EMNIST. To this end, we train and use a convolutional neural network.
2. A realistic language modeling task using the Stackoverflow dataset. For this, we train a recurrent neural network on a next word prediction task.

At each round, 50 clients are uniformly selected to update the model. Each client locally trains for $E = 20$ epochs on EMNIST and $E = 1$ epochs for Stackoverflow, using SGD. We train each model for 100 rounds.

Results and Discussion

Table 2: Results for convMNIST and Stackoverflow

	Accuracy		**Accuracy rate** (acc/byte)		**Overall comm.**(GB)		**Communication used**(%)	
Dataset	EMNIST	Stack	EMNIST	Stack	EMNIST	Stack	EMNIST	Stack
Baseline	**97.5%**	**13.07 %**	**29.3**	1.61	**33.27**	81.0	**100%**	100 %
FT	95.4%	**13.63 %**	**150**	2.34	**6.3**	58.3	**19%**	71.96 %
AT	**97.4%**	13.04 %	34.9	**2.99**	27.9	**43.6**	83%	**54 %**
Random	94.2 %	9.67 %	148.9	2.22	**6.3**	**43.6**	**19%**	**54 %**

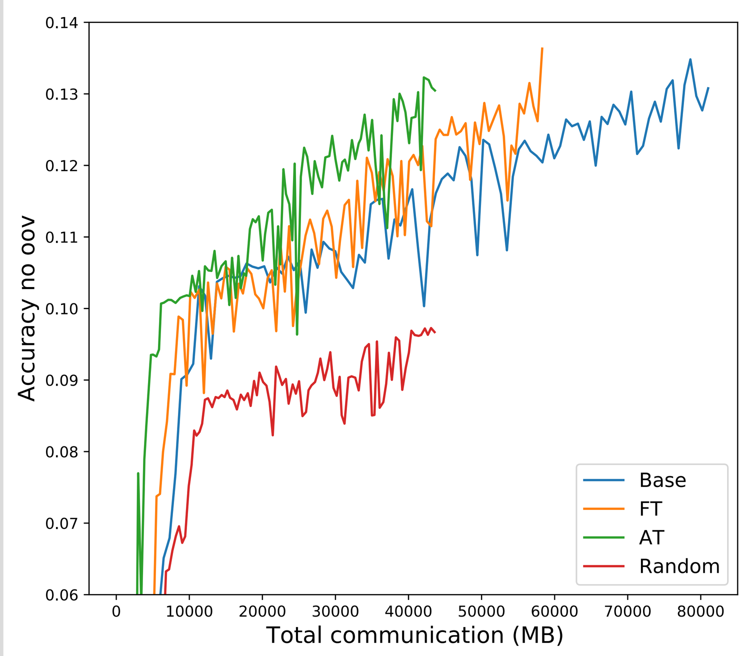


Figure 2. Accuracy rate per amount of communication on Stackoverflow.

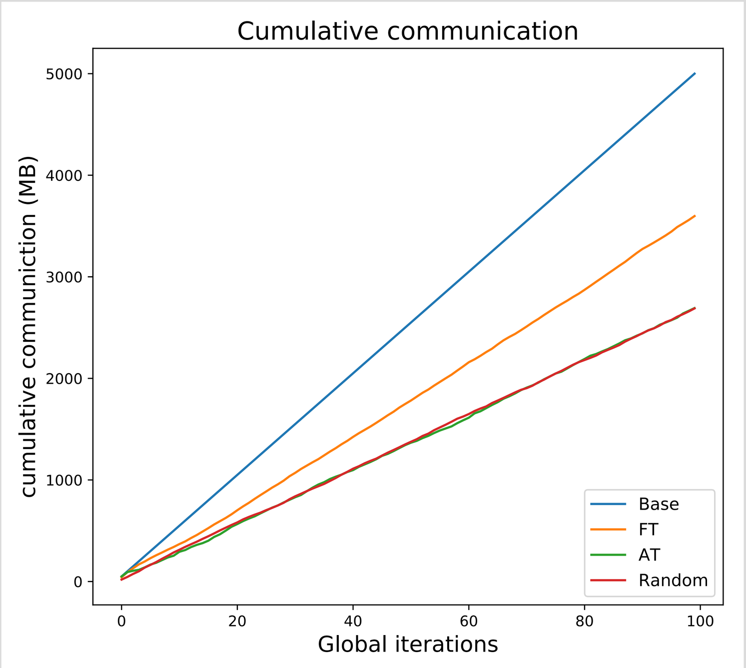


Figure 3.Overall communication on Stackoverflow.

- Randomly dropping clients reduces communication but deteriorates performance significantly.
- Thresholding techniques reduce communication while achieving better performance than random selection.
- The accuracy rate per amount of communication is higher for thresholding strategies than for either random selection or full communication.
- For both datasets, at least one thresholding strategy achieves the same performance as the full communication baseline while considerably reducing communication.
- Our approaches can be combined with compression strategies to lower the communication rates even further.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., andy Arcas, B. A. Communication-Efficient Learning ofDeep Networks from Decentralized Data. In Singh, A.and Zhu, J. (eds.),Proceedings of the 20th InternationalConference on Artificial Intelligence and Statistics, vol-ume 54 ofProceedings of Machine Learning Research, pp. 1273–1282, Fort Lauderdale, FL, USA, 20–22 Apr2017. PMLR. URL: http://proceedings.mlr.press/v54/mcmahan17a.html.
2. A.T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan. Distributed mean estimation with limited communication. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3329–3337. JMLR. org, 2017
3. J. Konečný, H.B. McMahan, F. X.Yu, P.Richtárik, A. T. Suresh, D. Bacon (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016.
4. H. Tang, X. Lian, T. Zhang, and J. Liu. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. arXiv preprint arXiv:1905.05957, 2019.
5. J. Konečný, and P. Richtárik. Randomized distributed mean estimation: Accuracy vs. communication.365Frontiers in Applied Mathematics and Statistics, 4:62, 2018.
6. D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. Qsgd: Communication-efficient SGD via gradient quantization and encoding. In Advances in Neural Information Processing Systems, pages3101709–1720, 2017.
7. S. Horvath, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtarik. Natural compression for356distributed deep learning.arXiv preprint arXiv:1905.10988, 2019.
8. N. Guo and V. Kostina. Optimal causal rate-constrained sampling for a class of continuous markov processes.arXiv preprint arXiv:2002.01581, 2020.