Federated Learning

A Signal Processing Perspective

WEN Hao

Main Problems

Scenario

Wireless cross-device

Main Problems

$$\theta_t^i \xrightarrow{\text{encode}} s_t^i \xrightarrow{\text{transmit}} x_t^i \xrightarrow{\text{aggregate}} y_t \xrightarrow{\text{combine}} \theta_t$$

local model | global model

[1]T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated Learning: A Signal Processing Perspective," *IEEE Signal Processing Magazine*, vol. 39, pp. 14–41, 5 2022

Encode

$$\text{encoded model} \rightarrow s_t^i = \phi_i(\theta_t^i)$$

Purpose and Methods

- Compression
  - sparsification
  - quantization
  - $\cdots$
- Privacy
  - multi-party encryption (computation, MPC)
  - homomorphic encryption (HE)
  - differential privacy (DP)
  - $\cdots$

Transmission

Wireless scenario

$$\text{channel input} \rightarrow x_t^i = \varphi_i(s_t^i \leftarrow \text{encoded model})$$

Key Points

- Learning-Aware Resource Allocation
- Over-the-Air Federated Learning (AirFL)

User Selection & Resource Management

RM: relevant only for users that communicate over the same media

$\rightarrow$ each user having a separate channel with the server.

- Random selection
- Delay Minimization With Probabilistic User Selection
  - Probabilistic User Selection

$$\rho_t^i = \alpha_t \frac{\|\theta_t^i - \theta_{t-E}^i\|}{\sum_{i=1}^{N} \|\theta_t^i - \theta_{t-E}^i\|} + (1-\alpha_t) \frac{\max_j d_j - d_i}{N \max_j d_j - \sum_{i=1}^{N} d_i}$$

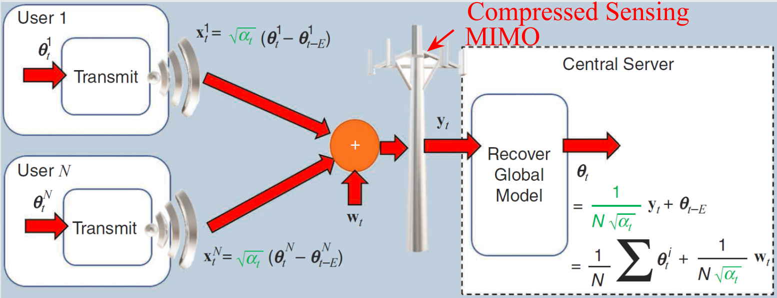
  - Delay-Minimizing Resource Division

$$\min_{\{\chi_{t,k}^i\}_{k=1}^K \in \{0,1\}^K} \max_{i \in \mathcal{G}_t} \frac{\beta_t^i}{\sum_{k=1}^{K} \chi_{t,k}^i R_{i,k}} \quad \text{s.t.} \sum_{i \in \mathcal{G}_t} \chi_{t,k}^i = \sum_{k=1}^{K} \chi_{t,k}^i = 1$$

d: distance to the access point; β: model size (bits); R: data rate over the channel.

Over-the-Air Federated Learning (AirFL)

Users simultaneously employ the complete temporal and spectral resources (complete reuse) of the uplink channel in a nonorthogonal manner.



Flat fading channel with Gaussian noise and interference:

$$y_t^i = h^i x_t^i + w_t^i + v_k$$

Analog Aggregation-Based FL[2]

$$y_t = \sum p_t^i \odot x_t^i \odot h_t^i + w_t$$

$$(p_t^i)_d = \frac{(\beta_t^i)_d K_i (b_t)_d}{(h_t^i)_d} \quad \text{power control vector}$$

Optimization Problem

$$R_t[d] = \frac{L\sigma^2}{2\left(\sum_{i=1}^U \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{K\rho_1 + 2KL\rho_2 \triangle_{t-1}}{2L\sum_{i=1}^U K_i \beta_{i,t}^d}, \quad \forall d,$$

$$R_t^{NC}[d] = \frac{L\sigma^2}{2\left(\sum_{i=1}^U \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{K\rho_1}{2L\sum_{i=1}^U K_i \beta_{i,t}^d}, \quad \forall d,$$

$$R_t^{SGD}[d] = \frac{L\sigma^2}{2\left(\sum_{i=1}^U \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{U(\rho_1 + 2L\rho_2 \triangle_{t-1})}{2L\sum_{i=1}^U K_i \beta_{i,t}^d}, \quad \forall d.$$

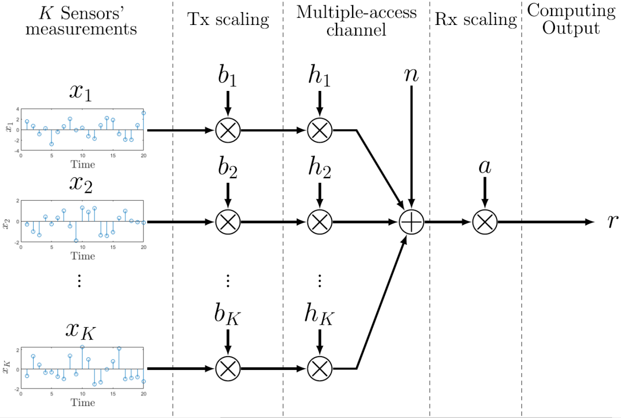
$$\min_{\{b_t, \beta_{i,t}\}_{i=1}^U} R_t$$

$$\text{s.t.} \quad \left|\frac{\beta_{i,t} K_i b_t}{h_{i,t}}\right|^2 (|w_{t-1}| + \eta)^2 \le P_i^{\max},$$

$$\beta_{i,t} \in \{0, 1\}, i \in \{1, 2, \ldots, U\},$$

[2]X. Fan, Y. Wang, Y. Huo, and Z. Tian, “Joint Optimization of Communications and Federated Learning Over the Air,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2021

AirComp[3]



[3]W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-Air Computation Systems: Optimization, Analysis and Scaling Laws," *IEEE Transactions on Wireless Communications*, vol. 19, pp. 5488–5502, 8 2020

Time-Varying Precoding for AirFL

$$y_t = \sum_{i=1}^{N} x_t^i + w_t, \quad x_t^i = \sqrt{\alpha_t}(\theta_t^i - \theta_{t-E}^i)$$

$$\theta_t = \frac{1}{\sqrt{\alpha_t}} y_t + \theta_{t-E} = \frac{1}{N}\sum_{i=1}^{N} \theta_t^i + \frac{1}{N\sqrt{\alpha_t}} w_t$$

$$\alpha_t = \frac{P}{\max_i \mathbb{E}\{\|\theta_t^i - \theta_{t-E}^i\|\}}$$

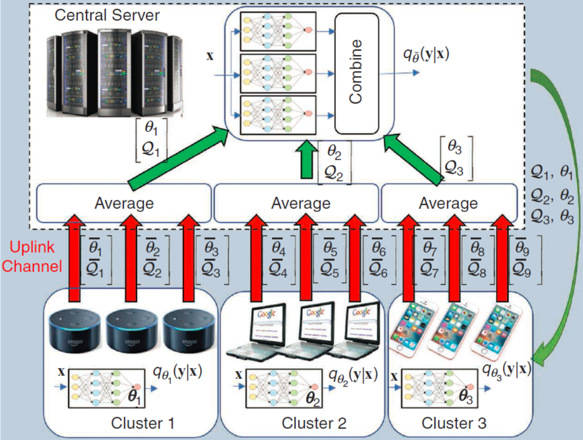
$$\theta_t^i \xrightarrow{\text{encode}} s_t^i \xrightarrow{\text{transmit}} x_t^i \xrightarrow{\text{aggregate}} y_t \xrightarrow{\text{combine}} \theta_t$$

local model

global model

Combining

Mixture of models



Security: Byzantine-Robust Combining

- Geometric median combining:

$$\theta_t = \arg\min_{\theta} \sum \|\theta - y_t^i\|_2$$

- Krum aggregation[4]:

$$\theta_t = KR(y_1, \ldots, y_N) = y_{i_*}, \quad i_* = \arg\min_{i} \sum_{i \to j} \|y_i - y_j\|^2$$

$i \to j$: the set of $N - f - 2$ nearest neighbors.

- Truncation mapping: discards "abnormal" subset of the model updates before averaging.

[4]P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, p. 118–128, Curran Associates, Inc., 2017.

References I

[1] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated Learning: A Signal Processing Perspective," *IEEE Signal Processing Magazine*, vol. 39, pp. 14–41, 5 2022.

[2] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint Optimization of Communications and Federated Learning Over the Air," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2021.

[3] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-Air Computation Systems: Optimization, Analysis and Scaling Laws," *IEEE Transactions on Wireless Communications*, vol. 19, pp. 5488–5502, 8 2020.

References II

[4] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, p. 118–128, Curran Associates, Inc., 2017.