# WAFeL: Weighted Over-the-Air Federated Learning

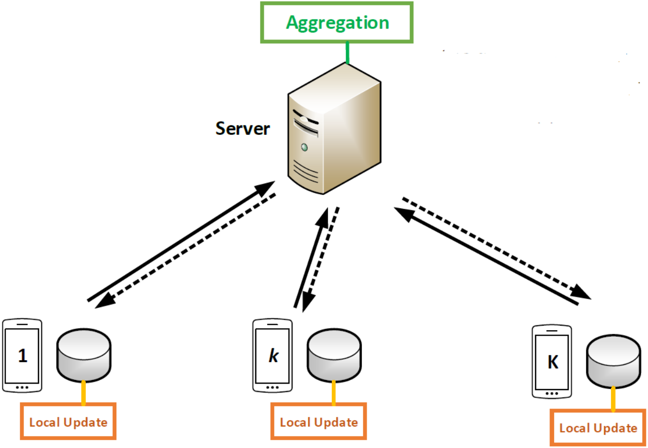
Seyed Mohammad Azimi-Abarghouyi

seyaa@kth.se

KTH Royal Institute of Technology

Jan. 2025

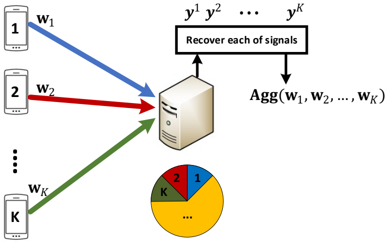
# Federated Learning



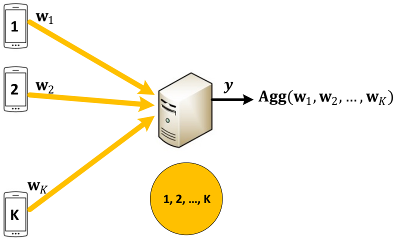

$$\textbf{GD: } \mathbf{w}_k \leftarrow \mathbf{w}_G - \underbrace{\mu}_{\text{learning rate}} \nabla \overbrace{F}^{\text{loss function}} (\mathbf{w}_G, \underbrace{\xi_k}_{\text{training batch}}), \forall k \stackrel{t}{\Leftrightarrow} \mathbf{w}_G = \mathbf{Agg}(\mathbf{w}_1, \ldots, \mathbf{w}_K)$$

# Wireless Federated Learning

Device $k$: $P_k$ (transmission power), $h_k$ (wireless channel to server)




a) "**Orthogonal**" Strategy: Computation after communication




b) "**Over-the-Air**" Strategy: Joint computation and communication

P2P: $\mathbf{y}^k = \sqrt{P_k} h_k \mathbf{w}_k, \forall k$

MAC: $\mathbf{y} = \sum_{k=1}^{K} \sqrt{P_k} h_k \mathbf{w}_k$

# Over-the-Air Federated Learning

- **CSIT-aware over-the-air federated learning**
  - Requires perfect Channel State Information at Transmitter (CSIT) and power control for each device

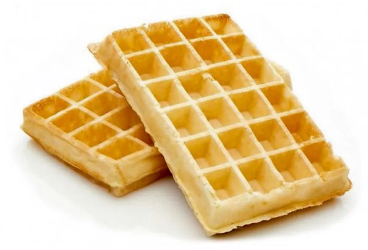
$$\mathbf{y} \xrightarrow[\text{channel inversion}]{\sqrt{P_k}=\frac{1}{h_k}} \frac{\mathbf{y}}{K} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k = \mathbf{w}_{\mathrm{G}}$$

- **Blind over-the-air federated learning**
  - Requires a server equipped with a large antenna array

$$\{\mathbf{y}_1, \ldots, \mathbf{y}_{N_{\mathrm{ant}}}\} \xrightarrow[\text{equalization}]{N_{\mathrm{ant}}\to\infty} \frac{\sum_{n=1}^{N_{\mathrm{ant}}}\left(\sum_{k=1}^{K} h_{k,n}\right)^{*}\mathbf{y}_n}{KN_{\mathrm{ant}}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k = \mathbf{w}_{\mathrm{G}}$$

# Motivation

- **Weighted over-the-Air Federated Learning (WAFeL)**
  - Eliminates the need for prior knowledge and power control
  - Operates efficiently in scenarios with a single-antenna server



# Aggregation

- Fixed aggregation with equal weights in each iteration $t$ of standard federated learning:

$$\mathbf{w}_{\mathrm{G},t} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_{k,t}.$$

Over-the-air schemes have consistently used this aggregation method, despite the presence of interference and noise.

- Proposed weighted aggregation in each iteration $t$ of WAFeL:

$$\mathbf{w}_{\mathrm{G},t} = \sum_{k=1}^{K}\alpha_{k,t}\mathbf{w}_{k,t},$$

where $\alpha_{k,t} \geq 0$ is the weight corresponding to device $k$, and the weight vector $\boldsymbol{\alpha}_t = [\alpha_{1,t}, \ldots, \alpha_{K,t}]^\top$, such that $\mathbf{1}^\top\boldsymbol{\alpha}_t = 1$.

# WAFeL - System Model

- Server with a single antenna
- $K$ devices with a single antenna
- Devices have no CSIT
- Server has Channel State Information at Receiver (CSIR)

Minimal requirements in any wireless system

# WAFeL - Transmission Scheme

- Each model parameter vector $\mathbf{w}_k = [w_{k,1}, \ldots, w_{k,s}]^\top$ undergos normalization as

$$\bar{\mathbf{w}}_k = \frac{(\mathbf{w}_k - \mu_k \mathbf{1})}{\sigma_k},$$

  where $\mu_k = \frac{1}{s}\sum_{i=1}^{s} w_{k,i}$ and $\sigma_k^2 = \frac{1}{s}\sum_{i=1}^{s}(w_{k,i} - \mu_k)^2$ are the mean and variance of model parameters.
- $(\mu_k, \sigma_k)$ are shared with the server.
- The device $k$ transmits

$$\mathbf{x}_k = \sqrt{P}\overline{\mathbf{w}}_k,$$

  where $P$ is the transmission power.

Consistent power transmission

# WAFeL - Transmission Scheme

- The received signal at the server is

$$\mathbf{y} = \sum_{k=1}^{K} \sqrt{P} h_k \bar{\mathbf{w}}_k + \mathbf{z},$$

where $\mathbf{z}$ is AWGN, where each entry has variance $\sigma_z^2$.

- Real-valued representation:

$$\mathbf{Y} = \sqrt{P}\mathbf{H}\bar{\mathbf{W}} + \mathbf{Z},$$

where $\bar{\mathbf{W}} = [\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_K]^\top$ and

$$\mathbf{Y} = \begin{bmatrix} \mathfrak{Re}\left\{\mathbf{y}^\top\right\} \\ \mathfrak{Im}\left\{\mathbf{y}^\top\right\} \end{bmatrix}, \ \mathbf{H} = \begin{bmatrix} \mathfrak{Re}\left\{\mathbf{h}^\top\right\} \\ \mathfrak{Im}\left\{\mathbf{h}^\top\right\} \end{bmatrix}, \ \mathbf{Z} = \begin{bmatrix} \mathfrak{Re}\left\{\mathbf{z}^\top\right\} \\ \mathfrak{Im}\left\{\mathbf{z}^\top\right\} \end{bmatrix},$$

where $\mathbf{h} = [h_1, \ldots, h_K]^\top$.

# WAFeL - Aggregation Scheme

- The server uses an equalization vector $\mathbf{b} \in \mathbb{R}^{2\times 1}$ to estimate as

$$\mathbf{w}_G^\top = \frac{1}{\sqrt{P}} \underbrace{\mathbf{b}^\top \mathbf{Y}}_{\text{equalized signals}} + \underbrace{\sum_{k=1}^{K} \alpha_k \mu_k \mathbf{1}^\top}_{\text{artificially added mean to ensure unbiasedness}} .$$

- The estimation error:

$$\mathbf{w}_G^\top = \underbrace{\sum_{k=1}^{K} \alpha_k \mathbf{w}_k^\top}_{\text{desired weighted aggregation}} + \underbrace{\underbrace{\left(\mathbf{b}^\top \mathbf{H} - (\boldsymbol{\alpha} \odot \boldsymbol{\sigma})^\top\right) \bar{\mathbf{W}}}_{\text{channel mismatch error}} + \underbrace{\frac{1}{\sqrt{P}} \mathbf{b}^\top \mathbf{Z}}_{\text{AWGN error}}}_{\text{effective estimation error}},$$

where $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_K]^\top$ and $(\boldsymbol{\alpha} \odot \boldsymbol{\sigma})^\top = [\alpha_1\sigma_1, \ldots, \alpha_K\sigma_K]$.

# WAFeL - Aggregation Scheme

- The optimal equalization vector is

$$\mathbf{b}_{\text{opt}}^\top = (\boldsymbol{\alpha} \odot \boldsymbol{\sigma})^\top \mathbf{H}^\top \left( \frac{1}{\text{SNR}} \mathbf{I}_2 + \mathbf{H}\mathbf{H}^\top \right)^{-1},$$

where $\text{SNR} = \frac{P}{\sigma_z^2}$.

- The estimation MSE:

$$\text{MSE}(\boldsymbol{\alpha}) = s\boldsymbol{\alpha}^\top \text{diag}(\boldsymbol{\sigma}) \left( \mathbf{I}_K + \text{SNR}\mathbf{H}^\top \mathbf{H} \right)^{-1} \text{diag}(\boldsymbol{\sigma})\boldsymbol{\alpha}.$$

# WAFeL - Convergence Analysis

- Assumption 1 (Lipschitz-Continuous Gradient):

$$F(\mathbf{w}_2) \leq F(\mathbf{w}_1) + \nabla F(\mathbf{w}_1)^T(\mathbf{w}_2 - \mathbf{w}_1) + \frac{L}{2}\|\mathbf{w}_2 - \mathbf{w}_1\|^2,$$
$$\|\nabla F(\mathbf{w}_2) - \nabla F(\mathbf{w}_1)\| \leq L\|\mathbf{w}_2 - \mathbf{w}_1\|,$$

  where $L > 0$ is the Lipschitz constant.
- Assumption 2 (Gradient Variance Bound):

$$\mathbb{E}\left\{\|\nabla F(\mathbf{w}_k, \boldsymbol{\xi}_k) - \nabla F(\mathbf{w}_k)\|^2\right\} \leq \frac{\sigma_{\mathrm{g}}^2}{B},$$

  where $|\boldsymbol{\xi}_k| = B$, the batch size, and $\sigma_{\mathrm{g}}^2$ denotes the gradient variance constant.

# WAFeL - Convergence Analysis

**Theorem**

Let $\eta \leq \frac{1}{L}$ and $\boldsymbol{\alpha}_t$ as the weight vector for each round $t \in \{0, \ldots, T-1\}$, then the convergence rate of `WAFeL` under Assumptions 1 and 2 is

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\{\|\nabla F(\mathbf{w}_{\mathrm{G},t})\|^2\right\} \leq \frac{2\left(F(\mathbf{w}_{\mathrm{G},0}) - F(\mathbf{w}^*)\right)}{\eta T}$$

$$+ \frac{L}{\eta T}\sum_{t=0}^{T-1}\mathcal{I}_t(\boldsymbol{\alpha}_t),$$

where

$$\mathcal{I}_t(\boldsymbol{\alpha}_t) = \eta^2\frac{\sigma_{\mathrm{g}}^2}{B}\|\boldsymbol{\alpha}_t\|^2 + \mathrm{MSE}_t(\boldsymbol{\alpha}_t).$$

# WAFeL - Convergence Analysis

- Under $\boldsymbol{\alpha}_t = \frac{1}{K}\mathbf{1}$ for each round $t \in \{0, \ldots, T-1\}$ and error-free transmission:

$$\mathcal{I}_t\left(\frac{1}{K}\mathbf{1}\right) = \eta^2 \frac{\sigma_{\mathrm{g}}^2}{B}\frac{1}{K} = \mathcal{I}.$$

- 

$$\mathcal{I}_t(\boldsymbol{\alpha}_t) - \mathcal{I} = \eta^2 \frac{\sigma_{\mathrm{g}}^2}{B} \overbrace{\underbrace{\left(\|\boldsymbol{\alpha}_t\|^2 - \frac{1}{K}\right)}_{\textit{mismatch}}}^{\textit{learning aspect}} + \overbrace{\underbrace{\mathrm{MSE}_t(\boldsymbol{\alpha}_t)}_{\textit{mse}}}^{\textit{communication aspect}}.$$

# WAFeL - Weight Selection

$$\boldsymbol{\alpha}_t = \arg \min_{\boldsymbol{\alpha} \geq \mathbf{0} \setminus \{\mathbf{0}\}} \|\boldsymbol{\alpha}\|^2 \leftarrow \text{mismatch}$$

subject to

$$\text{MSE}_t(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \text{diag}(\boldsymbol{\sigma}_t) \left(\mathbf{I}_K + \text{SNR}\mathbf{H}_t^\top \mathbf{H}_t\right)^{-1} \text{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha} \leq \text{th},$$

$$\mathbf{1}^\top \boldsymbol{\alpha} = 1 \leftarrow \text{due to averaging}$$

No physical constraints on selecting aggregation weights

# WAFeL - Weight Selection

**Algorithm**

Initialize $\lambda^{(0)}$ and $\boldsymbol{\alpha}^{(0)} = \frac{1}{K}\mathbf{1}$

Iterate

Update $\boldsymbol{\alpha}^{(j)} = \frac{\mathbf{G}_t(\lambda^{(j-1)})^{-1}\mathbf{1}}{\mathbf{1}^\top \mathbf{G}_t(\lambda^{(j-1)})^{-1}\mathbf{1}}$, where $\mathbf{G}_t(\lambda) = \mathbf{I}_K + \lambda \mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K + \mathrm{SNR}\mathbf{H}_t^\top \mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t)$.

Update $\lambda^{(j)} = \lambda^{(j-1)} + t\left({\boldsymbol{\alpha}^{(j)}}^\top \mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K + \mathrm{SNR}\mathbf{H}_t^\top \mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha}^{(j)} - \mathrm{th}\right)$.

Until $\left|\lambda\left(\boldsymbol{\alpha}^\top \mathrm{diag}(\boldsymbol{\sigma}_t)\left(\mathbf{I}_K + \mathrm{SNR}\mathbf{H}_t^\top \mathbf{H}_t\right)^{-1}\mathrm{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\alpha} - \mathrm{th}\right)\right| \leq \epsilon$.

Complexity order:

$$\mathcal{O}\left(K^3\right)$$

# WAFeL - Experiments

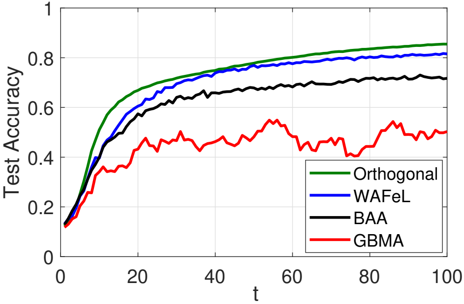



Figure: MNIST dataset

| $K$ | SNR |
|---|---|
| 30 | 10 |

- Orthogonal: Avoids interference using $K$ times more resources
- BAA: Employs power control with perfect CSIT
- GBMA: Compensates for the channel phase only


BAA: G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491-506, Jan. 2020.

GBMA: T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Sig. Process.*, vol. 68, pp. 2897-2911, 2020.

# Conclusions

- We proposed a new over-the-air federated learning scheme that focuses on optimizing aggregation weights.
- This scheme not only achieves significantly higher performance but also minimizes complexity and resource requirements.
- It provides a promising learning approach with potential applications that can be further explored in various setups.

Thank you!