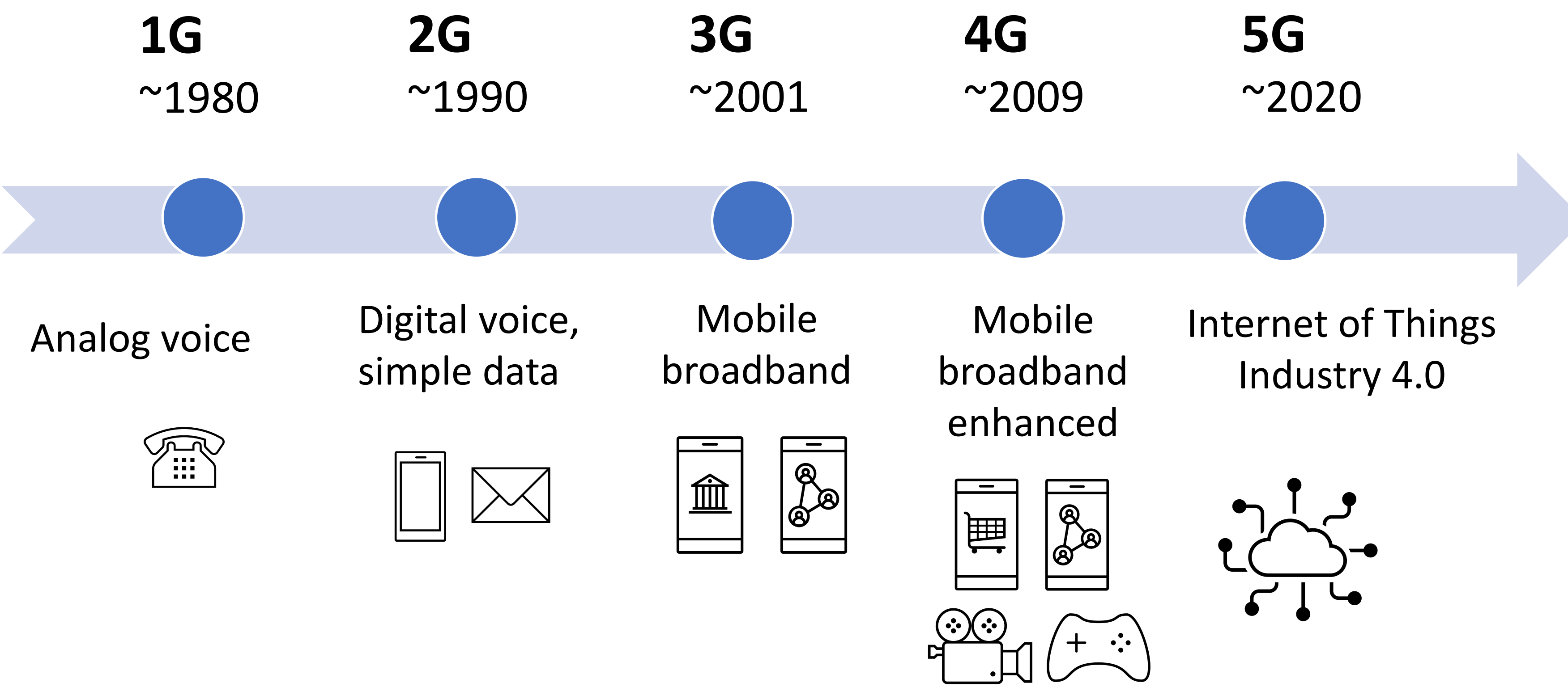
# Distributed Intelligence over Wireless Networks

IEEE SPS - EURASIP 6G Summer School
2022-09-01

LINKÖPING UNIVERSITY

Zheng Chen
Linköping University

# Evolution of Mobile Standards



# Demand-Driven Wireless Technology

New Application & Data Type → New Comm. Requirement → New Wireless Design

- Text messages
- Video streaming
- Social media
- Online gaming
- Remote surgery
- Status monitoring
- …

- Data rate
- Connectivity
- Spectral Efficiency
- Energy efficiency
- Latency
- Reliability
- Timing
- …

# Wireless Communications—Shannon's Legacy

- Shannon-Hartley theorem on channel capacity for memoryless point-to-point AWGN channel:

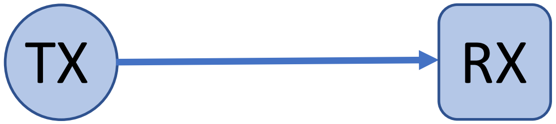
$$C = B \cdot \log_2\left(1 + \frac{S}{N}\right) \text{ bits per second}$$

- Shannon's channel coding theorem
  - If transmission rate $R < C$, there exists a coding scheme that achieves arbitrarily low error probability

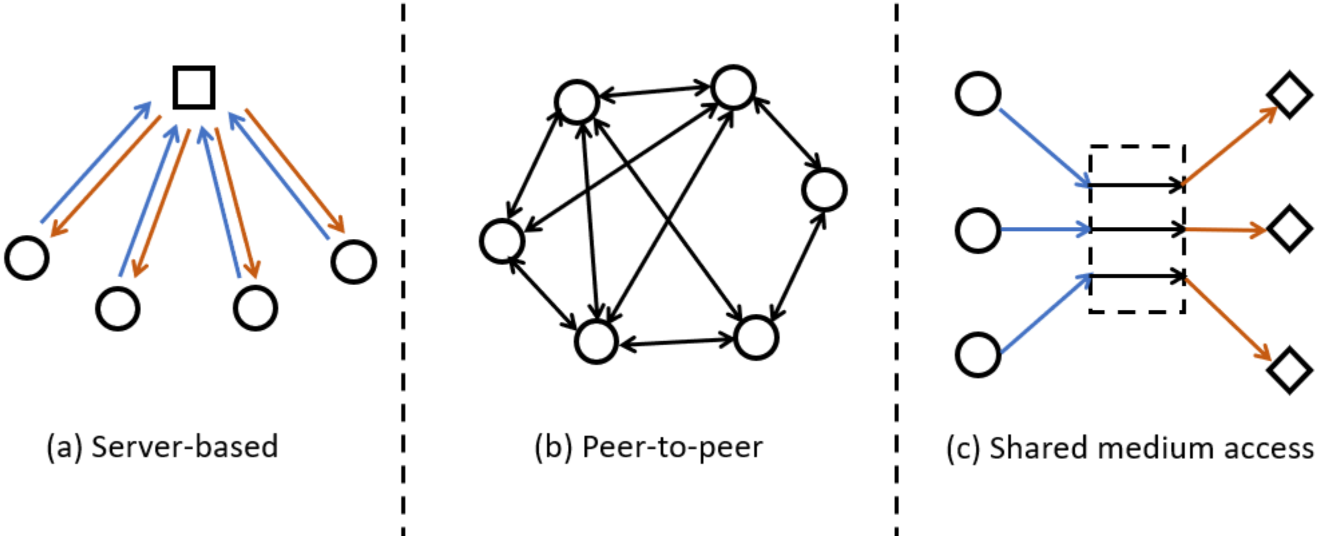
- Wireless communication system is usually designed to achieve the maximum number of bits that can be reliably transmitted under some resource constraints.

# Wireless Bottleneck in Distributed Systems

- Point to point



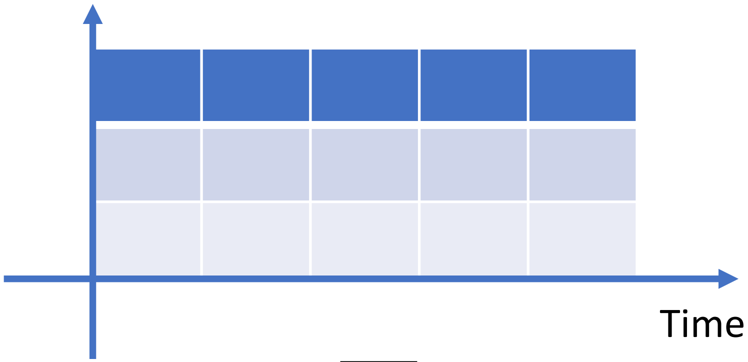

- Distributed systems



# Wireless Resource Limitations

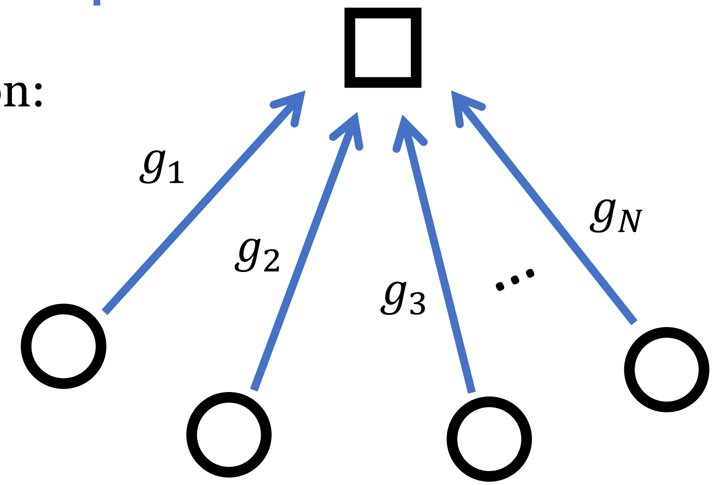
Wireless communication resources:

- Frequency
- Time
- Space




Rate-oriented design for resource allocation:

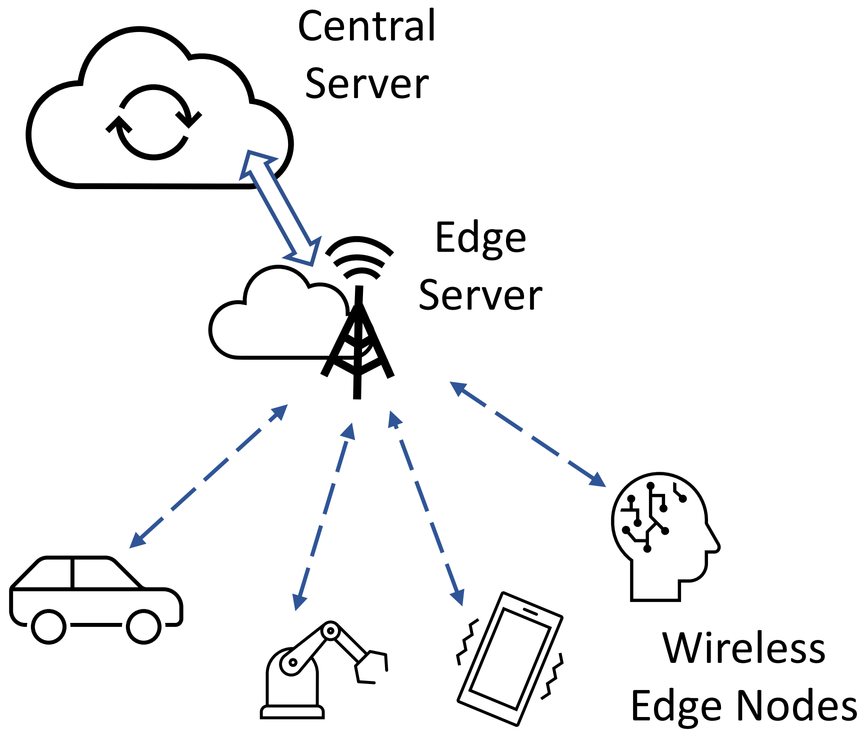
- Maximize $\sum_{i=1}^{N} R_i$
- Maximize $\min_i \{R_i\}$
- Others




$g_i$: channel gain of user $i$

# 6G: Era of Connected Intelligence

- For human perception
  - Messages carry specific meaning, must be correctly received
  - Higher rate implies higher QoE
- For machine perception
  - Error-free communication might not be needed
  - Performance depends on the tasks (training, inference, control…)
  - Goal-oriented comm. design



# Outline

- From centralized to distributed machine learning (ML)
- Optimization for ML
  - Stochastic gradient method
  - Distributed ML meets wireless
- Federated edge learning
  - Impact of resource allocation
  - Example on scheduling and aggregation design
- Model aggregation over distributed nodes
  - Over-the-Air (OtA) computation
  - OtA with multiple receivers
  - OtA FL with multiple antennas
  - Applications and challenges of OtA aggregation
- Summary

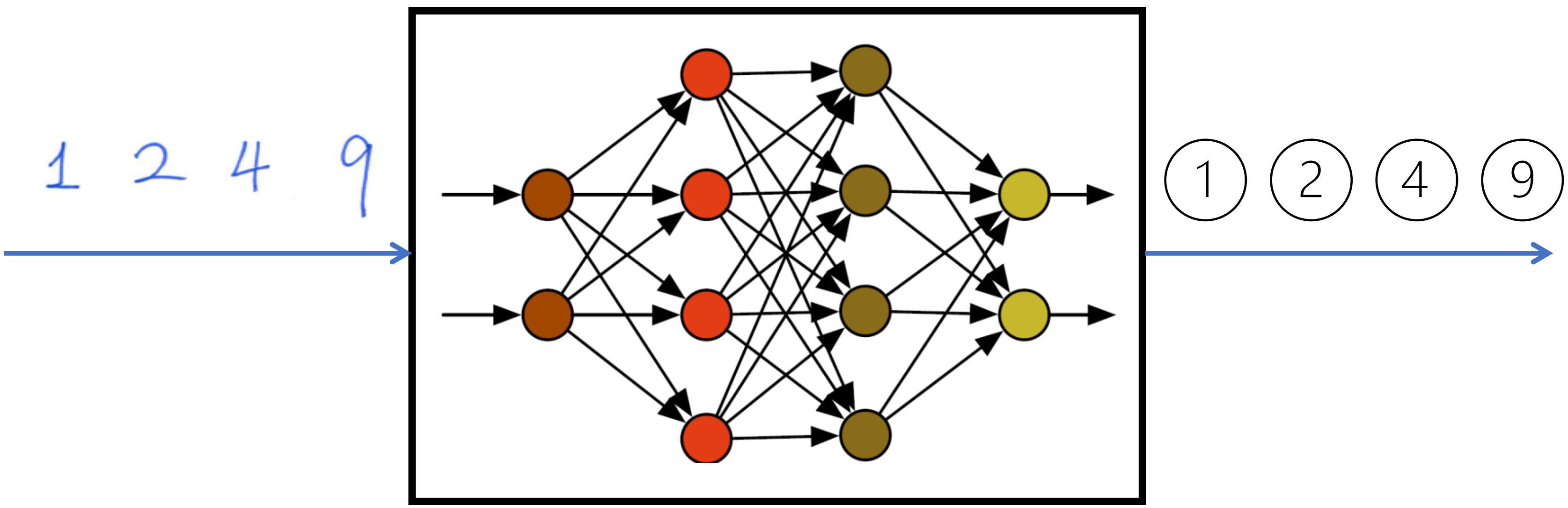
# Classes of ML Algorithms

- Supervised learning
  - Regression
  - Classification
- Unsupervised learning
  - Clustering
  - Anomaly detection
- Reinforcement learning

Focus of this lecture

# Example on Supervised Learning

Hand-written digits recognition with deep neural networks

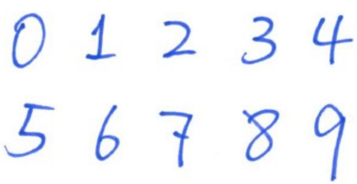


# Model Training

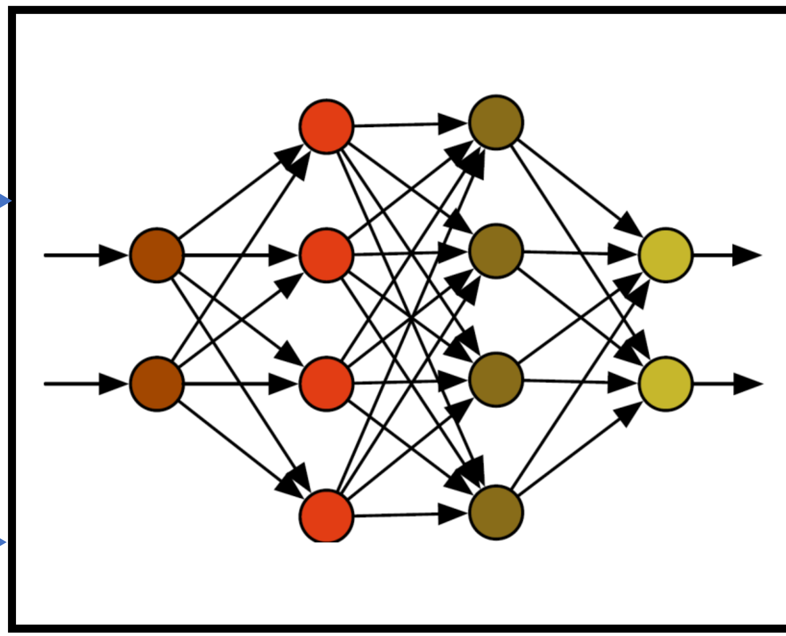
Training of deep learning algorithms

0 1 2 3 4
5 6 7 8 9

Output (label)

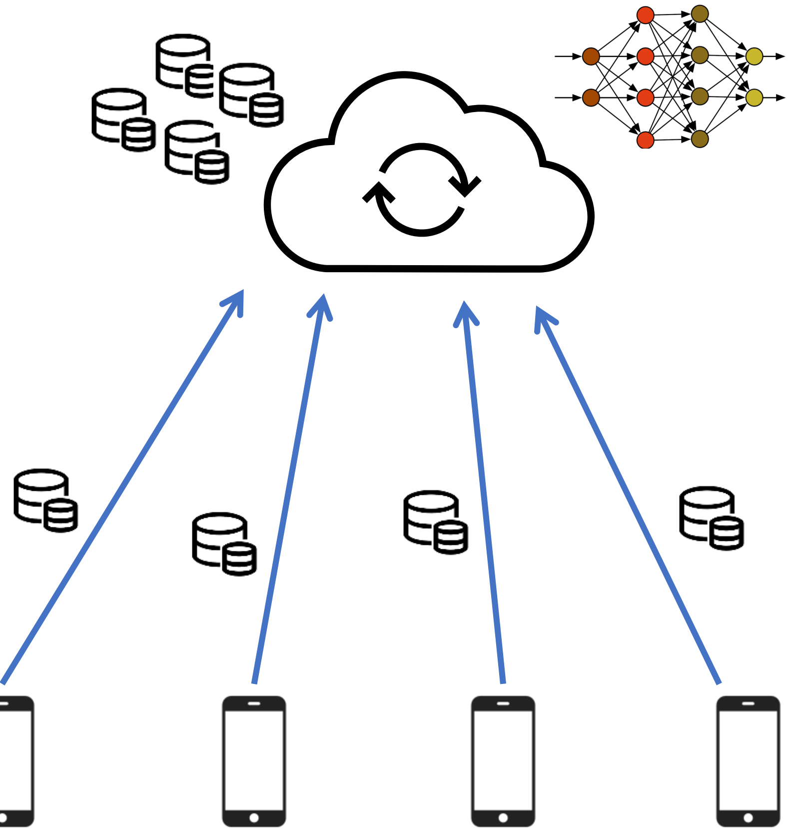



Input (image)

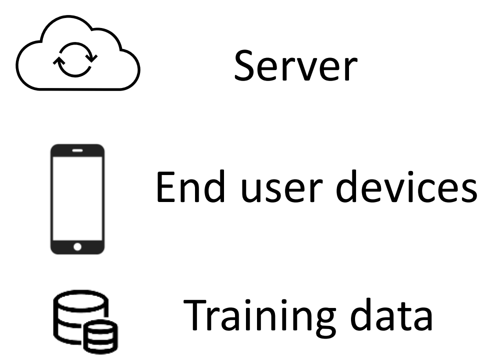


Trained Parameters

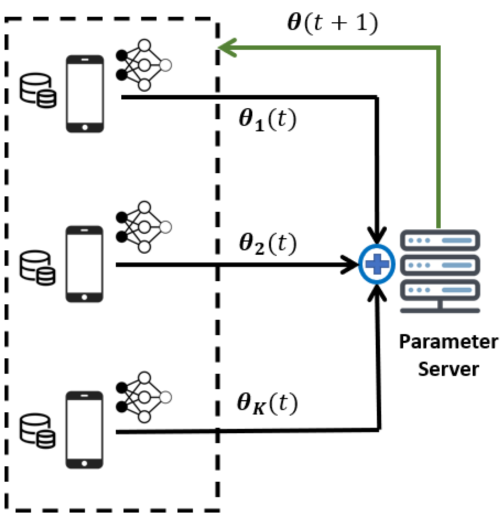
# Centralized Training

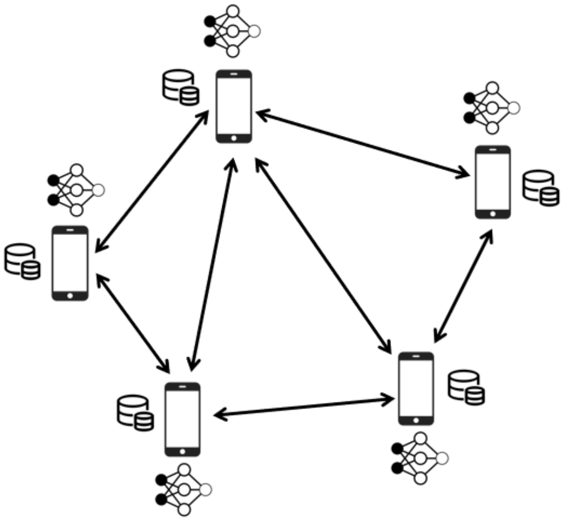


- Privacy issue
- Uploading cost
- Latency



# Collaborative Training with Decentralized Data



- A group of agents $\mathcal{N} = \{1, \ldots, N\}$ collaborate in training a common ML model, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$
- $\mathcal{S}_k$: local training data at device $k$. $\mathcal{S} = \bigcup_{k \in \mathcal{N}} \mathcal{S}_k$: entire dataset in the system
- Optimization objective: find $\boldsymbol{\theta}^*$ that minimizes the global loss function

$$F(\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} l(\boldsymbol{\theta}, x)$$



[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial intelligence and statistics*. PMLR, 2017.

# Optimization for ML

- Expected risk

$$R(\boldsymbol{\theta}) = \mathbb{E}[f(\boldsymbol{\theta}; \xi)]$$

$\boldsymbol{\theta}$: model parameters, $\xi$: randomness in data, $f: \mathbb{R}^d \to \mathbb{R}$

- Empirical risk

$$F(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{\theta})$$

- Objective: minimize$_{\boldsymbol{\theta}}$ $F(\boldsymbol{\theta})$
- Stochastic Gradient (SG) method

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \alpha_t g(t)$$

$g(t)$: stochastic gradient vector with $\mathbb{E}[g(t)] = \nabla F(\boldsymbol{\theta})$. Example:

$$g(t) = \begin{cases} \nabla f(\boldsymbol{\theta}(t), \xi_t) \\ \frac{1}{n}\sum_{i=1}^{n} \nabla f\left(\boldsymbol{\theta}(t), \xi_{t,i}\right) \end{cases}$$

[2] L. Bottou, Léon, F. E. Curtis, and J. Nocedal. "Optimization methods for large-scale machine learning." *SIAM Review* 60.2 (2018): 223-311.

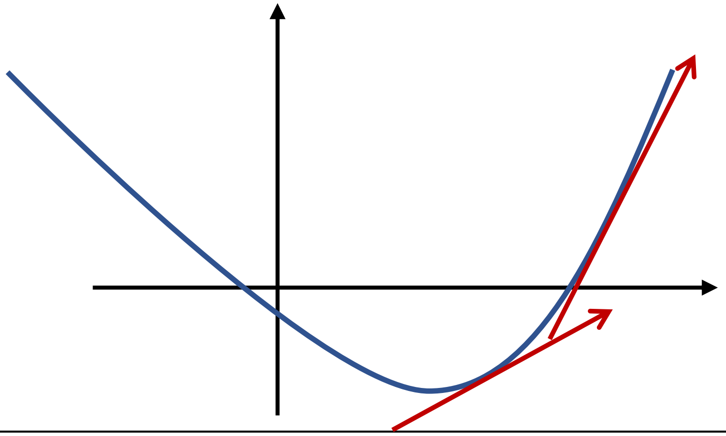
# Assumptions

## ➢L-smoothness

$F$ is continuously differentiable, and the gradient function $\nabla F(\boldsymbol{\theta})$ is Lipschitz continuous with constant $L > 0$, i.e.,

$$\|\nabla F(\boldsymbol{\theta}_1) - \nabla F(\boldsymbol{\theta}_2)\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

Intuitive explanation: the gradient function does not change too quickly



# Assumptions

## ➢Strong convexity

$F$ is strongly convex, i.e., there exists a constant $\mu > 0$ such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, we have

$$F(\boldsymbol{\theta}_1) - F(\boldsymbol{\theta}_1) \geq \nabla F(\boldsymbol{\theta}_1)^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + \frac{\mu}{2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2$$

Intuitive explanation: the objective function is not just convex, but also "curvy" enough

YES NO

# Assumptions

## ➢Bound on second-order moment

The second order moment of the stochastic gradient vector satisfies

$$\mathbb{E}_{\xi_t}[\|g(t)\|_2^2] \leq M + M_g \left\|\nabla F\big(\boldsymbol{\theta}(t)\big)\right\|^2$$

$$\Updownarrow$$

$$\mathbb{E}_{\xi_t}\left[\left\|g(t) - \nabla F\big(\boldsymbol{\theta}(t)\big)\right\|_2^2\right] \leq M + M_v \left\|\nabla F\big(\boldsymbol{\theta}(t)\big)\right\|^2$$

Standard assumption on bounded variance

$$\mathbb{E}\left[\left\|g(t) - \nabla F\big(\boldsymbol{\theta}(t)\big)\right\|^2\right] \leq \sigma^2$$

- not realistic unless parameter space is compact
- often not compatible with strong convexity assumption

# Convergence and Optimality Gap

Let $\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, F(\boldsymbol{\theta})$ and $F^* = F(\boldsymbol{\theta}^*)$, under these assumptions, with fixed step size $\alpha_t = \bar{\alpha}$, and $0 < \bar{\alpha} < \frac{1}{LM_G}$, then

$$\mathbb{E}\left[F\big(\boldsymbol{\theta}(t)\big) - F^*\right] \leq \frac{\bar{\alpha}LM}{2\mu} + (1 - \bar{\alpha}\mu)^t\Big(F\big(\boldsymbol{\theta}(0)\big) - F^*\Big)$$
$$\xrightarrow{t\to\infty} \frac{\bar{\alpha}LM}{2\mu}$$

- With fixed step size, SGD converges to some neighborhood of $\boldsymbol{\theta}^*$
- Less "noise" in stochastic gradient computation, smaller optimality gap

# Federated Learning: Collaborative Training over Decentralized Data

Federated Learning is one instance of distributed and parallel SGD with

- Massive and heterogeneous devices
- Non-IID and unbalanced data
- Limited communication

Two well-known problems:

- Straggler issue
- Communication constraints



[3] Stich, Sebastian U. "Local SGD converges fast and communicates little." ICLR 2019.
[4] Li, Xiang, et al. "On the convergence of FedAvg on non-iid data." ICLR 2020.

# Federated Learning at the Wireless Edge

During communication round $t$

Broadcast global model $\boldsymbol{\theta}(t)$

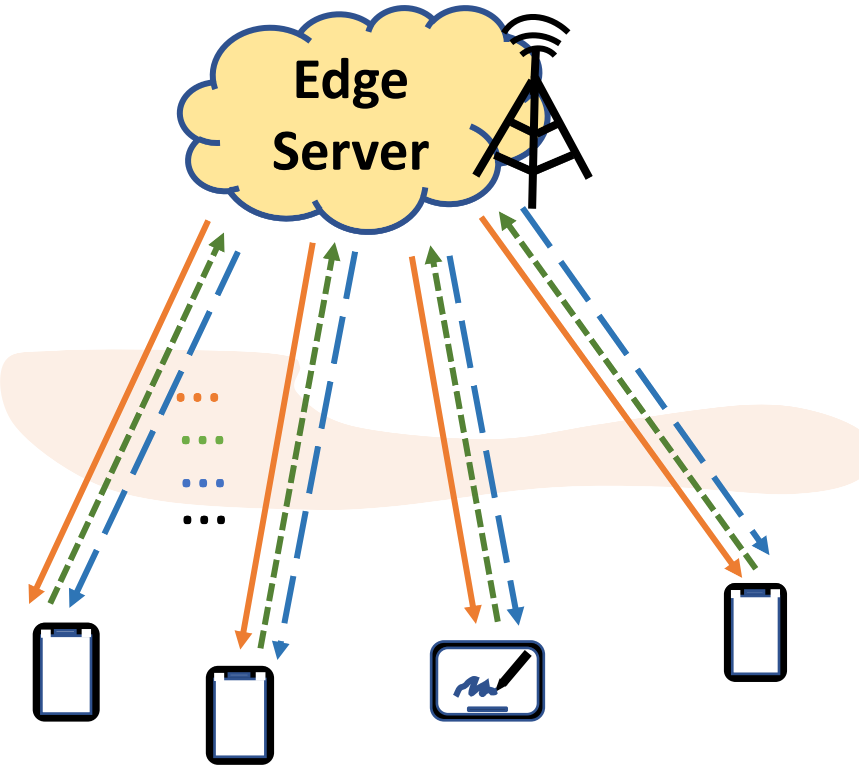
On-device training with local data ($E$ iterations)

$$\boldsymbol{\theta}_k(t, \tau+1) = \boldsymbol{\theta}_k(t, \tau) - \eta(t)\nabla F_k\big(\boldsymbol{\theta}(t, \tau)\big)$$

Upload model update $\boldsymbol{\theta}_k(t, E)$ to server

Broadcast new global model after weighted average of $\boldsymbol{\theta}_k(t, E)$

$$\boldsymbol{\theta}(t+1) = \sum_{\forall k} w_k \boldsymbol{\theta}_k(t, E)$$



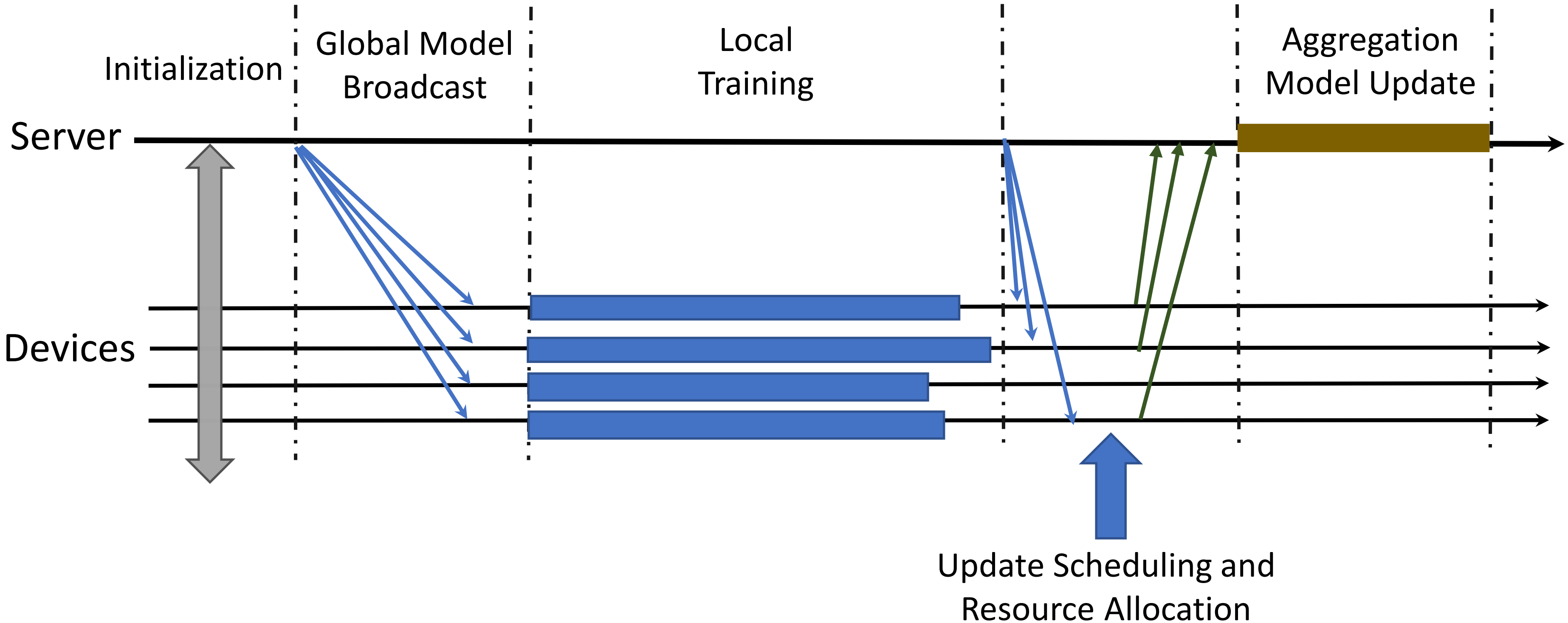

Downlink: model distribution

Uplink: update uploading (most affected by resource constraints)

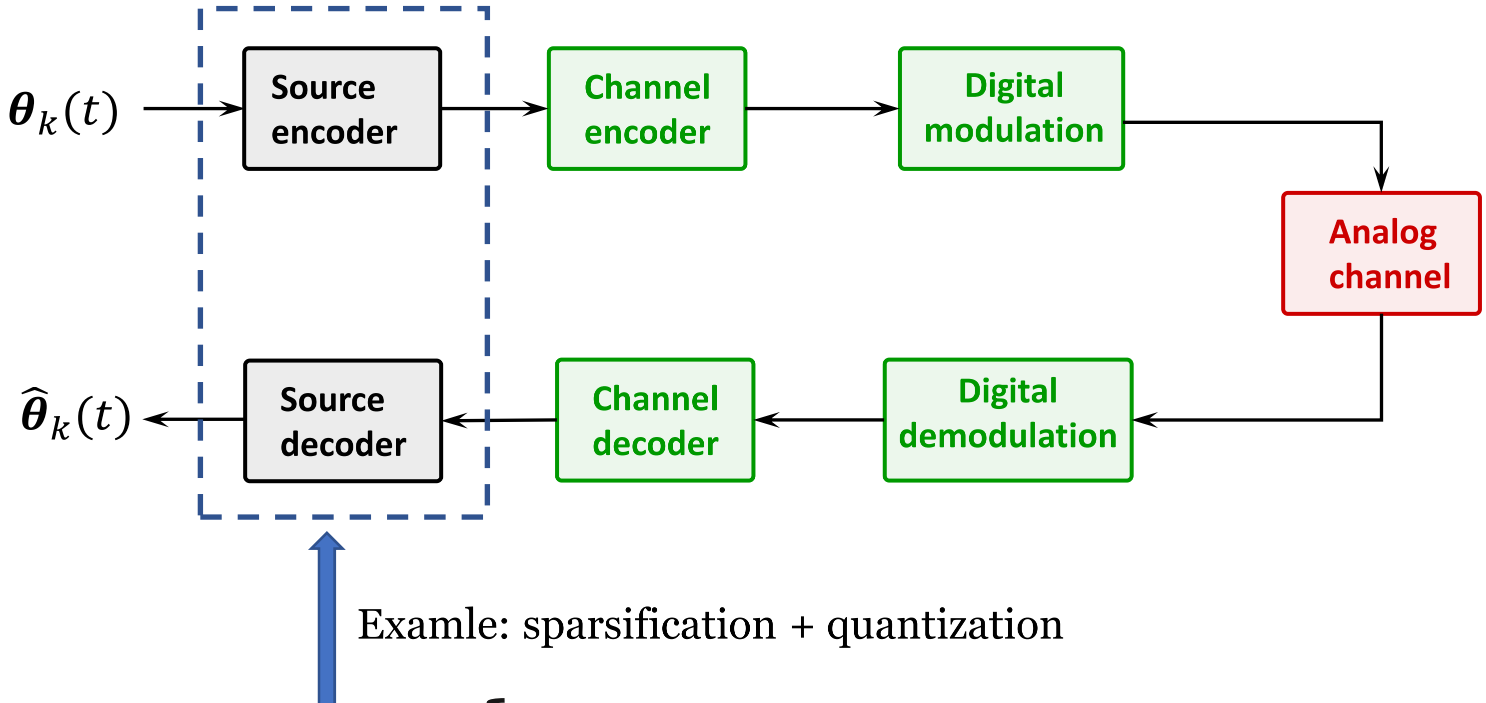


[3] T. Gafni, N. Shlezinger, K. Cohen, Y.C. Eldar, and H. V. Poor. Federated learning: A signal processing perspective, 2021.

# Training Procedure



# Federated Learning with Digital Transmission

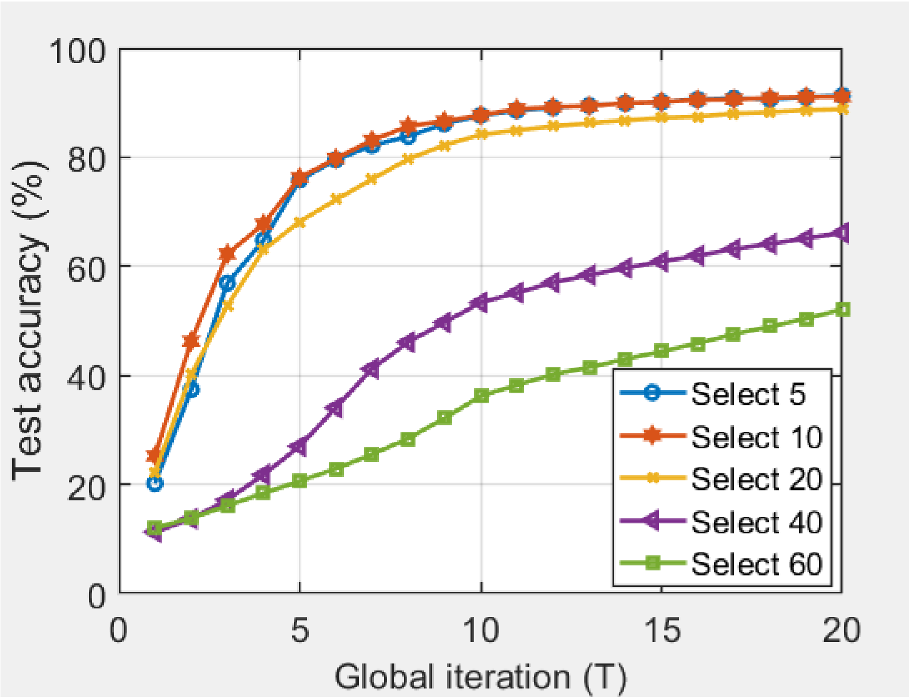



Examle: sparsification + quantization

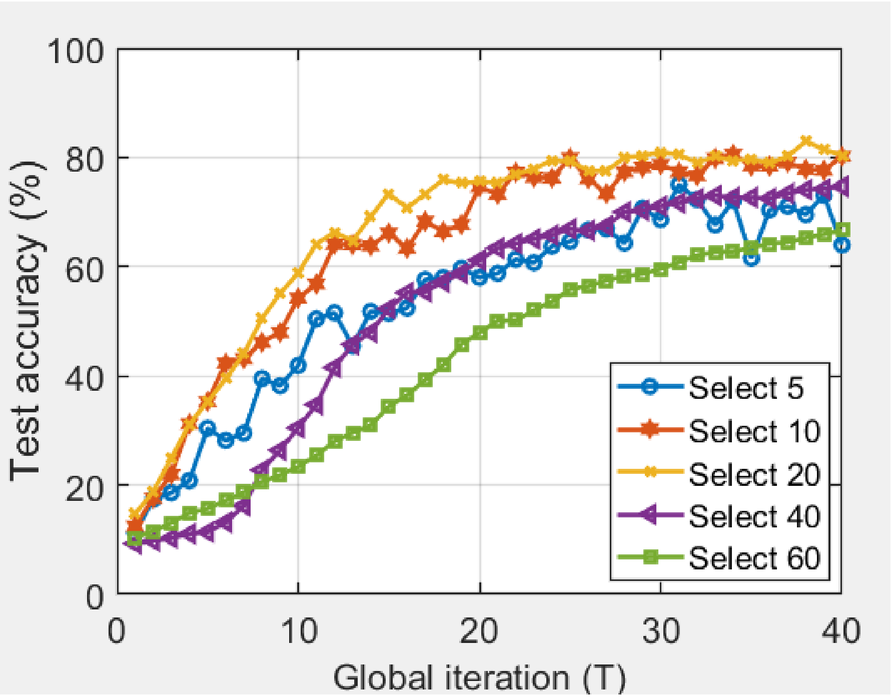
Resource limitation => $\begin{cases} \text{Limited number of scheduled users} \\ \text{Limited number of information bits => Compression loss} \end{cases}$

# Example: Impact of Resource Allocation

With IID data




With non-IID data

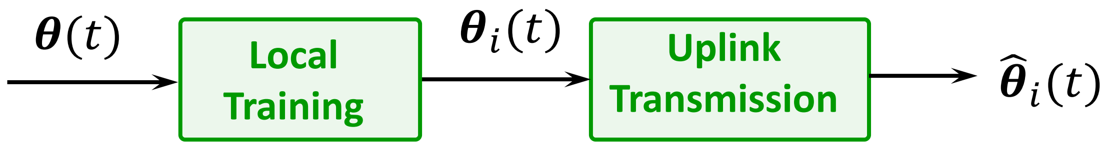



Ideally, schedule users with the most significant impact on the learning result.

# Distributed ML meets Wireless

During uplink transmission of model updates:

- Ideally, we want to obtain $\boldsymbol{\theta}(\mathcal{K}) = \frac{1}{|\mathcal{K}|}\sum_{i\in\mathcal{K}} \boldsymbol{\theta}_i\,(t)$, where $\mathcal{K}$ is the set of all participating agents
- In practice, we get $\widehat{\boldsymbol{\theta}}(\mathcal{K}_s) = \frac{1}{|\mathcal{K}_s|}\sum_{i\in\mathcal{K}_s} \widehat{\boldsymbol{\theta}}_i\,(t)$, where $\mathcal{K}_s$ is the set of scheduled agents



# Distributed ML meets Wireless

Mean Squared Error (MSE) of aggregated model

$$\mathrm{MSE} = \underbrace{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}(\mathcal{K}_s) - \mathbb{E}[\widehat{\boldsymbol{\theta}}(\mathcal{K}_s)]\right\|^2\right]}_{\text{variance}} + \underbrace{\left\|\mathbb{E}[\widehat{\boldsymbol{\theta}}(\mathcal{K}_s)] - \boldsymbol{\theta}(\mathcal{K})\right\|^2}_{\text{bias}}$$

- Bias: affected by the number of scheduled users and their data representation
- Variance: affected by compression loss (depending on how many bits we can reliably transmit)

# Joint Data- and Channel-aware Scheduling

Motivation:

- Schedule users with more homogeneous data representation to reduce bias
- Prioritize users with better channels to reduce variance

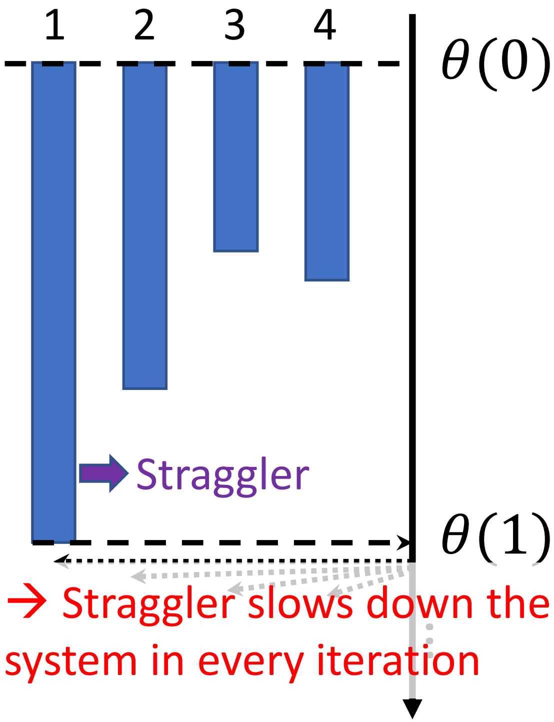
For supervised learning with labelled data, define $\boldsymbol{b}_k = [b_k^1, \dots, b_k^L]$ as the label distribution of local data set at user $k$.

- First, select $\Pi'(t) \in \mathcal{K}$ users with largest channel gain
- Then, select a subset $\Pi(t)$ that minimizes $\sum_{l=1}^{L} \left| \sum_{k \in \Pi(t)} \left( b_k^l - \bar{b} \right) \right|^2$

[4] C-H. Hu, Z. Chen, E. G. Larsson, " Scheduling and Aggregation Design for Asynchronous Federated Learning over Wireless Networks", submitted, 2022.

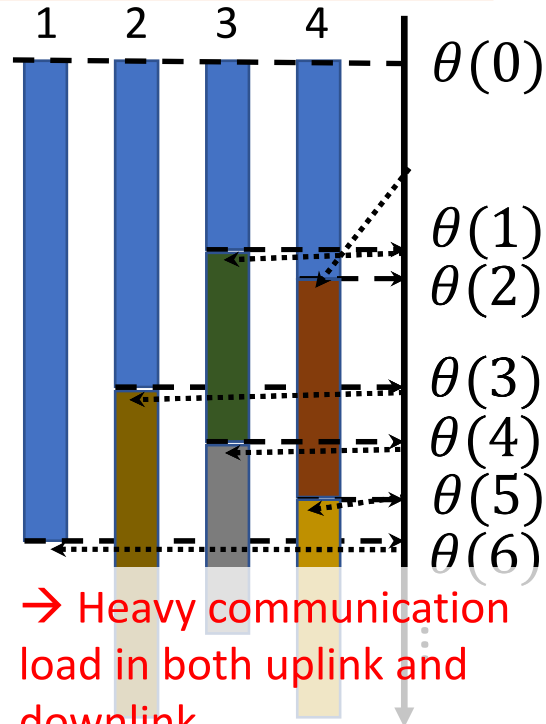
# Synchronous vs. Asynchronous Training

Synchronous FL



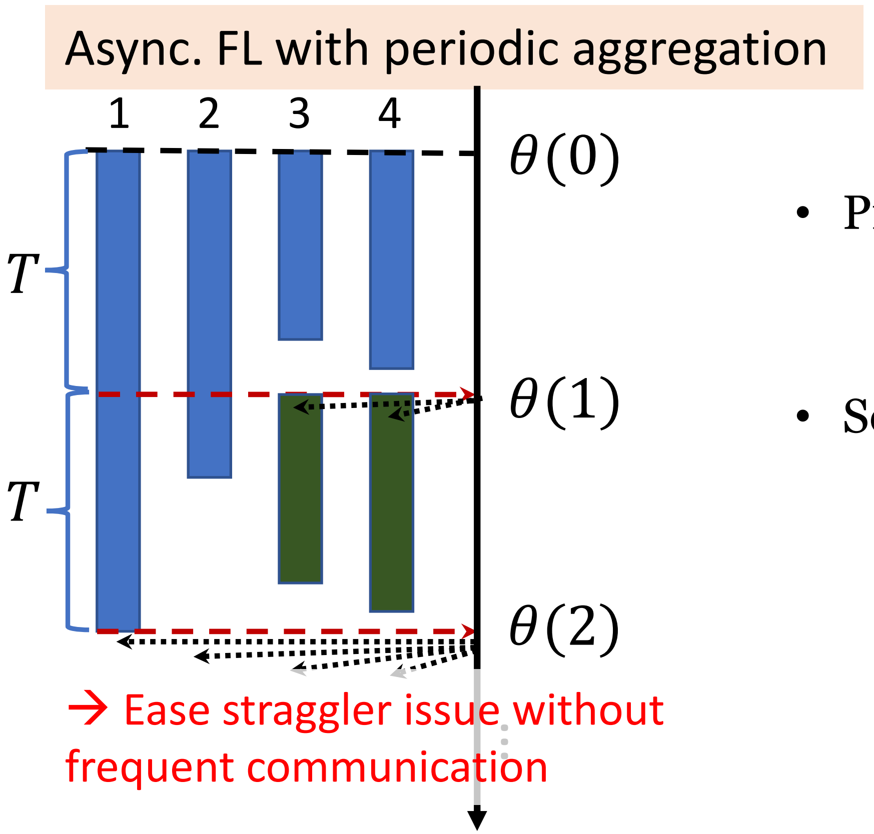

→ Straggler slows down the system in every iteration

Fully asynchronous FL




→ Heavy communication load in both uplink and downlink

# Asynchronous FL with Periodic Aggregation




→ Ease straggler issue without frequent communication

- Problem:
  - Different "age" of previously received global model
- Solution:
  - Age-aware aggregation policy to balance update freshness



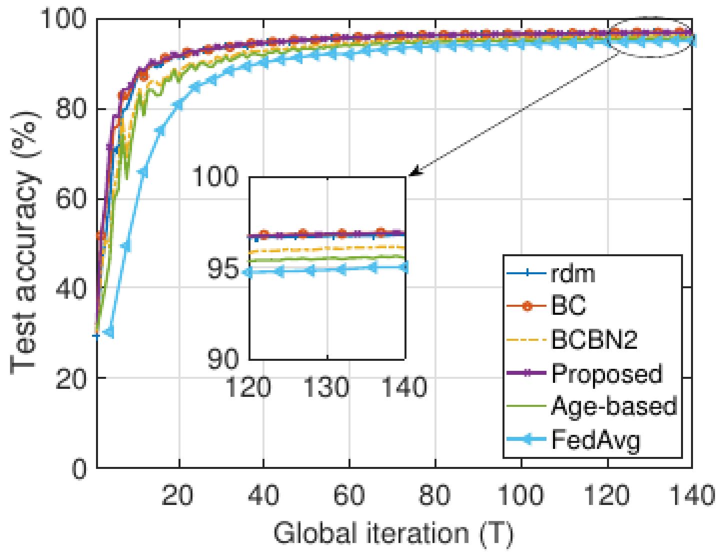
Slide credit: Chung-Hsuan Hu

# Simulation Results

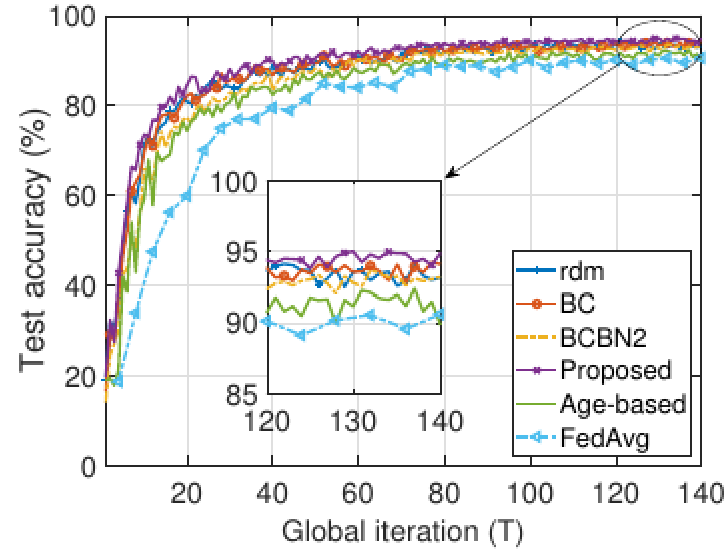
## Simulation setup:

MNIST dataset, CNN, model dimension $d = 21840$, Rayleigh fading, 40 users in total, scheduling 20% in every round

Comparison with alternative methods in [5] and [6]




(a) IID data distribution.

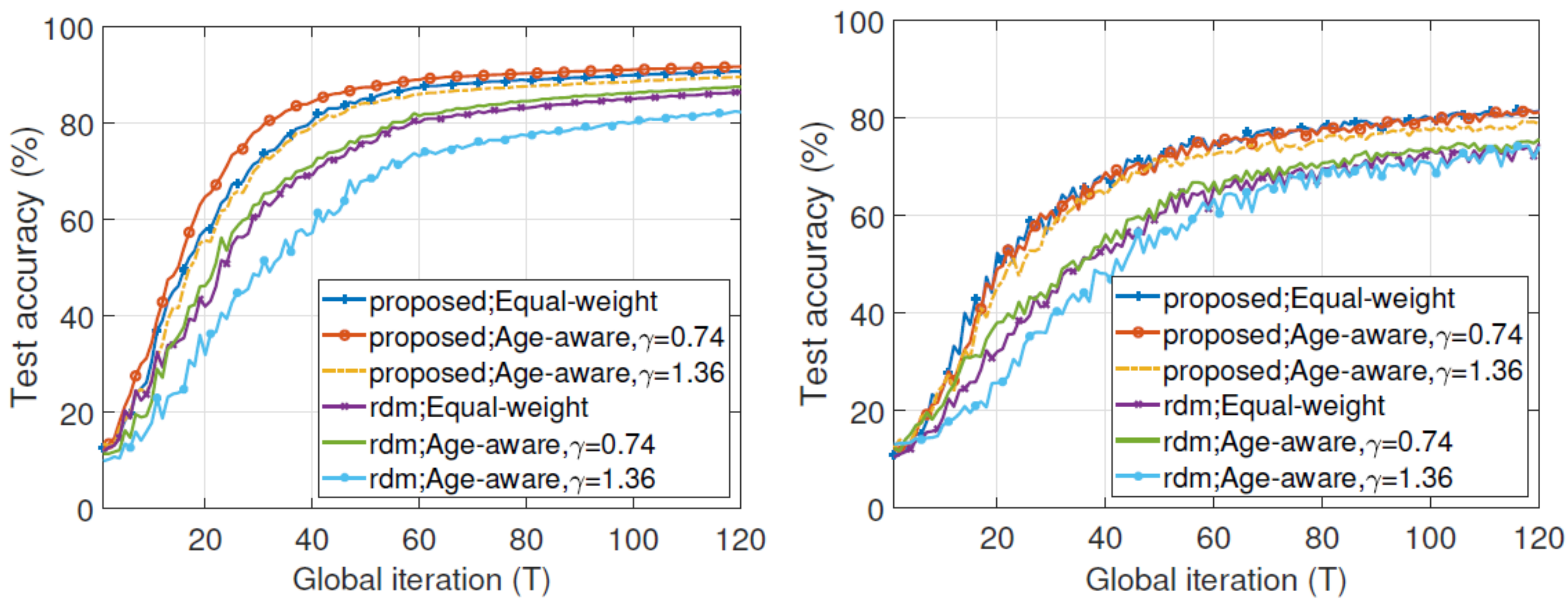



(b) Non-IID data distribution.



[5] M. M. Amiri, D. Gündüz, S. R. Kulkarni, and H. Vincent Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," IEEE Trans. on Wireless Commun., pp. 1–1, 2021.
[6] H. H. Yang, A. Arafa, T. Q. Quek, and H. V. Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in IEEE ICASSP, 2020.

# Simulation Results




(a) IID data distribution.

(b) Non-IID data distribution.

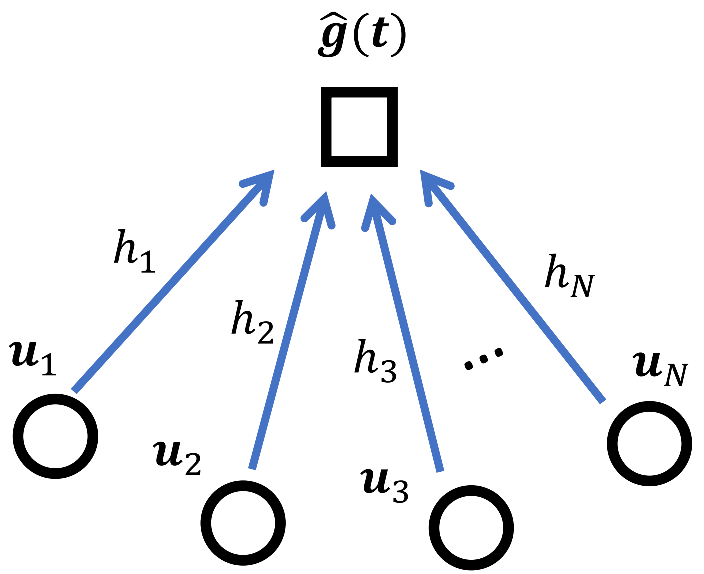
# From Higher Rate to Better Task Performance

Why do we need new wireless design for distributed intelligent systems?

Properties of AI data traffic:

- Higher data rate≠ higher impact on the task
- Temporal correlation/evolution of data
- Error-free transmission may not be necessary
- Importance matters more than fairness

# Model Aggregation over Distributed Nodes




$$\boldsymbol{u}_i(t) = \alpha(t)\nabla F_i\big(\boldsymbol{\theta}(\mathrm{t}), \mathrm{s}_i(\mathrm{t})\big)$$

Communication Goal:

Compute $\boldsymbol{g}(t) = \sum_{i=1}^{N} w_i \boldsymbol{u}_i(t)$

- Do we need to decode each stream correctly?
- What happens when N is very large?

# Goal-Oriented Communication Design

Data property:

- Stochastic gradient vector is a noisy measure itself

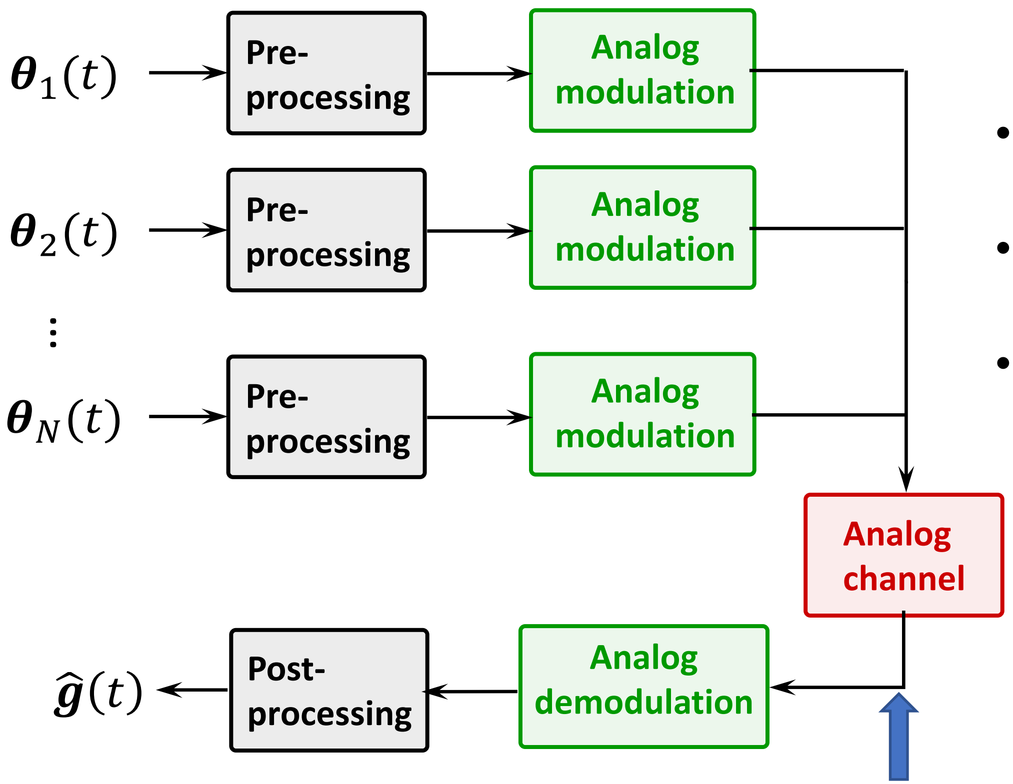
Objective:

- Received aggregated parameter vector $\hat{\boldsymbol{g}}(t)$ as close to the ground-truth $\boldsymbol{g}(t)$

Communication design:

- Exploit signal superposition in wireless channels
- Create a “linear” mapping between data value and received signal amplitude

# Over-the-Air Computation for Federated Learning




- All users share the same frequency-time resources
- Linear processing with analog modulation
- Noise and power limitation affect the reconstruction error

$$\|\hat{\boldsymbol{g}}(t) - \boldsymbol{g}(t)\|^2$$

Noise is directly added to the received superimposed signal

[7] G. Zhu, J. Xu, K. Huang and S. Cui, "Over-the-Air Computing for Wireless Data Aggregation in Massive IoT," in *IEEE Wireless Communications*, 2021.

# Computation of Nomographic Functions over Multiple Access Channels

A function $f$ of $N$ variables is nomographic if it can be represented in the form

$$f(s_1, \dots, s_N) = \phi\left(\sum_{i=1}^{N} \psi_i(s_i)\right)$$

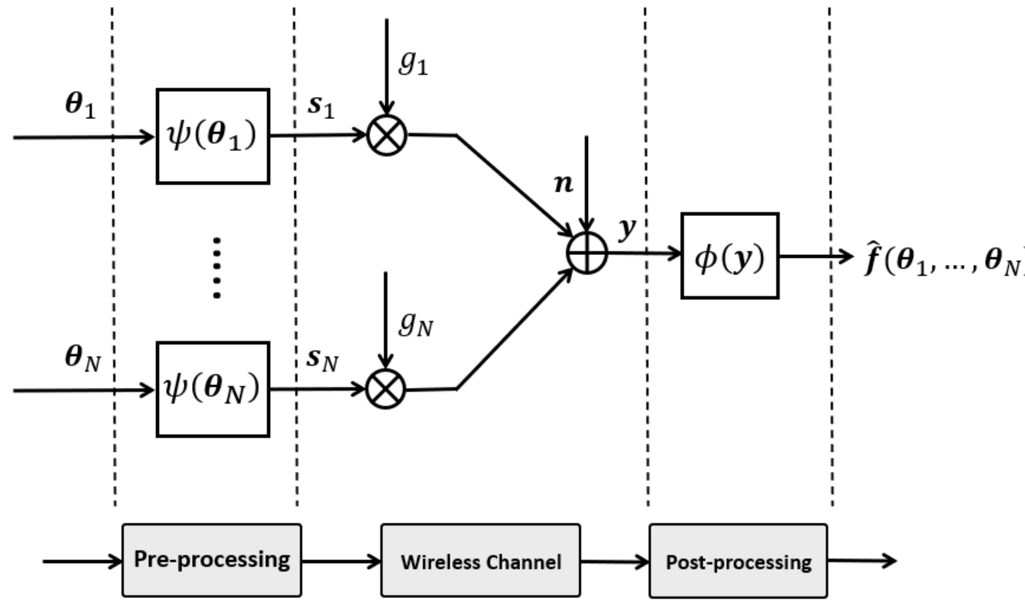
- Joint source-channel design
- Exploit signal superposition in MAC channels
- Communication is part of the computation process
- Harness interference instead of fighting it

LINKÖPING UNIVERSITY

[8] R. C. Buck, “Nomographic functions are nowhere dense,” Proceedings of the American Mathematical Society, 1982.
[9] B. Nazer and M. Gastpar, "Computation Over Multiple-Access Channels," in IEEE Transactions on Information Theory, Oct. 2007.

# OtA Computation over Fading Channels

- N senders, one receiver
- Channel gain of link $i$: $g_i$
- Power constraint: $P_{\max}$
- Computation objective
  $\hat{\boldsymbol{f}}(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N) = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}} w_i \boldsymbol{\theta}_i$




➢ Pre-processing: $\psi(\boldsymbol{\theta}_i) = \frac{w_i \eta}{g_i} \boldsymbol{\theta}_i$

➢ Wireless channel: $\boldsymbol{y} = \sum_{i=1}^{N} \psi(\boldsymbol{\theta}_i) \cdot g_i + \boldsymbol{n}$

➢ Post-processing: $\phi(\boldsymbol{y}) = {}^{\boldsymbol{y}}/_{\eta}$

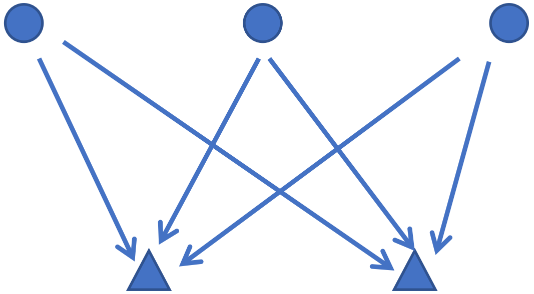
$$\eta = \sqrt{P_{\max}} \min_{i=1,\dots,N} \left\{ \frac{|g_i|}{w_i \sqrt{E_i}} \right\}$$

Bottleneck effect!

# OtA Computation with Multiple Receivers

- Set of senders: $\mathcal{S} = \{1,2,\dots,N_s\}$, set of receivers: $\mathcal{R} = \{1,2,\dots,N_r\}$
- Set of directed links: $\mathcal{E}$
- Set of senders connected to receiver $j$: $\mathcal{N}_j = \{i \in \mathcal{S} | (i,j) \in \mathcal{E}\}$
- Channel gain of link $(i,j)$: $h_{ij}$
- Data sample of sender $i$: $s_i$
- Computation objective at each receiver $j$:

$$f_j = \frac{1}{|\mathcal{N}_j|} \sum_{i \in \mathcal{N}_j} s_i$$

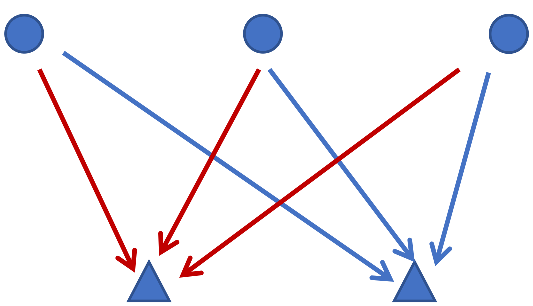


# OtA Computation with Multiple Receivers

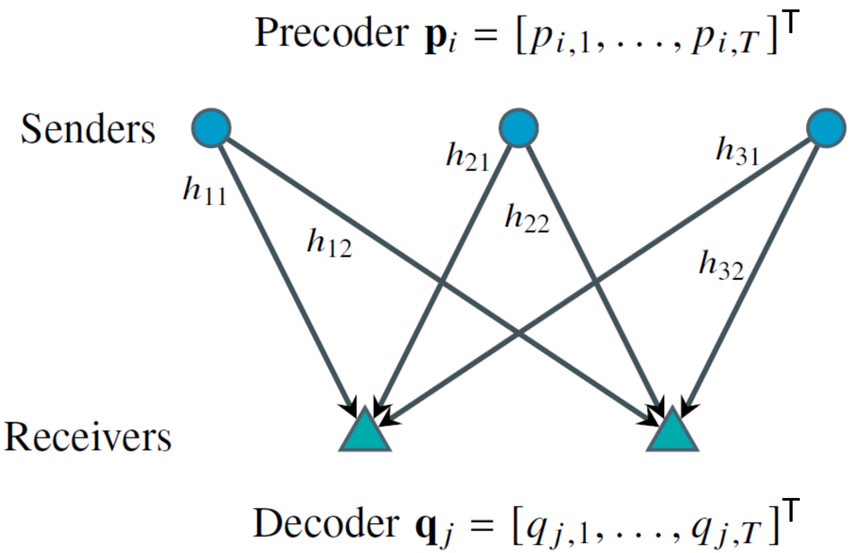
Baseline approach:

- Slot 1: first receiver active
- Slot 2: second receiver active

Number of required slots=number of receivers



# Multi-Slot Joint Precoding and Decoding Design




At slot $t$:

- Sender $i$ transmits $s_i p_{i,t}$
- Receiver $j$ receives

$$\sum_{i \in \mathcal{N}_j} s_i p_{i,t} h_{ij} + n_j$$

Multi-slot decoding:

$$\hat{\theta}_j = \sum_{t=1}^{T} \left( \sum_{i \in \mathcal{N}_j} s_i p_{i,t} h_{ij} + n_j \right) q_{j,t}$$



[10] Z. Chen, Y. Malitsky, ”Over-the-Air Computation with Multiple Receivers: A Space-Time Approach”, *arXiv preprint arXiv:2208.11751* (2022).

# Multi-Slot Joint Precoding and Decoding Design

Unbiased estimation if:

$$\sum_{t=1}^{T} p_{i,t} q_{j,t} = \frac{1}{|\mathcal{N}_j| h_{ij}} = w_{ij}$$

Conditioning on unbiased estimation, the MSE averaged over all receivers is

$$\begin{aligned}\mathrm{MSE} &= \frac{1}{N_r} \sum_{j=1}^{N_r} \mathbb{E}\left[|\hat{\theta}_j - \theta_j|^2\right] \\ &= \frac{\sigma^2}{N_r} \sum_{j=1}^{N_r} \sum_{t=1}^{T} q_{j,t}^2 .\end{aligned}$$

# Problem Formulation

Define $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_{N_s}]$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_{N_r}]$

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{N_r} \|\mathbf{q}_j\|^2 \\ \text{subject to} \quad & \mathbf{p}_i \cdot \overline{\mathbf{q}}_j = w_{ij}, \forall (i,j) \in \mathcal{E} \\ & \|\mathbf{p}_i\|^2 \le C_i, \forall i \in \mathcal{S}. \end{aligned}$$

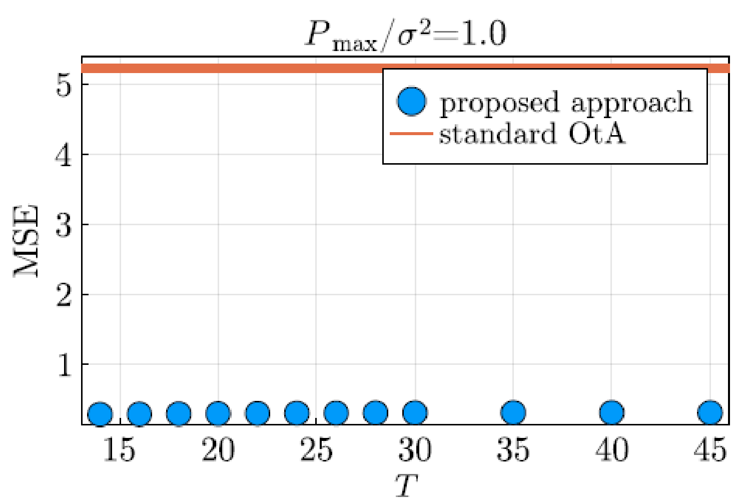
$$\begin{aligned} \underset{\mathbf{P},\mathbf{Q}}{\text{minimize}} \quad & \|\mathbf{Q}\|_F^2 \\ \text{subject to} \quad & \mathbf{P}^\top \mathbf{Q} = \mathbf{W} \\ & \|\mathbf{p}_i\|^2 \le C_i, i = 1, \dots, N_s. \end{aligned}$$

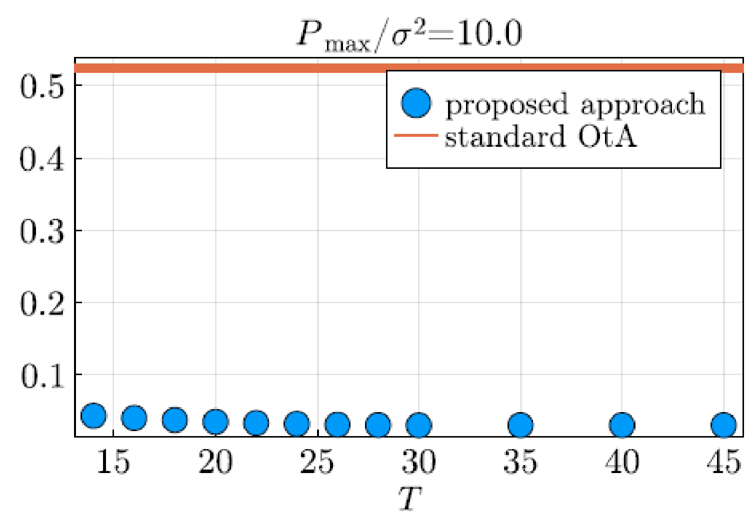
$$\begin{aligned} \underset{\mathbf{P},\mathbf{Q}}{\text{minimize}} \quad & \|\mathbf{P}^\top \mathbf{Q} - \mathbf{W}\|_F^2 + \lambda \|\mathbf{Q}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{p}_i\|^2 \le C_i, i = 1, \dots, N_s, \end{aligned}$$

# Simulation Results

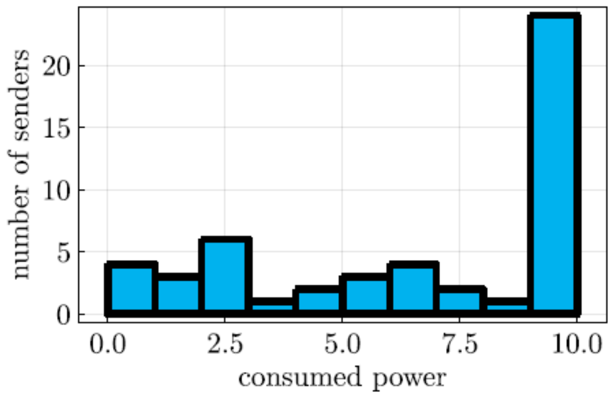
Simulation setup:

- 50 senders, 30 receivers
- Each sender randomly connected to 20 receivers
- Data and channel gain randomly generated from $\mathcal{CN}(0,1)$



# Simulation Results



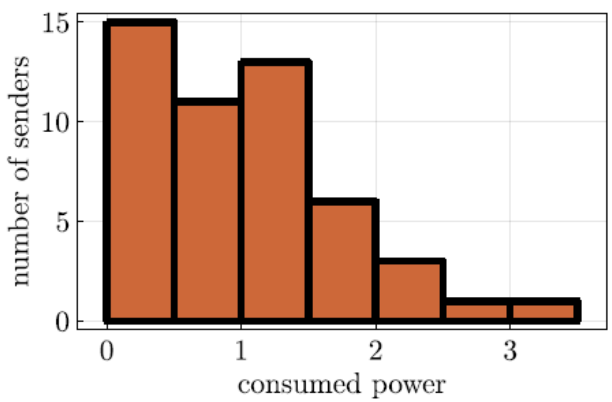




Fig. 3: Power consumption: proposed (above) and standard OtA (bottom) approaches.

Remark:

- With proposed approach, many senders reach the power limit
- With standard (parallel) OtA, much available power is unconsumed
- Standard OtA can be potentially optimized

Conclusion:

- With multi-slot joint precoding and decoding, we can save communication resources (e.g., time slots) as compared to standard parallel OtA approach

# OtA FL with multiple antennas

If no channel state information, with multiple-antenna receiver, what can we do?

- Users send orthogonal pilot signals and data
- Server estimates individual channels
- Server decodes the signal using estimated individual channels

Or

- Users send one common pilot and data
- Server estimates the sum channel
- Server decodes the signal using estimated sum channel



[11] E. Becirovic, Z.Chen, and E. G. Larsson. "Optimal MIMO Combining for Blind Federated Edge Learning with Gradient Sparsification." *SPAWC, 2022.*
[12] M. M. Amiri, T. M. Duman, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Blind federated edge learning," IEEE Transactions on Wireless Communications, 2021.

# Estimating Individual Channels vs. Estimating Sum Channel

Estimating $\hat{\boldsymbol{G}}$

$$\boldsymbol{Y}_{\text{p,orth}} = \sum_{k=1}^{K} \sqrt{\rho\tau_{\text{p}}}\boldsymbol{g}_k\boldsymbol{\phi}_k^{\text{H}} + \boldsymbol{N}_{\text{p}}$$

$$\hat{\boldsymbol{g}}_k = \frac{\sqrt{\rho\tau_{\text{p}}}\beta_k}{1+\rho\tau_{\text{p}}\beta_k}\boldsymbol{Y}_{\text{p,orth}}\boldsymbol{\phi}_k$$

$$\gamma_k = \frac{\rho\tau_{\text{p}}\beta_k^2}{1+\rho\tau_{\text{p}}\beta_k}$$

Estimating $\hat{\boldsymbol{h}}_{\text{sum}}$

$$\begin{aligned}\boldsymbol{Y}_{\text{p,sum}} &= \sum_{k=1}^{K} \sqrt{\rho\tau_{\text{p}}\frac{\beta_{\text{min}}}{\beta_k}}\boldsymbol{g}_k\boldsymbol{\phi}^{\text{H}} + \boldsymbol{N}_{\text{p}} \\ &= \sum_{k=1}^{K} \sqrt{\rho\tau_{\text{p}}\beta_{\text{min}}}\boldsymbol{h}_k\boldsymbol{\phi}^{\text{H}} + \boldsymbol{N}_{\text{p}} \\ &= \sqrt{\rho\tau_{\text{p}}\beta_{\text{min}}}\boldsymbol{h}_{\text{sum}}\boldsymbol{\phi}^{\text{H}} + \boldsymbol{N}_{\text{p}}\end{aligned}$$

$$\hat{\boldsymbol{h}}_{\text{sum}} = \frac{\sqrt{\rho\tau_{\text{p}}\beta_{\text{min}}}K}{1+\rho\tau_{\text{p}}\beta_{\text{min}}K}\boldsymbol{Y}_{\text{p,sum}}\boldsymbol{\phi}$$

$$\bar{\gamma} = \frac{\rho\tau_{\text{p}}\beta_{\text{min}}K^2}{1+\rho\tau_{\text{p}}\beta_{\text{min}}K}$$



Slide credit: Ema Becirovic

# Best Linear Unbiased Estimator for Signal Decoding

$$\boldsymbol{Y} = \sum_{k=1}^{K} \sqrt{\rho\eta_k}\boldsymbol{g}_k\boldsymbol{x}_k^{\mathsf{T}} + \boldsymbol{N}$$

## BLUE

$$[\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_k] = \left(\frac{1}{\sqrt{\rho}}\boldsymbol{D}_\eta^{-1/2}(\hat{\boldsymbol{G}}^{\mathsf{H}}\hat{\boldsymbol{G}})^{-1}\hat{\boldsymbol{G}}^{\mathsf{H}}\boldsymbol{Y}\right)^{\mathsf{T}}$$

$$\boldsymbol{D}_\eta = \mathrm{diag}(\eta_1, \ldots, \eta_K)$$

Full power!

## Benchmark

$$\hat{\boldsymbol{x}} = \widehat{\sum_{k=1}^{K} w_k\boldsymbol{x}_k} = c\,(\hat{\boldsymbol{h}}_{\mathsf{sum}}^{\mathsf{H}}\boldsymbol{Y})^{\mathsf{T}}$$

$$\max_k \|\sqrt{\eta_k}\boldsymbol{x}_k\|_2^2 = \max_k \eta\frac{w_k^2}{\beta_k}\|\boldsymbol{x}_k\|_2^2 = T$$

$$c = \frac{K}{M\sqrt{\eta\rho\bar{\gamma}}}$$



Slide credit: Ema Becirovic

# What Happens at Client Side

**Sparse BLUE**

1: **for** client $k \in \{1, \ldots, K\}$ **in parallel do**
2: receive (error free) $\boldsymbol{\theta}(t)$
3: obtain $\Delta\boldsymbol{\theta}_k(t)$ from stochastic gradient descent
4: $\boldsymbol{x}_k^{\text{full}} \leftarrow \mathrm{SPLIT}(\Delta\boldsymbol{\theta}_k(t))$
5: $\boldsymbol{x}_k^{\text{full}} \leftarrow \boldsymbol{x}_k^{\text{full}} + \boldsymbol{r}_k$
6: $\boldsymbol{x}_k^{\text{sparse}} \leftarrow \mathrm{SPARSE}(\boldsymbol{x}_k^{\text{full}}, S)$
7: $\boldsymbol{r}_k \leftarrow \boldsymbol{x}_k^{\text{full}} - \boldsymbol{x}_k^{\text{sparse}}$
8: $\boldsymbol{x}_k \leftarrow \boldsymbol{A}_k^t \boldsymbol{x}_k^{\text{sparse}}$
9: transmit $\phi_k$, $\sqrt{\eta_k}\boldsymbol{x}_k$
10: **end for**

**Sparse SUM**

1: **for** client $k \in \{1, \ldots, K\}$ **in parallel do**
2: receive (error free) $\boldsymbol{\theta}(t)$
3: obtain $\Delta\boldsymbol{\theta}_k(t)$ from stochastic gradient descent
4: $\boldsymbol{x}_k^{\text{full}} \leftarrow \mathrm{SPLIT}(\Delta\boldsymbol{\theta}_k(t))$
5: $\boldsymbol{x}_k^{\text{full}} \leftarrow \boldsymbol{x}_k^{\text{full}} + \boldsymbol{r}_k$
6: $\boldsymbol{x}_k^{\text{sparse}} \leftarrow \mathrm{SPARSE}(\boldsymbol{x}_k^{\text{full}}, S)$
7: $\boldsymbol{r}_k \leftarrow \boldsymbol{x}_k^{\text{full}} - \boldsymbol{x}_k^{\text{sparse}}$
8: $\boldsymbol{x}_k \leftarrow \boldsymbol{A}^t \boldsymbol{x}_k^{\text{sparse}}$
9: transmit $\phi$, $\sqrt{\eta_k}\boldsymbol{x}_k$
10: **end for**



Slide credit: Ema Becirovic

# What Happens at Server Side

**Sparse BLUE**

receive $\boldsymbol{Y}_{\text{p,orth}}$ and $\boldsymbol{Y}$

estimate $\hat{\boldsymbol{G}}$

**for** $k \in \{1, \ldots, K\}$ **do**

    $\hat{\boldsymbol{x}}_k \leftarrow \frac{1}{\sqrt{\eta_k \rho}} [\hat{\boldsymbol{G}}(\hat{\boldsymbol{G}}^{\mathsf{H}}\hat{\boldsymbol{G}})^{-1}]_k^{\mathsf{H}} \boldsymbol{Y}$

    solve sparsity problem to get $\hat{\boldsymbol{x}}_k^{\text{sparse}}$

    $\widehat{\Delta\boldsymbol{\theta}}_k(t, \tau) \leftarrow \text{UNSPLIT}(\hat{\boldsymbol{x}}_k^{\text{sparse}})$

**end for**

$\widehat{\Delta\boldsymbol{\theta}}(t) \leftarrow \sum_{k=1}^{K} w_k \widehat{\Delta\boldsymbol{\theta}}_k(t)$

$\boldsymbol{\theta}(t+1) \leftarrow \boldsymbol{\theta}(t) + \alpha_t^{\text{global}} \widehat{\Delta\boldsymbol{\theta}}(t)$

broadcast (error free) $\boldsymbol{\theta}(t+1)$

Individual processing of clients!

**Sparse SUM**

receive $\boldsymbol{Y}_{\text{p,sum}}$ and $\boldsymbol{Y}$

estimate $\hat{\boldsymbol{h}}_{\text{sum}}$

$\hat{\boldsymbol{x}} \leftarrow \frac{K}{M\sqrt{\rho\eta\bar{\gamma}}} \hat{\boldsymbol{h}}_{\text{sum}}^{\mathsf{H}} \boldsymbol{Y}$

solve sparsity problem to get $\widehat{\left(\sum_{k=1}^{K} w_k \boldsymbol{x}_k^{\text{sparse}}\right)}$

$\widehat{\Delta\boldsymbol{\theta}}(t) \leftarrow \text{UNSPLIT}\left(\widehat{\left(\sum_{k=1}^{K} w_k \boldsymbol{x}_k^{\text{sparse}}\right)}\right)$

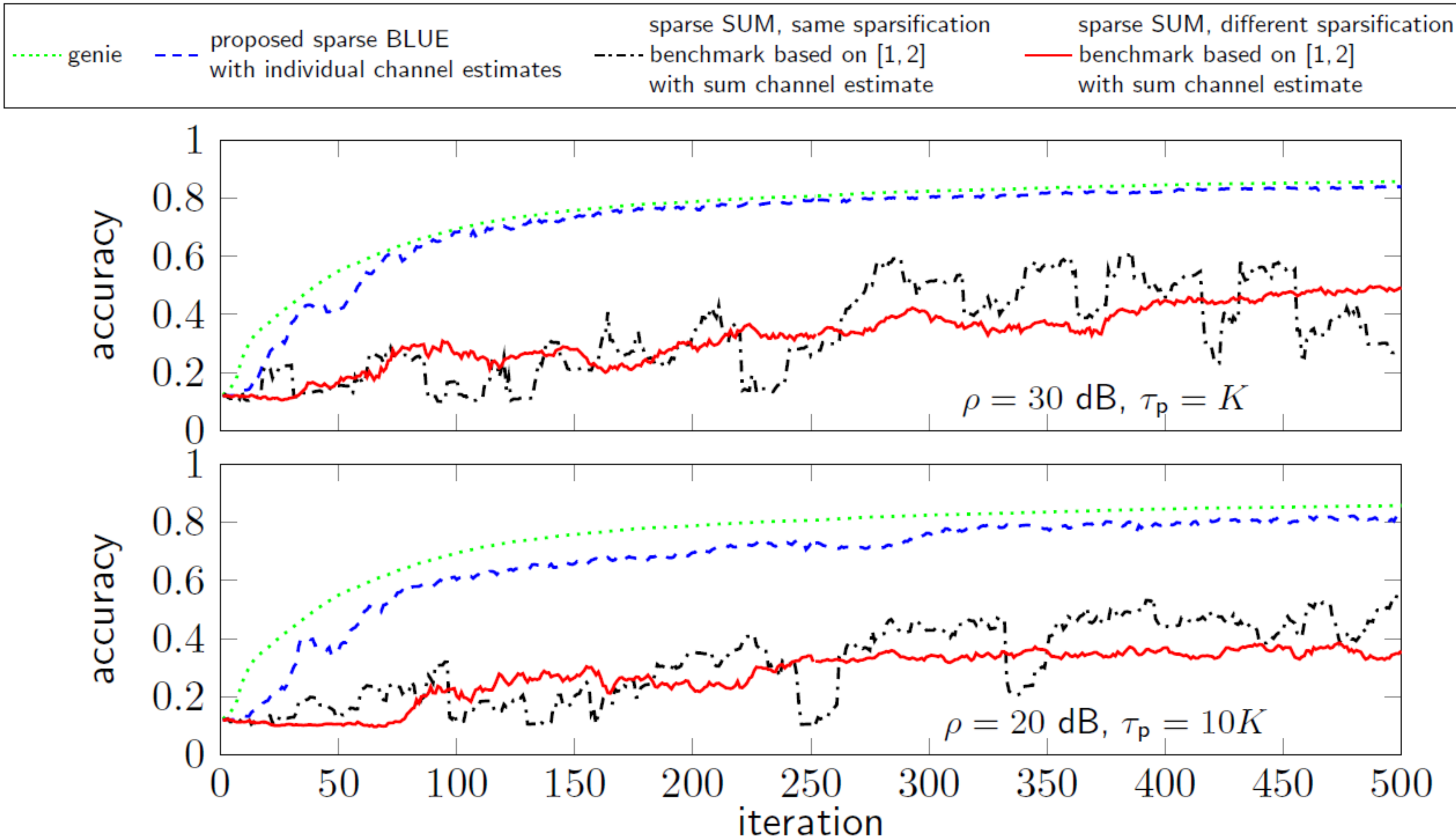
$\boldsymbol{\theta}(t+1) \leftarrow \boldsymbol{\theta}(t) + \alpha_t^{\text{global}} \widehat{\Delta\boldsymbol{\theta}}(t)$

broadcast (error free) $\boldsymbol{\theta}(t+1)$



Slide credit: Ema Becirovic

# Simulation Results






Slide credit: Ema Becirovic

# Applications of OtA Computation

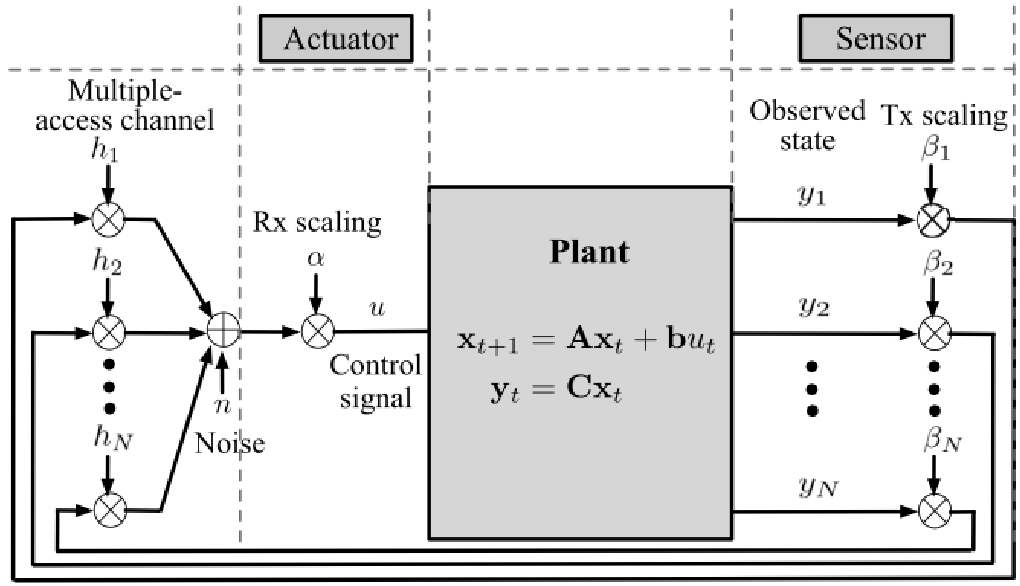
- Distributed learning
- Distributed control
- Distributed inference and estimation

- Key feature: the goal of communication is to compute some functions of data from a set of distributed nodes approximately correct, instead of receiving each independent data stream without errors



[13] T. Sery, N. Shlezinger, K. Cohen and Y. C. Eldar, "Over-the-Air Federated Learning From Heterogeneous Data," in *IEEE Transactions on Signal Processing*, 2021.
[14] P. Park, P. D. Marco and C. Fischione, "Optimized Over-the-Air Computation for Wireless Control Systems," in *IEEE Communications Letters*, Feb. 2022.

# Example: Wireless Control Systems




- $x_t$: system state
- $y_t$: observed output (measured at different sensors)
- $u_t$: control input signal



[14] P. Park, P. D. Marco and C. Fischione, "Optimized Over-the-Air Computation for Wireless Control Systems," in *IEEE Communications Letters*, Feb. 2022.

# Challenges of OtA Aggregation in FL

MSE of aggregated data may not be the most suitable performance metric, need task-specific metric design

| Scheduled users | 50 | 40 | 30 | 20 | 10 |
|---|---|---|---|---|---|
| Test accuracy | 92% | 71.65% | 56.86% | 42.45% | 27.37% |
| MSE | $7.75e-14$ | $3.70e-15$ | $7.54e-17$ | $7.60e-18$ | $1.81e-18$ |

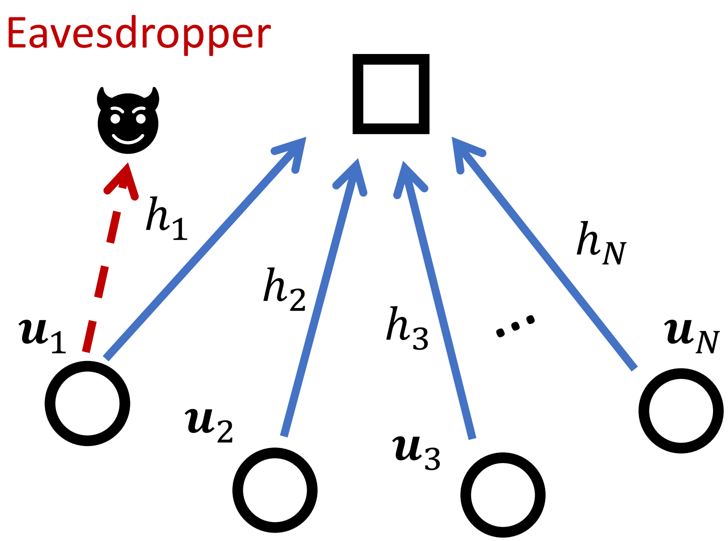
- Opting out bottleneck users can reduce the MSE of aggregated data
- Aggregated model might be biased (especially with non-IID training data) when opting out users



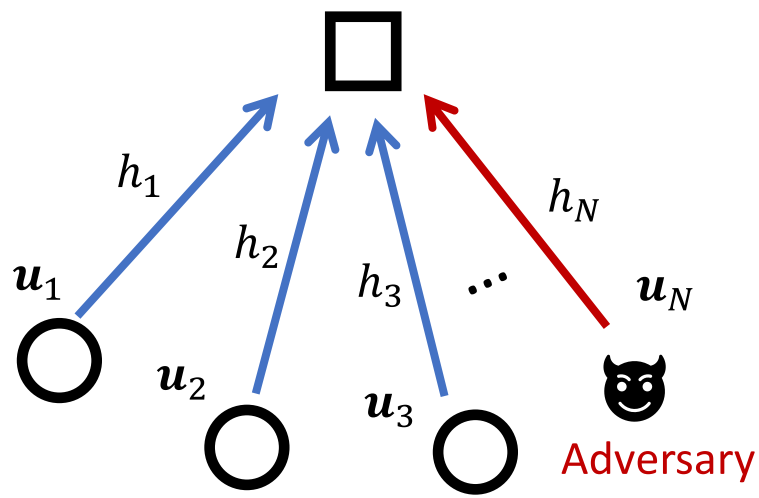
[15] Z. Chen, E. G. Larsson, C. Fischione, M. Johansson, Y. Malitsky, ”Over-the-Air Computation for Distributed Systems: Something Old and Something New”, submitted, 2022.

# Challenges of OtA Aggregation in FL

## Privacy, security and robustness




Membership inference attack



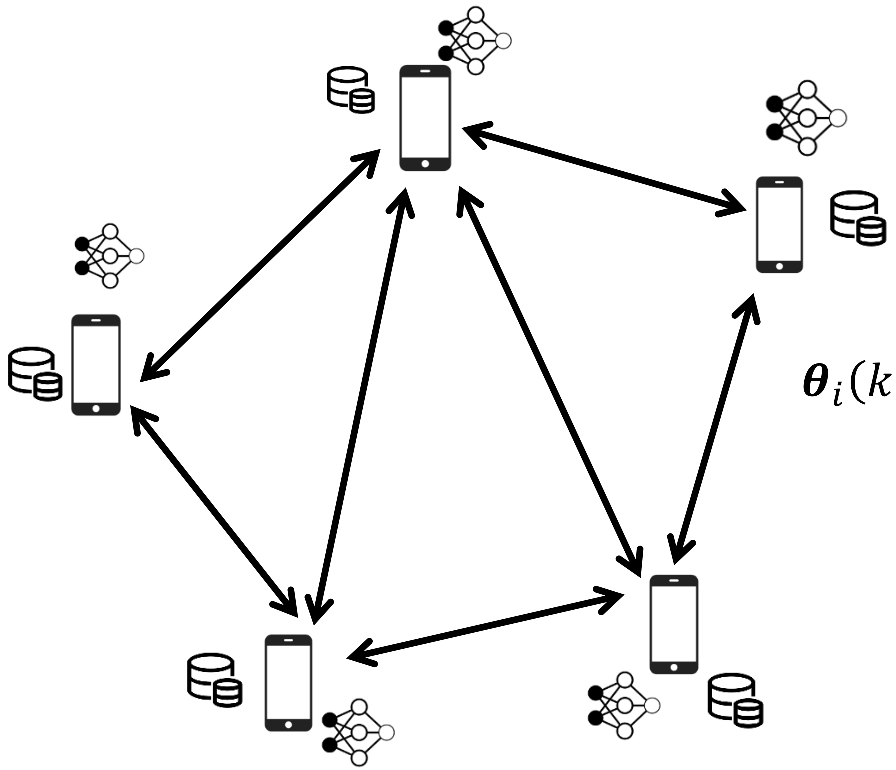

Model poisoning attack



[16] D. Liu and O. Simeone, "Privacy for Free: Wireless Federated Learning via Uncoded Transmission With Adaptive Power Control," in *IEEE Journal on Selected Areas in Communications*, Jan. 2021.
[17] J. Liao, Z. Chen, and E. G. Larsson, "Over-the-air federated learning with privacy protection via correlated additive perturbations," Allerton Conference on Communication, Control, and Computing, 2022.

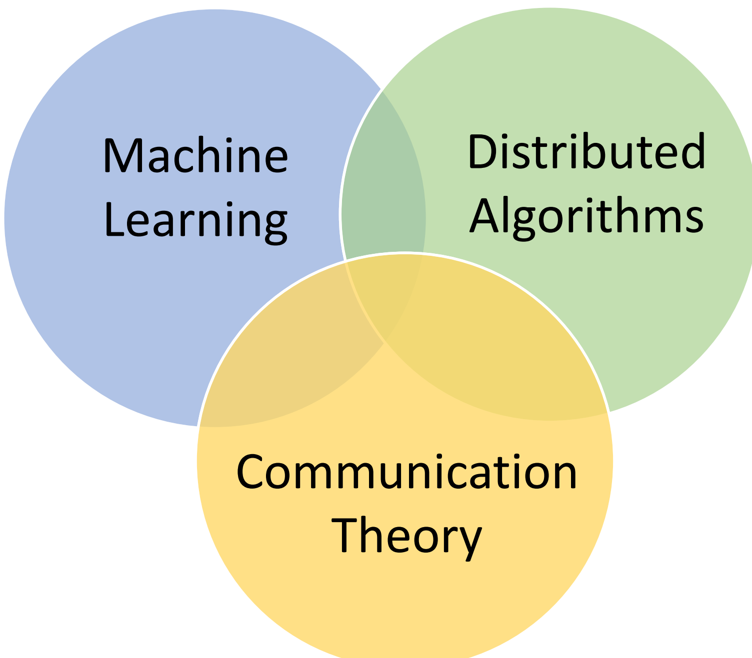
# Challenges of OtA Computation

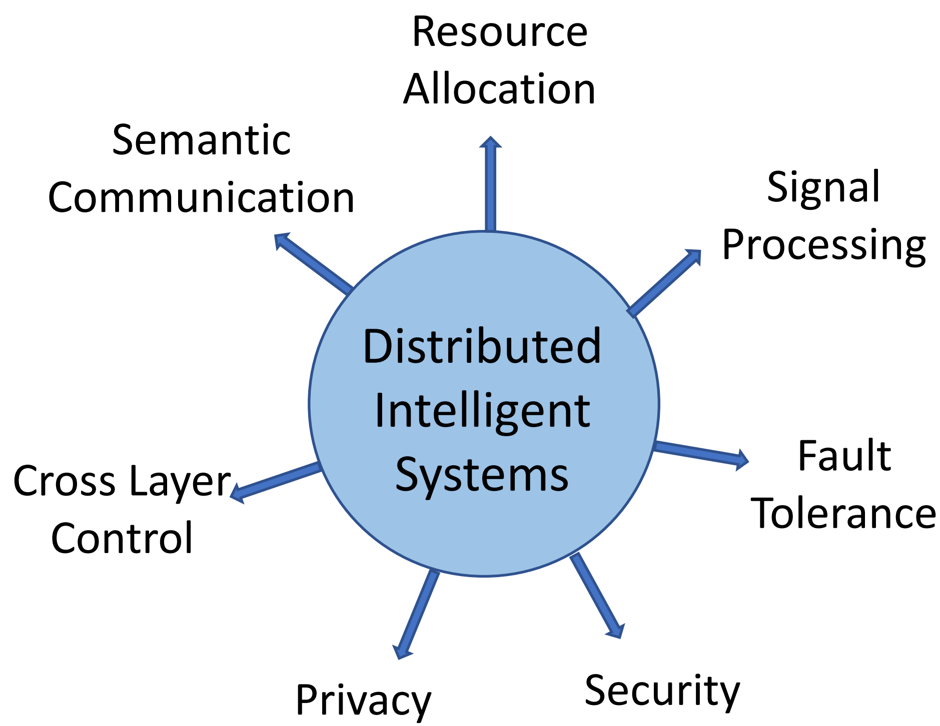
## Communication design in fully decentralized ML



$$\boldsymbol{\theta}_i(k+1) = \underbrace{\sum_{j=1}^{N} w_{ij}\,\boldsymbol{\theta}_j(k)}_{\text{Consensus}} - \underbrace{\alpha \nabla f_i(\boldsymbol{\theta}_i(k))}_{\text{Innovation}}$$

# Distributed Intelligence over Wireless Networks: An Interdisciplinary View



Thank you!

LINKÖPING
UNIVERSITY