Online Model Updating with Analog Aggregation in Wireless Edge Learning

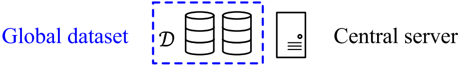
Juncheng Wang[*], Min Dong[†], Ben Liang[*], Gary Boudreau[‡], and Hatem Abou-Zeid[‡]

[*]Department of Electrical and Computer Engineering, University of Toronto, Canada
[†]Department of Electrical, Computer, and Software Engineering, Ontario Tech University, Canada
[‡]Ericsson Canada, Canada

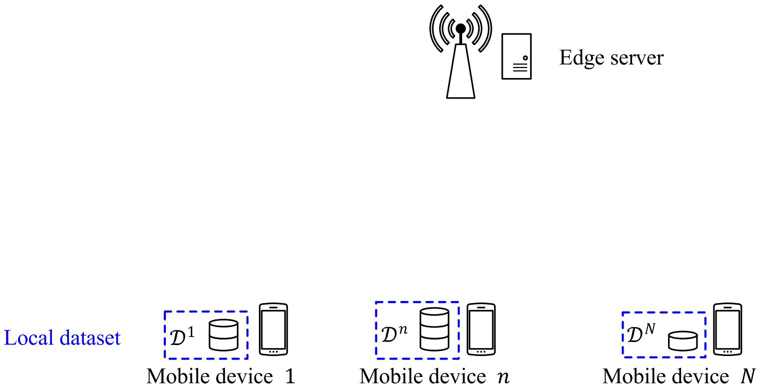
May 04, 2022

INFOCOM'22

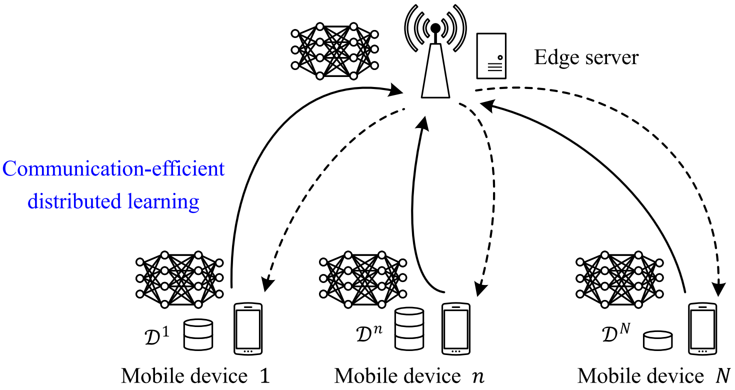
Wireless Edge Learning



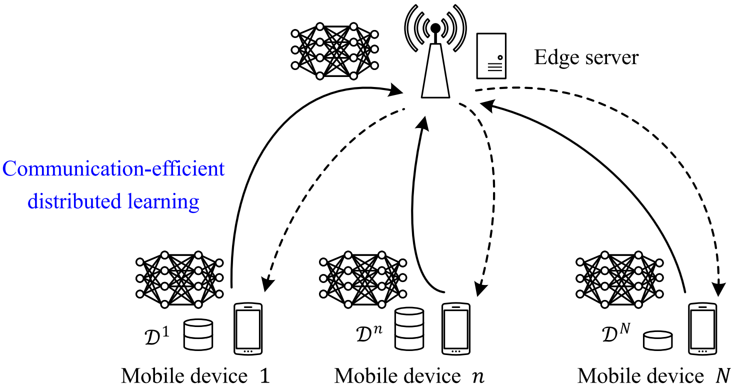
Wireless Edge Learning



Wireless Edge Learning



Wireless Edge Learning



- Integrate techniques from both machine learning and communications.

Federated Learning (FL) Objective

- N devices cooperate to find a global model $\mathbf{x}^\star$ from local datasets $\{\mathcal{D}^n\}$.

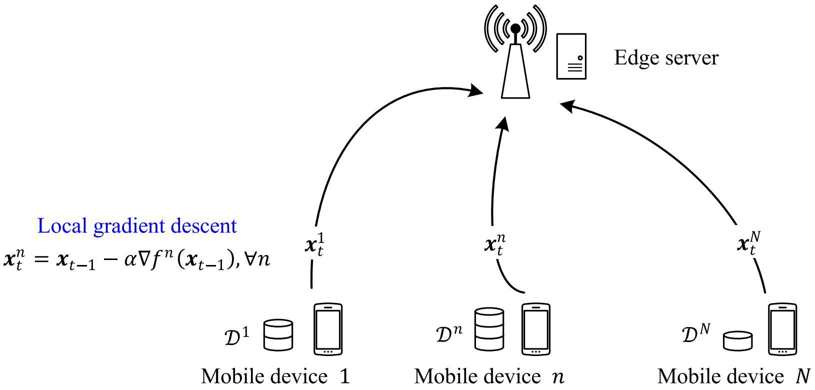
Federated Learning (FL) Objective

- N devices cooperate to find a global model $\mathbf{x}^\star$ from local datasets $\{\mathcal{D}^n\}$.

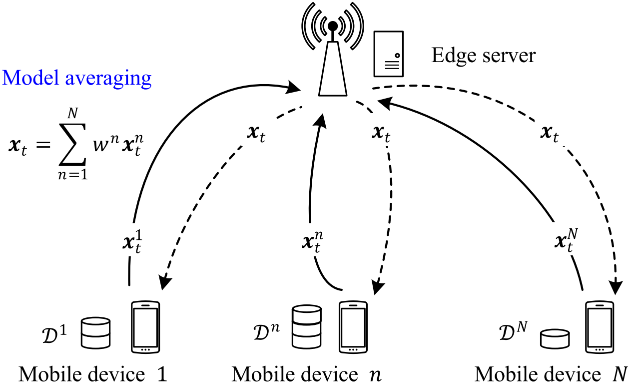
$$\min_{\mathbf{x}} \quad f(\mathbf{x}) = \sum_{n=1}^{N} w^n f^n(\mathbf{x}).$$

- $f^n(\mathbf{x}) = \frac{1}{|\mathcal{D}^n|} \sum_{i=1}^{|\mathcal{D}^n|} l(\mathbf{x}; \mathbf{u}^{n,i}, v^{n,i})$: local loss function.
- $l(\mathbf{x}; \mathbf{u}^{n,i}, v^{n,i})$: sample-wise loss function.
- $(\mathbf{u}^{n,i}, v^{n,i})$: data vector $\mathbf{u}^{n,i}$ with label $v^{n,i}$.
- $w^n = \frac{|\mathcal{D}^n|}{|\mathcal{D}|}$: weight on mobile device n.

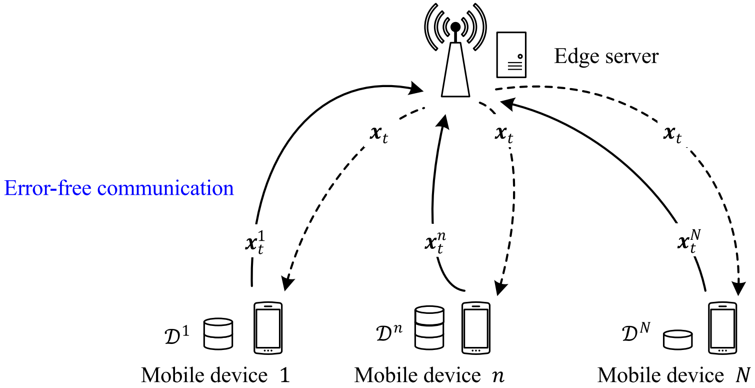
Error-Free FL Algorithm



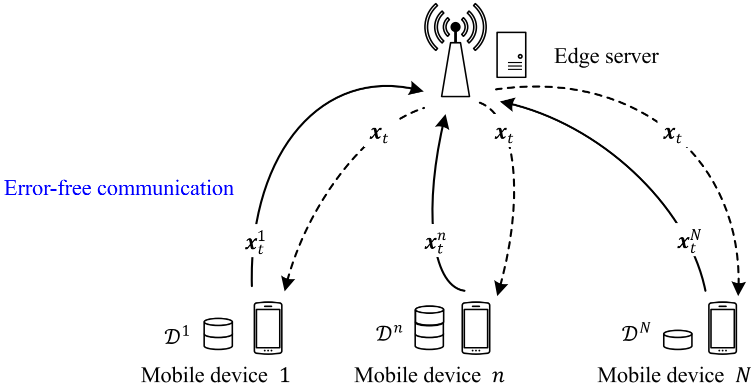
Error-Free FL Algorithm



Error-Free FL Algorithm



Error-Free FL Algorithm



- Does not consider the wireless communication layer.

Reducing Communication Overhead

- Machine learning
  - Quantization: certain levels to represent model values.

Reducing Communication Overhead

- Machine learning
  - Quantization: certain levels to represent model values.
  - Sparsification: set small model values to zero.

Reducing Communication Overhead

- Machine learning
  - Quantization: certain levels to represent model values.
  - Sparsification: set small model values to zero.
  - Local updates: multiple steps of gradient descent before aggregation.

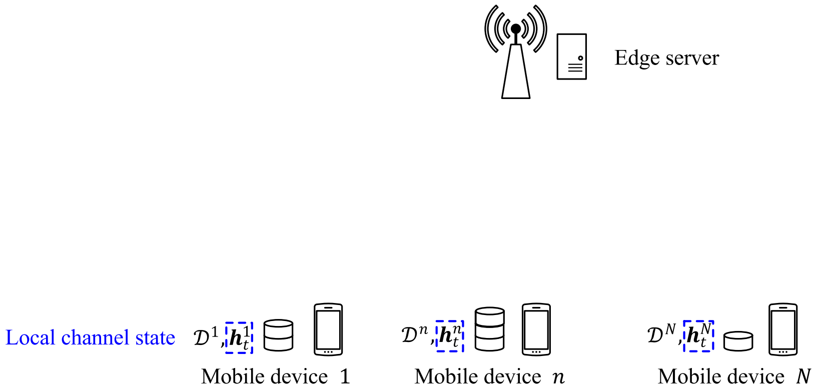
Reducing Communication Overhead

- Machine learning
  - Quantization: certain levels to represent model values.
  - Sparsification: set small model values to zero.
  - Local updates: multiple steps of gradient descent before aggregation.
- Wireless communication
  - Digital communication: entropy coding.

Reducing Communication Overhead

- Machine learning
  - Quantization: certain levels to represent model values.
  - Sparsification: set small model values to zero.
  - Local updates: multiple steps of gradient descent before aggregation.
- Wireless communication
  - Digital communication: entropy coding.
  - Analog communication: over-the-air (OTA) computation.

FL with OTA Analog Aggregation



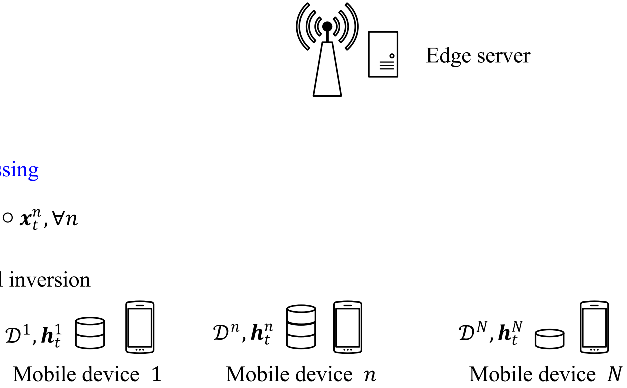
FL with OTA Analog Aggregation

Pre-processing

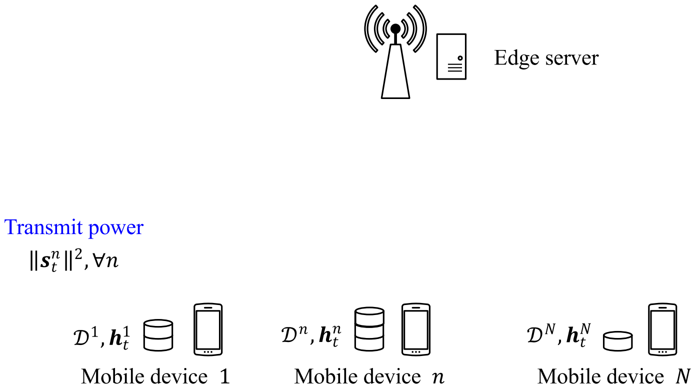
$$\boldsymbol{s}_t^n = \frac{1}{\lambda_t} w^n \boldsymbol{b}_t^n \circ \boldsymbol{x}_t^n, \forall n$$

Channel inversion

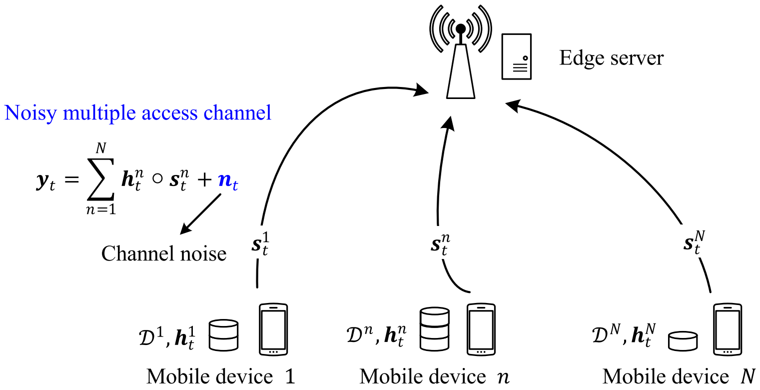
Power scaling



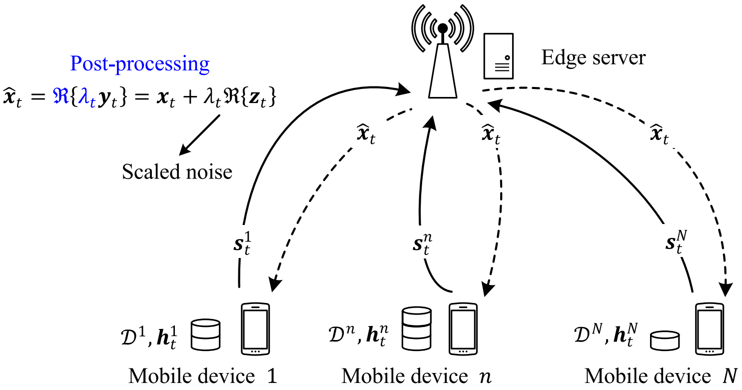
FL with OTA Analog Aggregation



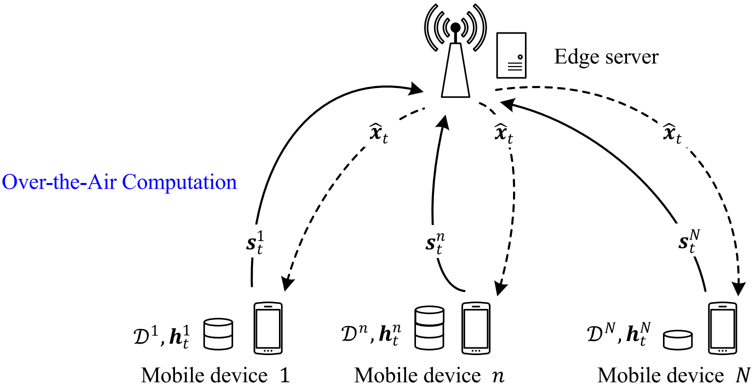
FL with OTA Analog Aggregation



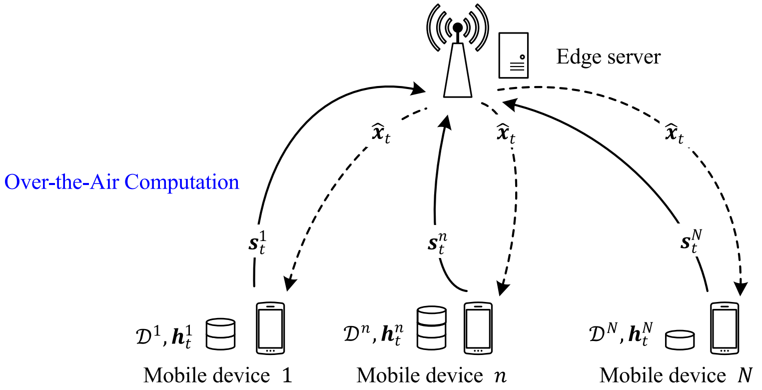
FL with OTA Analog Aggregation



FL with OTA Analog Aggregation

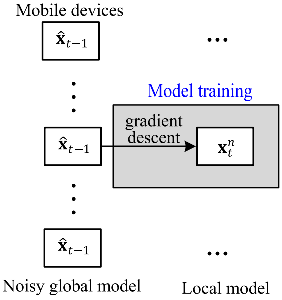


FL with OTA Analog Aggregation

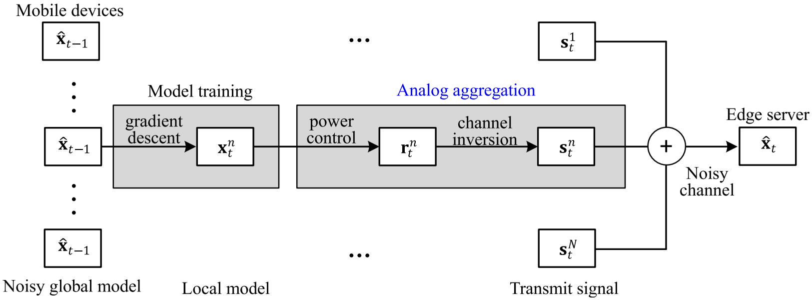


- Lower latency and bandwidth requirement than digital communication.

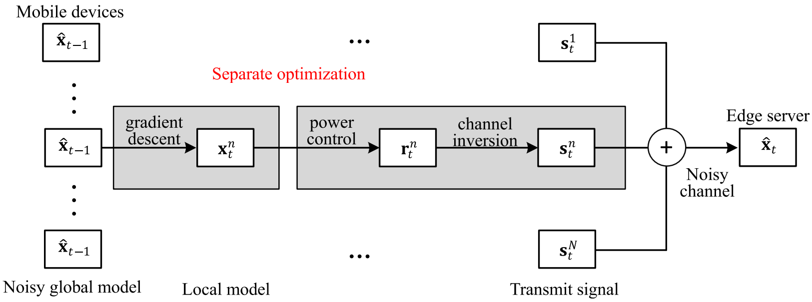
Existing Works



Existing Works

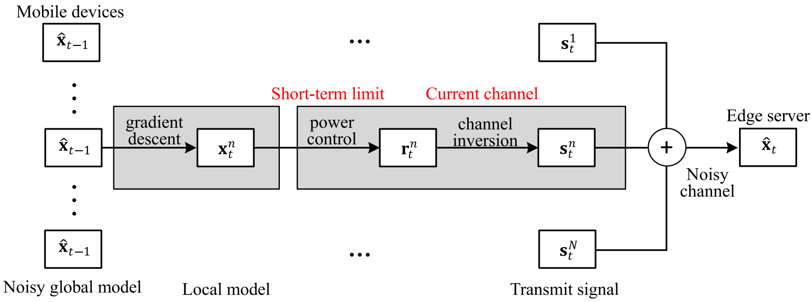


Existing Works



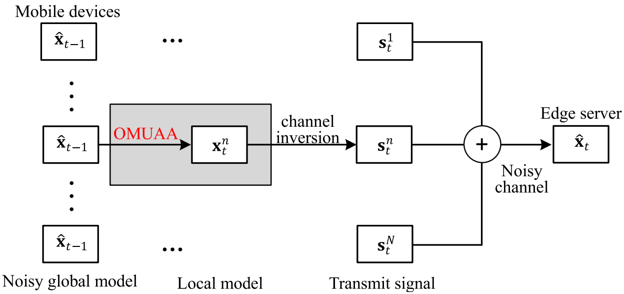
- All separately optimize model training and wireless transmission.

Existing Works



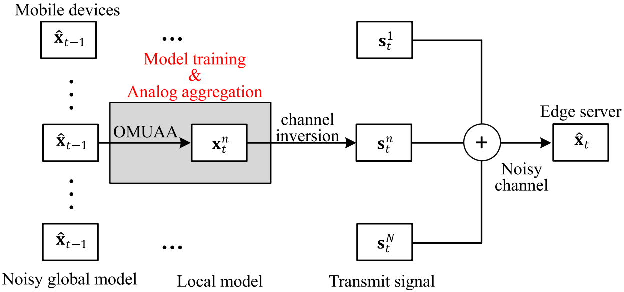
- All separately optimize model training and wireless transmission.
- All focus on per-iteration optimization.

Our Contributions



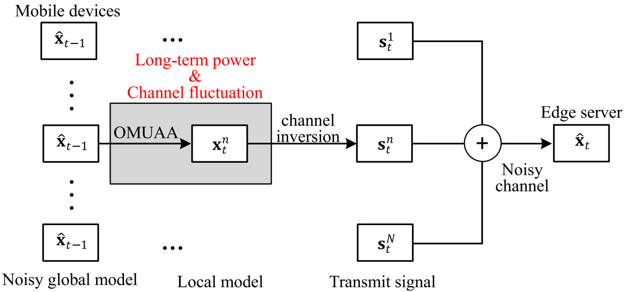
- Online Model Updating with Analog Aggregation (OMUAA) Algorithm

Our Contributions



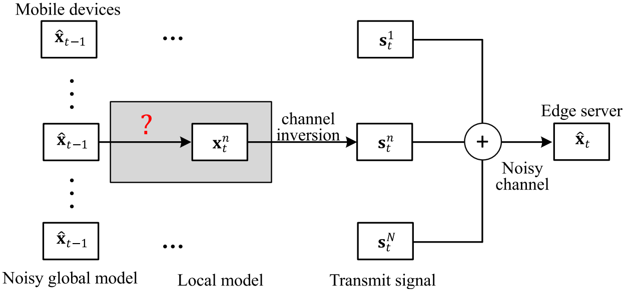
- Online Model Updating with Analog Aggregation (OMUAA) Algorithm
  - Jointly optimize model training and analog aggregation.

Our Contributions



- Online Model Updating with Analog Aggregation (OMUAA) Algorithm
  - Jointly optimize model training and analog aggregation.
  - Online solutions adapt to channel fluctuation under long-term power limits.

Our Contributions



- Online Model Updating with Analog Aggregation (OMUAA) Algorithm
  - Jointly optimize model training and analog aggregation.
  - Online solution adapts to channel fluctuation under long-term power limit.
  - Performance bounds on both computation and communication metrics.

Online Problem Formulation

$$
\begin{aligned}
\mathbf{P1}: \quad & \min_{\{\mathbf{x}_t^n \in \mathcal{X}\}} \quad \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{f(\hat{\mathbf{x}}_t)\} \\
& \text{s.t.} \quad \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{g_t^n(\mathbf{x}_t^n)\} \le 0, \quad \forall n.
\end{aligned}
$$

- $g_t^n(\mathbf{x}) = \frac{(w^n)^2}{\lambda_t^2}\|\mathbf{b}_t^n \circ \mathbf{x}\|^2 - \bar{P}^n$: long-term transmit power function.
- $\bar{P}^n$: average transmit power budget.
- $\mathcal{X} = \{\mathbf{x} : -\mathbf{x}_{\max} \preceq \mathbf{x} \preceq \mathbf{x}_{\max}\}$: possible short-term constraints.

Online Problem Formulation

$$\mathbf{P1}: \quad \min_{\{\mathbf{x}_t^n \in \mathcal{X}\}} \quad \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{f(\hat{\mathbf{x}}_t)\}$$

$$\text{s.t.} \quad \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{g_t^n(\mathbf{x}_t^n)\} \le 0, \quad \forall n.$$

- $g_t^n(\mathbf{x}) = \frac{(w^n)^2}{\lambda_t^2}\|\mathbf{b}_t^n \circ \mathbf{x}\|^2 - \bar{P}^n$: long-term transmit power function.
- $\bar{P}^n$: average transmit power budget.
- $\mathcal{X} = \{\mathbf{x} : -\mathbf{x}_{\max} \preceq \mathbf{x} \preceq \mathbf{x}_{\max}\}$: possible short-term constraints.

- Online algorithm with strong performance guarantees.

Algorithm Intuition

- Local virtual queue for long-term power constraint

$$Q_t^n = \max\{Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), 0\}.$$

Algorithm Intuition

- Local virtual queue for long-term power constraint

$$Q_t^n = \max\{Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), 0\}.$$

- Maintain queue stability

$$\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq Q_T^n.$$

Algorithm Intuition

- Local virtual queue for long-term power constraint

$$Q_t^n = \max\{Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), 0\}.$$

- Maintain queue stability

$$\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq Q_T^n.$$

- Minimize an upper bound of a drift-plus-penalty metric

$$\gamma\Big[\underbrace{\frac{1}{2}(Q_t^n)^2 - \frac{1}{2}(Q_{t-1}^n)^2}_{\text{Lyapunov drift}}\Big] + \underbrace{\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x} - \hat{\mathbf{x}}_{t-1} \rangle + \frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2}_{\text{penalty on training loss}}.$$

OMUAA: Mobile Device n's Algorithm

1: Update local model $\mathbf{x}_t^n$ by solving

$$\mathbf{P2}^n: \quad \min_{\mathbf{x}\in\mathcal{X}} \quad \overbrace{\underbrace{\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x} - \hat{\mathbf{x}}_{t-1} \rangle + \frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2}_{\text{training loss}} + \underbrace{\gamma Q_{t-1}^n g_t^n(\mathbf{x})}_{\text{power violation}}}^{\text{jointly optimize computation and communication}}.$$

OMUAA: Mobile Device n's Algorithm

1: Update local model $\mathbf{x}_t^n$ by solving

$$\mathbf{P2}^n: \quad \min_{\mathbf{x}\in\mathcal{X}} \quad \overbrace{\underbrace{\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x} - \hat{\mathbf{x}}_{t-1}\rangle + \frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2}_{\text{training loss}} + \underbrace{\gamma Q_{t-1}^n g_t^n(\mathbf{x})}_{\text{power violation}}}^{\text{jointly optimize computation and communication}}.$$

2: Update local virtual queue

$$Q_t^n = \max\{Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), 0\}.$$

OMUAA: Mobile Device n's Algorithm

1: Update local model $\mathbf{x}_t^n$ by solving

$$\mathbf{P2}^n: \quad \min_{\mathbf{x}\in\mathcal{X}} \quad \overbrace{\underbrace{\langle \nabla f^n(\hat{\mathbf{x}}_{t-1}), \mathbf{x} - \hat{\mathbf{x}}_{t-1} \rangle + \frac{1}{2\alpha}\|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2}_{\text{training loss}} + \underbrace{\gamma Q_{t-1}^n g_t^n(\mathbf{x})}_{\text{power violation}}}^{\text{jointly optimize computation and communication}}.$$

2: Update local virtual queue

$$Q_t^n = \max\{Q_{t-1}^n + g_t^n(\mathbf{x}_t^n), 0\}.$$

3: Transmit signals $\mathbf{s}_t^n = \frac{1}{\lambda_t} w^n \mathbf{b}_t^n \circ \mathbf{x}_t$ to the edge server.

OMUAA: Edge Server's Algorithm

Receive signals $\mathbf{y}_t$ over the air as

$$\mathbf{y}_t = \sum_{n=1}^{N} \mathbf{h}_t^n \circ \mathbf{s}_t^n + \mathbf{z}_t = \frac{1}{\lambda_t} \sum_{n=1}^{N} w^n \mathbf{x}_t^n + \mathbf{z}_t.$$

OMUAA: Edge Server's Algorithm

1: Receive signals $\mathbf{y}_t$ over the air as

$$\mathbf{y}_t = \sum_{n=1}^{N} \mathbf{h}_t^n \circ \mathbf{s}_t^n + \mathbf{z}_t = \frac{1}{\lambda_t} \sum_{n=1}^{N} w^n \mathbf{x}_t^n + \mathbf{z}_t.$$

2: Update noisy global model

$$\hat{\mathbf{x}}_t = \lambda_t \Re\{\mathbf{y}_t\} = \mathbf{x}_t + \lambda_t \Re\{\mathbf{z}_t\}.$$

OMUAA: Edge Server's Algorithm

1: Receive signals $\mathbf{y}_t$ over the air as

$$\mathbf{y}_t = \sum_{n=1}^{N} \mathbf{h}_t^n \circ \mathbf{s}_t^n + \mathbf{z}_t = \frac{1}{\lambda_t} \sum_{n=1}^{N} w^n \mathbf{x}_t^n + \mathbf{z}_t.$$

2: Update noisy global model

$$\hat{\mathbf{x}}_t = \lambda_t \Re\{\mathbf{y}_t\} = \mathbf{x}_t + \lambda_t \Re\{\mathbf{z}_t\}.$$

3: Broadcast $\hat{\mathbf{x}}_t$ to all mobile devices.

Online Model Solution

- Closed-form local model solution

$$\mathbf{x}_t^n = \left[\underbrace{\left(\mathbf{1} + \alpha \boldsymbol{\theta}_t^n\right)^{-1}}_{\text{scaling}} \circ \left(\underbrace{\hat{\mathbf{x}}_{t-1} - \alpha \nabla f^n(\hat{\mathbf{x}}_{t-1})}_{\text{local gradient descent}} \right) \right]_{-\mathbf{x}_{\max}}^{\mathbf{x}_{\max}}$$

where the i-th entry of $\boldsymbol{\theta}_t^n$ is

$$\theta_t^{n,i} = \frac{2\gamma Q_{t-1}^n (w^n)^2}{\lambda_t^2 |h_t^{n,i}|^2}.$$

Online Model Solution

- Closed-form local model solution

$$\mathbf{x}_t^n = \Big[\underbrace{(\mathbf{1} + \alpha \boldsymbol{\theta}_t^n)^{-1}}_{\text{scaling}} \circ \big(\underbrace{\hat{\mathbf{x}}_{t-1} - \alpha \nabla f^n(\hat{\mathbf{x}}_{t-1})}_{\text{local gradient descent}} \big) \Big]_{-\mathbf{x}_{\max}}^{\mathbf{x}_{\max}}$$

where the i-th entry of $\boldsymbol{\theta}_t^n$ is

$$\theta_t^{n,i} = \frac{2\gamma Q_{t-1}^n (w^n)^2}{\lambda_t^2 |h_t^{n,i}|^2}.$$

- Channel-aware: close to error-free when channel power $|h_t^{n,i}|^2$ is large.

Online Model Solution

- Closed-form local model solution

$$\mathbf{x}_t^n = \Big[\underbrace{(\mathbf{1} + \alpha \boldsymbol{\theta}_t^n)^{-1}}_{\text{scaling}} \circ \big(\underbrace{\hat{\mathbf{x}}_{t-1} - \alpha \nabla f^n(\hat{\mathbf{x}}_{t-1})}_{\text{local gradient descent}} \big) \Big]_{-\mathbf{x}_{\max}}^{\mathbf{x}_{\max}}$$

where the i-th entry of $\boldsymbol{\theta}_t^n$ is

$$\theta_t^{n,i} = \frac{2\gamma Q_{t-1}^n (w^n)^2}{\lambda_t^2 |h_t^{n,i}|^2}.$$

- Channel-aware: close to error-free when channel power $|h_t^{n,i}|^2$ is large.
- Power-aware: less power when violation measured by Q_{t-1}^n is large.

Performance Bounds

- Benchmark: optimal global solution $\{\mathbf{x}_t^\star\}$ to **P1** over noiseless channel.

Performance Bounds

- Benchmark: optimal global solution $\{\mathbf{x}_t^\star\}$ to **P1** over noiseless channel.

Theorem

For i.i.d. $\{\mathbf{h}_t\}$, given any $\epsilon > 0$, set $\alpha = \gamma = \epsilon$ and $\lambda_t = \epsilon^2, \forall t$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{f(\hat{\mathbf{x}}_t)\} \leq f^\star + \mathcal{O}\big((1+\rho^2+\Pi_T\rho)\epsilon\big), \quad \forall T \geq \frac{1}{\epsilon^2},$$

$$\frac{1}{T}\sum_{t=1}^{T} g_t^n(\mathbf{x}_t^n) \leq \mathcal{O}\big((1+\rho^2)\epsilon\big), \quad \forall n, \quad \forall T \geq \frac{1}{\epsilon^3},$$

where $\Pi_T \triangleq \sum_{t=1}^{T}\mathbb{E}\{\|\mathbf{x}_t^\star - \mathbf{x}_{t+1}^\star\|\}$ and $\|\Re\{\mathbf{z}_t\}\| \leq \rho, \forall t$.

Application to Image Classification Problem

- MNIST dataset (hand written digits 0-9)
  - $|\mathcal{D}| = 6 \times 10^4$ training and $|\mathcal{E}| = 1 \times 10^4$ testing data samples.
  - **u** represents an image of 784 pixel with label $v \in \{1, \ldots, 10\}$.

Application to Image Classification Problem

- MNIST dataset (hand written digits 0-9)
  - $|\mathcal{D}| = 6 \times 10^4$ training and $|\mathcal{E}| = 1 \times 10^4$ testing data samples.
  - $\mathbf{u}$ represents an image of 784 pixel with label $v \in \{1, \ldots, 10\}$.
- Cross-entropy training loss for multinomial logistic regression

$$l(\mathbf{x}; \mathbf{u}, v) = -\sum_{j=1}^{10} 1\{v = j\} \log \frac{\exp(\langle \mathbf{x}[j], \mathbf{u} \rangle)}{\sum_{k=1}^{10} \exp(\langle \mathbf{x}[k], \mathbf{u} \rangle)}$$

where $\mathbf{x} = [\mathbf{x}[1]^T, \ldots, \mathbf{x}[10]^T]^T$ is of size 7840.

System Setting

- Computation system
  - Non-i.i.d. data among devices.
  - Batch dataset with $|\mathcal{B}_t^n| = 20$ data samples each iteration.

System Setting

- Computation system
  - Non-i.i.d. data among devices.
  - Batch dataset with $|\mathcal{B}_t^n| = 20$ data samples each iteration.
- Communication system
  - $N = 10$ devices with 100 m to the edge server.
  - 500 subchannels over $\lceil \frac{7840}{500} \rceil = 16$ transmission frames.
  - Each subchannel is of bandwidth 15 kHz.

Performance Metrics & Benchmarks

- Performance metrics
  - Time-averaged test accuracy over $\mathcal{E}$

$$\bar{A}(T) = \frac{1}{T|\mathcal{E}|} \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{E}|} 1\left\{ \arg\max_j \left\{ \frac{\exp(\langle \hat{\mathbf{x}}_t[j], \mathbf{u}^i \rangle)}{\sum_{k=1}^{J} \exp(\langle \hat{\mathbf{x}}_t[k], \mathbf{u}^i \rangle)} \right\} = v^i \right\}.$$

  - Time-averaged training loss over $\{\mathcal{B}_t^n\}$

$$\bar{f}(T) = \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} \frac{1}{|\mathcal{B}_t^n|} \sum_{i=1}^{|\mathcal{B}_t^n|} w^n l(\hat{\mathbf{x}}_t; \mathbf{u}_t^{n,i}, v_t^{n,i}).$$

Performance Metrics & Benchmarks

- Performance metrics
  - Time-averaged test accuracy over $\mathcal{E}$

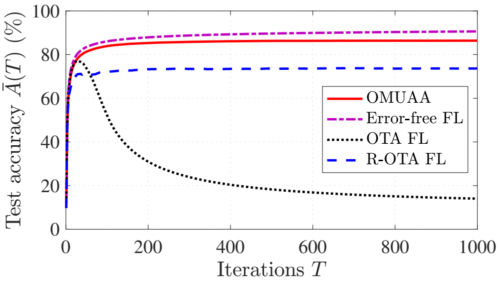
$$\bar{A}(T) = \frac{1}{T|\mathcal{E}|}\sum_{t=1}^{T}\sum_{i=1}^{|\mathcal{E}|} 1\left\{ \arg\max_j \left\{ \frac{\exp(\langle \hat{\mathbf{x}}_t[j], \mathbf{u}^i \rangle)}{\sum_{k=1}^{J} \exp(\langle \hat{\mathbf{x}}_t[k], \mathbf{u}^i \rangle)} \right\} = v^i \right\}.$$

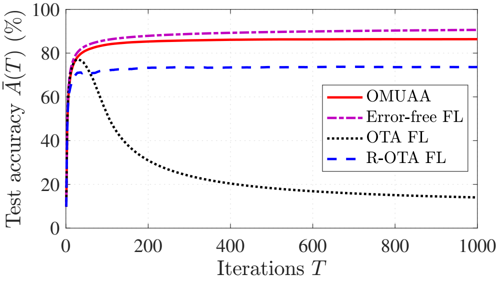
  - Time-averaged training loss over $\{\mathcal{B}_t^n\}$

$$\bar{f}(T) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{1}{|\mathcal{B}_t^n|}\sum_{i=1}^{|\mathcal{B}_t^n|} w^n l(\hat{\mathbf{x}}_t; \mathbf{u}_t^{n,i}, v_t^{n,i}).$$

- Performance benchmarks
  - Error-free FL: performance upper bound for OMUAA.
  - OTA FL: current best alternatives with long-term power constraints.
  - R-OTA FL: additional regularization $\kappa\|\mathbf{x}\|^2$ for OTA-FL.

Performance Comparison





- Error-free FL:

FL over noiseless channels.

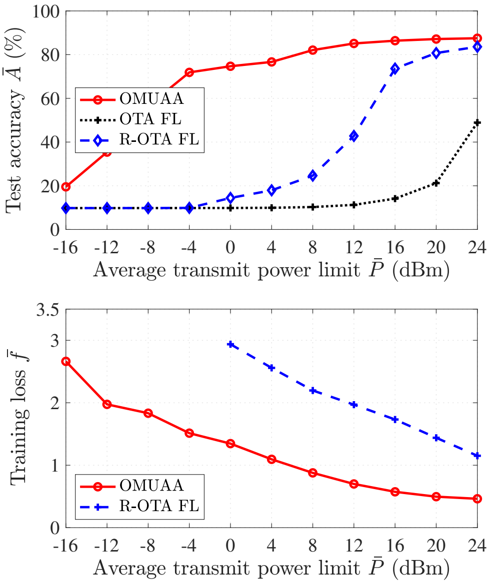
- OTA FL:

current best alternatives.

- R-OTA FL:

regularized current best.

Impact of Average Transmit Power Limit



- Error-free FL:

FL over noiseless channels.

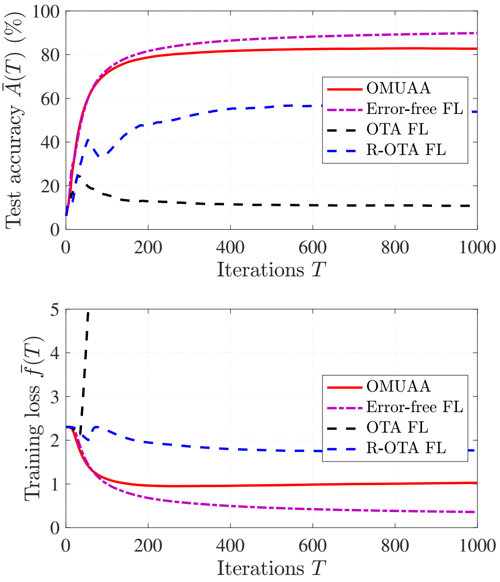
- OTA FL:

current best alternatives.

- R-OTA FL:

regularized current best.

Application to Neural Network



- Error-free FL:

FL over noiseless channels.

- OTA FL:

current best alternatives.

- R-OTA FL:

regularized current best.

Conclusions

- FL at the edge over noisy wireless channels
  - Joint online optimization of model training and analog aggregation.

Conclusions

- FL at the edge over noisy wireless channels
  - Joint online optimization of model training and analog aggregation.
  - Minimize accumulated training loss subject to long-term power constraints.

Conclusions

- FL at the edge over noisy wireless channels
  - Joint online optimization of model training and analog aggregation.
  - Minimize accumulated training loss subject to long-term power constraints.
- OMUAA algorithm
  - Integration of FL, OTA computation, and power allocation.

Conclusions

- FL at the edge over noisy wireless channels
  - Joint online optimization of model training and analog aggregation.
  - Minimize accumulated training loss subject to long-term power constraints.
- OMUAA algorithm
  - Integration of FL, OTA computation, and power allocation.
  - Both channel- and power-aware closed-form online solution.

Conclusions

- FL at the edge over noisy wireless channels
  - Joint online optimization of model training and analog aggregation.
  - Minimize accumulated training loss subject to long-term power constraints.
- OMUAA algorithm
  - Integration of FL, OTA computation, and power allocation.
  - Both channel- and power-aware closed-form online solution.
  - Performance bounds for both computation and communication metrics.

Conclusions

- FL at the edge over noisy wireless channels
  - Joint online optimization of model training and analog aggregation.
  - Minimize accumulated training loss subject to long-term power constraints.
- OMUAA algorithm
  - Integration of FL, OTA computation, and power allocation.
  - Both channel- and power-aware closed-form online solution.
  - Performance bounds for both computation and communication metrics.
- Simulation results
  - Substantial performance gain over the current best alternatives.