FLAGSHIP UNIVERSITY OF OULU

CWC cwc.oulu.fi

CENTRE FOR WIRELESS COMMUNICATIONS
University of Oulu

ICASSP2020 Barcelona

UNIVERSITY OF OULU

Q-GADMM: Quantized Group ADMM for Communication Efficient Decentralized Machine Learning

Presented by: Anis Elgabli

Joint work with:

Jihong Park (Deakin University, Australia)

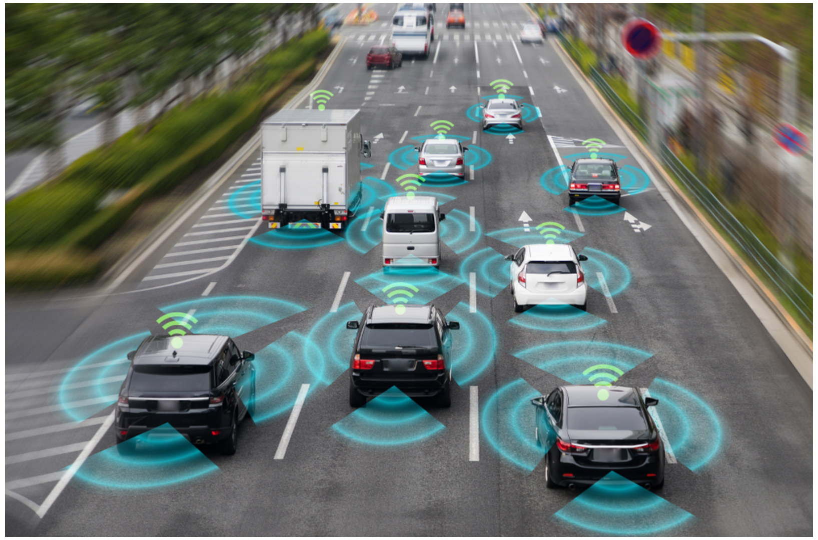
Chaouki Ben Issaid (CWC-OULU, Finland)

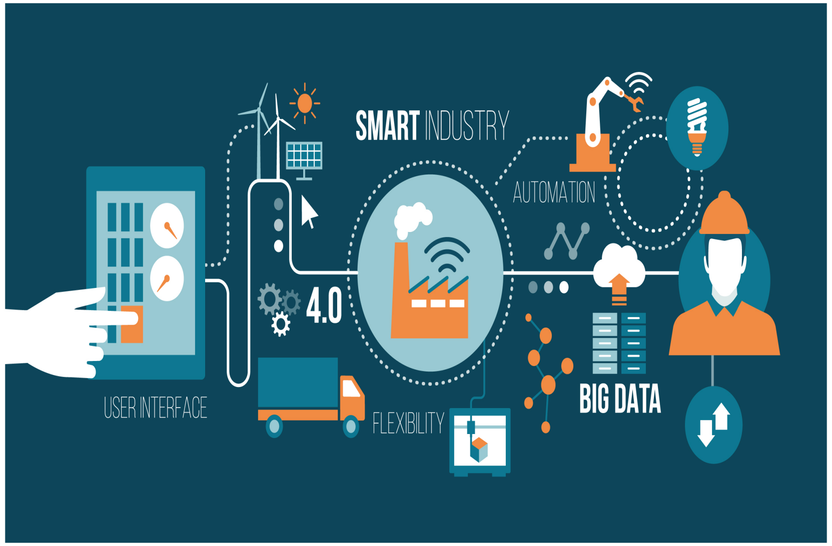
Mehdi Bennis (CWC-OULU, Finland)

Vaneet Aggarwal (Purdue University, IN, USA)

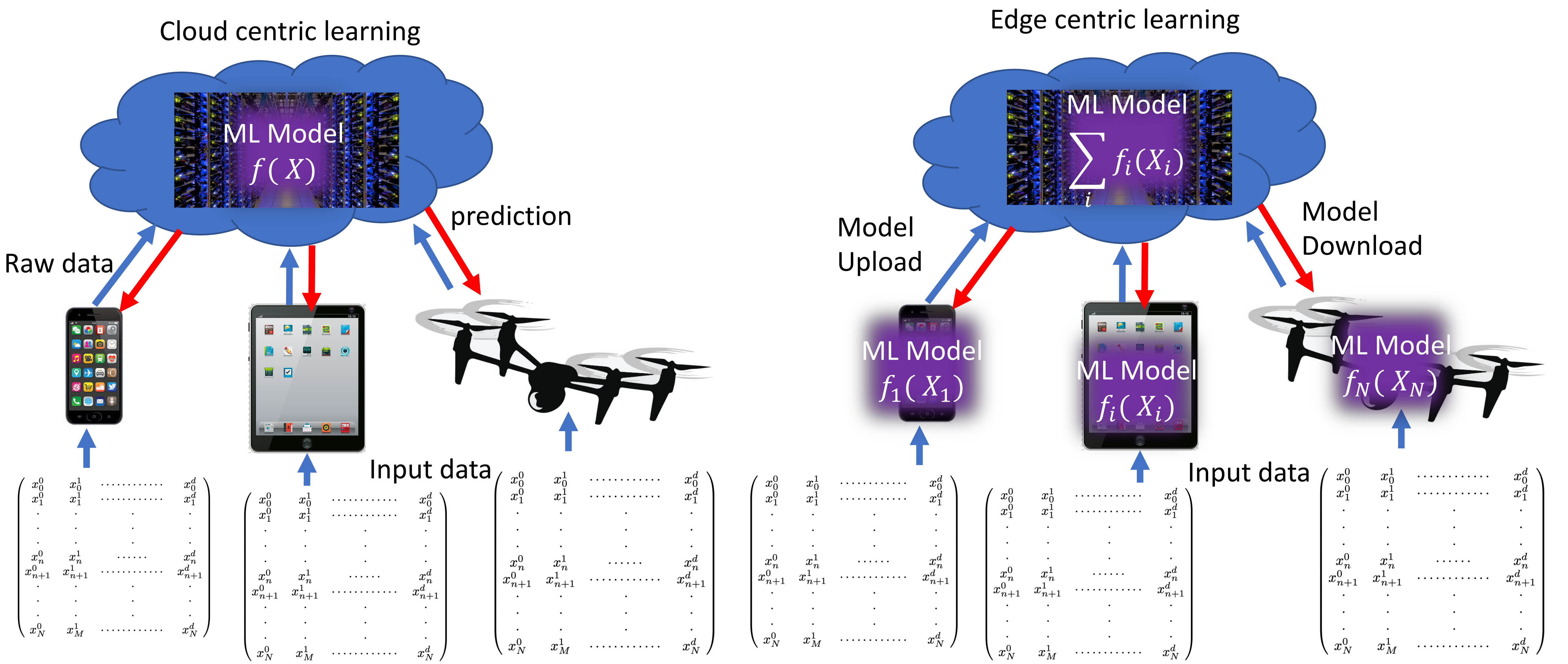
Amrit Bedi (IITK, India)

IoT: Huge amount of data is generated every where.





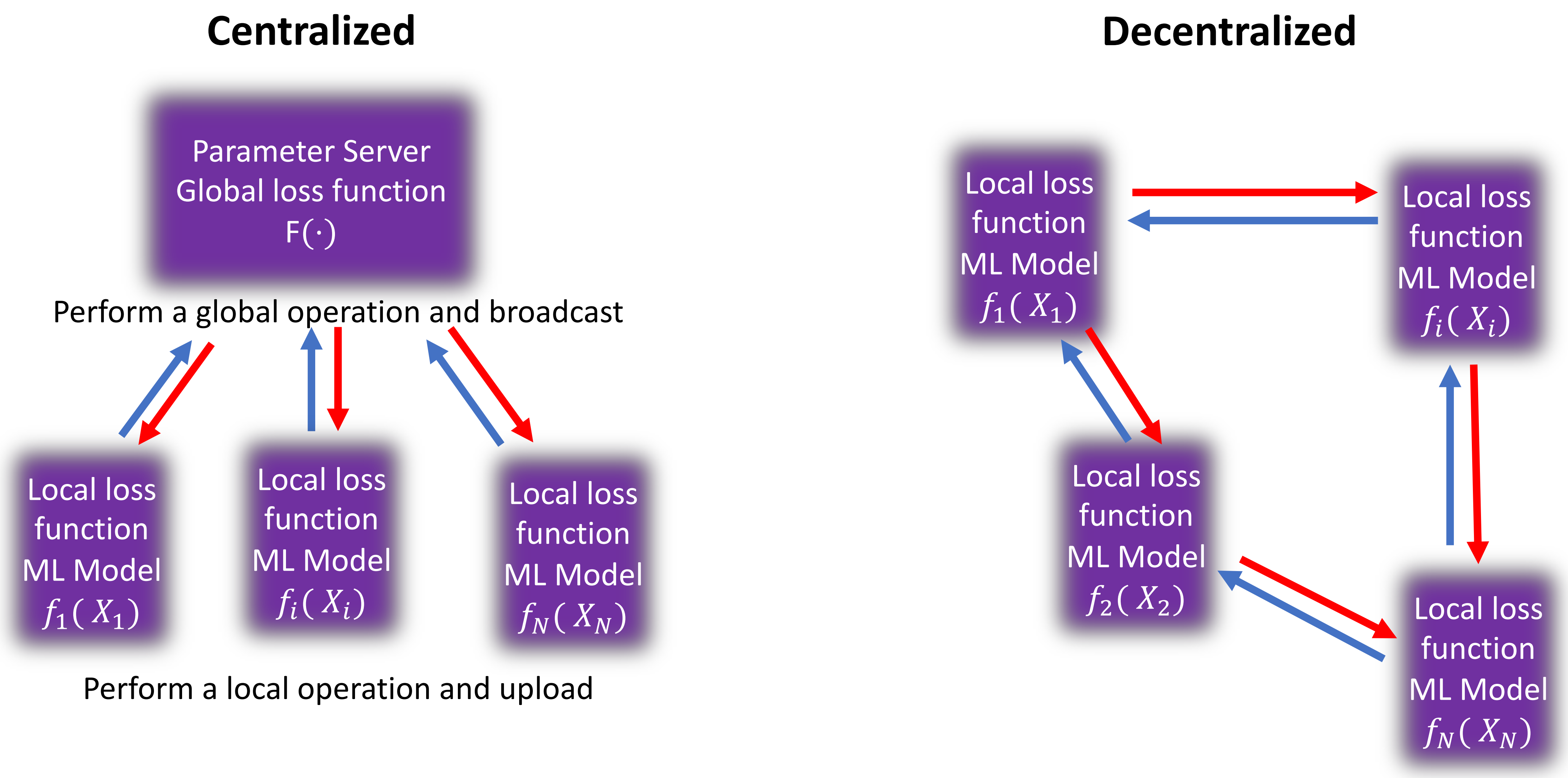
IoT: ML computations need to move to the Edge.



High Comm cost: transferring raw data, Limited comm resources
High Latency: send data and receive decision
Privacy issue: Data leaves its origin

Low Communication cost: transferring model parameters
Low latency: infer locally and respond
Privacy: No raw data exchange

Centralized vs. Decentralized Distributed Machine Learning



Communication-Efficient Distributed Machine Learning

- **Distributed ML: Every worker exchanges minimum number of variables per training iteration.**
  e.g. Exchange gradient vector, model parameter vector, model output.

- **Temporal-Sparsity: Minimum number of workers transmit per iteration**
  e.g. LAG (Lazily Aggregated Gradient) (Tianyi, Neurips 2018)

- **Model quantization: Minimum payload size per channel use**
  e.g. LAQ (Sun, Neurips 2019)

- **Spatial-Sparsity: Minimum number of neighbors to communicate with (Decentralization) and fast convergence.**
  e.g. GADMM (Group ADMM) (Anis, JMLR)

- **Contribution: Minimum number of neighbors plus minimum payload per channel use by utilizing model Quantization.**
  . Q-GADMM (Quantized Group ADMM)
  - Stochastic quantization, decreasing step size, Inherit convergence guarantees of GADMM for convex loss functions

Standard ADMM: The Parameter server (PS) based Learning problem:

Constrained formulation:

$$\min_{\boldsymbol{\Theta},\{\boldsymbol{\theta}_n\}_{n=1}^{N}} \quad \sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n)$$

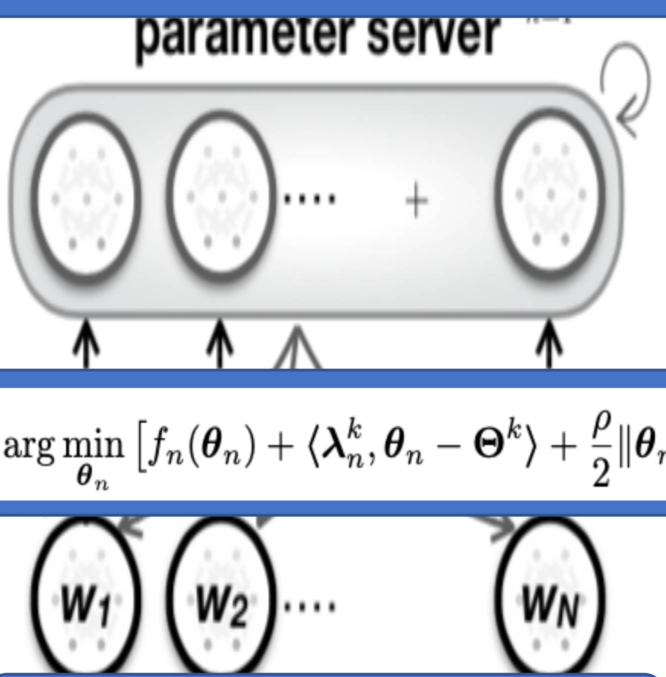
$$\text{s.t.} \quad \boldsymbol{\theta}_n = \boldsymbol{\Theta}, \quad \forall\, n.$$

The Augmented laggrangian:

$$\mathcal{L}_\rho(\boldsymbol{\Theta}, \{\boldsymbol{\theta}_n\}_{n=1}^{N}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n) + \sum_{n=1}^{N} \langle \boldsymbol{\lambda}_n, \boldsymbol{\theta}_n - \boldsymbol{\Theta} \rangle + \frac{\rho}{2} \sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\Theta}\|^2$$

ADMM updating steps of the primal and dual variables at iteration k+1

$$\boldsymbol{\Theta}^{k+1} = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{\theta}_n^{k+1} + \frac{1}{\rho} \boldsymbol{\lambda}_n^{k} \right)$$



$$\boldsymbol{\theta}_n^{k+1} = \arg\min_{\boldsymbol{\theta}_n} \left[f_n(\boldsymbol{\theta}_n) + \langle \boldsymbol{\lambda}_n^{k}, \boldsymbol{\theta}_n - \boldsymbol{\Theta}^{k} \rangle + \frac{\rho}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\Theta}^{k}\|^2 \right]$$

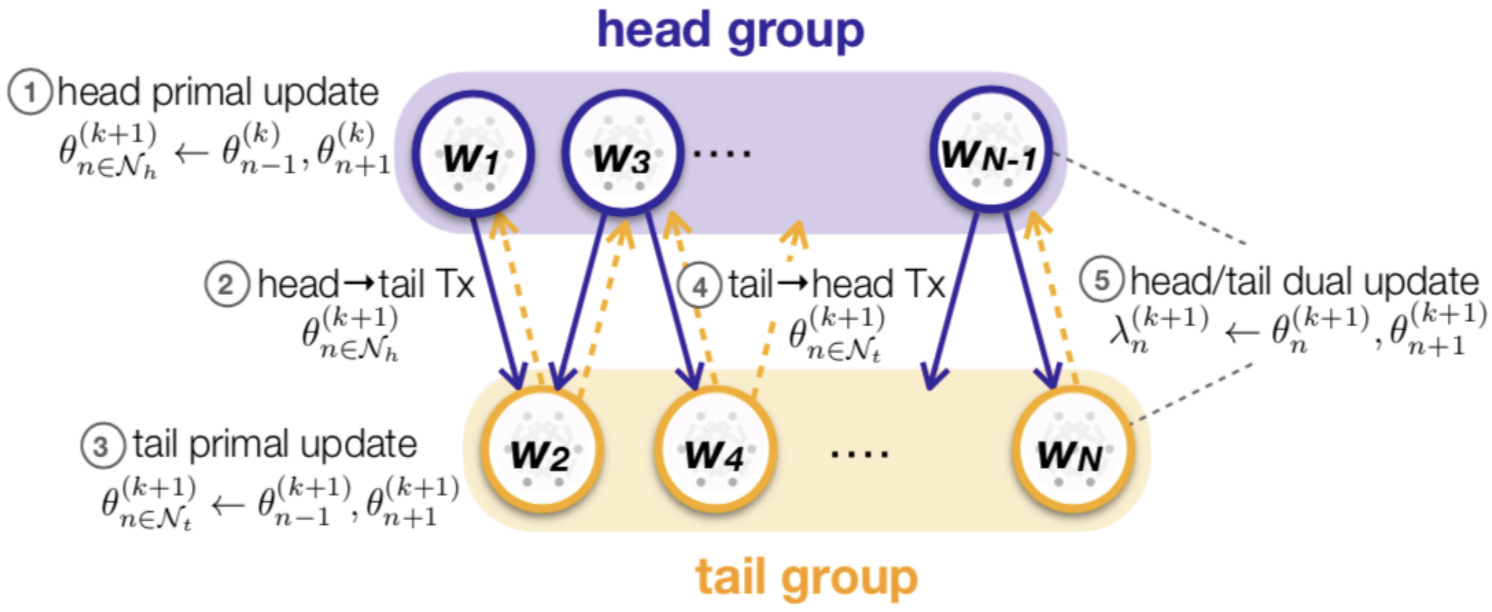
$$\boldsymbol{\lambda}_n^{k+1} = \boldsymbol{\lambda}_n^{k} + \rho(\boldsymbol{\theta}_n^{k+1} - \boldsymbol{\Theta}^{k+1})$$

Decentralized-based Learning problem: GADMM

Constrained formulation:

$$\boldsymbol{\theta}^\star := \arg \min_{\{\boldsymbol{\theta}_n\}_{n=1}^N} \sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n)$$

$$\text{s.t. } \boldsymbol{\theta}_n = \boldsymbol{\theta}_{n+1}, \quad n = 1, \cdots, N-1.$$



Group ADMM (GADMM)

- **Divide workers into two groups**
- **Workers belong to the same group update variables in parallel**
- **The two groups alternate in their updates**
- **Each worker communicates with at most two neighbors**

GADMM

Augmented Lagrangian:

$$\mathcal{L}_\rho(\{\boldsymbol{\theta}_n\}_{n=1}^N, \boldsymbol{\lambda}) = \sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n) + \sum_{n=1}^{N-1} \langle \boldsymbol{\lambda}_n, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1} \rangle + \frac{\rho}{2} \sum_{n=1}^{N-1} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2$$

Primal variables of head workers:

$$\boldsymbol{\theta}_n^{k+1} = \arg\min_{\boldsymbol{\theta}_n} \Big[f_n(\boldsymbol{\theta}_n) + \langle \boldsymbol{\lambda}_{n-1}^k, \boldsymbol{\theta}_{n-1}^k - \boldsymbol{\theta}_n \rangle + \langle \boldsymbol{\lambda}_n^k, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^k \rangle + \frac{\rho}{2} \|\boldsymbol{\theta}_{n-1}^k - \boldsymbol{\theta}_n\|^2 + \frac{\rho}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^k\|^2 \Big], n \in \mathcal{N}_h \setminus \{1\}$$

$$\boldsymbol{\theta}_n^{k+1} = \arg\min_{\boldsymbol{\theta}_n} \Big[f_n(\boldsymbol{\theta}_n) + \langle \boldsymbol{\lambda}_n^k, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^k \rangle + \frac{\rho}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^k\|^2 \Big], n = 1$$

Primal variables of tail workers:

$$\boldsymbol{\theta}_n^{k+1} = \arg\min_{\boldsymbol{\theta}_n} \Big[f_n(\boldsymbol{\theta}_n) + \langle \boldsymbol{\lambda}_{n-1}^k, \boldsymbol{\theta}_{n-1}^{k+1} - \boldsymbol{\theta}_n \rangle + \langle \boldsymbol{\lambda}_n^k, \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^{k+1} \rangle + \frac{\rho}{2} \|\boldsymbol{\theta}_{n-1}^{k+1} - \boldsymbol{\theta}_n\|^2 + \frac{\rho}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^{k+1}\|^2 \Big], n \in \mathcal{N}_t \setminus \{N\}.$$

$$\boldsymbol{\theta}_n^{k+1} = \arg\min_{\boldsymbol{\theta}_n} \Big[f_n(\boldsymbol{\theta}_n) + \langle \boldsymbol{\lambda}_{n-1}^k, \boldsymbol{\theta}_{n-1}^{k+1} - \boldsymbol{\theta}_n \rangle + \frac{\rho}{2} \|\boldsymbol{\theta}_{n-1}^{k+1} - \boldsymbol{\theta}_n\|^2 \Big], n = N.$$

Dual variables (updated locally, zero communication overhead):

$$\boldsymbol{\lambda}_n^{k+1} = \boldsymbol{\lambda}_n^k + \rho(\boldsymbol{\theta}_n^{k+1} - \boldsymbol{\theta}_{n+1}^{k+1}), n = \{1, \cdots, N-1\}.$$

Optimality of GADMM for convex loss functions

Necessary and sufficient optimality conditions (Boyd et al., 2011):

$$\boldsymbol{\theta}_n^\star = \boldsymbol{\theta}_{n-1}^\star, n \in \{2, \cdots, N\} \qquad \text{(primal feasibility)}$$

$$\mathbf{0} \in \partial f_n(\boldsymbol{\theta}_n^\star) - \boldsymbol{\lambda}_{n-1}^\star + \boldsymbol{\lambda}_n^\star, n \in \{2, \cdots, N-1\}$$

$$\mathbf{0} \in \partial f_n(\boldsymbol{\theta}_n^\star) + \boldsymbol{\lambda}_n^\star, n = 1 \qquad \text{(dual feasibility)}$$

$$\mathbf{0} \in \partial f_n(\boldsymbol{\theta}_n^\star) + \boldsymbol{\lambda}_{n-1}^\star, n = N$$

Primal residual

$$\mathbf{r}_{n,n+1}^{k+1} = \boldsymbol{\theta}_n^{k+1} - \boldsymbol{\theta}_{n+1}^{k+1}$$

Dual residual

$$\mathbf{s}_n^{k+1} = \begin{cases} \rho(\boldsymbol{\theta}_{n-1}^{k+1} - \boldsymbol{\theta}_{n-1}^{k}) + \rho(\boldsymbol{\theta}_{n+1}^{k+1} - \boldsymbol{\theta}_{n+1}^{k}), & \text{for } n \in \mathcal{N}_h \setminus \{1\} \\ \rho(\boldsymbol{\theta}_{n+1}^{k+1} - \boldsymbol{\theta}_{n+1}^{k}), & \text{for } n = 1. \end{cases}$$

Optimality of GADMM for convex loss functions

Lemma 1: *For the iterates $\boldsymbol{\theta}_n^{k+1}$ generated by Algorithm 1, we have*

(i) Upper bound on the optimality gap

$$\sum_{n=1}^{N}[f_n(\boldsymbol{\theta}_n^{k+1}) - f_n(\boldsymbol{\theta}^\star)] \leq -\sum_{n=1}^{N-1}\langle \boldsymbol{\lambda}_n^{k+1}, \mathbf{r}_{n,n+1}^{k+1}\rangle + \sum_{n\in\mathcal{N}_h}\langle \mathbf{s}_n^{k+1}, \boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_n^{k+1}\rangle.$$

(ii) Lower bound on the optimality gap

$$\sum_{n=1}^{N}[f_n(\boldsymbol{\theta}_n^{k+1}) - f_n(\boldsymbol{\theta}^\star)] \geq -\sum_{n=1}^{N-1}\langle \boldsymbol{\lambda}_n^\star, \mathbf{r}_{n,n+1}^{k+1}\rangle.$$

Optimality of GADMM for convex loss functions

Theorem 1: *When $f_n(\boldsymbol{\theta}_n)$ is closed, proper, and convex for all n, and the Lagrangian $\boldsymbol{\mathcal{L}}_0$ has a saddle point, for GADMM iterates, it holds that*

(i) the primal residual converges to zero as $k \to \infty$.i.e.,

$$\lim_{k\to\infty} \mathbf{r}_{n,n+1}^k = \mathbf{0}, n \in \{1, \cdots, N-1\}.$$

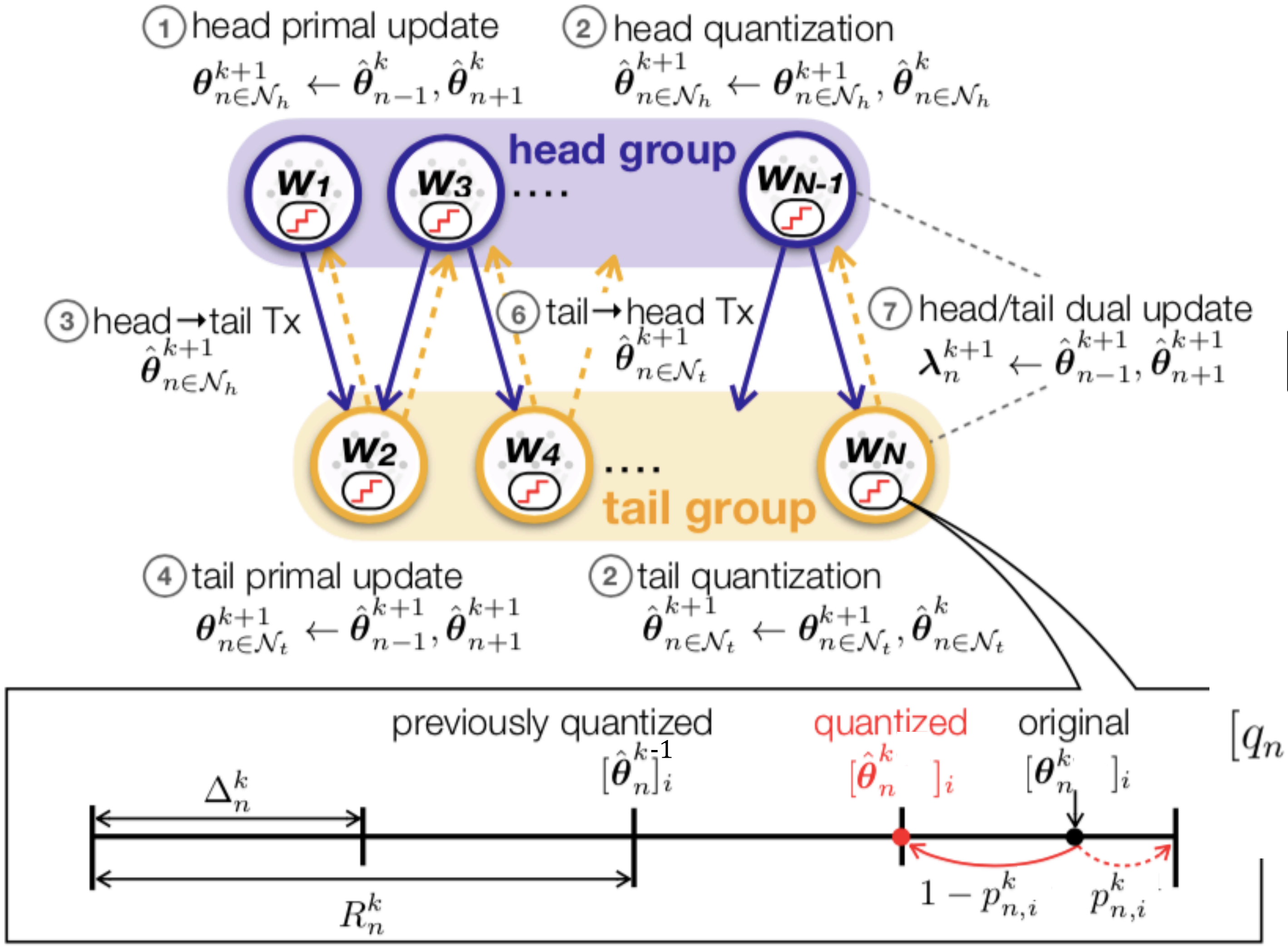
(ii) the dual residual converges to zero as $k \to \infty$.i.e.,

$$\lim_{k\to\infty} \mathbf{s}_n^k = \mathbf{0}, n \in \mathcal{N}_h.$$

(iii) the optimality gap converges to zero as $k \to \infty$.i.e.,

$$\lim_{k\to\infty} \sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n^k) = \sum_{n=1}^{N} f_n(\boldsymbol{\theta}^\star).$$

Q-GADMM: Quantized Group ADMM



$$R_n^k = \| \theta_n^k - \theta_n^{k-1} \|_\infty$$

$$\Delta_n^k = 2R_n^k/(2^{b_n^k} - 1)$$

$$[c_n(\boldsymbol{\theta}_n^k)]_i = \frac{1}{\Delta_n^k}\left([\boldsymbol{\theta}_n^k]_i - [\hat{\boldsymbol{\theta}}_n^{k-1}]_i + R_n^k\right)$$

Stochastic quantization

$$[q_n(\boldsymbol{\theta}_n^k)]_i = \begin{cases} \lceil [c_n(\boldsymbol{\theta}_n^k)]_i \rceil & \text{with probability } p_{n,i}^k \\ \lfloor [c_n(\boldsymbol{\theta}_n^k)]_i \rfloor & \text{with probability } 1 - p_{n,i}^k \end{cases}$$

Q-GADMM: Stochastic Quantization

To maintain unbiased error with zero mean:

$$p_{n,i}^k = \left(\lceil [c_n(\boldsymbol{\theta}_n^k)]_i \rceil - [c_n(\boldsymbol{\theta}_n^k)]_i \right) / \Delta_n^k$$

To guarantee convergence to the optimal solution as k → ∞, the following inequality should be satisfied:

$$\Delta_n^k \leq \Delta_n^{k-1} \quad \Longrightarrow \quad b_n^k \geq \left\lceil \log_2 \left(1 + (2^{b_n^{k-1}} - 1) R_n^k / R_n^{k-1} \right) \right\rceil$$

To transmit R_n^k, b_n^k and $q_n(\theta_n^k)$ using $\boldsymbol{b_n^k d} + \boldsymbol{b_R} + \boldsymbol{b_b}$ **bits**instead of **32d bits**

$b_R$and $b_b$needs at most **32x2 bits** to be transmitted

The quantized model is retrieved at the neighboring workers as follows:

$$\hat{\boldsymbol{\theta}}_n^k = \hat{\boldsymbol{\theta}}_n^{k-1} + \Delta_n^k q_n(\boldsymbol{\theta}_n^k) - R_n^k \mathbf{1}$$

Q-GADMM

Augmented Lagrangian:

$$\mathcal{L}_\rho = \sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n) + \sum_{n=1}^{N-1} \langle \boldsymbol{\lambda}_n, \boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_{n+1} \rangle + \frac{\rho}{2} \sum_{n=1}^{N-1} \|\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_{n+1}\|_2^2$$

The primal variables of the head workers:

$$\boldsymbol{\theta}_{n\in\mathcal{N}_h}^{k+1} = \underset{\theta_n}{\operatorname{argmin}} \{ f_n(\boldsymbol{\theta}_n) + \langle \boldsymbol{\lambda}_{n-1}^k, \hat{\boldsymbol{\theta}}_{n-1}^k - \boldsymbol{\theta}_n \rangle + \langle \boldsymbol{\lambda}_n^k, \boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_{n+1}^k \rangle$$
$$+ \frac{\rho}{2} \|\hat{\boldsymbol{\theta}}_{n-1}^k - \boldsymbol{\theta}_n\|_2^2 + \frac{\rho}{2} \|\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_{n+1}^k\|_2^2 \}, n \in \mathcal{N}_h \setminus \{1\}.$$

The primal variables of the tail workers:

$$\boldsymbol{\theta}_{n\in\mathcal{N}_t}^{k+1} = \underset{\boldsymbol{\theta}_n}{\operatorname{argmin}} \{ f_n(\boldsymbol{\theta}_n) + \langle \boldsymbol{\lambda}_{n-1}^k, \hat{\boldsymbol{\theta}}_{n-1}^{k+1} - \boldsymbol{\theta}_n \rangle + \langle \boldsymbol{\lambda}_n^k, \boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_{n+1}^{k+1} \rangle$$
$$+ \frac{\rho}{2} \|\hat{\boldsymbol{\theta}}_{n-1}^{k+1} - \boldsymbol{\theta}_n\|_2^2 + \frac{\rho}{2} \|\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_{n+1}^{k+1}\|_2^2 \}, n \in \mathcal{N}_h \setminus \{N\}.$$

The dual variables:

$$\boldsymbol{\lambda}_n^{k+1} = \boldsymbol{\lambda}_n^k + \rho(\hat{\boldsymbol{\theta}}_n^{k+1} - \hat{\boldsymbol{\theta}}_{n+1}^{k+1}),\ n = 1, \ldots, N-1$$

Q-GADMM: Convergence Analysis

Primal and dual visibility:

$$\boldsymbol{\theta}_n^* = \boldsymbol{\theta}_{n-1}^*, \forall n > 1, \quad \mathbf{0} \in \partial f_n(\boldsymbol{\theta}_n^*) - \boldsymbol{\lambda}_{n-1}^* + \boldsymbol{\lambda}_n^*$$

Primal residual

$$\mathbf{r}_{n,n+1}^{k+1} = \boldsymbol{\theta}_n^{k+1} - \boldsymbol{\theta}_{n+1}^{k+1}$$

Dual residual

$$\boldsymbol{s}_n^{k+1} = \begin{cases} \rho(\hat{\boldsymbol{\theta}}_{n-1}^{k+1} - \hat{\boldsymbol{\theta}}_{n-1}^{k}) + \rho(\hat{\boldsymbol{\theta}}_{n+1}^{k+1} - \hat{\boldsymbol{\theta}}_{n+1}^{k}), & \text{if } n \in \mathcal{N}_h \setminus \{1\} \\ \rho(\hat{\boldsymbol{\theta}}_{n+1}^{k+1} - \hat{\boldsymbol{\theta}}_{n+1}^{k}), & \text{if } n = 1. \end{cases}$$

Q-GADMM: Convergence Analysis

Lemma 2: At $k+1$ iteration of Q-GADMM, the optimality gap satisfies $\mathsf{LB}_1 \leq \mathbb{E}\left[\left\{\sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n^{k+1}) - \sum_{n=1}^{N} f_n(\boldsymbol{\theta}_n^{*})\right\}\right] \leq \mathsf{UB}_1$

$$\mathsf{LB}_1 = \mathbb{E}\left[\sum_{n=1}^{N-1} \langle \boldsymbol{\lambda}_n^{*}, \boldsymbol{r}_{n,n+1}^{k+1} \rangle\right]$$

$$\mathsf{UB}_1 = \mathbb{E}\left[-\sum_{n=1}^{N-1} \langle \boldsymbol{\lambda}_n^{k+1}, \boldsymbol{r}_{n,n+1}^{k+1} \rangle + \sum_{n=1}^{N-1} 2\rho \langle \boldsymbol{\epsilon}_n^{k+1}, \boldsymbol{\theta}_n^{*} - \boldsymbol{\theta}_n^{k+1} \rangle + \sum_{n \in \mathcal{N}_h} \langle \boldsymbol{s}_n^{k+1}, \boldsymbol{\theta}_n^{*} - \boldsymbol{\theta}_n^{k+1} \rangle\right]$$

$$V_k = 1/\rho \sum_{n=1}^{N-1} \left\| \boldsymbol{\lambda}_n^k - \boldsymbol{\lambda}_n^\star \right\|^2 + \rho \sum_{n \in \mathcal{N}_h \setminus \{1\}} \left\| \boldsymbol{\theta}_{n-1}^k - \boldsymbol{\theta}^\star \right\|^2 + \rho \sum_{n \in \mathcal{N}_h} \left\| \boldsymbol{\theta}_{n+1}^k - \boldsymbol{\theta}^\star \right\|^2$$

Lemma 3: When $f_n(\boldsymbol{\theta}_n)$ is closed, proper, and convex, and the Lagrangian $\boldsymbol{\mathcal{L}}_0$ has a saddle point, then the following inequality holds true at the $(k+1)$-th iteration of Q-GADMM:

$$\mathbb{E}\left[V^k - V^{k+1}\right] \geq \rho \sum_n \mathbb{E}\left[\| \boldsymbol{\epsilon}_n^{k+1} + \boldsymbol{\epsilon}_{n+1}^{k+1} \|_2^2\right] - \frac{\rho d}{2} \sum_{n \in \mathcal{N}_t} (\Delta_n^{k+1})^2 + H_v$$

$$H_v = \rho \sum_{n \in \mathcal{N}_h \setminus \{1\}} \mathbb{E}\left[\| (\boldsymbol{\theta}_{n-1}^{k+1} - \boldsymbol{\theta}_{n-1}^k) - (\boldsymbol{\epsilon}_{n-1}^{k+1} - \boldsymbol{\epsilon}_{n-1}^k) - \boldsymbol{r}_{n-1,n}^{k+1} \|_2^2\right] + \rho \sum_{n \in \mathcal{N}_h} \mathbb{E}\left[\| (\boldsymbol{\theta}_{n+1}^{k+1} - \boldsymbol{\theta}_{n+1}^k) - (\boldsymbol{\epsilon}_{n+1}^{k+1} - \boldsymbol{\epsilon}_{n+1}^k) + \boldsymbol{r}_{n,n+1}^{k+1} \|_2^2\right]$$
$$+ 2\rho \left(\mathbb{E}\left[\| \boldsymbol{\epsilon}_1^{k+1} \|_2^2\right] + \mathbb{E}\left[\| \boldsymbol{\epsilon}_N^{k+1} \|_2^2\right]\right) + 2\rho \sum_{n \in \mathcal{N}_t} \mathbb{E}\left[\| \boldsymbol{\epsilon}_n^k \|_2^2\right]$$

Theorem 2: For non-increasing quantization step sizes, *i.e.*, $\Delta_n^k \leq \Delta_n^{k-1}$ $\forall k$, and under the assumption of Lemma 2, as $k \to \infty$, the primal and dual residual converges to 0 with probability 1, *i.e.*, $\lim_{k\to\infty} \boldsymbol{r}_{n,n+1}^k \overset{a.s.}{=} \mathbf{0}$ and $\lim_{k\to\infty} \boldsymbol{s}_n^k \overset{a.s.}{=} \mathbf{0}$. Furthermore, the optimality gap converges to 0 with probability 1, *i.e.*, $\lim_{k\to\infty} \sum_{n=1}^N f_n(\boldsymbol{\theta}_n^k) = \sum_{n=1}^N f_n(\boldsymbol{\theta}^*)$.

Intuitively, when Δ_n^k is non-increasing, the RHS of (23) is always positive. For such positive $\mathbb{E}\left[V^k - V^{k+1}\right]$, iteration $k+1$ is one more step towards the optimal solution. Therefore, following the same proof of theorem 1 for GADMM [17], as $k \to \infty$, Q-GADMM converges to the optimal solution.

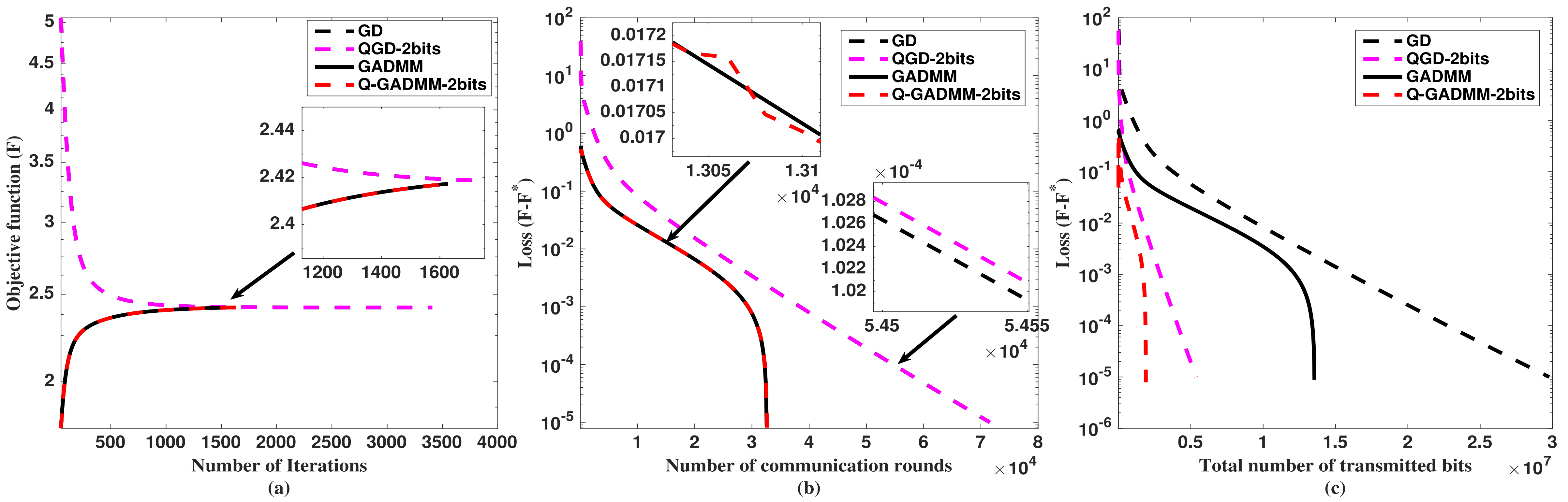
Simulation Setup

- Linear Regression (Convex loss function)
  - QGADMM

  Benchmarks:
    - GADMM
    - GD
    - QGD

- **Real data set:** California housing
- The number of workers, N=10.

Q-GADMM: Numerical Results



- Both GADMM and Q-GADMM significantly outperform the baselines in terms of the convergence speed.
- QGADMM requires the same number of iteration and communication round as GADMM to achieve 10E-5 objective error.
- However, Q-GADMM significantly outperforms GADMM in terms of minimizing the total payload size.

Concluding remarks

- Compared to the original GADMM, Q-GADMM enjoys the same convergence rate, but at significantly lower communication overhead.
- Numerical results in a convex linear regression task corroborate the advantages of Q-GADMM over GADMM, GD, and QGD.

- Future work
  - Q-GADMM over arbitrary and time varying topology.
  - Non-convex and stochastic problems (e.g classification using DNN).

To know more about GADMM:

Elgabli, Anis, et al. "GADMM: Fast and communication efficient framework for distributed machine learning." *arXiv preprint arXiv:1909.00047* (2019).
Elgabli, Anis, et al. "L-FGADMM: Layer-Wise Federated Group ADMM for Communication Efficient Decentralized Deep Learning." *arXiv preprint arXiv:1911.03654* (2019).

Questions

- For questions please do not hesitate to email me:
    - Anis.elgabli@oulu.fi