On-Device Federated Learning via Second-Order Optimization with Over-the-Air Computation

Sheng Hua, Kai Yang, Yuanming Shi

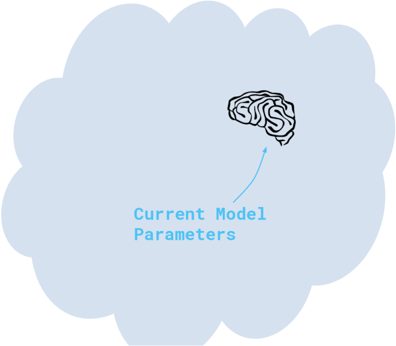
School of Information Science and Technology
ShanghaiTech University

上海科技大学
ShanghaiTech University

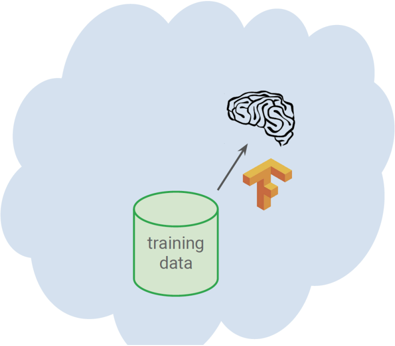
Outline

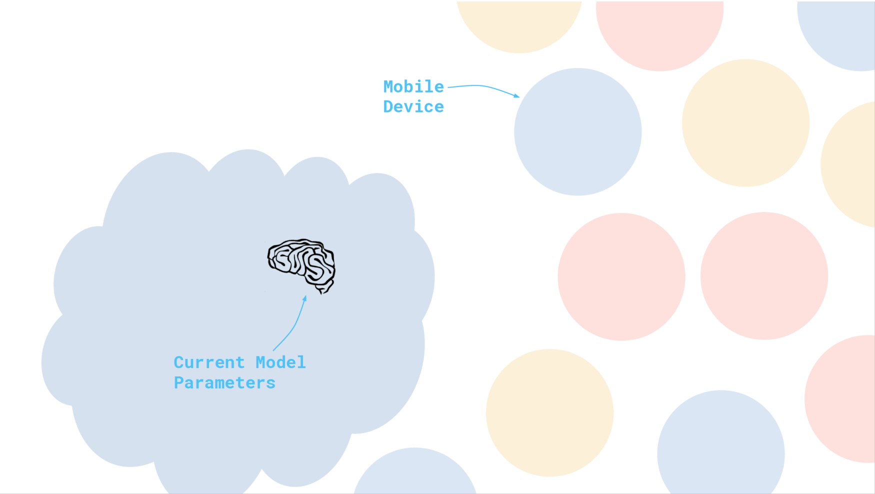
Cloud-Centric Machine Learning

The model lives in the cloud

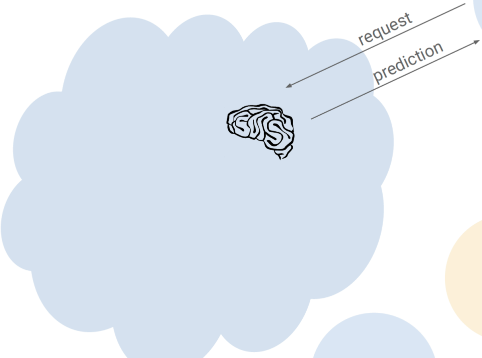


We train models in the cloud

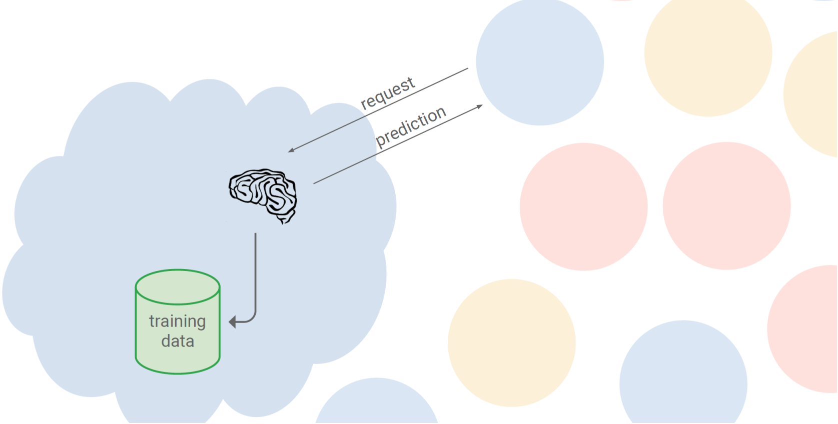



Mobile
Device
Current Model
Parameters

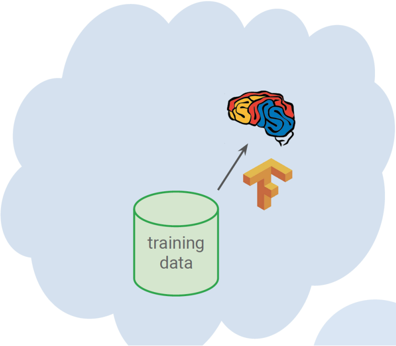
Make predictions in the cloud



Gather training data in the cloud



And make the models better



Why On-Device Learning?

- explosive growth in the volume of data on devices
- growing computation and storage capacity of devices
- privacy leakage, long delay
- ...

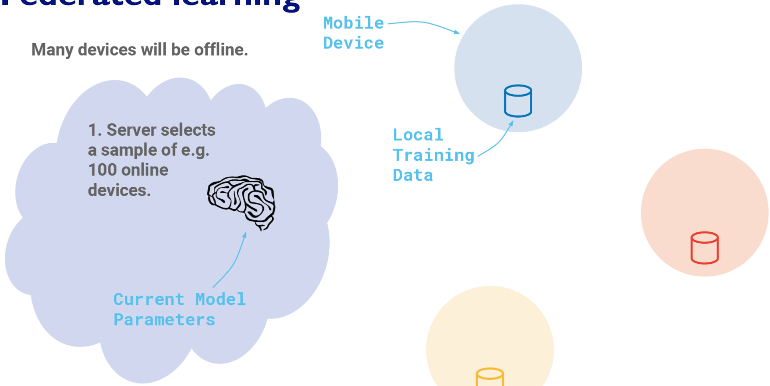
Why On-Device Learning?

- explosive growth in the volume of data on devices
- growing computation and storage capacity of devices
- privacy leakage, long delay
- ...

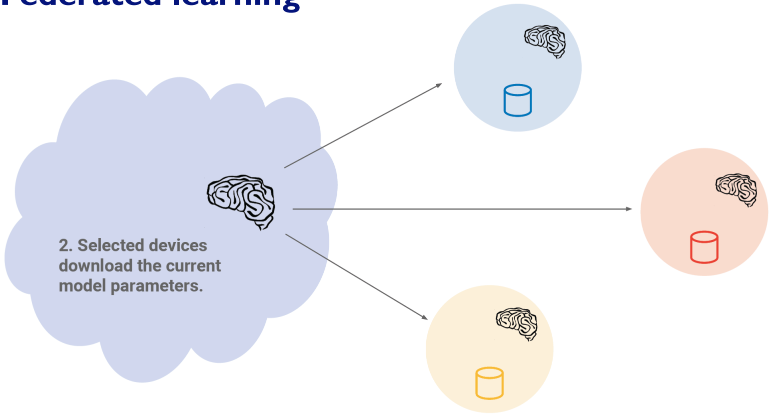
New Framework: **Federated Learning**

Federated learning

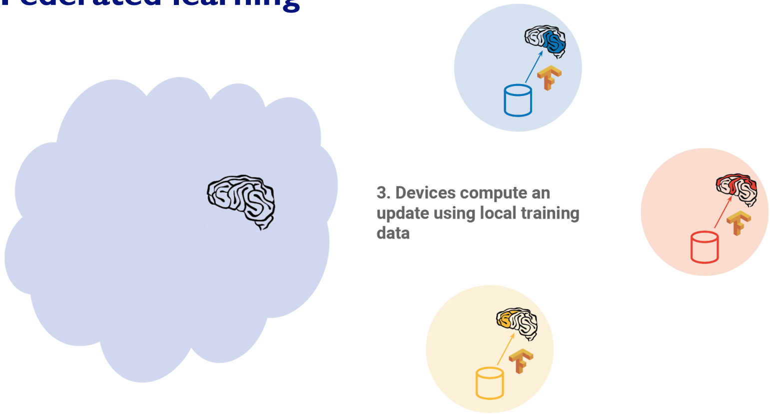
Many devices will be offline.



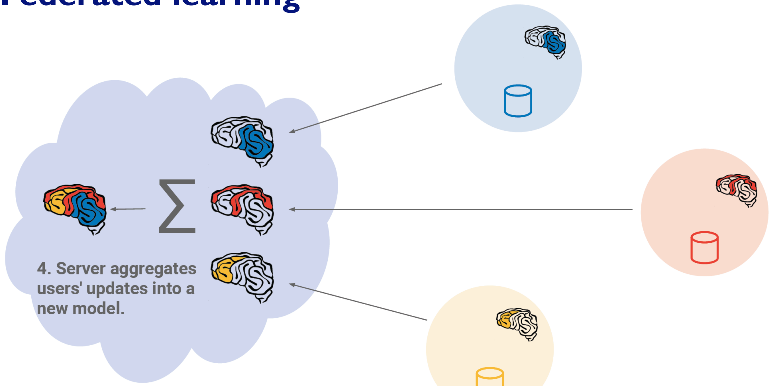
Federated learning



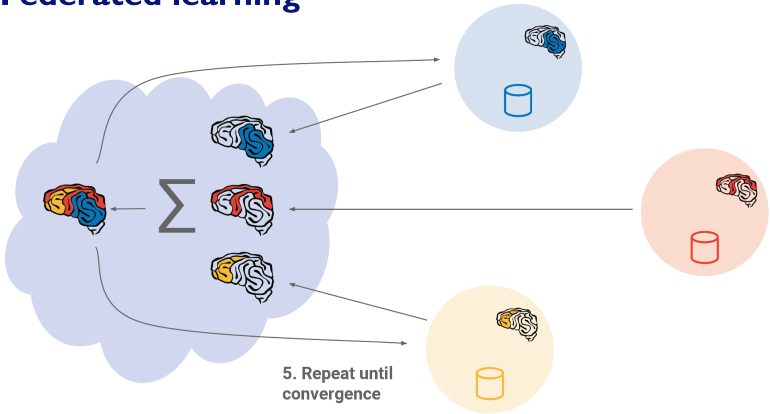
Federated learning



Federated learning

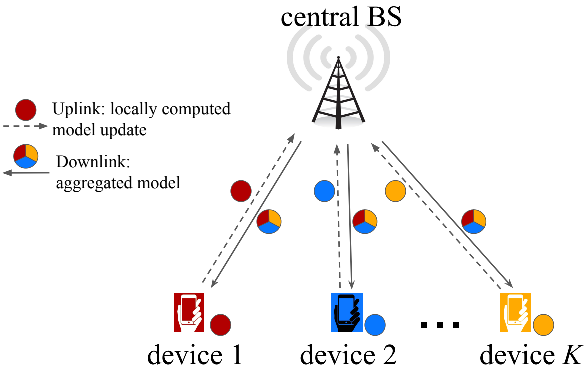


Federated learning



Federated Learning over Wireless Networks

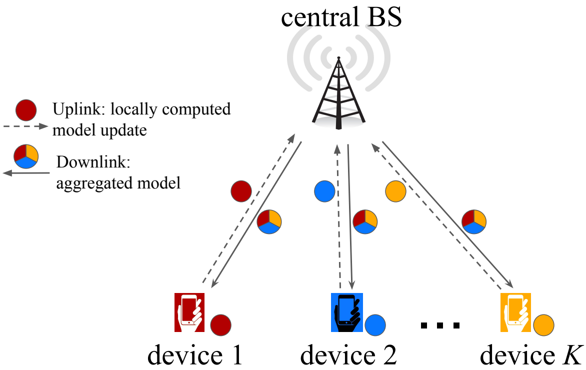
- **Goal**: train a shared global model via **wireless** federated computation



Q: How to efficiently aggregate models over wireless networks?

Federated Learning over Wireless Networks

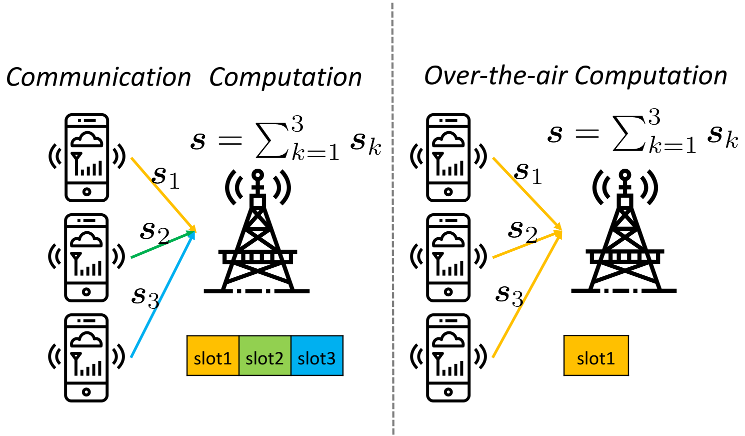
- **Goal**: train a shared global model via **wireless** federated computation



Q: How to efficiently aggregate models over wireless networks?

A: Via over-the-air computation.

Simple Illustration of Over-the-Air Computation



The Training Procedures

- **Challenges**
  - possible signal distortion in wireless communications
  - slow convergence because a large number of iterations is required to train a satisfactory model

The Training Procedures

- **Challenges**
  - possible signal distortion in wireless communications
  - slow convergence because a large number of iterations is required to train a satisfactory model
- **Our Work**
  - propose a difference-of-convex-functions (DC) algorithm to minimize signal distortion
  - fasten model convergence by adopting the canonical Newton's method for local model updates

Outline

Notations

- One N-antenna base station (BS), K single-antenna mobile devices
- $\mathcal{K}$ is the set of all devices
- the received signal at the BS after concurrent transmissions

$$\boldsymbol{y} = \sum_{k \in \mathcal{K}} \boldsymbol{h}_k b_k s_k + \boldsymbol{n}$$

  - s_k: the representative signal; b_k: allocated power
  - $\boldsymbol{h}_k$: the channel vector between the k-th device and the BS
  - $\boldsymbol{n} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$: the noise vector

Notations

- One N-antenna base station (BS), K single-antenna mobile devices
- $\mathcal{K}$ is the set of all devices
- the received signal at the BS after concurrent transmissions

$$\boldsymbol{y} = \sum_{k \in \mathcal{K}} \boldsymbol{h}_k b_k s_k + \boldsymbol{n}$$

  - s_k: the representative signal; b_k: allocated power
  - $\boldsymbol{h}_k$: the channel vector between the k-th device and the BS
  - $\boldsymbol{n} \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$: the noise vector
- the signal after decoding at the BS

$$\hat{\mathrm{w}} = \frac{1}{\sqrt{\eta}} \boldsymbol{a}^{\mathrm{H}} \boldsymbol{y} = \frac{1}{\sqrt{\eta}} \boldsymbol{a}^{\mathrm{H}} \sum_{k \in \mathcal{K}} \boldsymbol{h}_k b_k s_k + \frac{1}{\sqrt{\eta}} \boldsymbol{a}^{\mathrm{H}} \boldsymbol{n}$$

  - η: a normalizing factor
  - $\boldsymbol{a}$: the receive beamforming vector at the BS

Problem Formulation

- The distortion of decoded signal $\hat{\mathrm{w}}$ and ideal signal $\mathrm{w} = \sum_{k\in\mathcal{K}} s_k$ is measured by mean-square-error (MSE)

$$\begin{aligned}\mathrm{MSE}(\hat{\mathrm{w}}, \mathrm{w}; \boldsymbol{a}) &= \mathbb{E}\left(|\hat{\mathrm{w}} - \mathrm{w}|^2\right) \\ &= \sum_k \left|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k b_k/\sqrt{\eta} - 1\right|^2 + \sigma^2\|\boldsymbol{a}\|^2/\eta\end{aligned}$$

  which can be further simplified as

$$\mathrm{MSE}(\hat{\mathrm{w}}, \mathrm{w}; \boldsymbol{a}) = \frac{\|\boldsymbol{a}\|^2\sigma^2}{\eta} = \frac{\|\boldsymbol{a}\|^2\sigma^2}{P_0 \min_{k\in\mathcal{K}} \left\|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\right\|^2}$$

  by setting $b_k = \sqrt{\eta}\frac{\left(\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\right)^{\mathrm{H}}}{\|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\|^2}$ and $\eta = P_0 \min_{k\in\mathcal{K}} \left\|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\right\|^2$.

Problem Formulation

- **Goal**: find the optimal receive beamforming vector $\boldsymbol{a}$ to minimize MSE
- **Problem Formulation**

$$\underset{\boldsymbol{a}\in\mathbb{C}^N}{\text{minimize}} \left(\max_{k\in\mathcal{K}} \frac{\|\boldsymbol{a}\|^2}{\left\|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\right\|^2} \right),$$

  and it can be recast as

$$\begin{aligned} \underset{\boldsymbol{a}\in\mathbb{C}^N}{\text{minimize}} \quad & \|\boldsymbol{a}\|^2 \\ \text{subject to} \quad & \left\|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\right\|^2 \geq 1, \forall k \in \mathcal{K}. \end{aligned}$$

Problem Formulation

- **Goal**: find the optimal receive beamforming vector $\boldsymbol{a}$ to minimize MSE
- **Problem Formulation**

$$\underset{\boldsymbol{a}\in\mathbb{C}^N}{\text{minimize}} \left(\max_{k\in\mathcal{K}} \frac{\|\boldsymbol{a}\|^2}{\|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\|^2} \right),$$

  and it can be recast as

$$\begin{aligned} \underset{\boldsymbol{a}\in\mathbb{C}^N}{\text{minimize}} \quad & \|\boldsymbol{a}\|^2 \\ \text{subject to} \quad & \left\|\boldsymbol{a}^{\mathrm{H}}\boldsymbol{h}_k\right\|^2 \geq 1, \forall k \in \mathcal{K}. \end{aligned}$$

- **Challenges**: a nonconvex quadratically constrained quadratic programming (QCQP)
- **Proposal**: low-rank optimization after matrix lifting

Outline

Low-Rank Optimization

- define $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{A} \succeq \boldsymbol{0}$ with $\operatorname{rank}(\boldsymbol{A}) = 1$
- **Problem Rewrite**

$$\begin{aligned} \underset{\boldsymbol{A} \in \mathbb{C}^{N \times N}}{\text{minimize}} \quad & \operatorname{Tr}(\boldsymbol{A}) \\ \text{subject to} \quad & \operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{H}_k\right) \geq 1, \forall k \in \mathcal{K} \\ & \boldsymbol{A} \succeq \boldsymbol{0}, \operatorname{rank}(\boldsymbol{A}) = 1 \end{aligned}$$

Low-Rank Optimization

- define $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{A} \succeq \boldsymbol{0}$ with $\mathrm{rank}(\boldsymbol{A}) = 1$
- **Problem Rewrite**

$$
\begin{aligned}
\underset{\boldsymbol{A}\in\mathbb{C}^{N\times N}}{\text{minimize}} \quad & \mathrm{Tr}(\boldsymbol{A}) \\
\text{subject to} \quad & \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{H}_k\right) \geq 1, \forall k \in \mathcal{K} \\
& \boldsymbol{A} \succeq \boldsymbol{0}, \mathrm{rank}(\boldsymbol{A}) = 1
\end{aligned}
$$

- **Challenges**: the nonconvex rank-one constraint
- **Proposal**: a DC representation for the rank-one constraint

DC Reformulation

- **DC Representation**

$$\mathrm{rank}(\boldsymbol{A}) = 1 \Longleftrightarrow \mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\|_2 = 0,\ \mathrm{Tr}(\boldsymbol{A}) > 0$$

- **DC Reformulation**

$$\begin{aligned} \underset{\boldsymbol{A} \in \mathbb{C}^{N \times N}}{\text{minimize}} \quad & \mathrm{Tr}(\boldsymbol{A}) + \beta \left(\mathrm{Tr}(\boldsymbol{A}) - \|\boldsymbol{A}\|_2 \right) \\ \text{subject to} \quad & \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{H}_k \right) \geq 1, \forall k \in \mathcal{K} \\ & \boldsymbol{A} \succeq \mathbf{0}, \mathrm{Tr}(\boldsymbol{A}) > 0 \end{aligned}$$

DC Algorithm

- At iteration t, $\boldsymbol{A}^t$ is obtained by solving subproblem

$$\begin{array}{ll} \underset{\boldsymbol{A}\in\mathbb{C}^{N\times N}}{\text{minimize}} & (1+\beta)\,\mathrm{Tr}(\boldsymbol{A}) - \beta\left\langle \partial\left\|\boldsymbol{A}^t\right\|_2, \boldsymbol{A}\right\rangle \\ \text{subject to} & \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{H}_k\right) \geq 1, \forall k \in \mathcal{K} \\ & \boldsymbol{A} \succeq \mathbf{0}, \mathrm{Tr}(\boldsymbol{A}) > 0 \end{array},$$

where $\partial\left\|\boldsymbol{A}^t\right\|_2$ is one of subgradients of the spectral norm at point $\boldsymbol{A}^t$, and $\langle\cdot,\cdot\rangle$is the inner product of two matrices defined as $\langle\boldsymbol{X},\boldsymbol{Y}\rangle = \mathrm{Real}\left(\mathrm{Tr}\left(\boldsymbol{X}^{\mathrm{H}}\boldsymbol{Y}\right)\right)$

DC Algorithm

- At iteration t, $\boldsymbol{A}^t$ is obtained by solving subproblem

$$\begin{array}{ll}\underset{\boldsymbol{A}\in\mathbb{C}^{N\times N}}{\text{minimize}} & (1+\beta)\operatorname{Tr}(\boldsymbol{A})-\beta\left\langle\partial\left\|\boldsymbol{A}^t\right\|_2,\boldsymbol{A}\right\rangle\\ \text{subject to} & \operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{H}_k\right)\geq 1,\forall k\in\mathcal{K}\\ & \boldsymbol{A}\succeq\mathbf{0},\operatorname{Tr}(\boldsymbol{A})>0\end{array},$$

  where $\partial\left\|\boldsymbol{A}^t\right\|_2$ is one of subgradients of the spectral norm at point $\boldsymbol{A}^t$, and $\langle\cdot,\cdot\rangle$is the inner product of two matrices defined as $\langle\boldsymbol{X},\boldsymbol{Y}\rangle=\operatorname{Real}\left(\operatorname{Tr}\left(\boldsymbol{X}^{\mathrm{H}}\boldsymbol{Y}\right)\right)$
- Repeat the above DC algorithm until convergence for a feasible $\boldsymbol{A}$ with exact rank-one; then $\boldsymbol{a}$ is obtained via singular value decomposition (SVD).

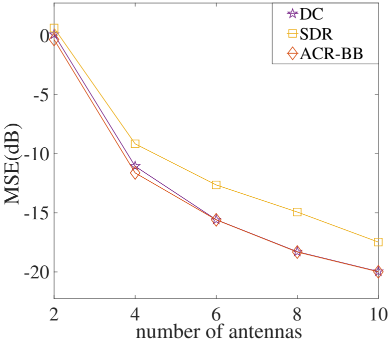
Outline

A Quick Review

- **Challenges**
  - possible signal distortion in wireless communications
  - slow convergence because a large number of iterations is required to train a satisfactory model
- **Our Work**
  - propose a difference-of-convex-functions (DC) algorithm to minimize signal distortion
  - fasten model convergence by adopting the canonical Newton's method for local model updates

Simulation Results

- $K = 5$, results averaged over 100 independently generated channel realizations



SDR[Sidiropoulos et al.'06]: convexify the nonconvex QCQP by simply dropping the rank-one constraint

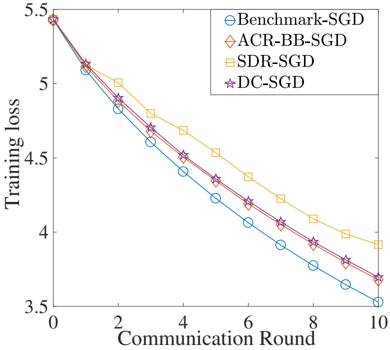
ACR-BB[Lu et al.'17]: a global approach for QCQP based on the branch-and-bound algorithm

Remark

- The proposed DC algorithm achieves nearly-optimal performance on minimizing MSE.

Simulation Results

- Classification experiment over CIFAR10 dataset
  - train a softmax classifier via the distributed stochastic gradient descent (SGD)



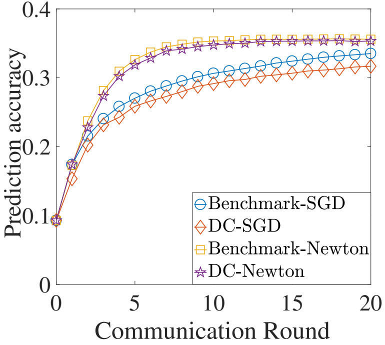
Benchmark: ideal transmission with no aggregation errors, i.e., MSE $= 0$dB

Remark

- The aggregation errors significantly degrade performance.

Simulation Results

- Classification experiment over CIFAR10 dataset
  - train a softmax classifier via the canonical Newton's method



Remark

- The Newton's method significantly fasten the convergence of trained model and is much more robust to the aggregation errors.

Outline

Concluding Remarks

- We propose a **second-order based** model update method for on-device federated learning.
- We develop a low-rank approach to support over-the-air computation, followed by **a novel DC algorithm**.
- We demonstrate the connection between aggregation errors and model convergence behaviors through experiments, i.e., large aggregation error $\Longrightarrow$ slow convergence
- Second order methods benefit from two aspects:
  1. reduce the communication burden because much less communication rounds are required for convergence
  2. be **much more robust to aggregation errors** therefore errors result in very limited performance loss

Thanks !

huasheng@shanghaitech.edu.cn