MARINA: Faster Non-Convex Distributed Learning with Compression

Eduard Gorbunov
MIPT, Yandex.Research

Konstantin Burlachenko
KAUST

Zhize Li
KAUST

Peter Richtárik
KAUST

Moscow Conference on Combinatorics and Applications

The talk is based on the paper accepted to ICML 2021

June 2, 2021

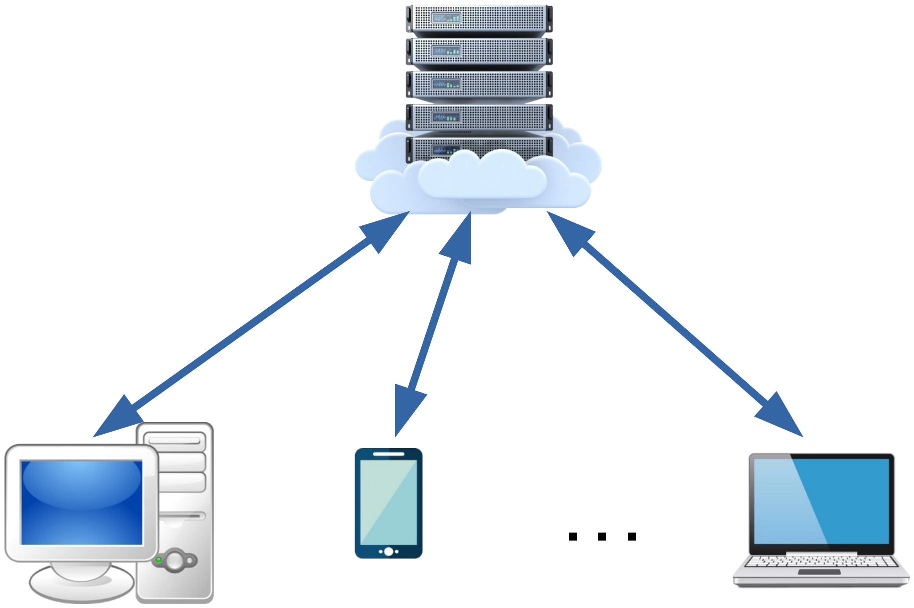
Konstantin Burlachenko
PhD student
KAUST

Zhize Li
Research Scientist
KAUST

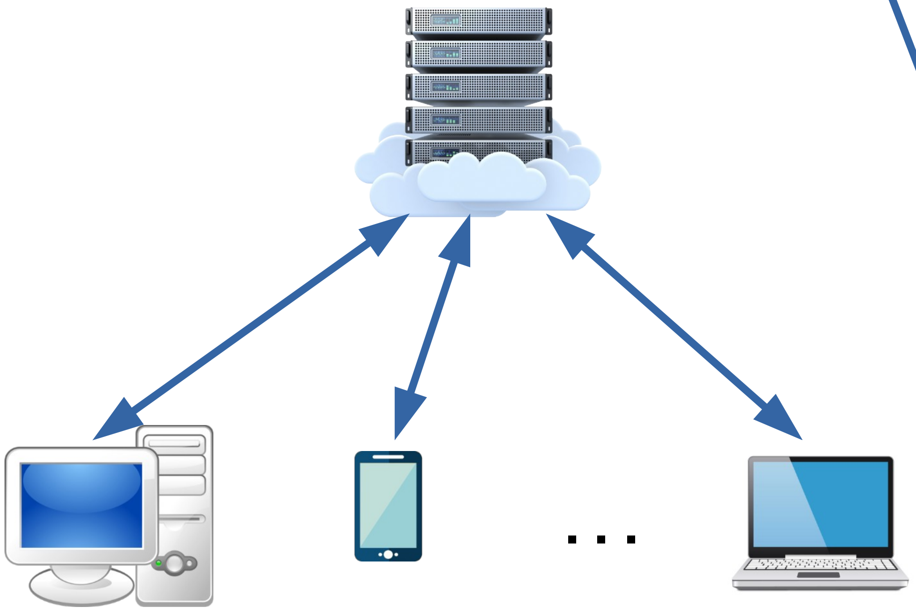
Peter Richtárik
Professor of Computer Science
KAUST

Outline

1. The Problem



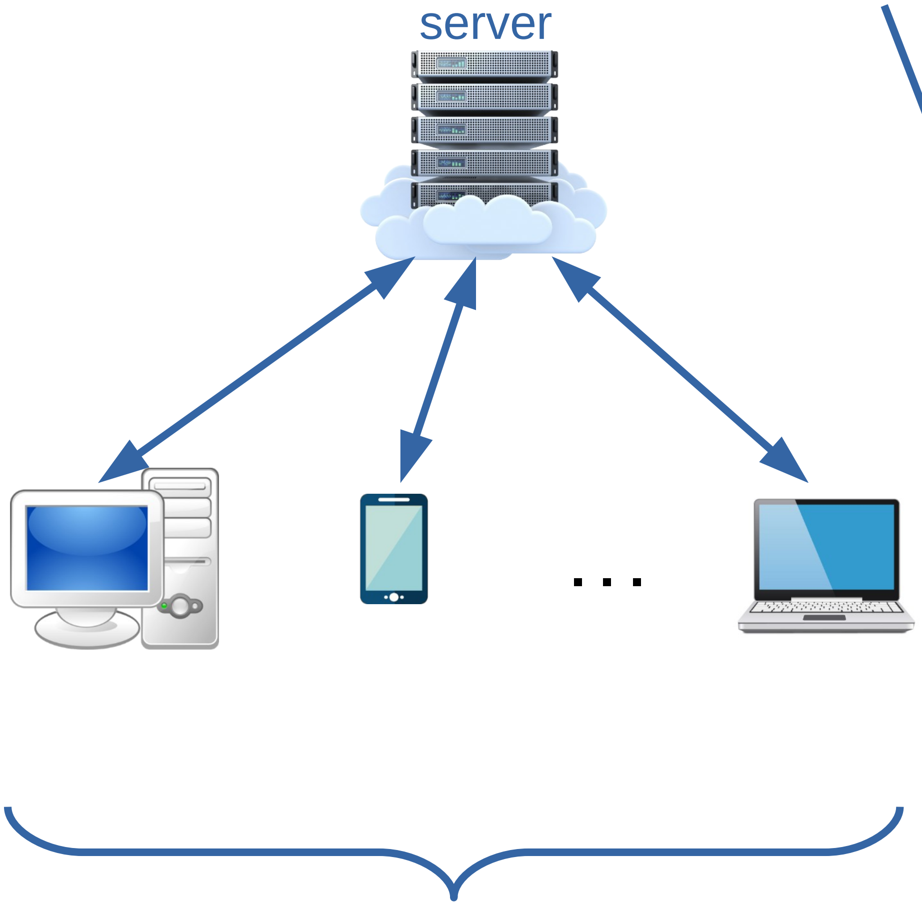
$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}$$



parameters of the model

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

# of parameters



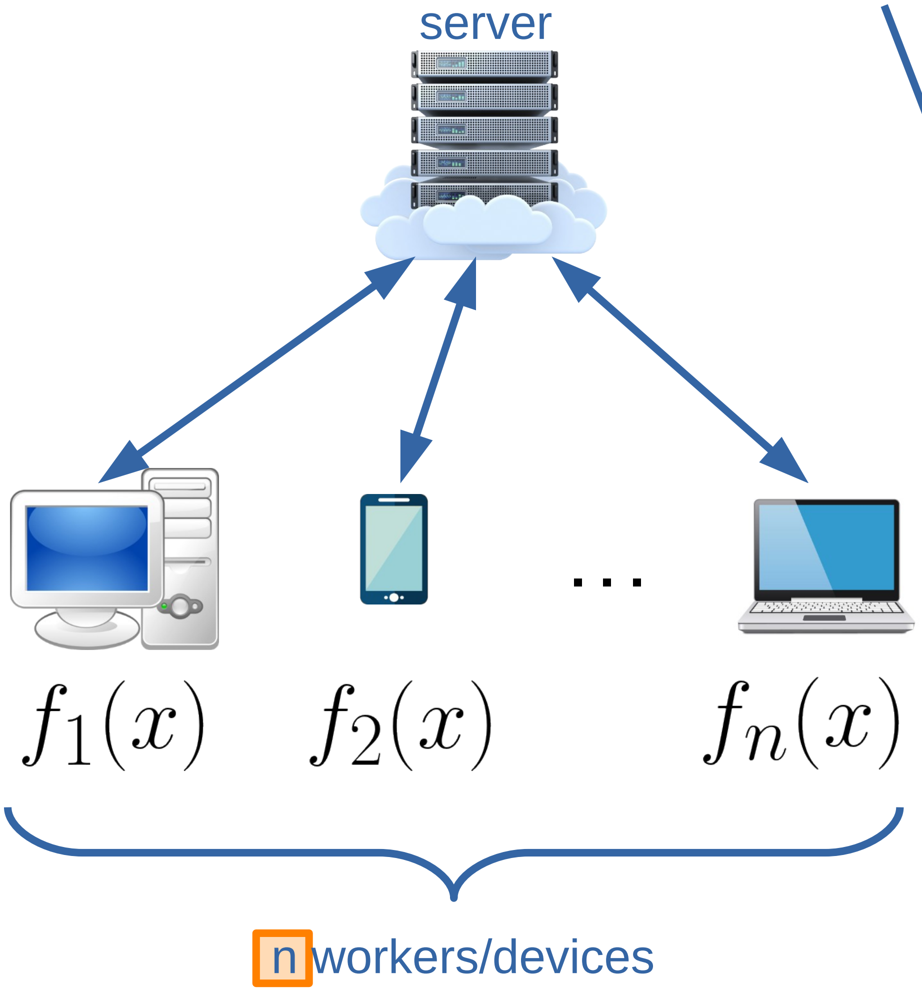
$\boxed{n}$ workers/devices

parameters of the model

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{\boxed{n}} \sum_{i=1}^{\boxed{n}} f_i(x) \right\}$$

# of parameters

# of workers/devices



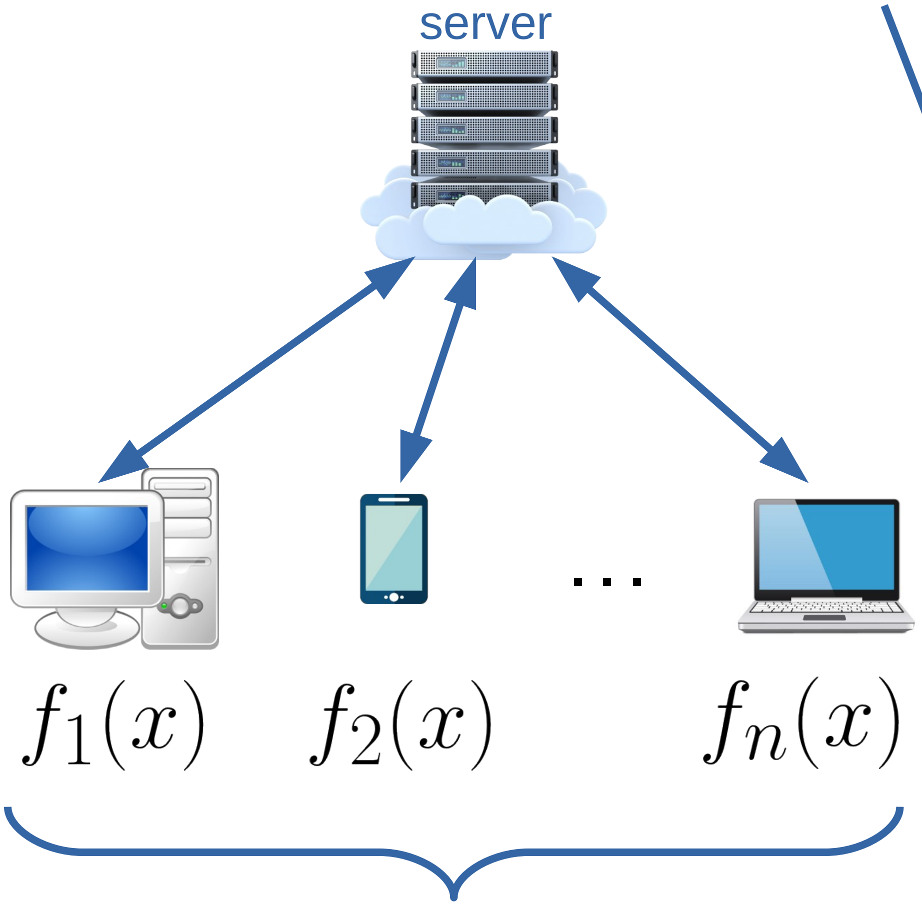
parameters of the model

$$\min_{x\in\mathbb{R}^d}\left\{ f(x) := \frac{1}{\boxed{n}}\sum_{i=1}^{\boxed{n}} f_i(x)\right\}$$

# of parameters

# of workers/devices

loss on the data accessible by worker *i*



$\boxed{n}$ workers/devices

parameters of the model

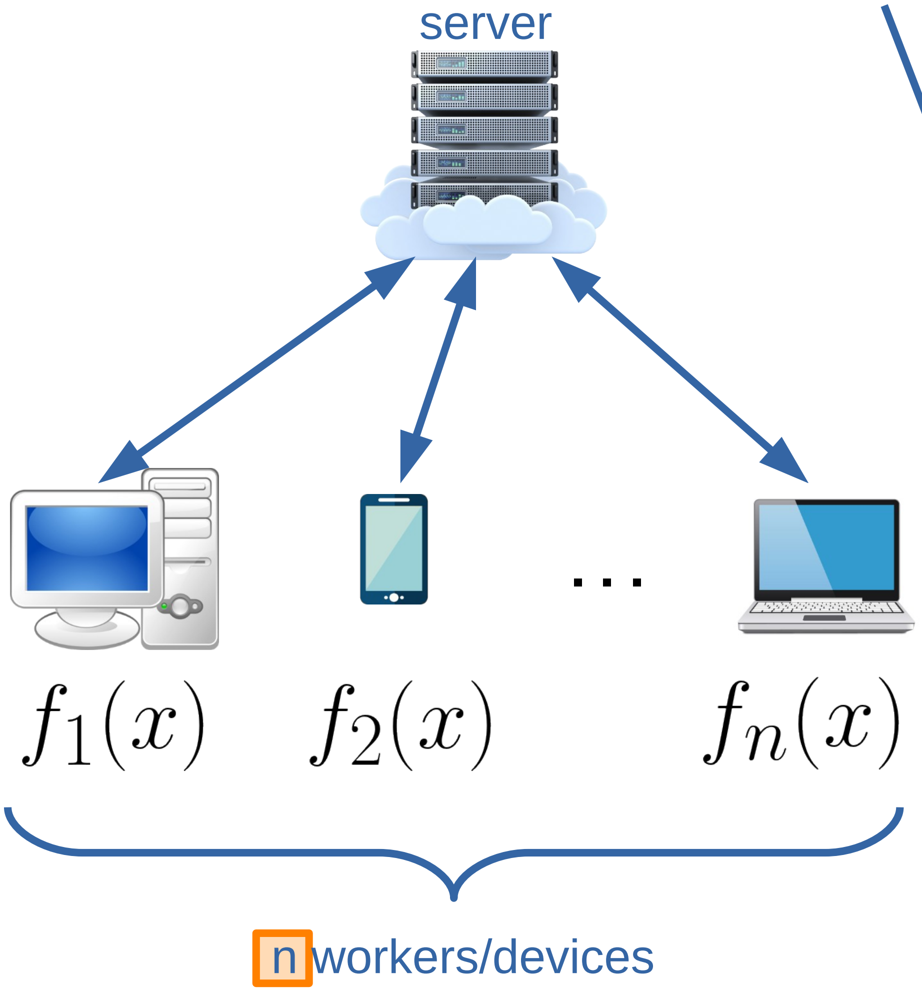
$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{\boxed{n}} \sum_{i=1}^{\boxed{n}} f_i(x) \right\}$$

\# of parameters

\# of workers/devices

loss on the data accessible by worker *i*

- $f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i}\left[f_{\xi_i}(x)\right]$



parameters of the model

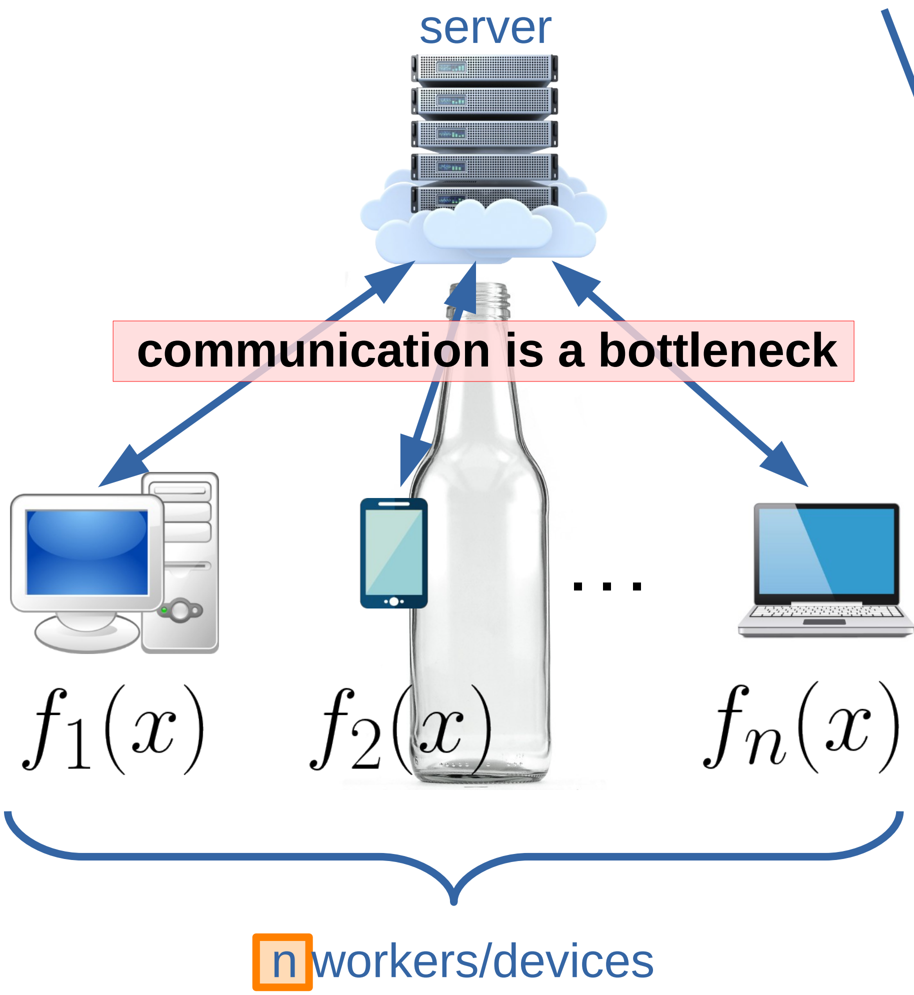
$$\min_{x\in\mathbb{R}^d}\left\{ f(x) := \frac{1}{\boxed{n}}\sum_{i=1}^{\boxed{n}} f_i(x)\right\}$$

\# of parameters

\# of workers/devices

loss on the data accessible by worker *i*

- $f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right]$
- $f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$



parameters of the model

$$\min_{x\in\mathbb{R}^d}\left\{ f(x) := \frac{1}{\boxed{n}}\sum_{i=1}^{\boxed{n}} f_i(x)\right\}$$

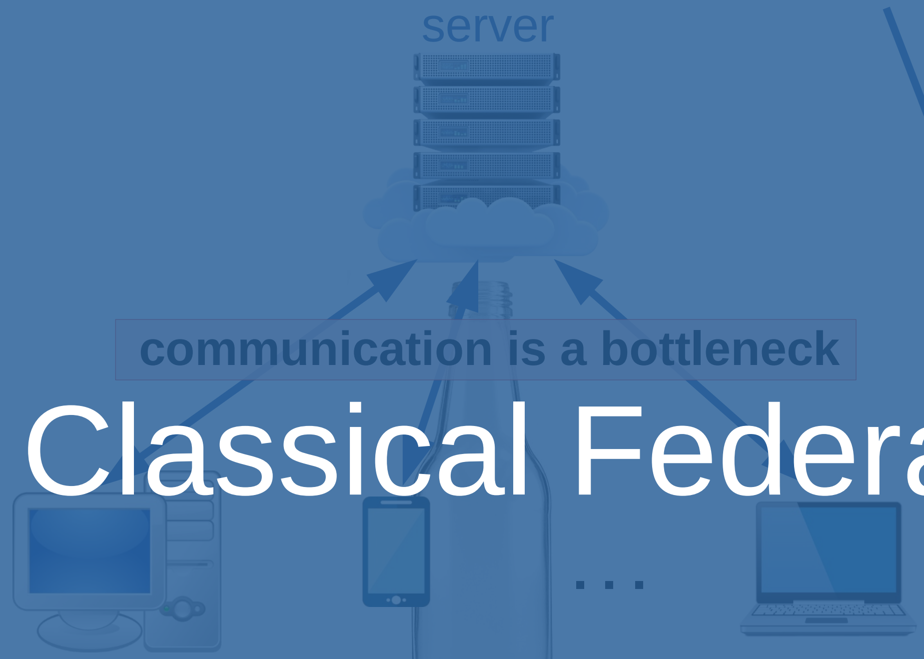
# of parameters

# of workers/devices

loss on the data accessible by worker *i*

- $$f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right]$$
- $$f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$$

Classical Federated Learning Setup



$f_1(x) \quad f_2(x) \quad \dots \quad f_n(x)$

$\boxed{n}$ workers/devices

parameters of the model

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{\boxed{n}} \sum_{i=1}^{\boxed{n}} f_i(x) \right\}$$

# of parameters

# of workers/devices

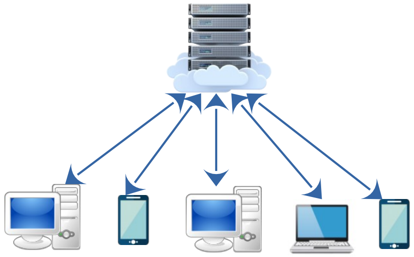
loss on the data accessible by worker *i*

- $f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i} \left[f_{\xi_i}(x) \right]$
- $f_i(x) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(x)$

How to Handle Communication Bottleneck?

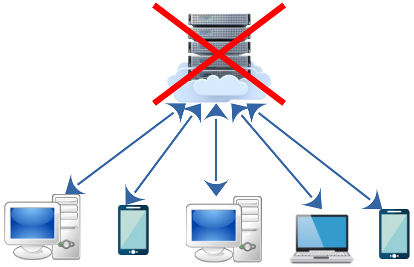
How to Handle Communication Bottleneck?

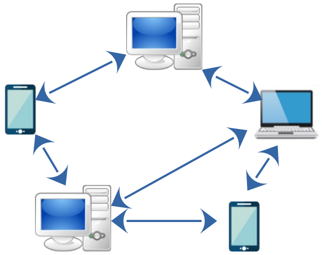
- Change the topology of the network



How to Handle Communication Bottleneck?

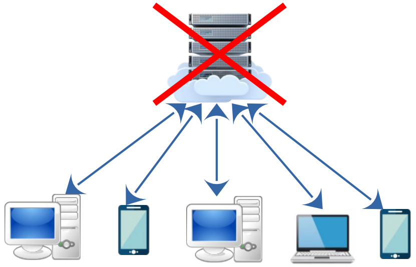
- Change the topology of the network → Decentralized optimization

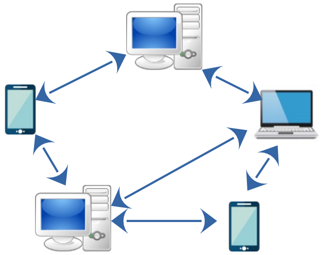




How to Handle Communication Bottleneck?

- Change the topology of the network → Decentralized optimization

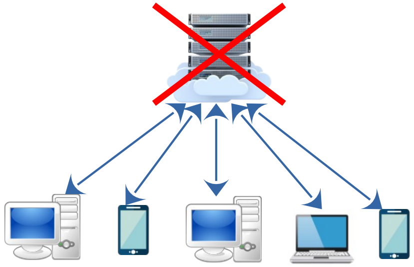


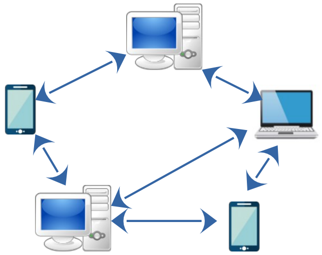


- Do more work on each worker
  in the hope of communicating less

How to Handle Communication Bottleneck?

- Change the topology of the network → Decentralized optimization

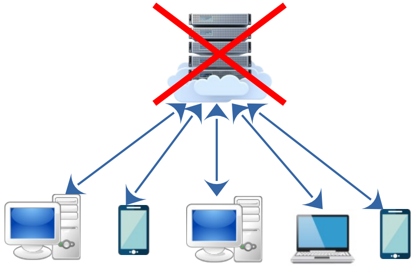




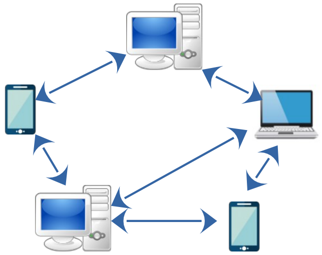
- Do more work on each worker in the hope of communicating less → Local-SGD/Federated Averaging

How to Handle Communication Bottleneck?

- Change the topology of the network → Decentralized optimization



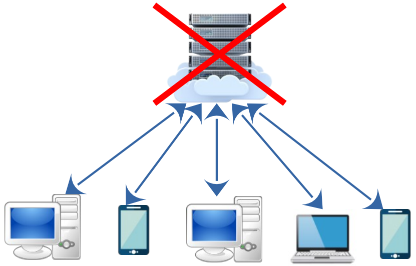
→



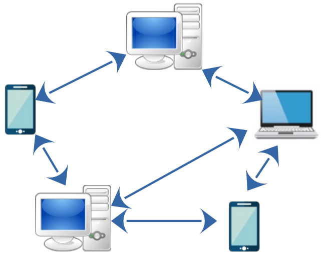
- Do more work on each worker in the hope of communicating less → Local-SGD/Federated Averaging
- Send less information to reduce the communication cost

How to Handle Communication Bottleneck?

- Change the topology of the network → Decentralized optimization



→



- Do more work on each worker in the hope of communicating less → Local-SGD/Federated Averaging

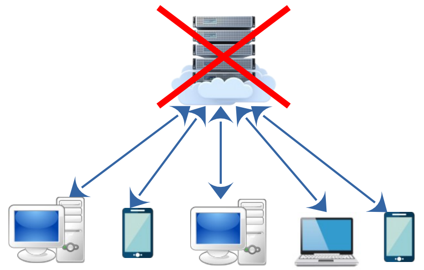
- Send less information to reduce the communication cost

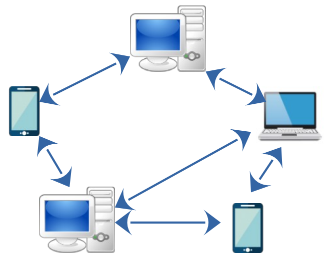
Workers send dense vectors

$$g = \begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix}$$

How to Handle Communication Bottleneck?

- Change the topology of the network → Decentralized optimization





- Do more work on each worker in the hope of communicating less → Local-SGD/Federated Averaging

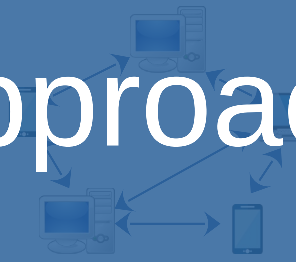
- Send less information to reduce the communication cost

Workers send dense vectors → Workers send compressed/sparse vectors

$$g = \begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \quad \longrightarrow \quad \mathcal{Q}(g) = \frac{5}{2} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -7 \\ 0 \end{pmatrix}$$

How to Handle Communication Bottleneck?

- Change the topology of the network → Decentralized optimization



We study this approach

- Do more work on each worker in the hope of communicating less → Local-SGD/Federated Averaging

- Send less information to reduce the communication cost

Workers send dense vectors → Workers send compressed/sparse vectors

$$g = \begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \quad \longrightarrow \quad \mathcal{Q}(g) = \frac{5}{2} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -7 \\ 0 \end{pmatrix}$$

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x)$$

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega \|x\|^2$$

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega\|x\|^2$$

Example: RandK (for K = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow$$

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega \|x\|^2$$

Example: RandK (for K = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow$$

Pick K = 2 components uniformly at random

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega \|x\|^2$$

Example: RandK (for K = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \frac{5}{2} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -7 \\ 0 \end{pmatrix}$$

Pick K = 2 components uniformly at random

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \le \omega \|x\|^2$$

Example: RandK (for K = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \underbrace{\frac{5}{2}}_{\text{for unbiasedness}} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -7 \\ 0 \end{pmatrix}$$

Pick K = 2 components uniformly at random

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega \|x\|^2$$

Example: RandK (for K = $\boxed{2}$)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \frac{5}{\boxed{2}} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -7 \\ 0 \end{pmatrix}$$

for unbiasedness

Pick K = $\boxed{2}$ components uniformly at random

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega \|x\|^2$$

Example: RandK (for K = $\boxed{2}$)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \frac{5}{\boxed{2}} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -7 \\ 0 \end{pmatrix} \qquad \omega = \frac{d}{K} - 1$$

for unbiasedness

Pick K = $\boxed{2}$ components uniformly at random

Unbiased compression (quantization)

$$x \to \mathcal{Q}(x) \qquad \mathbb{E}[\mathcal{Q}(x)] = x$$

$$\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega\|x\|^2$$

Example: RandK (for K = $\boxed{2}$)

$$\boxed{d = 5}\left\{\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix}\right. \Rightarrow \frac{\boxed{5}}{\boxed{2}} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ -7 \\ 0 \end{pmatrix} \qquad \omega = \frac{\boxed{d}}{\boxed{K}} - 1$$

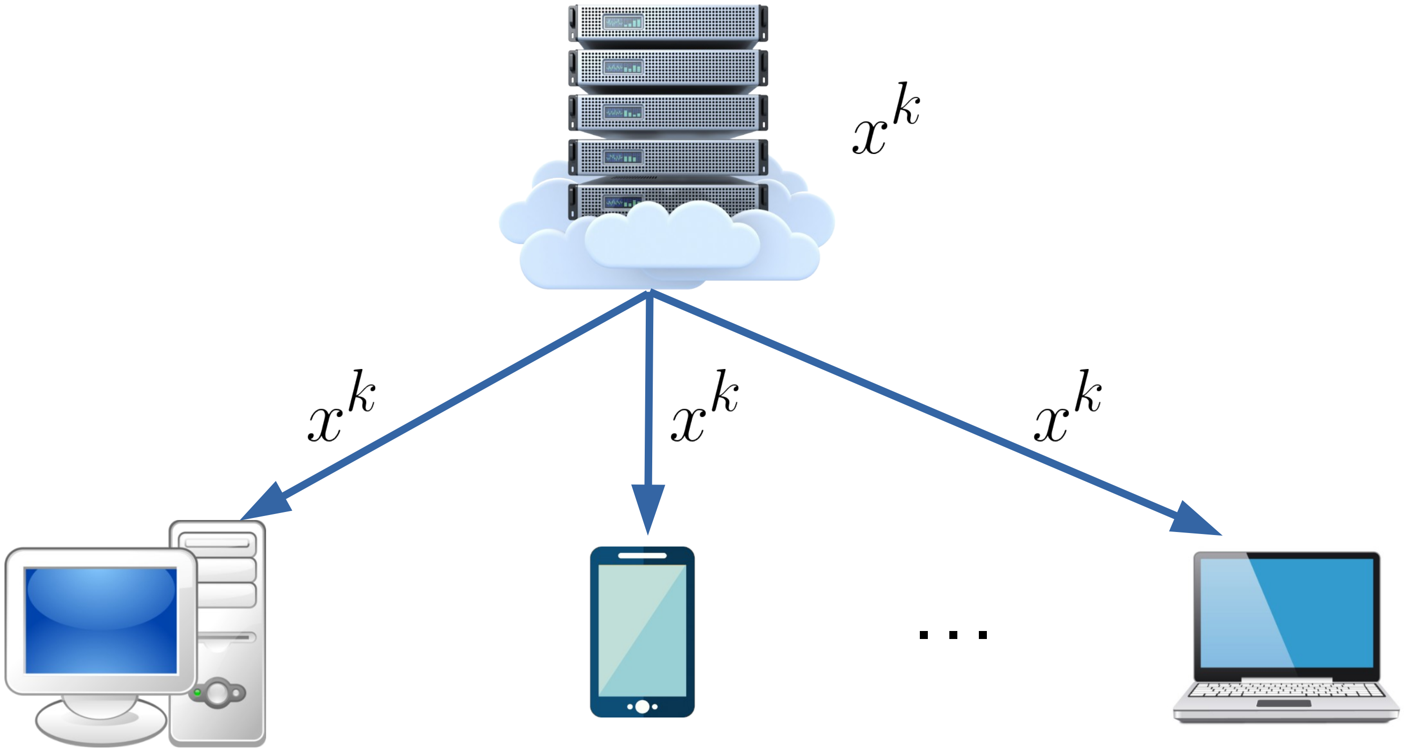
for unbiasedness

Pick K = $\boxed{2}$ components uniformly at random

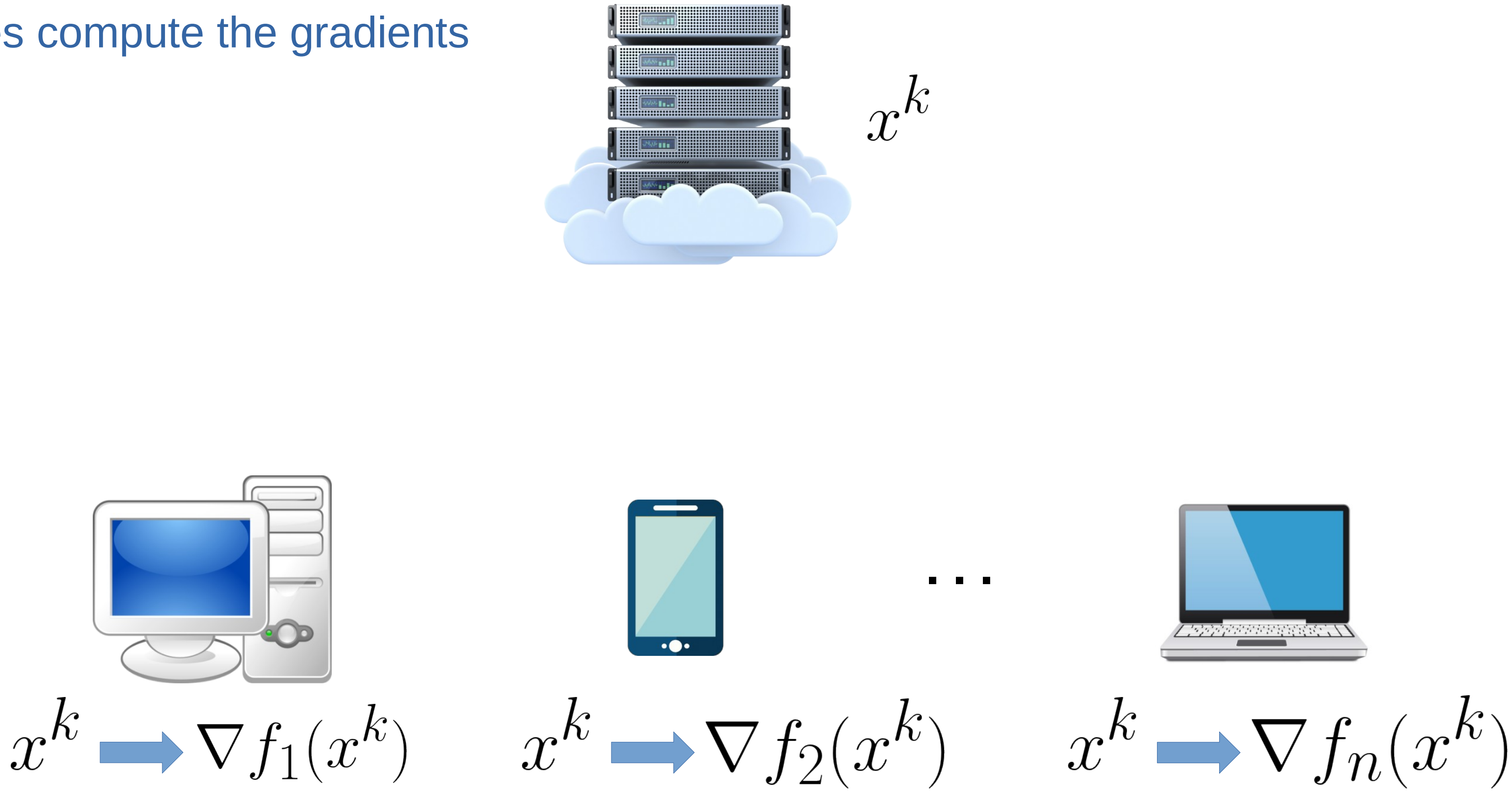
2. Quantized Gradient Descent (QGD)

Alistarh, Dan, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. **"QSGD: Communication-efficient SGD via gradient quantization and encoding."** *In Advances in Neural Information Processing Systems*, pp. 1709-1720. 2017.

1 Server broadcasts the parameters

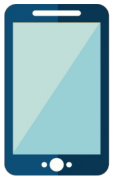


1. Server broadcasts the parameters
2. Devices compute the gradients



1. Server broadcasts the parameters
2. Devices compute the gradients
3. Devices quantize the gradients

x^k

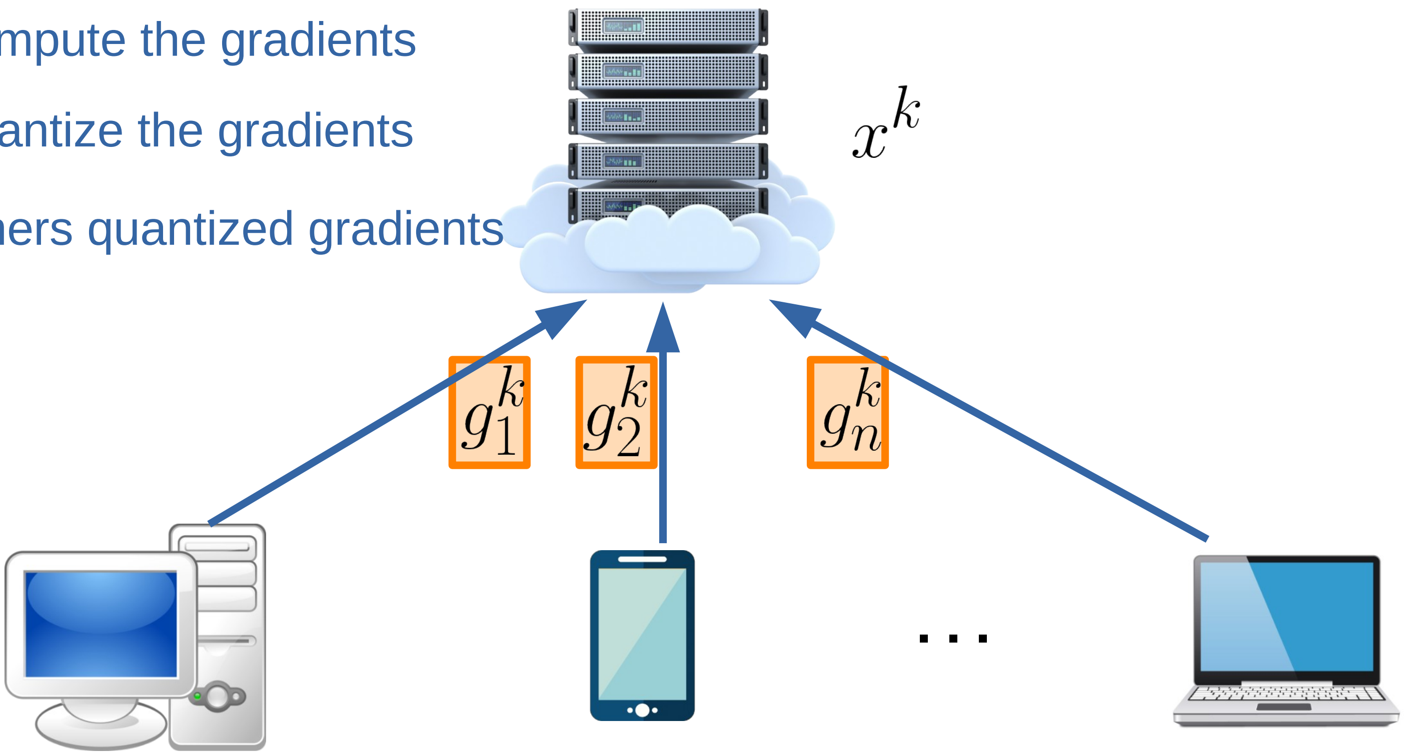


$\cdots$

$x^k \rightarrow \nabla f_1(x^k)$ $x^k \rightarrow \nabla f_2(x^k)$ $x^k \rightarrow \nabla f_n(x^k)$

$\boxed{g_1^k = \mathcal{Q}\left(\nabla f_1(x^k)\right)}$ $\boxed{g_2^k = \mathcal{Q}\left(\nabla f_2(x^k)\right)}$ $\boxed{g_n^k = \mathcal{Q}\left(\nabla f_n(x^k)\right)}$

1. Server broadcasts the parameters
2. Devices compute the gradients
3. Devices quantize the gradients
4. Server gathers quantized gradients

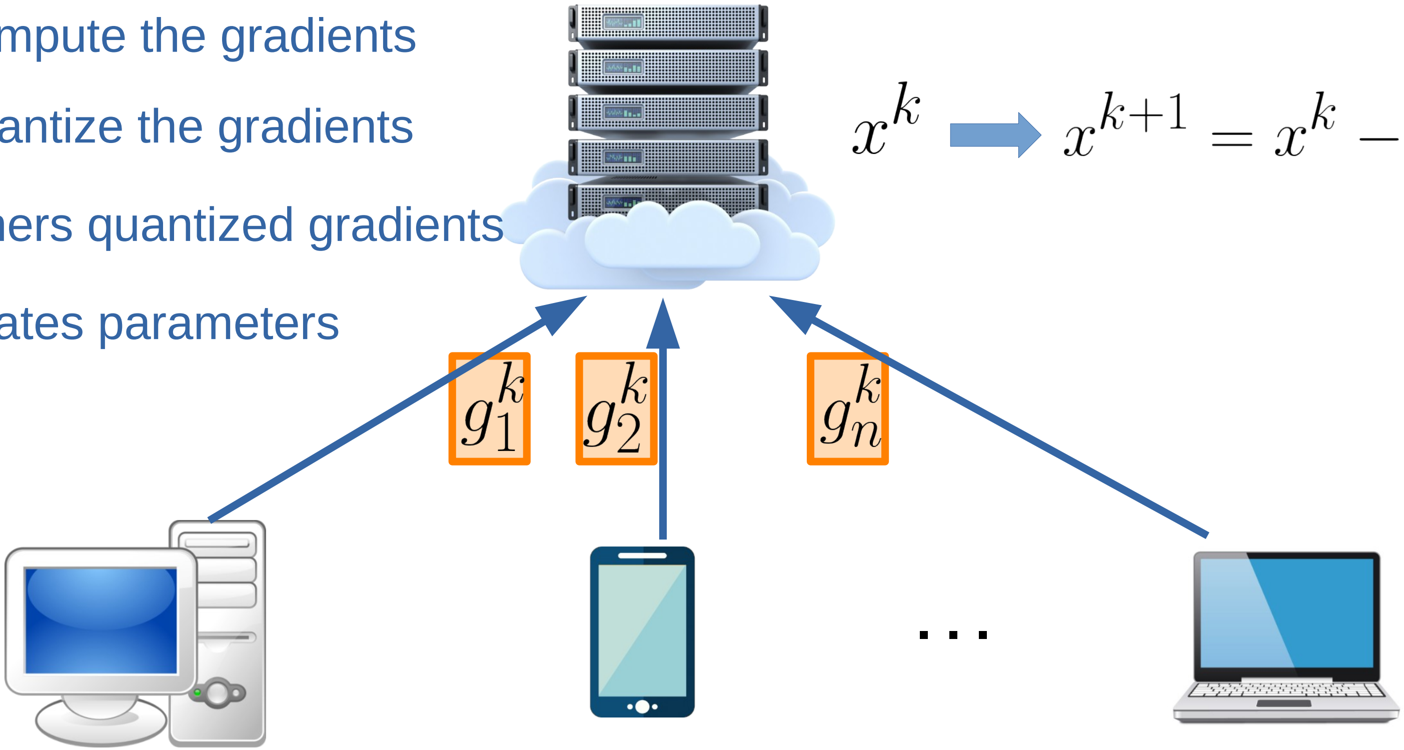


$x^k \rightarrow \nabla f_1(x^k)$ $\quad$ $x^k \rightarrow \nabla f_2(x^k)$ $\quad$ $x^k \rightarrow \nabla f_n(x^k)$

$$\boxed{g_1^k = \mathcal{Q}\left(\nabla f_1(x^k)\right)} \quad \boxed{g_2^k = \mathcal{Q}\left(\nabla f_2(x^k)\right)} \quad \boxed{g_n^k = \mathcal{Q}\left(\nabla f_n(x^k)\right)}$$

1. Server broadcasts the parameters
2. Devices compute the gradients
3. Devices quantize the gradients
4. Server gathers quantized gradients
5. Server updates parameters

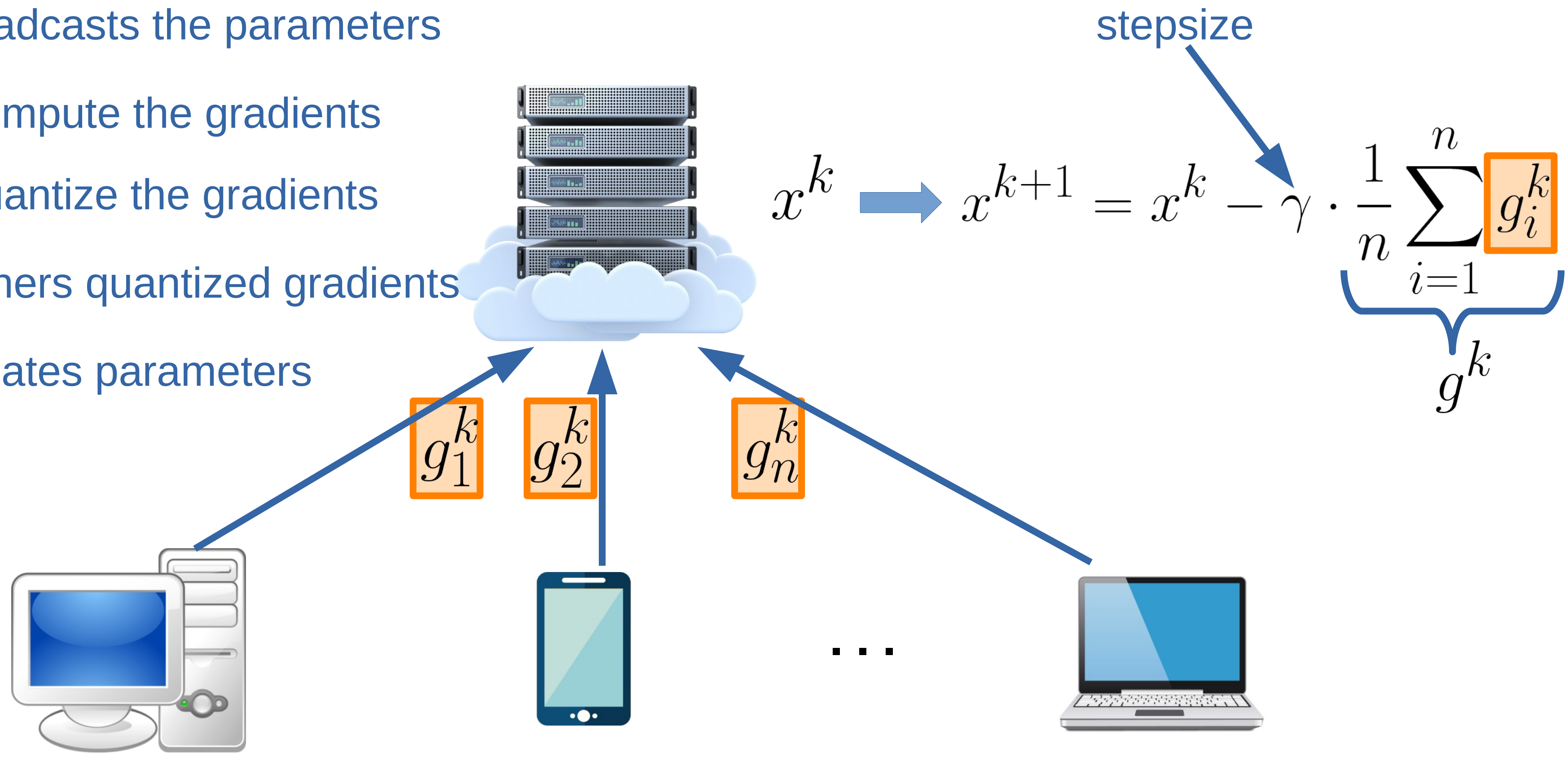
$$x^k \rightarrow x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} \boxed{g_i^k}$$



$$x^k \rightarrow \nabla f_1(x^k) \qquad x^k \rightarrow \nabla f_2(x^k) \qquad x^k \rightarrow \nabla f_n(x^k)$$

$$\boxed{g_1^k = \mathcal{Q}\left(\nabla f_1(x^k)\right)} \quad \boxed{g_2^k = \mathcal{Q}\left(\nabla f_2(x^k)\right)} \quad \boxed{g_n^k = \mathcal{Q}\left(\nabla f_n(x^k)\right)}$$

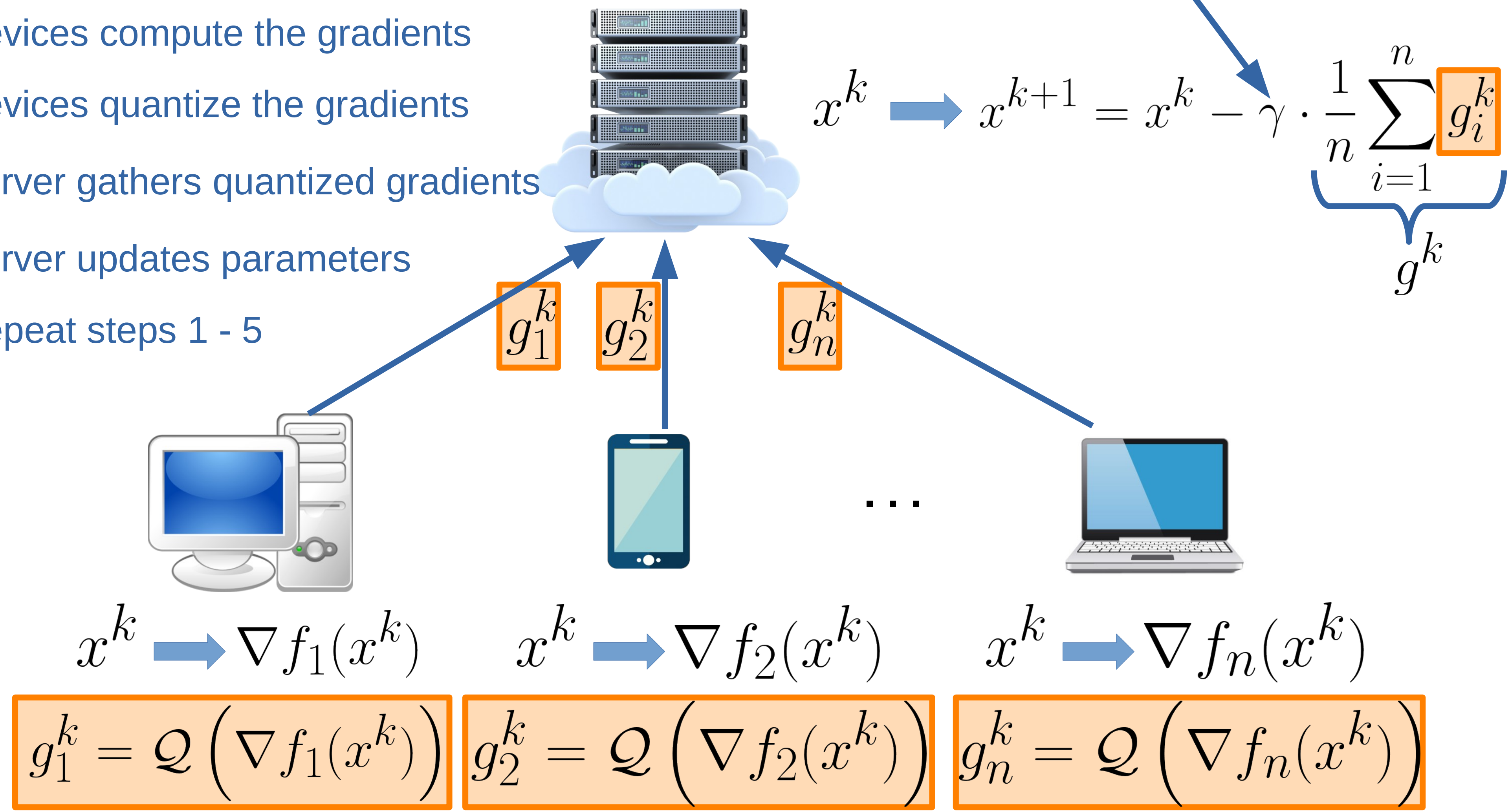
1. Server broadcasts the parameters
2. Devices compute the gradients
3. Devices quantize the gradients
4. Server gathers quantized gradients
5. Server updates parameters



$$x^k \Rightarrow \nabla f_1(x^k) \qquad x^k \Rightarrow \nabla f_2(x^k) \qquad x^k \Rightarrow \nabla f_n(x^k)$$

$$\boxed{g_1^k = \mathcal{Q}\left(\nabla f_1(x^k)\right)} \quad \boxed{g_2^k = \mathcal{Q}\left(\nabla f_2(x^k)\right)} \quad \boxed{g_n^k = \mathcal{Q}\left(\nabla f_n(x^k)\right)}$$

1. Server broadcasts the parameters
2. Devices compute the gradients
3. Devices quantize the gradients
4. Server gathers quantized gradients
5. Server updates parameters
6. Repeat steps 1 - 5



Assumptions

1 Uniform lower bound:

Assumptions

1 Uniform lower bound: $\boxed{\exists f_* \in \mathbb{R} : \ \forall x \in \mathbb{R}^d \quad f(x) \geq f_*}$

Assumptions

1. Uniform lower bound: $\boxed{\exists f_* \in \mathbb{R} : \ \forall x \in \mathbb{R}^d \quad f(x) \geq f_*}$

2. Smoothness:

Assumptions

1. Uniform lower bound: $\exists f_* \in \mathbb{R} : \ \forall x \in \mathbb{R}^d \quad f(x) \geq f_*$

2. Smoothness: $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$

Complexity Bound for QGD

Khaled, Ahmed, and Peter Richtárik. **"Better theory for SGD in the nonconvex world."** arXiv preprint arXiv:2002.03329 (2020).

QGD finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

Complexity Bound for QGD

Khaled, Ahmed, and Peter Richtárik. **"Better theory for SGD in the nonconvex world."** arXiv preprint arXiv:2002.03329 (2020).

QGD finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

$$\mathcal{O}\left(\frac{\Delta_0}{\varepsilon^2} + \frac{(1+\omega)\Delta_0^2}{\varepsilon^4 n} + \frac{(1+\omega)\Delta_0\Delta_f^*}{\varepsilon^4 n}\right)$$

communication rounds

Complexity Bound for QGD

Khaled, Ahmed, and Peter Richtárik. **"Better theory for SGD in the nonconvex world."** arXiv preprint arXiv:2002.03329 (2020).

QGD finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\Delta_0}{\varepsilon^2} + \frac{(1+\omega)\Delta_0^2}{\varepsilon^4 n} + \frac{(1+\omega)\Delta_0\Delta_f^*}{\varepsilon^4 n}\right)$$

communication rounds

Complexity Bound for QGD

Khaled, Ahmed, and Peter Richtárik. **"Better theory for SGD in the nonconvex world."** arXiv preprint arXiv:2002.03329 (2020).

QGD finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\Delta_0}{\boxed{\varepsilon^2}} + \frac{(1+\omega)\Delta_0^2}{\boxed{\varepsilon^4}n} + \frac{(1+\omega)\Delta_0\Delta_f^*}{\boxed{\varepsilon^4}n}\right)$$

communication rounds

Complexity Bound for QGD

Khaled, Ahmed, and Peter Richtárik. **"Better theory for SGD in the nonconvex world."** arXiv preprint arXiv:2002.03329 (2020).

QGD finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\Delta_0}{\boxed{\varepsilon^2}} + \frac{(1+\boxed{\omega})\Delta_0^2}{\boxed{\varepsilon^4} n} + \frac{(1+\boxed{\omega})\Delta_0\Delta_f^*}{\boxed{\varepsilon^4} n}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|Q(x) - x\|^2 \leq \omega \|x\|^2}$$

Complexity Bound for QGD

Khaled, Ahmed, and Peter Richtárik. **"Better theory for SGD in the nonconvex world."** arXiv preprint arXiv:2002.03329 (2020).

QGD finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants $\longrightarrow$

$$\mathcal{O}\left(\frac{\boxed{\Delta_0}}{\boxed{\varepsilon^2}} + \frac{(1+\boxed{\omega})\boxed{\Delta_0^2}}{\boxed{\varepsilon^4}n} + \frac{(1+\boxed{\omega})\boxed{\Delta_0}\Delta_f^*}{\boxed{\varepsilon^4}n}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega\|x\|^2}$$

$$\boxed{\Delta_0 = f(x^0) - f_*}$$

Complexity Bound for QGD

Khaled, Ahmed, and Peter Richtárik. **"Better theory for SGD in the nonconvex world."** arXiv preprint arXiv:2002.03329 (2020).

QGD finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants $\longrightarrow$ $\mathcal{O}\left(\frac{\boxed{\Delta_0}}{\boxed{\varepsilon^2}} + \frac{(1+\boxed{\omega})\boxed{\Delta_0^2}}{\boxed{\varepsilon^4} n} + \frac{(1+\boxed{\omega})\boxed{\Delta_0}\boxed{\Delta_f^*}}{\boxed{\varepsilon^4} n}\right)$ communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega \|x\|^2}$$

$$\boxed{\Delta_0 = f(x^0) - f_*}$$

$$\boxed{\Delta_f^* = f_* - \frac{1}{n}\sum_{i=1}^{n} f_{i,*}}$$

3. DIANA

Mishchenko, Konstantin, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. **"Distributed learning with compressed gradient differences."** arXiv preprint arXiv:1901.09269 (2019).

Horváth, Samuel, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. **"Stochastic distributed learning with gradient quantization and variance reduction."** arXiv preprint arXiv:1904.05115 (2019).

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n} \sum_{i=1}^{n} g_i^k$$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n} \sum_{i=1}^{n} g_i^k$$

QGD: $$g_i^k = \mathcal{Q}\left(\nabla f_i(x^k)\right)$$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n} \sum_{i=1}^{n} g_i^k$$

QGD: $$g_i^k = \mathcal{Q}\left(\nabla f_i(x^k)\right)$$

DIANA: $$g_i^k = h_i^k + \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n} \sum_{i=1}^{n} g_i^k$$

QGD: $$g_i^k = \mathcal{Q}\left(\nabla f_i(x^k)\right)$$

DIANA: $$g_i^k = \boxed{h_i^k} + \mathcal{Q}\left(\nabla f_i(x^k) - \boxed{h_i^k}\right)$$

learnable local shifts

$$\boxed{h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)}$$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

QGD: $$g_i^k = \boxed{\mathcal{Q}\left(\nabla f_i(x^k)\right)}$$

vectors that devices have to send

DIANA: $$g_i^k = \boxed{h_i^k} + \boxed{\mathcal{Q}\left(\nabla f_i(x^k) - \boxed{h_i^k}\right)}$$

learnable local shifts

$$\boxed{h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)}$$

Complexity Bound for DIANA

DIANA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

Complexity Bound for DIANA

DIANA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

$$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\omega)\sqrt{\omega/n}\right)}{\varepsilon^2}\right)$$

communication rounds

Complexity Bound for DIANA

DIANA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

$$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\omega)\sqrt{\omega/n}\right)}{\varepsilon^2}\right)$$

Hides numerical factors and smoothness constants

communication rounds

Complexity Bound for DIANA

DIANA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \le \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\omega)\sqrt{\omega/n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

Complexity Bound for DIANA

DIANA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\boxed{\omega})\sqrt{\boxed{\omega}/n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x)-x\|^2 \leq \omega\|x\|^2}$$

Complexity Bound for DIANA

DIANA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\boxed{\Delta_0}\left(1 + (1 + \boxed{\omega})\sqrt{\boxed{\omega}/n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega\|x\|^2}$$

$$\boxed{\Delta_0 = f(x^0) - f_*}$$

Complexity Bounds for DIANA and QGD

QGD: $$\mathcal{O}\left(\frac{\Delta_0}{\varepsilon^2} + \frac{(1+\omega)\Delta_0^2}{\varepsilon^4 n} + \frac{(1+\omega)\Delta_0\Delta_f^*}{\varepsilon^4 n}\right)$$

DIANA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\omega)\sqrt{\omega/n}\right)}{\varepsilon^2}\right)$$

Complexity Bounds for DIANA and QGD

QGD: $$\mathcal{O}\left(\frac{\Delta_0}{\boxed{\varepsilon^2}} + \frac{(1+\omega)\Delta_0^2}{\boxed{\varepsilon^4}n} + \frac{(1+\omega)\Delta_0\Delta_f^*}{\boxed{\varepsilon^4}n}\right)$$

DIANA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\omega)\sqrt{\omega/n}\right)}{\boxed{\varepsilon^2}}\right)$$

Complexity Bounds for DIANA and QGD

QGD: $$\mathcal{O}\left(\frac{\Delta_0}{\boxed{\varepsilon^2}} + \frac{(1+\boxed{\omega})\Delta_0^2}{\boxed{\varepsilon^4}n} + \frac{(1+\boxed{\omega})\Delta_0\Delta_f^*}{\boxed{\varepsilon^4}n}\right)$$

DIANA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\boxed{\omega})\sqrt{\boxed{\omega}/n}\right)}{\boxed{\varepsilon^2}}\right)$$

Complexity Bound for DIANA

QGD: $$\mathcal{O}\left(\frac{\Delta_0}{\varepsilon^2} + \frac{(1+\omega)\Delta_0^2}{\varepsilon^4 n} + \frac{(1+\omega)\Delta_0\Delta_f^*}{\varepsilon^4 n}\right)$$

Is it possible to get better rates?

DIANA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\omega)\sqrt{\omega/n}\right)}{\varepsilon^2}\right)$$

4. MARINA

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

DIANA: $g_i^k = h_i^k + \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$

$$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n} \sum_{i=1}^{n} g_i^k$$

DIANA: $g_i^k = h_i^k + \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$

$$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$$

MARINA: $g_i^k = \begin{cases} \nabla f_i\left(x^k\right) & \text{w.p. } p \\ g^{k-1} + \mathcal{Q}\left(\nabla f_i\left(x^k\right) - \nabla f_i\left(x^{k-1}\right)\right) & \text{w.p. } 1-p \end{cases}$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

DIANA: $g_i^k = h_i^k + \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$

$$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$$

typically small

MARINA: $g_i^k = \begin{cases} \nabla f_i\left(x^k\right) & \text{w.p. } p \\ g^{k-1} + \mathcal{Q}\left(\nabla f_i\left(x^k\right) - \nabla f_i\left(x^{k-1}\right)\right) & \text{w.p. } 1-p \end{cases}$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

DIANA: $g_i^k = h_i^k + \mathcal{Q}\left(\nabla f_i(x^k) - \boxed{h_i^k}\right)$

$$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$$

typically small

MARINA: $g_i^k = \begin{cases} \nabla f_i\left(x^k\right) & \text{w.p. } p \\ g^{k-1} + \mathcal{Q}\left(\nabla f_i\left(x^k\right) - \boxed{\nabla f_i\left(x^{k-1}\right)}\right) & \text{w.p. } 1-p \end{cases}$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

DIANA: $g_i^k = h_i^k + \boxed{\mathcal{Q}\left(\nabla f_i(x^k) - \boxed{h_i^k}\right)}$

$$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$$

MARINA: $g_i^k = \begin{cases} \boxed{\nabla f_i\left(x^k\right)} & \text{w.p. } p \\ g^{k-1} + \boxed{\mathcal{Q}\left(\nabla f_i\left(x^k\right) - \boxed{\nabla f_i\left(x^{k-1}\right)}\right)} & \text{w.p. } 1-p \end{cases}$

vectors that devices have to send

typically small

$$x^{k+1} = x^k - \gamma \cdot \boxed{\frac{1}{n}\sum_{i=1}^{n} g_i^k} = x^k - \gamma \boxed{g^k}$$

DIANA: $g_i^k = h_i^k + \boxed{\mathcal{Q}\left(\nabla f_i(x^k) - \boxed{h_i^k}\right)}$

$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right)$

vectors that devices have to send

typically small

MARINA: $g_i^k = \begin{cases} \boxed{\nabla f_i\left(x^k\right)} & \text{w.p. } p \\ g^{k-1} + \boxed{\mathcal{Q}\left(\nabla f_i\left(x^k\right) - \boxed{\nabla f_i\left(x^{k-1}\right)}\right)} & \text{w.p. } 1-p \end{cases}$

$$x^{k+1} = x^k - \gamma \cdot \boxed{\frac{1}{n}\sum_{i=1}^{n} g_i^k} = x^k - \gamma \boxed{g^k}$$

DIANA: $g_i^k = h_i^k + \boxed{\mathcal{Q}\left(\nabla f_i(x^k) - \boxed{h_i^k}\right)}$

vectors that devices have to send

$$h_i^{k+1} = h_i^k + \alpha \mathcal{Q}\left(\nabla f_i(x^k) - h_i^k\right) \quad \mathbb{E}\left[\boxed{g^k} \mid x^k\right] = \nabla f(x^k)$$

typically small

MARINA: $g_i^k = \begin{cases} \boxed{\nabla f_i\left(x^k\right)} & \text{w.p. } p \\ g^{k-1} + \boxed{\mathcal{Q}\left(\nabla f_i\left(x^k\right) - \boxed{\nabla f_i\left(x^{k-1}\right)}\right)} & \text{w.p. } 1-p \end{cases}$

$$\mathbb{E}\left[\boxed{g^k} \mid x^k\right] \neq \nabla f(x^k)$$

Complexity Bound for MARINA

MARINA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

Complexity Bound for MARINA

MARINA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$ after

$$\mathcal{O}\left(\frac{\Delta_0 \left(1 + \omega/\sqrt{n}\right)}{\varepsilon^2}\right)$$

communication rounds

Complexity Bound for MARINA

MARINA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\Delta_0\left(1+\omega/\sqrt{n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

Complexity Bound for MARINA

MARINA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\Delta_0\left(1+\boxed{\omega}/\sqrt{n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x)-x\|^2 \leq \omega\|x\|^2}$$

Complexity Bound for MARINA

MARINA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\boxed{\Delta_0}\left(1+\boxed{\omega}/\sqrt{n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x)-x\|^2 \leq \omega\|x\|^2}$$

$$\boxed{\Delta_0 = f(x^0) - f_*}$$

Complexity Bound for MARINA

MARINA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\boxed{\Delta_0}\left(1+\boxed{\omega}/\sqrt{n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x)-x\|^2 \leq \omega\|x\|^2}$$

$$\boxed{\Delta_0 = f(x^0) - f_*}$$

$$p = \frac{1}{\omega+1} = \Theta\left(\frac{\zeta_{\mathcal{Q}}}{d}\right)$$

Complexity Bound for MARINA

MARINA finds such $\hat{x}$ that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \boxed{\varepsilon^2}$ after

Hides numerical factors and smoothness constants

$$\mathcal{O}\left(\frac{\boxed{\Delta_0}\left(1+\boxed{\omega}/\sqrt{n}\right)}{\boxed{\varepsilon^2}}\right)$$

communication rounds

$$\boxed{\mathbb{E}\|\mathcal{Q}(x)-x\|^2 \leq \omega\|x\|^2}$$

$$\boxed{\Delta_0 = f(x^0) - f_*}$$

$$p = \frac{1}{\omega+1} = \Theta\left(\frac{\boxed{\zeta_{\mathcal{Q}}}}{d}\right)$$

$$\boxed{\zeta_{\mathcal{Q}}} = \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\|\mathcal{Q}(x)\|_0\right]$$

assumption (holds for RandK, l2-quantization)

expected density

Complexity Bounds for MARINA and DIANA

DIANA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\omega)\sqrt{\omega/n}\right)}{\varepsilon^2}\right)$$

MARINA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+\omega/\sqrt{n}\right)}{\varepsilon^2}\right)$$

Complexity Bounds for MARINA and DIANA

DIANA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+(1+\boxed{\omega})\sqrt{\boxed{\omega}/n}\right)}{\varepsilon^2}\right)$$

MARINA: $$\mathcal{O}\left(\frac{\Delta_0\left(1+\boxed{\omega}/\sqrt{n}\right)}{\varepsilon^2}\right)$$

5. Experiments

The Problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{N} \sum_{t=1}^{N} \left(1 - \frac{1}{1 + \exp(-y_t a_t^\top x)} \right)^2 \right\}$$

The Problem

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\frac{1}{N}\sum_{t=1}^{N}\left(1-\frac{1}{1+\exp(-y_t a_t^\top x)}\right)^2\right\}$$

$y_t \in \{-1,1\}$ $\qquad$ $a_t \in \mathbb{R}^d$

The Problem

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\frac{1}{N}\sum_{t=1}^{N}\left(1-\frac{1}{1+\exp(-y_t a_t^\top x)}\right)^2\right\}$$

$$y_t \in \{-1,1\} \qquad a_t \in \mathbb{R}^d$$

- The dataset was split into 5 equal parts among 5 clients

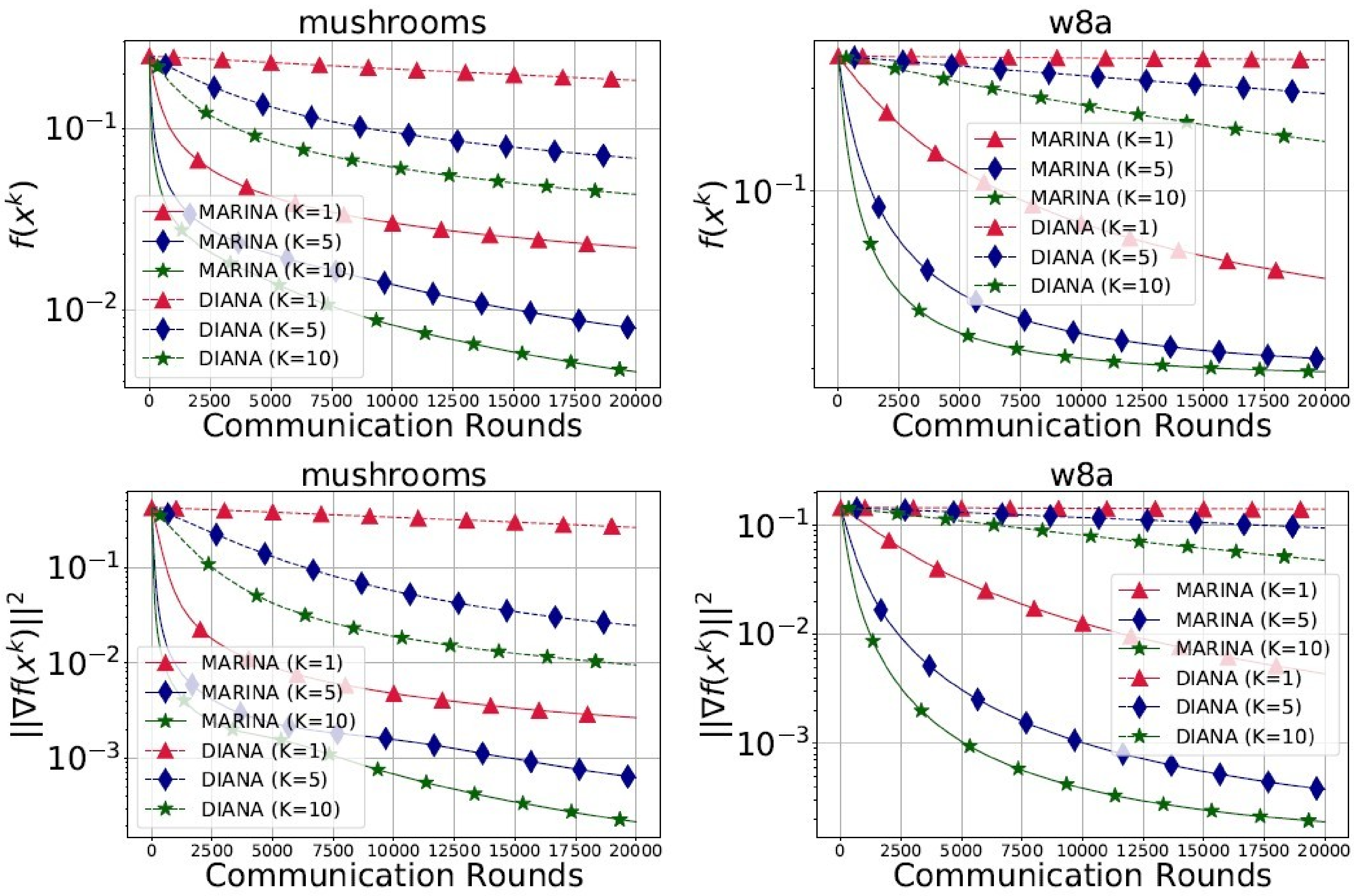
The Problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{N} \sum_{t=1}^{N} \left(1 - \frac{1}{1 + \exp(-y_t a_t^\top x)} \right)^2 \right\}$$

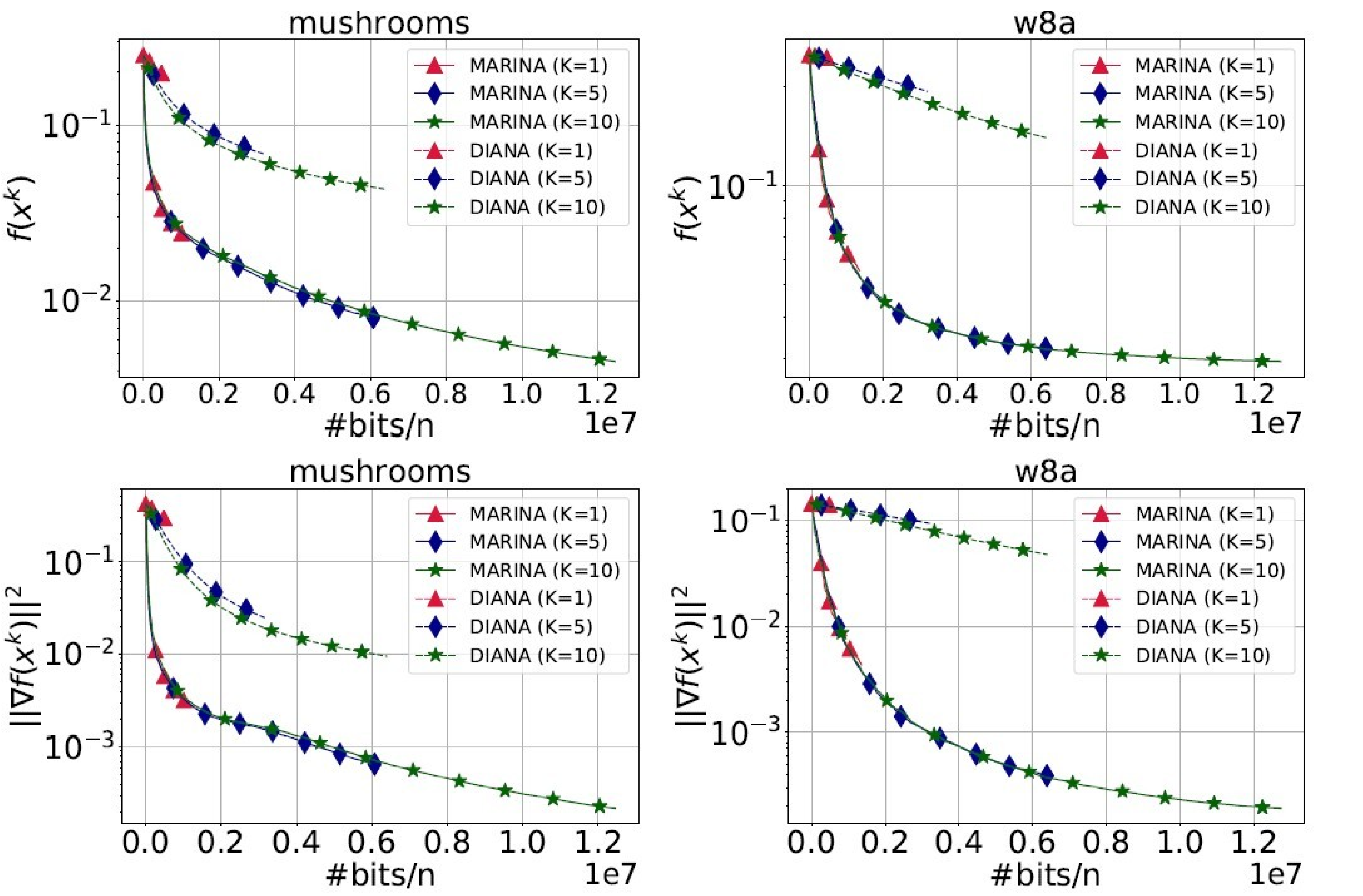
$$y_t \in \{-1, 1\} \qquad a_t \in \mathbb{R}^d$$

- The dataset was split into 5 equal parts among 5 clients
- Theoretical stepsizes

MARINA vs DIANA (RandK)



MARINA vs DIANA (RandK)



6. Extra Results

In the paper, we also have:

- Variance Reduced MARINA for the problems with $f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$

In the paper, we also have:

- Variance Reduced MARINA for the problems with $f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$
- Variance Reduced MARINA for the problems with $f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i}\left[f_{\xi_i}(x)\right]$

In the paper, we also have:

- Variance Reduced MARINA for the problems with $f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$
- Variance Reduced MARINA for the problems with $f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i}\left[f_{\xi_i}(x)\right]$
- Rates under Polyak- Lojasiewicz Condition

In the paper, we also have:

- Variance Reduced MARINA for the problems with $f_i(x) = \frac{1}{m}\sum\limits_{j=1}^{m} f_{ij}(x)$
- Variance Reduced MARINA for the problems with $f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i}\left[f_{\xi_i}(x)\right]$
- Rates under Polyak- Lojasiewicz Condition
- Explicit dependencies on smoothness constants, non-uniform sampling

In the paper, we also have:

- Variance Reduced MARINA for the problems with $f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$
- Variance Reduced MARINA for the problems with $f_i(x) = \mathbf{E}_{\xi_i \sim \mathcal{D}_i}\left[f_{\xi_i}(x)\right]$
- Rates under Polyak- Lojasiewicz Condition
- Explicit dependencies on smoothness constants, non-uniform sampling
- Simple proofs

https://arxiv.org/pdf/2102.07845.pdf