MARINA: Faster Non-Convex Distributed Learning with Compression

Eduard Gorbunov
MIPT, Yandex.Research

Konstantin Burlachenko
KAUST

Zhize Li
KAUST

Peter Richtárik
KAUST

ICML
International Conference On Machine Learning

Research

MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

KAUST

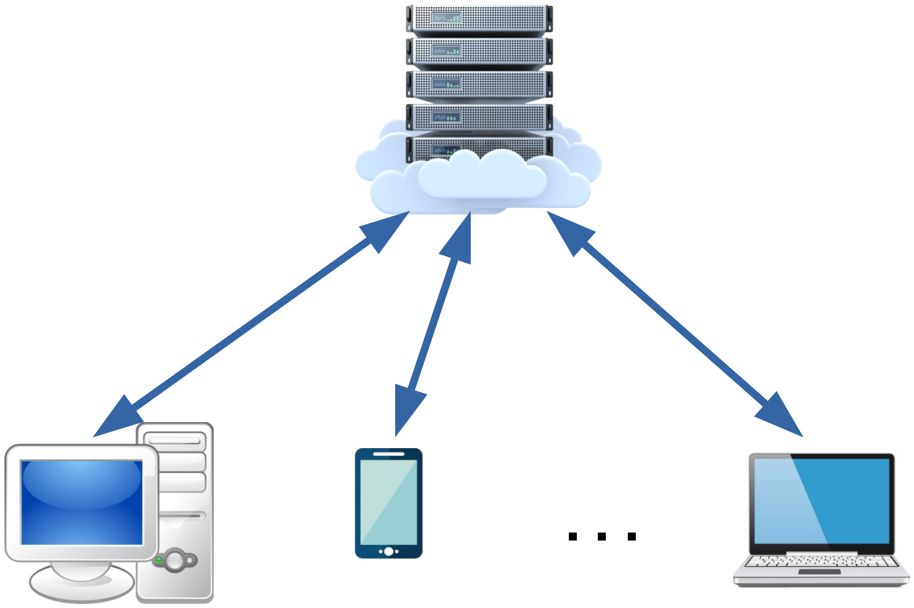
July 18-24, 2021

Konstantin Burlachenko
PhD student
KAUST

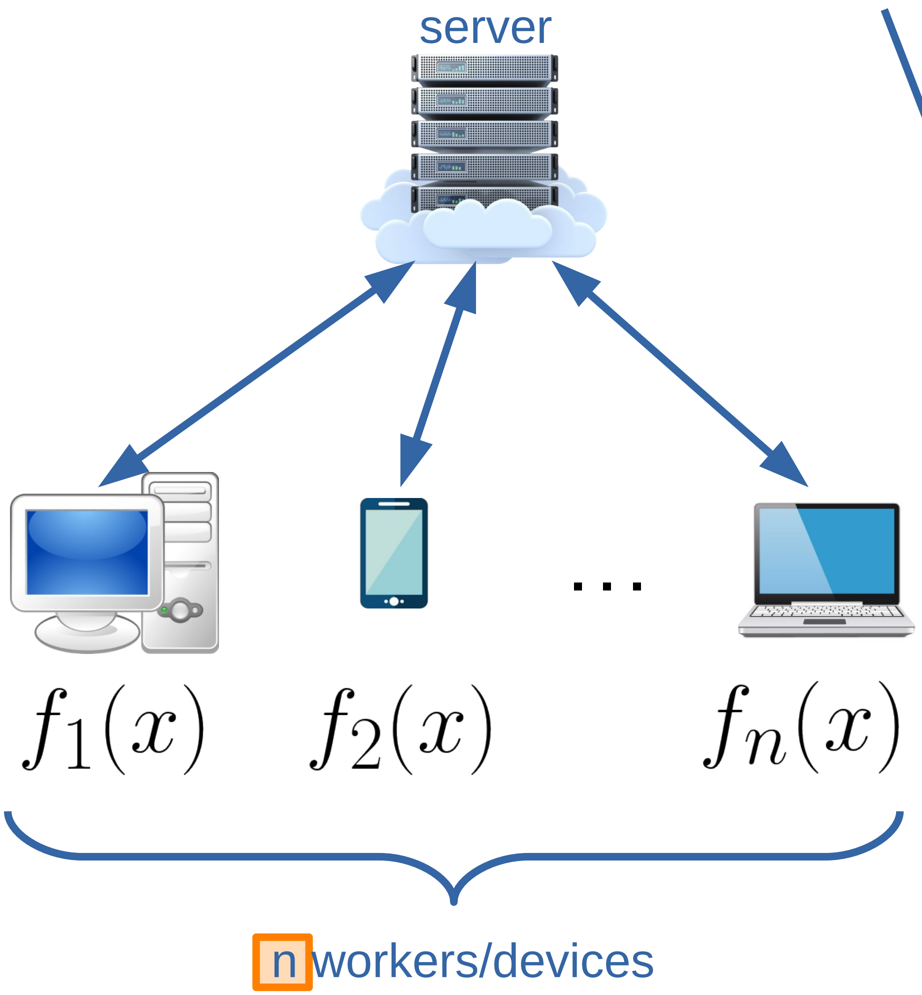
Zhize Li
Research Scientist
KAUST

Peter Richtárik
Professor of Computer Science
KAUST

1. The Problem



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



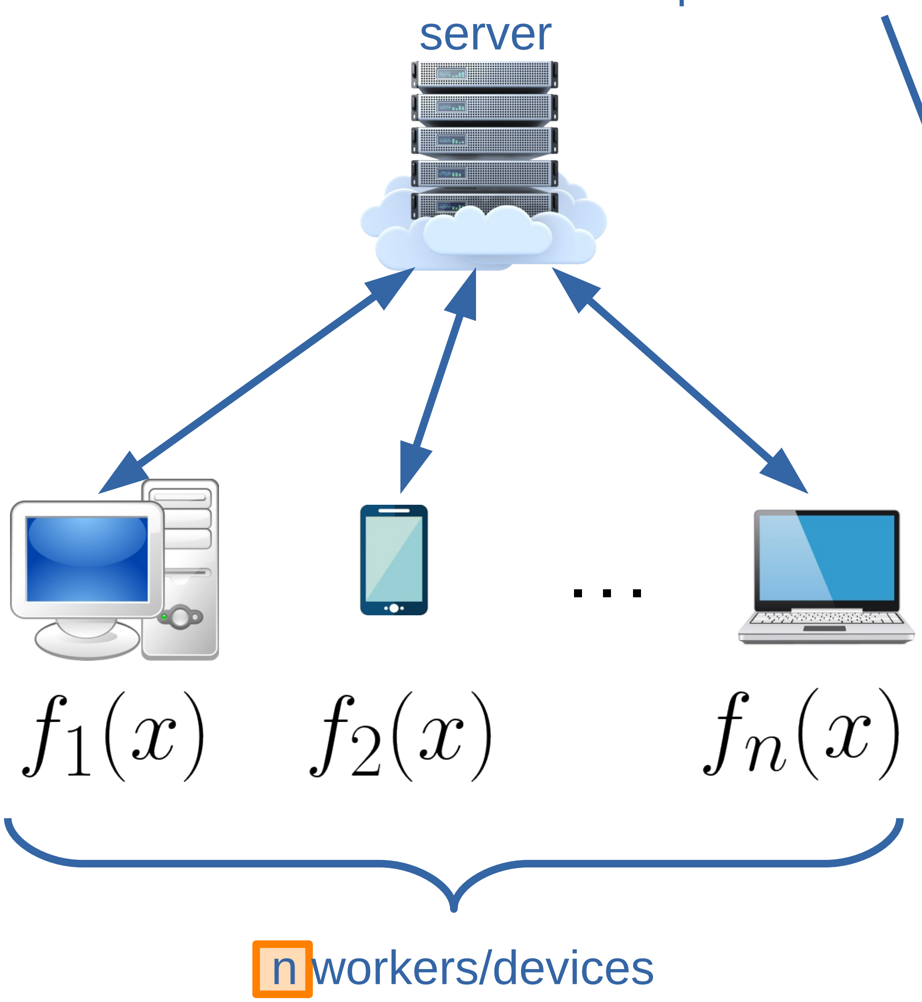
parameters of the model

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

# of parameters

# of workers/devices

loss on the data accessible by worker *i*



parameters of the model

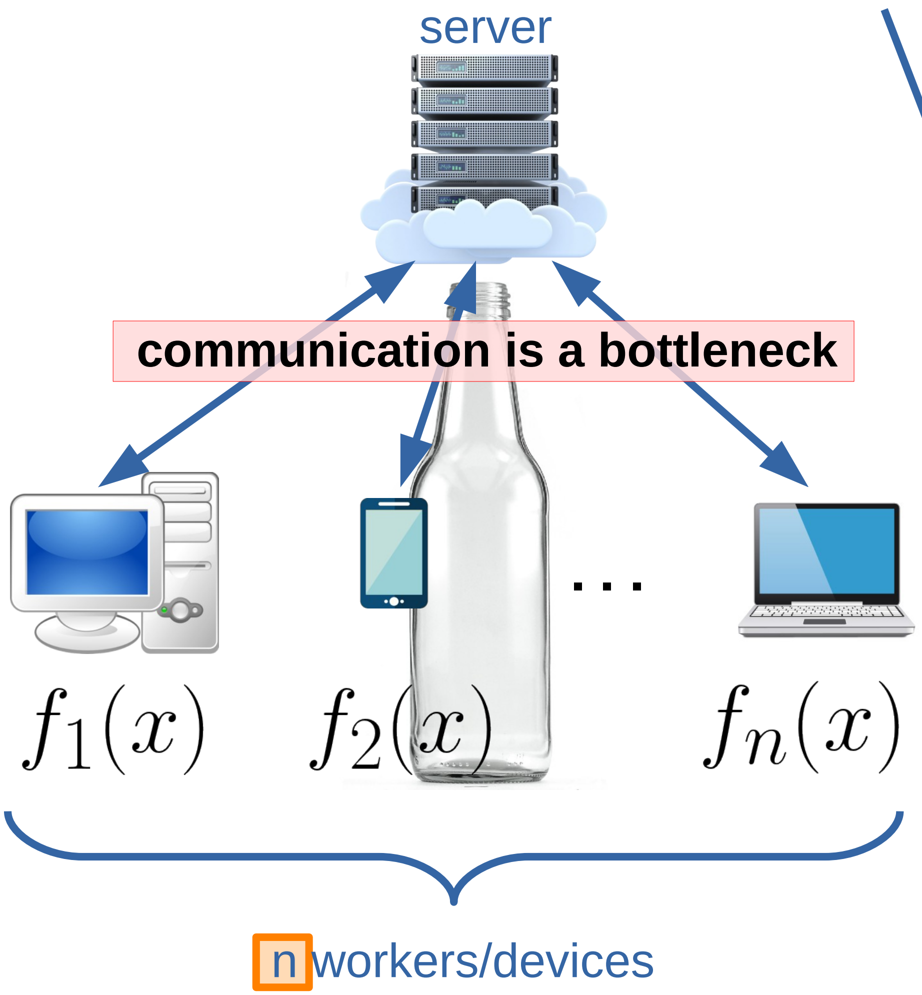
$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

# of parameters

# 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

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

# 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?

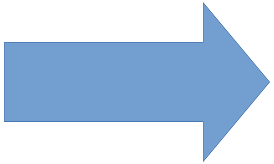
One of the possible solutions: **send less information at communication rounds**

How to Handle Communication Bottleneck?

One of the possible solutions: **send less information at communication rounds**

Workers send dense vectors

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



How to Handle Communication Bottleneck?

One of the possible solutions: **send less information at communication rounds**

Workers send dense vectors

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

Workers send compressed/sparse vectors

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

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)

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

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

Example: RandK (for K = 2)

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

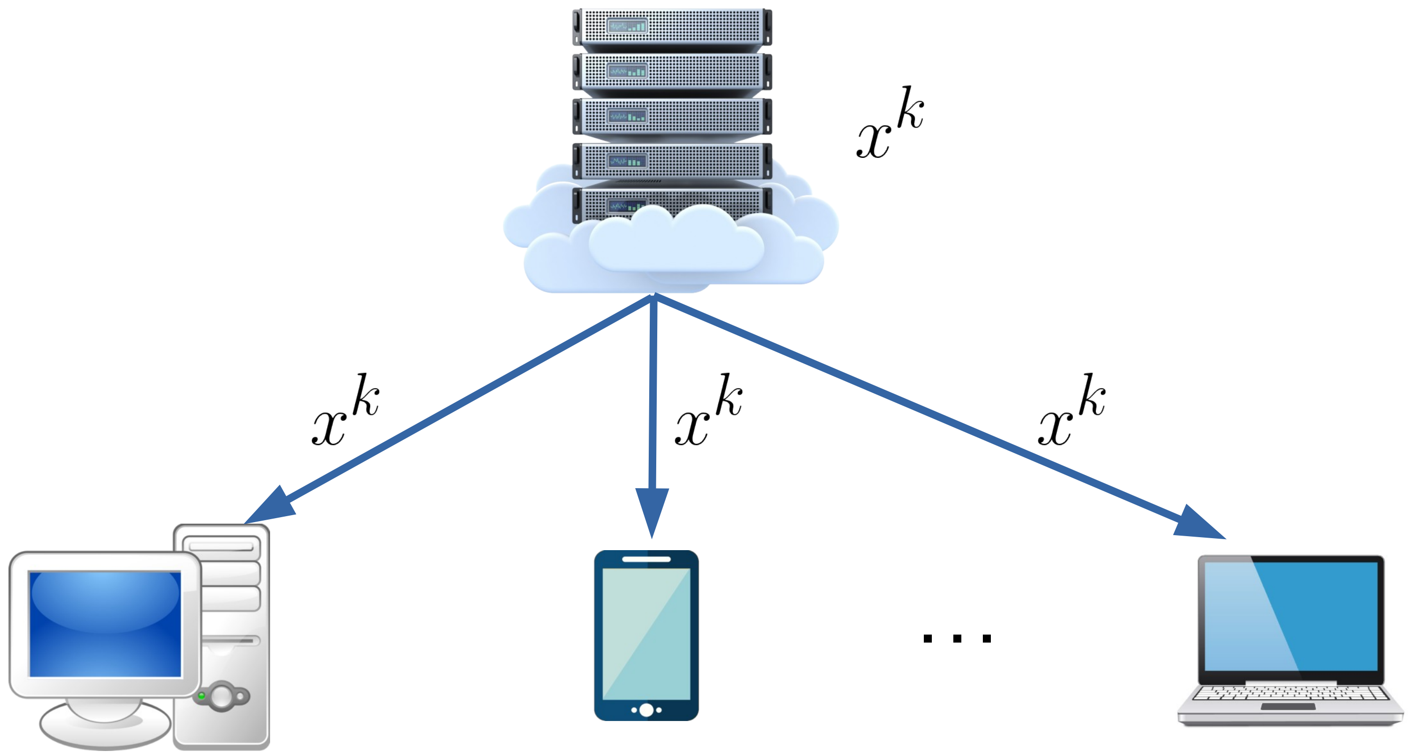
for unbiasedness

Pick K = 2 components uniformly at random

2. Quantized Gradient Descent (QGD)

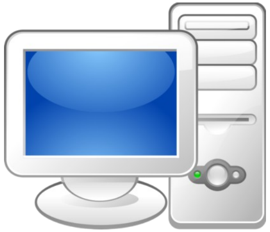
Alistarh, Dan, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. **"QSGD: Communication-efficient SGD via gradient quantization and encoding."** *In Advances in Neural Information Processing Systems*, pp. 1709-1720. 2017.

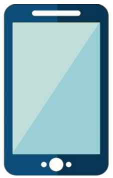
1 Server broadcasts the parameters



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

x^k



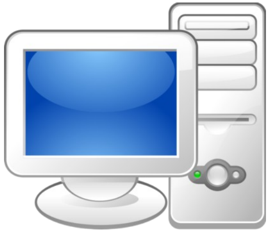


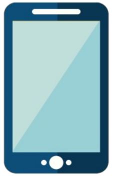
…

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

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

x^k



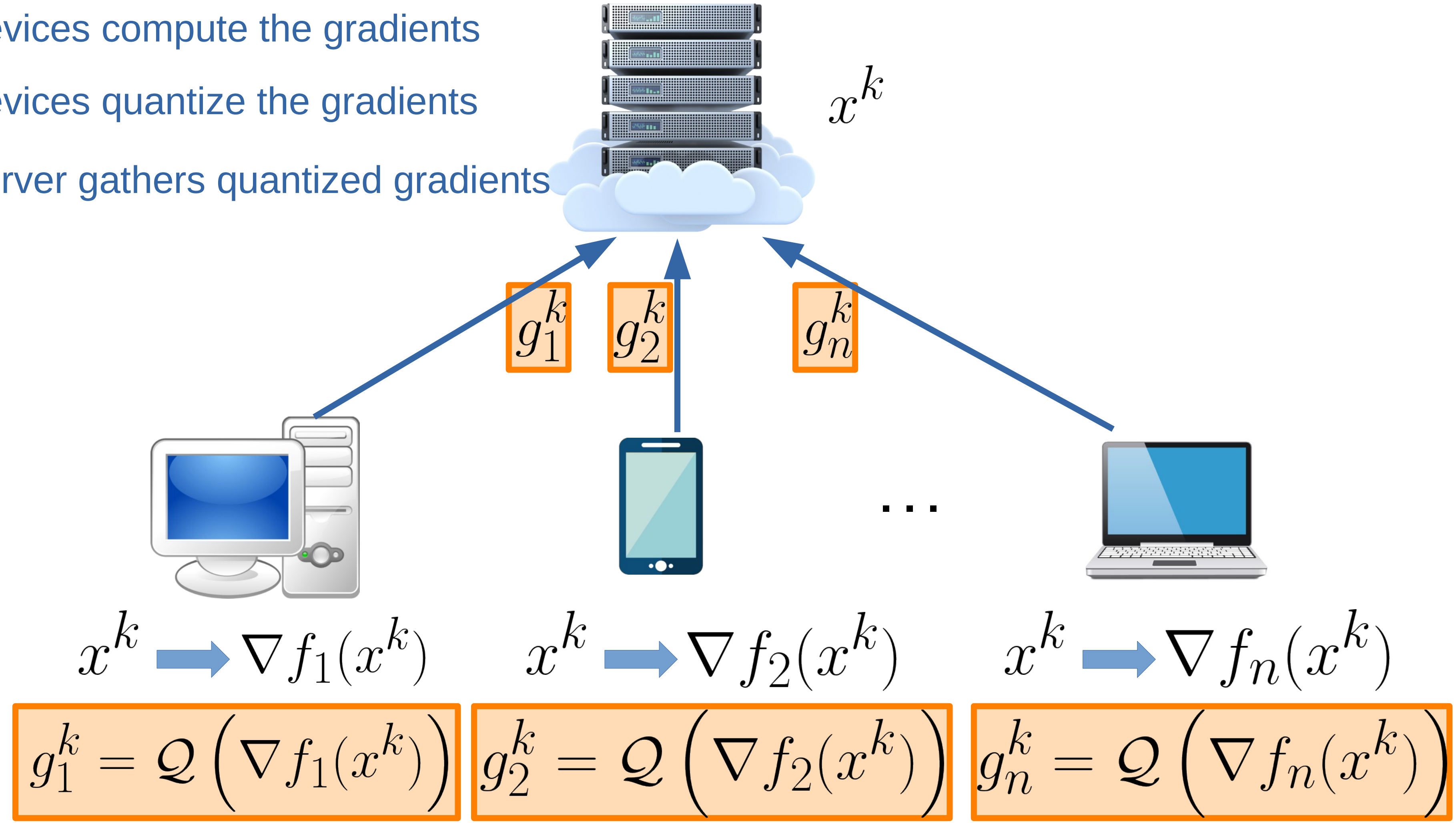


$\cdots$

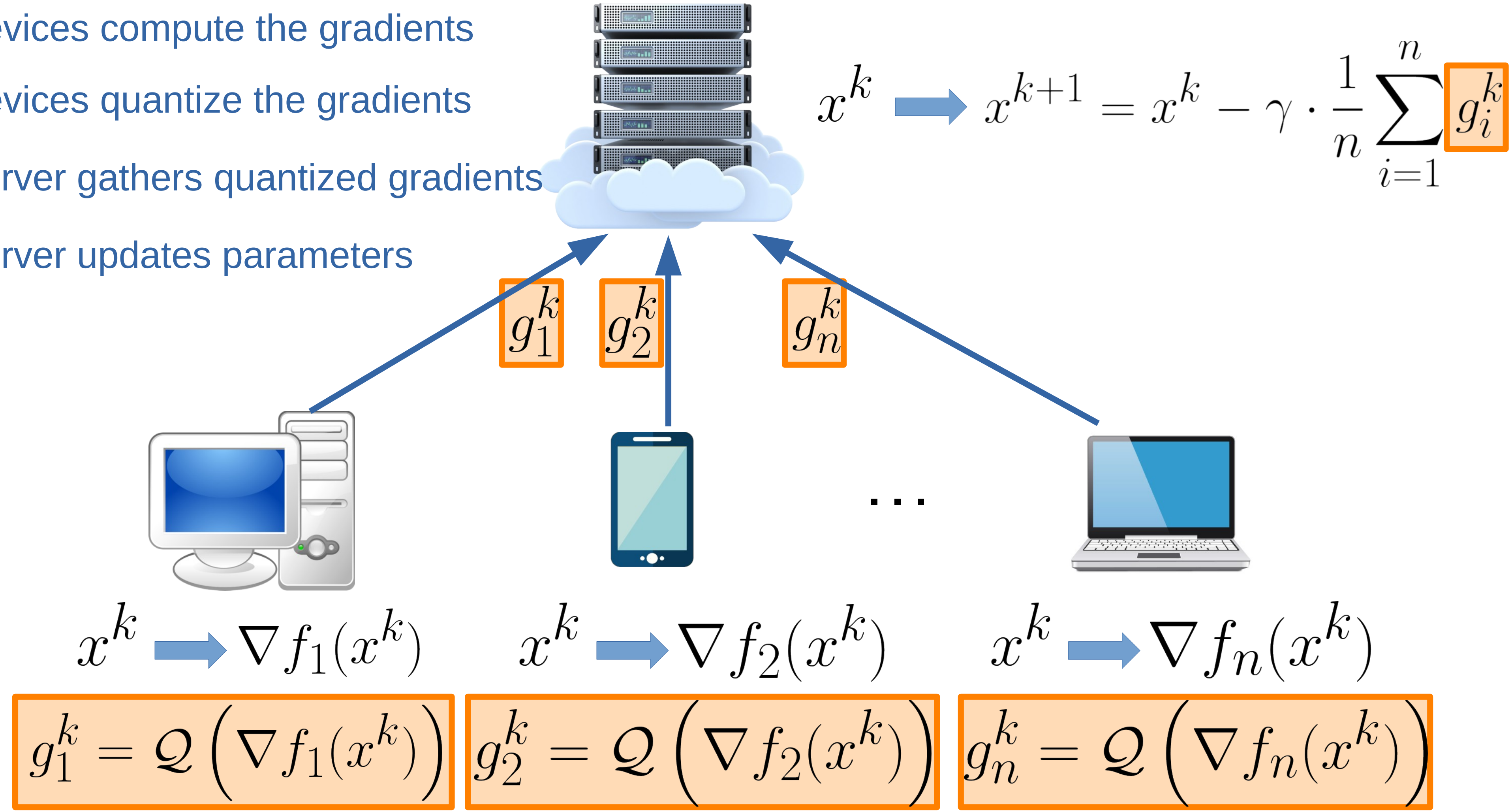
$x^k \rightarrow \nabla f_1(x^k)$ $x^k \rightarrow \nabla f_2(x^k)$ $x^k \rightarrow \nabla f_n(x^k)$

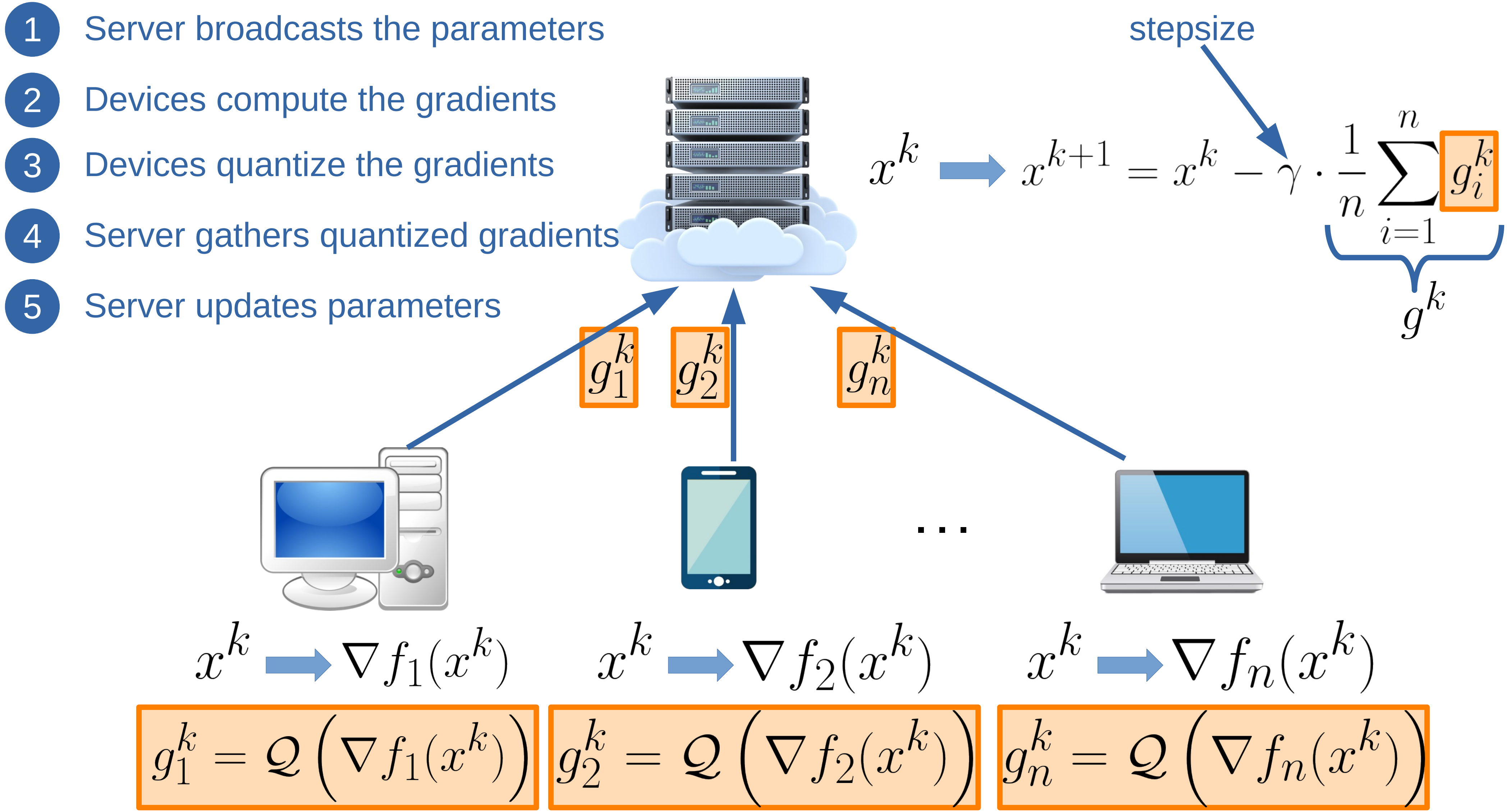
$g_1^k = \mathcal{Q}\left(\nabla f_1(x^k)\right)$ $g_2^k = \mathcal{Q}\left(\nabla f_2(x^k)\right)$ $g_n^k = \mathcal{Q}\left(\nabla f_n(x^k)\right)$

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

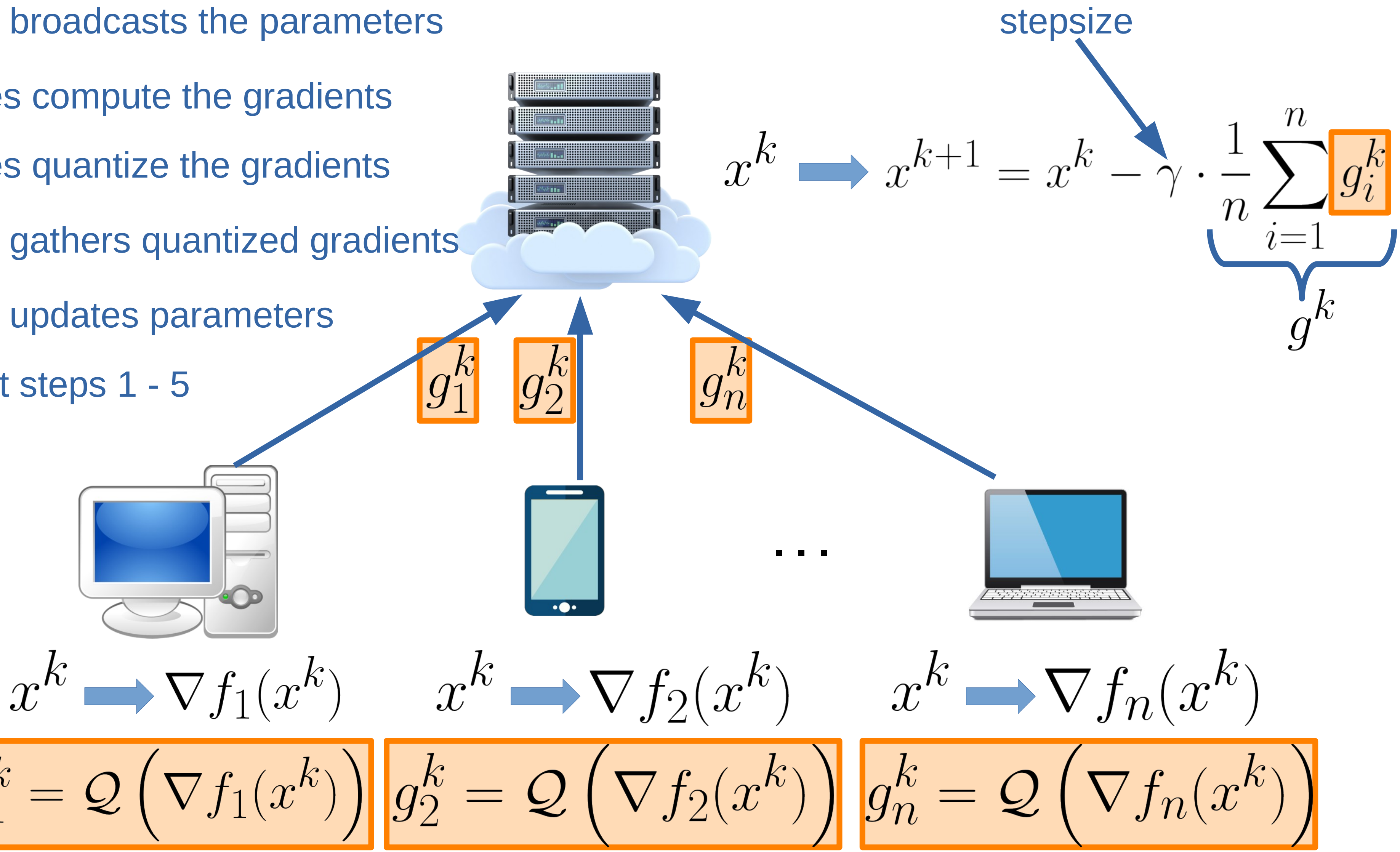


1. Server broadcasts the parameters
2. Devices compute the gradients
3. Devices quantize the gradients
4. Server gathers quantized gradients
5. Server updates parameters




1 Server broadcasts the parameters
2 Devices compute the gradients
3 Devices quantize the gradients
4 Server gathers quantized gradients
5 Server updates parameters
stepsize
$x^k \rightarrow x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} g_i^k$
g^k
g_1^k
g_2^k
g_n^k
$\cdots$
$x^k \rightarrow \nabla f_1(x^k)$
$x^k \rightarrow \nabla f_2(x^k)$
$x^k \rightarrow \nabla f_n(x^k)$
$g_1^k = \mathcal{Q}\left(\nabla f_1(x^k)\right)$
$g_2^k = \mathcal{Q}\left(\nabla f_2(x^k)\right)$
$g_n^k = \mathcal{Q}\left(\nabla f_n(x^k)\right)$

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

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

Hides numerical factors and smoothness constants

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

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

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

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

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

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

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

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

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

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

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

Not optimal!

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

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

learnable local shifts

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

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

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

vectors that devices have to send

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

learnable local shifts

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

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 Bound for DIANA

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

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

Is it possible to get better rates?

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

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

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

vectors that devices have to send

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

typically small

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

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

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

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

vectors that devices have to send

typically small

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

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

Complexity Bound for MARINA

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

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

communication rounds

Hides numerical factors and smoothness constants

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

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

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

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

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

5. Extra Results

In the paper, we also have:

- Variance Reduced MARINA (uses stochastic gradients instead of full gradients)
- MARINA with partial participation of clients
- Rates under Polyak- Lojasiewicz Condition
- Explicit dependencies on smoothness constants, non-uniform sampling
- Simple proofs
- Numerical experiments with generalized linear models and neural networks

If you have any questions, feel free to write me on my email eduard.gorbunov@phystech.edu
…or just find me at the poster session :-)