Rising Stars in AI Symposium 2022

Shifted Compression Framework for Distributed Learning

Egor Shulgin
Optimization and Machine Learning Lab

Joint work with Peter Richtárik

Distributed Learning

% of training time spent in communication

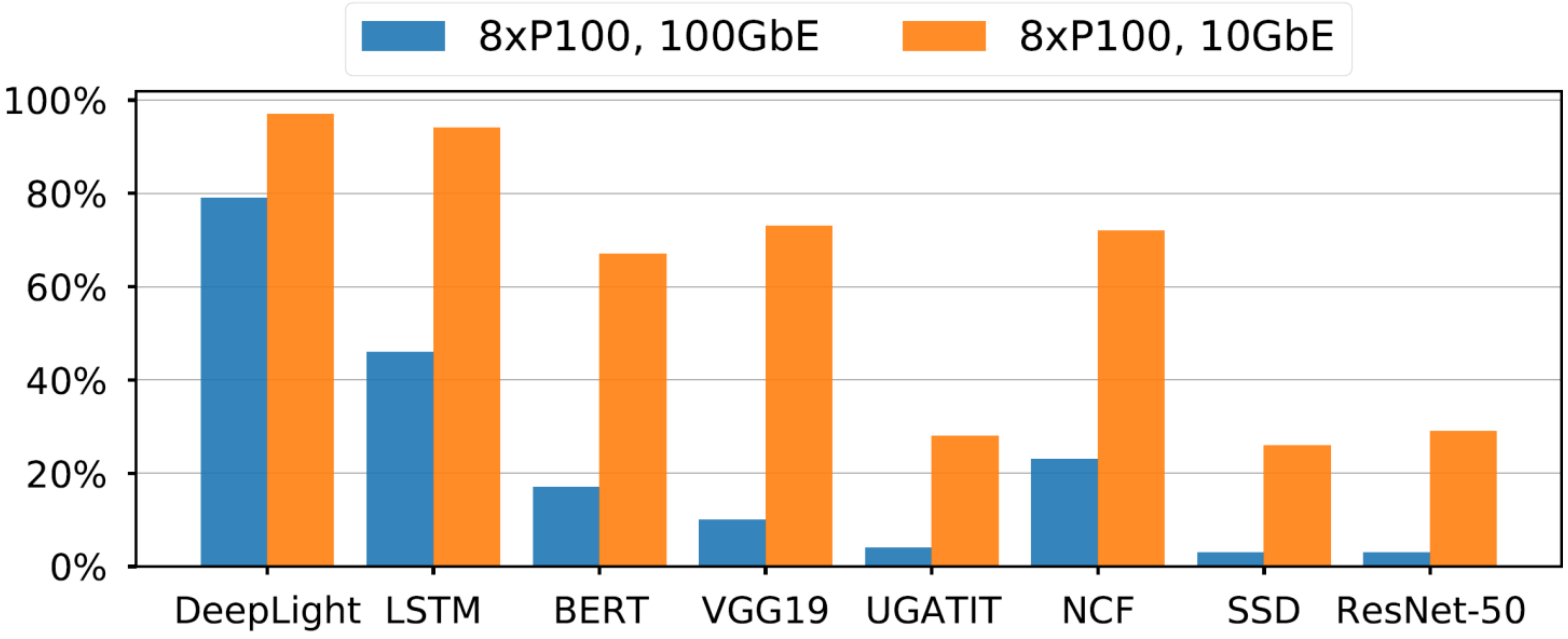


Image credit to Sapio et al., NSDI '21 presentation

Problem Formulation

Model weights

Number of nodes

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

Number of features

Loss on local data of node i

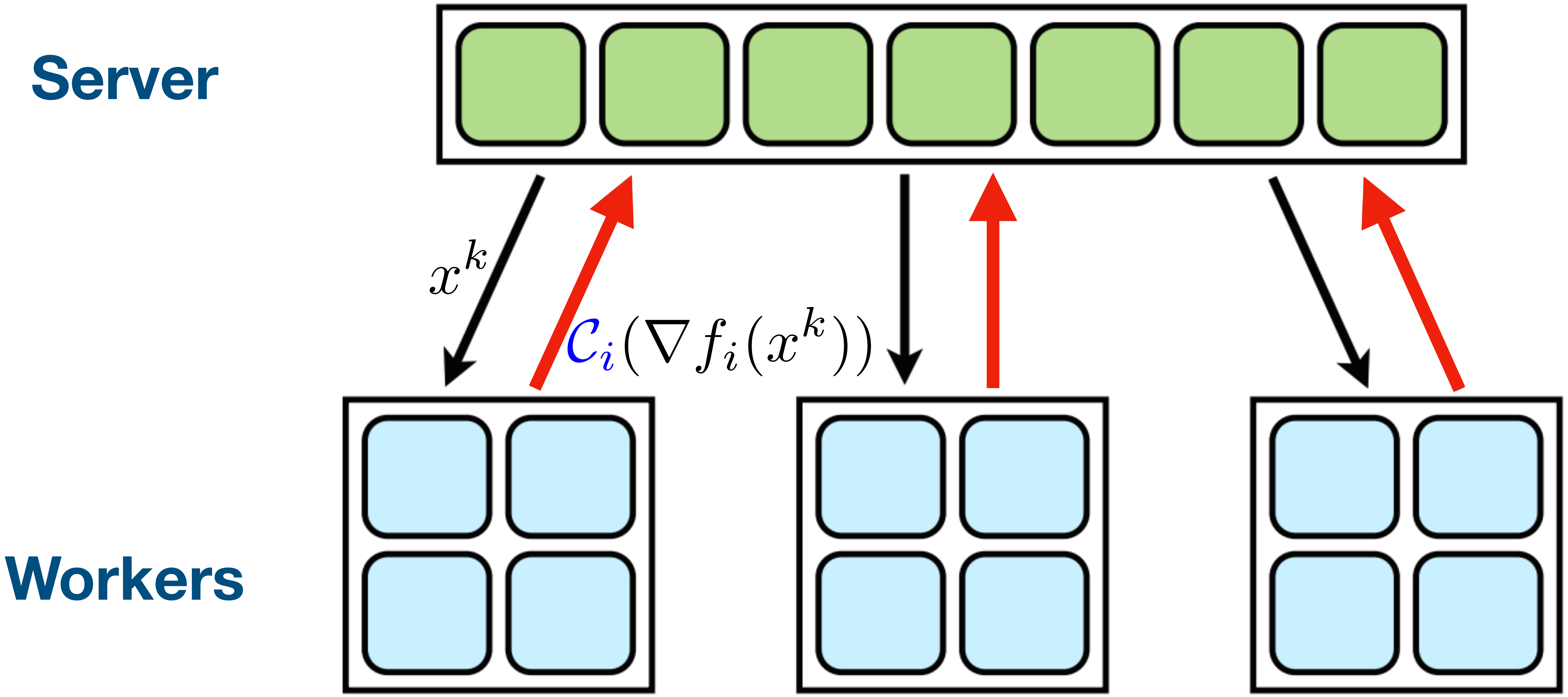
Assumptions:

- μ-strong convexity
- L-smoothness

Distributed Learning

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

$$x^{k+1} = x^k - \frac{\gamma}{n} \sum_{i=1}^{n} \mathcal{C}_i(\nabla f_i(x^k))$$



Communication is the Bottleneck

Distributed Compressed Gradient Descent (DCGD) scheme

Solution: compress the transmitted updates

Compression Operators $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$

Contractive

$$\mathbf{E}\|\mathcal{C}(x) - x\|^2 \leq (1-\delta)\|x\|^2$$

Top-K (for K=2)

$$\begin{bmatrix} 2 \\ -1 \\ 5 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 \\ 0 \\ 5 \end{bmatrix}$$

Picks components with largest absolute value

Unbiased

$$\mathbf{E}\mathcal{Q}(x) = x, \quad \mathbf{E}\|\mathcal{Q}(x) - x\|^2 \leq \omega\|x\|^2$$

Rand-K (for K=2)

$$\begin{bmatrix} 2 \\ -1 \\ 5 \end{bmatrix} \Rightarrow \frac{3}{2} \cdot \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}$$

Picks components uniformly at random

Convergence of DCGD

Expected distance to the solution

Problem

$$\mathbf{E}\left\|x^k - x^\star\right\|^2 \leq (1-\gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{2\gamma\omega}{\mu n} \cdot \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^\star)\right\|^2$$

Linear convergence term

Neighborhood term
(due to compression)

Communication complexity:
(in the interpolation regime $\nabla f_i(x^\star) = 0$)

$$\tilde{O}\left(\kappa\left(1 + \frac{\omega}{n}\right)\right)$$

Comes from compression

Condition number: $\kappa = L/\mu$

Shifted Compression Solution

Shifted compressor: $\mathbf{E}\mathcal{Q}_h(x) = x, \quad \mathbf{E}\|\mathcal{Q}_h(x) - x\|^2 \leq \omega\|x - \mathbf{h}\|^2$

shift vector

Any $\mathcal{Q}_h$ arises by a shift of unbiased operator $\mathcal{Q}$: $\mathcal{Q}_h(x) = h + \mathcal{Q}(x - h)$

Method: $$x^{k+1} = x^k - \gamma \frac{1}{n}\sum_{i=1}^{n}\left[h_i^k + \mathcal{Q}(\nabla f_i(x) - h_i^k)\right]$$

$$\mathbf{E}\left\|x^k - x^\star\right\|^2 \leq (1 - \gamma\mu)^k \left\|x^0 - x^\star\right\|^2 + \frac{2\gamma\omega}{\mu n} \cdot \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^\star) - h_i\|^2$$

Neighborhood term

Imaginary situation: we know optimal shifts $h_i^\star = \nabla f_i(x^\star)$

Practical Solution

Goal: learn the optimal shifts: $h_i^k \to \nabla f_i(x^\star)$

Via loopless mechanism: $$h_i^{k+1} = \begin{cases} \nabla f_i(x^k) & \text{with probability } p \\ h_i^k & \text{with probability } 1-p \end{cases}$$

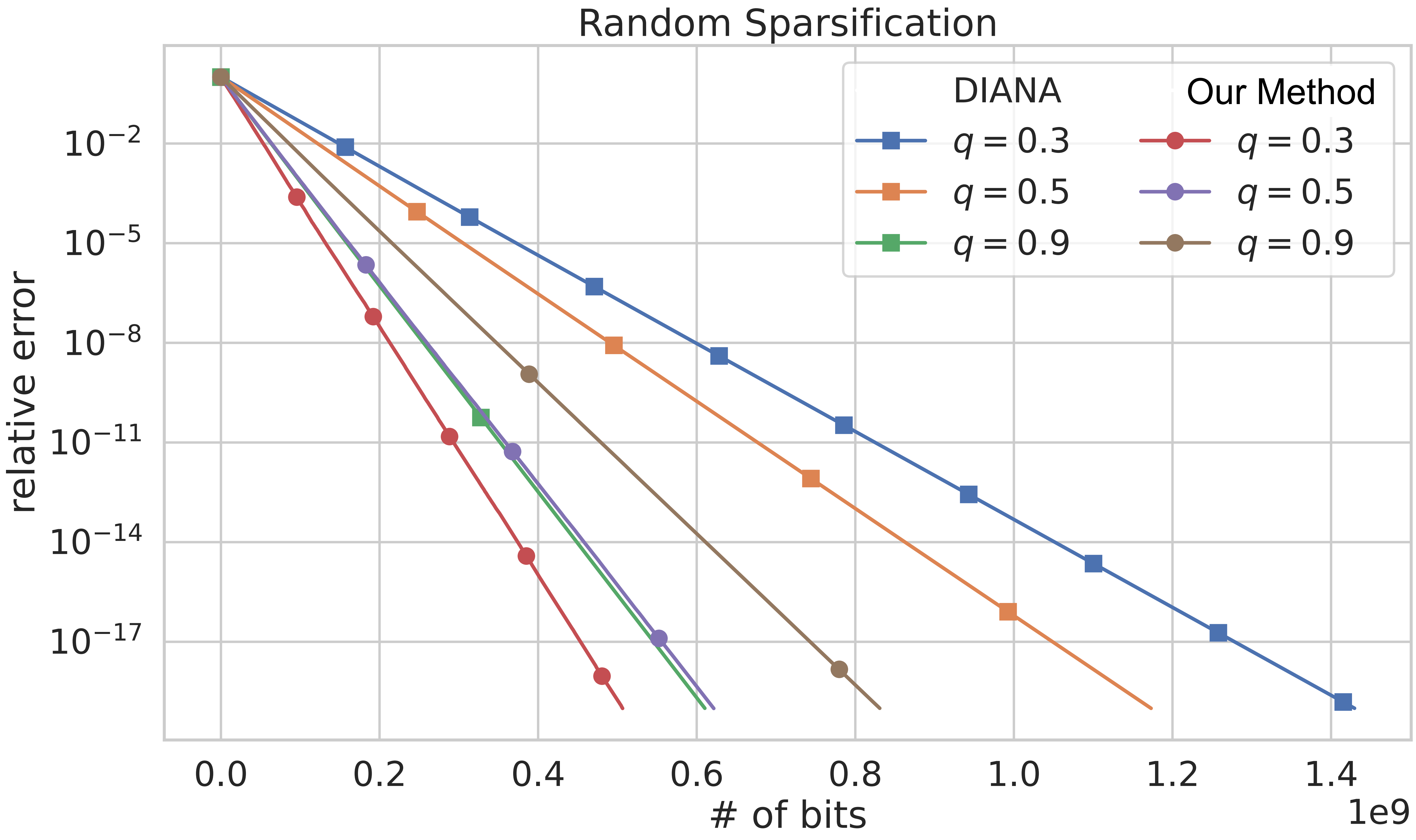
Convergence result: $$\mathbf{E}V^k \leq \max\left\{(1-\gamma\mu)^k, \left(1-p+\frac{2\omega}{nM}\right)^k\right\} V^0$$

Lyapunov function: $$V^k = \left\|x^k - x^\star\right\|^2 + \omega M \gamma^2 \cdot \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \nabla f_i(x^\star)\right\|^2$$

Communication complexity: $$\tilde{O}\left(\max\left\{\kappa\left(1+\frac{\omega}{n}\right), \frac{1}{p}\right\}\right)$$

Empirical perfomance

Numerical results for Regularized Logistic Regression



$q = k/d$

Comparison of DIANA and our Algorithm

Contributions Summary

- **Generalizations** of existing distributed methods to allow using both biased and unbiased compressors

- **Improved rates** for methods with compressed iterates with and without variance-reduction

$$\kappa^{2}\left(1+\frac{\omega}{n}\right) \to \kappa\left(1+\frac{\omega}{n}\right)$$

- **New** loopless algorithm with simpler approach to reduction of variance coming from compression

Any Questions?

More details
in the paper

Contacts
egor.shulgin@kaust.edu.sa

Egor Shulgin
shulgin-egor.github.io