北京大学
PEKING UNIVERSITY

Lower Bounds and Accelerated Algorithms in Distributed Optimization with Communication Compression

Kun Yuan (袁 坤)

Center for Machine Learning Research @ Peking University

Nov 23, 2023

Joint work with

Yutong He
(Peking U.)

Xinmeng Huang
(UPenn)

Yiming Chen
(Alibaba)

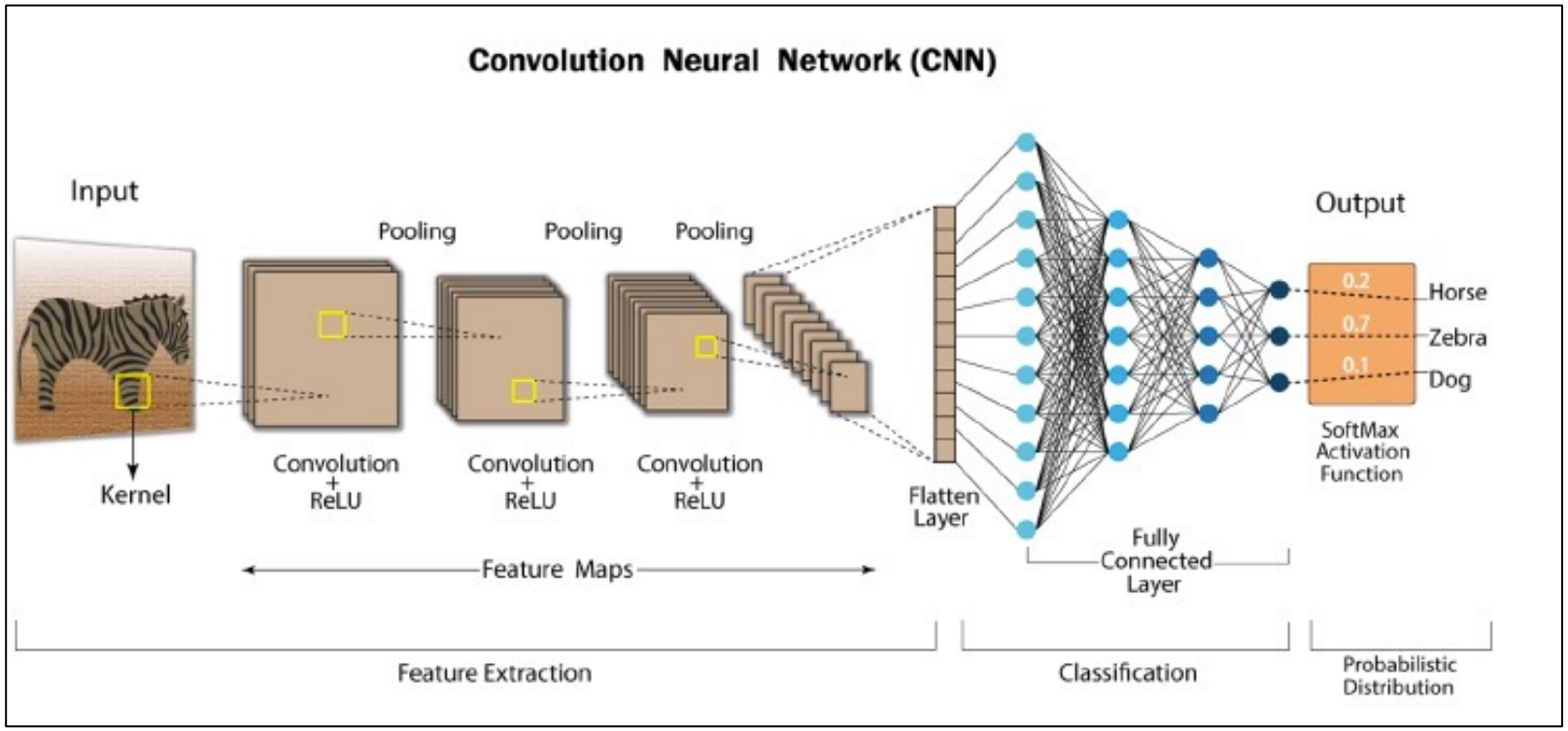
Wotao Yin
(Alibaba)

北京大学
PEKING UNIVERSITY

PART 01

Basics and Motivation

Training deep neural network is notoriously difficult



DNN training = non-convexity + **massive dataset** + huge models

Distributed learning

北京大学 PEKING UNIVERSITY

- Training deep neural networks typically requires **massive** datasets; efficient and scalable distributed optimization algorithms are in urgent need

- A network of n nodes (devices such as GPUs) collaborate to solve the problem:

$$\min_{x\in\mathbb{R}^d} \quad f(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad \text{where} \quad f_i(x)=\mathbb{E}_{\xi_i\sim D_i} F(x;\xi_i)$$

  - Each component $f_i : \mathbb{R}^d \to \mathbb{R}$ is local and private to node i
  - Random variable ξ_i denotes the local data that follows distribution D_i
  - Each local distribution D_i is different; data heterogeneity exists

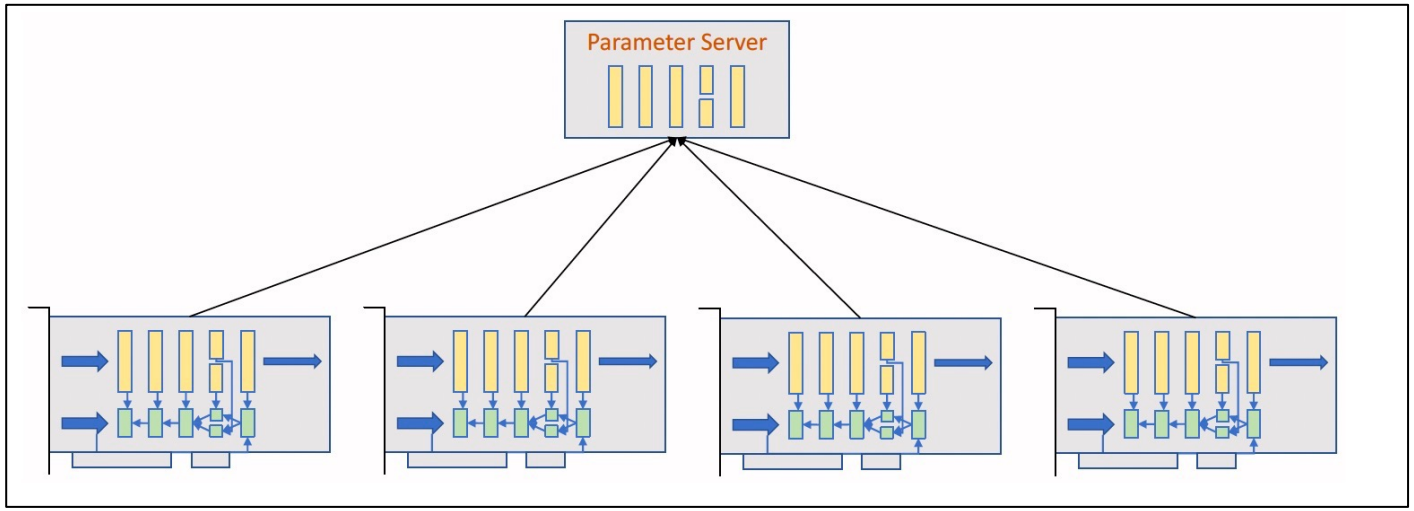
Vanilla parallel stochastic gradient descent (PSGD)

$$g_i^{(k)} = \nabla F(x^{(k)}; \xi_i^{(k)}) \qquad \text{(Local compt.)}$$

$$x^{(k+1)} = x^{(k)} - \frac{\gamma}{n} \sum_{i=1}^{n} g_i^{(k)} \qquad \text{(Global comm.)}$$

- Each node i samples data $\xi_i^{(k)}$ and computes gradient $\nabla F(x^{(k)}; \xi_i^{(k)})$
- All nodes synchronize (i.e. globally average) to update model x per iteration

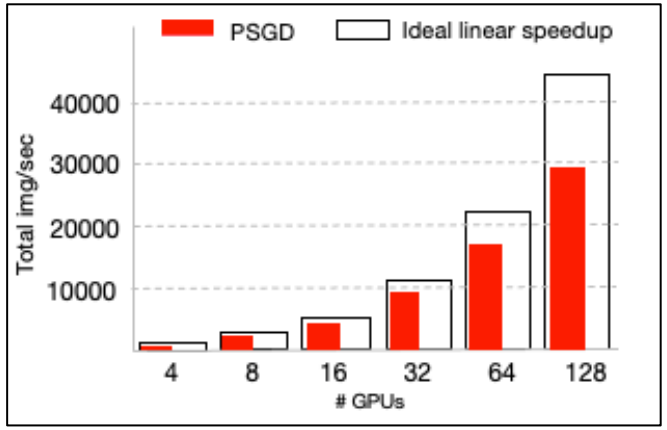
Vanilla parallel stochastic gradient descent (PSGD)



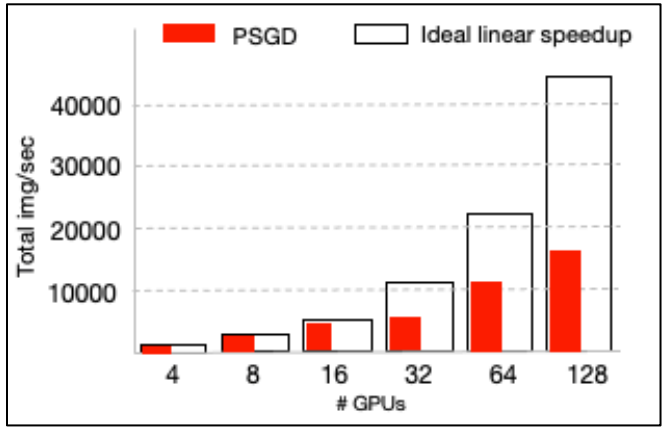
- Each node sends a full model (or gradient) to the server; proportional to dimension d
- When model dimension d is large, PSGD suffers severe communication overhead

PSGD cannot achieve linear speedup due to comm. overhead

- PSGD cannot achieve ideal linear speedup in throughput due to comm. overhead
- Larger comm-to-compt ratio leads to worse performance in PSGD



Small comm.-to-compt. ratio



Large comm.-to-compt. ratio

- How can we accelerate PSGD? **Distributed optimization with communication compression**

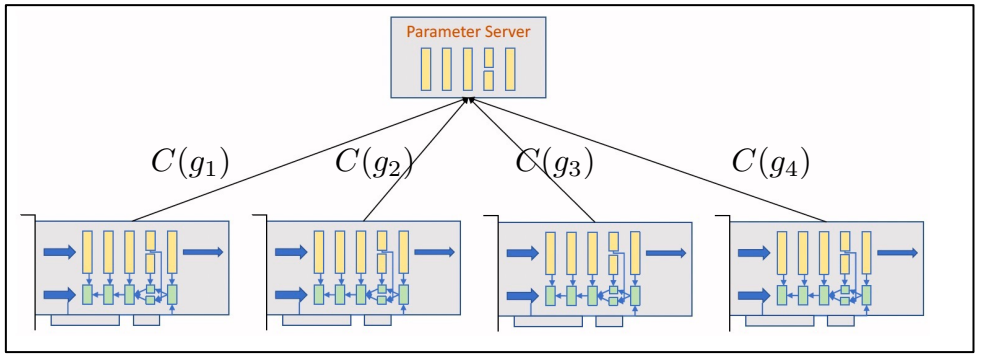
B. Ying, K. Yuan, H. Hu, Y. Chen and W. Yin, “BlueFog: Make decentralized algorithms practical for optimization and deep learning”, arXiv: 2111. 04287, 2021

Communication compression

- A basic (but not state-of-the-art) algorithm is QSGD [Alistarh et. al., 2017]

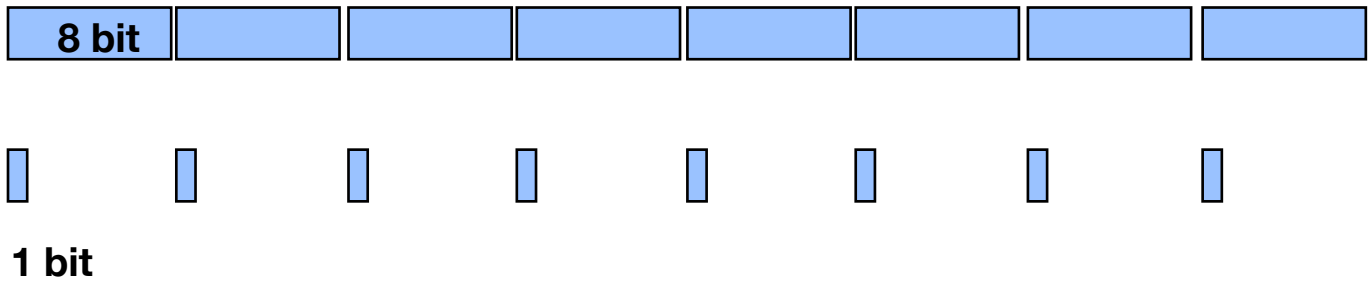
$$g_i^{(k)} = \nabla F(x_i^{(k)}; \xi_i^{(k)})$$

$$x_i^{(k+1)} = x_i^{(k)} - \frac{\gamma}{n} \sum_{j=1}^{n} C(g_j^{(k)})$$



- $C(\cdot)$ is a compressor. It can quantize or sparsify the full gradient

Quantization

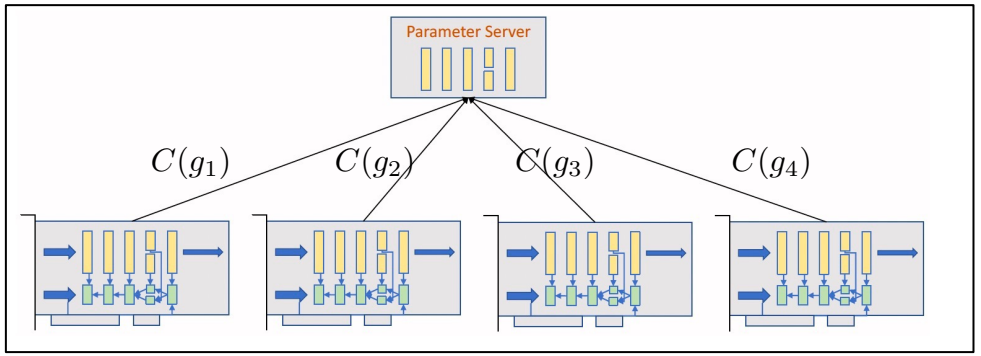


Communication compression

- A basic (but not state-of-the-art) algorithm is QSGD [Alistarh et. al., 2017]

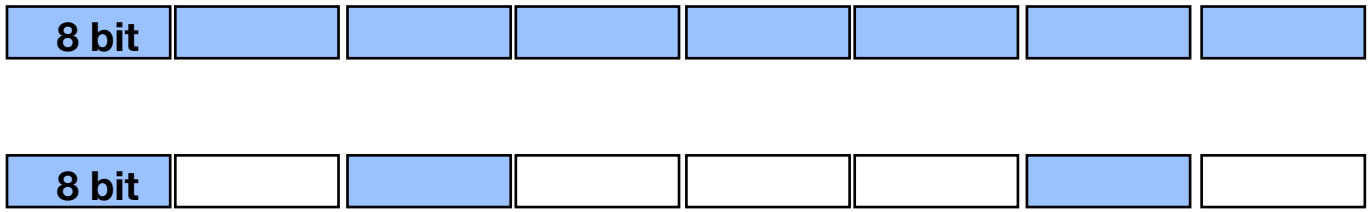
$$g_i^{(k)} = \nabla F(x_i^{(k)}; \xi_i^{(k)})$$

$$x_i^{(k+1)} = x_i^{(k)} - \frac{\gamma}{n} \sum_{j=1}^{n} C(g_j^{(k)})$$



- $C(\cdot)$ is a compressor. It can quantize or sparsify the full gradient

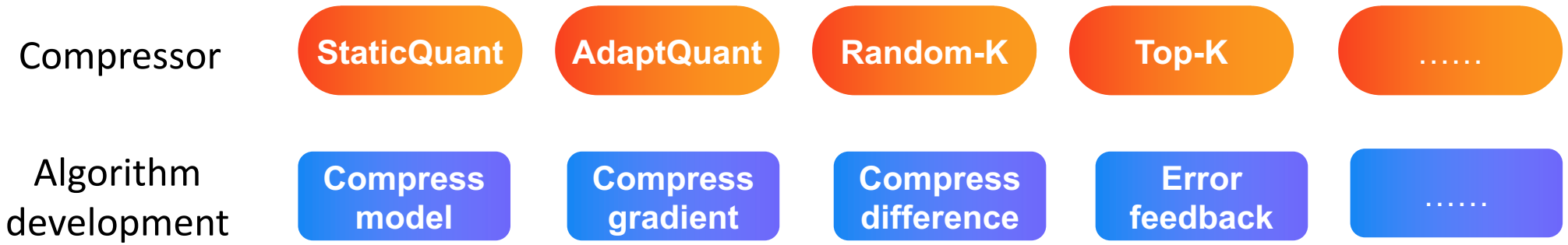
Sparsification



Communication compression algorithms

北京大学 PEKING UNIVERSITY

- There are extensive studies in distributed learning with communication compression



- The combination of different compressors, algorithms, and strategies gives rise to

Q-SGD [Alistarh et. al., 2017], Mem-SGD [Stich et. al., 2018], EF21-SGD [Fatkhullin et. al., 2021], CSER [Xie et.al., 2020], Double Squeeze [Tang et. al., 2019], Artemis [Philippenko et.al. 2022], etc.

Fundamental open questions

北京大学 PEKING UNIVERSITY

- **What is the optimal complexity in distributed optimization with communication compression?**

- **Can we develop effective algorithms that can attain the optimal complexity?**

北京大学
PEKING UNIVERSITY

PART 02

Optimal Complexity Formulation

Function class and gradient oracle class

- Function class. We let $\mathcal{F}_{\mu,L}$ denote the set of all functions satisfying Assumption 1

Assumption 1 (Smoothness and Strong Convexity) Each local objective f_i has L-Lipschitz gradient and μ-strongly convex. Moreover, we assume that $f(x^{(0)}) - \inf_{x\in\mathbb{R}^d} f(x) \leq \Delta$ with $f = \frac{1}{n}\sum_{i=1}^n f_i$.

- Gradient oracle class. Each worker accesses local gradient $\nabla f_i(x)$ via a stochastic oracle

Assumption 2 (Stochastic gradient) The gradient oracles $\{O_i : 1 \leq i \leq n\}$ satisfy

$$\mathbb{E}_{\zeta_i}[O_i(x;\zeta_i)] = \nabla f_i(x) \quad \text{and} \quad \mathbb{E}_{\zeta_i}[\|O_i(x;\zeta_i) - \nabla f_i(x)\|^2] \leq \sigma^2, \quad \forall\, x \in \mathbb{R}^d.$$

Compressor Class

- Compressor class. Most compressors in literature are either **unbiased** or **contractive**

- We let $\mathcal{U}_\omega$ denote the set of unbiased compressors satisfying Assumption 3

Assumption 3 (Unbiased compressor) The compression operator $C : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

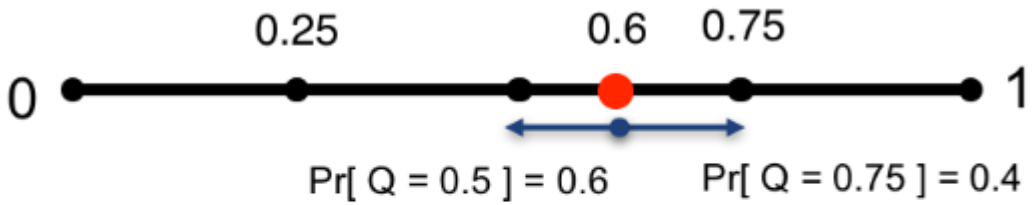
$$\mathbb{E}[C(x)] = x, \quad \mathbb{E}[\|C(x) - x\|^2] \leq \omega \|x\|^2, \quad \forall\, x \in \mathbb{R}^d$$

for constant $\omega \geq 0$, where the expectation is taken over the randomness of the compression operator C.

- Identity operator I (i.e. no compression) is an unbiased compressor with $\omega = 0$.

Example: random quantization

- Random quantization with 5 levels:



E[Q] = 0.6 × 0.5 + 0.4 * 0.75 = 0.6

Unbiased: $$\mathbb{E}[Q(x)] = \frac{Q_+(x) - x}{Q_+(x) - Q_-(x)} \cdot Q_-(x) + \frac{x - Q_-(x)}{Q_+(x) - Q_-(x)} \cdot Q_+(x) = x$$

Compressor Class

北京大学 PEKING UNIVERSITY

- Compressor class. Most compressors in literature are either **unbiased** or **contractive**
- We let $\mathcal{C}_\delta$ denote the set of unbiased compressors satisfying Assumption 4

Assumption 4 (Contractive compressor) The compression operator $C : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

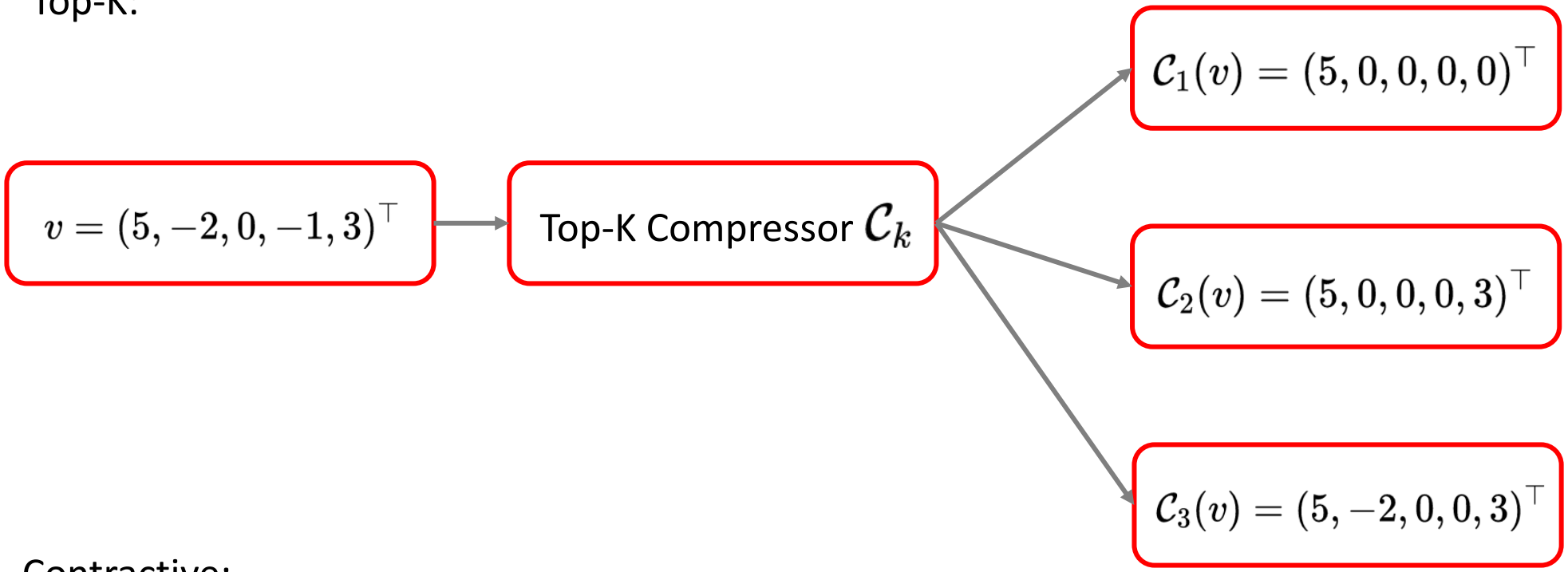
$$\mathbb{E}[\|C(x) - x\|^2] \leq (1 - \delta)\|x\|^2, \quad \forall\, x \in \mathbb{R}^d$$

for constant $\delta \in (0, 1]$, where the expectation is taken over the randomness of the compression operator C.

- Identity operator I (i.e. no compression) is a contractive compressor with $\delta = 1$.

Example: Top-K compressor

- Top-K:



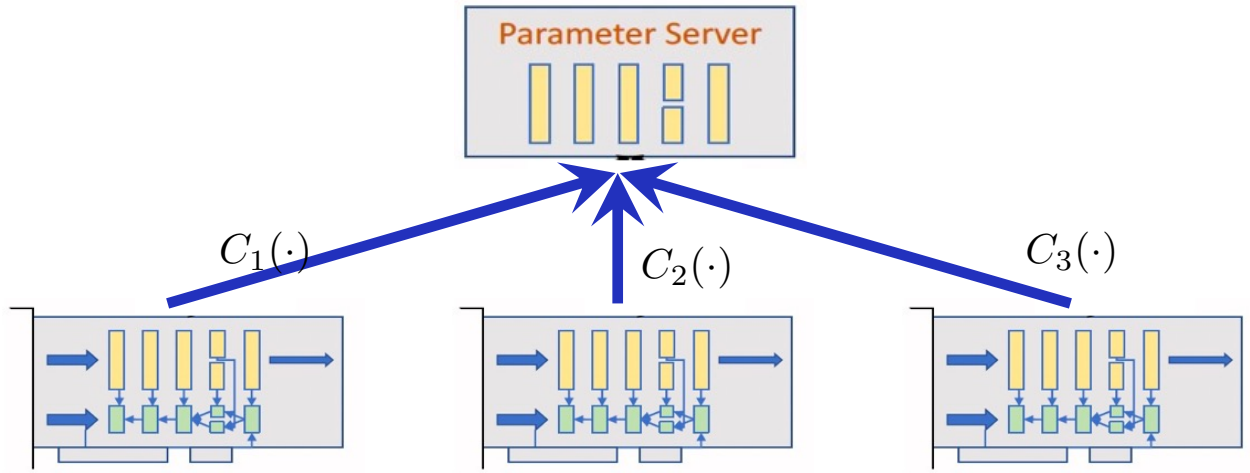
Contractive:

$$\|\mathcal{C}_k(v) - v\|^2 = \|v\|^2 - \|\mathcal{C}_k(v)\|^2 \leq (1 - k/d)\|v\|^2, \text{ i.e., } \delta = k/d \in (0, 1]$$

Algorithm Class

北京大学 PEKING UNIVERSITY

- Workers communicate directly with a central server. All iterations are synchronized.
- Each worker $i \in \{1, \cdots, n\}$ is endowed with C_i .
- Zero-respecting property: # non-zeros increase only by local update or comm. with the server



Complexity metric

- Let $\hat{x}_A^t$ denote the output of algorithm A after t communication rounds

- We define the complexity metric as

$$T_\epsilon(A, \{(f_i, C_i)\}_{i=1}^n) = \min\left\{t \in \mathbb{N} : \mathbb{E}[f(\hat{x}_A^t)] - \min_x f(x) \leq \epsilon\right\}$$

- It indicates the **smallest** number of iterations to achieve an ϵ - accurate solution

Find optimal complexity

- The problem to find optimal complexity can be formulated into the problem

$$\inf_{A\in\mathcal{A}} \sup_{\{C_i\}_{i=1}^n\subseteq\mathcal{C}} \sup_{\{O_i\}_{i=1}^n\subseteq\mathcal{O}_{\sigma^2}} \sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_{\mu,L}} T_\epsilon(A,\{(f_i,C_i)\}_{i=1}^n)$$

- In other words, given a class of functions $\mathcal{F}_{\mu,L}$, gradient oracles $\mathcal{O}_{\sigma^2}$, compressors $\mathcal{C}$ (being $\mathcal{C}_\delta$ or $\mathcal{U}_\omega$), the formulation seeks the optimal algorithm and the convergence complexity it has.

Find optimal complexity

北京大学 PEKING UNIVERSITY

- The problem to find optimal complexity can be formulated into the problem

$$\inf_{A\in\mathcal{A}} \sup_{\{C_i\}_{i=1}^n\subseteq\mathcal{C}} \sup_{\{O_i\}_{i=1}^n\subseteq\mathcal{O}_{\sigma 2}} \sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_{\mu,L}} T_\epsilon(A,\{(f_i,C_i)\}_{i=1}^n)$$

- Very challenging to directly solve the above problem
- We will establish the lower bound first, then the upper bound, and finally show they match each other
- The algorithm to achieve the lower bound is optimal

北京大学
PEKING UNIVERSITY

PART 03

Find Lower Bound

Lower bound with unbiased compressors

Theorem 1 (Unbiased compression)

For every $\Delta, L > 0$, $n \geq 2$, $\omega \geq 0$, $\sigma > 0$, $T \geq (1+\omega)^2$, there exists a set of local loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_{\mu,L}$, stochastic gradient oracles $\{O_i\}_{i=1}^n \subseteq \mathcal{O}_{\sigma^2}$, ω-unbiased compressors $\{C_i\}_{i=1}^n \subseteq \mathcal{U}_\omega$, such that for any algorithm $A \in \mathcal{A}$ starting from a given constant $x^{(0)}$, the lower bound complexity is given by

$$\Omega\left(\frac{\sigma^2}{\mu n \epsilon} + (1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right),$$

where ϵ is the desired accuracy.

Consistency with existing lower bound

$$\Omega\left(\frac{\sigma^2}{\mu n\epsilon}+(1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$$

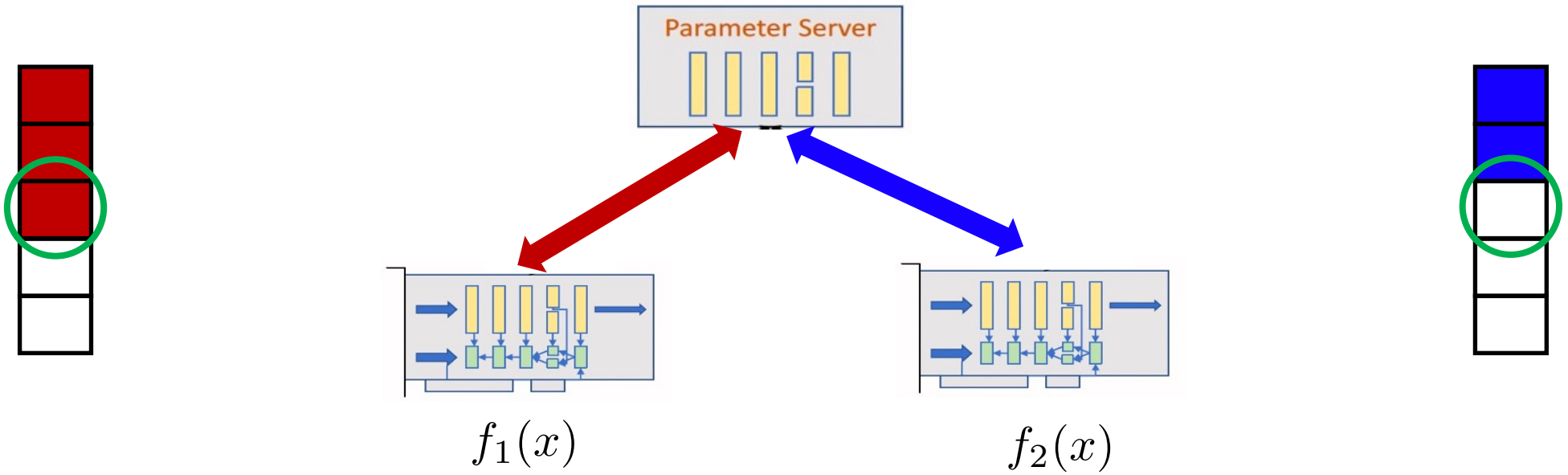
σ^2 is the gradient noise, n is the number of nodes, and ω gauges the compression ratio

- When $n=1$ and $\omega=0$, it recovers the lower bound in stochastic strongly-convex optimization
- When $n=1$, $\omega=0$ and $\sigma^2=0$, it recovers Nesterov lower bound

$$(n=1,\omega=0)\quad \Omega\left(\frac{\sigma^2}{\mu\epsilon}+\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right) \qquad (n=1,\omega=0,\sigma=0)\quad \Omega\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$$

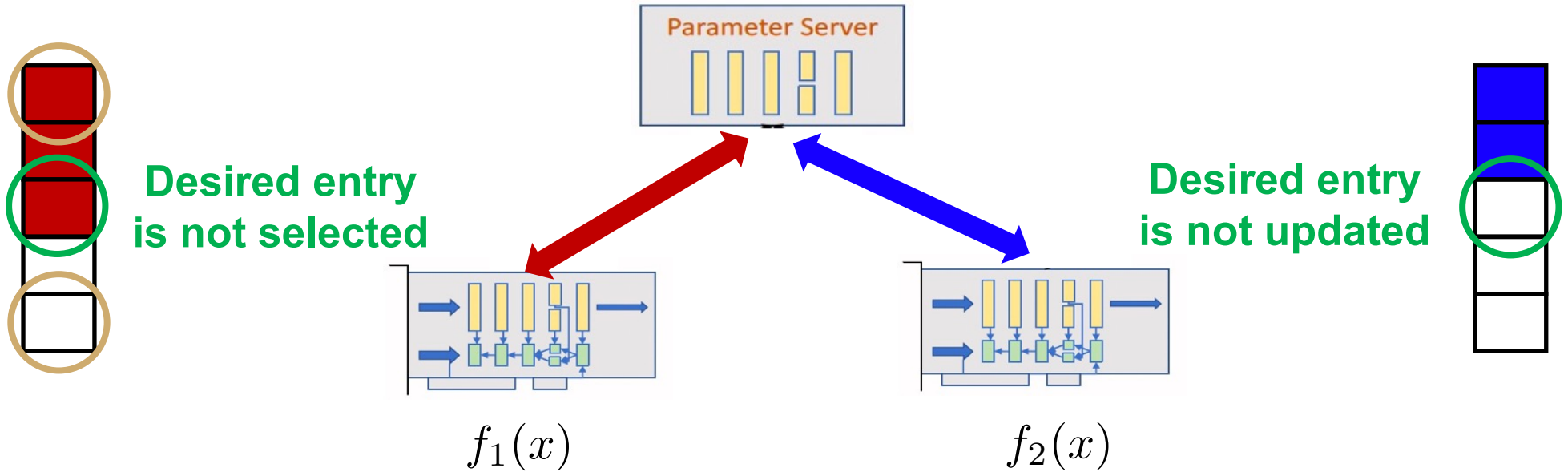
Distributed optimization with compression

- We construct adversarial objective functions
- The odd element in $f_2(x)$ can only be updated by receiving corresponding element from worker 1



Intuition behind lower bound

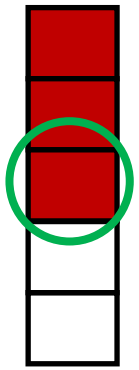
- With compression, the desired entry may not be able to communicate



- It explains why communication compression converges slower than vanilla distributed SGD

Intuition behind lower bound

- What's the probability that the desired entry is sent?



Desired entry is not selected

- We constructed an adversarial compressor: **rand-s**
- In rand-s compressor, s elements will be uniformly randomly chosen from all d elements
- Rand-s is an unbiased compressor (after scale) with $\omega = \frac{d}{s} - 1$
- The desired entry is sent with probability $s/d = (1+\omega)^{-1}$

Intuition behind lower bound

- In expectation, the desired entry will experience $(1+\omega)$ rounds to be sent

- Recall the lower bound complexity of standard distributed optimization as

$$\Omega\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$$

- Since sending the desired entry require $(1+\omega)$ rounds, the lower bound with compression is

$$\Omega\left((1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$$

Lower bound with contractive compressor

北京大学 PEKING UNIVERSITY

Theorem 2 (Contractive compression)

For every $\Delta, L > 0, n \geq 2, \delta \in (0, 1), \sigma > 0, T \geq (1+\omega)^2$, there exists a set of local loss functions $\{f_i\}_{i=1}^n \subseteq \mathcal{F}_{\mu,L}$, stochastic gradient oracles $\{O_i\}_{i=1}^n \subseteq \mathcal{O}_{\sigma^2}$, δ-contractive compressors $\{C_i\}_{i=1}^n \subseteq \mathcal{C}_\delta$, such that for any algorithm $A \in \mathcal{A}$ starting from a given constant $x^{(0)}$, it holds that

$$\mathbb{E}[f(\hat{x}) - f^\star] = \Omega\left(\frac{\sigma^2}{\mu n \epsilon} + \frac{1}{\delta}\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$$

北京大学
PEKING UNIVERSITY

PART 04

Algorithm to Attain Lower Bounds

Optimal algorithm for distributed optimization

- Recall the problem $\min_{x\in\mathbb{R}^d} \quad f(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad \text{where} \quad f_i(x)=\mathbb{E}_{\xi_i\sim D_i}F(x;\xi_i).$
- We first consider the scenario in which the true gradient $\nabla f_i(x)$ is known and no compression is used
- The optimal algorithm is **Nesterov Acceleration**. For k=1,2,…,T

$$
\left.\begin{aligned} z^k &= \frac{1}{\gamma}x^k + (\frac{1}{p}-\frac{1}{\gamma})x^{k-1} + (1-\frac{1}{p})z^{k-1} \\ y^k &= (1-\frac{\gamma}{p})x^k + \frac{\gamma}{p}z^k \end{aligned}\right\} \text{(extrapolation)}
$$

$$
g_i^k = \nabla f_i(y^k) \qquad \text{(true gradient)}
$$

$$
g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k \qquad \text{(precise communication)}
$$

$$
x^{k+1} = y^k - \frac{\eta}{p}g^k \qquad \text{(gradient descent)}
$$

Extension to stochastic and compressed scenario

- A trivial extension: For k = 1,2,…,T

$$
\left.
\begin{aligned}
z^k &= \frac{1}{\gamma}x^k + (\frac{1}{p} - \frac{1}{\gamma})x^{k-1} + (1 - \frac{1}{p})z^{k-1} \\
y^k &= (1 - \frac{\gamma}{p})x^k + \frac{\gamma}{p}z^k
\end{aligned}
\right\} \text{(extrapolation)}
$$

$$
g_i^k = \nabla F(y^k; \xi_i^k) \qquad \text{(stochastic gradient)}
$$

$$
g^k = \frac{1}{n}\sum_{i=1}^{n} C_i(g_i^k) \qquad \text{(communication compression)}
$$

$$
x^{k+1} = y^k - \frac{\eta}{p}g^k \qquad \text{(gradient descent)}
$$

- Does not work well. Stochastic gradient and compression introduces too much noise.

Upper bound and nearly-optimal algorithms

- A refined algorithm: For k = 1,2,…,T/R

$$
\left.
\begin{aligned}
z^k &= \frac{1}{\gamma}x^k + (\frac{1}{p} - \frac{1}{\gamma})x^{k-1} + (1 - \frac{1}{p})z^{k-1} \\
y^k &= (1 - \frac{\gamma}{p})x^k + \frac{\gamma}{p}z^k
\end{aligned}
\right\} \quad \text{(extrapolation)}
$$

$$
g_i^k = \frac{1}{R}\sum_{r=1}^{R} \nabla F(y^k; \xi_i^{(k,r)}) \qquad \text{(large-batch stochastic gradient)}
$$

$$
g^k = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSC}(g_i^k, R) \qquad \text{(multi-step compression)}
$$

$$
x^{k+1} = y^k - \frac{\eta}{p}g^k \qquad \text{(gradient descent)}
$$

- The communication and sample budget are the same as previous algorithms

Upper bound and nearly-optimal algorithms

北京大学 PEKING UNIVERSITY

- Multi-step compression (MSC) protocol:

Input: Origical vector v and compression round R

Initialize $v^{(0)} = 0$;

for $r = 0, \cdots, R-1$ **do**

compress $v - v^{(r)}$ into $c^{(r)} = C(v - v^{(r)})$

send $c^{(r)}$ to the target

update $v^{(r+1)} = v^{(r)} + c^{(r)}$

end for

- With MSC, we can prove

$$v^{(R)} = \sum_{r=0}^{R-1} c^{(r)}$$

$$\mathbb{E}\|v^{(R)} - v\|^2 \leq (1-\delta)^R \|v\|^2$$

- It holds that

$$v^{(R)} \to v \text{ as } R \to \infty$$

Upper bound and nearly-optimal algorithms

北京大学 PEKING UNIVERSITY

- A refined algorithm NEOLITHIC: For k = 1,2,…,T/R

$$
\left.
\begin{aligned}
z^k &= \frac{1}{\gamma}x^k + (\frac{1}{p} - \frac{1}{\gamma})x^{k-1} + (1 - \frac{1}{p})z^{k-1} \\
y^k &= (1 - \frac{\gamma}{p})x^k + \frac{\gamma}{p}z^k
\end{aligned}
\right\} \quad \text{(extrapolation)}
$$

$$
g_i^k = \frac{1}{R}\sum_{r=1}^{R} \nabla F(y^k; \xi_i^{(k,r)}) \qquad \text{(large-batch stochastic gradient)}
$$

$$
g^k = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSC}(g_i^k, R) \qquad \text{(multi-step compression)}
$$

$$
x^{k+1} = y^k - \frac{\eta}{p}g^k \qquad \text{(gradient descent)}
$$

- As $R \to \infty$, the refined algorithm will approach to vanilla Nesterov; the introduced noise can be controlled

Upper bound and nearly-optimal algorithms

北京大学 PEKING UNIVERSITY

Theorem 3 (Upper bound)

(Informal) When utilizing unbiased compressors, the algorithm converges at

$$\mathbb{E}[f(x^K) - f^\star] = \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n \epsilon} + (1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logrithm terms that are independent of ϵ. When using contractive compressors, the algorithm converges at

$$\mathbb{E}[f(x^K) - f^\star] = \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n \epsilon} + \frac{1}{\delta}\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right).$$

Upper bound and nearly-optimal algorithms

$$\inf_{A\in\mathcal{A}} \sup_{\{C_i\}_{i=1}^n\subseteq\mathcal{C}} \sup_{\{O_i\}_{i=1}^n\subseteq\mathcal{O}_{\sigma^2}} \sup_{\{f_i\}_{i=1}^n\subseteq\mathcal{F}_{\mu,L}} T_\epsilon(A,\{(f_i,C_i)\}_{i=1}^n)$$

- The lower bound and upper bound are nearly matched

$$\begin{aligned}\mathbb{E}[f(\hat{x})-f^\star] &= \Omega\left(\frac{\sigma^2}{\mu n\epsilon}+(1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)\\ &= \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n\epsilon}+(1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)\end{aligned}$$

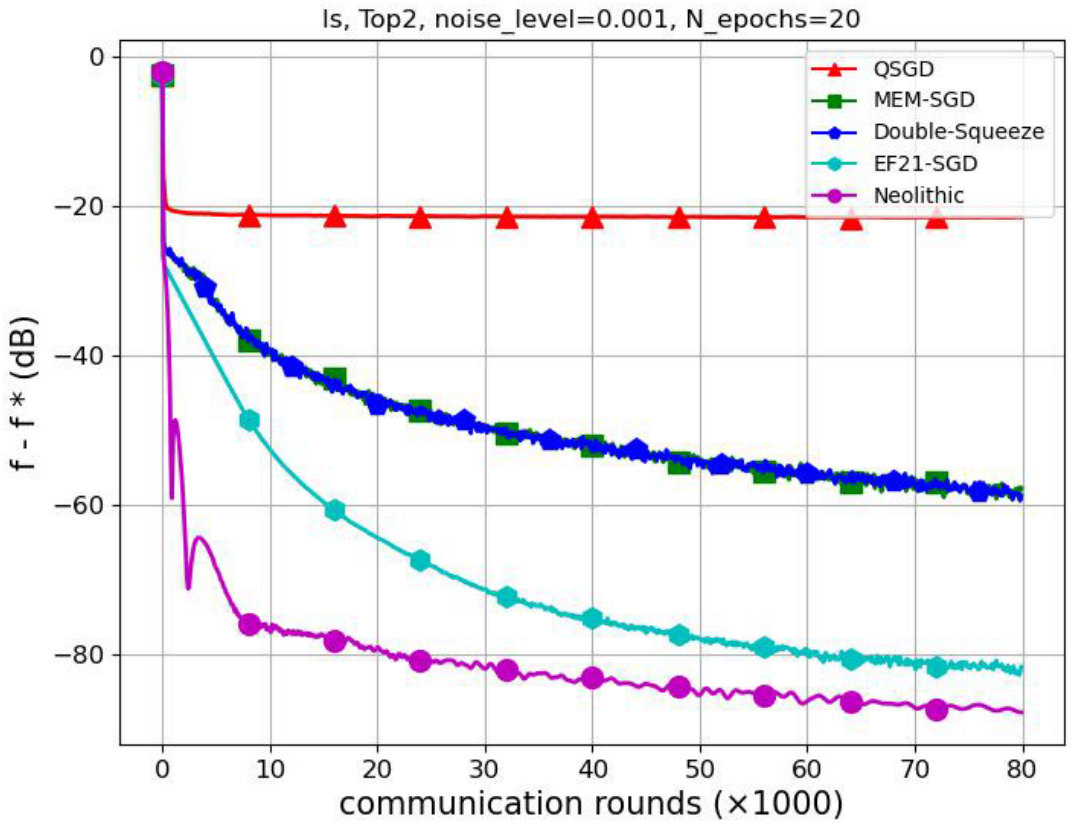
- **The established lower bound is tight. Our algorithm is optimal.**

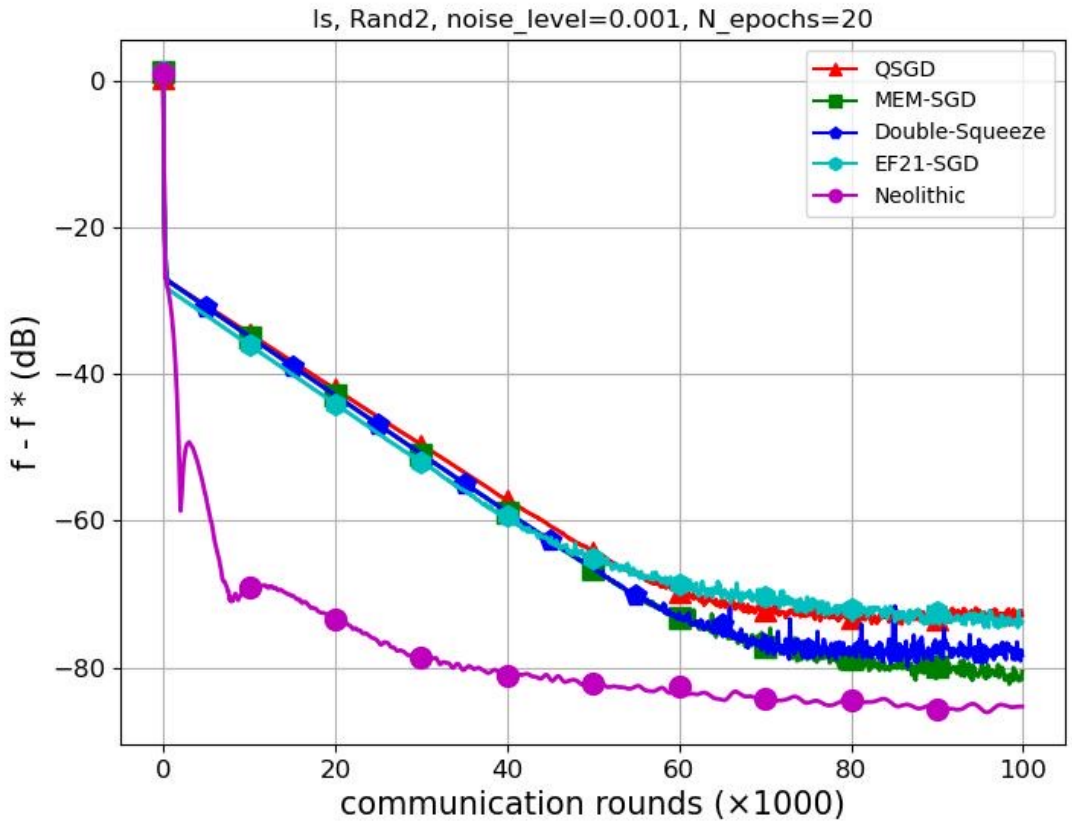
北京大学
PEKING UNIVERSITY

PART 05

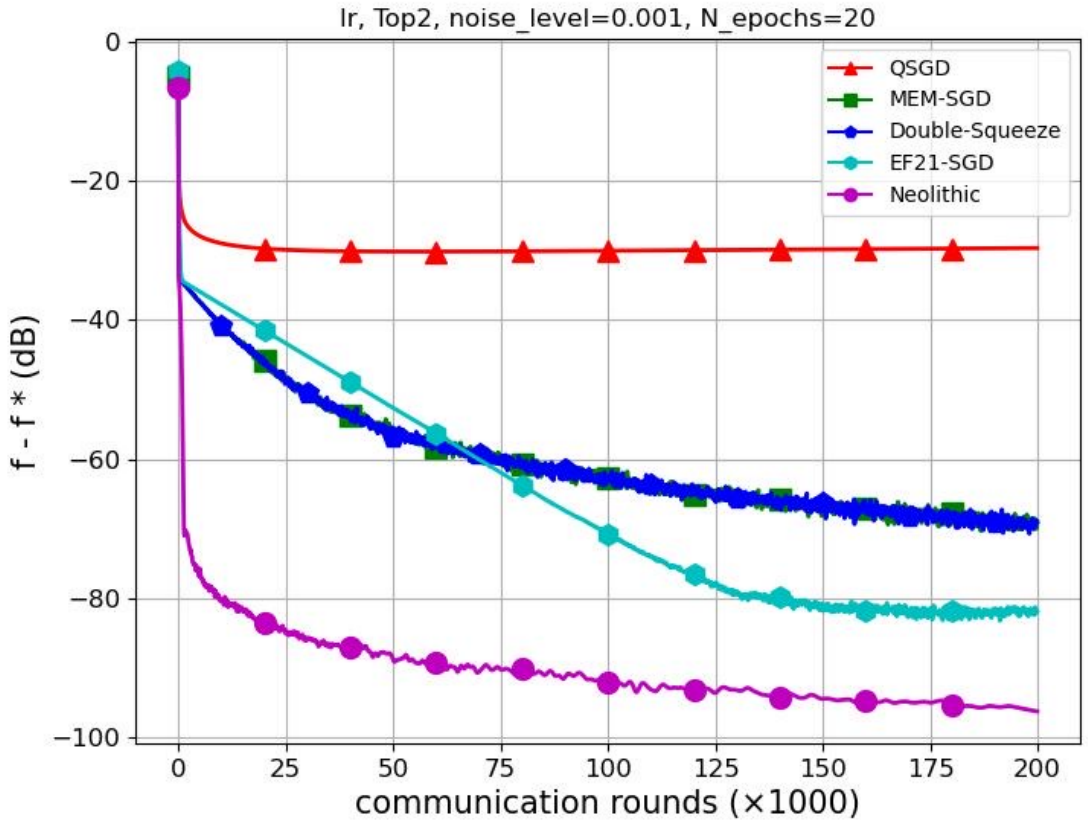
Empirical Studies

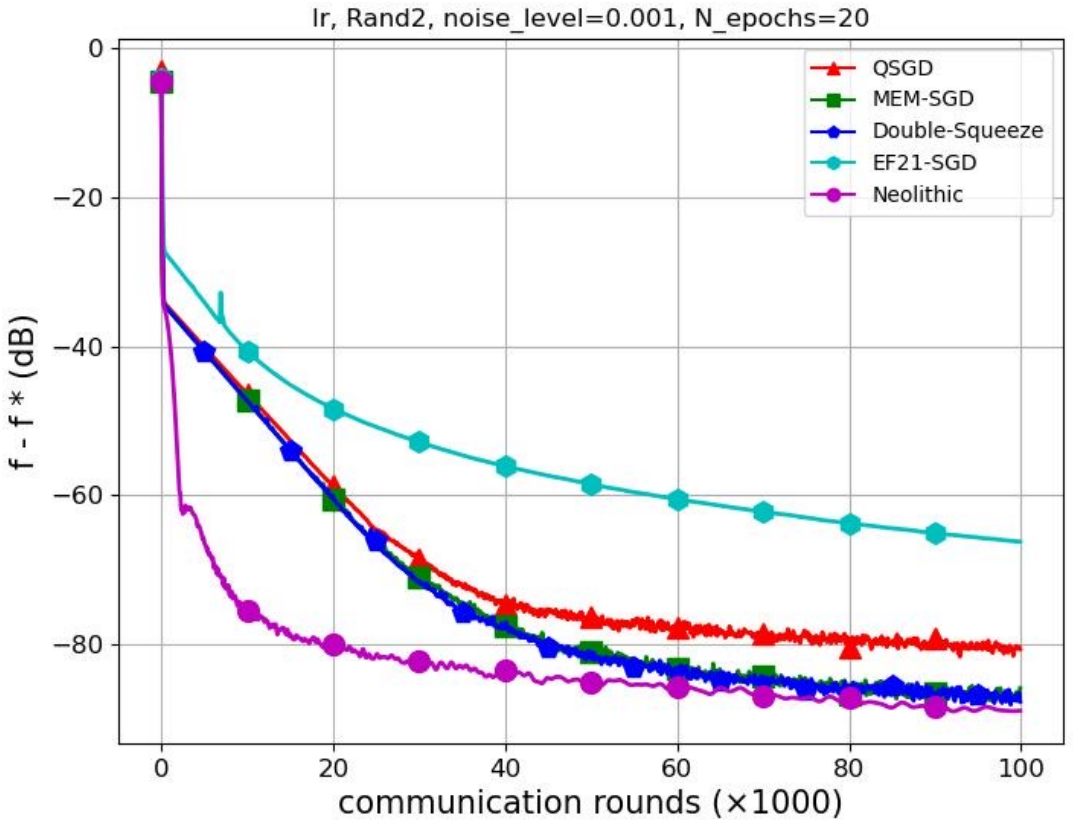
Least square with synthetic data





Logistic regression with synthetic data





北京大学
PEKING UNIVERSITY

PART 05

More results

Unbiased compression on generally-convex and non-convex scenarios

北京大学 PEKING UNIVERSITY

- We also study generally-convex and non-convex scenarios

Unbiased compressor

Method	NC	GC	SC
L.B.	$\Omega\left(\frac{L\Delta_f\sigma^2}{n\epsilon^2}+\frac{(1+\omega)L\Delta_f}{\epsilon}\right)$	$\Omega\left(\frac{\Delta_x\sigma^2}{n\epsilon^2}+\frac{(1+\omega)\sqrt{L\Delta_x}}{\sqrt{\epsilon}}\right)$	$\Omega\left(\frac{\sigma^2}{\mu n\epsilon}+(1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu\Delta_x}{\epsilon}\right)\right)$
Ours	$\tilde{\mathcal{O}}\left(\frac{L\Delta_f\sigma^2}{n\epsilon^2}+\frac{(1+\omega)L\Delta_f}{\epsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{\Delta_x\sigma^2}{n\epsilon^2}+\frac{(1+\omega)\sqrt{L\Delta_x}}{\sqrt{\epsilon}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n\epsilon}+(1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$

- Lower bound and upper bound are **nearly matched** (up to logarithm terms)

- **The lower bound is tight. Our proposed NEOLITHIC is nearly optimal**

Contractive compression on generally-convex and non-convex scenarios

北京大学 PEKING UNIVERSITY

- We also study generally-convex and non-convex scenarios

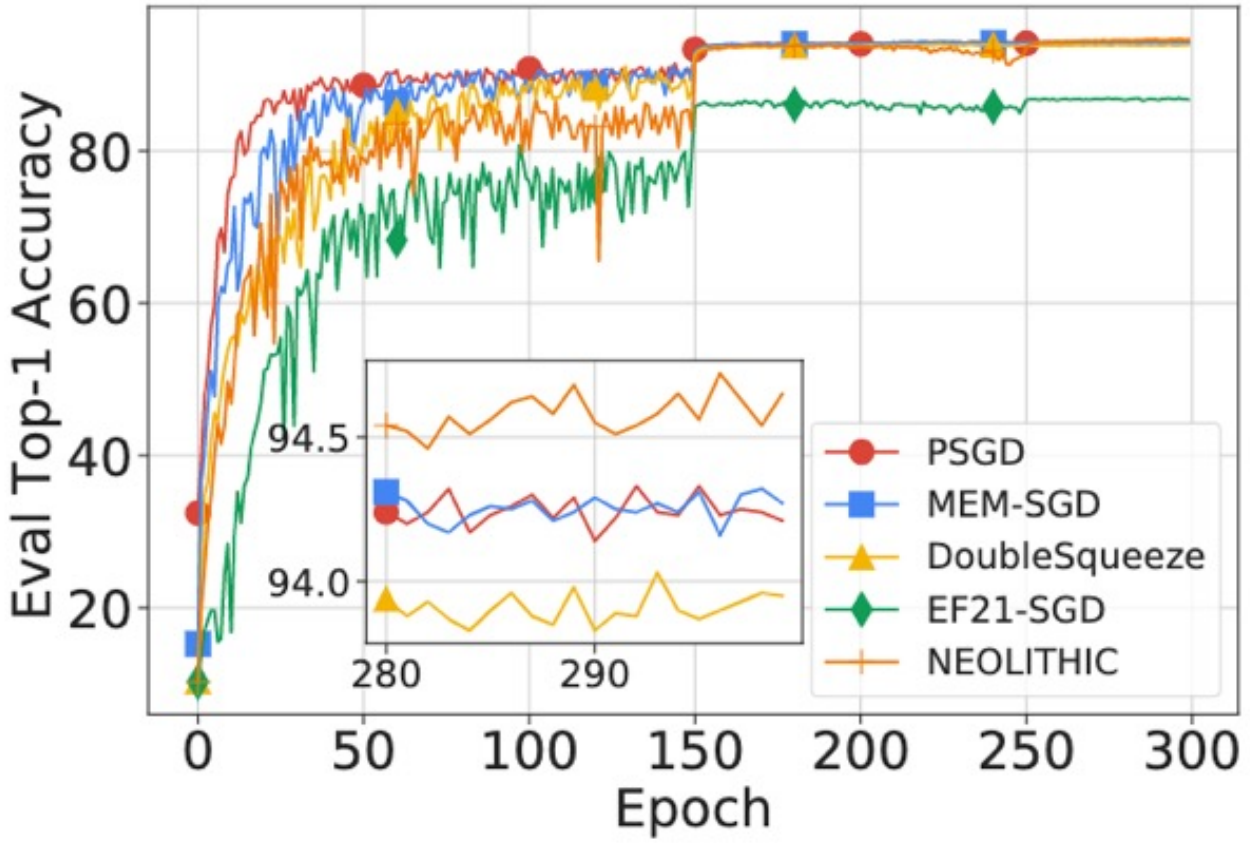
Contractive compressor

Method	NC	GC	SC
L.B.	$\Omega\left(\frac{L\Delta_f\sigma^2}{n\epsilon^2}+\frac{L\Delta_f}{\delta\epsilon}\right)$	$\Omega\left(\frac{\Delta_x\sigma^2}{n\epsilon^2}+\frac{\sqrt{L\Delta_x}}{\delta\sqrt{\epsilon}}\right)$	$\Omega\left(\frac{\sigma^2}{\mu n\epsilon}+\frac{1}{\delta}\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu\Delta_x}{\epsilon}\right)\right)$
Ours	$\tilde{\mathcal{O}}\left(\frac{L\Delta_f\sigma^2}{n\epsilon^2}+\frac{L\Delta_f}{\delta\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)^{\ddagger}$	$\tilde{\mathcal{O}}\left(\frac{\Delta_x\sigma^2}{n\epsilon^2}+\frac{\sqrt{L\Delta_x}}{\delta\sqrt{\epsilon}}\ln\left(\frac{1}{\epsilon}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n\epsilon}+\frac{1}{\delta}\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$

- Lower bound and upper bound are **nearly matched** (up to logarithm terms)
- **The lower bound is tight. Our proposed NEOLITHIC is nearly optimal**

Deep learning experiments

- 8 workers, 1% compression ratio (top-k compressors), minibatch=128, R=2, ResNet18/ResNet20



Deep learning experiments

北京大学 PEKING UNIVERSITY

- 8 workers; 4-bit quantization (unbiased); minibatch=128; R=2

METHODS	RESNET18	RESNET20
PSGD	93.99 ± 0.52	91.62 ± 0.13
MEM-SGD	94.35 ± 0.01	91.27 ± 0.08
DOUBLE-SQUEEZE	94.11 ± 0.14	90.73 ± 0.02
EF21-SGD	87.37 ± 0.49	65.82 ± 4.86
NEOLITHIC	$\mathbf{94.63 \pm 0.09}$	$\mathbf{91.43 \pm 0.10}$

A brief summary

北京大学 PEKING UNIVERSITY

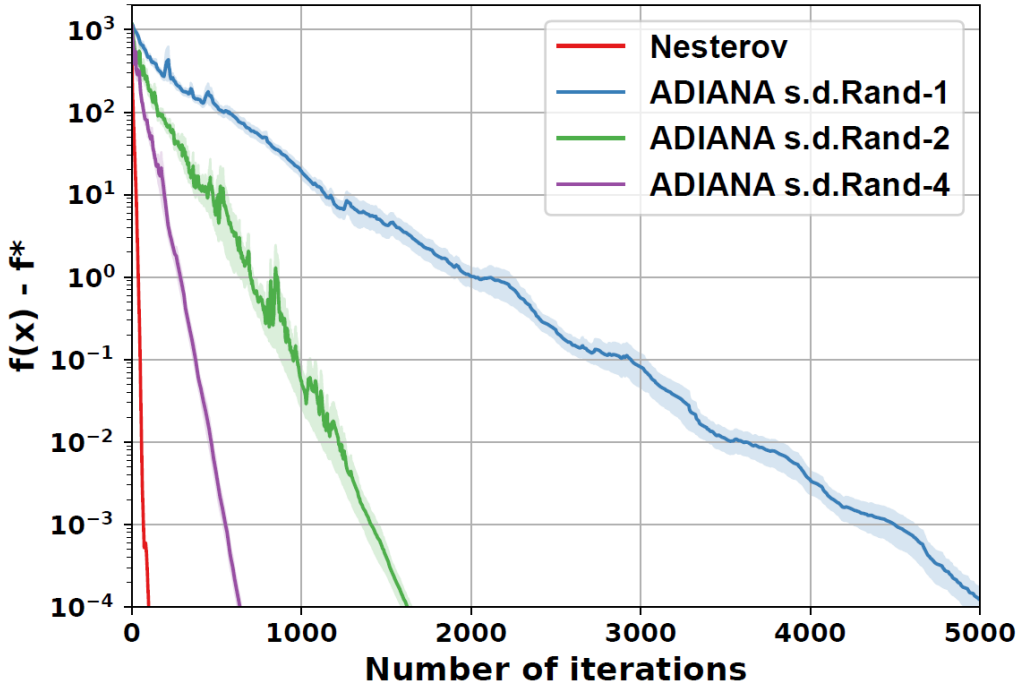
- Compression is critical to save communication in distributed learning
- We establish the optimal complexity for distributed learning with communication compression
- We propose NEOLITHIC to nearly attain this optimal complexity
- Our results hold for non-convex, generally-convex, and strongly-convex scenarios

北京大学
PEKING UNIVERSITY

The end of the story? Not Yet !

Some known facts

- Unbiased compression reduces communication cost per iteration
- However, it introduces **information distortion**
  - Make algorithms converge slower $\mathcal{O}\left((1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$
  - Incurs more communication rounds to achieve the desired solution
- Not sure whether the extra rounds of communication can outweigh the saving in per-iteration communication



北京大学
PEKING UNIVERSITY

Can compression REALLY save communication?

北京大学
PEKING UNIVERSITY

Unbiased Compression Saves Communication in Distributed Optimization: When and How Much?

Yutong He Xinmeng Huang Kun Yuan

Optimal total communication cost of non-compressed algorithms

- For simplicity, we consider deterministic and strongly-convex problems

$$\min \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

- For non-compressed algorithm, **the number of communication rounds** is lower bounded by

$$\Omega(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right))$$

A tight bound; NAG can touch it

- If the optimization variable is of **dimension** d , the total communication cost (TCC) is tightly evaluated as

comm. cost per round

$$\mathrm{TCC} = \Theta(d \cdot \sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right))$$

optimal comm. rounds

Optimal total communication cost of non-compressed algorithms

Proposition 1 (Non-compression)

Suppose each $f_i(x)$ is smooth and strongly-convex, the optimal total communication cost of non-compressed algorithms is given by

$$\mathrm{TCC} = \Theta(d\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right))$$

Nesterov accelerated gradient (NAG) achieves the above total communication cost

Optimal total communication cost of compressed algorithms

北京大学 PEKING UNIVERSITY

- For compressed algorithm with unbiased compression, when solving strongly-convex and smooth problems, **the number of communication rounds** is lower bounded by [1]

$$\Omega\left((1+\omega)\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right)$$

where $\omega > 0$ characterizes the compression error $\mathbb{E}\|C(x)-x\|^2 \le \omega\|x\|^2$

- This lower bound is **tight**. The compressed method NEOLITHIC [1] can match this lower bound

- If a compressor incurs larger compression error ω , the algorithm needs more extra communication rounds

[1] Y. He, X. Huang, Y. Chen, W. Yin, and K. Yuan, , “Lower Bounds and Accelerated Algorithms in Distributed Optimization with Communication Compression”, arXiv 2305.07612, 2023

Optimal total communication cost of compressed algorithms

北京大学 PEKING UNIVERSITY

- For unbiased compression, the lower bound of the **per-iteration communication cost** [2]

Proposition 2

Let d be the deminsion of the input variable x. For any unbiased compressor C satisfying Assumption 3, the per-iteration communication cost of $C(x)$ is lower bounded by

$$\Omega\left(\frac{d}{1+\omega}\right)$$

This lower bound is **tight** and can be achieved by random-K compressor

- The more aggressive the compressor is, the more per-round communication cost it saves

[2] Y. He, X. Huang, and K. Yuan, , "Unbiased Compression Saves Communication in Distributed Optimization: When and How Much?", NeurIPS, 2023

Optimal total communication cost of compressed algorithms

北京大学 PEKING UNIVERSITY

- Optimal total communication cost

Theorem 4 (Unbiased compression)

Suppose each $f_i(x)$ is smooth and strongly-convex, and the unbiased compressor is used in communication compression, the optimal total communication cost of compressed algorithms with unbiased compression is given by

$$\mathrm{TCC} = \Theta\left(\underbrace{\frac{d}{1+\omega}}_{\text{optimal per-round comm. cost}} \cdot \underbrace{(1+\omega)\sqrt{\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)}}_{\text{optimal comm. rounds}}\right) = \Theta\left(d\sqrt{\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)}\right)$$

- The per-round communication saving is **fully compensated** by the extra rounds of communication.

Compression v.s. non-compression

北京大学 PEKING UNIVERSITY

- Recall the optimal total communication cost of algorithms with compression and non-compression

Unbiased compression	Non-compression
$\Theta\left(d\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\Theta\left(d\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$

- Compression **cannot save communication**. They have the same total communication cost

北京大学 PEKING UNIVERSITY

Q: Can unbiased compression reduce the total communication cost ?

A: It cannot save communication for deterministic, strongly-convex and smooth problems.

Our work [2] also shows that unbiased compression cannot save communication for deterministic, convex and smooth problems.

---

[2] Y. He, X. Huang, and K. Yuan, , “Unbiased Compression Saves Communication in Distributed Optimization: When and How Much?”, NeurIPS, 2023

北京大学
PEKING UNIVERSITY

Very frustrating! Let's abandon compression!!

北京大学
PEKING UNIVERSITY

Wait!

There exist scenarios when compression works!

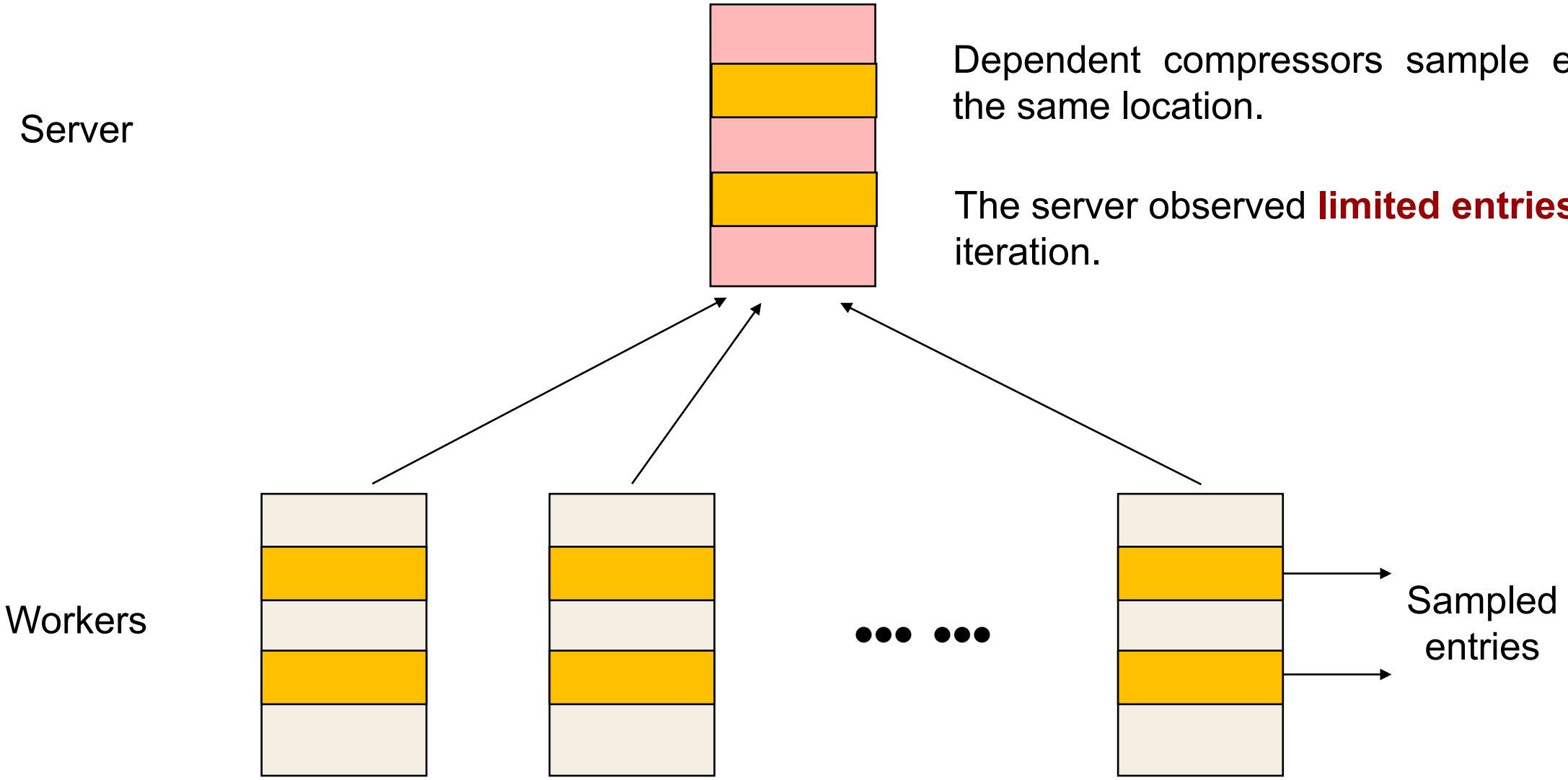
Independent compressor

Assumption 5 (Independent Compressor) We assume all compressors $\{C_i\}_{i=1}^n$ are mutually independent, i.e., outputs $\{C_i(x_i)\}_{i=1}^n$ are mutually independent random variables for any $\{x_i\}_{i=1}^n$.

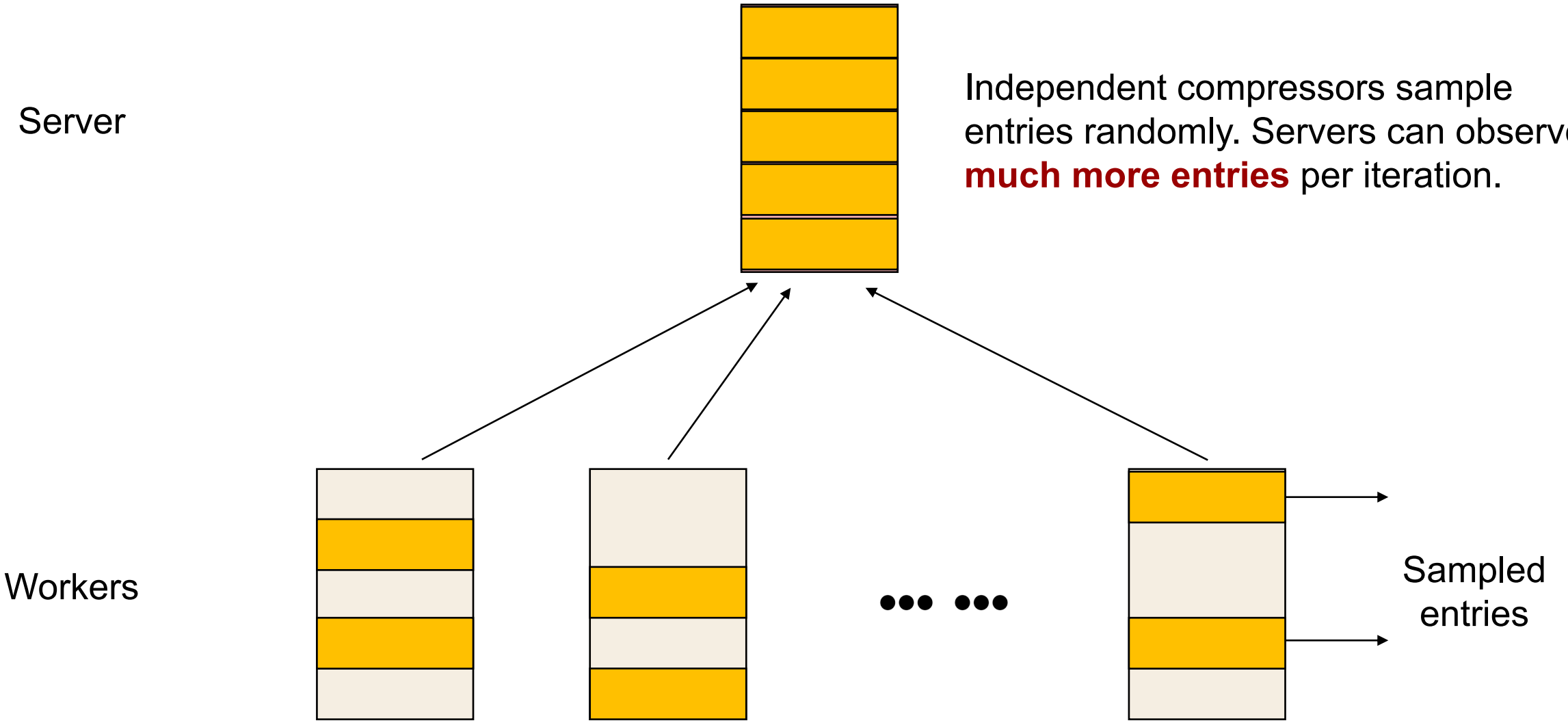
- As we discussed in previous slides, unbiased compression **alone** cannot save communication

- However, **independent unbiased compression** can save communication compared to non-compression

Independent compressor saves communication



Intuition on why independent compressor can save communication



Optimal total communication cost of independent compressed algorithms

Theorem 5 (Independent unbiased compressor)

Under Assumptions 1-3 and 5, the optimal total communication cost of compressed algorithms with independent unbiased compression is lower bounded by (when $\sqrt{L/\mu}$ is sufficiently large)

$$\mathrm{TCC} = \Omega\left(\underbrace{\frac{d}{1+\omega}}_{\text{optimal per-round comm. cost}} \cdot \underbrace{(1+\frac{\omega}{\sqrt{n}})\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)}_{\text{comm. round lower bound}}\right) = \Omega\left(d\left(\frac{1+\frac{\omega}{\sqrt{n}}}{1+\omega}\right)\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$$

This lower bound is tight. We show a refined ADIANA algorithm can match this lower bound

Comparison between various compression strategies

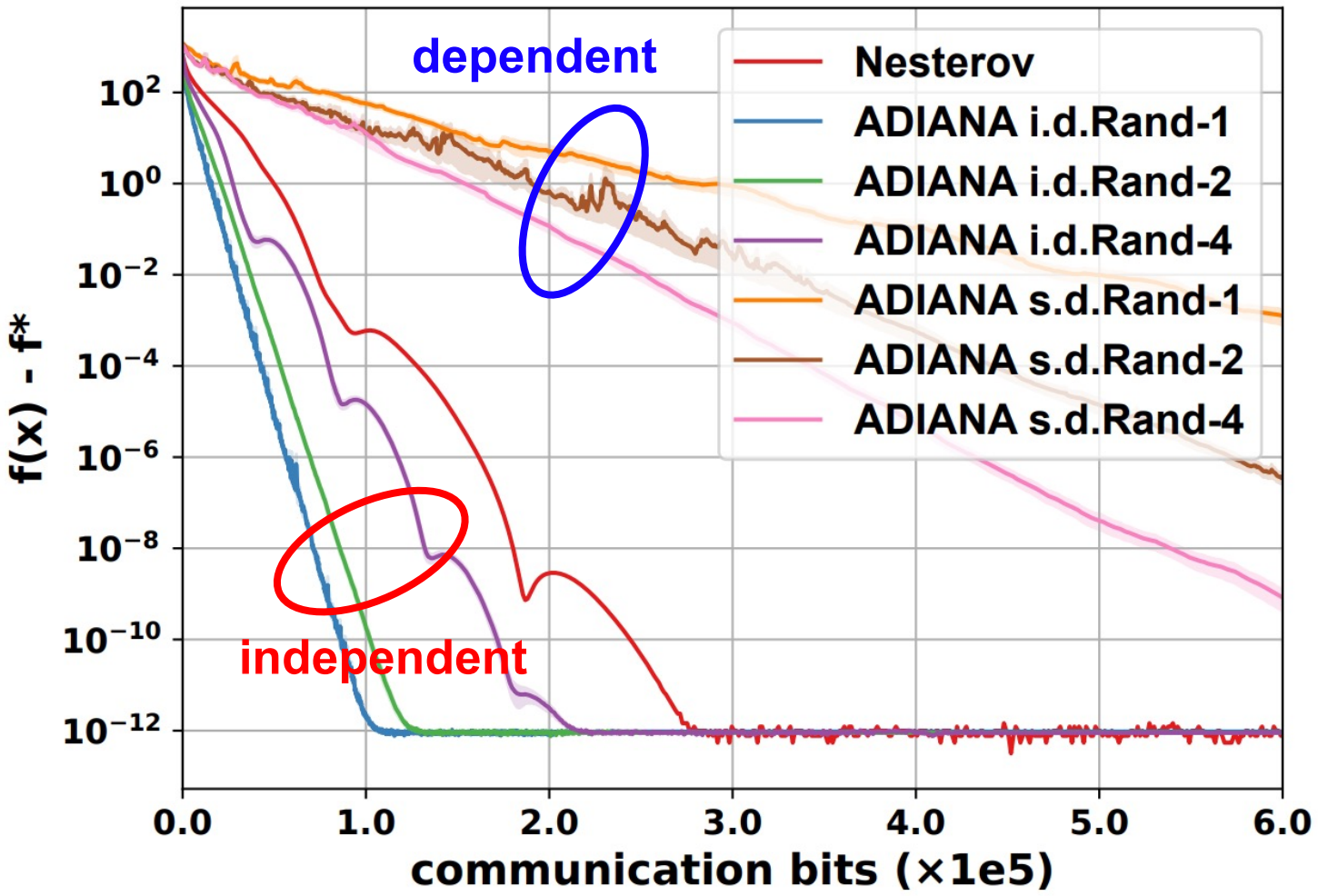
北京大学 PEKING UNIVERSITY

- Compare the total communication cost of various compression strategies

Non-compression	Unbiased compression	Independent unbiased compression
$\Theta\left(d\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\Theta\left(d\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\Theta\left(d\left(\frac{1+\frac{\omega}{\sqrt{n}}}{1+\omega}\right)\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$

- Unbiased compression alone **cannot save communication**
- Independent unbiased compression **can save communication** at most by $1/\sqrt{n}$ when $\omega \gg 1$
- Communication saving can not beyond $1/\sqrt{n}$. It is the **optimal communication saving**!

Numerical studies



Independent compressor **saves** communication

Dependent compressor **does not save** communication

Summary

For deterministic and convex problems

- Unbiased compressor alone cannot save communication
- Independent unbiased compressor can save communication
- Communication saving can be up to $1/\sqrt{n}$

Future work

- Examine non-convex and stochastic problems
- Examine contractive compressors

References

Y. He, X. Huang, Y. Chen, W. Yin, K. Yuan, “Unbiased compression saves communication in distributed optimization: when and how much ? ”, NeurIPS, 2023

Y. He, X. Huang, Y. Chen, W. Yin, K. Yuan, “Lower Bounds and Accelerated Algorithms in Distributed Optimization with Communication Compression”, arXiv 2305.07612, 2023

X. Huang, Y. Chen, W. Yin, K. Yuan, "Lower Bounds and Nearly Optimal Algorithms in Distributed Learning with Communication Compression", NeurIPS 2022

北京大学
PEKING UNIVERSITY

Thank you!

Kun Yuan homepage: https://kunyuan827.github.io/

We have openings for PostDocs