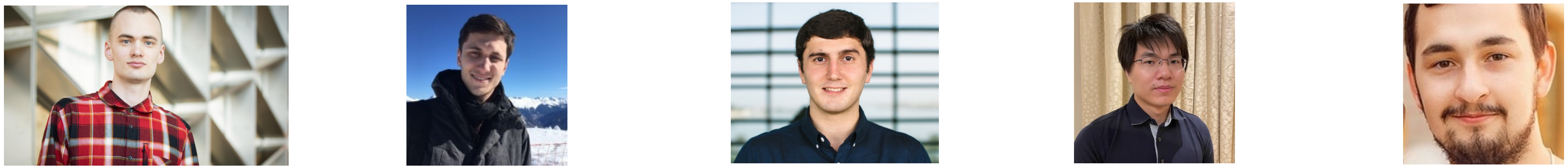
# 3PC: Three Point Compressors for Communication-Efficient Distributed Training

**Peter Richtárik (KAUST)**

## ICML 2022, Baltimore, Maryland, USA

*joint work with*

**Igor Sokolov** (KAUST), **Ilyas Fatkhullin** (ETH), **Elnur Gasanov** (KAUST), **Zhize Li** (KAUST), **Eduard Gorbunov** (MIPT)



# Distributed Nonconvex Optimization

$$\min_{x \in \mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

# Distributed Nonconvex Optimization

$$\min_{x\in\mathbb{R}^{d}} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

# model parameters / features

# Distributed Nonconvex Optimization

# devices / machines

$$\min_{x \in \mathbb{R}^{d}} f(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

# model parameters / features

# Distributed Nonconvex Optimization

# devices / machines

$$\min_{x \in \mathbb{R}^{d}} f(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

# model parameters / features

Lipschitz gradient:

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

# Distributed Compressed Gradient Descent

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

# Distributed Compressed Gradient Descent

# devices / machines

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

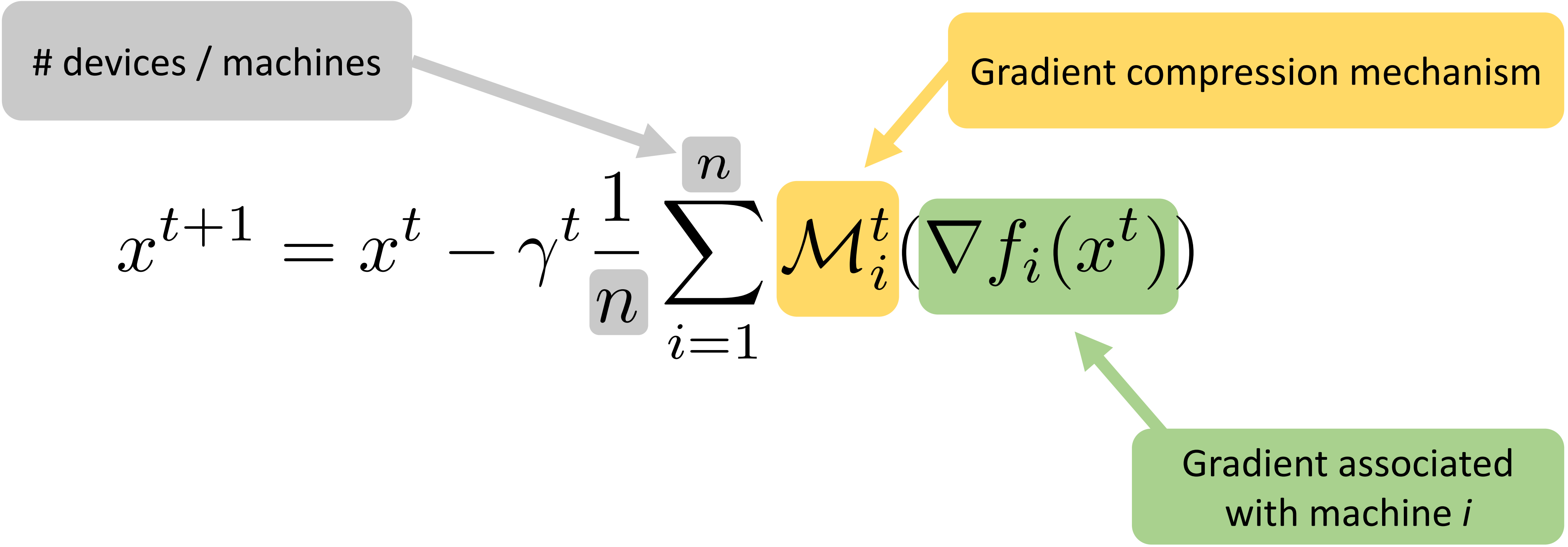
# Distributed Compressed Gradient Descent

# devices / machines

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

Gradient associated with machine *i*

# Distributed Compressed Gradient Descent



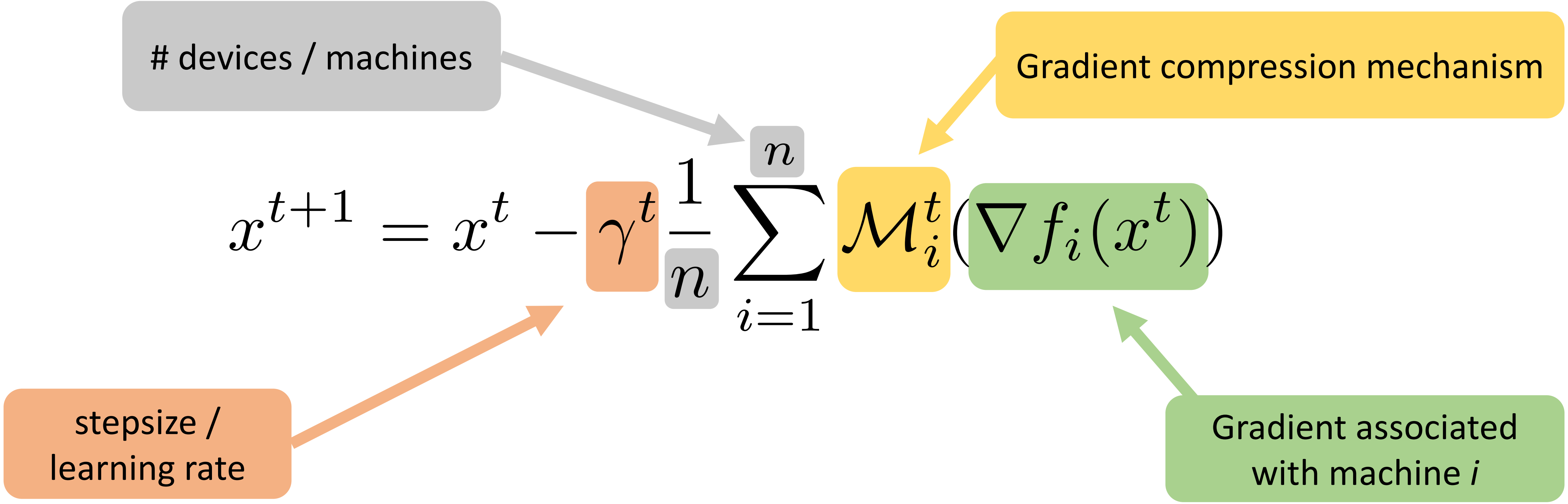
# devices / machines

Gradient compression mechanism

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

Gradient associated with machine *i*

# Distributed Compressed Gradient Descent



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

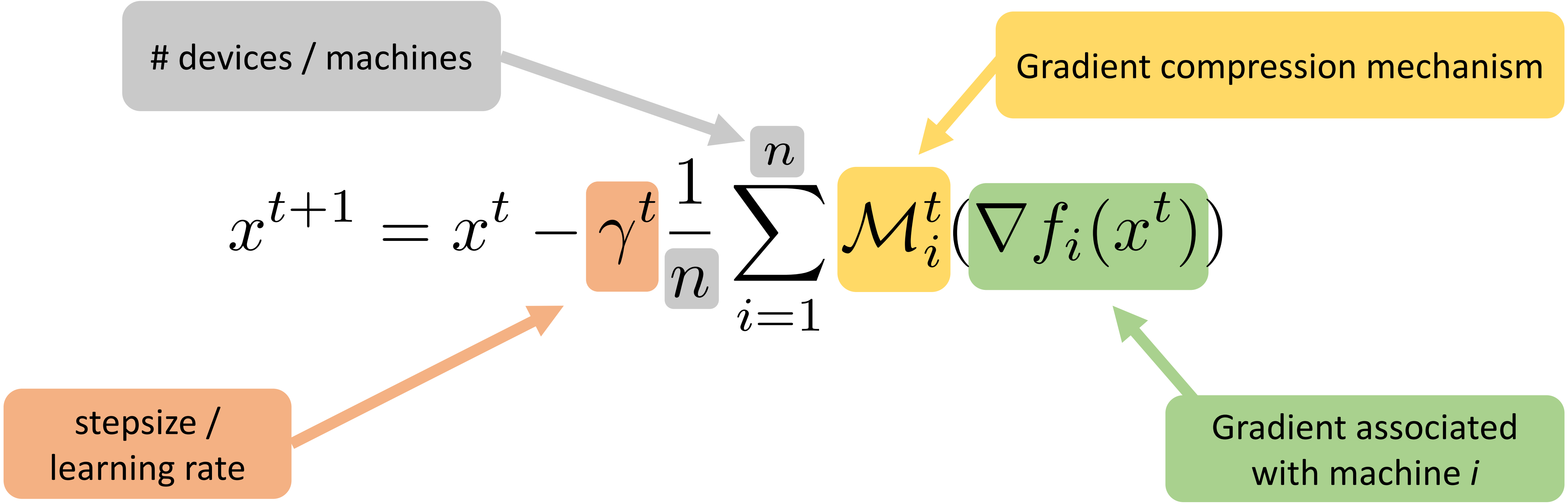
# devices / machines

Gradient compression mechanism

stepsize / learning rate

Gradient associated with machine *i*

# Distributed Compressed Gradient Descent



# devices / machines

Gradient compression mechanism

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

stepsize / learning rate

Gradient associated with machine *i*

**?** How to design a good compression mechanism?

# 3PC: Three Point Compressor

$$
\begin{aligned}
g_i^t &= \mathcal{M}_i^t \left( \nabla f_i(x^t) \right) \\
&= \mathcal{M} \left( g_i^{t-1} , \nabla f_i(x^{t-1}) , \nabla f_i(x^t) \right) \\
&= \mathcal{M}_{\, g_i^{t-1} , \nabla f_i(x^{t-1})} \left( \nabla f_i(x^t) \right)
\end{aligned}
$$

# 3PC: Three Point Compressor

Current compressed gradient

$$
\begin{aligned}
g_i^t &= \mathcal{M}_i^t\left(\nabla f_i(x^t)\right) \\
&= \mathcal{M}\left(g_i^{t-1}, \nabla f_i(x^{t-1}), \nabla f_i(x^t)\right) \\
&= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)
\end{aligned}
$$

Current compressed gradient

# 3PC: Three Point Compressor

$$
\begin{aligned}
g_i^t &= \mathcal{M}_i^t\left(\nabla f_i(x^t)\right) \\
&= \mathcal{M}\left(g_i^{t-1}, \nabla f_i(x^{t-1}), \nabla f_i(x^t)\right) \\
&= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)
\end{aligned}
$$

Current gradient

# 3PC: Three Point Compressor

Current compressed gradient

Current gradient

$$
\begin{aligned}
g_i^t &= \mathcal{M}_i^t\left(\nabla f_i(x^t)\right) \\
&= \mathcal{M}\left(\boxed{g_i^{t-1}, \nabla f_i(x^{t-1})}, \nabla f_i(x^t)\right) \\
&= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)
\end{aligned}
$$

# 3PC: Three Point Compressor

Current compressed gradient

$$\begin{aligned} g_i^t &= \mathcal{M}_i^t\left(\nabla f_i(x^t)\right) \\ &= \mathcal{M}\left(g_i^{t-1}, \nabla f_i(x^{t-1}), \nabla f_i(x^t)\right) \\ &= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right) \end{aligned}$$

Current gradient

Previous gradient

# 3PC: Three Point Compressor

Current compressed gradient

$$g_i^t = \mathcal{M}_i^t\left(\nabla f_i(x^t)\right)$$

Current gradient

$$= \mathcal{M}\left(g_i^{t-1}, \nabla f_i(x^{t-1}), \nabla f_i(x^t)\right)$$

Previous compressed gradient

Previous gradient

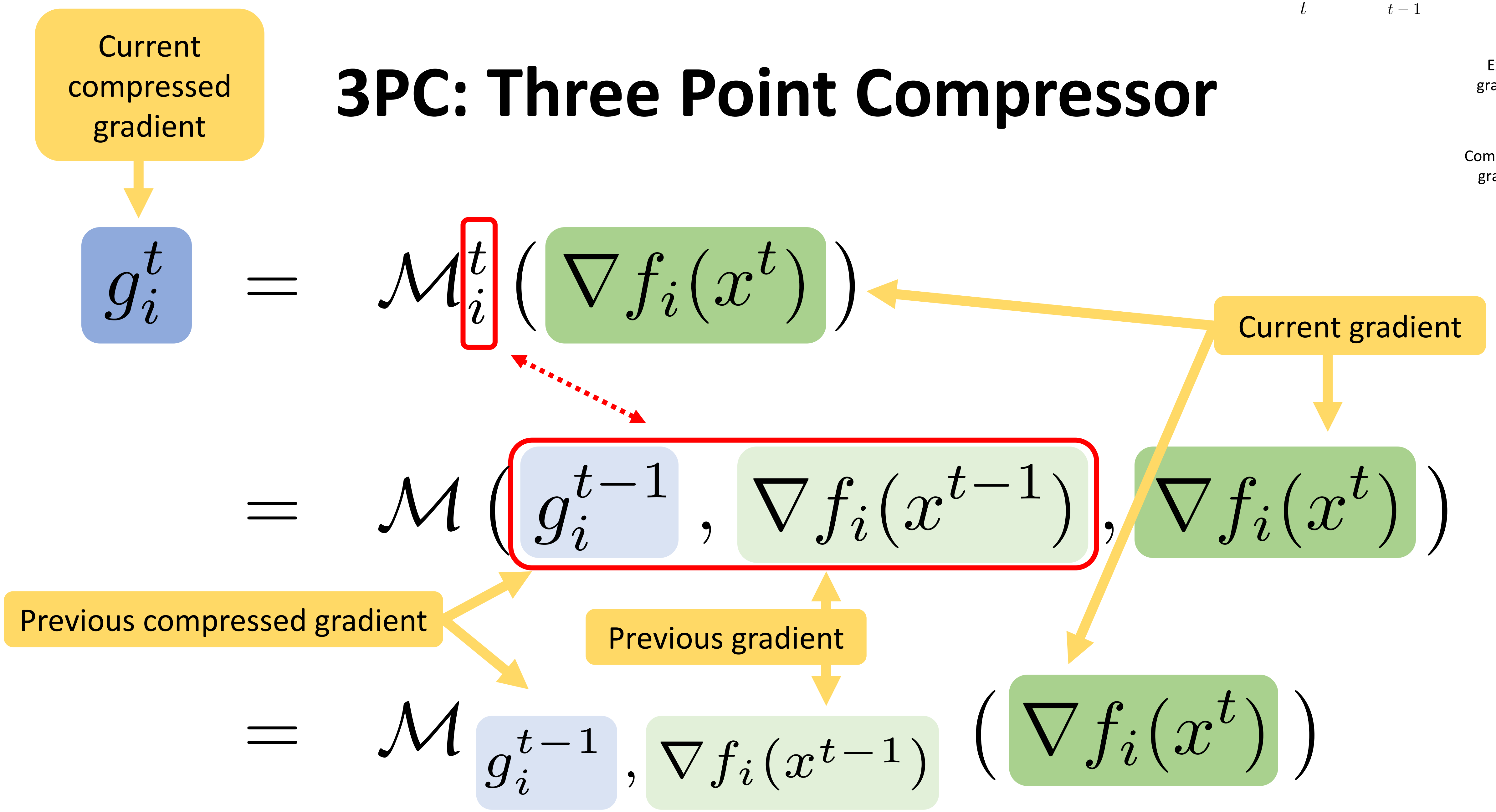
$$= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)$$

$t$ $t-1$

Exact gradient

Compressed gradient

# 3PC: Three Point Compressor



Current compressed gradient

$$g_i^t = \mathcal{M}\boxed{{}_i^t}\left(\nabla f_i(x^t)\right)$$

Current gradient

$$= \mathcal{M}\left(\boxed{g_i^{t-1}, \nabla f_i(x^{t-1})}, \nabla f_i(x^t)\right)$$

Previous compressed gradient

Previous gradient

$$= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)$$

# 3PC: Three Point Compressor

Current compressed gradient

$t$ $t-1$

Exact gradient

$g_i^t$

Compressed gradient

$$g_i^t = \mathcal{M}\boxed{{}_i^t}\left(\nabla f_i(x^t)\right)$$

Current gradient

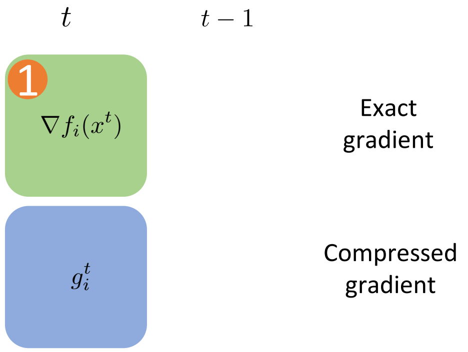
$$= \mathcal{M}\left(\boxed{g_i^{t-1}, \nabla f_i(x^{t-1})}, \nabla f_i(x^t)\right)$$

Previous compressed gradient

Previous gradient

$$= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)$$

# 3PC: Three Point Compressor



Current compressed gradient

$$g_i^t = \mathcal{M}_i^t\left(\nabla f_i(x^t)\right)$$

Current gradient

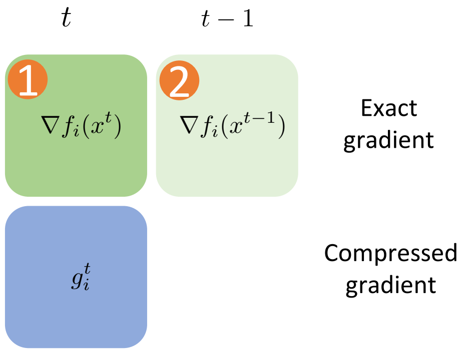
$$= \mathcal{M}\left(g_i^{t-1}, \nabla f_i(x^{t-1}), \nabla f_i(x^t)\right)$$

Previous compressed gradient

Previous gradient

$$= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)$$

# 3PC: Three Point Compressor



|  | $t$ | $t-1$ |  |
|---|---|---|---|
|  | ① $\nabla f_i(x^t)$ | ② $\nabla f_i(x^{t-1})$ | Exact gradient |
|  | $g_i^t$ |  | Compressed gradient |

Current compressed gradient

$$g_i^t = \mathcal{M}_i^t\left(\nabla f_i(x^t)\right)$$

Current gradient

$$= \mathcal{M}\left(g_i^{t-1}, \nabla f_i(x^{t-1}), \nabla f_i(x^t)\right)$$

Previous compressed gradient

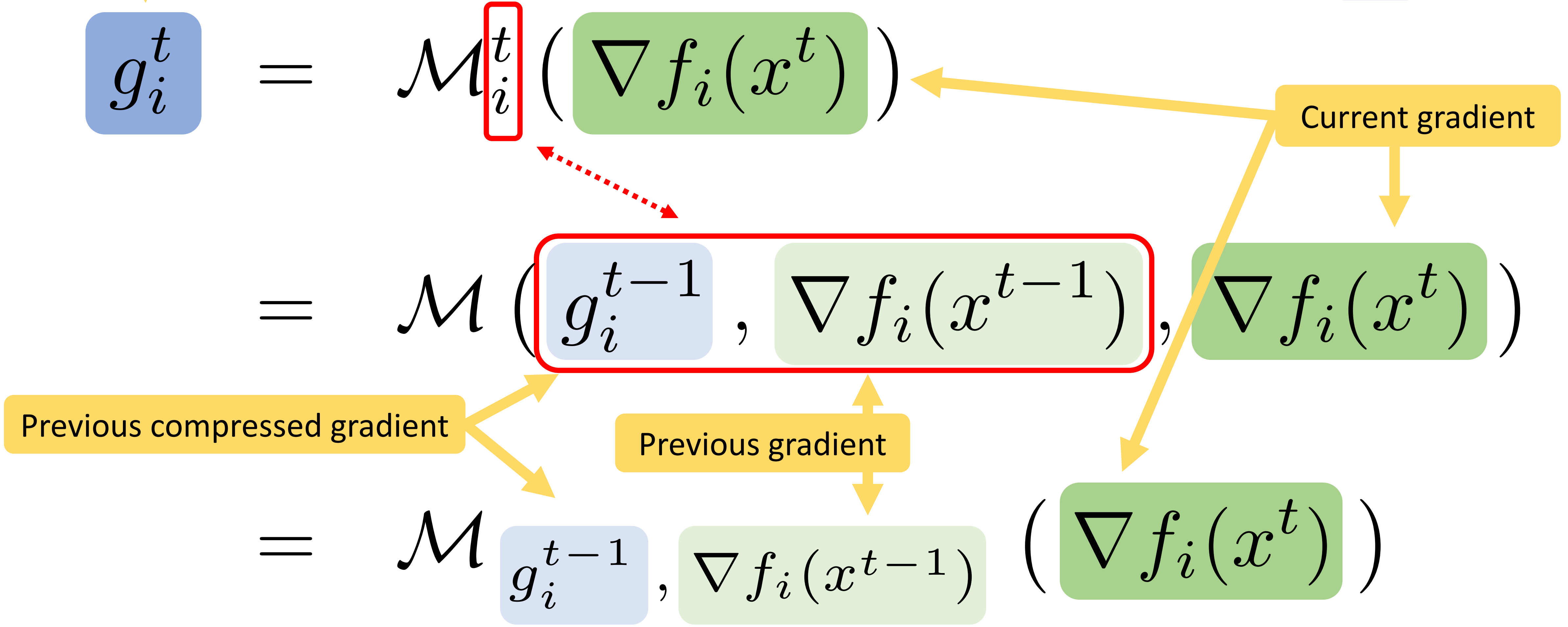
Previous gradient

$$= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)$$

# 3PC: Three Point Compressor

| | $t$ | $t-1$ | |
|---|---|---|---|
| | ① $\nabla f_i(x^t)$ | ② $\nabla f_i(x^{t-1})$ | Exact gradient |
| | $g_i^t$ | ③ $g_i^{t-1}$ | Compressed gradient |

Current compressed gradient



$$g_i^t = \boxed{\mathcal{M}_i^t}\left(\nabla f_i(x^t)\right)$$

$$= \mathcal{M}\left(\boxed{g_i^{t-1}, \nabla f_i(x^{t-1})}, \nabla f_i(x^t)\right)$$

$$= \mathcal{M}_{g_i^{t-1}, \nabla f_i(x^{t-1})}\left(\nabla f_i(x^t)\right)$$

Current gradient

Previous compressed gradient

Previous gradient

# 3PC Inequality

# 3PC Inequality

**Inequality characterizing a Contractive Compressor:**

# 3PC Inequality

**Inequality characterizing a Contractive Compressor:**

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

# 3PC Inequality

**Inequality characterizing a Contractive Compressor:**

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

# 3PC Inequality

**Inequality characterizing a Contractive Compressor:**

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

# 3PC Inequality

**Inequality characterizing a Contractive Compressor:**

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

# 3PC Inequality

**Inequality characterizing a Contractive Compressor:**

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

**Inequality characterizing a 3PC:**

**Inequality characterizing a Contractive Compressor:**

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-A)\|g_i^{t-1} - \nabla f_i(x^{t-1})\|^2 + B\|\nabla f_i(x^t) - \nabla f_i(x^{t-1})\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

$$\mathrm{E}\left[\|g_i^t - \overset{\textcircled{1}}{\nabla f_i(x^t)}\|^2\right] \leq (1-A)\|g_i^{t-1} - \nabla f_i(x^{t-1})\|^2 + B\|\overset{\textcircled{1}}{\nabla f_i(x^t)} - \nabla f_i(x^{t-1})\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

$$\mathrm{E}\left[\|g_i^t - \overset{\textcircled{1}}{\nabla f_i(x^t)}\|^2\right] \leq (1-A)\|g_i^{t-1} - \overset{\textcircled{2}}{\nabla f_i(x^{t-1})}\|^2 + B\|\overset{\textcircled{1}}{\nabla f_i(x^t)} - \overset{\textcircled{2}}{\nabla f_i(x^{t-1})}\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

$$\mathrm{E}\left[\|g_i^t - \overset{\text{①}}{\nabla f_i(x^t)}\|^2\right] \leq (1-A)\|\overset{\text{③}}{g_i^{t-1}} - \overset{\text{②}}{\nabla f_i(x^{t-1})}\|^2 + B\|\overset{\text{①}}{\nabla f_i(x^t)} - \overset{\text{②}}{\nabla f_i(x^{t-1})}\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1-\lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

$$\mathrm{E}\left[\|g_i^t - \overset{①}{\nabla f_i(x^t)}\|^2\right] \leq (1 - A)\|\overset{③}{g_i^{t-1}} - \overset{②}{\nabla f_i(x^{t-1})}\|^2 + B\|\overset{①}{\nabla f_i(x^t)} - \overset{②}{\nabla f_i(x^{t-1})}\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1 - \lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1 - \lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

*A*, *B* : Parameters characterizing a 3PC

$$\mathrm{E}\left[\|g_i^t - \overset{\textcircled{1}}{\nabla f_i(x^t)}\|^2\right] \leq (1 - A)\|\overset{\textcircled{3}}{g_i^{t-1}} - \overset{\textcircled{2}}{\nabla f_i(x^{t-1})}\|^2 + B\|\overset{\textcircled{1}}{\nabla f_i(x^t)} - \overset{\textcircled{2}}{\nabla f_i(x^{t-1})}\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1 - \lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1 - \lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

*A*, *B* : Parameters characterizing a 3PC

$$\mathrm{E}\left[\underbrace{\|g_i^t - \overset{\text{①}}{\nabla f_i(x^t)}\|^2}_{\text{Compression error at iteration } t}\right] \leq (1 - A)\|\overset{\text{③}}{g_i^{t-1}} - \overset{\text{②}}{\nabla f_i(x^{t-1})}\|^2 + B\|\overset{\text{①}}{\nabla f_i(x^t)} - \overset{\text{②}}{\nabla f_i(x^{t-1})}\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1 - \lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1 - \lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

*A*, *B* : Parameters characterizing a 3PC

$$\mathrm{E}\left[\underbrace{\|g_i^t - \overset{①}{\nabla f_i(x^t)}\|^2}_{\text{Compression error at iteration } t}\right] \leq (1 - A)\underbrace{\|\overset{③}{g_i^{t-1}} - \overset{②}{\nabla f_i(x^{t-1})}\|^2}_{\text{Compression error at iteration } t \text{ - } 1} + B\|\overset{①}{\nabla f_i(x^t)} - \overset{②}{\nabla f_i(x^{t-1})}\|^2$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1 - \lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1 - \lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

# 3PC Inequality

## Inequality characterizing a 3PC:

*A*, *B* : Parameters characterizing a 3PC

$$\mathrm{E}\left[\underbrace{\|g_i^t - \overset{\text{①}}{\nabla f_i(x^t)}\|^2}_{\text{Compression error at iteration } t}\right] \leq (1 - A)\underbrace{\|\overset{\text{③}}{g_i^{t-1}} - \overset{\text{②}}{\nabla f_i(x^{t-1})}\|^2}_{\text{Compression error at iteration } t \text{ - } 1} + B\|\overset{\text{①}}{\nabla f_i(x^t)} - \overset{\text{②}}{\nabla f_i(x^{t-1})}\|^2$$

$$\mathrm{E}\left[\|\mathcal{M}_{h,y}(x) - x\|^2\right] \leq (1 - A)\|h - y\|^2 + B\|x - y\|^2$$

$$\forall h, y, x \in \mathbb{R}^d$$

## Inequality characterizing a Contractive Compressor:

$$\mathrm{E}\left[\|g_i^t - \nabla f_i(x^t)\|^2\right] \leq (1 - \lambda)\|\nabla f_i(x^t)\|^2$$

$$\mathrm{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1 - \lambda)\|x\|^2 \qquad \forall x \in \mathbb{R}^d$$

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | |
|---|---|
| | |
| | |
| | |
| | |
| | |

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| $g_i^t \equiv \nabla f_i(x^t)$ | Gradient Descent (GD) |
|---|---|
| | |
| | |
| | |
| | |

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | | |
|---|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$ | + fast $O$(1/t) convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| | | |
| | | |
| | | |
| | | |

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | |
|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$<br>+ fast $O$(1/t) convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$ | **Compressed Gradient Descent (CGD)** |
| | |
| | |
| | |

$\mathrm{E}\,\|\mathcal{C}(u) - u\|^2 \leq (1-\lambda)\|u\|^2, \quad \forall u \in \mathbb{R}^d$

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | |
|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$<br>+ fast $O(1/t)$ convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$<br>- diverges!<br>+ communicates little<br>$\mathrm{E}\,\|\mathcal{C}(u)-u\|^2 \le (1-\lambda)\|u\|^2, \quad \forall u \in \mathbb{R}^d$ | **Compressed Gradient Descent (CGD)** |
| | |
| | |
| | |

$$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$$

# Examples of Compression Mechanisms

| | | |
|---|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$ | + fast *O*(1/t) convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$ | - diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$ | | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| | | |
| | | |

$$\mathrm{E}\,\|\mathcal{C}(u) - u\|^2 \le (1-\lambda)\|u\|^2, \quad \forall u \in \mathbb{R}^d$$

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | | |
|---|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$ | + fast *O*(1/t) convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$ | - diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$ | + fast *O*(1/t) convergence<br>+ communicates little | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| | | |
| | | |

$\mathrm{E}\,\|\mathcal{C}(u) - u\|^2 \leq (1-\lambda)\|u\|^2, \quad \forall u \in \mathbb{R}^d$

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | | |
|---|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$ | + fast *O*(1/t) convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$ | - diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$ | + fast *O*(1/t) convergence<br>+ communicates little | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| $g_i^t \equiv \begin{cases} \nabla f_i(x^t) & \text{if} \quad \|\nabla f_i(x^t) - g_i^{t-1}\|^2 > \zeta\|\nabla f_i(x^t) - \nabla f_i(x^{t-1})\|^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | | **Lazily Aggregated Gradient (LAG)**<br>[Chen, Giannakis, Sun, Yin @ NeurIPS 2018] |
| | | |

$$\mathrm{E}\,\|\mathcal{C}(u) - u\|^2 \leq (1-\lambda)\|u\|^2, \quad \forall u \in \mathbb{R}^d$$

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | |
|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$<br>+ fast $O(1/t)$ convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$<br>- diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$<br>+ fast $O(1/t)$ convergence<br>+ communicates little | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| $g_i^t \equiv \begin{cases} \nabla f_i(x^t) & \text{if} \quad \Vert\nabla f_i(x^t) - g_i^{t-1}\Vert^2 > \zeta \Vert\nabla f_i(x^t) - \nabla f_i(x^{t-1})\Vert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | **Lazily Aggregated Gradient (LAG)**<br>[Chen, Giannakis, Sun, Yin @ NeurIPS 2018] |
| | |

$\mathrm{E}\, \Vert\mathcal{C}(u) - u\Vert^2 \le (1-\lambda)\Vert u\Vert^2, \quad \forall u \in \mathbb{R}^d$

Communicates when a trigger is fired!

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | | |
|---|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$ | + fast $O(1/t)$ convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$ | - diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$ | + fast $O(1/t)$ convergence<br>+ communicates little | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| $g_i^t \equiv \begin{cases} \nabla f_i(x^t) & \text{if} \quad \lVert \nabla f_i(x^t) - g_i^{t-1} \rVert^2 > \zeta \lVert \nabla f_i(x^t) - \nabla f_i(x^{t-1}) \rVert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | - convergence not understood<br>+ communication savings unclear | **Lazily Aggregated Gradient (LAG)**<br>[Chen, Giannakis, Sun, Yin @ NeurIPS 2018] |
| | | |

$\mathrm{E}\, \lVert \mathcal{C}(u) - u \rVert^2 \leq (1-\lambda) \lVert u \rVert^2, \quad \forall u \in \mathbb{R}^d$

Communicates when a trigger is fired!

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | | |
|---|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$ | + fast $O(1/t)$ convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$ | - diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$ | + fast $O(1/t)$ convergence<br>+ communicates little | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| $g_i^t \equiv \begin{cases} \nabla f_i(x^t) & \text{if} \quad \lVert \nabla f_i(x^t) - g_i^{t-1} \rVert^2 > \zeta \lVert \nabla f_i(x^t) - \nabla f_i(x^{t-1}) \rVert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | - convergence not understood<br>+ communication savings unclear | **Lazily Aggregated Gradient (LAG)**<br>[Chen, Giannakis, Sun, Yin @ NeurIPS 2018] |
| $g_i^t \equiv \begin{cases} g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right) & \text{if} \quad \lVert \nabla f_i(x^t) - g_i^{t-1} \rVert^2 > \zeta \lVert \nabla f_i(x^t) - \nabla f_i(x^{t-1}) \rVert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | | **CLAG = EF21 + LAG**<br>**(Lazily Aggregated EF21)**<br>[NEW @ ICML 2022] |

$\mathrm{E}\, \lVert \mathcal{C}(u) - u \rVert^2 \leq (1-\lambda) \lVert u \rVert^2, \quad \forall u \in \mathbb{R}^d$

Communicates when a trigger is fired!

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | |
|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$<br>+ fast *O*(1/t) convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$<br>- diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$<br>+ fast *O*(1/t) convergence<br>+ communicates little | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| $g_i^t \equiv \begin{cases} \nabla f_i(x^t) & \text{if} \quad \lVert \nabla f_i(x^t) - g_i^{t-1} \rVert^2 > \zeta \lVert \nabla f_i(x^t) - \nabla f_i(x^{t-1}) \rVert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$<br>- convergence not understood<br>+ communication savings unclear | **Lazily Aggregated Gradient (LAG)**<br>[Chen, Giannakis, Sun, Yin @ NeurIPS 2018] |
| $g_i^t \equiv \begin{cases} g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right) & \text{if} \quad \lVert \nabla f_i(x^t) - g_i^{t-1} \rVert^2 > \zeta \lVert \nabla f_i(x^t) - \nabla f_i(x^{t-1}) \rVert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | **CLAG = EF21 + LAG**<br>**(Lazily Aggregated EF21)**<br>[NEW @ ICML 2022] |

$\mathrm{E}\,\|\mathcal{C}(u) - u\|^2 \leq (1-\lambda)\|u\|^2, \quad \forall u \in \mathbb{R}^d$

Communicates when a trigger is fired!

$g_i^t \equiv \mathcal{M}_i^t(\nabla f_i(x^t))$

# Examples of Compression Mechanisms

| | | |
|---|---|---|
| $g_i^t \equiv \nabla f_i(x^t)$ | + fast $O(1/t)$ convergence<br>- communicates a lot | **Gradient Descent (GD)** |
| $g_i^t \equiv \mathcal{C}(\nabla f_i(x^t))$ | - diverges!<br>+ communicates little | **Compressed Gradient Descent (CGD)** |
| $g_i^t \equiv g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right)$ | + fast $O(1/t)$ convergence<br>+ communicates little | **Error Feedback 2021 (EF21)**<br>[R., Sokolov, Fatkhullin @ NeurIPS 2021] |
| $g_i^t \equiv \begin{cases} \nabla f_i(x^t) & \text{if} \quad \lVert \nabla f_i(x^t) - g_i^{t-1} \rVert^2 > \zeta \lVert \nabla f_i(x^t) - \nabla f_i(x^{t-1}) \rVert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | - convergence not understood<br>+ communication savings unclear | **Lazily Aggregated Gradient (LAG)**<br>[Chen, Giannakis, Sun, Yin @ NeurIPS 2018] |
| $g_i^t \equiv \begin{cases} g_i^{t-1} + \mathcal{C}\left(\nabla f_i(x^t) - g_i^{t-1}\right) & \text{if} \quad \lVert \nabla f_i(x^t) - g_i^{t-1} \rVert^2 > \zeta \lVert \nabla f_i(x^t) - \nabla f_i(x^{t-1}) \rVert^2 \\ g_i^{t-1} & \text{otherwise} \end{cases}$ | + combines benefits of EF21 and LAG | **CLAG = EF21 + LAG**<br>**(Lazily Aggregated EF21)**<br>[NEW @ ICML 2022] |

$\mathrm{E}\,\|\mathcal{C}(u) - u\|^2 \le (1-\lambda)\|u\|^2, \quad \forall u \in \mathbb{R}^d$

Communicates when a trigger is fired!

# Three Point Compressors: Theory

# Three Point Compressors: Theory

**Theorem**

If $\mathcal{M}_i^t$ is a 3PC with parameters $A$ and $B$, then the method

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

satisfies

# Three Point Compressors: Theory

**Theorem**

If $\mathcal{M}_i^t$ is a 3PC with parameters $A$ and $B$, then the method

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

satisfies

$$\mathrm{E}\|\nabla f(\hat{x}^t)\|^2 \leq \frac{2(f(x^0) - \inf_x f(x))}{\gamma t} + \frac{\frac{1}{n}\sum\limits_{i=1}^{n} \mathrm{E}\|g_i^0 - \nabla f_i(x^0)\|^2}{At}$$

# Three Point Compressors: Theory

**Theorem**

If $\mathcal{M}_i^t$ is a 3PC with parameters $A$ and $B$, then the method

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

satisfies

$$\mathrm{E}\|\nabla f(\hat{x}^t)\|^2 \leq \frac{2(f(x^0) - \inf_x f(x))}{\gamma t} + \frac{\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\|g_i^0 - \nabla f_i(x^0)\|^2}{At}$$

# Three Point Compressors: Theory

**Theorem**

If $\mathcal{M}_i^t$ is a 3PC with parameters $A$ and $B$, then the method

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

$g_i^0 = \mathcal{M}_i^0(\nabla f_i(x^0))$

satisfies

$$\mathrm{E}\|\nabla f(\hat{x}^t)\|^2 \leq \frac{2(f(x^0) - \inf_x f(x))}{\gamma t} + \frac{\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\|g_i^0 - \nabla f_i(x^0)\|^2}{At}$$

# Three Point Compressors: Theory

**Theorem**

If $\mathcal{M}_i^t$ is a 3PC with parameters $A$ and $B$, then the method

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

$g_i^0 = \mathcal{M}_i^0(\nabla f_i(x^0))$

satisfies

$$\mathrm{E}\|\nabla f(\hat{x}^t)\|^2 \leq \frac{2(f(x^0) - \inf_x f(x))}{\gamma t} + \frac{\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\|g_i^0 - \nabla f_i(x^0)\|^2}{At}$$

stepsize $0 < \gamma \leq \left(L_- + L_+\sqrt{\frac{B}{A}}\right)^{-1}$

$$\|\nabla f(x) - \nabla f(y)\| \leq L_- \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_+^2 \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d$$

# Three Point Compressors: Theory

**Theorem**

If $\mathcal{M}_i^t$ is a 3PC with parameters $A$ and $B$, then the method

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

$$g_i^0 = \mathcal{M}_i^0(\nabla f_i(x^0))$$

satisfies

$$\mathrm{E}\|\nabla f(\hat{x}^t)\|^2 \leq \frac{2(f(x^0) - \inf_x f(x))}{\gamma t} + \frac{\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\|g_i^0 - \nabla f_i(x^0)\|^2}{At}$$

stepsize $0 < \gamma \leq \left(L_- + L_+ \sqrt{\frac{B}{A}}\right)^{-1}$

$$\|\nabla f(x) - \nabla f(y)\| \leq L_- \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_+^2 \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d$$

**Corollary**

$$t = \mathcal{O}\left(\frac{(f(x^0) - \inf_x f(x))\left(L_- + L_+\sqrt{\frac{B}{A}}\right)}{\varepsilon}\right) \quad \Rightarrow \quad \mathrm{E}\|\nabla f(\hat{x}^t)\|^2 \leq \varepsilon$$

# Better Rates for Lazy Aggregation

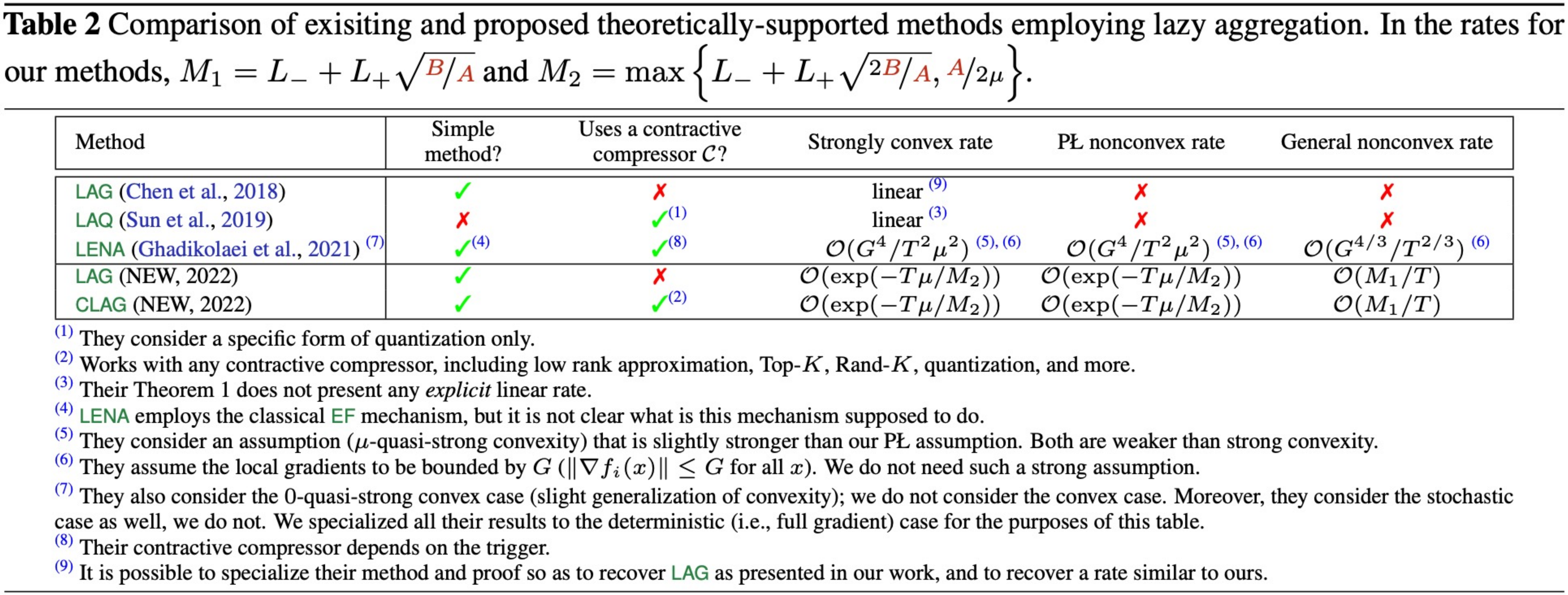


**Table 2** Comparison of exisiting and proposed theoretically-supported methods employing lazy aggregation. In the rates for our methods, $M_1 = L_- + L_+\sqrt{B/A}$ and $M_2 = \max\left\{L_- + L_+\sqrt{2B/A}, A/2\mu\right\}$.

| Method | Simple method? | Uses a contractive compressor $\mathcal{C}$? | Strongly convex rate | PŁ nonconvex rate | General nonconvex rate |
|---|---|---|---|---|---|
| LAG (Chen et al., 2018) | ✓ | ✗ | linear (9) | ✗ | ✗ |
| LAQ (Sun et al., 2019) | ✗ | ✓(1) | linear (3) | ✗ | ✗ |
| LENA (Ghadikolaei et al., 2021) (7) | ✓(4) | ✓(8) | $\mathcal{O}(G^4/T^2\mu^2)$ (5), (6) | $\mathcal{O}(G^4/T^2\mu^2)$ (5), (6) | $\mathcal{O}(G^{4/3}/T^{2/3})$ (6) |
| LAG (NEW, 2022) | ✓ | ✗ | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(M_1/T)$ |
| CLAG (NEW, 2022) | ✓ | ✓(2) | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(M_1/T)$ |

(1) They consider a specific form of quantization only.

(2) Works with any contractive compressor, including low rank approximation, Top-$K$, Rand-$K$, quantization, and more.

(3) Their Theorem 1 does not present any *explicit* linear rate.

(4) LENA employs the classical EF mechanism, but it is not clear what is this mechanism supposed to do.

(5) They consider an assumption ($\mu$-quasi-strong convexity) that is slightly stronger than our PŁ assumption. Both are weaker than strong convexity.

(6) They assume the local gradients to be bounded by $G$ ($\|\nabla f_i(x)\| \leq G$ for all $x$). We do not need such a strong assumption.

(7) They also consider the 0-quasi-strong convex case (slight generalization of convexity); we do not consider the convex case. Moreover, they consider the stochastic case as well, we do not. We specialized all their results to the deterministic (i.e., full gradient) case for the purposes of this table.

(8) Their contractive compressor depends on the trigger.

(9) It is possible to specialize their method and proof so as to recover LAG as presented in our work, and to recover a rate similar to ours.

# Better Rates for Lazy Aggregation

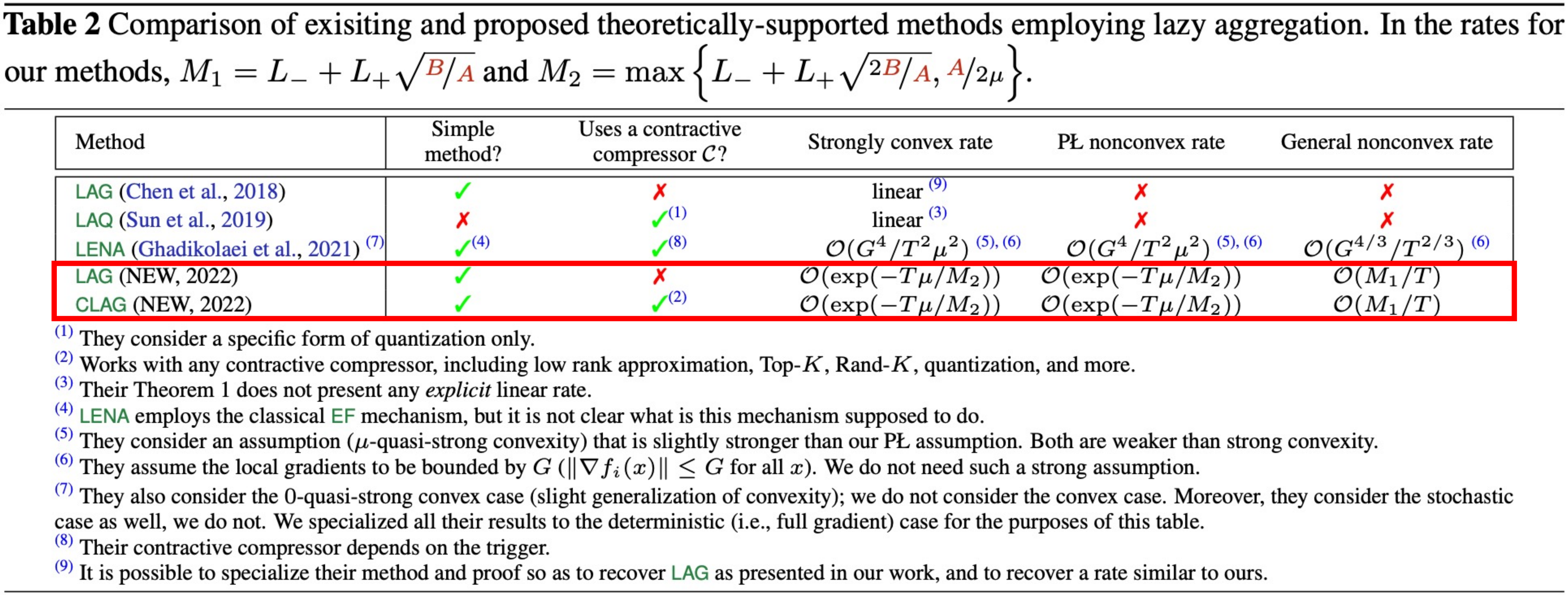


**Table 2** Comparison of exisiting and proposed theoretically-supported methods employing lazy aggregation. In the rates for our methods, $M_1 = L_- + L_+\sqrt{B/A}$ and $M_2 = \max\left\{L_- + L_+\sqrt{2B/A}, A/2\mu\right\}$.

| Method | Simple method? | Uses a contractive compressor $\mathcal{C}$? | Strongly convex rate | PŁ nonconvex rate | General nonconvex rate |
|---|---|---|---|---|---|
| LAG (Chen et al., 2018) | ✓ | ✗ | linear (9) | ✗ | ✗ |
| LAQ (Sun et al., 2019) | ✗ | ✓(1) | linear (3) | ✗ | ✗ |
| LENA (Ghadikolaei et al., 2021) (7) | ✓(4) | ✓(8) | $\mathcal{O}(G^4/T^2\mu^2)$ (5), (6) | $\mathcal{O}(G^4/T^2\mu^2)$ (5), (6) | $\mathcal{O}(G^{4/3}/T^{2/3})$ (6) |
| LAG (NEW, 2022) | ✓ | ✗ | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(M_1/T)$ |
| CLAG (NEW, 2022) | ✓ | ✓(2) | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(\exp(-T\mu/M_2))$ | $\mathcal{O}(M_1/T)$ |

(1) They consider a specific form of quantization only.

(2) Works with any contractive compressor, including low rank approximation, Top-$K$, Rand-$K$, quantization, and more.

(3) Their Theorem 1 does not present any *explicit* linear rate.

(4) LENA employs the classical EF mechanism, but it is not clear what is this mechanism supposed to do.

(5) They consider an assumption ($\mu$-quasi-strong convexity) that is slightly stronger than our PŁ assumption. Both are weaker than strong convexity.

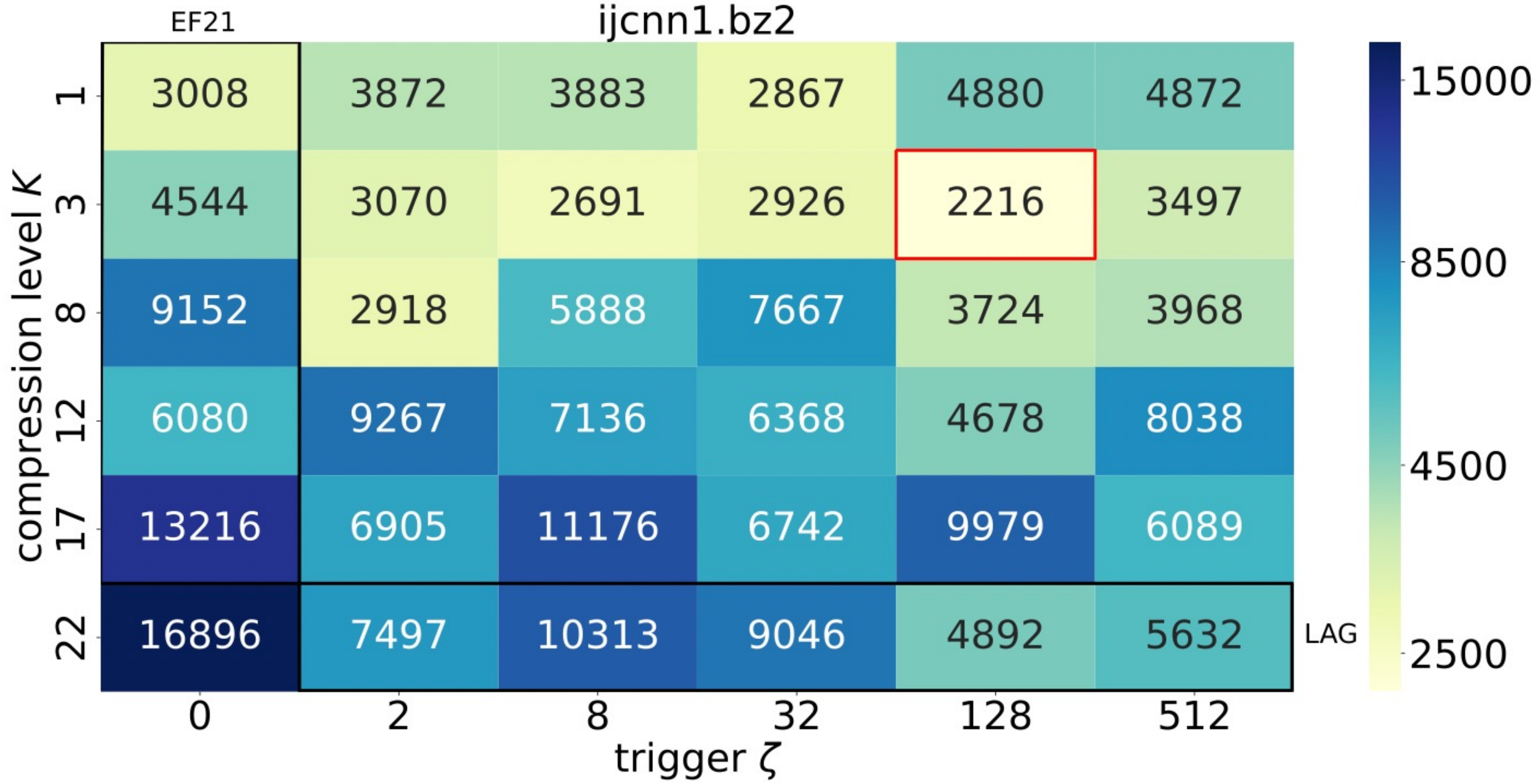
(6) They assume the local gradients to be bounded by $G$ ($\|\nabla f_i(x)\| \leq G$ for all $x$). We do not need such a strong assumption.

(7) They also consider the 0-quasi-strong convex case (slight generalization of convexity); we do not consider the convex case. Moreover, they consider the stochastic case as well, we do not. We specialized all their results to the deterministic (i.e., full gradient) case for the purposes of this table.

(8) Their contractive compressor depends on the trigger.

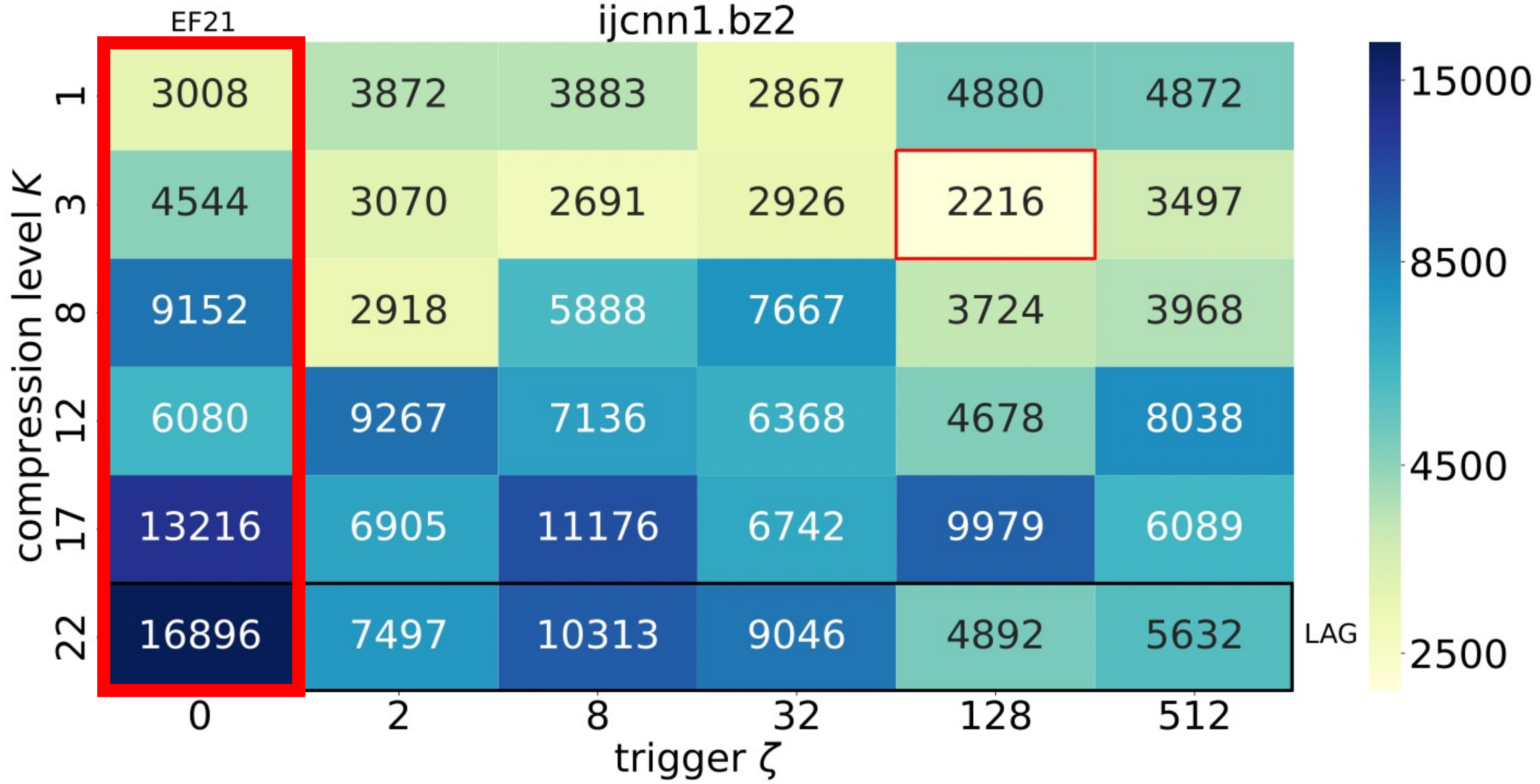
(9) It is possible to specialize their method and proof so as to recover LAG as presented in our work, and to recover a rate similar to ours.

# CLAG: Combining the Benefits of EF21 and LAG



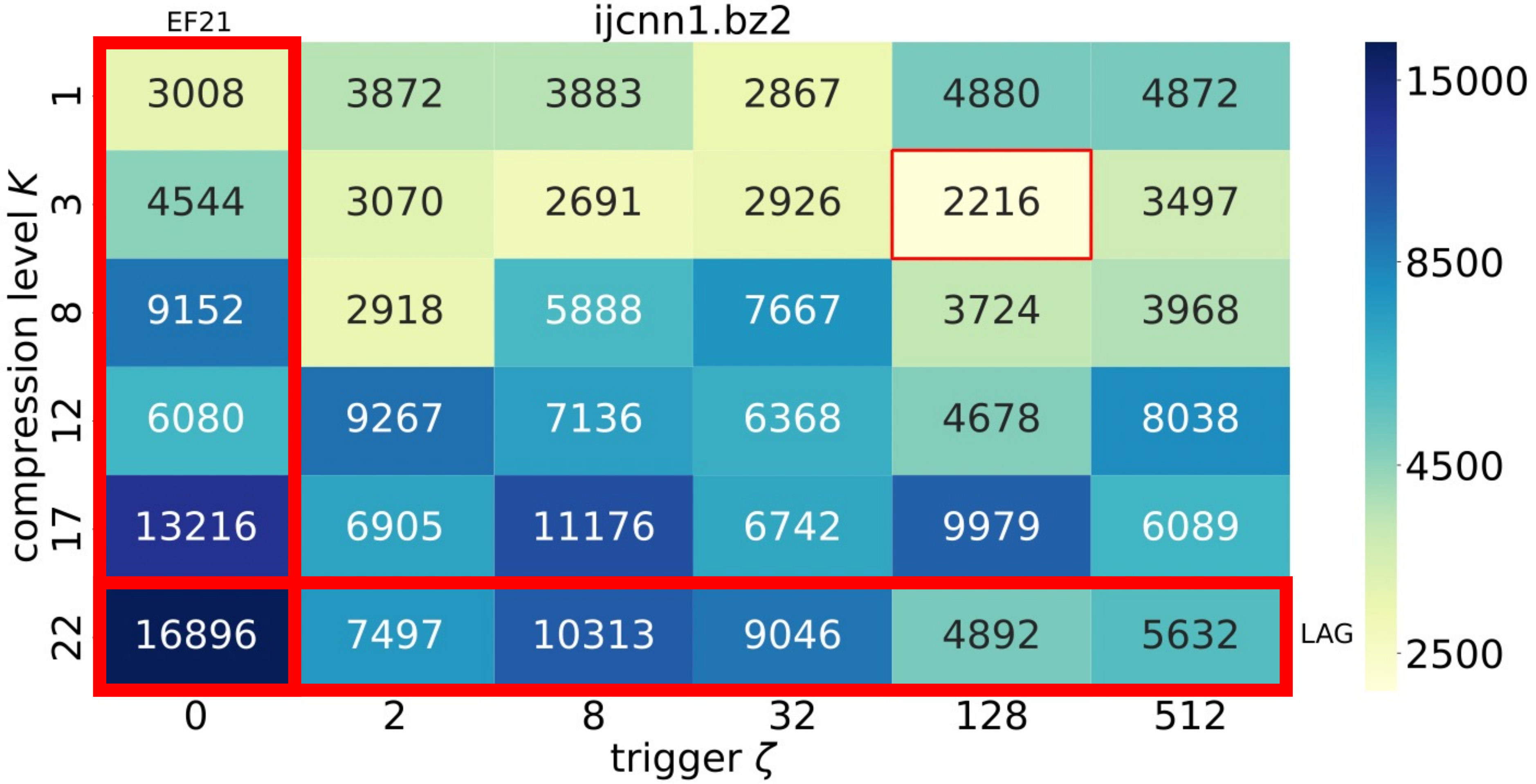
|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  | 0 | 2 | 8 | 32 | 128 | 512 |
| 1 | 3008 | 3872 | 3883 | 2867 | 4880 | 4872 |
| 3 | 4544 | 3070 | 2691 | 2926 | 2216 | 3497 |
| 8 | 9152 | 2918 | 5888 | 7667 | 3724 | 3968 |
| 12 | 6080 | 9267 | 7136 | 6368 | 4678 | 8038 |
| 17 | 13216 | 6905 | 11176 | 6742 | 9979 | 6089 |
| 22 | 16896 | 7497 | 10313 | 9046 | 4892 | 5632 |

# CLAG: Combining the Benefits of EF21 and LAG



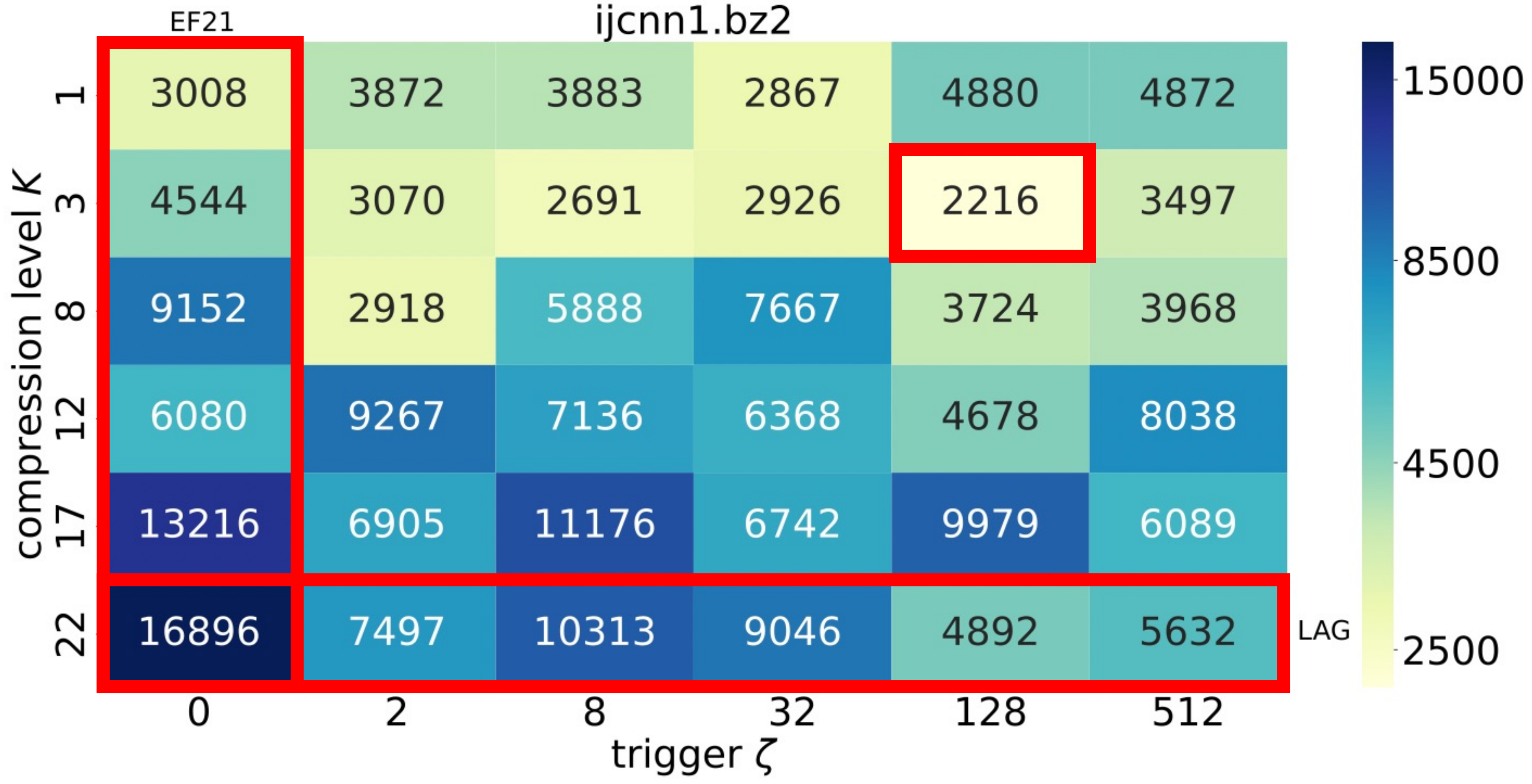
| | 0 | 2 | 8 | 32 | 128 | 512 |
|---|---|---|---|---|---|---|
| 1 | 3008 | 3872 | 3883 | 2867 | 4880 | 4872 |
| 3 | 4544 | 3070 | 2691 | 2926 | 2216 | 3497 |
| 8 | 9152 | 2918 | 5888 | 7667 | 3724 | 3968 |
| 12 | 6080 | 9267 | 7136 | 6368 | 4678 | 8038 |
| 17 | 13216 | 6905 | 11176 | 6742 | 9979 | 6089 |
| 22 | 16896 | 7497 | 10313 | 9046 | 4892 | 5632 |

# CLAG: Combining the Benefits of EF21 and LAG



| | | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 2 | 8 | 32 | 128 | 512 |
| 1 | 3008 | 3872 | 3883 | 2867 | 4880 | 4872 |
| 3 | 4544 | 3070 | 2691 | 2926 | 2216 | 3497 |
| 8 | 9152 | 2918 | 5888 | 7667 | 3724 | 3968 |
| 12 | 6080 | 9267 | 7136 | 6368 | 4678 | 8038 |
| 17 | 13216 | 6905 | 11176 | 6742 | 9979 | 6089 |
| 22 | 16896 | 7497 | 10313 | 9046 | 4892 | 5632 |

# CLAG: Combining the Benefits of EF21 and LAG



| | | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 2 | 8 | 32 | 128 | 512 |
| 1 | 3008 | 3872 | 3883 | 2867 | 4880 | 4872 |
| 3 | 4544 | 3070 | 2691 | 2926 | 2216 | 3497 |
| 8 | 9152 | 2918 | 5888 | 7667 | 3724 | 3968 |
| 12 | 6080 | 9267 | 7136 | 6368 | 4678 | 8038 |
| 17 | 13216 | 6905 | 11176 | 6742 | 9979 | 6089 |
| 22 | 16896 | 7497 | 10313 | 9046 | 4892 | 5632 |

# Summary & Extensions

- 3PC recovers
  - gradient descent (GD) and its rate
  - error feedback (EF21) and its rate
  - lazy aggregation (LAG) & gets a better rate
- 3PC uncovers a hidden link between error feedback and lazy aggregation mechanisms & literature
- 3PC includes
  - new method (CLAG) which combines the benefits of EF21 and LAG
  - several additional new methods (not mentioned in this talk)
- We prove
  - Fast O(1/$t$) rate for smooth nonconvex functions
  - linear rate under the Polyak-Łojasiewicz condition (not mentioned in this talk)