Controlling Delay in Asynchronous SGD: Optimality for Any Data Regime

Artavazd Maranjyan
https://artomaranjyan.github.io

Protocol Learning: Decentralized Collaborative Learning at Scale
ICLR 2026 Workshop
26 April 2026

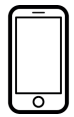
Distributed Learning (Data Parallelism)

$\mathcal{D}_1$

$$f_1(x) := \mathbb{E}_{\xi_1 \sim \mathcal{D}_1}\left[f_1(x;\xi_1)\right]$$

$\mathcal{D}_2$

$$f_2(x) := \mathbb{E}_{\xi_2 \sim \mathcal{D}_2}\left[f_2(x;\xi_2)\right]$$



$\mathcal{D}_3$

$$f_3(x) := \mathbb{E}_{\xi_3 \sim \mathcal{D}_3}\left[f_3(x;\xi_3)\right]$$

Homogenous Data Setting

$$\mathcal{D}_1 = \mathcal{D}_2 = \cdots = \mathcal{D}_n = \mathcal{D}$$

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \left\{ f(x) := \mathbb{E}_{\xi\in\mathcal{D}}\left[f(x;\xi)\right] \right\}$$

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

Server

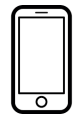
Compute Heterogeneous System

$\nabla f_1(x;\xi_1)$

Compute time = τ_1

$\nabla f_2(x;\xi_2)$

Compute time = τ_2



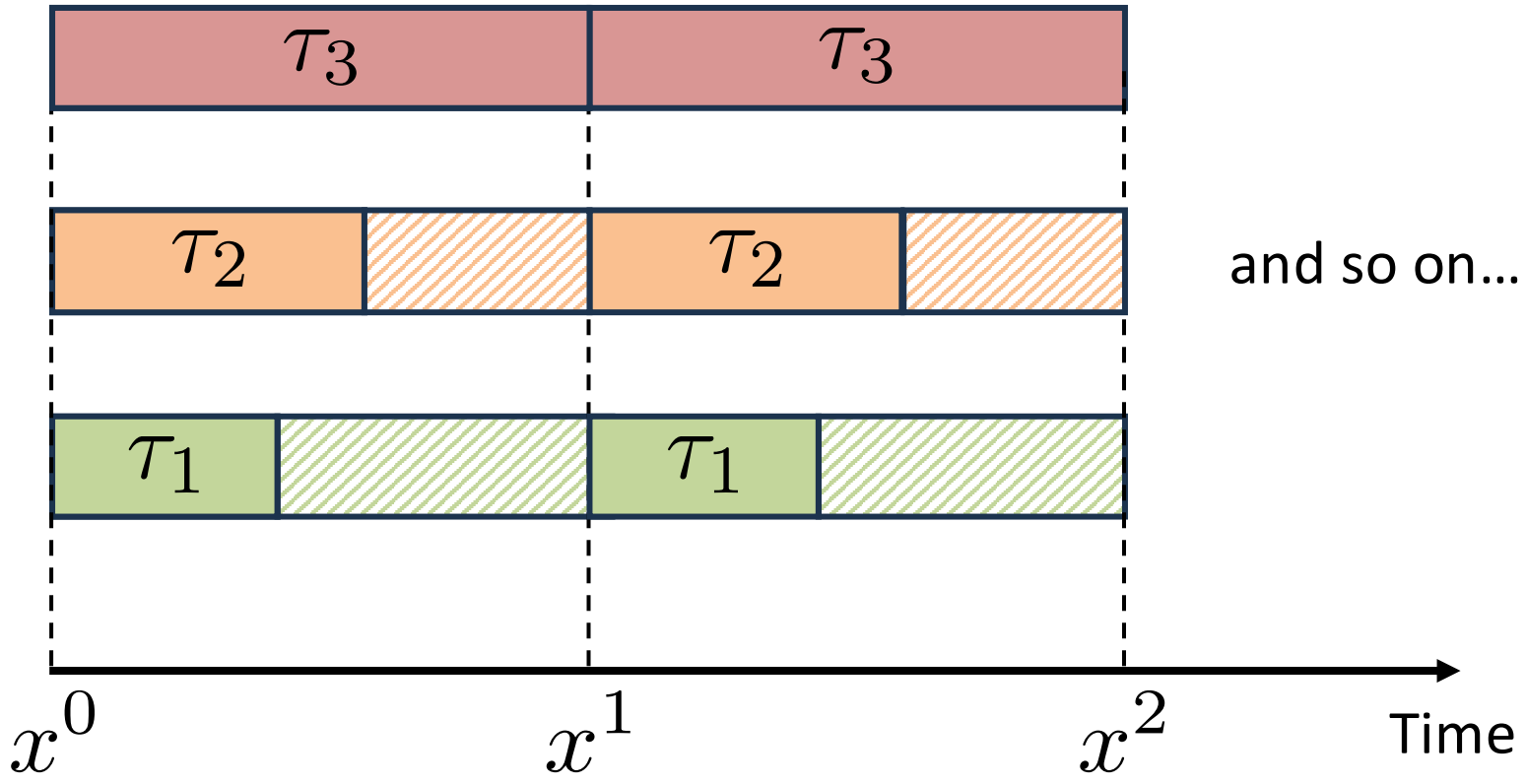
$\nabla f_3(x;\xi_3)$

Compute time = τ_3

Server

Naive Minibatch SGD:
Each worker does one job only

$n = 3$



and so on…

$$x^{k+1} = x^k - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla f_i\left(x^k; \xi_i^k\right)$$

The First Asynchronous SGD

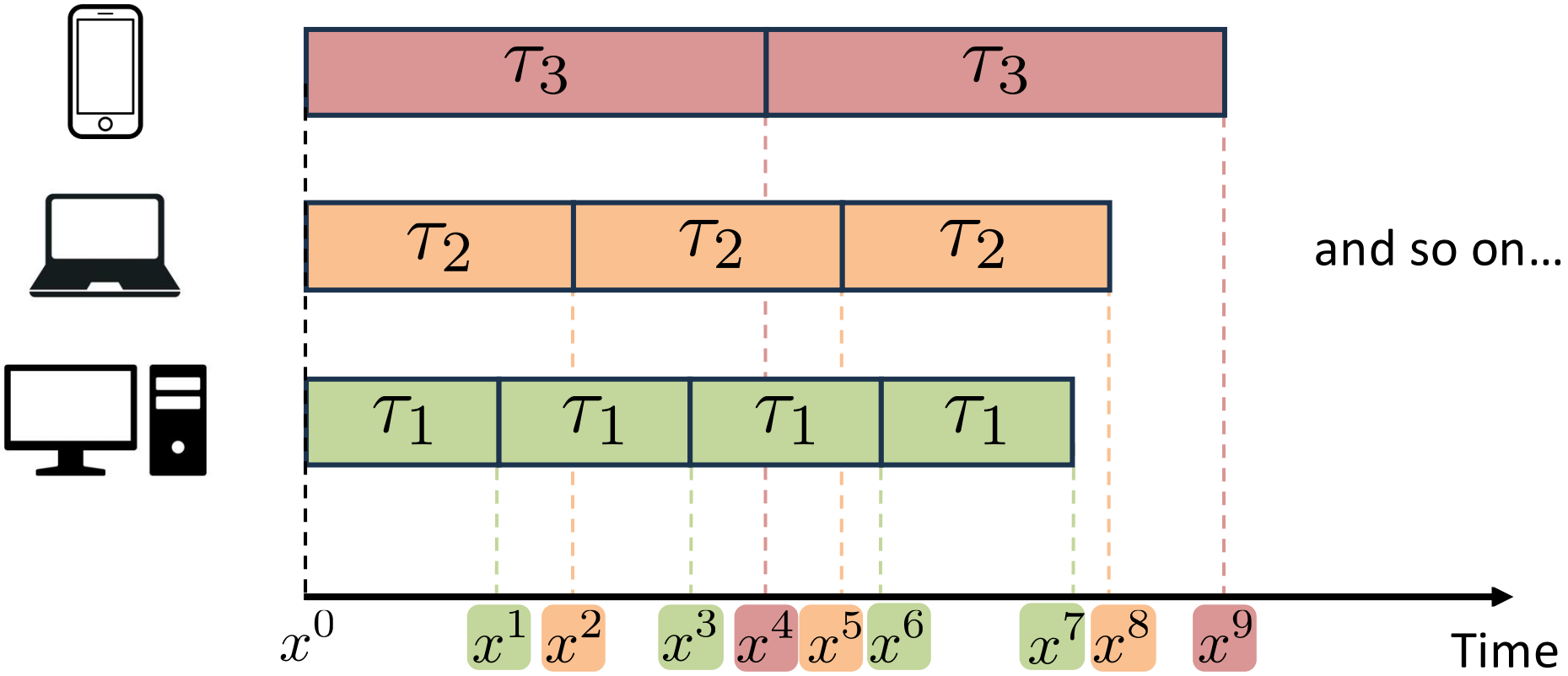
Benjamin Recht, Christopher Re, Stephen Wright, Feng Niu

HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent

Conference on Neural Information Processing Systems (2011)

NeurIPS Test of Time Award, 2020

Asynchronous SGD: Remove the synchronization



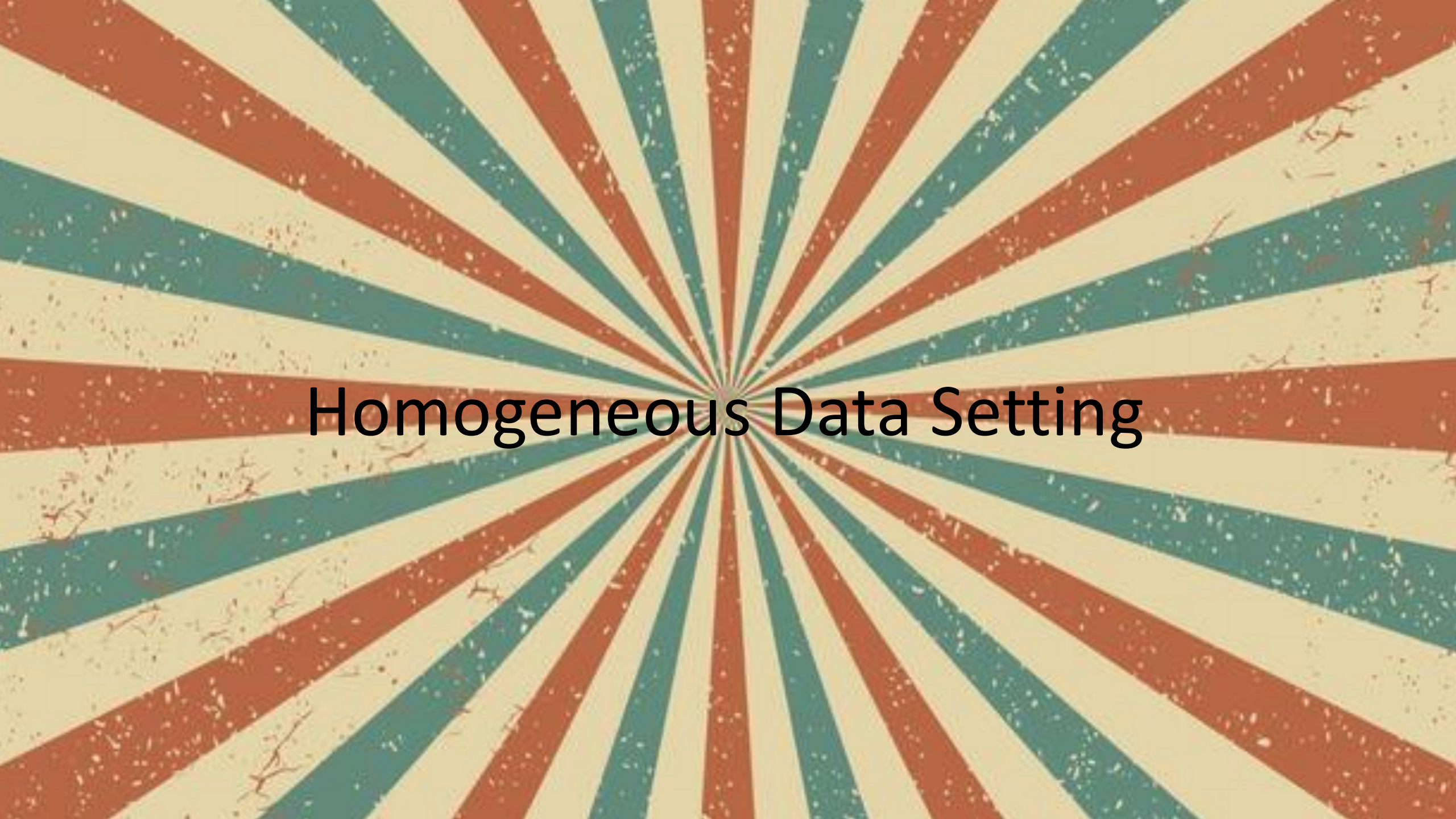
Updates of Asynchronous SGD has delayed stochastic gradients

Delay of worker i^k at iteration k

$$x^{k+1} = x^k - \gamma \nabla f_{i^k}\left(x^{k-\delta^k_{i^k}}; \xi^{k-\delta^k_{i^k}}\right)$$

Uses a gradient from a single worker

Server


Homogeneous Data Setting

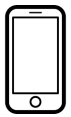
Homogenous Data Setting

$\mathcal{D}$

$$f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x;\xi)]$$

$\mathcal{D}$

$$f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x;\xi)]$$



$\mathcal{D}$

$$f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x;\xi)]$$

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \{f(x) := \mathbb{E}_{\xi \in \mathcal{D}}[f(x;\xi)]\}$$

Server

Asynchronous SGD is too wild: Ringmaster ASGD ***tames*** it

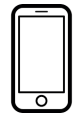
Artavazd Maranjyan, Alexander Tyurin, and Peter Richtárik

Ringmaster ASGD: The first Asynchronous SGD with optimal time complexity

International Conference on Machine Learning (2025).

ICML
International Conference
On Machine Learning

Ringmaster ASGD:
Have a threshold on delays



$$\nabla f\left(x^{k-\delta^k};\xi_i^{k-\delta^k}\right)$$

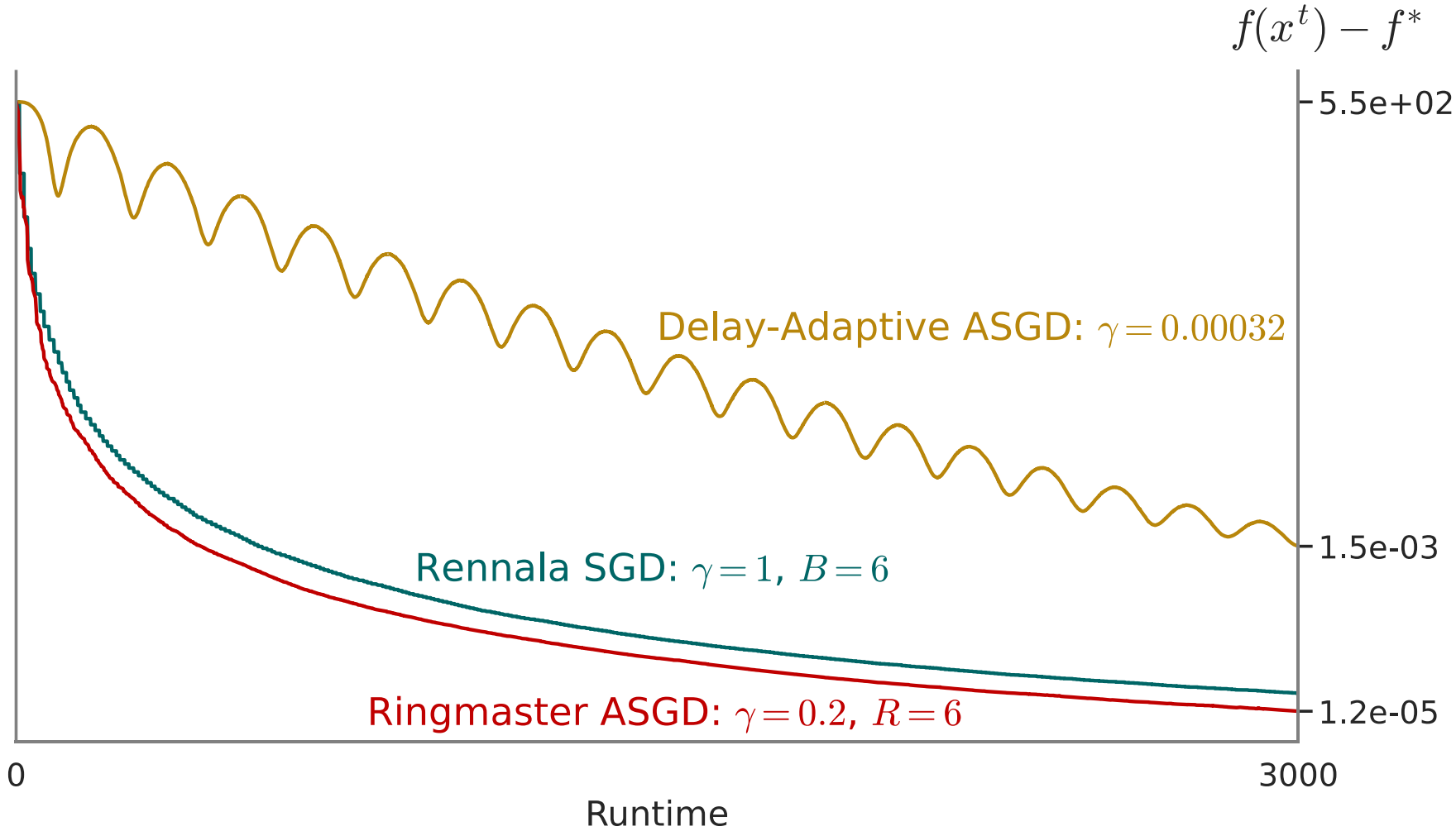
If: $\delta^k < R$

$$x^{k+1} = x^k - \gamma \nabla f\left(x^{k-\delta^k};\xi_i^{k-\delta^k}\right)$$

Else: Ignore the gradient and send the current point x^k to the worker

Server

Ringmaster ASGD outperforms existing baselines




Heterogenous Data Setting

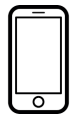
Heterogenous Data Setting

$\mathcal{D}_1$

$$f_1(x) := \mathbb{E}_{\xi_1 \sim \mathcal{D}_1}\left[f_1(x;\xi_1)\right]$$

$\mathcal{D}_2$

$$f_2(x) := \mathbb{E}_{\xi_2 \sim \mathcal{D}_2}\left[f_2(x;\xi_2)\right]$$



$\mathcal{D}_3$

$$f_3(x) := \mathbb{E}_{\xi_3 \sim \mathcal{D}_3}\left[f_3(x;\xi_3)\right]$$

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

Server

Problems with the Naive Approach

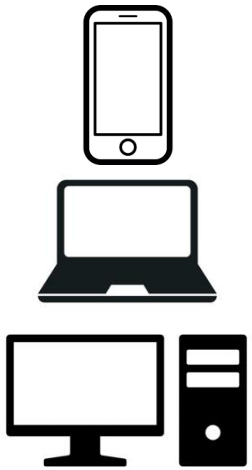
Delay of worker i^k at iteration k

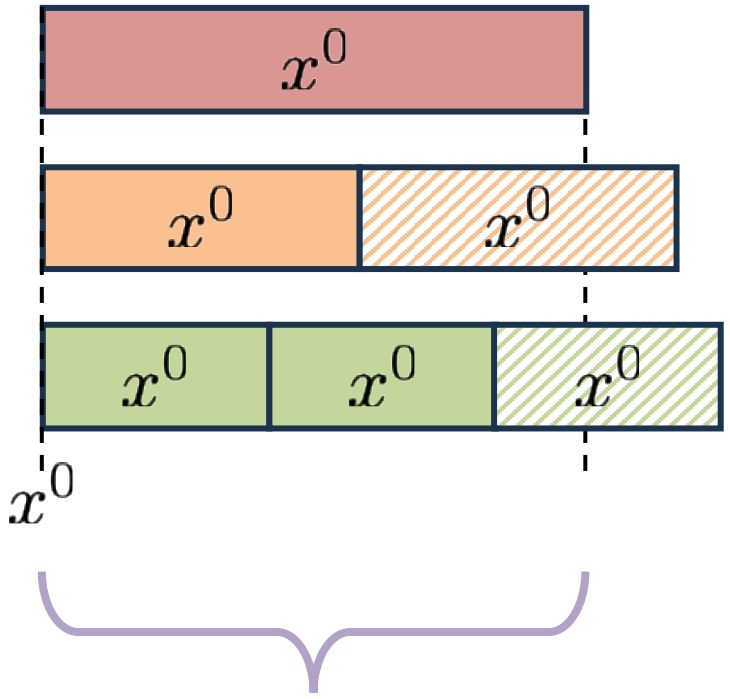
$$x^{k+1} = x^k - \gamma \nabla f_{i^k}\left(x^{k-\delta^k_{i^k}}; \xi^{k-\delta^k_{i^k}}\right)$$

Uses a gradient from a single worker

Server

Ringleader ASGD (Phase 1)





$$G_3 = \nabla f_3\left(x^0; \xi_3^{0,1}\right)$$

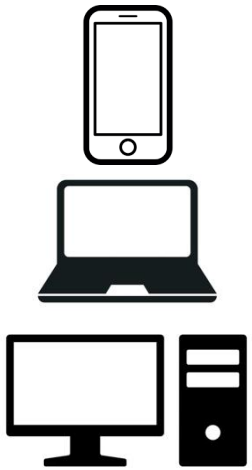
$$G_2 = \nabla f_2\left(x^0; \xi_2^{0,1}\right)$$

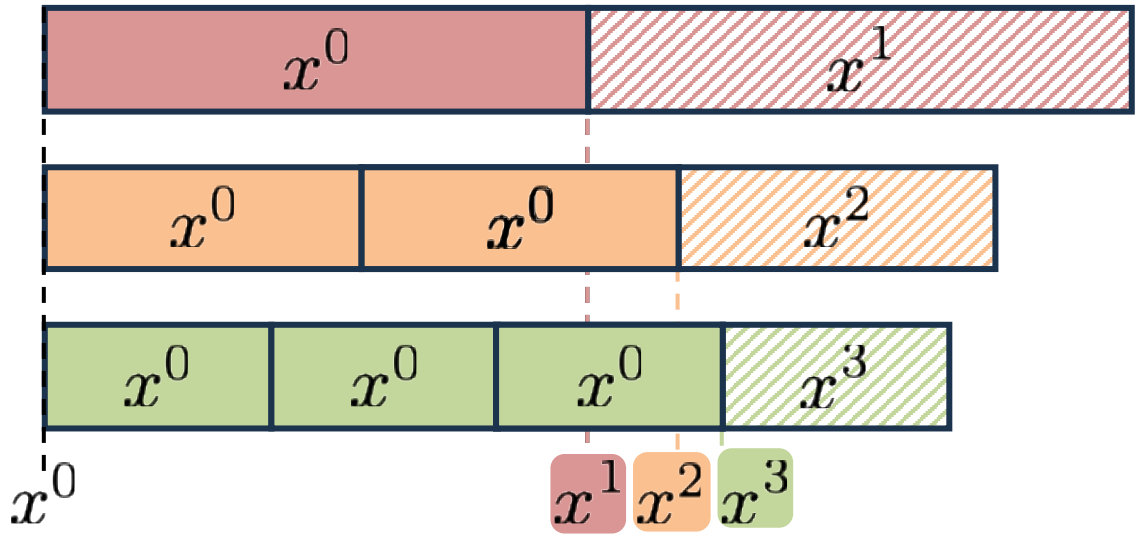
$$G_1 = \sum_{j=1}^{2} \nabla f_1\left(x^0; \xi_1^{0,j}\right)$$

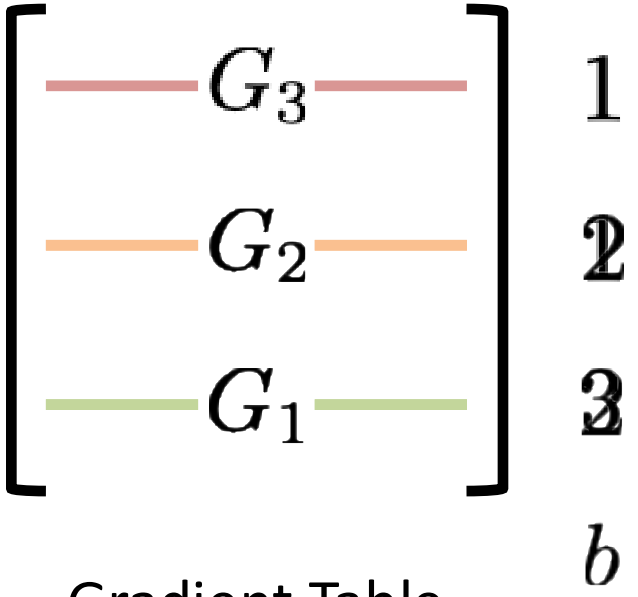
$$\begin{bmatrix} G_3 \\ G_2 \\ G_1 \end{bmatrix} \begin{matrix} \\ \\ \\ b \end{matrix}$$

Gradient Table

Ringleader ASGD (Phase 2)



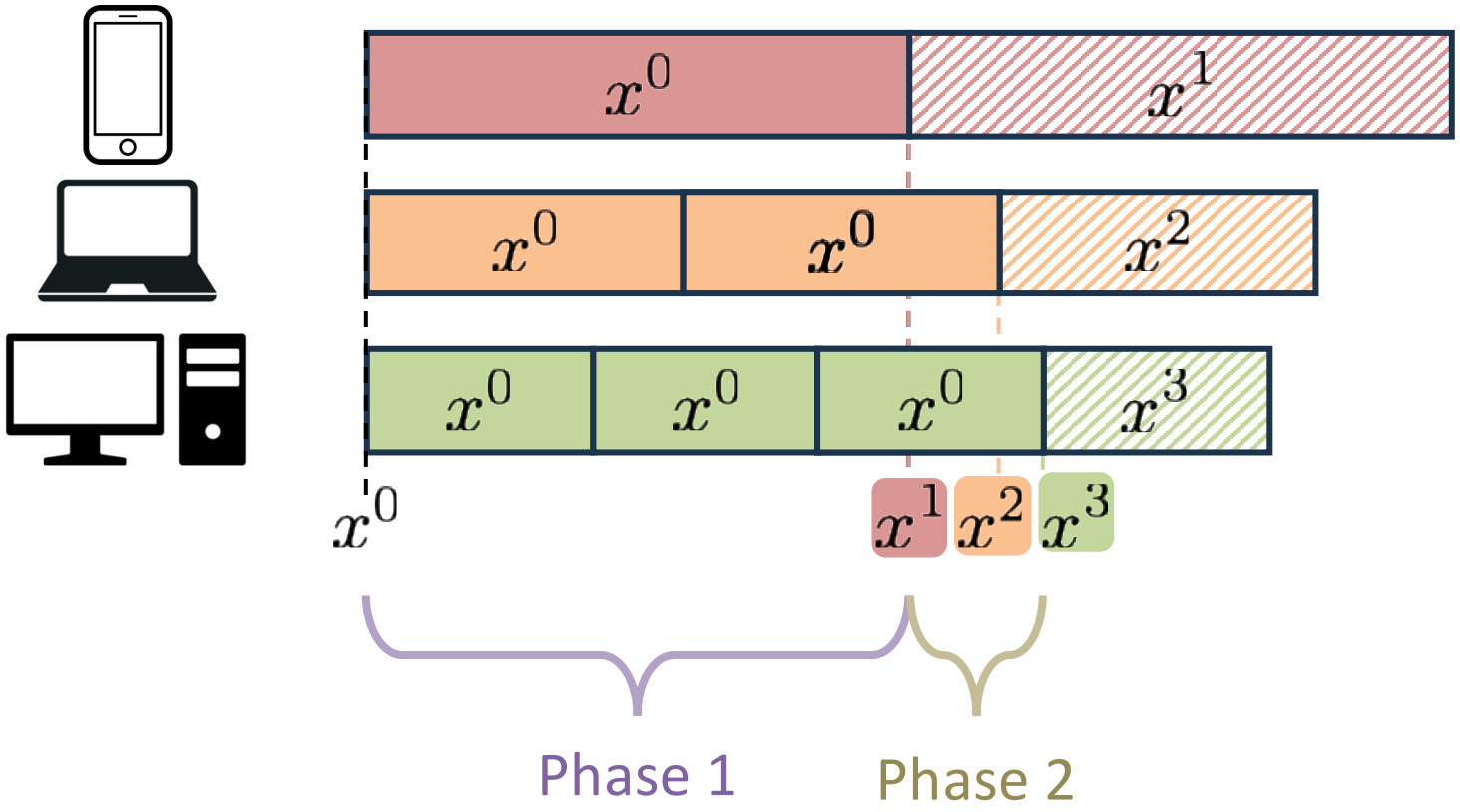




Gradient Table

$$x^3 = x^2 - \gamma \frac{1}{n} \sum_{i=1}^{n} \frac{G_i}{b_i}$$

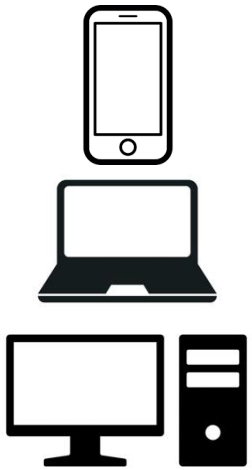
Ringleader ASGD (Transition Phase)

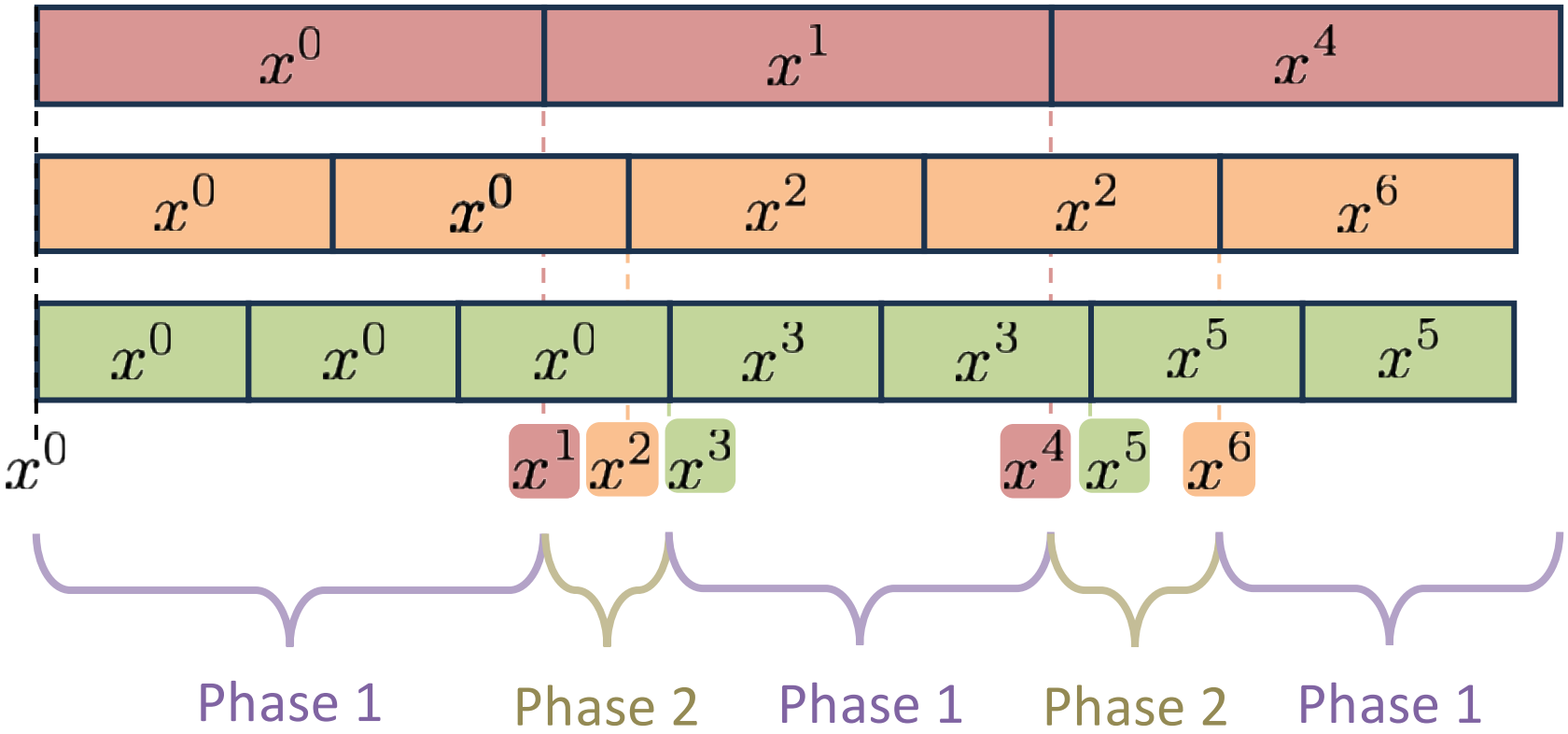


$$\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \begin{matrix} 0 \\ 0 \\ 0 \\ b \end{matrix}$$

Gradient Table

Ringleader ASGD





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

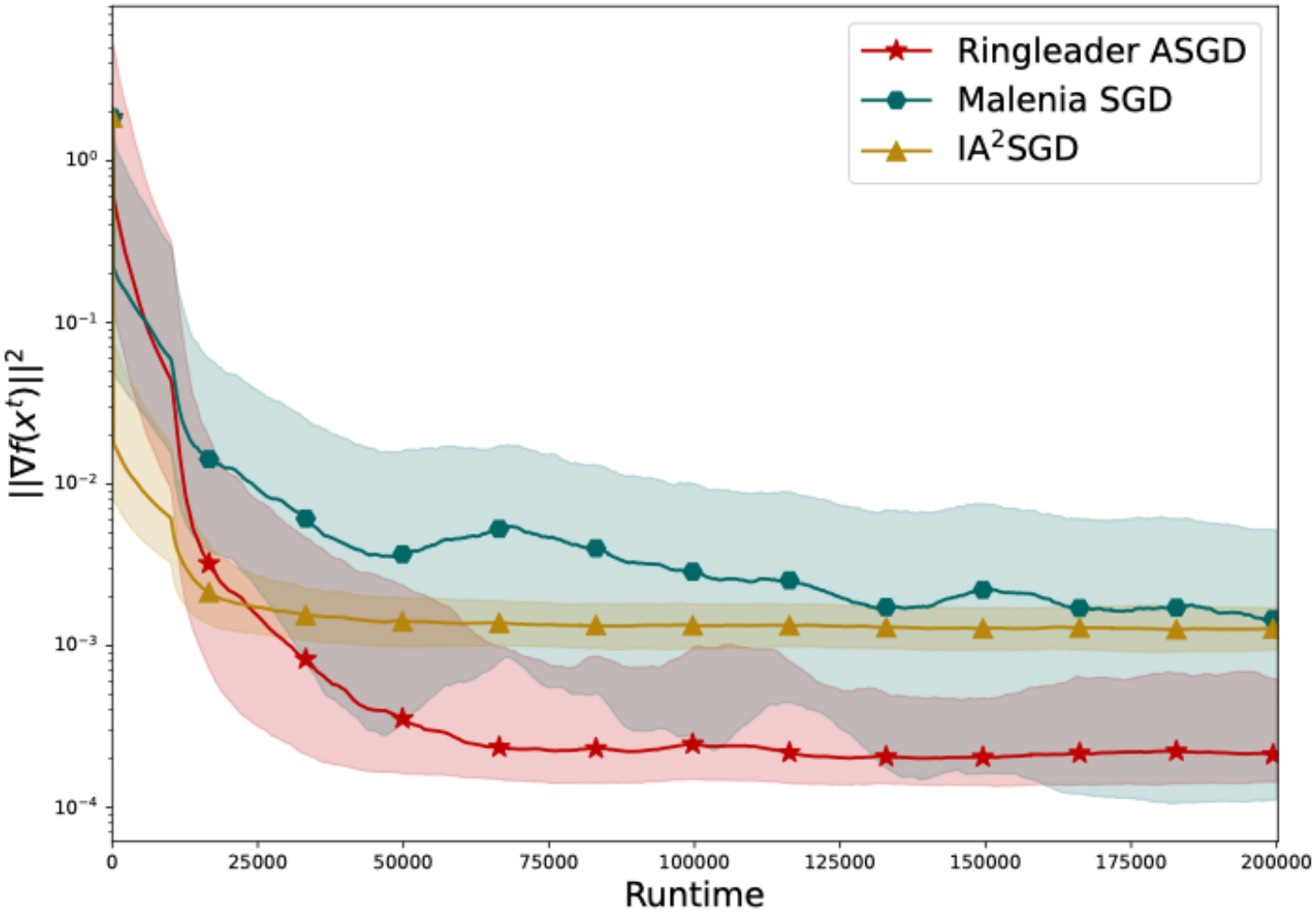
$$\delta_i^k \leq 2n - 2$$

Ringleader ASGD outperforms existing baselines

Two-layer MLP

MNIST

$$n = 100$$
$$\tau_i = i + |\eta_i|$$
$$\eta_i \sim \mathcal{N}(0, i)$$



KHALAS