Ringmaster ASGD: *The First Asynchronous SGD with Optimal Time Complexity*

Artavazd Maranjyan,
Alexander Tyurin, Peter Richtárik

Problem setup

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

Loss of a data sample ξ

The distribution of the training dataset

We have n workers available to work in parallel, all having access to compute stochastic gradients $f(x;\xi)$.
We consider the *fixed computation model*:

worker i takes no more than τ_i seconds to compute a single stochastic gradient.

Without loss of generality, $0 < \tau_1 \le \tau_2 \le \cdots \le \tau_n$

Assumptions

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x - y\|, \ \forall x, y \in \mathbb{R}^d$$

$$\mathbb{E}_\xi [\nabla f(x;\xi)] = \nabla f(x), \ \forall x \in \mathbb{R}^d,$$

$$\mathbb{E}_\xi \left[\|\nabla f(x;\xi) - \nabla f(x)\|^2 \right] \le \sigma^2, \ \forall x \in \mathbb{R}^d$$

$$f(x) \ge f^{\inf} \text{ for all } x \in \mathbb{R}^d$$

Iteration Complexity

$$K = \mathcal{O}\left(\frac{R}{\varepsilon} + \frac{\sigma^2}{\varepsilon^2} \right) \qquad \gamma = \min\left\{ \frac{1}{2RL}, \frac{\varepsilon}{4L\sigma^2} \right\}$$

$$\frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}\left[\|\nabla f(x^k)\|^2 \right] \le \varepsilon$$

Choice of Threshold

$$R = \max\left\{ 1, \left\lceil \frac{\sigma^2}{\varepsilon} \right\rceil \right\}$$

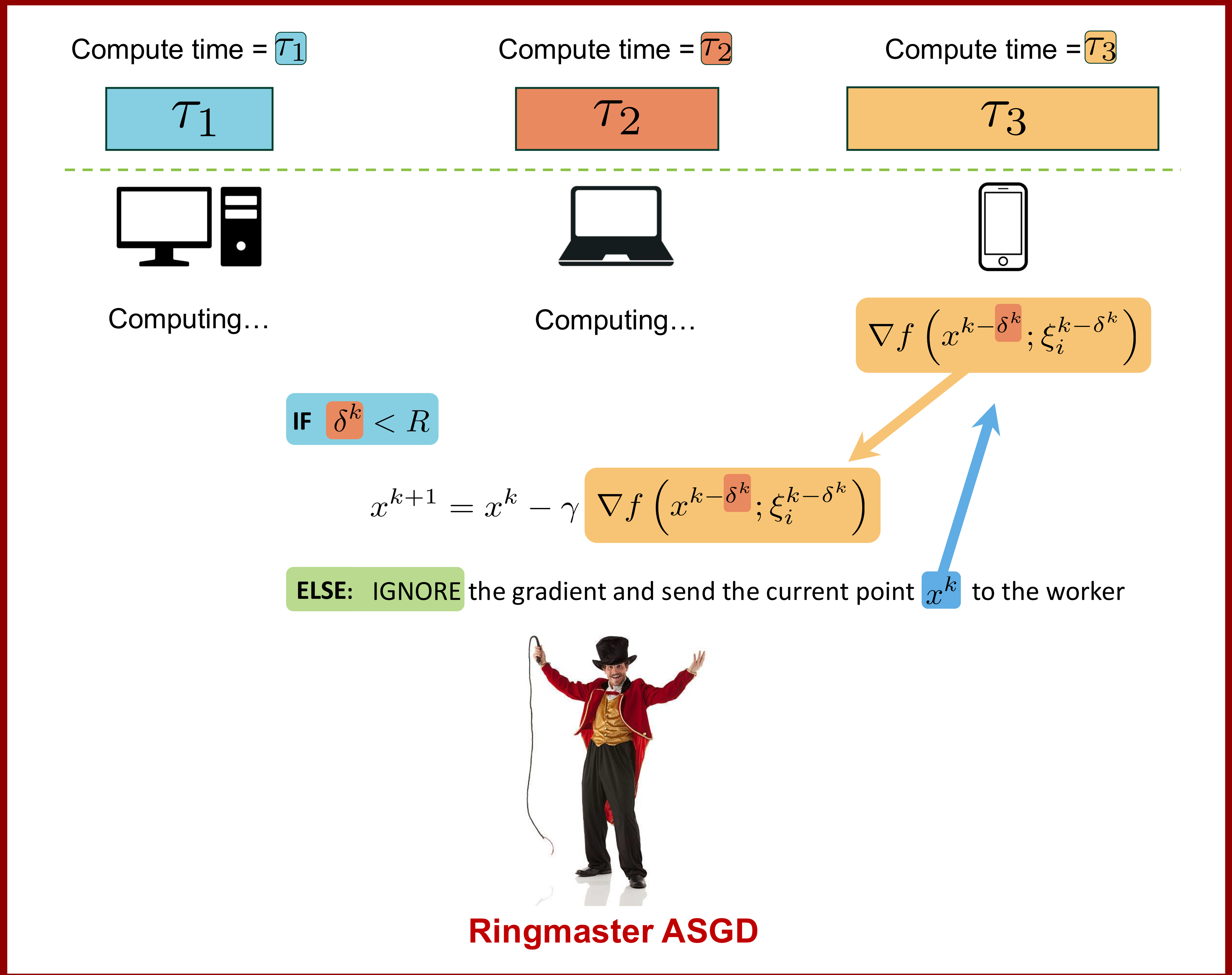
$R = 1$	Optimal Choice	$R = \infty$
Hero SGD		HOGWILD!

Time Complexity

$$\mathcal{O}\left(\min_{m \in [n]} \left[\left(\frac{1}{m} \sum_{i=1}^{m} \frac{1}{\tau_i} \right)^{-1} \left(\frac{1}{\varepsilon} + \frac{\sigma^2}{m\varepsilon^2} \right) \right] \right)$$

non-decreasing | decreasing

First Optimal Asynchronous SGD: Tame the Wild, Ignore Old Gradients, Achieve Optimality



Comparison

Method	Time Complexity
Asynchronous SGD (Koloskova et al., 2022) (Mishchenko et al., 2022)	$\tau_h^n \left(\frac{1}{\varepsilon} + \frac{\sigma^2}{m\varepsilon^2} \right)$
Ringmaster ASGD	$\min_{m \in [n]} \left\{ \tau_h^m \left(\frac{1}{\varepsilon} + \frac{\sigma^2}{m\varepsilon^2} \right) \right\}$
Lower Bound (Tyurin & Richtárik, 2024)	$\min_{m \in [n]} \left\{ \tau_h^m \left(\frac{1}{\varepsilon} + \frac{\sigma^2}{m\varepsilon^2} \right) \right\}$

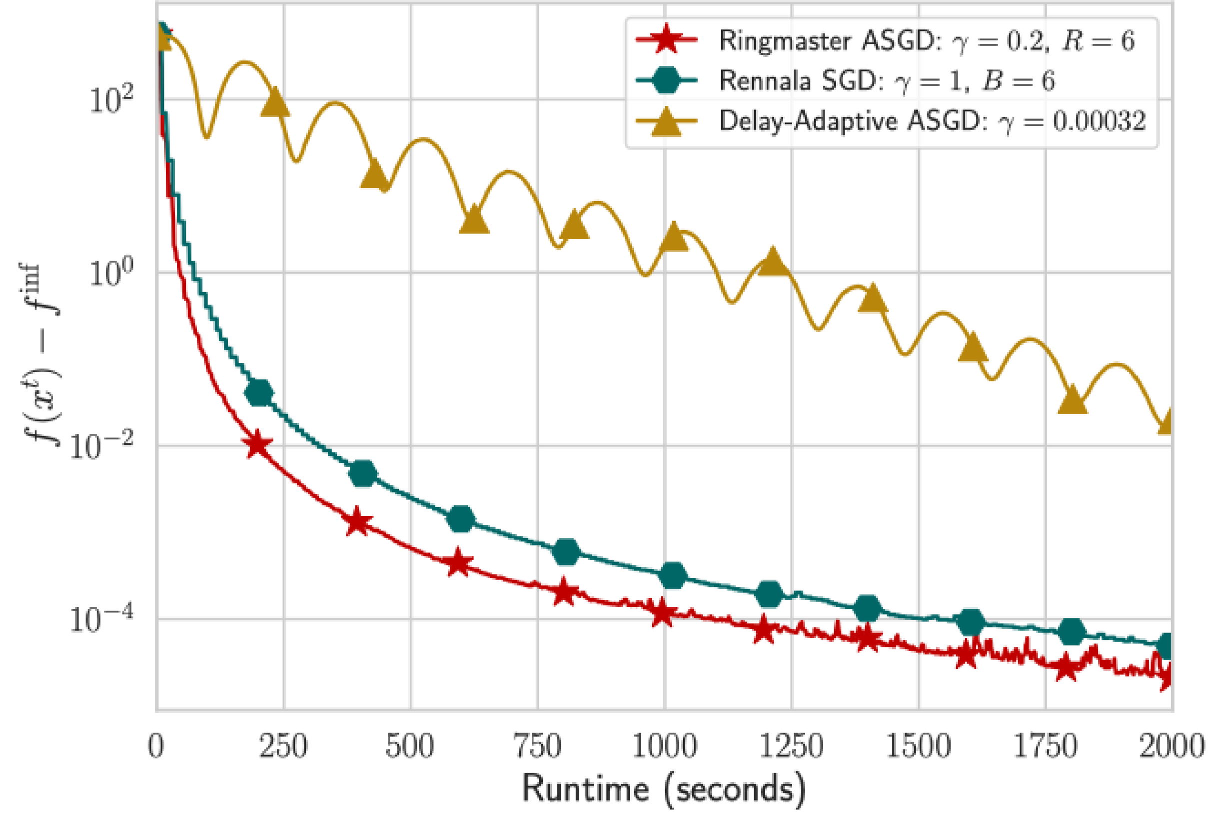
$$\tau_h^m := \left(\frac{1}{m} \sum_{i=1}^{m} \frac{1}{\tau_i} \right)^{-1}$$

Experiments

$$f(x) = \frac{1}{2} x^\top \mathbf{A} x - b^\top x \qquad \forall x \in \mathbb{R}^d$$

$$\tau_i = i + |\eta_i| \text{ for all } i \in [n], \text{ where } \eta_i \sim \mathcal{N}(0, i)$$

$$n = 6174 \qquad d = 1729$$



ICML
International Conference
On Machine Learning

King Abdullah University of
Science and Technology
جامعة الملك عبدالله
للعلوم والتقنية