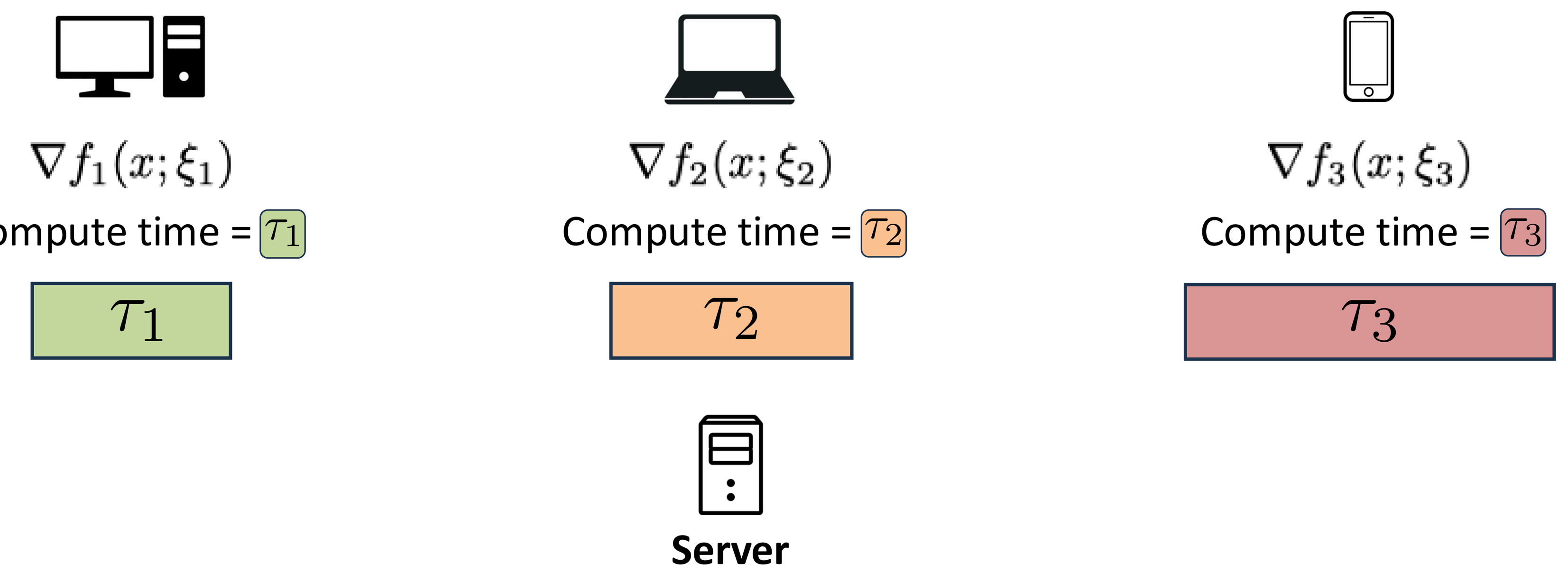
Ringleader ASGD: The First Asynchronous SGD with Optimal Time Complexity under Data Heterogeneity

Artavazd Maranjyan, Peter Richtárik

Problem setup



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

$$f_i(x) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[f_i(x;\xi_i)\right]$$

Assumptions

$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[\nabla f_i(x;\xi_i)\right] = \nabla f_i(x)$$

$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_i(x;\xi_i) - \nabla f_i(x)\|_2^2\right] \le \sigma^2$$

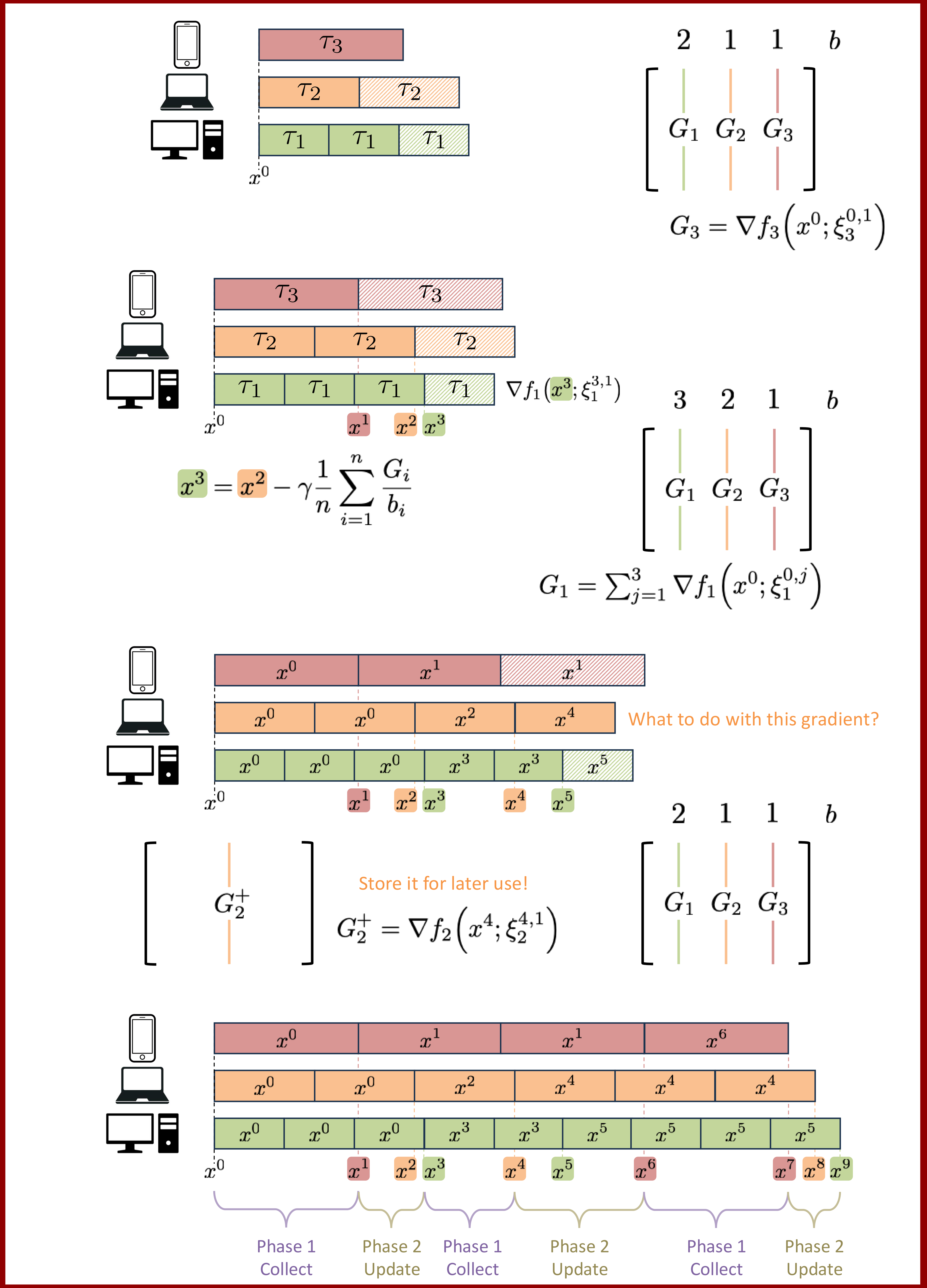
$$f(x) \ge f^*, \quad \forall x \in \mathbb{R}^d$$

$$\left\|\nabla f(x) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i)\right\|^2 \le \frac{L^2}{n}\sum_{i=1}^{n}\|x - y_i\|^2$$

Ringleader ASGD

Repeat until convergence

- Phase 1 (collect):
  Accumulate gradients until the gradient table is full
- Phase 2 (update):
  Perform a single model update per worker
  (store the intermediate gradients in a temporary table)
- Transition Step:
  Remove old gradients from the table



Comparison

Method	Time Complexity	Optimal	No sync.	No idle workers	No discarded work
Naive Minibatch SGD	$\frac{L_f\Delta}{\varepsilon}\left(\tau_n + \tau_n\frac{\sigma^2}{n\varepsilon}\right)$	✘	✘	✘	✔
IA^{2}SGD	$\frac{L_{\max}\Delta}{\varepsilon}\left(\tau_n + \tau_n\frac{\sigma^2}{n\varepsilon}\right)$	✘	✔	✔	✔
Malenia SGD	$\frac{L_f\Delta}{\varepsilon}\left(\tau_n + \tau_{\text{avg}}\frac{\sigma^2}{n\varepsilon}\right)$	✔	✘	✔	✘
Ringleader ASGD	$\frac{L\Delta}{\varepsilon}\left(\tau_n + \tau_{\text{avg}}\frac{\sigma^2}{n\varepsilon}\right)$	✔	✔	✔	✔
Lower Bound	$\frac{L_f\Delta}{\varepsilon}\left(\tau_n + \tau_{\text{avg}}\frac{\sigma^2}{n\varepsilon}\right)$	—	—	—	—

Experiments

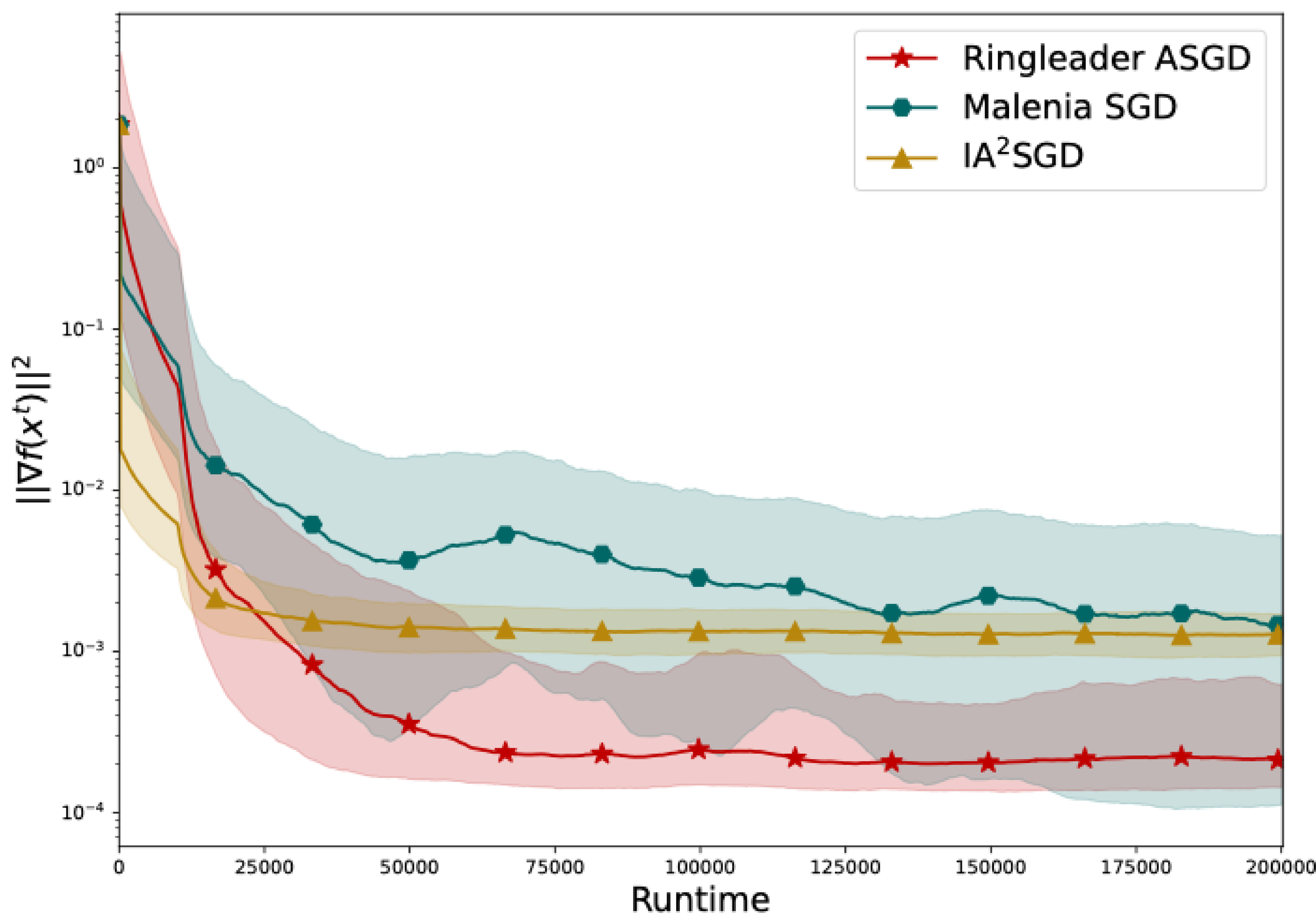
Two-layer MLP

MNIST

$n = 100$

$\tau_i = i + |\eta_i|$

$\eta_i \sim \mathcal{N}(0, i)$



ICLR

King Abdullah University of Science and Technology

جامعة الملك عبدالله للعلوم والتقنية