Trading Redundancy for Communication: Speeding up Distributed SGD for Non-convex Optimization

Farzin Haddadpour

PENNSTATE
1855

Joint work with

Mohammad Mahdi Kamani

Mehrdad Mahdavi

Viveck Cadambe

PENNSTATE

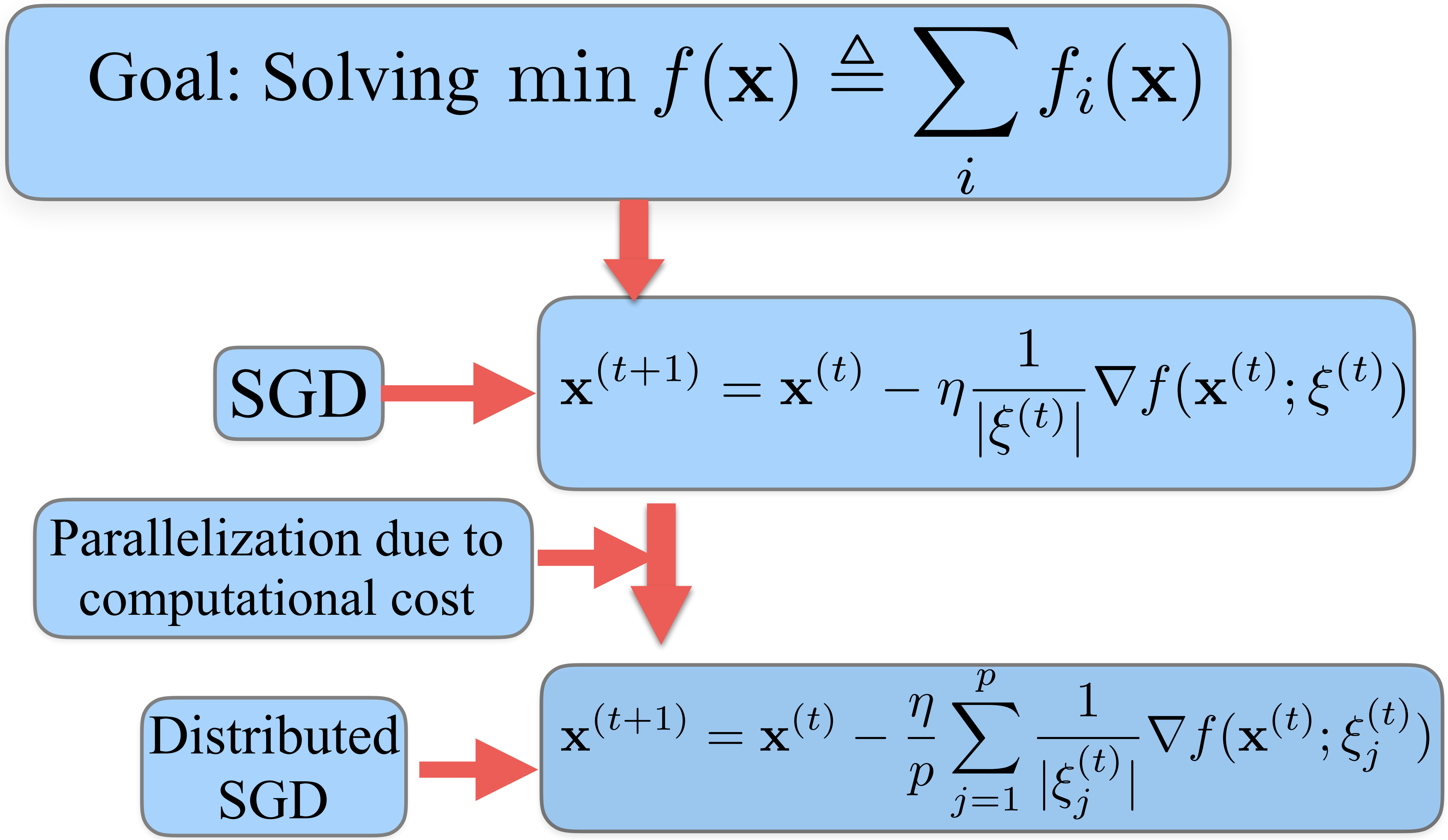
1 8 5 5

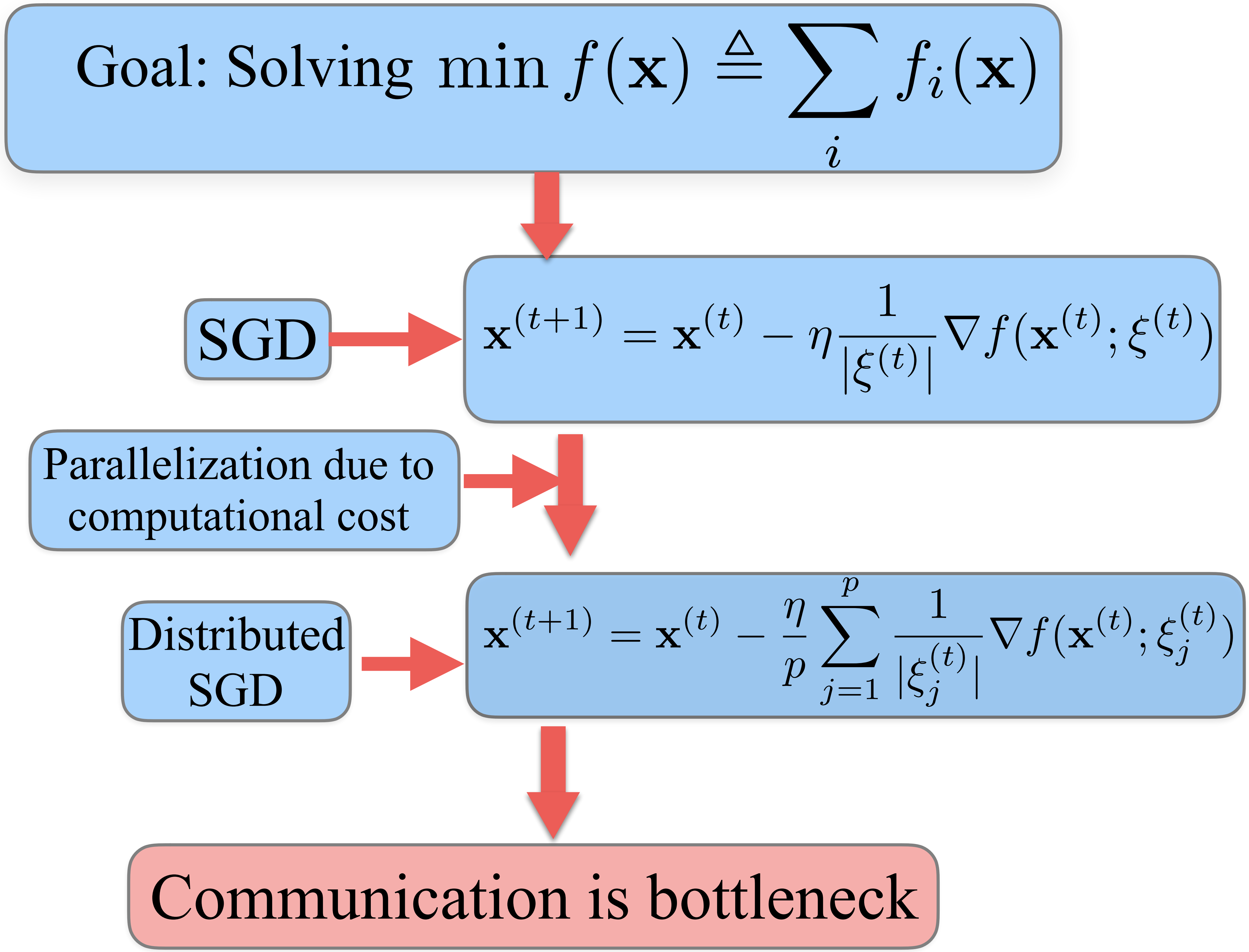
Goal: Solving $\min f(\mathbf{x}) \triangleq \sum_i f_i(\mathbf{x})$

Goal: Solving $\min f(\mathbf{x}) \triangleq \sum_i f_i(\mathbf{x})$

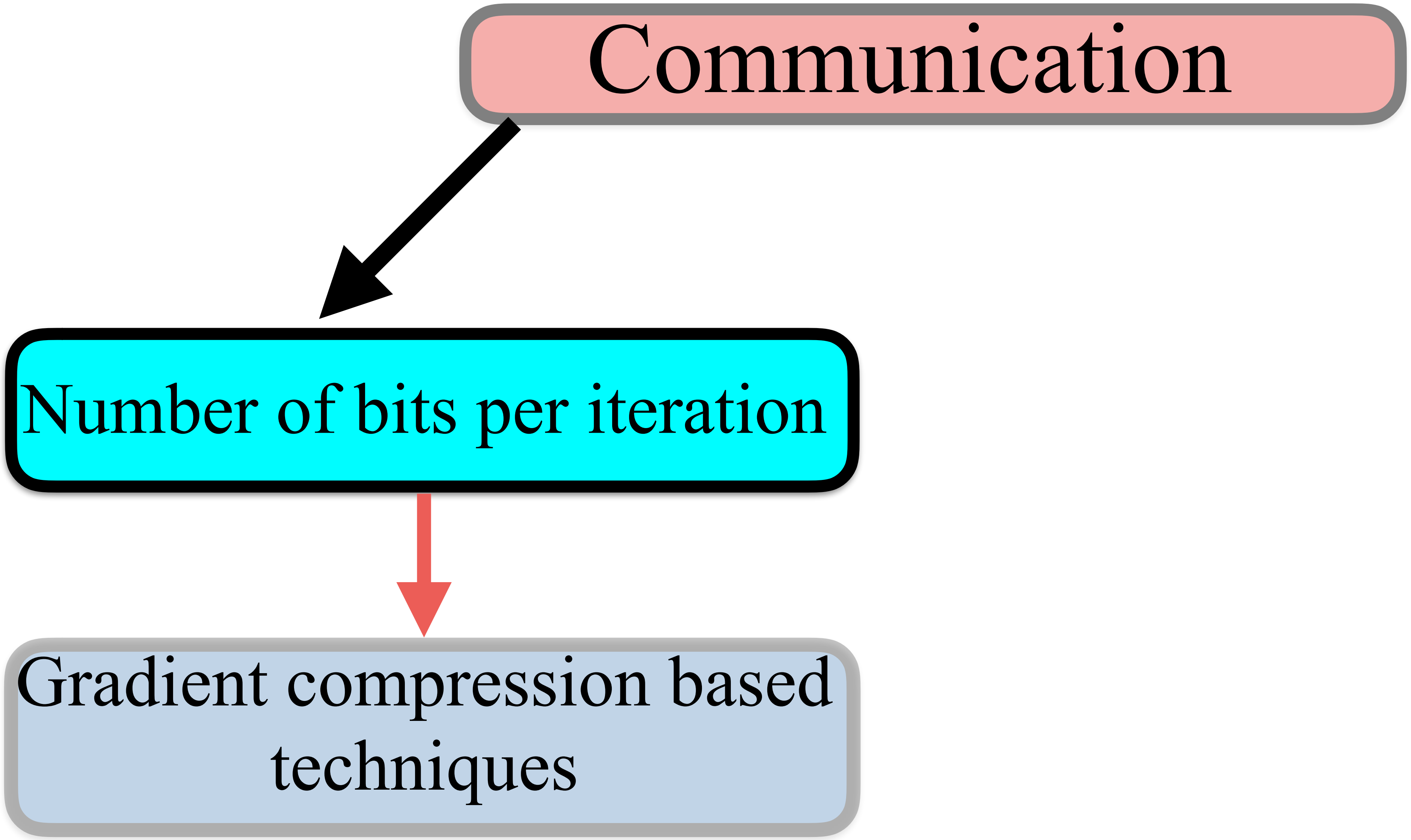
SGD

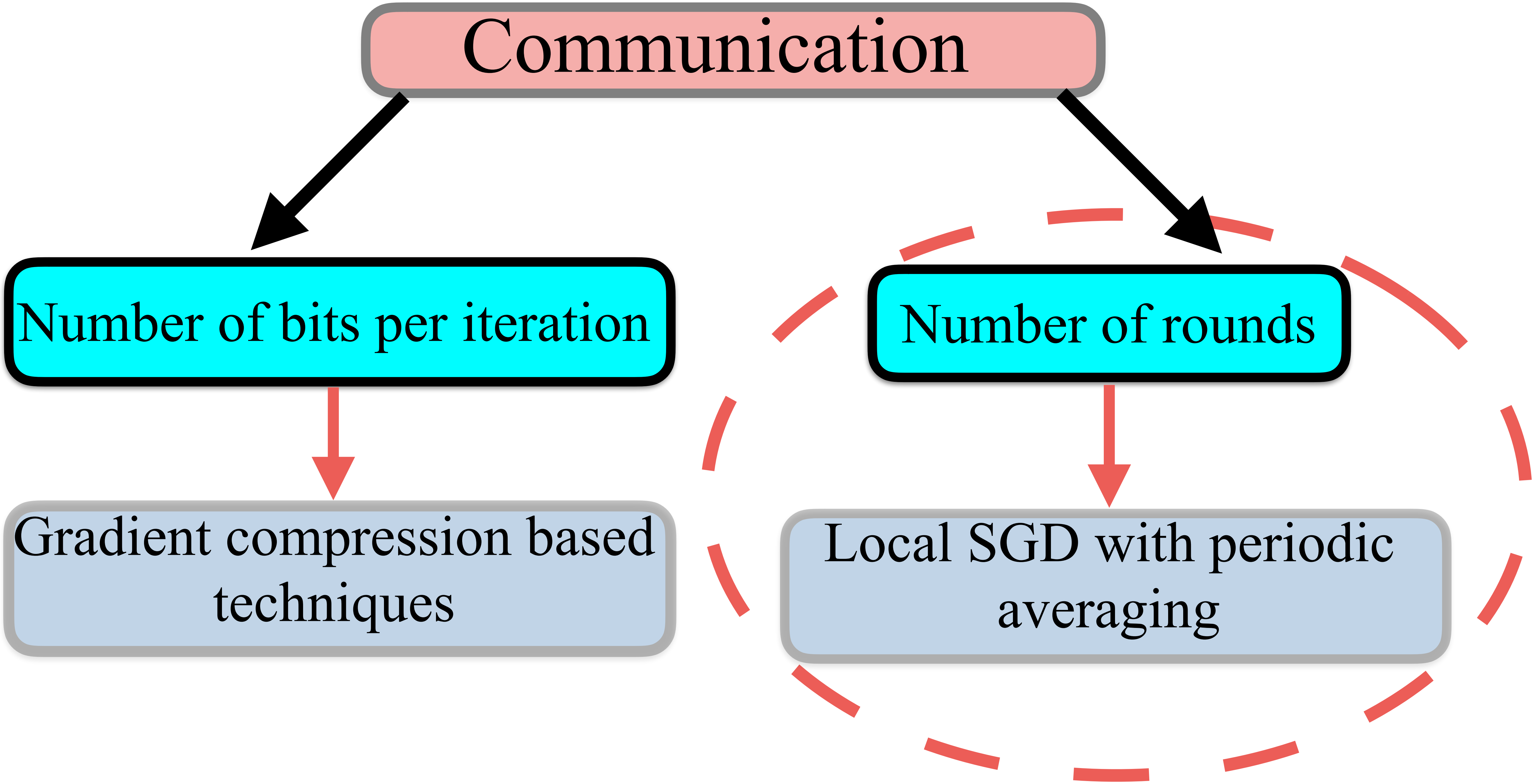
$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \frac{1}{|\xi^{(t)}|} \nabla f(\mathbf{x}^{(t)}; \xi^{(t)})$$


Goal: Solving $\min f(\mathbf{x}) \triangleq \sum_i f_i(\mathbf{x})$
SGD
$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \frac{1}{|\xi^{(t)}|} \nabla f(\mathbf{x}^{(t)}; \xi^{(t)})$
Parallelization due to computational cost
Distributed SGD
$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\eta}{p} \sum_{j=1}^{p} \frac{1}{|\xi_j^{(t)}|} \nabla f(\mathbf{x}^{(t)}; \xi_j^{(t)})$



Goal: Solving $\min f(\mathbf{x}) \triangleq \sum_i f_i(\mathbf{x})$
SGD
$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \frac{1}{|\xi^{(t)}|} \nabla f(\mathbf{x}^{(t)}; \xi^{(t)})$
Parallelization due to computational cost
Distributed SGD
$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\eta}{p} \sum_{j=1}^{p} \frac{1}{|\xi_j^{(t)}|} \nabla f(\mathbf{x}^{(t)}; \xi_j^{(t)})$
Communication is bottleneck


Communication
Number of bits per iteration
Gradient compression based techniques


Communication
Number of bits per iteration
Number of rounds
Gradient compression based techniques
Local SGD with periodic averaging

Local SGD with periodic averaging

$$\mathbf{x}_j^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left[\mathbf{x}_j^{(t)} - \eta\, \tilde{\mathbf{g}}_j^{(t)}\right] \text{ if } \tau | T$$

$$\mathbf{x}_j^{(t+1)} = \mathbf{x}_j^{(t)} - \eta\, \tilde{\mathbf{g}}_j^{(t)} \text{ otherwise,}$$

Averaging step (a)

Local update (b)

Local SGD with periodic averaging

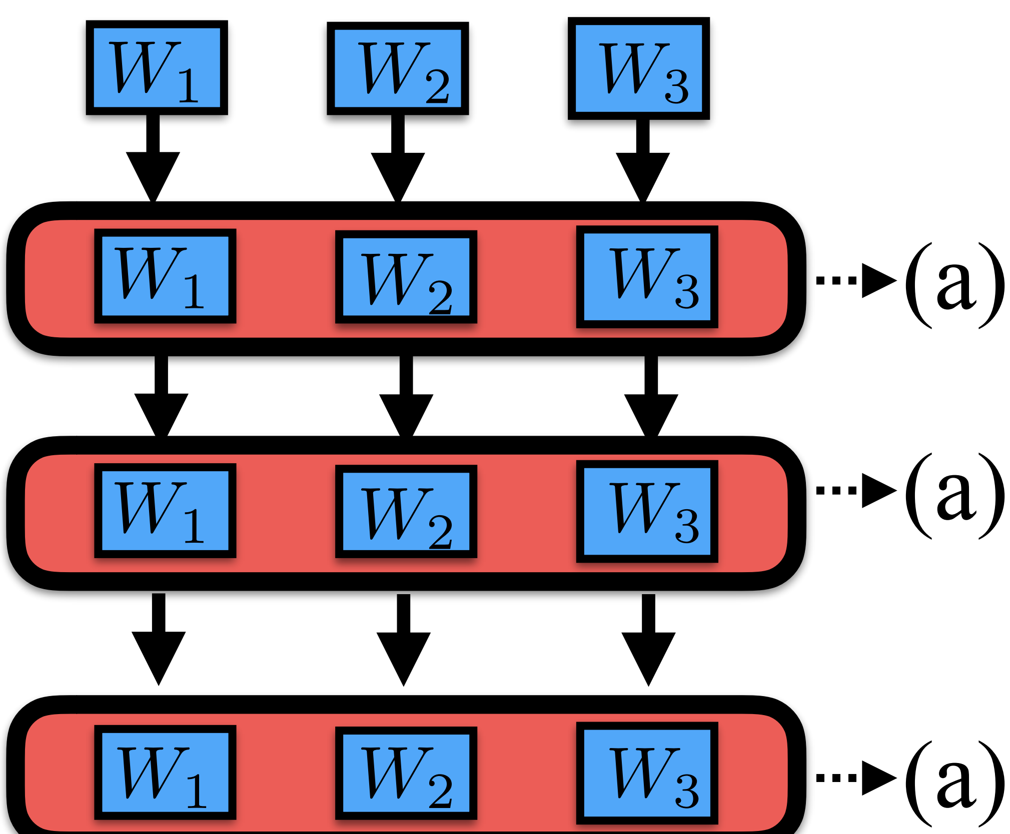
$$\mathbf{x}_j^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left[\mathbf{x}_j^{(t)} - \eta\,\tilde{\mathbf{g}}_j^{(t)}\right] \text{ if } \tau | T$$

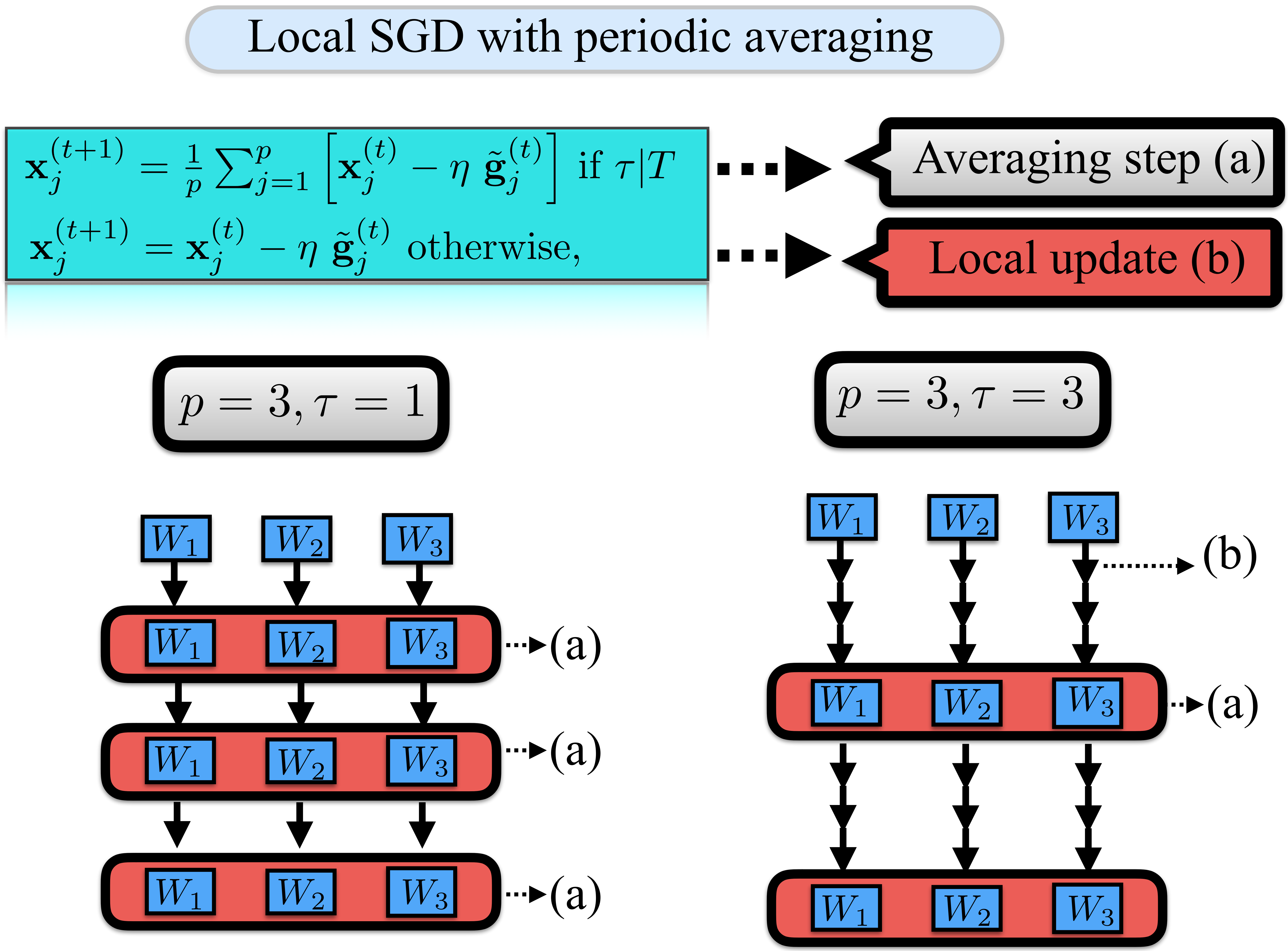
$$\mathbf{x}_j^{(t+1)} = \mathbf{x}_j^{(t)} - \eta\,\tilde{\mathbf{g}}_j^{(t)} \text{ otherwise,}$$

Averaging step (a)

Local update (b)

$p = 3, \tau = 1$




Local SGD with periodic averaging
$\mathbf{x}_j^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left[\mathbf{x}_j^{(t)} - \eta\, \tilde{\mathbf{g}}_j^{(t)}\right]$ if $\tau | T$
$\mathbf{x}_j^{(t+1)} = \mathbf{x}_j^{(t)} - \eta\, \tilde{\mathbf{g}}_j^{(t)}$ otherwise,
Averaging step (a)
Local update (b)
$p = 3, \tau = 1$
$p = 3, \tau = 3$
W_1 W_2 W_3
(a)
(b)

Convergence Analysis of Local SGD with periodic averaging

Table 1: Comparison of different SGD based algorithms.

Strategy	Convergence error	Assumptions	Com-round(T/τ)
SGD	$O(1/\sqrt{pT})$	i.i.d. & b.g	T
[Yu *et.al.*]	$O(1/\sqrt{pT})$	i.i.d. & b.g	$O(p^{\frac{3}{4}}T^{\frac{1}{4}})$
[Wang & Joshi]	$O(1/\sqrt{pT})$	i.i.d.	$O(p^{\frac{3}{2}}T^{\frac{1}{2}})$

b.g: Bounded gradient $\|\mathbf{g}_i\|_2^2 \leq G$

Unbiased gradient estimation $\mathbb{E}[\tilde{\mathbf{g}}_j] = \mathbf{g}_j$

Convergence Analysis of Local SGD with periodic averaging

Table 1: Comparison of different SGD based algorithms.

Strategy	Convergence error	Assumptions	Com-round(T/τ)
SGD	$O(1/\sqrt{pT})$	i.i.d. & b.g	T
[Yu *et.al.*]	$O(1/\sqrt{pT})$	i.i.d. & b.g	$O(p^{\frac{3}{4}}T^{\frac{1}{4}})$
[Wang & Joshi]	$O(1/\sqrt{pT})$	i.i.d.	$O(p^{\frac{3}{2}}T^{\frac{1}{2}})$

b.g: Bounded gradient $\|\mathbf{g}_i\|_2^2 \leq G$

Unbiased gradient estimation $\mathbb{E}[\tilde{\mathbf{g}}_j] = \mathbf{g}_j$

A. Residual error is observe in practice but theoretical understanding is missing?
B. How we can capture this in convergence analysis?
C. Any solution to improve it?

Insufficiency of convergence analysis

A. Residual error is observe in practice but theoretical understanding is missing?

Unbiased gradient estimation does not hold

Insufficiency of convergence analysis

A. Residual error is observe in practice but theoretical understanding is missing?

Unbiased gradient estimation does not hold

B. How to capture this in convergence analysis?

Our work

Analysis based on biased gradients

Insufficiency of convergence analysis

A. Residual error is observe in practice but theoretical understanding is missing?

Unbiased gradient estimation does not hold

B. How to capture this in convergence analysis?

Our work — Analysis based on biased gradients

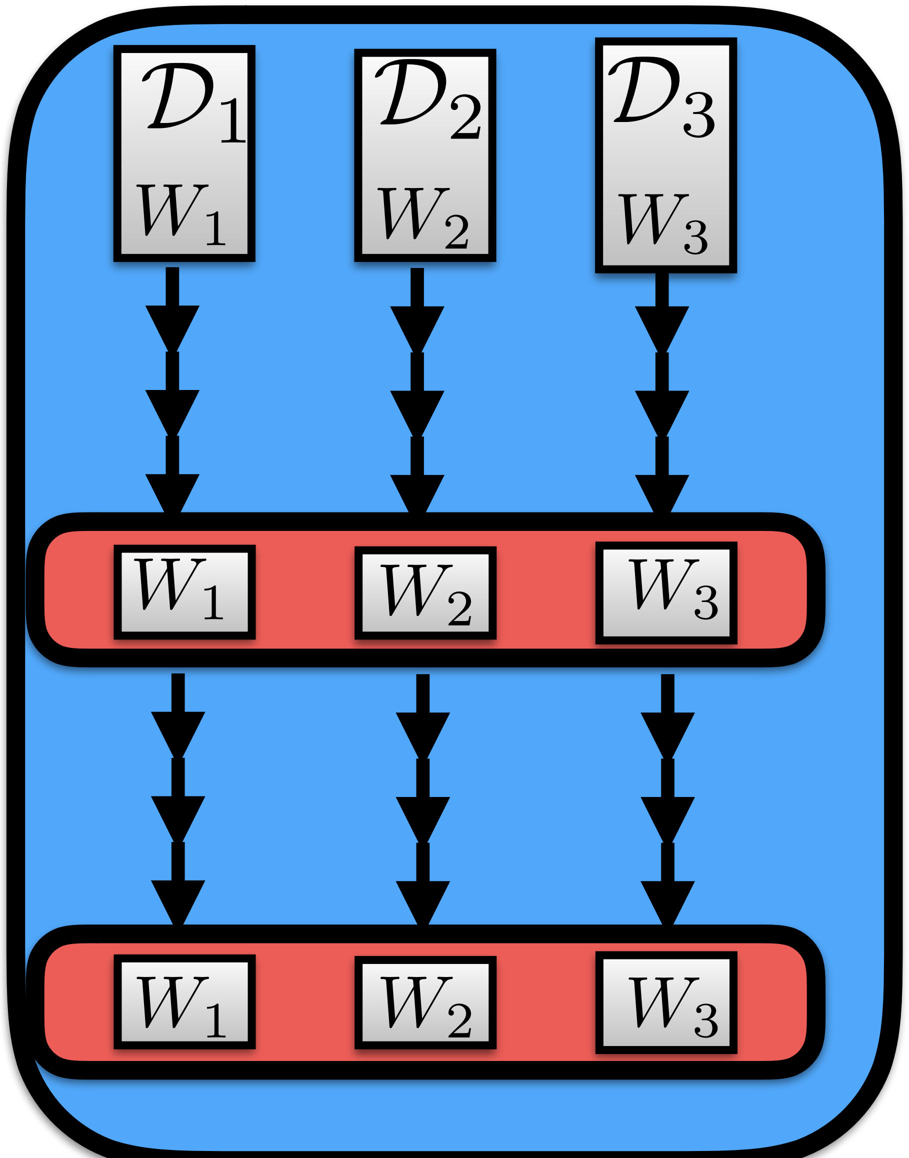
C. Any solution to improve it?

Our work — Redundancy

Redundancy infused local SGD (RI-SGD)

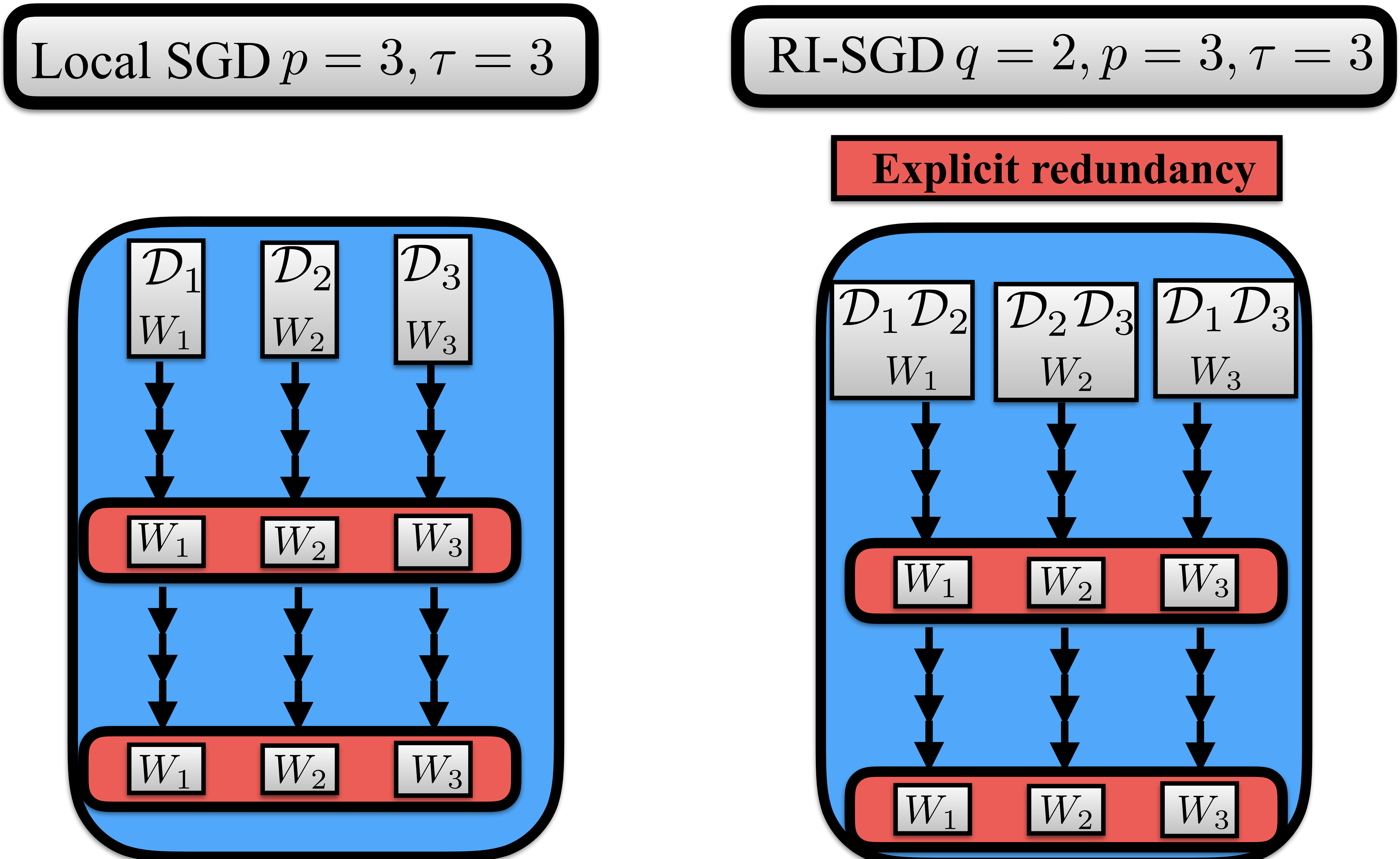
$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$$

Local SGD $p = 3, \tau = 3$



Redundancy infused local SGD (RI-SGD)

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$$



Comparing RI-SGD with other schemes

Assumption → b.d: Bounded inner product of gradients $\langle \mathbf{g}_i, \mathbf{g}_j \rangle \leq \beta$

Biased gradients

Redundancy → q: Number of data chunks at each worker node

Comparing RI-SGD with other schemes

Assumption → b.d: Bounded inner product of gradients $\langle \mathbf{g}_i, \mathbf{g}_j \rangle \leq \beta$

Biased gradients

Redundancy → q: Number of data chunks at each worker node

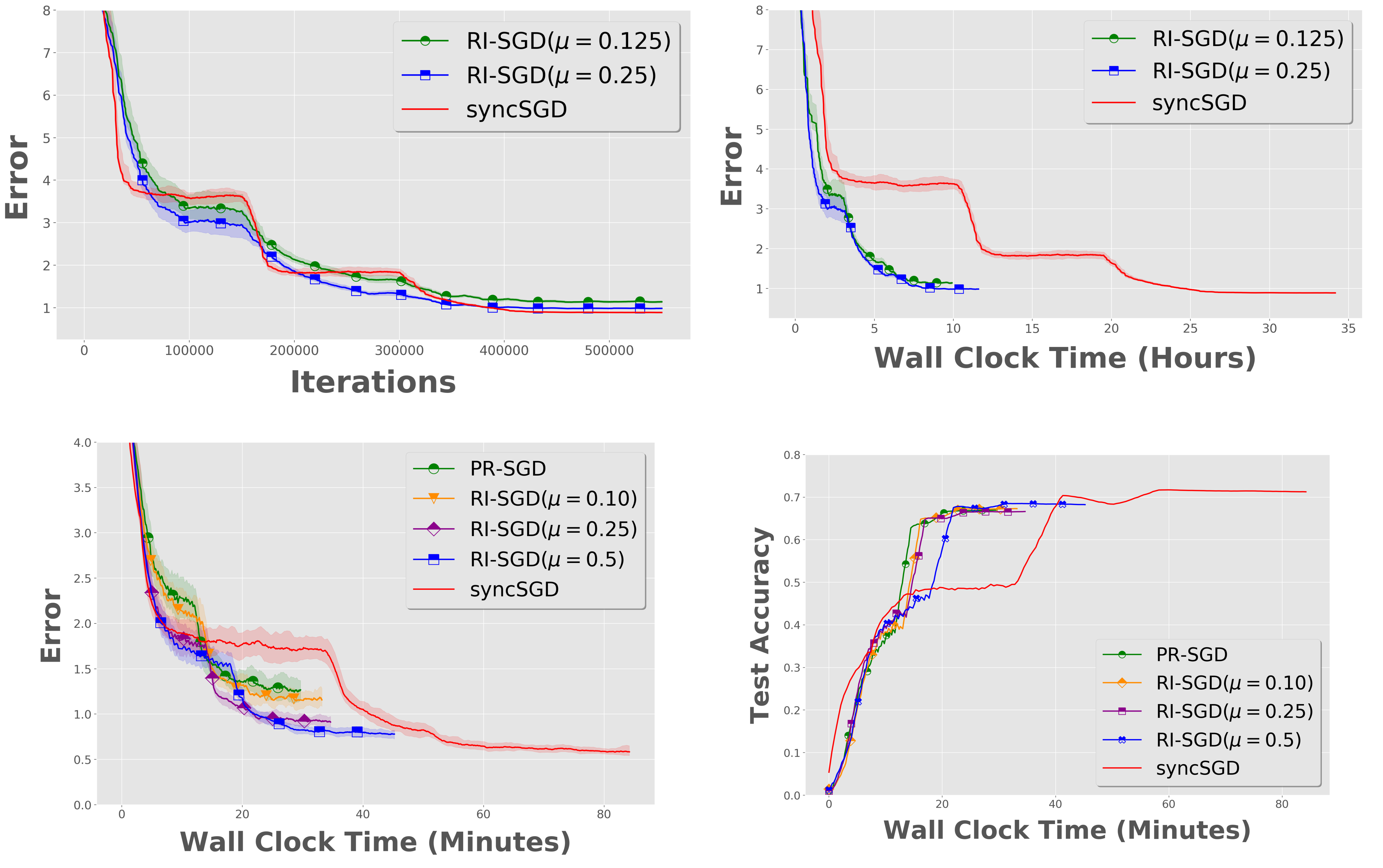
Table 1: Comparison of different SGD based algorithms.

Strategy	Convergence error	Assumptions	Com-round(T/τ)
SGD	$O(1/\sqrt{pT})$	i.i.d. & b.g	T
[Yu *et.al.*]	$O(1/\sqrt{pT})$	i.i.d. & b.g	$O(p^{\frac{3}{4}}T^{\frac{1}{4}})$
[Wang & Joshi]	$O(1/\sqrt{pT})$	i.i.d.	$O(p^{\frac{3}{2}}T^{\frac{1}{2}})$
RI-SGD (τ, q)	$O(1/\sqrt{pT}) + O((1-q/p)\beta)$	non-i.i.d. & b.d.	$O(p^{\frac{3}{2}}T^{\frac{1}{2}})$

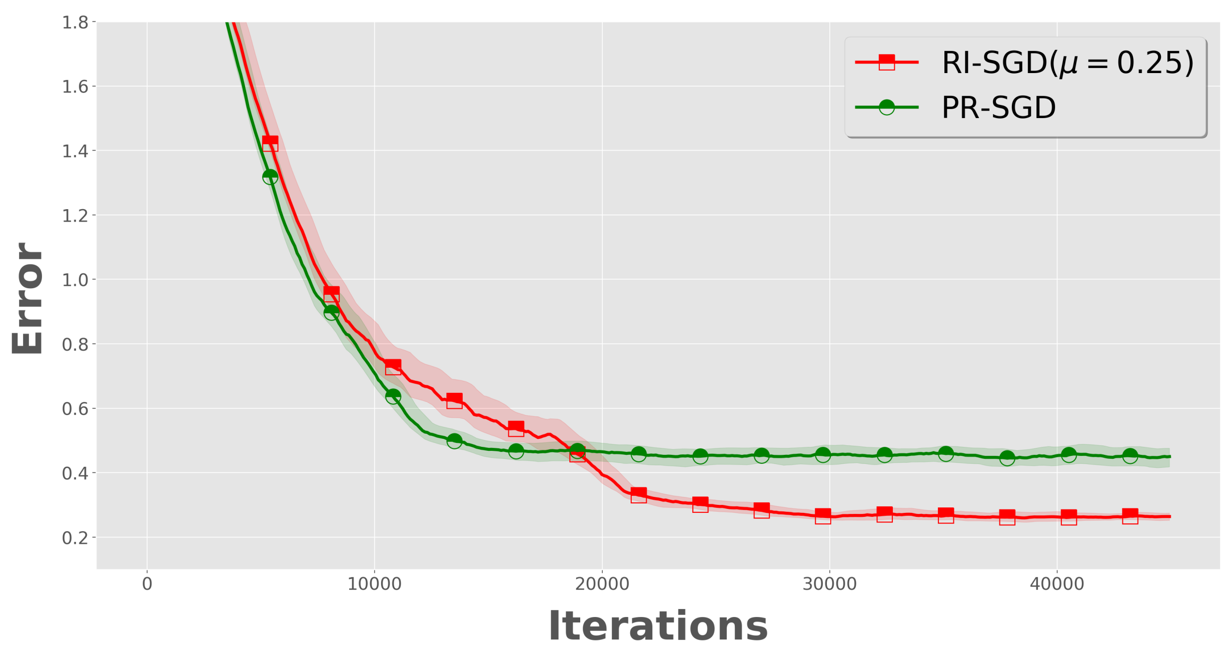
Advantages of RI-SGD:

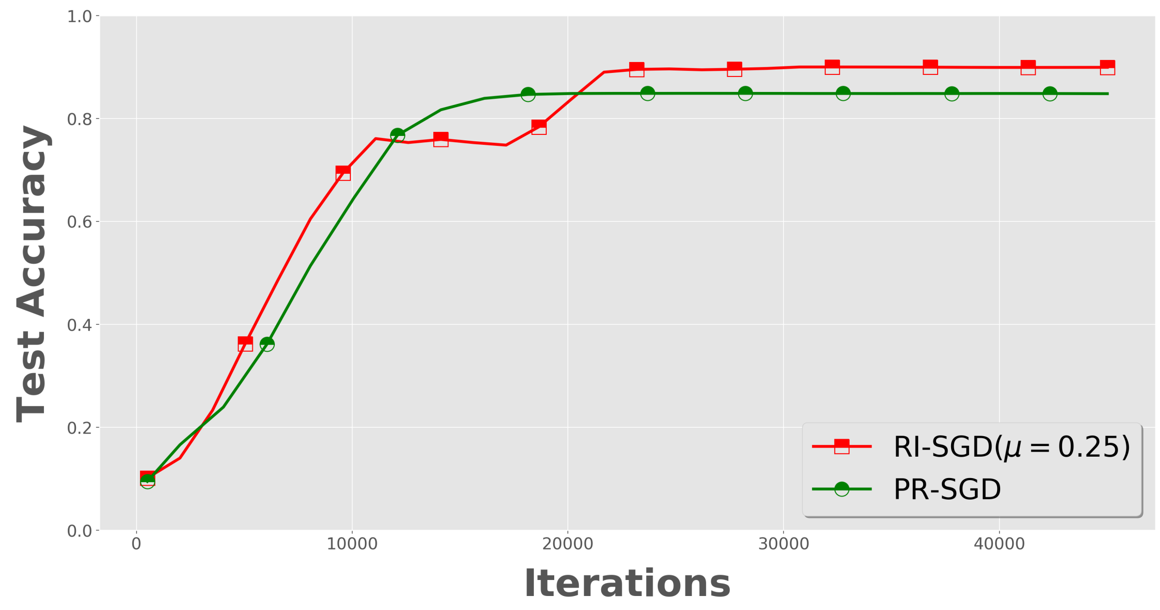
1. **Speed up not only due to larger effective mini-batch size, but also due to increasing intra-gradient diversity.**
2. **Fault-tolerance.**
3. **Extension to heterogeneous mini-batch size and possible application to federated optimization.**

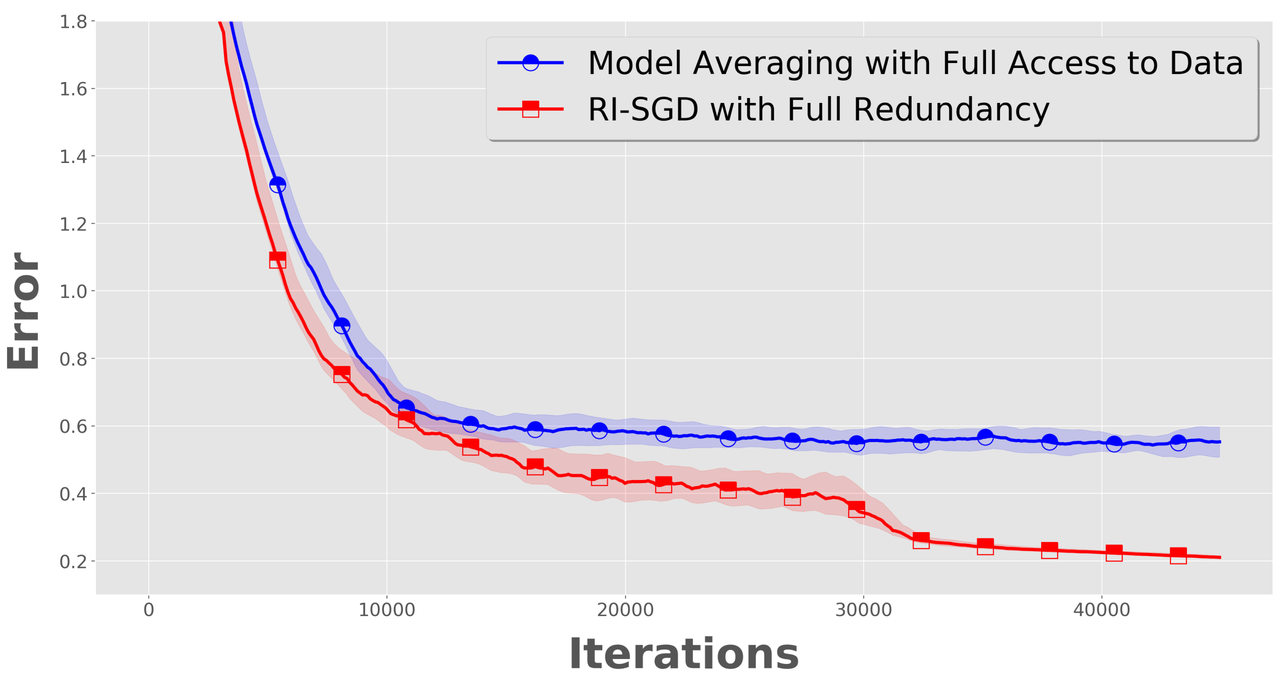
Faster convergence: Experiments over Image-net (top figures) and Cifar-100 (bottom figures)

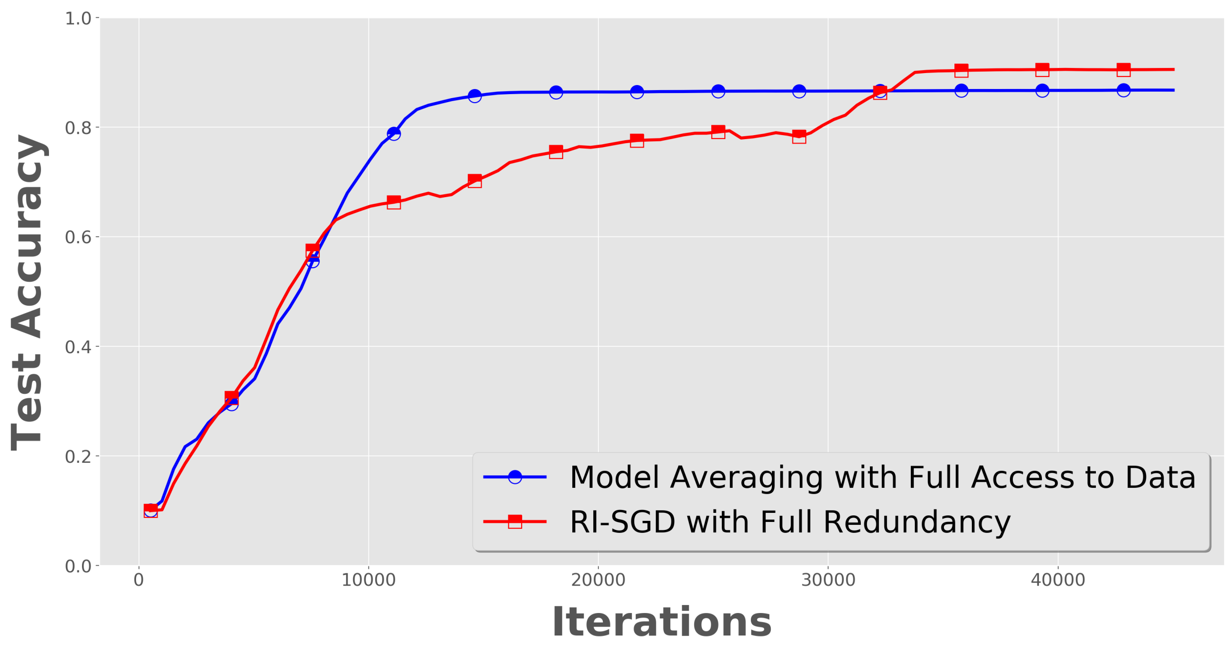


Increasing intra-gradient diversity: Experiments over Cifar-10

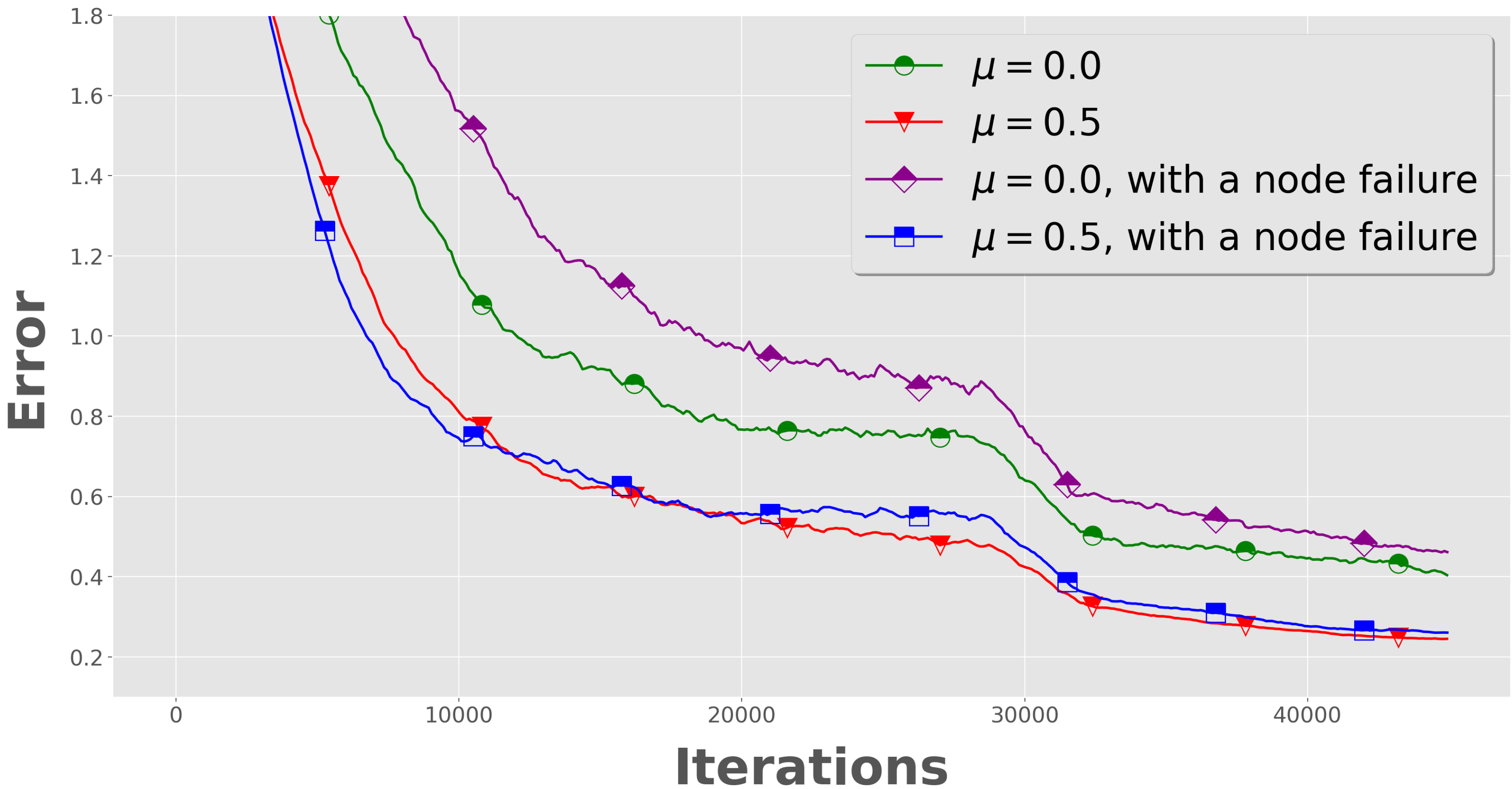


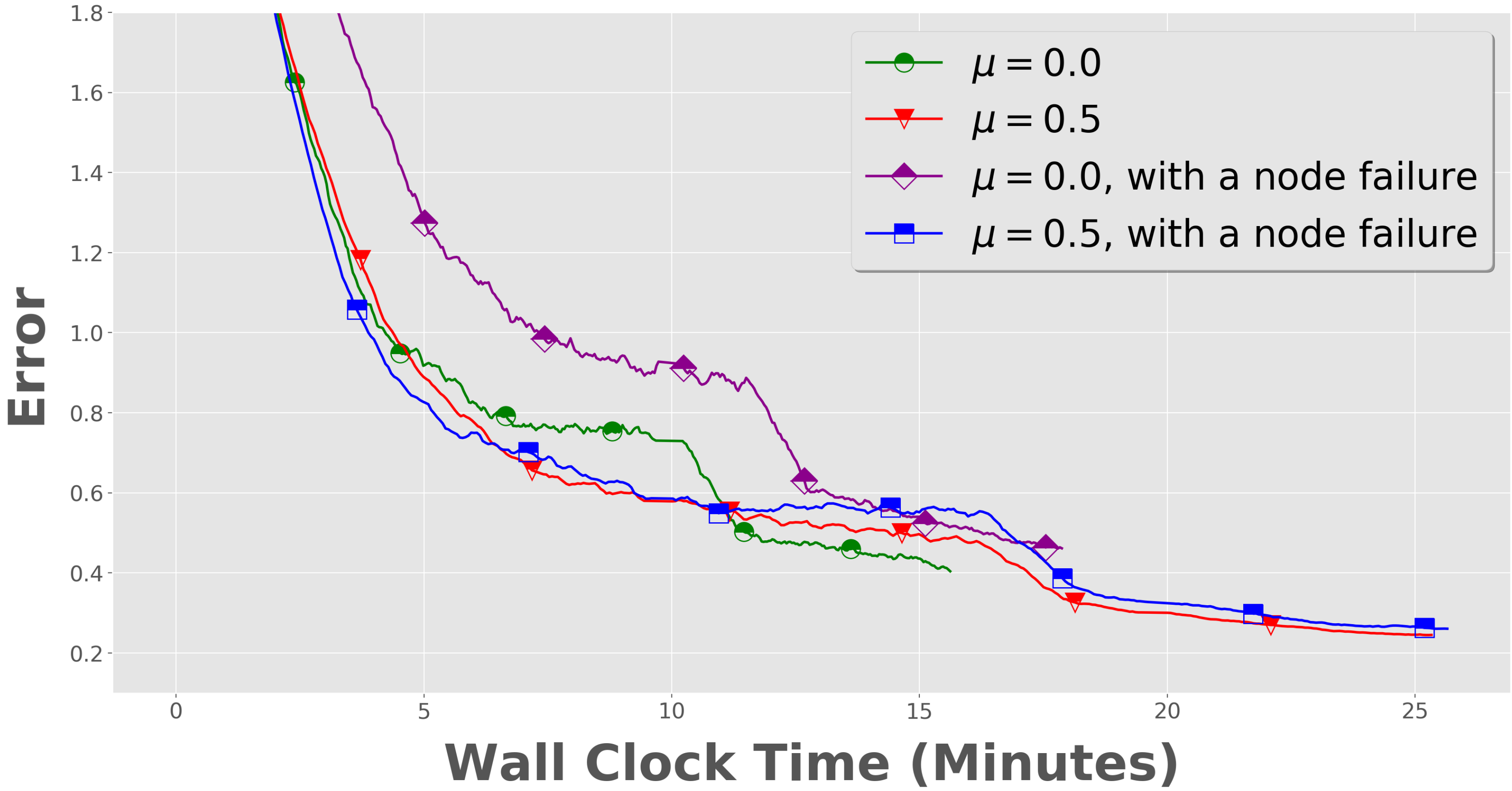






Fault-Tolerance: Experiments over Cifar-10





For more details please come to my poster session **Wed Jun 12th 06:30 -- 09:00 PM @ Pacific Ballroom #185**

Thanks for your attention!