# Private and Communication-Efficient Federated Learning via Sketches

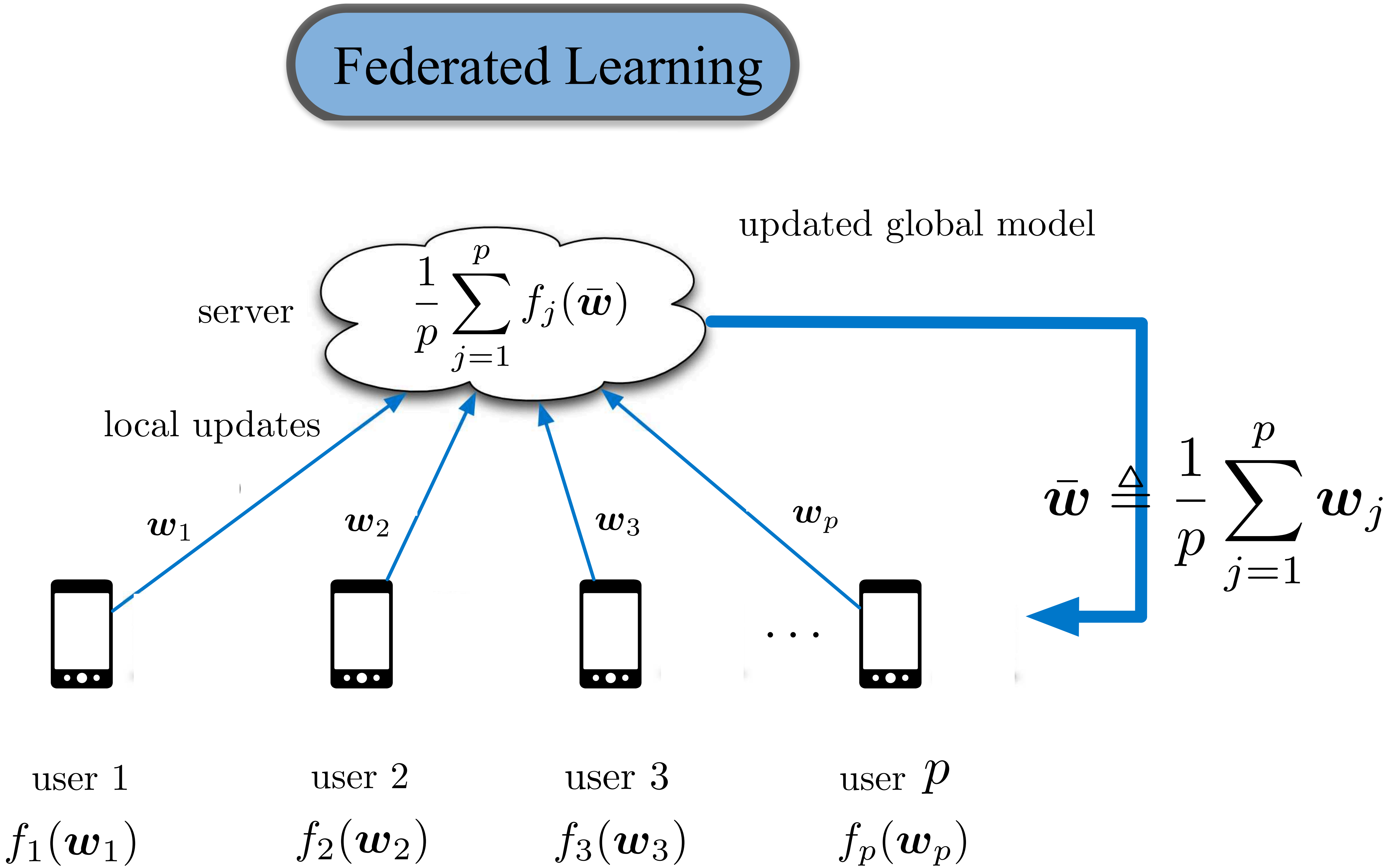
Farzin Haddadpour    Ping Li

Bai du 百度

# Outline:

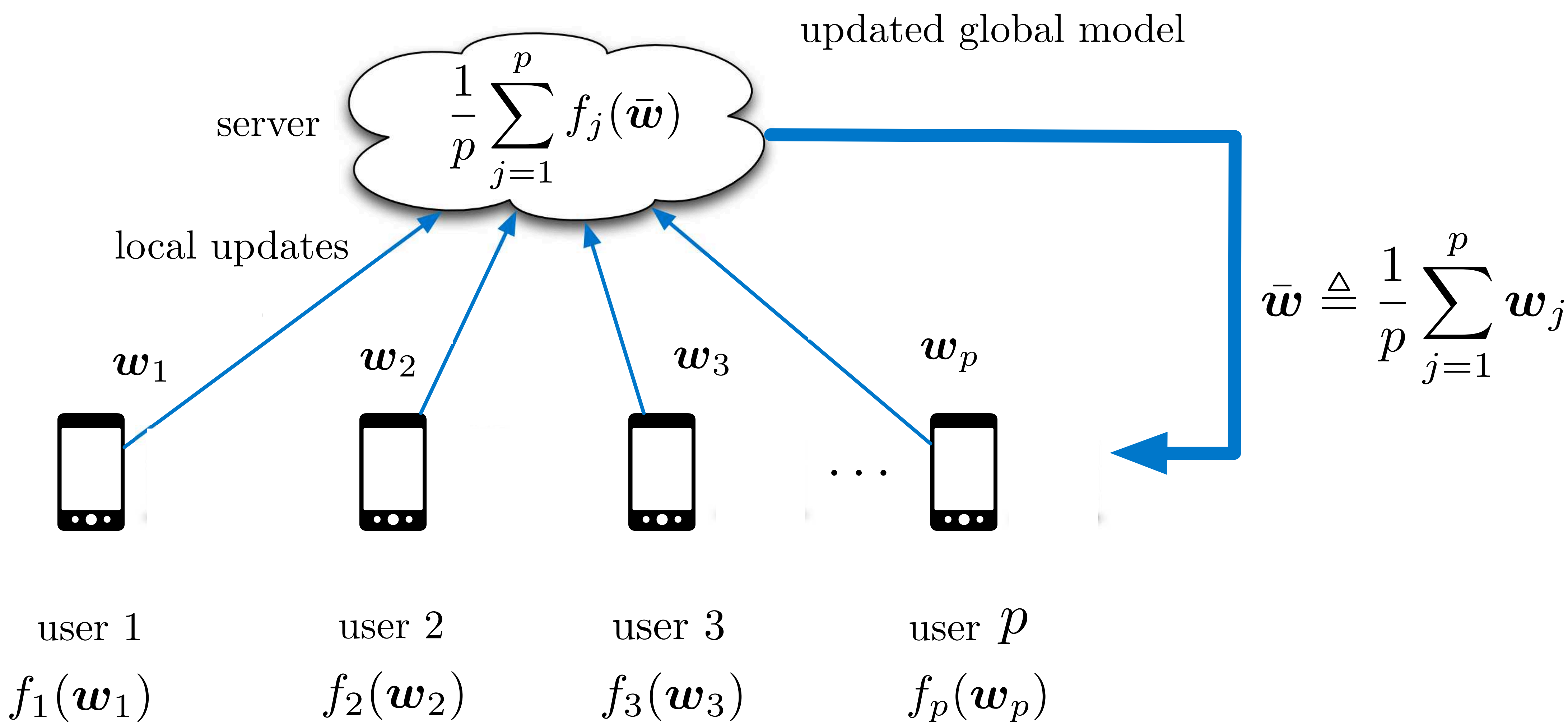
1. Federated learning review
2. Approaches to deal with communication cost
3. Sketches
4. Ongoing research

1. **Federated learning review**
2. **Approaches to deal with communication cost**
3. **Sketches**
4. **Ongoing research**



Federated Learning
updated global model
server
$\frac{1}{p}\sum_{j=1}^{p} f_j(\bar{\boldsymbol{w}})$
local updates
$\boldsymbol{w}_1$
$\boldsymbol{w}_2$
$\boldsymbol{w}_3$
$\boldsymbol{w}_p$
$\bar{\boldsymbol{w}} \triangleq \frac{1}{p}\sum_{j=1}^{p} \boldsymbol{w}_j$
user 1
$f_1(\boldsymbol{w}_1)$
user 2
$f_2(\boldsymbol{w}_2)$
user 3
$f_3(\boldsymbol{w}_3)$
user $p$
$f_p(\boldsymbol{w}_p)$

# Federated Learning




**Goal:** $$\bar{\boldsymbol{w}} = \underset{\bar{\boldsymbol{w}} \in \mathbb{R}^d}{\arg\min} \left[ \frac{1}{p}\sum_{j=1}^{p} f_j(\boldsymbol{w}) \right]$$

# Three bottlenecks for federated learning:

1. Communication cost/complexity
2. Privacy
3. Robustness against data heterogeneity

# Three bottlenecks for federated learning:

1. **Communication cost/complexity**
2. **Privacy**
3. **Robustness against data heterogeneity**

Goal: Improving all aspects

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

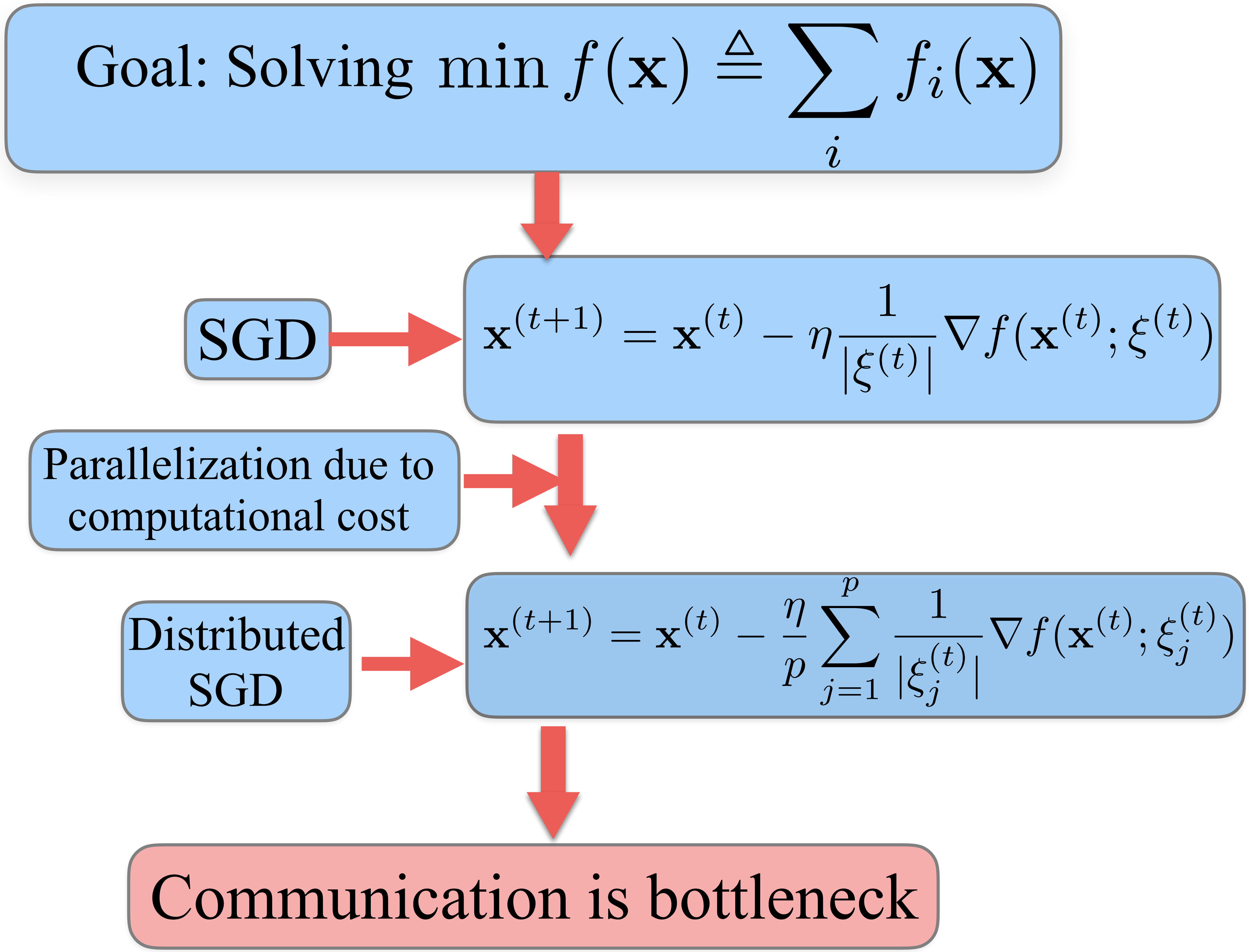
SGD

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

Parallelization due to computational cost

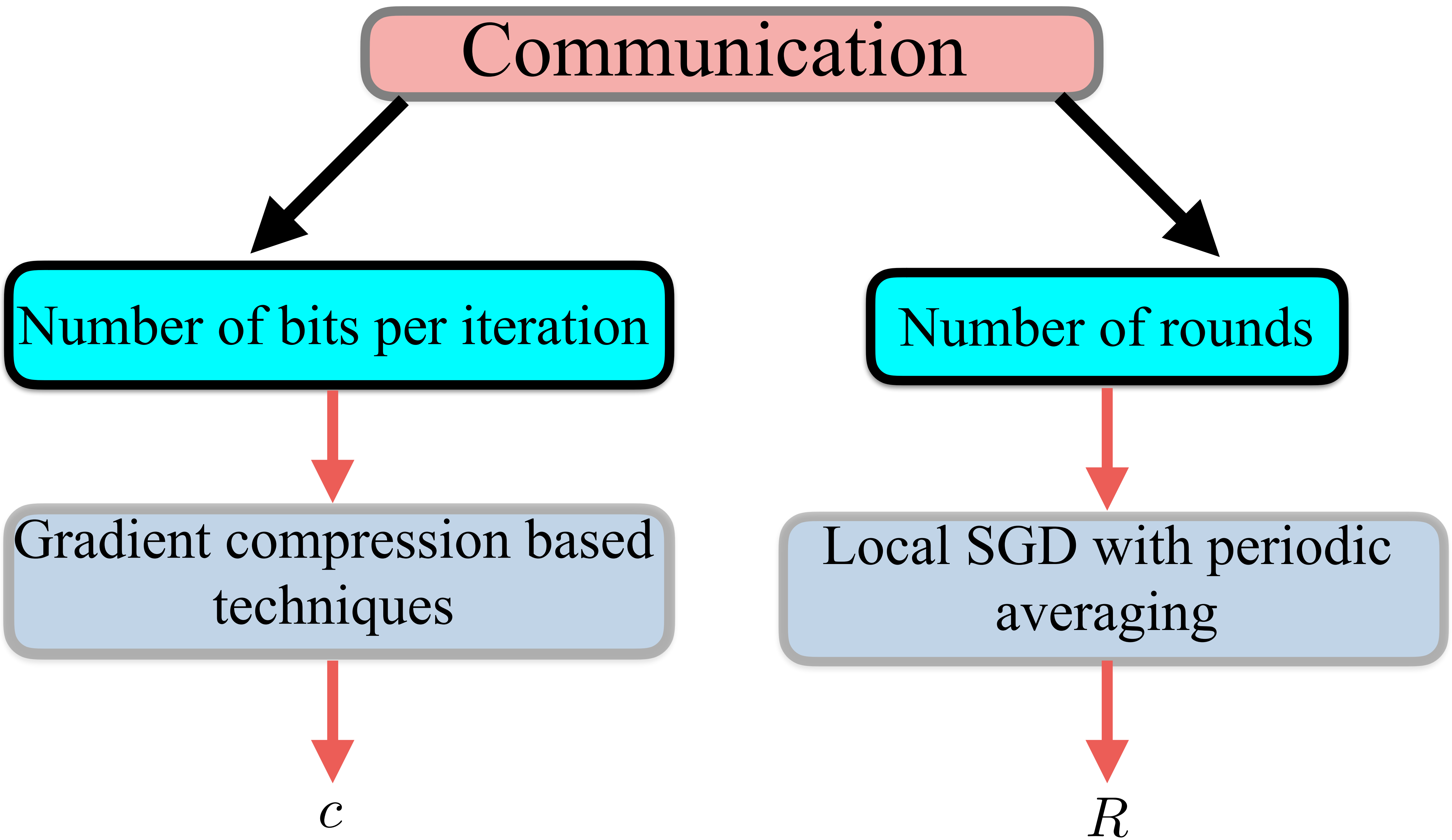
Distributed SGD

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

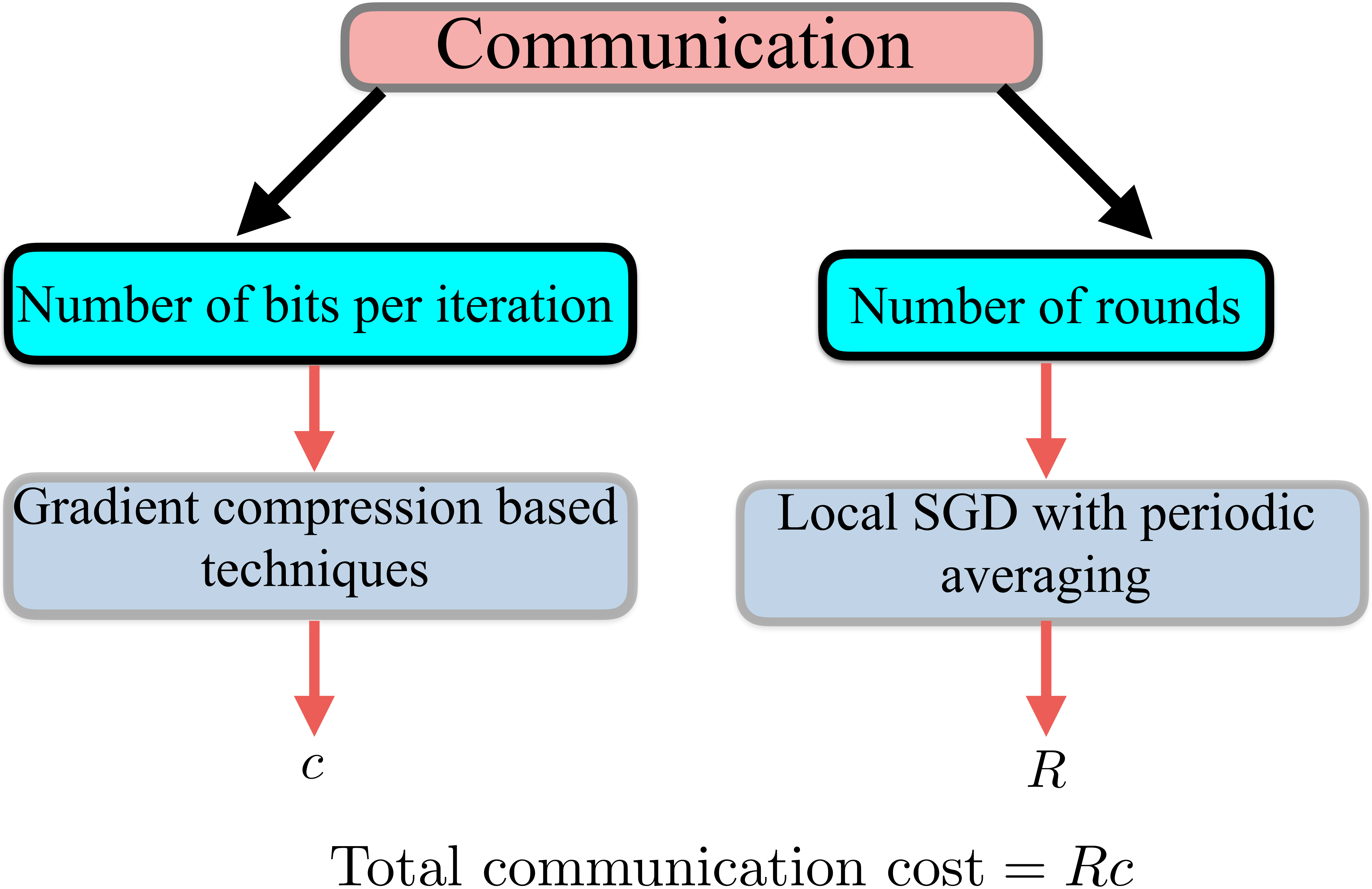

Goal: Solving $\min f(\mathbf{x}) \triangleq \sum_i f_i(\mathbf{x})$
SGD
$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \frac{1}{|\xi^{(t)}|} \nabla f(\mathbf{x}^{(t)}; \xi^{(t)})$
Parallelization due to computational cost
Distributed SGD
$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\eta}{p} \sum_{j=1}^{p}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

# Outline:

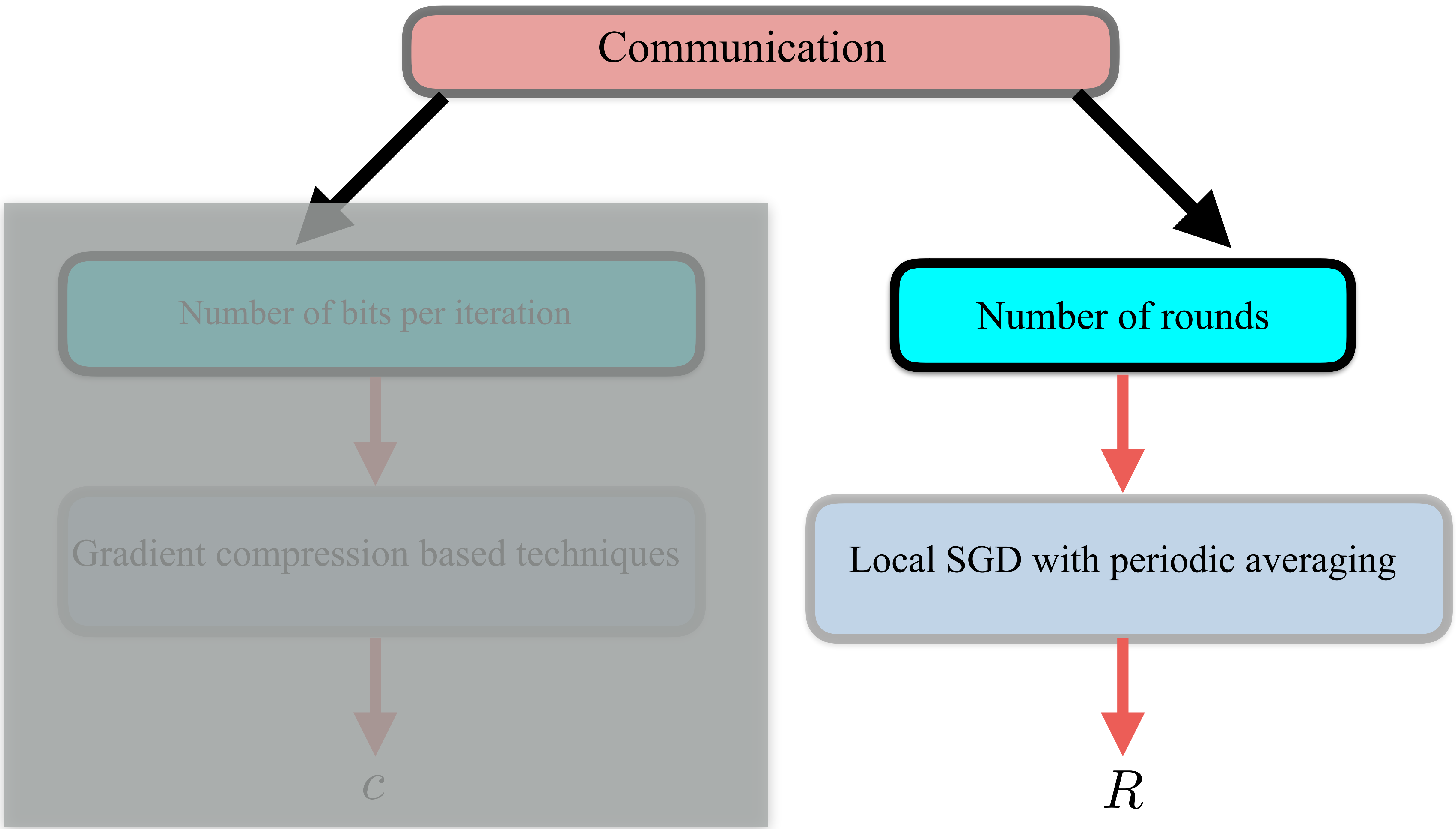
1. Federated learning review
2. **Approaches to deal with communication cost**
3. Sketches
4. Ongoing research

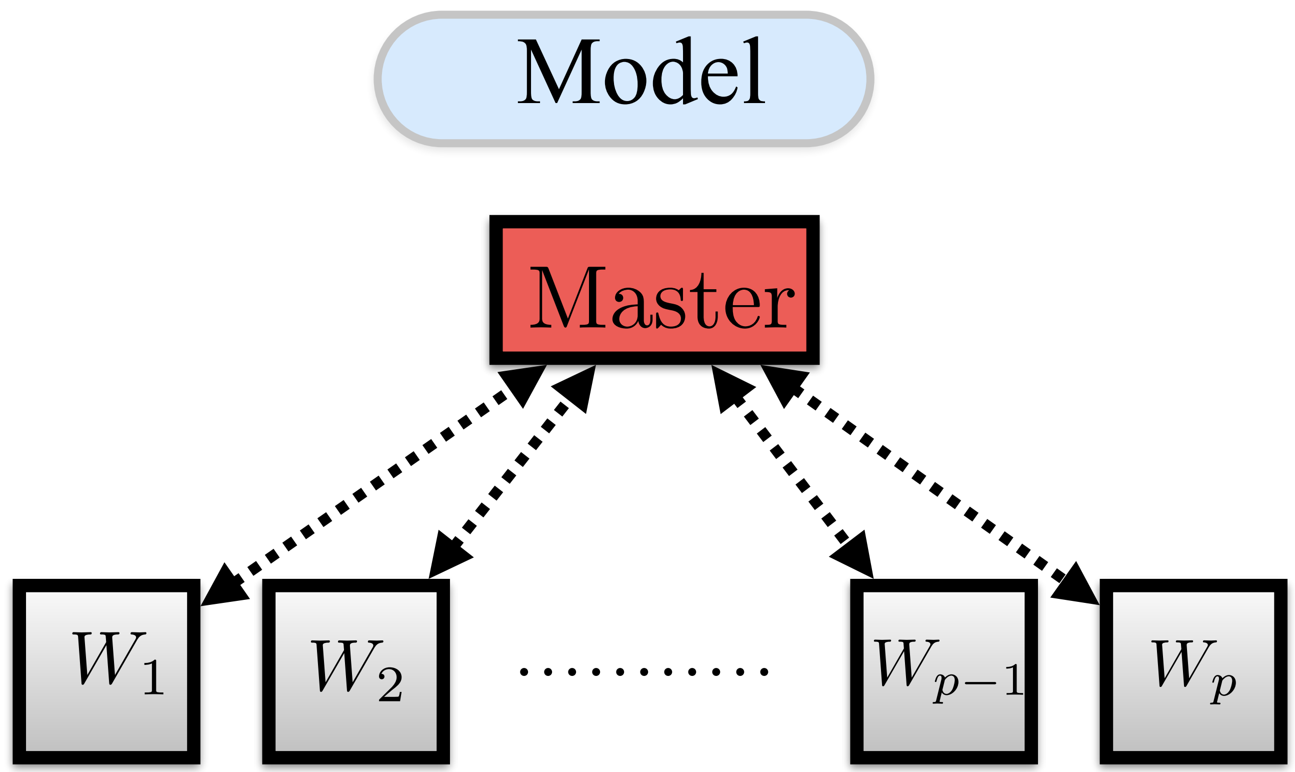


Communication
Number of bits per iteration
Number of rounds
Gradient compression based techniques
Local SGD with periodic averaging
$c$
$R$



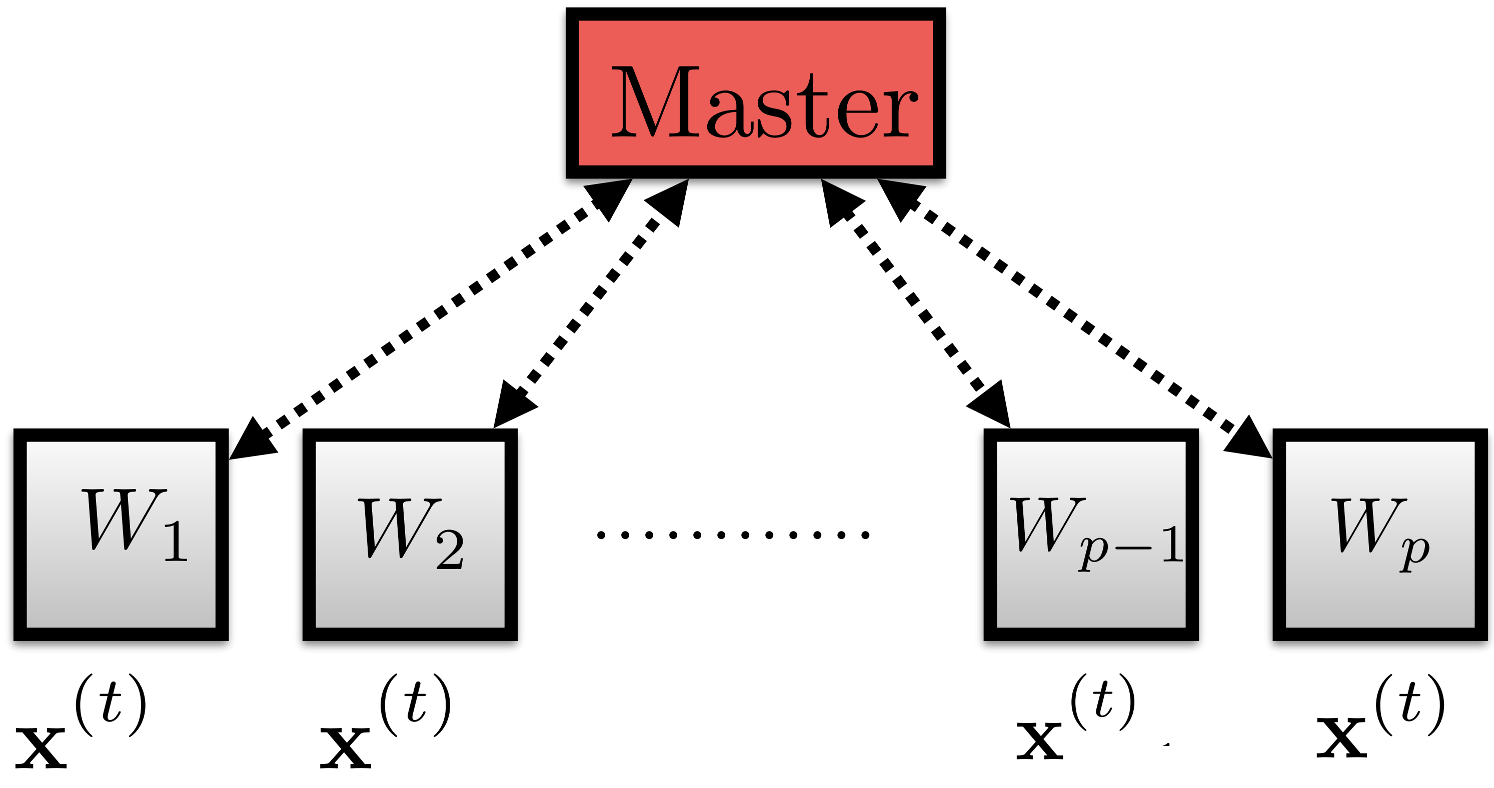
Communication
Number of bits per iteration
Number of rounds
Gradient compression based techniques
Local SGD with periodic averaging
$c$
$R$
Total communication cost $= Rc$


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

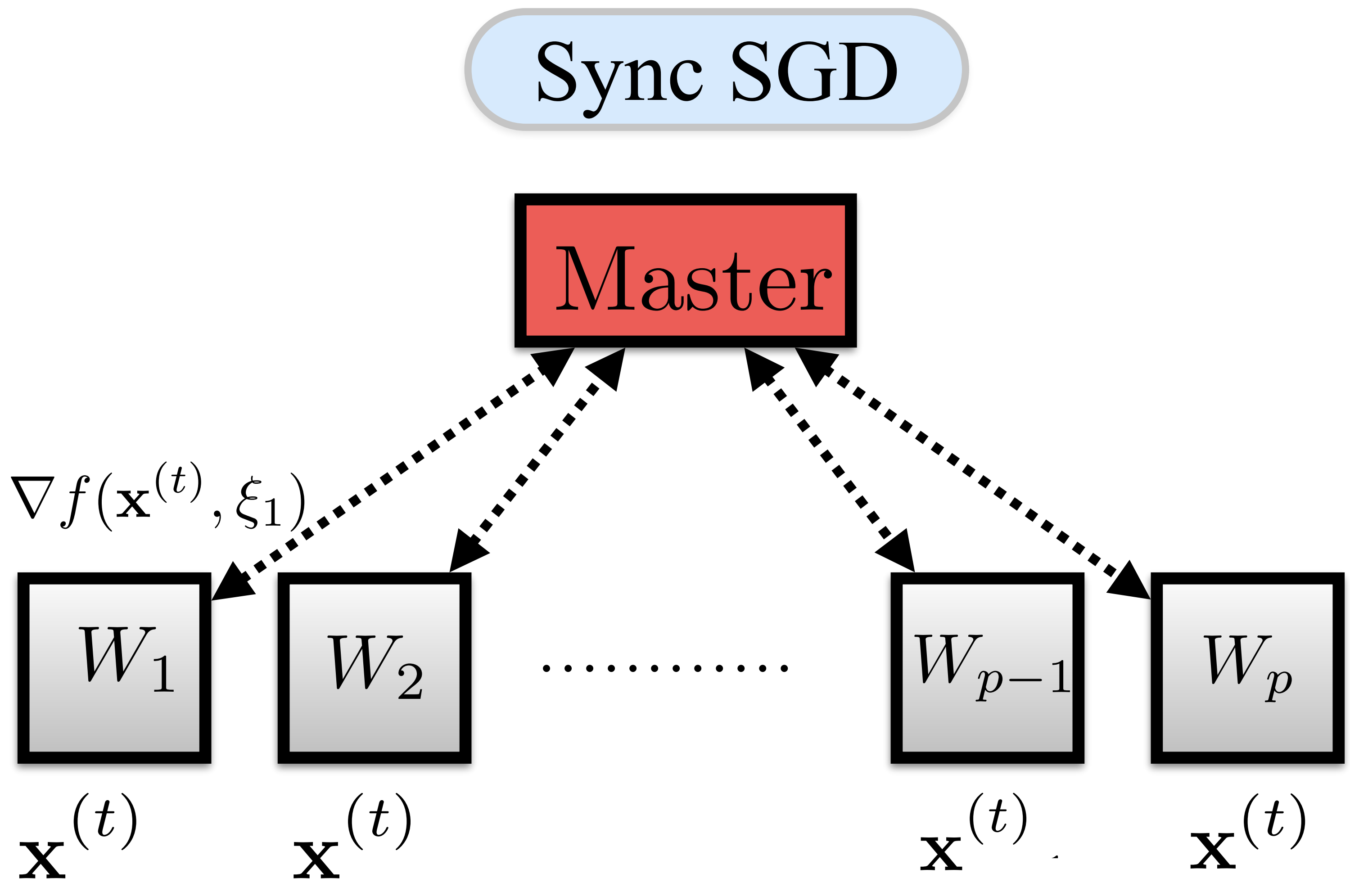


Model
Master
$W_1$
$W_2$
$W_{p-1}$
$W_p$

## Sync SGD




Device $j$ computes : $\nabla f_j(\mathbf{x}^{(t)}, \xi_j^{(t)}) \in \mathbb{R}^d$

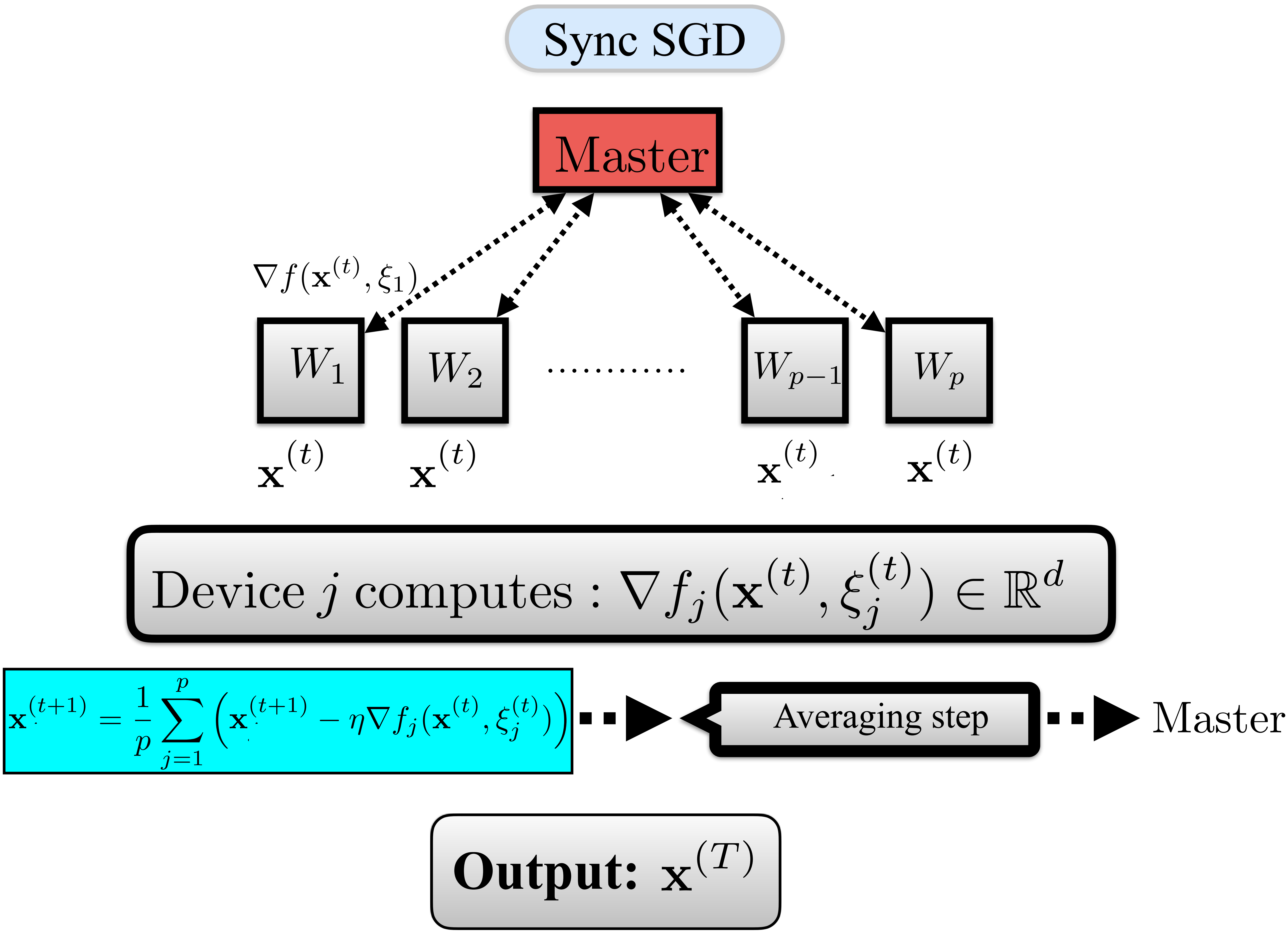



Device $j$ computes : $\nabla f_j(\mathbf{x}^{(t)}, \xi_j^{(t)}) \in \mathbb{R}^d$

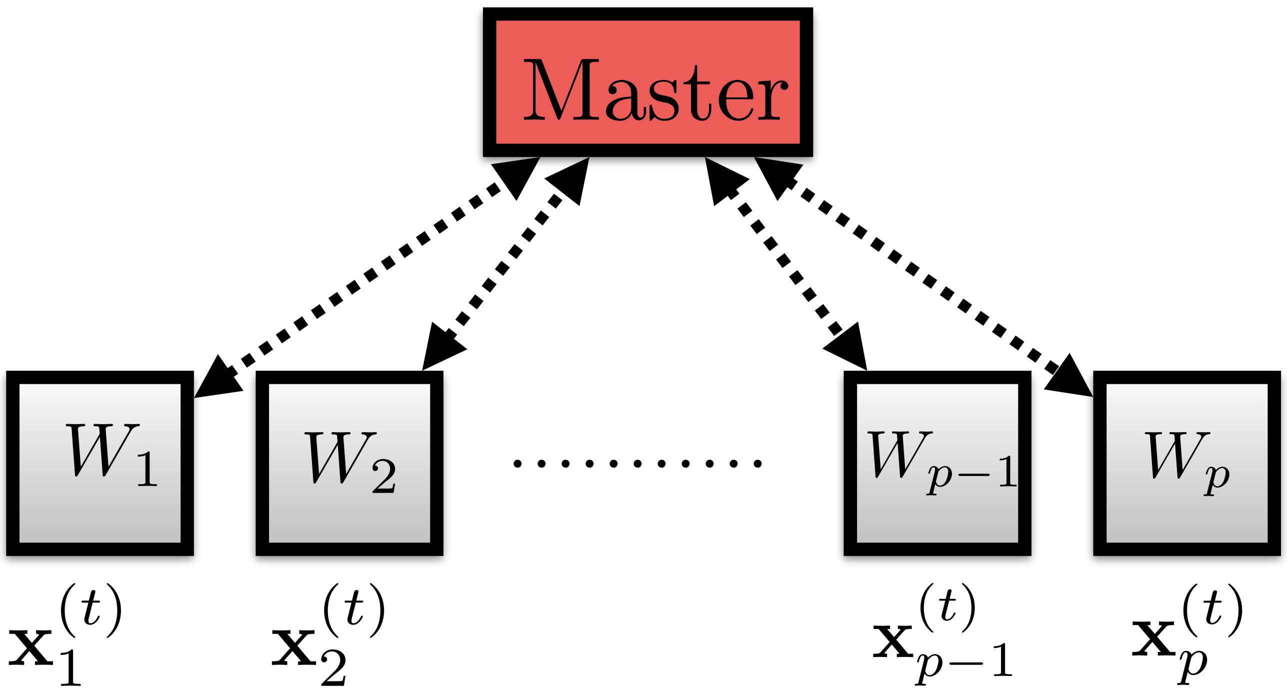
$$\boxed{\mathbf{x}^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x}^{(t+1)} - \eta \nabla f_j(\mathbf{x}^{(t)}, \xi_j^{(t)})\right)}$$

Averaging step → Master

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \frac{1}{p}\sum_{j=1}^{p} \nabla f_j(\mathbf{x}^{(t)}, \xi_j^{(t)})$$

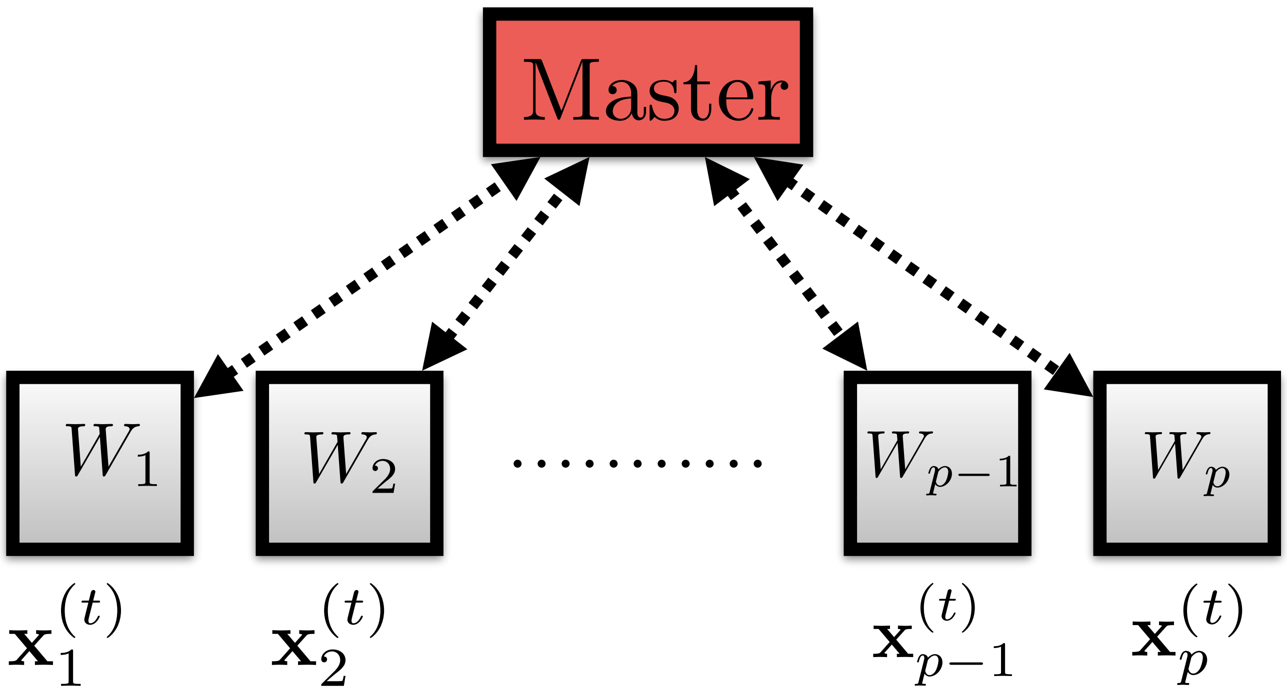

Sync SGD
Master
$\nabla f(\mathbf{x}^{(t)}, \xi_1)$
$W_1$
$W_2$
$W_{p-1}$
$W_p$
$\mathbf{x}^{(t)}$
$\mathbf{x}^{(t)}$
$\mathbf{x}^{(t)}$
$\mathbf{x}^{(t)}$
Device $j$ computes : $\nabla f_j(\mathbf{x}^{(t)}, \xi_j^{(t)}) \in \mathbb{R}^d$
$\mathbf{x}^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{x}^{(t+1)} - \eta\nabla f_j(\mathbf{x}^{(t)}, \xi_j^{(t)})\right)$
Averaging step
Master
Output: $\mathbf{x}^{(T)}$

# Local SGD with periodic averaging




$$\tilde{\mathbf{g}}_j^{(t)} = \nabla f(\mathbf{x}_j^{(t)}, \xi_j)$$

# Local SGD with periodic averaging




$$\tilde{\mathbf{g}}_j^{(t)} = \nabla f(\mathbf{x}_j^{(t)}, \xi_j)$$

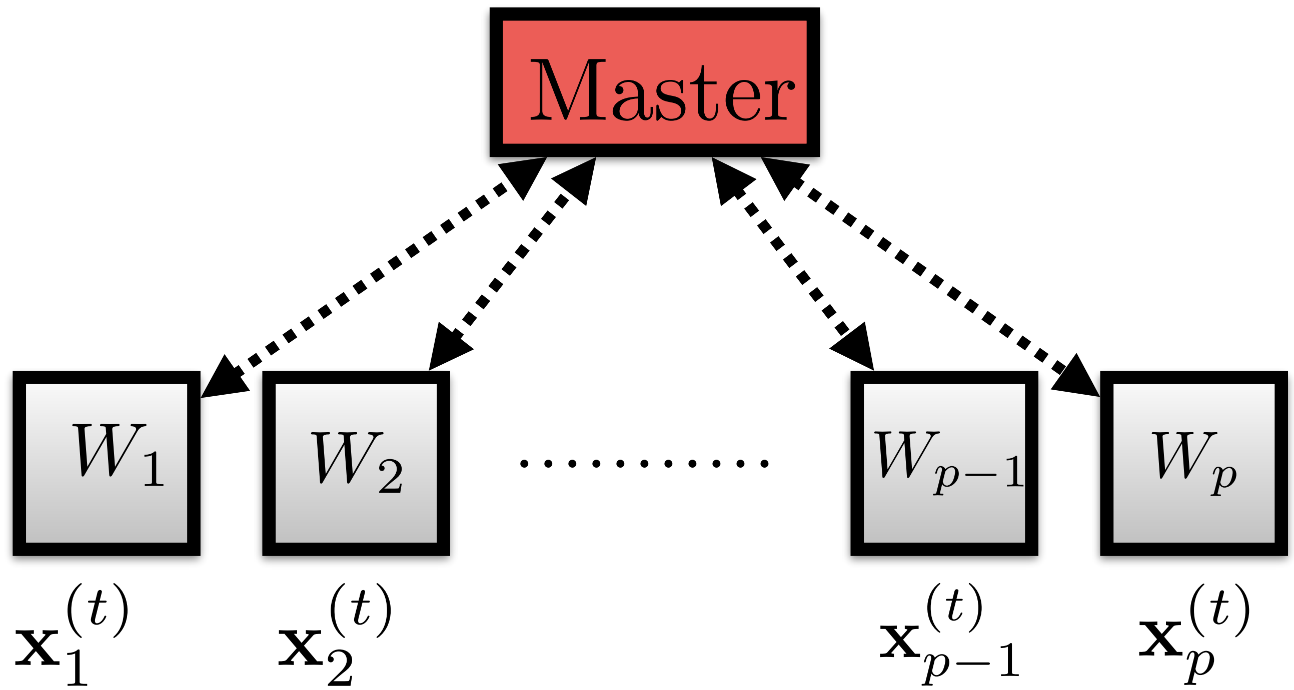
$$\mathbf{x}_j^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left[\mathbf{x}_j^{(t)} - \eta\,\tilde{\mathbf{g}}_j^{(t)}\right] \text{ if } \; t|\tau$$

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

Averaging step (a) ▶ Master

Local update (b) ▶ $W_j$

# Local SGD with periodic averaging




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

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

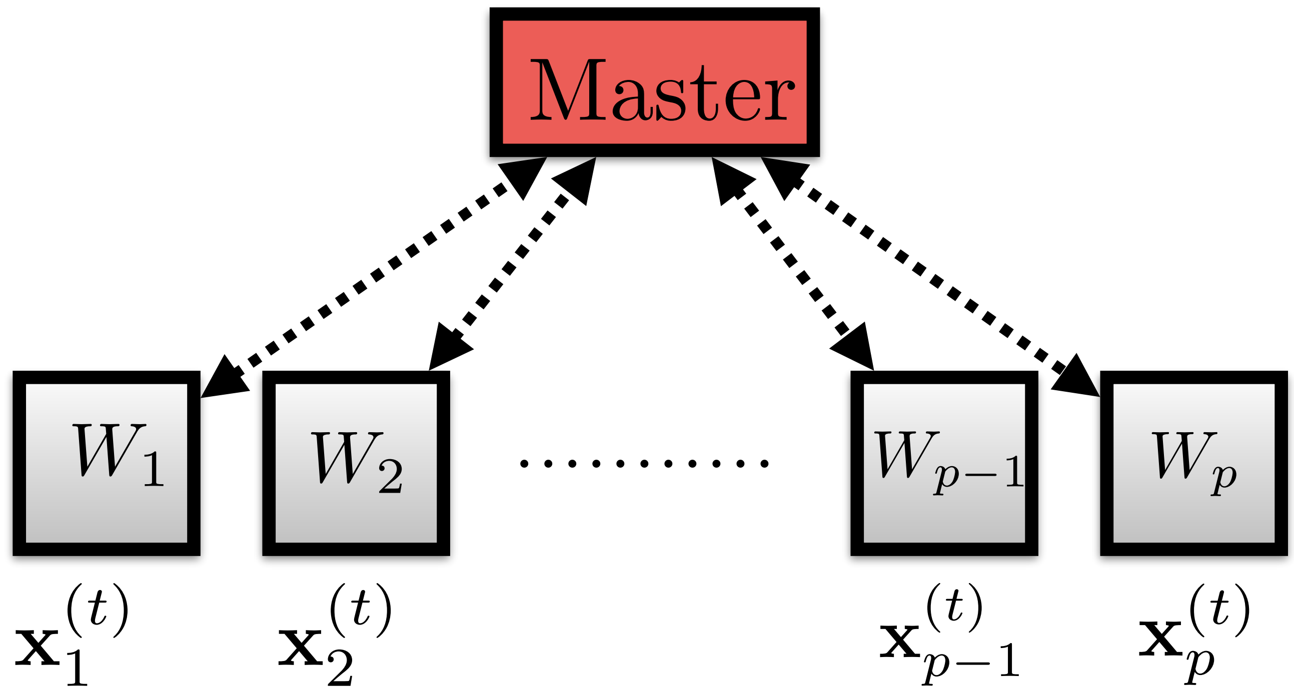
Averaging step (a) → Master

Local update (b) → $W_j$

$$\tilde{\mathbf{g}}_j^{(t)} = \nabla f(\mathbf{x}_j^{(t)}, \xi_j)$$

if $t|\tau$: $\bar{\mathbf{x}}^{(t)} = \mathbf{x}_j^{(t)}$ for $1 \leq j \leq p$

# 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 $\tau | T$ → Averaging step (a) → Master

$\mathbf{x}_j^{(t+1)} = \mathbf{x}_j^{(t)} - \eta\, \tilde{\mathbf{g}}_j^{(t)}$ otherwise, → Local update (b) → $W_j$

**Output:** $\bar{\mathbf{x}}^{(T)} = \dfrac{1}{p}\displaystyle\sum_{j=1}^{p} \mathbf{x}_j^{(T)}$

# Local SGD with periodic averaging

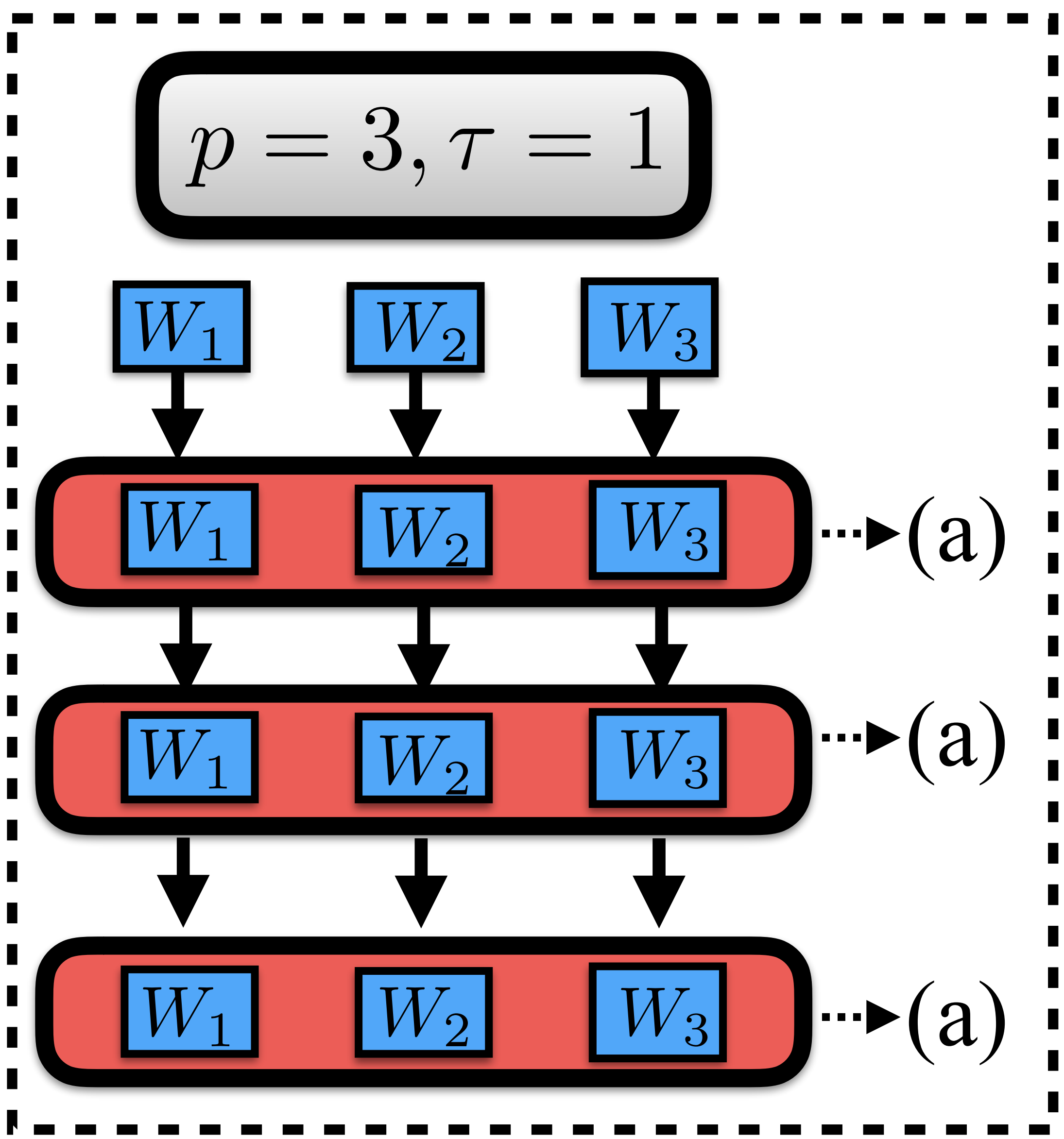
$$\mathbf{x}_j^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left[\mathbf{x}_j^{(t)} - \eta\, \tilde{\mathbf{g}}_j^{(t)}\right] \text{ if } \tau | T$$

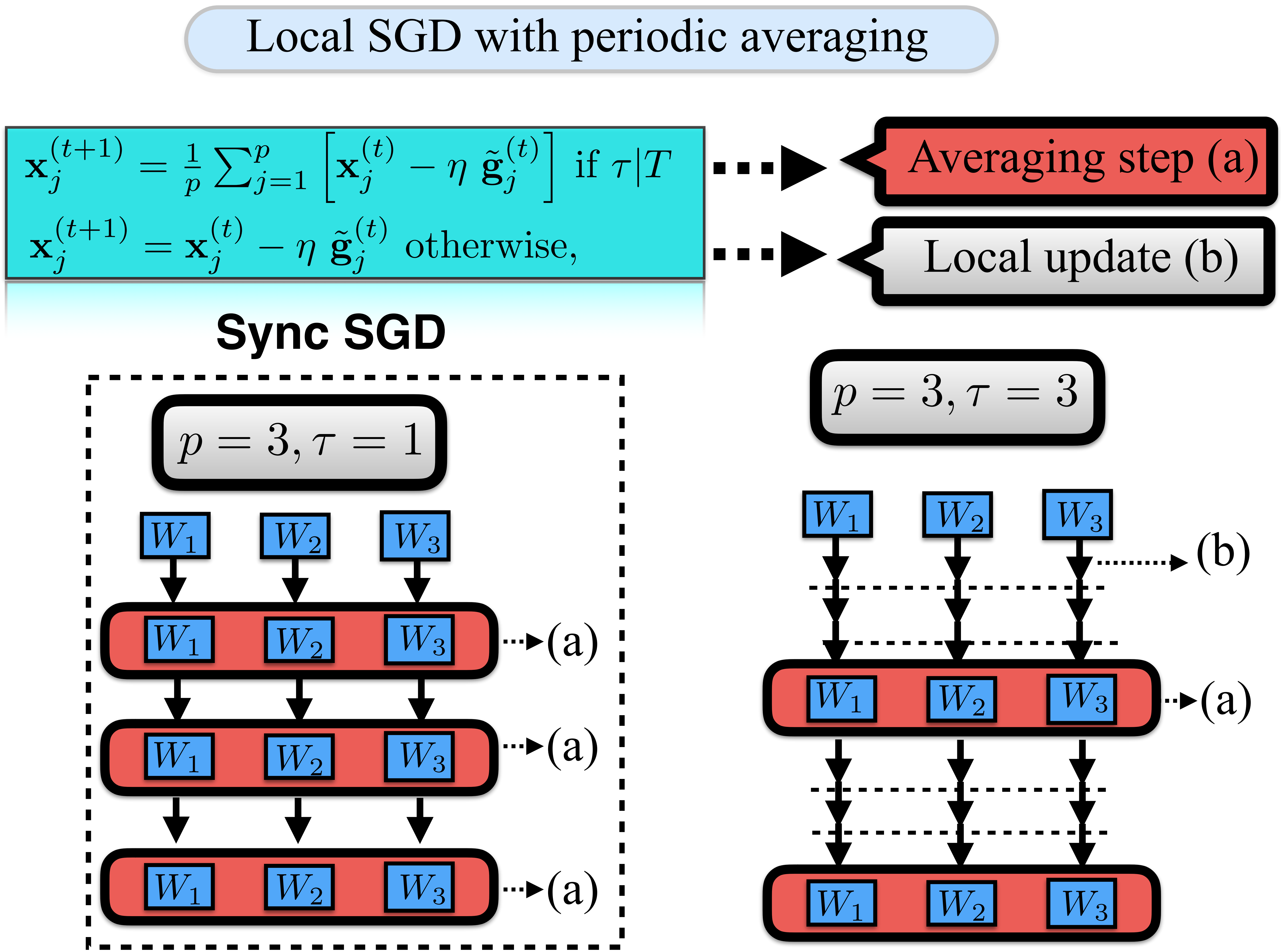
Averaging step (a)

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

Local update (b)

**Sync SGD**



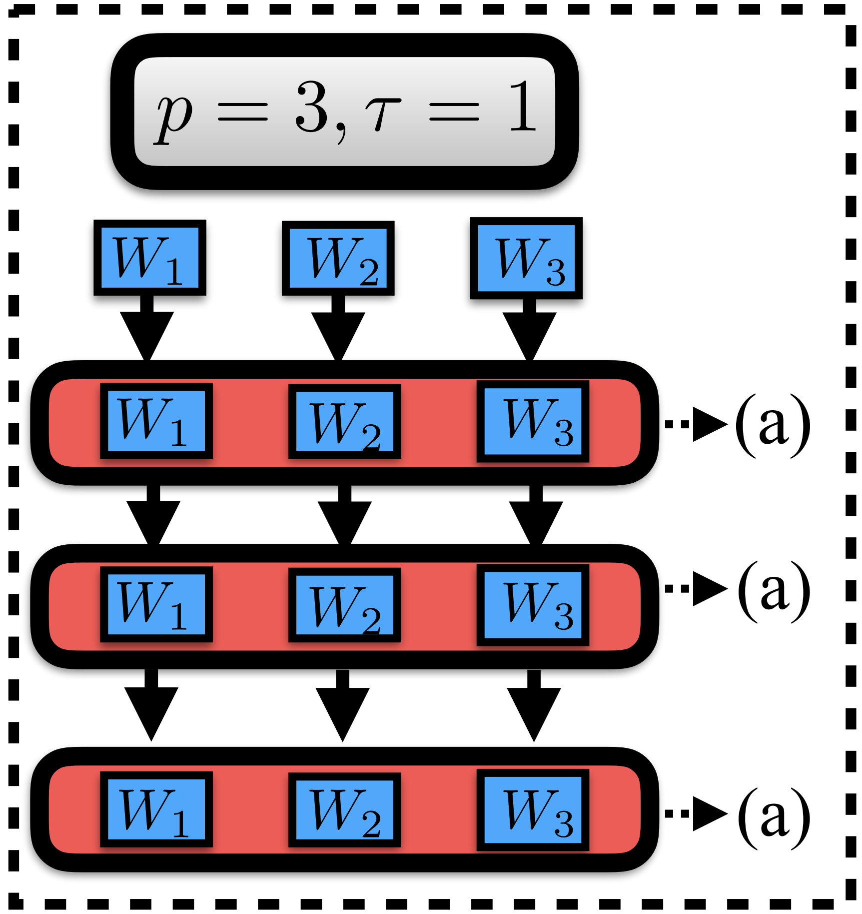


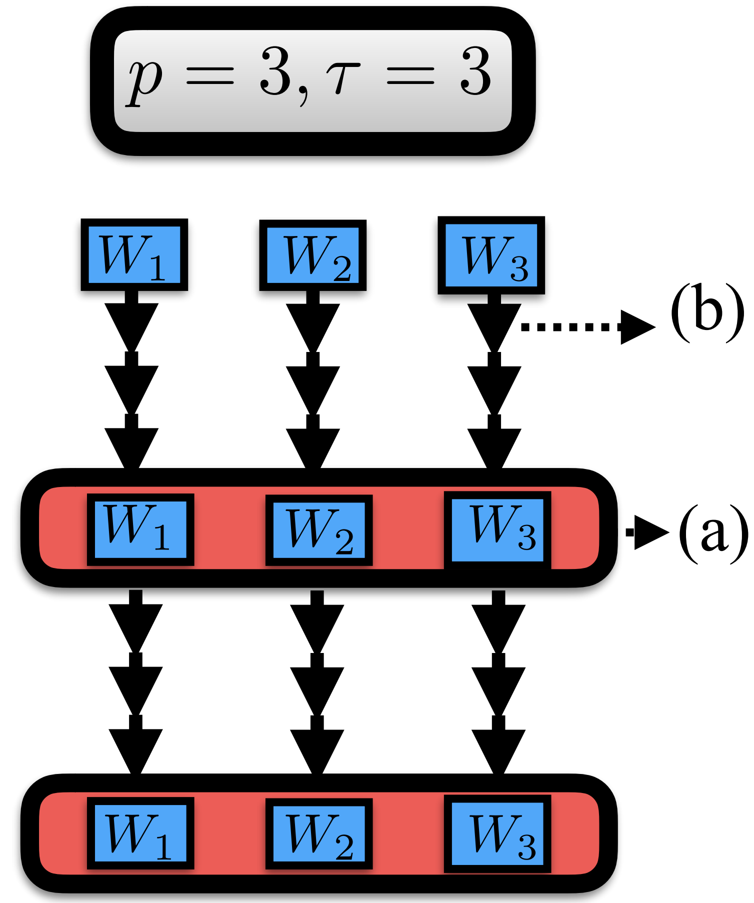
Local SGD with periodic averaging
$\mathbf{x}_j^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left[\mathbf{x}_j^{(t)} - \eta\,\tilde{\mathbf{g}}_j^{(t)}\right]$ if $\tau | T$
$\mathbf{x}_j^{(t+1)} = \mathbf{x}_j^{(t)} - \eta\,\tilde{\mathbf{g}}_j^{(t)}$ otherwise,
Averaging step (a)
Local update (b)
Sync SGD
$p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

# Local SGD with periodic averaging

$$R = \frac{T}{\tau}$$

**Sync SGD**




| | | |
|---|---|---|
| **Convergence error** | $O\left(\frac{1}{pT}\right) = O\left(\frac{1}{3T}\right)$ | $O\left(\frac{1}{pT}\right) = O\left(\frac{1}{3T}\right)$ |
| **Communication round** | $\frac{T}{\tau} = T$ | $\frac{T}{\tau} = \frac{T}{3}$ |

# State-of-the-art for R

$$\frac{1}{R}\sum_{r=1}^{R}\|\nabla f(\bar{\boldsymbol{w}}^{(r)})\|_2^2 \leq \epsilon$$

Number of communication rounds to achieve a stationary point with $\epsilon$ error.

# State-of-the-art for R

$$\frac{1}{R}\sum_{r=1}^{R}\|\nabla f(\bar{\boldsymbol{w}}^{(r)})\|_2^2 \leq \epsilon$$

Number of communication rounds to achieve a stationary point with $\epsilon$ error.

SCAFFOLD [Karimireddy et al, 2019]

$$R(\epsilon) = O\left(\frac{1}{\epsilon}\right)$$

# State-of-the-art for R

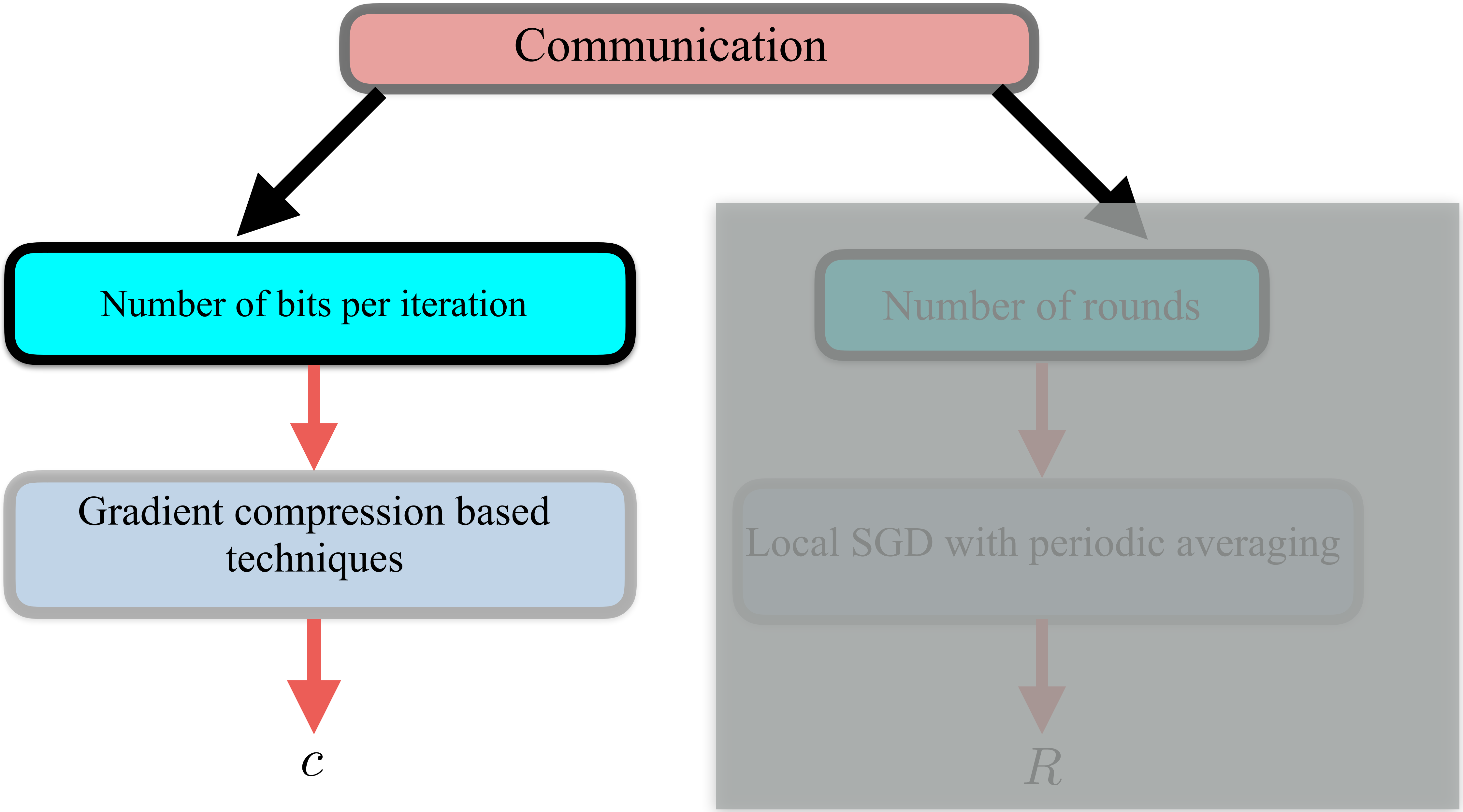
$$\frac{1}{R}\sum_{r=1}^{R}\left\|\nabla f(\bar{\boldsymbol{w}}^{(r)})\right\|_2^2 \leq \epsilon$$

Number of communication rounds to achieve a stationary point with $\epsilon$ error.

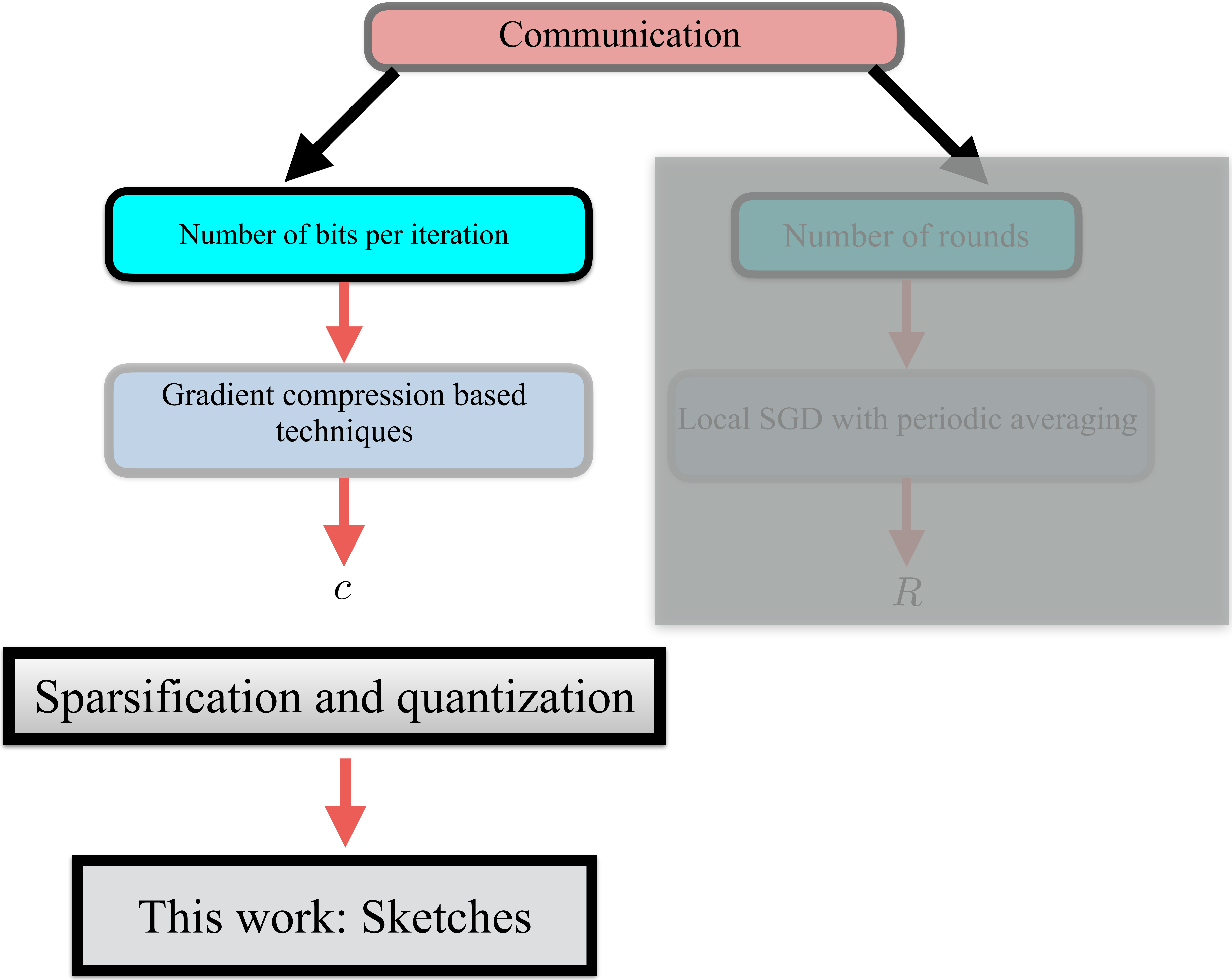
SCAFFOLD [Karimireddy et al, 2019]

$$R(\epsilon) = O\left(\frac{1}{\epsilon}\right) \overset{\mathbf{g}_i \in \mathbb{R}^d}{\Longrightarrow} Rc = O\left(\frac{d}{\epsilon}\right)$$

1. Federated learning review
2. Approaches to deal with communication cost
3. **Sketches**
4. Ongoing research




Sparsification and quantization



Communication
Number of bits per iteration
Gradient compression based techniques
$c$
Number of rounds
Local SGD with periodic averaging
$R$
Sparsification and quantization
This work: Sketches

**Sparsification or quantization**

State-of-the-art

**[Ivkin, Nikita, et al., 2019]**
**“Communication-efficient distributed sgd with sketching”**

$$\mathbf{g} \in \mathbb{R}^d \to \tilde{\mathbf{g}} \in \mathbb{R}^{\dim(S)}$$

# State-of-the-art

**[Ivkin, Nikita, et al., 2019]**
**"Communication-efficient distributed sgd with sketching"**

$$\mathbf{g} \in \mathbb{R}^d \to \tilde{\mathbf{g}} \in \mathbb{R}^{\dim(S)}$$

with probability at least$1 - \delta$,

$$c = O\left(k \log\left(\frac{d}{\epsilon\delta}\right)\right)$$

# State-of-the-art

**[Ivkin, Nikita, et al., 2019]**
**"Communication-efficient distributed sgd with sketching"**

$$\mathbf{g} \in \mathbb{R}^d \to \tilde{\mathbf{g}} \in \mathbb{R}^{\dim(S)}$$

$$\text{with probability at least} 1-\delta, \quad R = O(\frac{1}{\epsilon^2})$$

$$c = O\left(k \log\left(\frac{d}{\epsilon^2\delta}\right)\right), \text{and } Rc = O\left(\frac{k}{\epsilon^2}\log\left(\frac{d}{\epsilon^2\delta}\right)\right)$$

# Short-comings

**[Ivkin, Nikita, et al., 2019]**
**"Communication-efficient distributed SGD with sketching"**

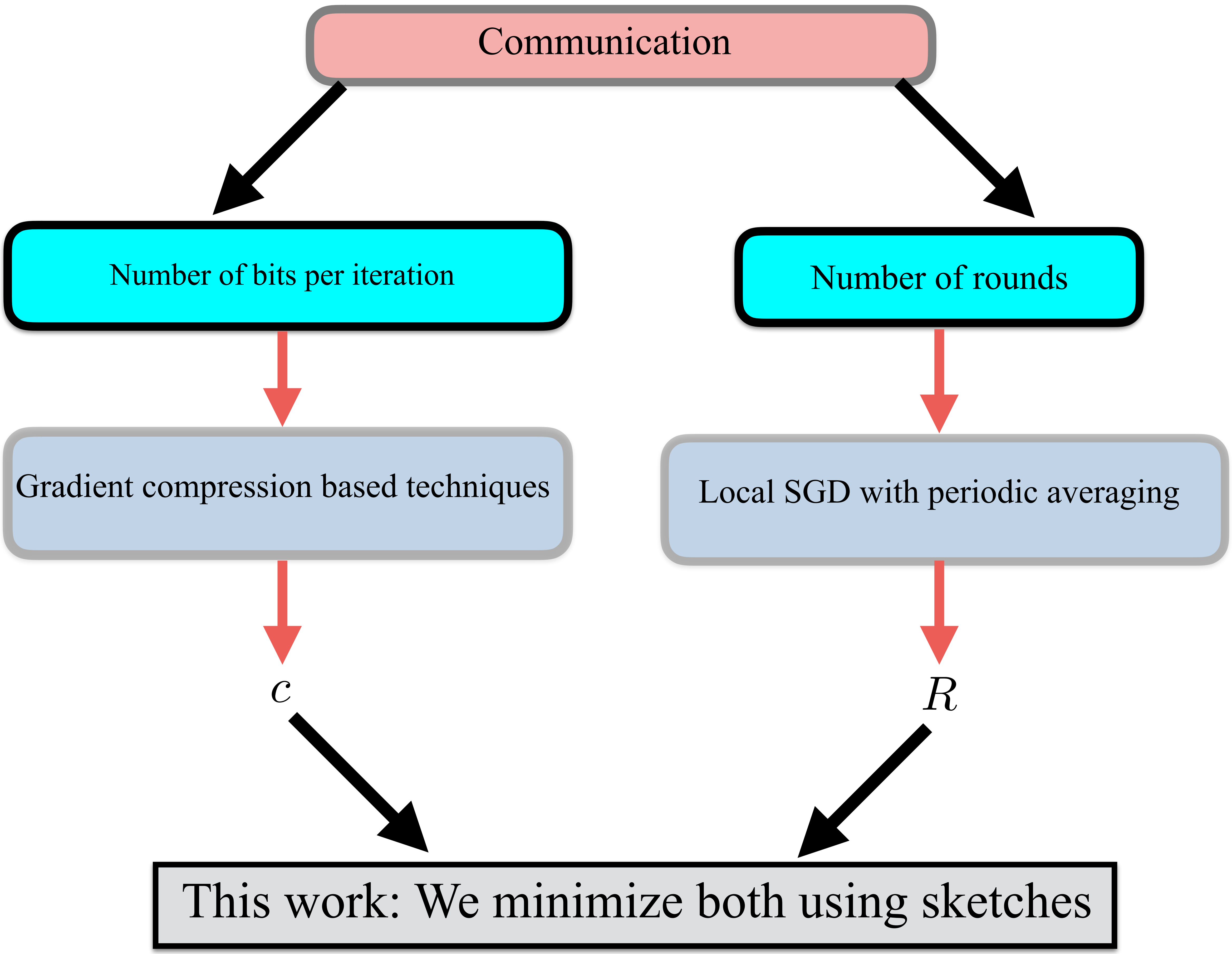
- **Higher communication rounds**
- **Not private**
- **One machine analysis**
- **Strong assumptions**
- **Only for homogenous setting**

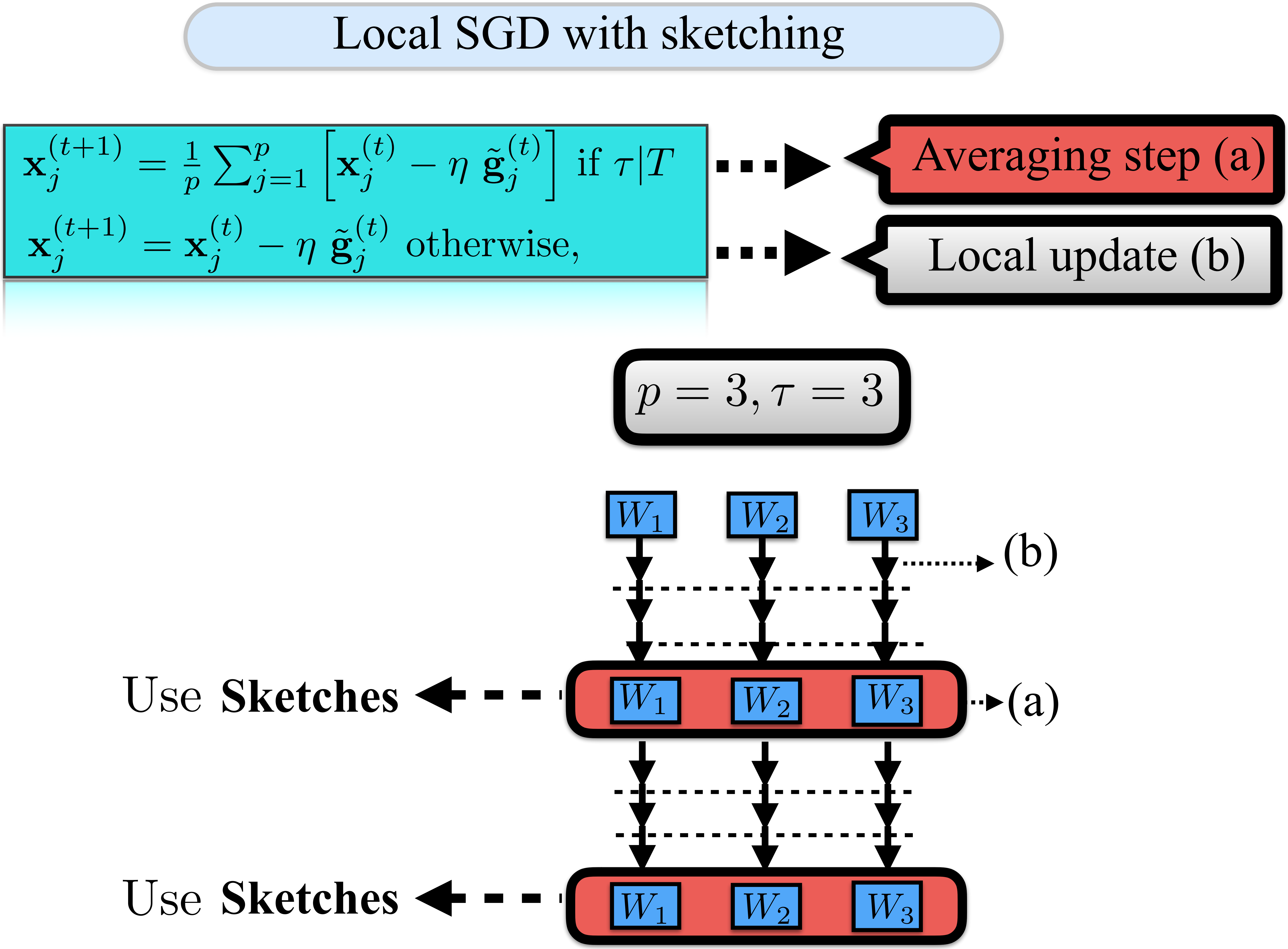
# Short-comings

**[Ivkin, Nikita, et al., 2019]**
**"Communication-efficient distributed SGD with sketching"**

- **Higher communication rounds**
- **Not private**
- **One machine analysis**
- **Strong assumptions**
- **Only for homogenous setting**

**How to improve? This paper!**


Communication
Number of bits per iteration
Number of rounds
Gradient compression based techniques
Local SGD with periodic averaging
$c$
$R$
This work: We minimize both using sketches



Local SGD with sketching
$\mathbf{x}_j^{(t+1)} = \frac{1}{p}\sum_{j=1}^{p}\left[\mathbf{x}_j^{(t)} - \eta\,\tilde{\mathbf{g}}_j^{(t)}\right]$ 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)
$p = 3, \tau = 3$
$W_1$
$W_2$
$W_3$
(b)
(a)
Use **Sketches**
Use **Sketches**

# Our result for homogenous setting and general non-convex

$$\mathbf{g} \in \mathbb{R}^d \rightarrow \tilde{\mathbf{g}} \in \mathbb{R}^{\dim(S)}$$

with probability at least$1-\delta$, $\quad R = O(\frac{1}{\epsilon})$

$$\boxed{c = O\left(k \log\left(\frac{d}{\epsilon\delta}\right)\right), \text{and } Rc = O\left(\frac{k}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right)}$$

# General non-convex

| Scheme | $Rc$ | Differentially Privacy | Hetregenous Distribution |
|---|---|---|---|
| [Ivkin, Nikita, et al., 2019] | $O\left(\frac{k}{\epsilon^2}\log\left(\frac{d}{\epsilon^2\delta}\right)\right)$ | ✗ | ✗ |
| [Li, Tian,2019] | - | ✓ | ✓ |
| Scaffold [Karimireddy,19] | $O\left(\frac{d}{\epsilon}\right)$ | ✗ | ✓ |
| **FedSketch** | $O\left(\frac{k}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right)$ | ✓ | ✗ |

**Interesting Observation: Improvement for non-convex is much better than strongly convex objectives**

# References


- Ivkin, N., Rothchild, D., Ullah, E., Stoica, I., & Arora, R. (2019). Communication-efficient distributed sgd with sketching. In *Advances in Neural Information Processing Systems* (pp. 13144-13154).

- Li, T., Liu, Z., Sekar, V., & Smith, V. (2019). Privacy for Free: Communication-Efficient Learning with Differential Privacy Using Sketches. *arXiv preprint arXiv:1911.00972*.

- Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2019). SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*.

1. Federated learning review
2. Approaches to deal with communication cost
3. Sketches
4. **Ongoing research**

## Ongoing Directions:

1. Extension to hetregenous setting
2. Improving communication efficiency using different algorithms
3. Using different sketching

Thanks for your attention!