Distributed Random Reshuffling Methods over Networks

香港中文大學（深圳） The Chinese University of Hong Kong, Shenzhen

深圳市大数据研究院 Shenzhen Research Institute of Big Data

Kun Huang[1,2], Linli Zhou[1] & Shi Pu[1,2]

[1]School of Data Science (SDS), The Chinese University of Hong Kong, Shenzhen

[2]Shenzhen Research Institute of Big Data (SRIBD)

Problem

Consider

$$\min_{x\in\mathbb{R}^p} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \text{ with } f_i(x) = \frac{1}{m}\sum_{\ell=1}^{m} f_{i,\ell}(x). \tag{1}$$

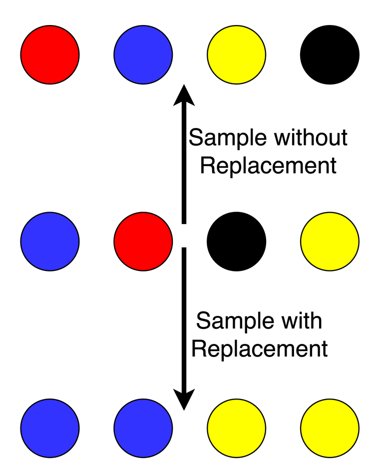


Figure 1. Sample with or without replacement

- RR permits the utilization of all the data points in every epoch
- RR can provably reduce the variance of stochastic gradient.

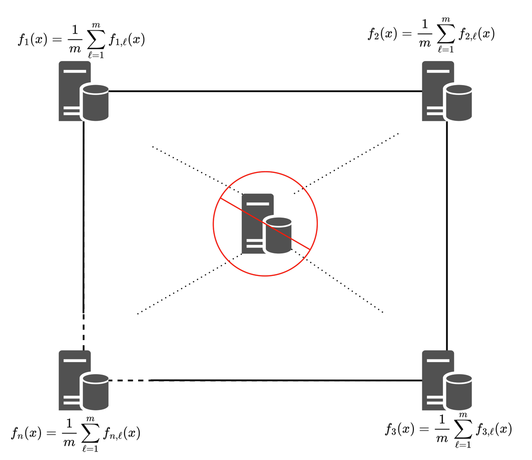


Figure 2. A decentralized architecture

- It reduces the **latency** and improves the algorithmic **robustness** compared to centralized learning algorithms with a central server.
- Assumption 1: The underlying graph is undirected and strongly connected. The mixing matrix W is symmetric and stochastic.

- **Q1**: Can we design an efficient distributed RR algorithm over networks with similar convergence guarantees as centralized RR?
- **Q2**: Can we minimize the impact of the network topology on the convergence rate compared to the existing algorithms, while keeping the goal of Q1?

Algorithms

Initialize $x_{i,0}^0$ and $\{\alpha_t\}$.
for *Epoch* $t \leftarrow 0$ *to* $T-1$ **do**
  for *Agent* i *in parallel* **do**
    Independently sample a random permutation $\{\boldsymbol{\pi_0^i, \pi_1^i, \ldots, \pi_{m-1}^i}\}$ of $\{1,2,\ldots,m\}$.
    for $\ell = 0,1,\ldots,m-1$ **do**

- D-RR:

$$x_{i,t}^{\ell+1} = \underbrace{\sum_{j\in\mathcal{N}_i} w_{ij}}_{\text{Communication}} \underbrace{\left(x_{j,t}^{\ell} - \alpha_t \nabla f_{j,\pi_\ell^j}(x_{j,t}^{\ell})\right)}_{\text{Computation}}$$

- GT-RR: $y_{i,t}^0 = \nabla f_{i,\pi_0^i}(x_{i,t}^0)$

$$x_{i,t}^{\ell+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\left(x_{j,t}^{\ell} - \alpha_t y_{j,t}^{\ell}\right)$$
$$y_{i,t}^{\ell+1} = \sum_{j\in\mathcal{N}_i} w_{ij} y_{j,t}^{\ell} + \nabla f_{i,\pi_{\ell+1}^i}(x_{i,t}^{\ell+1}) - \nabla f_{i,\pi_\ell^i}(x_{i,t}^{\ell}), \ell \neq m-1$$

- ED-RR:

$$x_{i,t}^{1} = \sum_{j\in\mathcal{N}_i} w_{ij}\left(x_{j,t}^{0} - \alpha_t \nabla f_{j,\pi_0^j}(x_{j,t}^{0})\right)$$
$$x_{i,t}^{\ell+1} = \sum_{j\in\mathcal{N}_i} w_{ij}\left(2x_{i,t}^{\ell} - x_{i,t}^{\ell-1} - \alpha_t(\nabla f_{i,\pi_\ell^i}(x_{i,t}^{\ell}) - \nabla f_{i,\pi_\ell^i}(x_{i,t}^{\ell-1}))\right)$$

  Set $x_{i,t+1}^0 = x_{i,t}^m$.
Output $x_{i,T}^0$

Unified form: $\mathbf{x}_t^{\ell} : (x_{1,t}^{\ell}, x_{2,t}^{\ell}, \ldots, x_{n,t}^{\ell})^T$

$$\mathbf{x}_t^{\ell+1} = A(C\mathbf{x}_t^{\ell} - \alpha_t \nabla F_{\pi_\ell}(\mathbf{x}_t^{\ell})) - B\mathbf{z}_t^{\ell}$$
$$\mathbf{z}_t^{\ell+1} = \mathbf{z}_t^{\ell} + B\mathbf{x}_t^{\ell+1}, \ell = 0,1,\ldots,m-1$$
$$\mathbf{x}_{t+1}^0 = \mathbf{x}_t^m$$

GT-RR:	ED-RR:	D-RR:
$A = W$, $B = I - W$	$A = W$, $B = (I-W)^{1/2}$	$A = W$, $B = 0$
$C = W$	$C = I$	$C = W$
$\mathbf{z}_t^0 = -W\mathbf{x}_t^0$	$\mathbf{z}_t^0 = 0$	

Intuition

- Averaged over all the agents, due to $\mathbf{1}^\intercal W = \mathbf{1}^\intercal$,

$$\bar{x}_t^{\ell+1} = \bar{x}_t^{\ell} - \alpha_t \frac{1}{n}\sum_{i=1}^{n} \nabla f_{i,\pi_\ell^i}(x_{i,t}^{\ell}),\ \bar{x}_t^{\ell} := \frac{1}{n}\sum_{i=1}^{n} x_{i,t}^{\ell} \tag{2}$$

- The original problem can be rewritten as

$$\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{\ell=1}^{m} f_{i,\ell}(x)\right) \longrightarrow \min_{x\in\mathbb{R}^p} \frac{1}{m}\sum_{\ell=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n} f_{i,\pi_\ell^i}(x)\right) \tag{3}$$

- The update (2) can be seen as approximately implementing the centralized RR method for solving Problem (3).

$$\bar{x}_t^{\ell+1} = \bar{x}_t^{\ell} - \frac{\alpha_t}{n}\sum_{i=1}^{n} \nabla f_{i,\pi_\ell^i}(\bar{x}_t^{\ell}) + \alpha_t \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \nabla f_{i,\pi_\ell^i}(\bar{x}_t^{\ell}) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_{i,\pi_\ell^i}(x_{i,t}^{\ell})\right)}_{}$$

Assumption 2: Each $f_{i,\ell}$ is lower bounded and has Lipschitz continuous gradient.

- One of the key ingredients is to estimate $\sum_{\ell=0}^{m-1}\sum_{i=1}^{n}\|x_{i,t}^{\ell} - \bar{x}_t^{\ell}\|^2$ and $\sum_{i=1}^{n}\|x_{i,t}^{0} - \bar{x}_t^{0}\|^2$.
- One challenge is that $\mathbb{E}[\nabla f_{i,\pi_\ell^i}(x_{i,t}^{\ell})|\mathcal{F}_t^{\ell-1}] \neq \nabla f_i(x_{i,t}^{\ell})$, where the filtration $\mathcal{F}_t^{\ell}$ $(\ell = 0,1,\ldots,m-1)$ is generated by $\{x_{i,p}^{j} | i\in[n], j = 0,1,\ldots,\ell, p = 0,1,\ldots,t\}$. However, we observe that $\mathbb{E}[\nabla f_{i,\pi_\ell^i}(x_{i,t}^{0})|\mathcal{F}_t^{0}] = \nabla f_i(x_{i,t}^{0})$.

Results: Smooth Nonconvex Objective Functions

Algorithm	Convergence Rate
Centralized RR	$\mathcal{O}\left(\frac{1}{\boldsymbol{m^{1/3}T^{2/3}}}\right)$
DSGD	$\mathcal{O}\left(\frac{1}{\sqrt{mnT}} + \frac{n^2}{(1-\lambda)^2 mT}\right)$
D-RR	$\mathcal{O}\left(\frac{1}{(1-\lambda)T^{2/3}}\right)$ [1] (a)
D-RR (**New**)	$\mathcal{O}\left(\frac{1}{\boldsymbol{(1-\lambda)^{2/3}m^{1/3}T^{2/3}}} + \frac{1}{(1-\lambda)T}\right)$
DSGT	$\mathcal{O}\left(\frac{1}{\sqrt{mnT}} + \frac{n}{(1-\lambda)mT} + \frac{n}{(1-\lambda)^4 m^2T^2}\right)$
GT-RR	$\mathcal{O}\left(\frac{1}{\boldsymbol{(1-\lambda)^{1/3}m^{1/3}T^{2/3}}} + \frac{1}{(1-\lambda)T} + \frac{1}{m^{7/3}(1-\lambda)^{7/3}T^{5/3}}\right)$
ED/ D^2	$\mathcal{O}\left(\frac{1}{\sqrt{mnT}} + \frac{n}{(1-\lambda)mT} + \frac{n}{(1-\lambda)^2 m^2T^2}\right)$
ED-RR	$\mathcal{O}\left(\frac{1}{\boldsymbol{(1-\lambda)^{1/3}m^{1/3}T^{2/3}}} + \frac{1}{(1-\lambda)T} + \frac{1}{m^{7/3}(1-\lambda)^{4/3}T^{5/3}}\right)$

(a) The result is obtained by minimizing over the arbitrary constant η in the original result $\mathcal{O}(1/(\eta T^{2/3}) + \eta^2/[(1-\lambda)^3 T^{2/3}])$ in [1].

Table 1. A summary of related the theoretical results using a constant stepsize.

Results: the PL Condition Case

Algorithm	Final Error Bound (Constant Stepsize α)	Convergence Rate (Decreasing Stepsize)
Centralized RR	$\mathcal{O}(\boldsymbol{m\alpha^2})$	$\mathcal{O}\left(\frac{\log(T)}{\boldsymbol{mT^2}}\right)$
DSGD	$\mathcal{O}\left(\frac{\alpha}{n} + \frac{\alpha^2}{(1-\lambda)^2}\right)$ (b)	$\mathcal{O}\left(\frac{1}{mnT} + \frac{1}{(1-\lambda)^2 m^2T^2}\right)$ (b)
D-RR [1]	$\mathcal{O}\left(\frac{m\alpha^2}{(1-\lambda)^3}\right)$ (b)	$\mathcal{O}\left(\frac{1}{(1-\lambda)^3 mT^2}\right)$ (b)
DSGT	$\mathcal{O}\left(\frac{\alpha}{n} + \frac{\alpha^2}{1-\lambda} + \frac{\alpha^4}{n(1-\lambda)^4}\right)$	$\mathcal{O}\left(\frac{1}{mnT} + \frac{1}{(1-\lambda)^3 m^2T^2}\right)$ (b)
GT-RR	$\mathcal{O}\left(\frac{\boldsymbol{m\alpha^2}}{\boldsymbol{1-\lambda}} + \frac{m^4\alpha^4}{(1-\lambda)^2}\right)$	$\mathcal{O}\left(\frac{1}{\boldsymbol{(1-\lambda)mT^2}} + \frac{1}{(1-\lambda)^2 T^4}\right)$
ED/ D^2	$\mathcal{O}\left(\frac{\alpha}{n} + \frac{\alpha^2}{(1-\lambda)} + \frac{\alpha^4}{n(1-\lambda)^3}\right)$	$\mathcal{O}\left(\frac{1}{mnT} + \frac{1}{(1-\lambda)m^2T^2}\right)$ (b)
ED-RR	$\mathcal{O}\left(\frac{\boldsymbol{m\alpha^2}}{\boldsymbol{1-\lambda}} + \frac{m^4\alpha^4}{(1-\lambda)^2}\right)$	$\mathcal{O}\left(\frac{1}{\boldsymbol{(1-\lambda)mT^2}} + \frac{1}{(1-\lambda)^2 T^4}\right)$

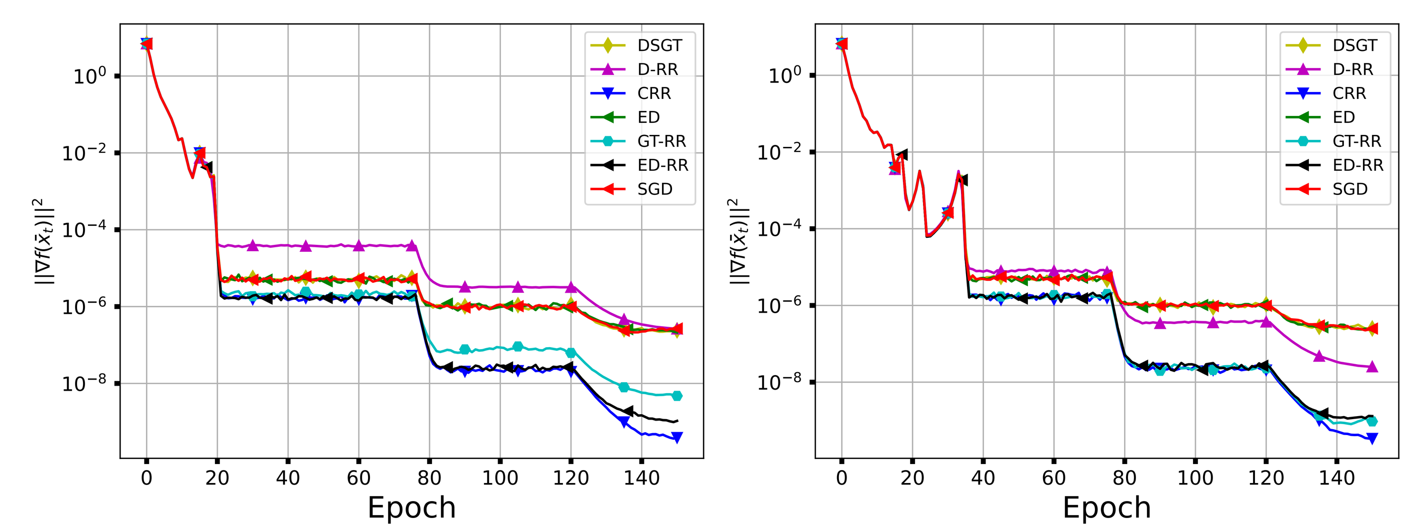
(b) The results are obtained for smooth strongly convex objective functions.

Table 2. A summary of the related theoretical results under smooth objective functions satisfying the PL condition.

Take-away Messages

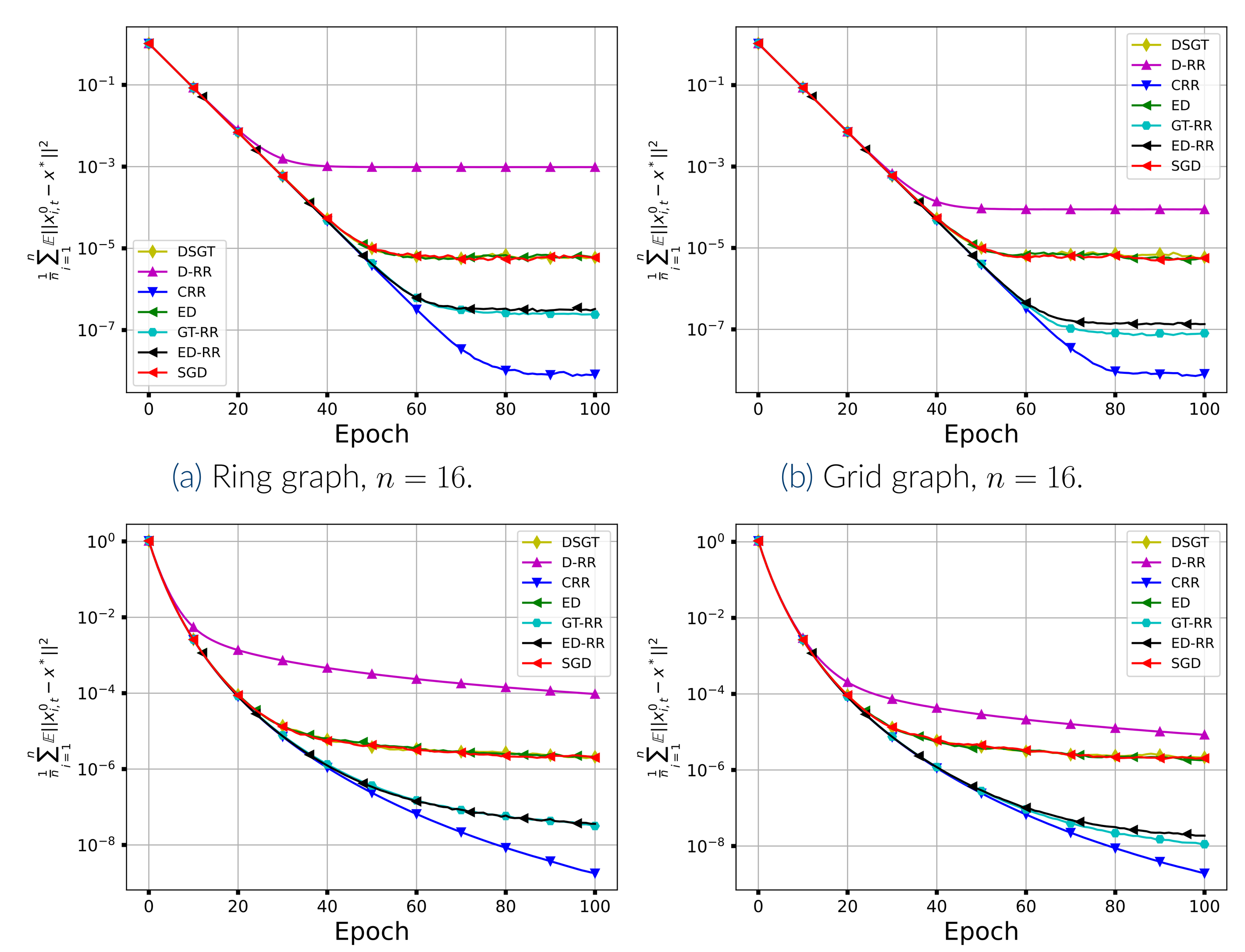
- We can design distributed RR methods that achieve comparable convergence rate to centralized RR while reducing the influence of the networks.
- In light of the unified framework and techniques in this work, many previous distributed stochastic gradient methods can also be equipped with RR updates similarly.

Simulations: Data Heterogeneous Setting



(a) Ring graph, $n = 16$. (b) Grid graph, $n = 16$.

Figure 3. Comparison among ED-RR, GT-RR, D-RR, ED, DSGT, SGD, and centralized RR for solving logistic regression with nonconvex regularizer on the CIFAR-10 dataset using a constant stepsize. The stepsizes are sequentially set as $1/50$, $1/250$, and $1/1000$.



(a) Ring graph, $n = 16$. (b) Grid graph, $n = 16$.

(c) Ring graph, $n = 16$. (d) Grid graph, $n = 16$.

Figure 4. Comparison among ED-RR, GT-RR, D-RR, ED, DSGT, SGD, and centralized RR for solving logistic regression on the CIFAR-10 dataset. **First row:** The stepsize is set as $\alpha = 0.001$ for all the methods. **Second row:** The stepsize is set as $\alpha_t = 1/(30t + 300)$ for all the methods.

References

[1] K. Huang, X. Li, A. Milzarek, S. Pu, and J. Qiu. *Distributed random reshuffling over networks*, IEEE Transactions on Signal Processing, 71 (2023), pp. 1143–1158.

[2] K. Huang, L. Zhou, and S. Pu. *Distributed random reshuffling methods with improved convergence*, 2023. https://arxiv.org/abs/2306.12037.

{kunhuang, linlizhou}@link.cuhk.edu.cn