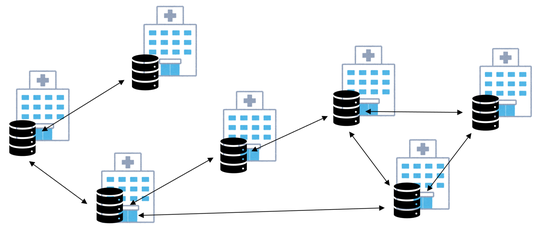
Beyond Exponential Graph: Communication-Efficient Topologies for Decentralized Learning via Finite-time Convergence

Yuki Takezawa[1,2], Ryoma Sato[1,2], Han Bao[1,2], Kenta Niwa[3], Makoto Yamada[2]

[1]Kyoto University, [2]OIST, [3]NTT Communication Science Laboratories

Background

- **Decentralized learning**
  - Decentralized learning can preserve privacy because it does not need to aggregate all training data into one server.



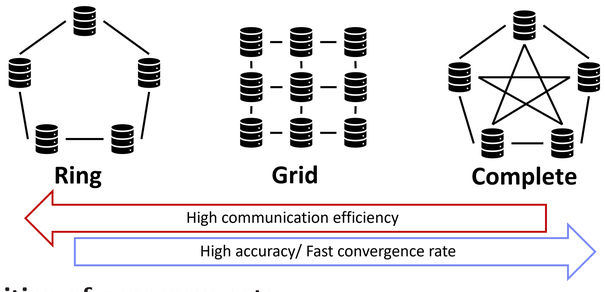
  - Let the number of nodes be n and the loss function of node i be $f_i(x)$, decentralized learning is formulated as follows:

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{1}$$

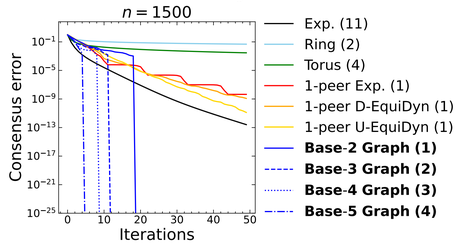
- **Decentralized SGD**
  - W is the mixing matrix that satisfies $\sum_i W_{ij} = \sum_j W_{ij} = 1$.

$$x_i^{(r+1)} = \sum_{j=1}^{n} W_{ij}\left(x_j^{(r)} - \eta \nabla F_j(x_j^{(r)};\xi)\right) \tag{2}$$

- **Trade-off between communication cost and convergence rate**
  - The smaller the maximum degree of an underlying network topology is, the fewer the communication costs become.
  - The faster the consensus rate (a.k.a. spectral gap) of a topology is, the faster the convergence rate of decentralized learning becomes.



- **Definition of consensus rate**
  - Let x_i be the parameters that node i has.
  - Consensus rate is the speed that x_i reaches the average $\frac{1}{n}\sum_{j=1}^{n} x_i$ when x_i is updated as $x_i \leftarrow \sum_{j=1}^{n} W_{ij}x_j$.



Contribution

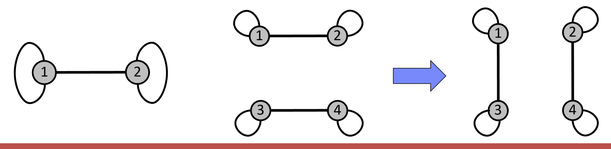
We propose **Base-$(k+1)$ Graph**.

- It is finite-time convergence for any number of nodes n and maximum degree k.
- It can endow Decentralized SGD with a faster convergence rate and more communication efficiency than existing graphs.

Topology	Convergence Rate ↓	Max Degree ↓	#Nodes n
Ring	$\mathcal{O}\left(\frac{1}{n\epsilon^2}+\frac{n^2}{\epsilon^{3/2}}+\frac{n^2}{\epsilon}\right)$	2	$\forall n \in \mathbb{N}$
Torus	$\mathcal{O}\left(\frac{1}{n\epsilon^2}+\frac{n}{\epsilon^{3/2}}+\frac{n}{\epsilon}\right)$	4	$\forall n \in \mathbb{N}$
Exp.	$\mathcal{O}\left(\frac{1}{n\epsilon^2}+\frac{\log_2(n)}{\epsilon^{3/2}}+\frac{\log_2(n)}{\epsilon}\right)$	$\lceil\log_2(n)\rceil$	$\forall n \in \mathbb{N}$
1-peer Exp.	$\mathcal{O}\left(\frac{1}{n\epsilon^2}+\frac{\log_2(n)}{\epsilon^{3/2}}+\frac{\log_2(n)}{\epsilon}\right)$	1	A power of 2
1-peer Hypercube	$\mathcal{O}\left(\frac{1}{n\epsilon^2}+\frac{\log_2(n)}{\epsilon^{3/2}}+\frac{\log_2(n)}{\epsilon}\right)$	1	A power of 2
Base-$(k+1)$ Graph	$\mathcal{O}\left(\frac{1}{n\epsilon^2}+\frac{\log_{k+1}(n)}{\epsilon^{3/2}}+\frac{\log_{k+1}(n)}{\epsilon}\right)$	k	$\forall n \in \mathbb{N}$

Related Work

- **1-peer Hypercube Graph & 1-peer Exp. Graph**
  - They are finite-time convergence only when the number of nodes n is a power of two.
  - In the figure below, all edge weights are $\frac{1}{2}$.

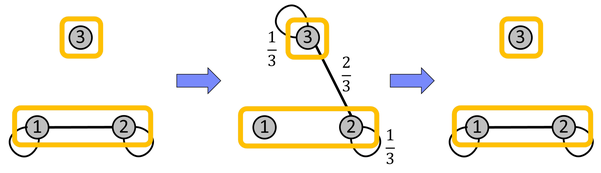


Proposed Method

Base-$(k+1)$ Graph

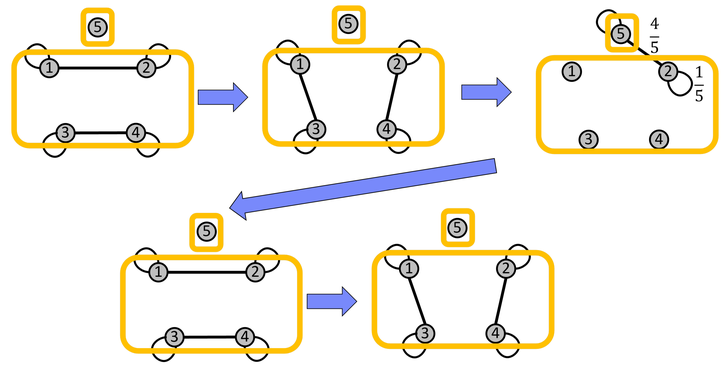
- Edge weights are omitted when they are $\frac{1}{2}$.

- **Example of Base-2 Graph with $n = 3(= 2+1)$**



	Node 1	Node 2	Node 3
Initial value	x_1	x_2	x_3
	$\frac{x_1+x_2}{2}$	$\frac{x_1+x_2}{2}$	x_3
	$\frac{x_1+x_2}{2}$	$\frac{x_1+x_2+4x_3}{6}$	$\frac{x_1+x_2+x_3}{3}$
	$\frac{x_1+x_2+x_3}{3}$	$\frac{x_1+x_2+x_3}{3}$	$\frac{x_1+x_2+x_3}{3}$

- **Example of Base-2 Graph with $n = 5(= 2^2+1)$**



Theorem (Length of Base-$(k+1)$ Graph)

For any n and k, the length of Base-$(k+1)$ Graph is less than or equal to $2\log_{k+1}(n)+2$.

Results

- Results with $n = 25$.

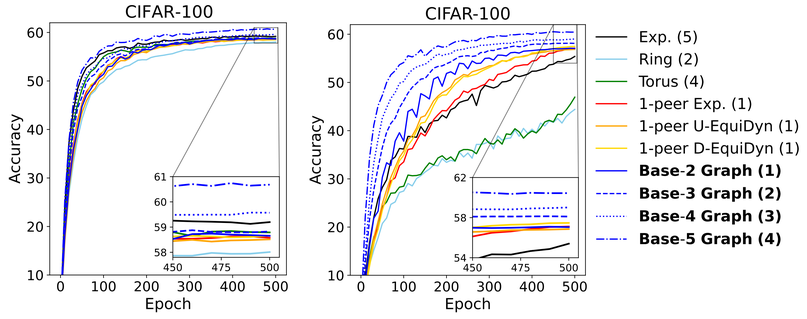


Figure: Left: I.I.D. Right: non-I.I.D.

Conclusions

We proposed **Base-$(k+1)$ Graph**.

- **Theoretically:** It can achieve a faster convergence rate and more communication efficiency.
- **Experimentally:** It can achieve stable training and high accuracy.