Communication-Efficient Topologies for Decentralized Learning with $O(1)$ Consensus Rate

Advances in Neural Information Processing Systems (2022)

Alice H. Oh and Alekh Agarwal and Danielle Belgrave and Kyunghyun Cho
Reporter: Fengjiao Gong

April 19, 2023

Outline

1. Background
2. Proposed topology — EquiTopo
3. Applying EquiTopo to decentralized learning
4. Experiments

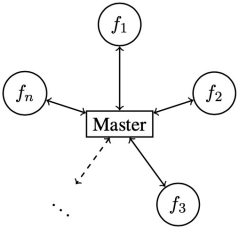
Background

Consider the following distributed problem over a network of n computing nodes:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \text{ where } f_i(\boldsymbol{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[F(\boldsymbol{x};\xi_i)\right].$$

- ξ_i is the local data following local distribution $\mathcal{D}_i$
- $f_i(\boldsymbol{x})$ is kept at node i
- $\boldsymbol{x}_i^{(t)}$ is node i's local model at iteration t
- $\overline{\boldsymbol{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^{(t)}$.

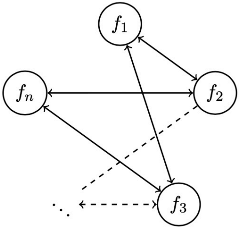
Background



Centralized learning — simple & useful

- The master node may become *a communication bottleneck* if it has limited communication resources, as the number of nodes n grows large.
- The master node may become *a robustness bottleneck* in the sense that if the master node fails then the entire network fails.
- Impractical to have a single master node that communicates with all agents, or to have all nodes within the required proximity of the master.

Background



Decentralized learning — connected

- Lower overhead in per-iteration communication
- Less effective in mixing information and slower convergence

Notation

Given graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V}$ and directed edges $\mathcal{E}$,

- An edge $(j, i) \in \mathcal{E}$ means node j can directly send information to node i.
- Node i's degree is the number of its in-neighbors $|\{j \mid (j, i) \in \mathcal{E}\}|$.
- A one-peer graph means that the degree for each node is at most 1 .
- For undirected graphs, $(j, i) \in \mathcal{E}$ if and only if $(i, j) \in \mathcal{E}$.

Background – Directed graph

Equal-neighbor update rule in a directed graph: at step k,

- node i broadcasts the value x_i^k to its out-neighbors
- receives values x_j^k from its in-neighbors
- sets x_i^{k+1} to be the average of the messages it has received

$$x_i^{k+1} = \frac{1}{d_i^{\text{in},k}} \sum_{j \in N_i^{\text{in},k}} x_j^k$$

Node i repeatedly revises it's opinion vector x_i^k by averaging the opinions of it's neighbors.

Background — Undirected graph

Over undirected graphs, an alternative popular choice of update rule is to set

$$x_i^{k+1} = x_i^k + \epsilon \sum_{j \in N_i^k} \left(x_j^k - x_i^k \right)$$

where $\epsilon > 0$ is sufficiently small.

Background — Linear consensus process

Both can be written in the form of the **linear consensus process** defined as

$$x^{k+1} = A^k x^k, \quad k = 0, 1, \ldots$$

by stacking up the variables x_i^k into the vector x^k, where the matrices $A^k \in \mathbb{R}^{n \times n}$ are stochastic, and the initial vector $x^0 \in \mathbb{R}^n$ is given.

Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. Proceedings of the IEEE, 106(5):953–976, 2018.

Notation

Each graph is associated with a nonnegative weight matrix $\boldsymbol{W} = [w_{ij}] \in \mathbb{R}^{n \times n}$

- Nonnegative

$$w_{ij} \text{ is non-zero only if } (j, i) \in \mathcal{E} \text{ or } i = j$$

- Doubly stochastic

$$\boldsymbol{W}\mathbb{1}_n = \boldsymbol{W}^T\mathbb{1}_n = \mathbb{1}_n$$

- Symmetric — for undirected graph
- Time-varying — for the dynamic pattern between iterations.

Network topology selection

Topology evaluation

- **Maximum graph degree** — communication cost
- **Consensus rate** — effectiveness to mix information

A densely-connected topology enables decentralized methods to converge faster but results in less efficient communication since each node needs to average with more neighbors.

Consensus rate

For weight matrices $\left\{\boldsymbol{W}^{(t)}\right\}_{t\geq 0} \subseteq \mathbb{R}^{n\times n}$, minimum nonnegative number β

$$\mathbb{E}\left[\left\|\boldsymbol{W}^{(t)}\boldsymbol{x} - \bar{x}\cdot \mathbb{1}_n\right\|^2\right] \leq \beta^2 \left\|\boldsymbol{x} - \bar{x}\cdot \mathbb{1}_n\right\|^2, \forall \boldsymbol{t} \geq 0$$

is the **consensus rate**, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\mathbb{1}_n \in \mathbb{R}^n$ is the all-ones vector.

Consensus rate

Denote

- $\boldsymbol{J} = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$
- $\boldsymbol{\Pi} = \boldsymbol{I} - \boldsymbol{J}$, where $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ is the identity matrix

Equivalently,

$$\mathbb{E}\left[\left\|\boldsymbol{\Pi}\boldsymbol{W}^{(t)}\boldsymbol{x}\right\|^2\right] \le \beta^2\|\boldsymbol{\Pi}\boldsymbol{x}\|^2$$

If $\boldsymbol{W}^{(t)} \equiv \boldsymbol{W}$, then

$$\beta \equiv \|\boldsymbol{\Pi}\boldsymbol{W}\|_2$$

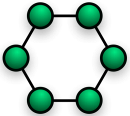
where $\|\boldsymbol{A}\|_2$ is the spectral norm (maximum singular value of $A^HA)^{\frac{1}{2}}$.

Commonly-used topologies

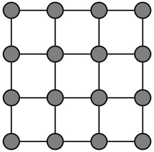
- Ring
- Grid
- Torus
- Hypercube
- Exponential graph
- Random graph

Commonly-used Topologies

- **Ring** — A ring network is a network topology in which each node connects to exactly two other nodes, forming a single continuous pathway for signals through each node — a ring.
  Rings can be unidirectional, either clockwise or anticlockwise around the ring, or bidirectional.

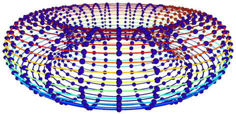


- **Grid** — A regular grid is a tessellation of n-dimensional Euclidean space by congruent parallelotopes (e.g. bricks).
  `https://en.wikipedia.org/wiki/Regular_grid`

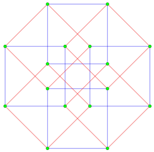


Commonly-used Topologies

- **Torus** — Topologically, a torus is a closed surface defined as the product of two circles.
  `https://en.wikipedia.org/wiki/Torus`



- **Hypercube** — In topology, a hypercube, also known as an n-cube or an n-dimensional cube, is a geometric shape that generalizes the concept of a square (a 2-dimensional cube) and a cube (a 3-dimensional cube) to higher dimensions
  `https://en.wikipedia.org/wiki/Hypercube`



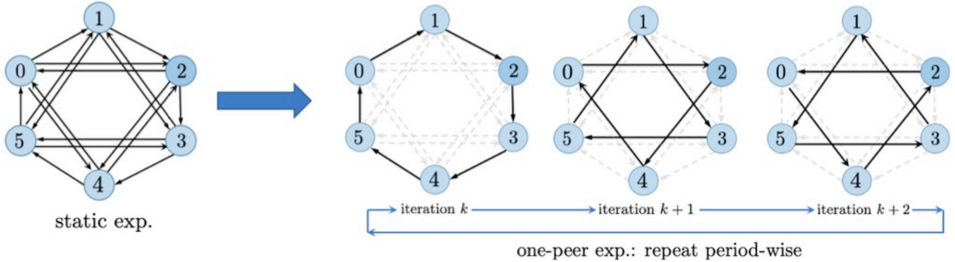
Grid Network

In a regular grid topology, each node in the network is connected with ***two neighbors*** *along one or more dimensions.*

- If the network is *one-dimensional*, and the chain of nodes is connected to form a circular loop, the resulting topology is known as a **ring**.
- In general, when an *n-dimensional* grid network is connected circularly in more than one dimension, the resulting network topology is a **torus**, and the network is called "**toroidal**".
- When *the number of nodes* along each dimension of a toroidal network *is 2*, the resulting network is called a **hypercube**.

`https://en.wikipedia.org/wiki/Grid_network`

Exponential graph



- "Static Exp." — static exponential graph
  each node communicates to $\lceil \log_2(n) \rceil$ neighbors.
- "O.-P. Exp." — one-peer exponential graph
  each node cycles through all its neighbors, communicating only to a single neighbor per iteration.

Random graph

"E.-R. Rand": *Erdos-Renyi random graph $G(n,p)$*

1. A symmetric adjacency matrix A whose $\binom{n}{2}$ distinct off-diagonal entries are independent Bernoulli random variables taking the value 1 with probability p.
2. Here $p = (1+a)\log(n)/n$ and $a > 0$, for which it is known that the random graph is connected with high probability.

"Geo. Rand": *Geometric random graph G(n, r)*

1. n nodes are placed uniformly and independently in the unit square $[0,1]^2$ and two nodes are connected with an edge if their distance is at most r_n.
2. Here $r_n^2 = (1+a)\log(n)/n$ for some $a > 0$, for which it is known that the random graph is connected with high probability.

Comparison between different commonly-used topologies

Topology	Connection	Pattern	Degree	Consensus Rate	size n
Ring	undirect.	static	$\Theta(1)$	$1-\Theta(1/n^2)$	arbitrary
Grid	undirect.	static	$\Theta(1)$	$1-\Theta(1/(n\ln(n)))$	arbitrary
Torus	undirect.	static	$\Theta(1)$	$1-\Theta(1/n)$	arbitrary
Hypercube	undirect.	static	$\Theta(\ln(n))$	$1-\Theta(1/\ln(n))$	power of 2
Static Exp.	directed	static	$\Theta(\ln(n))$	$1-\Theta(1/\ln(n))$	arbitrary
O.-P. Exp.	directed	dynamic	1	finite-time conv.†	power of 2
E.-R. Rand	undirect.	static	$\Theta(\ln(n))^\diamond$	$\Theta(1)$	arbitrary
Geo. Rand	undirect.	static	$\Theta(\ln(n))$	$1-\Theta(\ln(n)/n)$	arbitrary
D-EquiStatic	directed	static	$\Theta(\ln(n))$	$\rho \in (0,1)^\ddagger$	arbitrary
U-EquiStatic	undirect.	static	$\Theta(\ln(n))$	$\rho \in (0,1)^\ddagger$	arbitrary
OD-EquiDyn	directed	dynamic	1	$\sqrt{(1+\rho)/2}$	arbitrary
OU-EquiDyn	undirect.	dynamic	1	$\sqrt{(2+\rho)/3}$	arbitrary

† One-peer exponential graph has finite-time exact convergence only when n is the power of 2.

$^\diamond$ $\Theta(\ln(n))$ is the averaged degree; its maximum degree can be $O(n)$ with a non-zero probability.

‡ Constant $\rho = \Theta(1)$ is independent of network-size n.

Existing topologies are not good enough.

Develop topology — EquiTopo

EquiTopo:

- **Directed EquiTopo Graphs**
- **Undirected EquiTopo Graphs**

Pros:

1. network-size-independent consensus rate
2. (almost) constant graph degrees

Develop topology — EquiTopo

Directed EquiTopo Graphs

1. Directed static EquiTopo graphs (**D-EquiStatic**)
2. One-peer directed EquiTopo graphs (**OD-EquiDyn**)

Directed EquiTopo Graphs

- Mod operation — returns a value in $[n] = \{1, \cdots, n\}$

$$i \bmod n = \begin{cases} \ell & \text{if } i = kn + \ell \text{ for some } k \in \mathbb{Z} \text{ and } \ell \in [n-1] \\ n & \text{if } i = kn \text{ for some } k \in \mathbb{Z} \end{cases} \tag{1}$$

- Doubly stochastic basis matrix $\boldsymbol{A}^{(u,n)} = \left[a_{ij}^{(u,n)}\right] \in \mathbb{R}^{n\times n}$

$$a_{ij}^{(u,n)} = \begin{cases} \frac{n-1}{n}, & \text{if } i = (j+u) \bmod n \\ \frac{1}{n}, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

- Basis weight graphs $\mathcal{G}\left(\boldsymbol{A}^{(u,n)}\right)$
  1. Degree one
  2. Same label difference $(i-j) \bmod n$ for all edges (j, i).

Directed EquiTopo Graphs

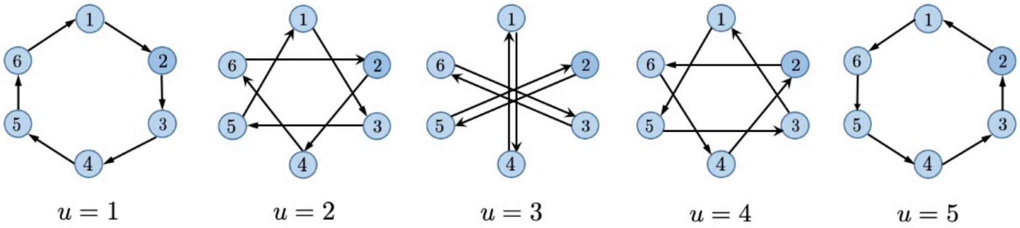


Figure 1: The set of five basis graphs $\left\{\mathcal{G}\left(\boldsymbol{A}^{(u,6)}\right)\right\}_{u=1}^{5}$ for $n=6$

Given a graph of size n, the basis graphs are $\left\{\mathcal{G}\left(\boldsymbol{A}^{(u,n)}\right)\right\}_{u=1}^{n-1}$.

Since n is clear from the context, we omit it and write $\boldsymbol{A}^{(u)}$ *instead.*

D-EquiStatic

Directed static EquiTopo graphs (D-EquiStatic)

- Weight matrix
$$\boldsymbol{W} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{A}^{(u_i)} \tag{3}$$
where $u_i \in [n-1]$ and $M > 0$ is the number of basis graphs we will sample.
- $\boldsymbol{W}$ is doubly stochastic
- All nodes of the directed graph $\mathcal{G}(\boldsymbol{W})$ have the same degree that is no more than M.

D-EquiStatic

Theorem 1 Let $A^{(u)}$ be defined by (2) for any $u \in [n-1]$. For any constant $\rho \in (0,1)$, we can choose a sequence of $u_1, \cdots, u_M$ from $[n-1]$ with $M = \Theta\left(\ln(n)/\rho^2\right)$ and construct the D-EquiStatic weight matrix $\boldsymbol{W}$ as in (3) such that **the consensus rate of $\boldsymbol{W}$ is** ρ, i.e.,

$$\|\boldsymbol{\Pi}\boldsymbol{W}\boldsymbol{x}\| \leq \rho\|\boldsymbol{\Pi}\boldsymbol{x}\|, \forall \boldsymbol{x} \in \mathbb{R}^n \tag{4}$$

Here,

- Graph $\mathcal{G}(\boldsymbol{W})$ has degree at most M , so we just say that **the degree is** $\Theta(\ln(n))$.
- Degree can be easily predefined by specifying M.
- Consensus rate is independent of the network size n.
- ρ is tunable, and is chosen to be a constant, e.g., $\rho = 0.5$.

One-peer directed EquiTopo graphs (OD-EquiDyn)

Algorithm 1: OD-EquiDyn weight matrix generation at iteration t

Input: constant $\eta \in (0,1)$; basis index $\{u_1, u_2, \ldots, u_M\}$ from a weight matrix $\boldsymbol{W}$ of form $(\boxed{3})$;
Pick v_t from uniform distribution over the basis index $\{u_1, u_2, \ldots, u_M\}$;
Produce basis matrix $\boldsymbol{A}^{(v_t)}$ according to $(\boxed{2})$;
Output: $\boldsymbol{W}^{(t)} = (1-\eta)\boldsymbol{I} + \eta \boldsymbol{A}^{(v_t)}$

- One peer — the degree for each node is at most 1
- To further reduce the degree to one

OD-EquiDyn

Theorem 2 Let the one-peer directed weight matrix $W^{(t)}$ be generated by *Alg*.1 It holds that

$$\mathbb{E}\left[\left\|\boldsymbol{\Pi}\boldsymbol{W}^{(t)}\boldsymbol{x}\right\|^2\right] \le (1-2\eta(1-\eta)(1-\rho))\|\boldsymbol{\Pi}\boldsymbol{x}\|^2, \quad \forall \boldsymbol{x}\in\mathbb{R}^n$$

where ρ is the consensus rate of the weight matrix $\boldsymbol{W}$ (which can be tuned freely as in *Theorem 1*) .

- Let $\eta = 1/2$, then it holds that

$$\mathbb{E}\left\|\boldsymbol{\Pi}\boldsymbol{W}^{(t)}\boldsymbol{x}\right\|^2 \le (1+\rho)/2\|\boldsymbol{\Pi}\boldsymbol{x}\|^2$$

- Let $\boldsymbol{W}=\boldsymbol{J}$, then $\|\boldsymbol{\Pi}\boldsymbol{J}\boldsymbol{x}\| = 0$, which implies $\rho = 0$, so

$$\mathbb{E}\left\|\boldsymbol{\Pi}\boldsymbol{W}^{(t)}\boldsymbol{x}\right\|^2 \le \frac{1}{2}\|\boldsymbol{\Pi}\boldsymbol{x}\|^2$$

Develop topology — EquiTopo

Undirected EquiTopo Graphs

1. Undirected static EquiTopo graphs (**U-EquiStatic**)
2. One-peer undirected EquiTopo graphs (**OU-EquiDyn**)

Undirected EquiTopo Graphs

Undirected static EquiTopo graphs (U-EquiStatic)

- Weight matrix

$$\widetilde{\boldsymbol{W}} = \frac{1}{2}\left(\boldsymbol{W} + \boldsymbol{W}^T\right) = \frac{1}{2M}\sum_{i=1}^{M}\left(\boldsymbol{A}^{(u_i)} + \left[\boldsymbol{A}^{(u_i)}\right]^T\right) \tag{5}$$

- Basis index are $\{u_i, -u_i\}_{i=1}^{M}$ because $\boldsymbol{A}^{(-u)} = \left[\boldsymbol{A}^{(u)}\right]^T$

U-EquiStatic

Theorem 3 Let W be a D-EquiStatic matrix with consensus rate ρ and $\widetilde{W}$ be the U-EquiStatic matrix defined by (5). It holds that

$$\|\mathbf{\Pi}\widetilde{\boldsymbol{W}}\boldsymbol{x}\| \leq \rho\|\mathbf{\Pi}\boldsymbol{x}\|, \forall \boldsymbol{x} \in \mathbb{R}^n \tag{6}$$

One-peer undirected EquiTopo graphs (OU-EquiDyn)

Algorithm 2: OU-EquiDyn weight matrix generation at iteration t

---

Input: $\eta \in (0,1)$; basis index $\{u_i, -u_i\}_{i=1}^{M}$ from a symmetric weight matrix $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{n\times n}$ of form (5);

Pick $v_t \in \{u_i, -u_i\}_{i=1}^{M}$ and $s_t \in [n]$ uniformly at random;

Initialize $\boldsymbol{A} = [a_{ij}] = \boldsymbol{I}$ and $b_i = 0, \forall i \in [n]$;

for $j = (s_t : s_t + n - 1 \mod n)$ **do**

  $i = (j + v_t) \mod n$;

  if $b_i = 0$ *and* $b_j = 0$ **then**

    $a_{ij} = a_{ji} = (n-1)/n$;

    $a_{ii} = a_{jj} = 1/n$;

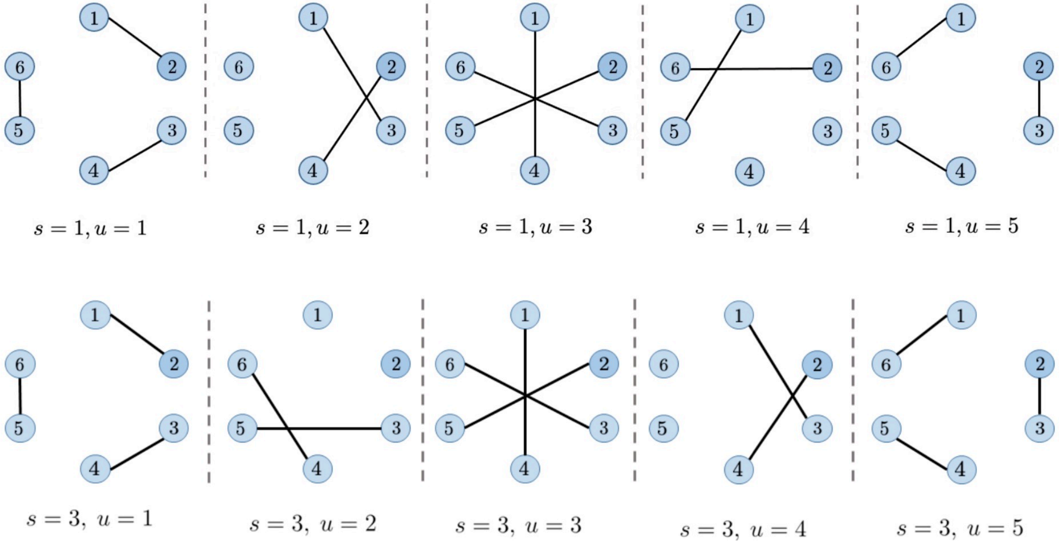
    $b_i = 1, b_j = 1$;

  end

end

Output: $\widetilde{\boldsymbol{W}}^{(t)} = (1-\eta)I + \eta\boldsymbol{A}$

OU-EquiDyn



OU-EquiDyn

Theorem 4 Let $\widetilde{\boldsymbol{W}}$ be a U-EquiStatic matrix with consensus rate ρ, and $\widetilde{\boldsymbol{W}}^{(t)}$ be an OU-EquiDyn matrix generated by Alg. 2, it holds that

$$\mathbb{E}\left[\left\|\boldsymbol{\Pi}\widetilde{\boldsymbol{W}}^{(t)}\boldsymbol{x}\right\|^2\right] \le \left(1-\frac{4}{3}\eta(1-\eta)(1-\rho)\right)\|\boldsymbol{\Pi}\boldsymbol{x}\|^2, \quad \forall \boldsymbol{x}\in\mathbb{R}^n$$

- When $\eta = 1/2$, it holds that

$$\mathbb{E}\left\|\boldsymbol{\Pi}\boldsymbol{W}^{(t)}\boldsymbol{x}\right\|^2 \le [(2+\rho)/3]\|\boldsymbol{\Pi}\boldsymbol{x}\|^2.$$

- When $\widetilde{W} = \boldsymbol{J}$ and the basis index $\{1,\cdots,n-1\}$ are input to Alg. 2, we obtain an OU-EquiDyn sequence $\widetilde{\boldsymbol{W}}^{(t)}$ such that

$$\mathbb{E}\left\|\boldsymbol{\Pi}\widetilde{\boldsymbol{W}}^{(t)}\boldsymbol{x}\right\|^2 \le (2/3)\|\boldsymbol{\Pi}\boldsymbol{x}\|^2.$$

Comparison between different commonly-used topologies

Topology	Connection	Pattern	Degree	Consensus Rate	size n
Ring	undirect.	static	$\Theta(1)$	$1-\Theta(1/n^2)$	arbitrary
Grid	undirect.	static	$\Theta(1)$	$1-\Theta(1/(n\ln(n)))$	arbitrary
Torus	undirect.	static	$\Theta(1)$	$1-\Theta(1/n)$	arbitrary
Hypercube	undirect.	static	$\Theta(\ln(n))$	$1-\Theta(1/\ln(n))$	power of 2
Static Exp.	directed	static	$\Theta(\ln(n))$	$1-\Theta(1/\ln(n))$	arbitrary
O.-P. Exp.	directed	dynamic	1	finite-time conv.†	power of 2
E.-R. Rand	undirect.	static	$\Theta(\ln(n))^\diamond$	$\Theta(1)$	arbitrary
Geo. Rand	undirect.	static	$\Theta(\ln(n))$	$1-\Theta(\ln(n)/n)$	arbitrary
D-EquiStatic	directed	static	$\Theta(\ln(n))$	$\rho \in (0,1)^\ddagger$	arbitrary
U-EquiStatic	undirect.	static	$\Theta(\ln(n))$	$\rho \in (0,1)^\ddagger$	arbitrary
OD-EquiDyn	directed	dynamic	1	$\sqrt{(1+\rho)/2}$	arbitrary
OU-EquiDyn	undirect.	dynamic	1	$\sqrt{(2+\rho)/3}$	arbitrary

† One-peer exponential graph has finite-time exact convergence only when n is the power of 2.
$^\diamond$ $\Theta(\ln(n))$ is the averaged degree; its maximum degree can be $O(n)$ with a non-zero probability.
‡ Constant $\rho = \Theta(1)$ is independent of network-size n.

Remark EquiTopo family (especially the one-peer variants) has achieved the best balance between maximum graph degree and consensus rate.

Apply EquiTopo to decentralized learning

Two well-known decentralized algorithms

- Decentralized stochastic gradient descent —**DSGD**
- Decentralized stochastic gradient tracking algorithm —**DSGT**

Assumptions

- **A.1** Each local cost function $f_i(x)$ is differentiable, and there exists a constant $L > 0$ such that $\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.
- **A.2** Let $\boldsymbol{g}_i^{(t)} = \nabla F\left(\boldsymbol{x}_i^{(t)}; \xi_i^{(t)}\right)$. There exists $\sigma^2 > 0$ such that for any t and i

$$\mathbb{E}_{\xi_i^{(t)} \sim \mathcal{D}_i} \boldsymbol{g}_i^{(t)} = \nabla f_i\left(\boldsymbol{x}_i^{(t)}\right), \text{ and } \mathbb{E}_{\xi_i^{(t)} \sim \mathcal{D}_i}\left[\left\|\boldsymbol{g}_i^{(t)} - \nabla f_i\left(\boldsymbol{x}_i^{(t)}\right)\right\|^2\right] \leq \sigma^2.$$

- **A.3** (For DSGD only) There exists b^2 such that $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\| \leq b^2$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Decentralized stochastic gradient descent — DSGD

Algorithm 1 DECENTRALIZED SGD (DSGD)

input for each node $i \in [n]$ initialize $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$,
stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations T,
mixing matrix distributions $\mathcal{W}^{(t)}$ for $t \in [0, T]$

for t **in** $0 \ldots T$ **do**
  Sample $W^{(t)} \sim \mathcal{W}^{(t)}$
  In parallel (task for worker $i, i \in [n]$)
  Sample $\xi_i^{(t)}$, compute $\mathbf{g}_i^{(t)} := \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$
  $\mathbf{x}_i^{(t+\frac{1}{2})} = \mathbf{x}_i^{(t)} - \eta_t \mathbf{g}_i^{(t)}$ ▷ stochastic gradient updates
  $\mathbf{x}_i^{(t+1)} := \sum_{j \in \mathcal{N}_i^t} w_{ij}^{(t)} \mathbf{x}_j^{(t+\frac{1}{2})}$ ▷ gossip averaging
end for

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized sgd with changing topology and local updates. In International Conference on Machine Learning (ICML), pages 1–12, 2020.

Decentralized stochastic gradient descent — DSGD

The **d**ecentralized **s**tochastic **g**radient **d**escent (DSGD) is given by

$$\boldsymbol{x}_i^{(t+1)} = \sum_{j=1}^{n} w_{ij}^{(t)} \left(\boldsymbol{x}_j^{(t)} - \gamma \boldsymbol{g}_j^{(t)} \right) \tag{8}$$

where the weight matrix $\boldsymbol{W}^{(t)} = \left[w_{ij}^{(t)} \right]$ can be time-varying and random.

Decentralized stochastic gradient descent — DSGD

Theorem 5 Consider the DSGD algorithm (8). Under Assumptions A.1-A.3, it holds that

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla f\left(\bar{\boldsymbol{x}}^{(t)}\right)\right\|^2\right] = \mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \frac{\beta^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}} + \frac{\beta^{\frac{2}{3}}\boldsymbol{b}^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{2}{3}}} + \frac{\beta}{T(1-\beta)}\right)$$

where error $\frac{1}{(T+1)}\sum_{t=0}^{T}\left(\mathbb{E}f\left(\bar{\mathbf{x}}^{(t)}\right) - f^{\star}\right) \leq \epsilon$

- $\beta = \rho$ with D-EquiStatic $\boldsymbol{W}$ or U -EquiStatic $\widetilde{\boldsymbol{W}}$
- $\beta = \sqrt{(1+\rho)/2}$ for OD-EquiDyn $\boldsymbol{W}^{(t)}$
- $\beta = \sqrt{(2+\rho)/3}$ for OU-EquiDyn $\widetilde{W}^{(t)}$

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized sgd with changing topology and local updates. In International Conference on Machine Learning (ICML), pages 1–12, 2020.

Decentralized stochastic gradient descent — DSGD

Theorem 5 Consider the DSGD algorithm (8). Under Assumptions A.1-A.3, it holds that

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla f\left(\bar{\boldsymbol{x}}^{(t)}\right)\right\|^2\right] = \mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \frac{\beta^{\frac{2}{3}}\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{1}{3}}} + \frac{\beta^{\frac{2}{3}}\boldsymbol{b}^{\frac{2}{3}}}{T^{\frac{2}{3}}(1-\beta)^{\frac{2}{3}}} + \frac{\beta}{T(1-\beta)}\right)$$

- For *a sufficiently large* T, the term $\mathcal{O}(1/\sqrt{nT})$ dominates the rate, and we say the algorithm *reaches the linear speedup stage*.
- The *transient iterations* are referred to as those iterations *before an algorithm reaches the linear-speedup stage*.
- The smaller the transient iteration complexity is, the faster the algorithm converges.

Decentralized stochastic gradient descent — DSGD

For **strongly convex** cost functions

Topology	Per-iter Comm.	Convergence Rate	Trans. Iters.
Ring	$\Theta(1)$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{nT}+\frac{\kappa n^2\sigma^2}{T^2}+\frac{\kappa n^4 b^2}{T^2}\right)$	$\tilde{\mathcal{O}}(\kappa n^5)$
Torus	$\Theta(1)$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{nT}+\frac{\kappa n\sigma^2}{T^2}+\frac{\kappa n^2 b^2}{T^2}\right)$	$\tilde{\mathcal{O}}(\kappa n^3)$
Static Exp.	$\Theta(\ln(n))$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{nT}+\frac{\kappa \ln(n)\sigma^2}{T^2}+\frac{\kappa \ln^2(n) b^2}{T^2}\right)$	$\tilde{\mathcal{O}}(\kappa n\ln^2(n))$
O.-P. Exp.	1	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{nT}+\frac{\kappa \ln(n)\sigma^2}{T^2}+\frac{\kappa \ln^2(n) b^2}{T^2}\right)$	$\tilde{\mathcal{O}}(\kappa n\ln^2(n))$
D(U)-EquiStatic	$\Theta(\ln(n))$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{nT}+\frac{\kappa\sigma^2}{T^2}+\frac{\kappa b^2}{T^2}\right)$	$\tilde{\mathcal{O}}(\kappa n)$
OD (OU)-EquiDyn	1	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{nT}+\frac{\kappa\sigma^2}{T^2}+\frac{\kappa b^2}{T^2}\right)$	$\tilde{\mathcal{O}}(\kappa n)$

It is observed that OD/OU-EquiDyn endows DSGD with the lightest communication, fastest convergence rate, and smallest transient iteration complexity.

Decentralized stochastic gradient tracking algorithm — DSGT

Algorithm 1 GRADIENT TRACKING (DSGT)

input Initial values $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$ on each node $i \in [n]$, communication graph $G = ([n], E)$ and mixing matrix W, stepsize γ, initialize $\mathbf{y}_i^{(0)} = \nabla F_i(\mathbf{x}_i^{(0)}, \xi_i^{(0)})$, $\mathbf{g}_i^{(0)} = \mathbf{y}_i^{(0)}$ in parallel for $i \in [n]$.

1: **in parallel on all workers** $i \in [n]$**, for** $t = 0, \dots, T-1$ **do**

2: each node i sends $\left(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}\right)$ to is neighbors

3: $\mathbf{x}_i^{(t+1)} = \sum_{j:\{i,j\}\in E} w_{ij} \left(\mathbf{x}_j^{(t)} - \gamma \mathbf{y}_j^{(t)}\right)$ ▷ update model parameters

4: Sample $\xi_i^{(t+1)}$, compute gradient $\mathbf{g}_i^{(t+1)} = \nabla F_i\left(\mathbf{x}_i^{(t+1)}, \xi_i^{(t+1)}\right)$

5: $\mathbf{y}_i^{(t+1)} = \sum_{j:\{i,j\}\in E} w_{ij} \mathbf{y}_j^{(t)} + \left(\mathbf{g}_i^{(t+1)} - \mathbf{g}_i^{(t)}\right)$ ▷ update tracking variable

6: **end parallel for**

Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. Advances in Neural Information Processing Systems, 34, 2021.

Decentralized stochastic gradient tracking algorithm — DSGT

- Each agent i keeps an estimate of the minimizer $\boldsymbol{x}_i(t) \in \mathbb{R}^{1\times N}$
- $\boldsymbol{y}^{(t)} \in \mathbb{R}^{1\times N}$ to estimate the average gradient $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i\left(\boldsymbol{x}_i(t)\right)$.

$$
\begin{aligned}
\boldsymbol{x}_i^{(t+1)} &= \textstyle\sum_{j=1}^{n} w_{ij}^{(t)} \left(\boldsymbol{x}_j^{(t)} - \gamma \boldsymbol{y}_j^{(t)}\right) \\
\boldsymbol{y}_i^{(t+1)} &= \textstyle\sum_{j=1}^{n} w_{ij}^{(t)} \boldsymbol{y}_j^{(t)} + \boldsymbol{g}_i^{(t+1)} - \boldsymbol{g}_i^{(t)}, \boldsymbol{y}_i^{(0)} = \boldsymbol{g}_i^{(0)}.
\end{aligned}
\tag{9}
$$

Improved convergence rate over *asymmetric* or *time-varying* weight matrices.

G.QuandN.Li, "Harnessing smoothness to accelerate distributed optimization," IEEETrans- actions on Control of Network Systems, vol. 5, pp. 1245–1260, Sept. 2018.

Decentralized stochastic gradient tracking algorithm — DSGT

Theorem 6 Consider the DSGT algorithm in (9). If $\left\{\boldsymbol{W}^{(t)}\right\}_{t\geq 0}$ have consensus rate β , then under Assumptions A.1-A.2, it holds for $T \geq \frac{1}{1-\beta}$ that

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla f\left(\overline{\boldsymbol{x}}^{(t)}\right)\right\|^2\right] = \mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}}{(1-\beta)T^{\frac{2}{3}}} + \frac{1}{(1-\beta)^2 T}\right).$$

When utilizing the EquiTopo matrices, the corresponding β is specified in Theorem 5.

DSGT achieves linear speedup for large T.

Decentralized stochastic gradient tracking algorithm — DSGT

Topology	Per-iter Comm.	Convergence Rate	Trans. Iters.
Ring	$\Theta(1)$	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \frac{n^2\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{n^4}{T}\right)$	$\mathcal{O}(n^{15})$
Torus	$\Theta(1)$	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \frac{n\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{n^2}{T}\right)$	$\mathcal{O}(n^9)$
Static Exp.	$\Theta(\ln(n))$	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \frac{\ln(n)\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\ln^2(n)}{T}\right)$	$\mathcal{O}(n^3\ln^6(n))$
O.-P. Exp.	1	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \frac{\ln(n)\sigma^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\ln^2(n)}{T}\right)$	$\mathcal{O}(n^3\ln^6(n))$
D(U)-EquiStatic	$\Theta(\ln(n))$	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \left(\frac{\sigma}{T}\right)^{\frac{2}{3}} + \frac{1}{T}\right)$	$\mathcal{O}(n^3)$
OD (OU)-EquiDyn	1	$\mathcal{O}\left(\frac{\sigma}{\sqrt{nT}} + \left(\frac{\sigma}{T}\right)^{\frac{2}{3}} + \frac{1}{T}\right)$	$\mathcal{O}(n^3)$

OD/OU-EquiDyn endows DSGT with the lightest communication, fastest convergence rate, and smallest transient iteration complexity.

Experiments

1. Consensus rate
   - Network-size independent consensus rate.
   - Comparison with other topologies.
2. DSGD with EquiTopo
   - Distributed least-square problem
   - Distributed deep learning

Experiments — Network-size independent consensus rate

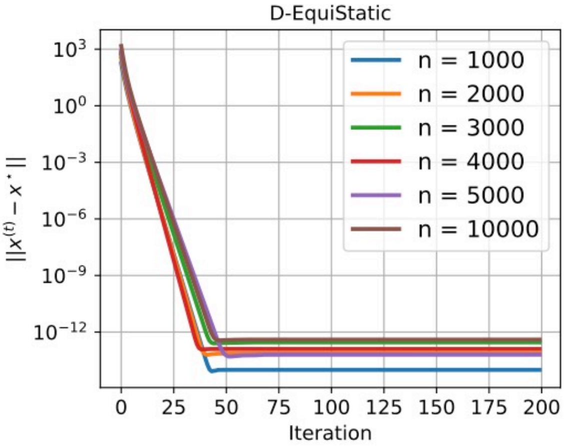


Figure 2: The D-EquiStatic topology can achieve network-size independent consensus rate

Experiments — Comparison with other topologies

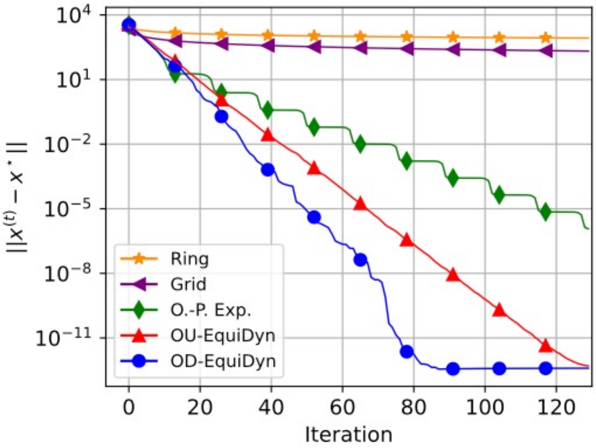


Figure 3: OD/OU-EquiDyn is faster than other topologies (i.e., ring, grid, and one-peer exponen- tial graph) with $\mathbb{O}(1)$ degree in consensus rate.

Experiments — DSGD with EquiTopo on Least-square

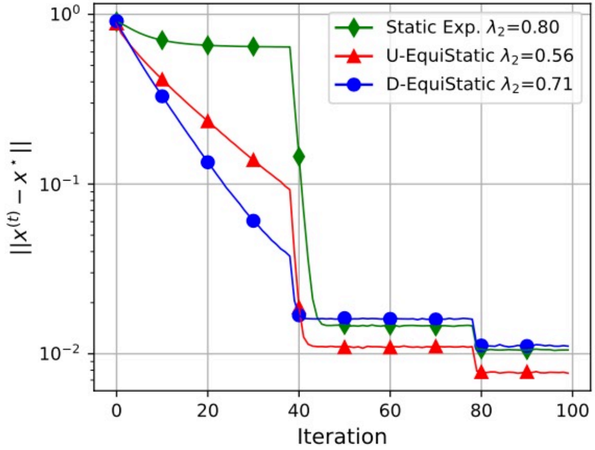


Figure 4: D/U-EquiStatic in DSGD. λ_2 is the second largest eigenvalue.

Experiments — DSGD with EquiTopo on Deep Learning

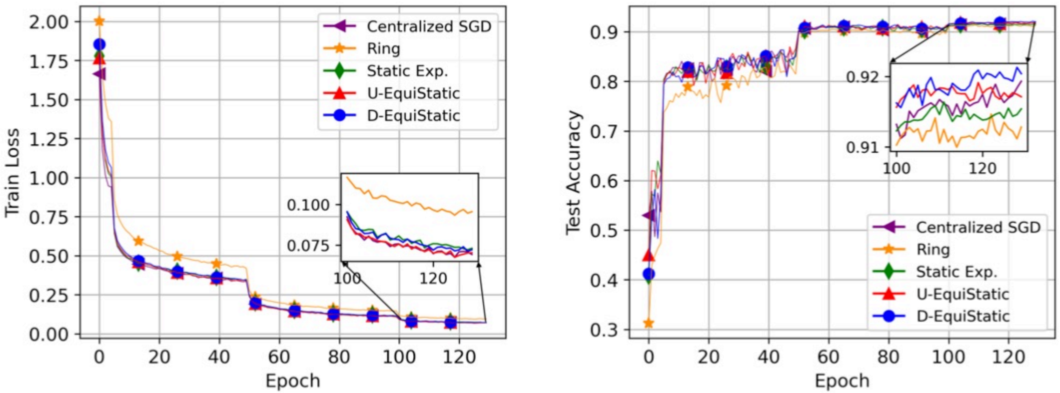


Figure 5: Train loss and test accuracy comparisons among different topologies for ResNet-20 on CIFAR-10.

Conclusion

- This paper develops several novel graphs built upon a set of basis graphs in which the label difference between any pair of connected nodes are equivalent.
- With a general name EquiTopo, these new graphs can achieve network-size-independent consensus rates while maintaining (almost) constant graph degrees.

Thanks!