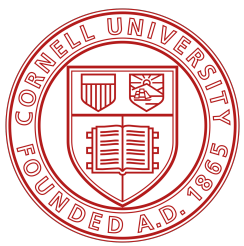
Distributed Optimization for Machine Learning

Lecture 16 - Decentralized SGD and Gradient tracking

Tianyi Chen

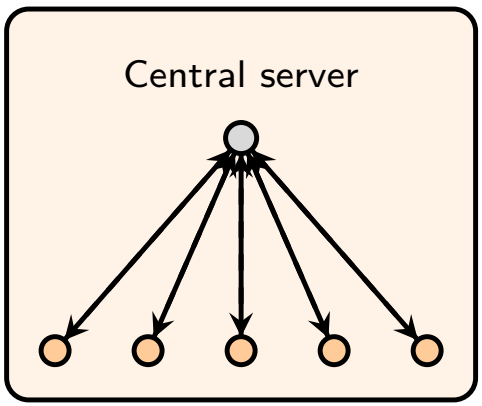
School of Electrical and Computer Engineering
Cornell Tech, Cornell University

October 22, 2025



Review of Parallel SGD: synchronize every time

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

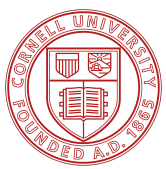


Local data on nodes

$$g_i^k = \nabla F(\boldsymbol{x}^k;\, \xi_i^k) \qquad \text{(Local compt.)}$$
$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \frac{\eta}{n}\sum_{i=1}^{n} g_i^k \qquad \text{(Global comm.)}$$

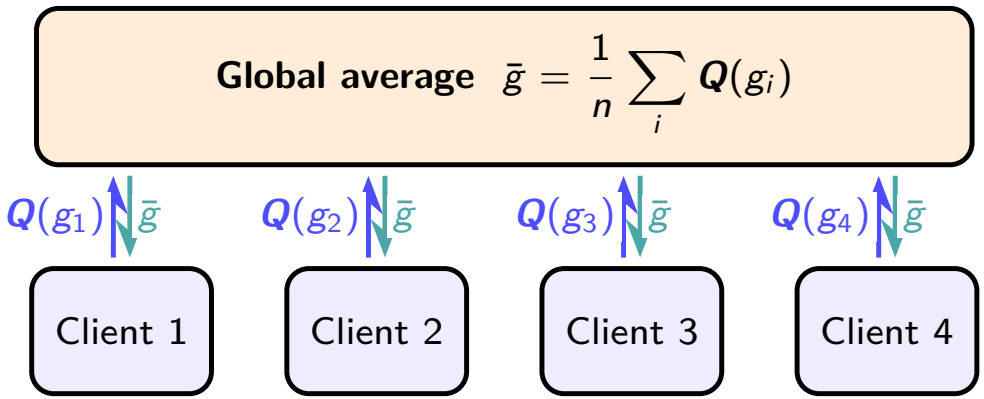
- All nodes synchronize (i.e., *globally average*) every time.
- Global average protocol: Ring-AllReduce or central server

Communication cost is $\mathcal{O}(n)$ which is high when n is large.



Last-lecture: Compressed SGD

Compressed SGD: Each node i communicates its compressed gradient.

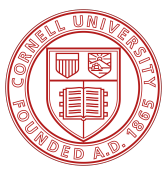


$$g_i^k = \nabla F(x^k; \xi_i^k) \qquad \text{(Local compt.)}$$

$$x^{k+1} = x^k - \frac{\eta}{n}\sum_{i=1}^{n} Q(g_i^k) \qquad \text{(Global comm.)}$$

- Compressed methods: quantization, sparsification

Reduce the communication cost by reducing the dimensionality.

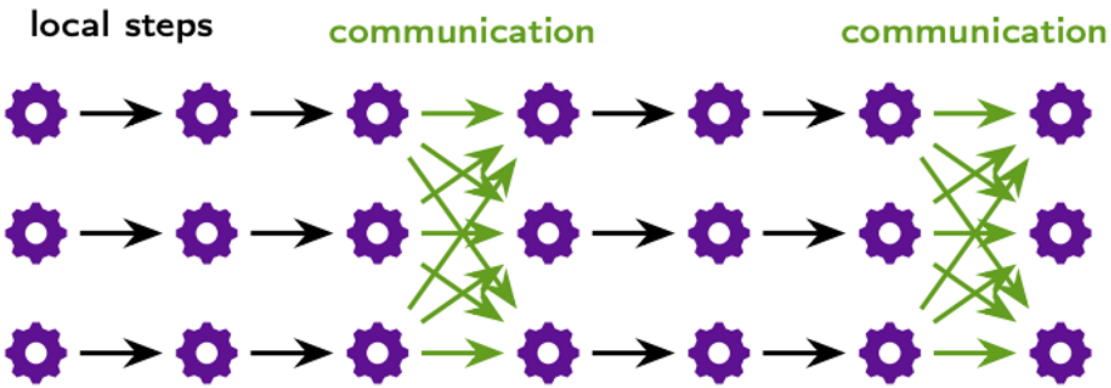


Last-lecture: Local SGD

Local SGD: Each node i performs τ-step SGD before averaging.

$$\boldsymbol{x}_i^{(s+1)} = \boldsymbol{x}_i^{(s)} - \eta \nabla F\big(\boldsymbol{x}_i^{(s)}; \boldsymbol{\xi}_i^{(s)}\big), \ s = 1, \cdots, \tau \quad \text{(Local updates)}$$

$$\boldsymbol{x}^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i^{(\tau)} \quad \text{(Global comm.)}$$



Reduce the temporal communication cost by τ.

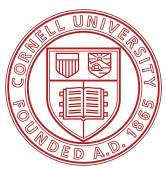
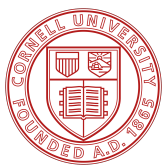
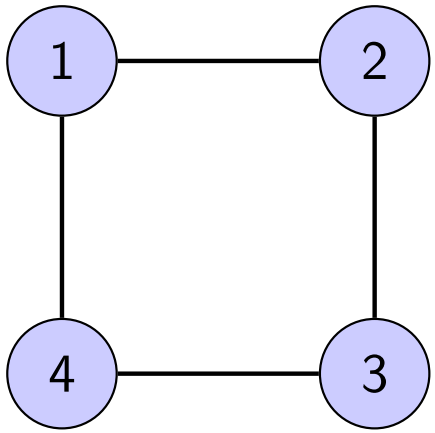


Table of Contents



Global averaging protocol: averaging over a graph

- **Setup:** A network of n nodes connected by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
- **Goal:** Each node i has a copy of the average $\frac{1}{n}\sum_{i=1}^{n} x_i(0)$.



- **Node set $\mathcal{V}$:**

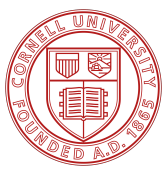
$$\mathcal{V} = \{1, 2, 3, 4\}$$

- **Edge set $\mathcal{E}$:**

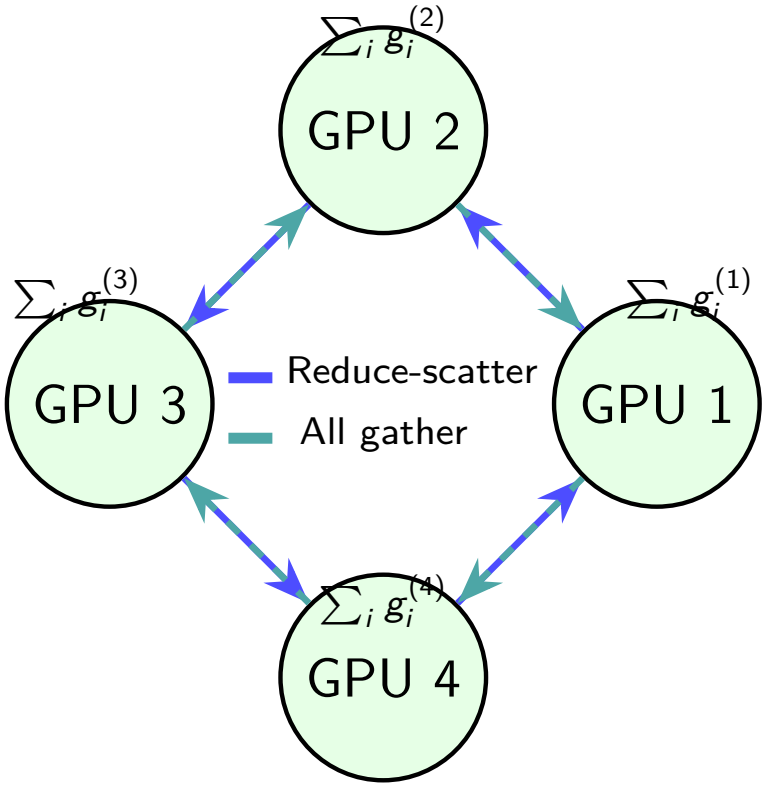
$$\mathcal{E} = \{(1, 2), (1, 4), (2, 3), (3, 4)\}$$

- **Average consensus protocol**:

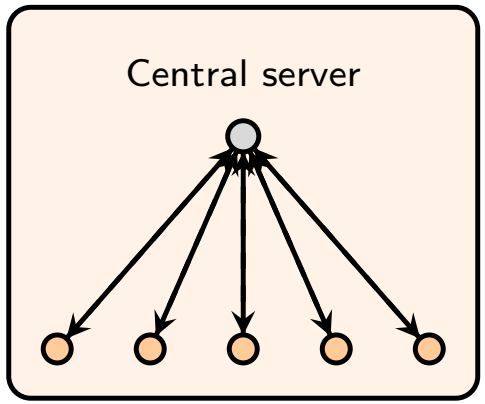
$$x_i(k+1) = \sum_{\{j:(i,j)\in\mathcal{E}\}} w_{ij} x_j(k)$$



Global averaging protocol: averaging over a graph



Ring-AllReduce: $\mathcal{O}(n)$ latency



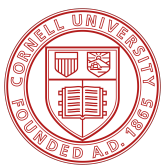
Local data on nodes

Central server: $\mathcal{O}(n)$ comm. cost.

Both global averaging protocols require $\mathcal{O}(n)$ complexity.

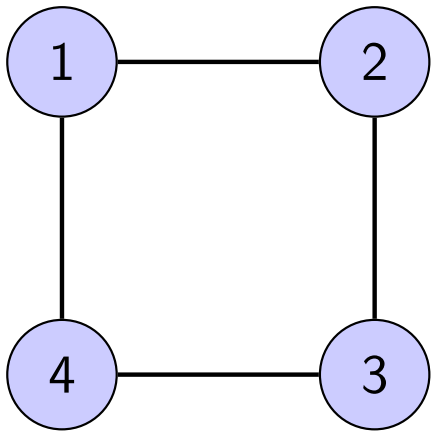
Moving beyond global averaging

Can we leverage **Gossip** to reduce the spatial communication cost?



Neighbor set

- **Setup:** A network of n nodes connected by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.



- **Node set $\mathcal{V}$:**

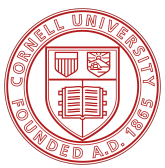
$$\mathcal{V} = \{1, 2, 3, 4\}$$

- **Edge set $\mathcal{E}$:**

$$\mathcal{E} = \{(1, 2), (1, 4), (2, 3), (3, 4)\}$$

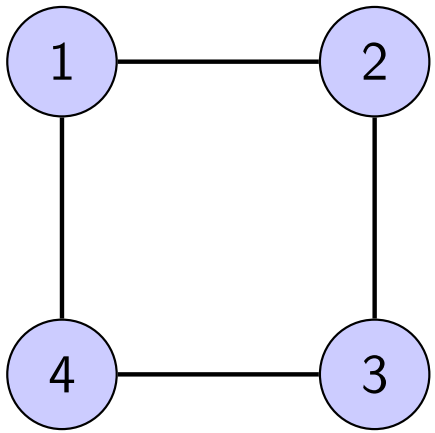
- **Neighbor set $\mathcal{N}_i$ for node i:**

$$\mathcal{N}_i = i \cup \{j : (i, j) \in \mathcal{E}\}$$



Neighbor set

- **Setup:** A network of n nodes connected by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.



- **Neighbor set $\mathcal{N}_1$ for node** 1:

$$\mathcal{N}_1 = \{1, 2, 4\}$$

- **Neighbor set $\mathcal{N}_2$ for node** 2:

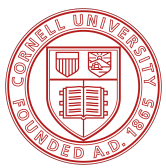
$$\mathcal{N}_2 = \{1, 2, 3\}$$

- **Neighbor set $\mathcal{N}_3$ for node** 3:

$$\mathcal{N}_3 = \{2, 3, 4\}$$

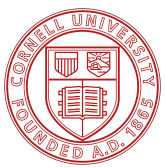
- **Neighbor set $\mathcal{N}_4$ for node** 4:

$$\mathcal{N}_4 = \{1, 3, 4\}$$



Key idea of decentralized SGD

Instead of global averaging, partial averaging (gossip) with neighbors



Example of decentralized SGD on the ring graph

- Consider the 4-node ring graph given above.

Decentralized SGD on the 4-node ring graph:

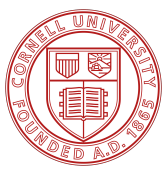
1. At time k, each node i runs the local SGD:

$$\boldsymbol{x}_i^{(k+\frac{1}{2})} = \boldsymbol{x}_i^{(k)} - \eta \nabla F(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k)})$$

2. At time k, every node i sent its current state to its neighbors $j \in \mathcal{N}_i$.
3. Every node i update its state by partial averaging with its neighbors:

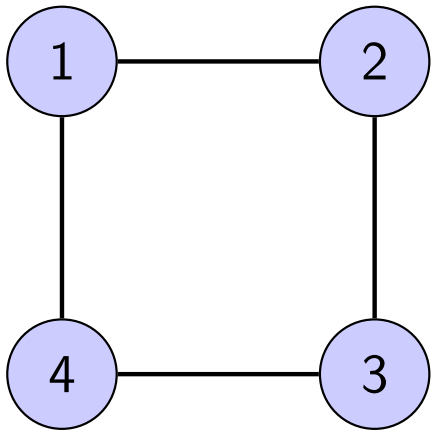
$$\text{For } \{l, m, i\} \in \mathcal{N}_i: \quad x_i^{(k+1)} = \frac{1}{3} x_i^{(k+\frac{1}{2})} + \frac{1}{3} x_l^{(k+\frac{1}{2})} + \frac{1}{3} x_m^{(k+\frac{1}{2})}$$

- Reduce the per-iteration comm. cost of PSGD from $\mathcal{O}(n)$ to $\mathcal{O}(1)$.



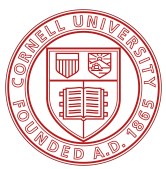
Review of weight matrix for the ring graph

- The weight matrix for the 4-node ring graph is static



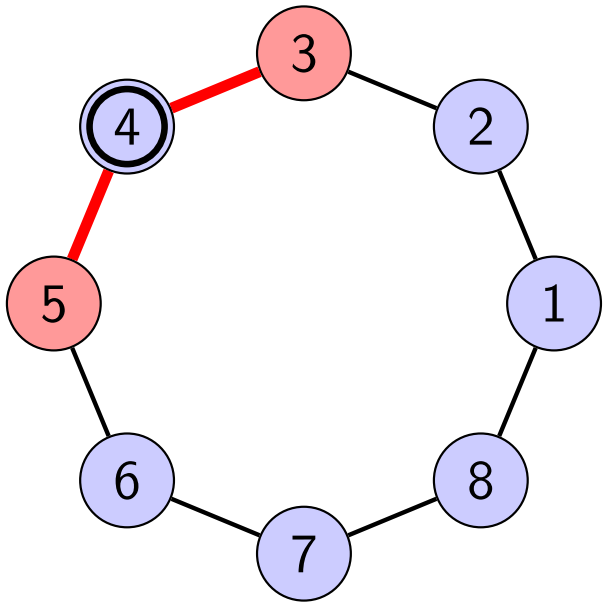
$$\mathbf{W} = \begin{bmatrix} 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \end{bmatrix}$$

- $\mathbf{W}$ is **doubly stochastic** ($\mathbf{1}^T\mathbf{W} = \mathbf{1}^T$ and $\mathbf{W1} = \mathbf{1}$).
- (Deterministic) Gossip: **reach consensus**
- Adding more edges to the graph can improve the consensus rate

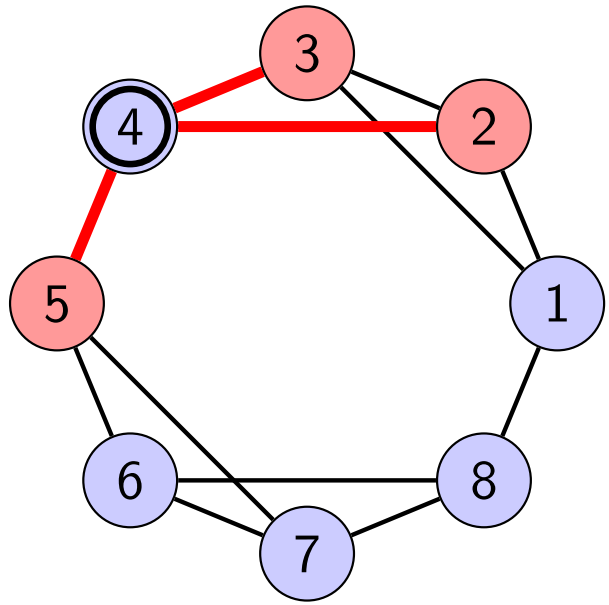


More graph topologies

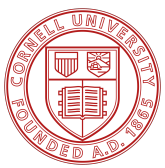
- Decentralized SGD can be defined for all undirected graphs.



Ring graph: every node has 2 neighbors



Every node has 3 neighbors.



Decentralized SGD: partial averaging via gossip

- DSGD: local SGD + partial averaging with neighbors

Decentralized SGD (DSGD)

Given a communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, at each round k,

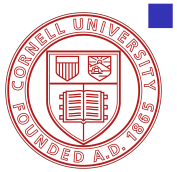
1. Each node i runs the local SGD:

$$\boldsymbol{x}_i^{(k+\frac{1}{2})} = \boldsymbol{x}_i^{(k)} - \eta \nabla F(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k)})$$

2. Each node i broadcasts its current state to its neighbors $j \in \mathcal{N}_i$.
3. All nodes update their states by partial averaging with neighbors

$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{x}_j^{k+\frac{1}{2}}$$

- **DSGD** reduces the spatial comm. cost when the graph is sparse.



DSGD has lower per-iter comm. cost than PSGD

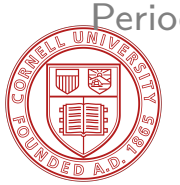
- Experiment on a **256-GPU** cluster. Use Ring-AllReduce as global averaging protocol for PSGD.

Model	**Ring-AllReduce (ms)**	**Partial Averaging (ms)**
ResNet-50 (25.5M)	278	**150**

Table: Per-iteration *communication* runtime (lower is better [1]).

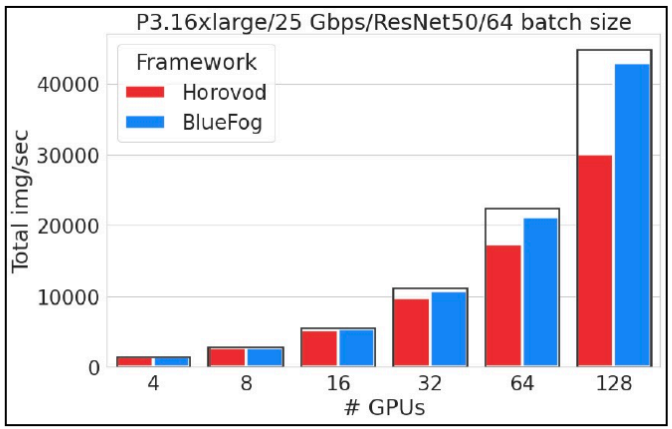
- **46% reduction** vs. PSGD.

[1] Y. Chen, K. Yuan, Y. Zhang, P. Pan, Y. Xu, and W. Yin, "Accelerating Gossip SGD with Periodic Global Averaging , ICML 2021.

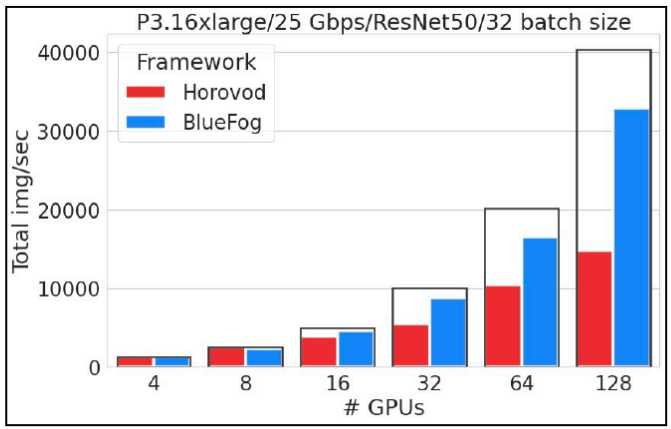


DSGD is more communication-efficient than PSGD

- DSGD (BlueFog) has better scalability than PSGD (Horovod) due to its small comm. overhead.

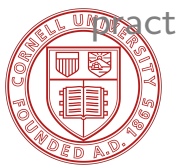


Small comm.-to-compt. ratio



Large comm.-to-compt. ratio

[2] B. Ying, K. Yuan, H. Hu, Y. Chen and W. Yin, BlueFog: Make decentralized algorithms practical for optimization and deep learning, arXiv: 2111. 04287, 2021



Summary of DSGD

Key idea: Instead of global averaging, partial averaging with neighbors

DSGD

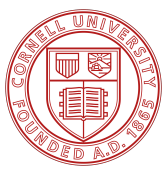
$$\boldsymbol{x}_i^{k+\frac{1}{2}} = \boldsymbol{x}_i^k - \eta \nabla F(\boldsymbol{x}_i^k; \xi_i^k)$$

$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{x}_j^{k+\frac{1}{2}}$$

- Each node i samples mini-batch ξ_i^k and computes $\nabla F\left(\boldsymbol{x}_i^k; \xi_i^k\right)$.
- Nodes synchronize with its neighbors.

Benefit: Reduce comm. cost to $O\left(d_{\max}\right)$ where $d_{\max} = \max_i \{|\mathcal{N}_i|\}$.

How about its convergence?



Quadratic objectives: setup and average dynamics

Consider local quadratic objectives

$$f_i(\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}_i \boldsymbol{x} - \boldsymbol{b}_i^\top \boldsymbol{x},$$

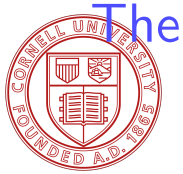
and the decentralized update (no noise):

$$\boldsymbol{x}_i^{k+1} = \sum_{j=1}^{n} w_{ij}\Big(\boldsymbol{x}_j^k - \eta(\boldsymbol{A}_j \boldsymbol{x}_j^k - \boldsymbol{b}_j)\Big), \quad \text{where } w_{ij} = 0 \text{ if } j \notin \mathcal{N}_i.$$

Define the (virtual) average $\bar{\boldsymbol{x}}^k = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^k$, and if $\mathbf{W}$ is **column-stochastic** ($\sum_i w_{ij} = 1$), then

$$\bar{\boldsymbol{x}}^{k+1} = \bar{\boldsymbol{x}}^k - \eta\,\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{A}_j \boldsymbol{x}_j^k - \boldsymbol{b}_j).$$

The average iterate follows a global gradient-like step on local gradients.



Quadratic objectives: gradient decomposition

We can decompose each local gradient around the average point:

$$\boldsymbol{A}_j \boldsymbol{x}_j^k - \boldsymbol{b}_j = \boldsymbol{A}_j \bar{\boldsymbol{x}}^k - \boldsymbol{b}_j + \boldsymbol{A}_j(\boldsymbol{x}_j^k - \bar{\boldsymbol{x}}^k).$$

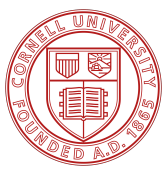
Substituting into the update gives

$$\bar{\boldsymbol{x}}^{k+1} = \bar{\boldsymbol{x}}^k - \eta \left[\tfrac{1}{n} \sum_{j=1}^n (\boldsymbol{A}_j \bar{\boldsymbol{x}}^k - \boldsymbol{b}_j) \right] - \eta \, \tfrac{1}{n} \sum_{j=1}^n \boldsymbol{A}_j (\boldsymbol{x}_j^k - \bar{\boldsymbol{x}}^k).$$

Define the **consensus error term:**

$$\boldsymbol{E}_k := \tfrac{1}{n} \sum_{j=1}^n \boldsymbol{A}_j (\boldsymbol{x}_j^k - \bar{\boldsymbol{x}}^k).$$

Then $$\bar{\boldsymbol{x}}^{k+1} = \bar{\boldsymbol{x}}^k - \eta \, \tfrac{1}{n} \sum_{i=1}^n \nabla f_i(\bar{\boldsymbol{x}}^k) - \eta \, \boldsymbol{E}_k.$$



Quadratic objectives: interpretation

Hence the average iterate evolves as

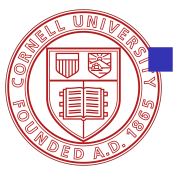
$$\bar{\boldsymbol{x}}^{k+1} = \bar{\boldsymbol{x}}^k - \eta \left(\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\bar{\boldsymbol{x}}^k) + \boldsymbol{E}_k \right).$$

If the matrices $\boldsymbol{A}_i$ are uniformly bounded, say $\|\boldsymbol{A}_i\| \le L$, then

$$\|\boldsymbol{E}_k\| \le \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{A}_i\| \, \|\boldsymbol{x}_i^k - \bar{\boldsymbol{x}}^k\| \le L \max_i \|\boldsymbol{x}_i^k - \bar{\boldsymbol{x}}^k\|.$$

The virtual consensus point follows a gradient step on the *global objective* $\frac{1}{n}\sum_i f_i$, up to an error proportional to the consensus disagreement.

- Column-stochasticity ($\sum_i w_{ij} = 1$) preserves the average.
- The consensus error $\|\boldsymbol{E}_k\|$ decays with the spectral gap $1 - \lambda_2(W)$.
- Faster consensus $\Rightarrow$ smaller error $\Rightarrow$ near-centralized convergence.



Does DSGD ensure convergence?

- **Convergence of DSGD** depends on the consensus rate of the partial averaging and the convergence of PSGD.
- The consensus rate depends on the graph topology.
  - The fully connected graph reaches the consensus after one-step averaging.
  - For a ring graph, although not exact consensus, still closer.
- PSGD can tolerate inexact consensus if the error is decreasing.
- **Convergence guarantee? Speed?**

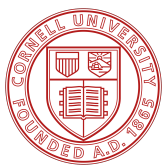
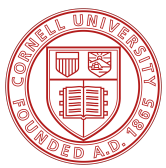


Table of Contents



Partial averaging as one-step average consensus

- Write the DSGD in the matrix form.

DSGD

$$\boldsymbol{x}_i^{k+\frac{1}{2}} = \boldsymbol{x}_i^k - \eta \nabla F(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k)})$$

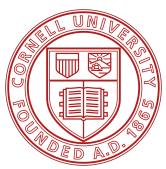
$$\boldsymbol{x}^{k+1} = W \boldsymbol{x}^{k+\frac{1}{2}}$$

where $\boldsymbol{x}^l = [\boldsymbol{x}_1^l, \cdots, \boldsymbol{x}_n^l], \quad l = k, k+\frac{1}{2}.$

- Recall average consensus algorithm (fixed weight matrix)

$$\mathbf{x}(t+1) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}^t\mathbf{x}(0)$$

- The partial averaging in DSGD is one-step ($t = 1$) version of average consensus algorithm with varying $x(0) = \boldsymbol{x}^{k+\frac{1}{2}}$



Review: Convergence rate of consensus

- Recall the average consensus algorithm (fixed weight matrix)

$$\mathbf{x}(t+1) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}^t\mathbf{x}(0)$$

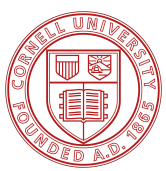
Theorem 1 (Convergence rate of average consensus)

If $\mathbf{W}$ is doubly stochastic, it holds for the average consensus protocol that

$$\left\| \mathbf{x}(t) - \frac{\mathbf{1}\mathbf{1}^T\mathbf{x}(t)}{n} \right\| \leq \rho^t \left\| \mathbf{x}(0) - \frac{\mathbf{1}\mathbf{1}^T\mathbf{x}(0)}{n} \right\|,$$

where $\rho = \max_{i\geq 2} |\lambda_i(\mathbf{W})| < 1$.

Q: Is one-step consensus enough for DSGD's convergence?

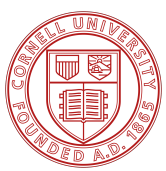


Evolution of DSGD

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

- DSGD converges if partial averaging reaches consensus (global average) asymptotically
- Letting $\bar{\boldsymbol{x}}^{k+l} = \frac{\mathbf{1}\mathbf{1}^T\boldsymbol{x}^{k+l}}{n}, \ell = 1/2, 0$, partial averaging in DSGD:

$$\begin{aligned}\left\|\boldsymbol{x}^{k+1} - \bar{\boldsymbol{x}}^{k+1}\right\| &\le \rho \left\|\boldsymbol{x}^{k+\frac{1}{2}} - \bar{\boldsymbol{x}}^{k+\frac{1}{2}}\right\| \\ &\le \rho\eta \sum_{i=1}^{n} \left\|\nabla f_i(\boldsymbol{x}_i^k) - \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}_i^k)\right\| + \rho\eta n\sigma\end{aligned}$$

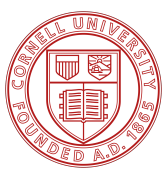


Evolution of DSGD

- Letting $\bar{\boldsymbol{x}}^{k+l} = \frac{\mathbf{1}\mathbf{1}^T \boldsymbol{x}^{k+l}}{n}, \ell = 1/2, 0$, partial averaging in DSGD:

$$
\begin{aligned}
&\left\|\boldsymbol{x}^{k+1} - \bar{\boldsymbol{x}}^{k+1}\right\| \\
\leq\ & \rho\eta \sum_{i=1}^{n} \left\| \nabla f_i(\boldsymbol{x}_i^k) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}_i^k) \right\| + \rho\eta n\sigma \\
\leq\ & \rho\eta \sum_{i=1}^{n} \left\| \nabla f_i(\bar{\boldsymbol{x}}_i^k) - \nabla f_i(\boldsymbol{x}_i^k) \right\| + \rho\eta \sum_{i=1}^{n} \left\| \nabla f_i(\bar{\boldsymbol{x}}_i^k) - \nabla f(\bar{\boldsymbol{x}}_i^k) \right\| \\
&+ \rho\eta \sum_{i=1}^{n} \left\| \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}_i^k) - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\bar{\boldsymbol{x}}_i^k) \right\| + \rho\eta n\sigma \\
\leq\ & 2\rho\eta L \left\| \bar{\boldsymbol{x}}^k - \boldsymbol{x}^k \right\| + \rho\eta b + \rho\eta n\sigma
\end{aligned}
$$

where $b^2 \triangleq \frac{1}{n}\sum_{i=1}^{n} \left\|\nabla f_i(x) - \nabla f(x)\right\|^2$ and L is the smoothness.

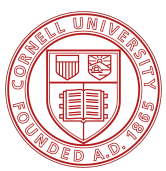


Evolution of DSGD

- Letting $\bar{\boldsymbol{x}}^{k+l} = \frac{\mathbf{1}\mathbf{1}^T \boldsymbol{x}^{k+l}}{n}, \ell = 1/2, 0$, partial averaging in DSGD:

$$\left\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^k\right\| \leq (2\rho\eta L)^k \left\|\bar{\boldsymbol{x}}^k - \boldsymbol{x}^k\right\| + \frac{\rho\eta(b + n\sigma)}{1 - 2\rho\eta L} \to 0 \quad \text{when} \quad \eta \to 0$$

where $b^2 \triangleq \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x) - \nabla f(x)\right\|^2$ and L is the smoothness.



Convergence rate of DSGD

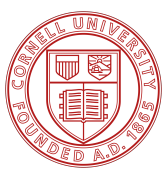
Assumptions:

- f_i: L-smooth
- $\nabla F(\boldsymbol{x}; \xi_i)$ is an unbiased estimate of $\nabla f_i(\boldsymbol{x})$, with bounded variance
- Bounded data heterogeneity: $\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq b^2$.

Theorem 1 (Convergence rate of DSGD)

Suppose above assumptions hold and let $\eta = 1/\sqrt{K}$. Then

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[||\nabla f(\bar{\boldsymbol{x}}^k)||_2^2] \leq \mathcal{O}\left(\frac{\sigma}{\sqrt{nK}} + \frac{\rho^{2/3}\sigma^{2/3}}{K^{2/3}(1-\rho)^{1/3}} + \frac{\rho^{2/3} b^{2/3}}{K^{2/3}(1-\rho)^{2/3}}\right)$$



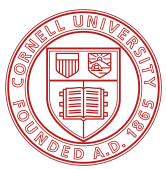
Convergence of DSGD

Theorem 1 (Convergence rate of DSGD)

Suppose above assumptions hold and let $\eta = 1/\sqrt{K}$. Then

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[||\nabla f(\bar{\boldsymbol{x}}^k)||_2^2] \leq \mathcal{O}\left(\frac{\sigma}{\sqrt{nK}} + \frac{\rho^{2/3}\sigma^{2/3}}{K^{2/3}(1-\rho)^{1/3}} + \frac{\rho^{2/3}b^{2/3}}{K^{2/3}(1-\rho)^{2/3}}\right)$$

- The first term dominates when $K \to \infty$, which suggests DSGD achieves *linear speedup* asymptotically.
- Convergence of DSGD depends on the *network topology*: sparse topology ($\rho \to 1$) results in slower convergence.
- Convergence of DSGD depends on the *data heterogeneity*: large heterogeneity b results in slower convergence.

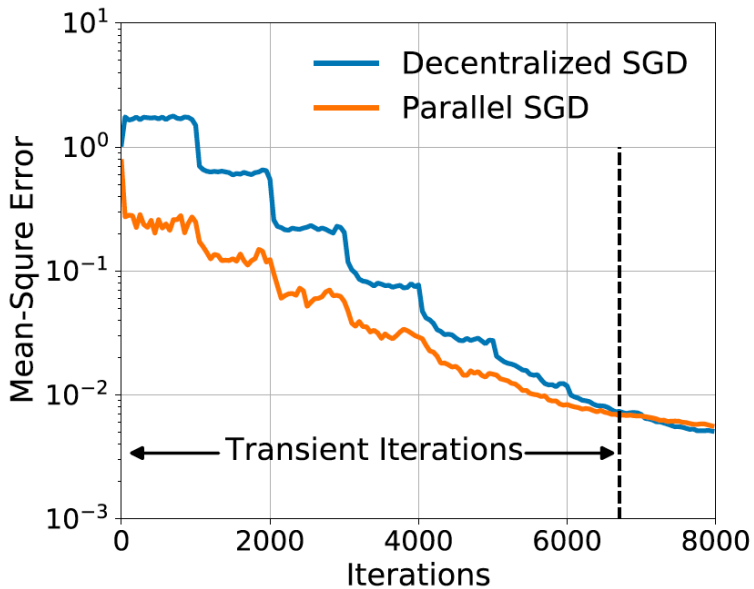


PSGD v.s. DSGD

PSGD: $\mathcal{O}\left(\frac{\sigma}{\sqrt{nK}}\right)$ Transient iterations: Extra overhead $\leq \frac{\sigma}{\sqrt{nK}}$

DSGD: $\mathcal{O}\Big(\frac{\sigma}{\sqrt{nK}} + \boxed{\frac{\rho^{2/3}\sigma^{2/3}}{K^{2/3}(1-\rho)^{1/3}} + \frac{\rho^{2/3}b^{2/3}}{K^{2/3}(1-\rho)^{2/3}}}\Big)$

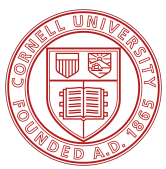
Extra overhead



- DSGD can asymptotically converge as fast as P-SGD
- Transient iterations:

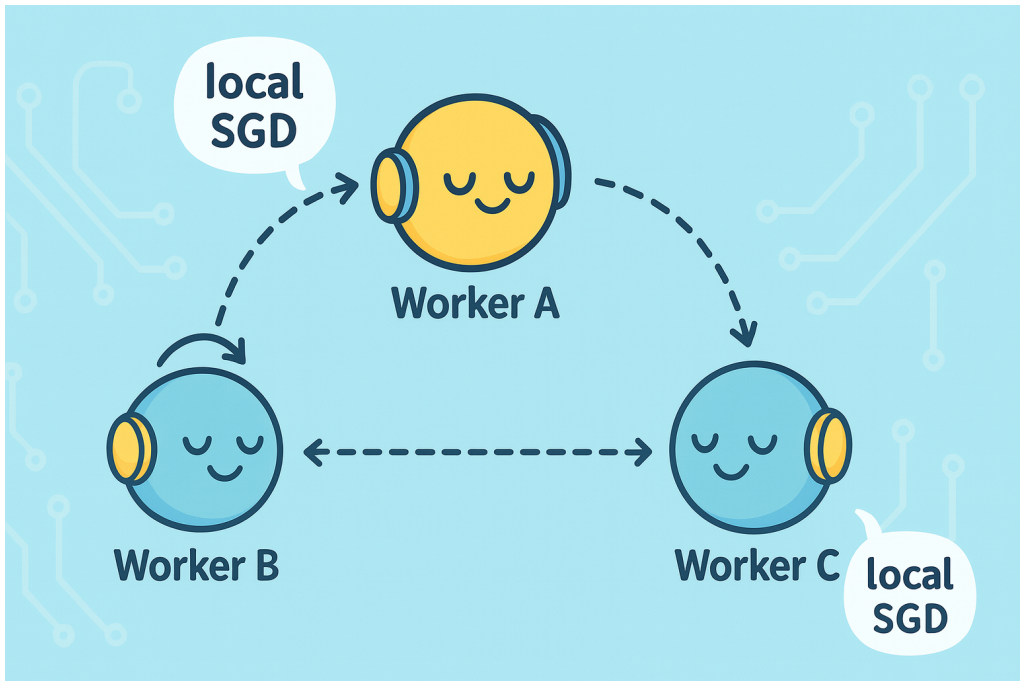
$$\mathcal{O}\left(\frac{\rho^4 n^3}{\sigma^2(1-\rho)^2} + \frac{\rho^4 n^3 b^4}{\sigma^6(1-\rho)^4}\right)$$

affected by *network topology* and *data heterogeneity*.



Summary of DSGD

- **Principle:** In each time step, nodes update its state by local SGD and communicate with its neighbors to perform partial averaging.
- **Benefit:** Low per-iter. communication cost when graph is sparse; achieve consensus asymptotically.
- **Drawback:** Cannot handle heterogeneous data setting.



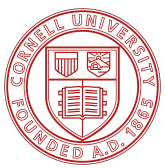
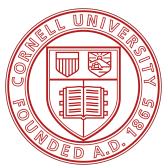


Table of Contents



Why DSGD suffers from data heterogeneity?

DSGD

$$\boldsymbol{x}_i^{k+\frac{1}{2}} = \boldsymbol{x}_i^k - \eta \nabla F(\boldsymbol{x}_i^{(k)}; \boldsymbol{\xi}_i^{(k)})$$

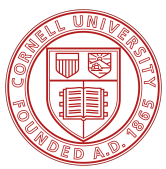
$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{x}_j^{k+\frac{1}{2}}$$

- Consider the setting without gradient noise, DSGD can be written as

$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \left(\boldsymbol{x}_i^k - \eta \nabla f_i(\boldsymbol{x}_i^{(k)}) \right)$$

- When achieving stationary: $\boldsymbol{x}_i^k = \boldsymbol{x}^*$ for all i. Then the next step is

$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \left(\boldsymbol{x}^* - \eta \nabla f_i(\boldsymbol{x}^*) \right) = \boldsymbol{x}^* - \eta \sum_{j \in \mathcal{N}_i} w_{ij} \nabla f_i(\boldsymbol{x}^*)$$



Why DSGD suffers from data heterogeneity?

- In homogeneous scenario: $\nabla f_i(\boldsymbol{x}^*) = 0$ for all i. Then the stationary point $\boldsymbol{x}^*$ is stable because

$$\boldsymbol{x}_i^{k+1} = \boldsymbol{x}^* - \eta \sum_{j \in \mathcal{N}_i} w_{ij} \nabla f_i(\boldsymbol{x}^*) = \boldsymbol{x}^*$$

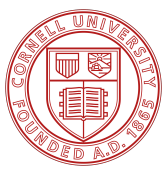
- In heterogeneous setting: $\nabla f_i(\boldsymbol{x}^*) \neq \frac{1}{n} \sum_{i=1}^n \nabla f_i(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*) = 0$.

$$\boldsymbol{x}_i^{k+1} = \boldsymbol{x}^* - \eta \sum_{j \in \mathcal{N}_i} w_{ij} \nabla f_i(\boldsymbol{x}^*) \neq \boldsymbol{x}^*$$

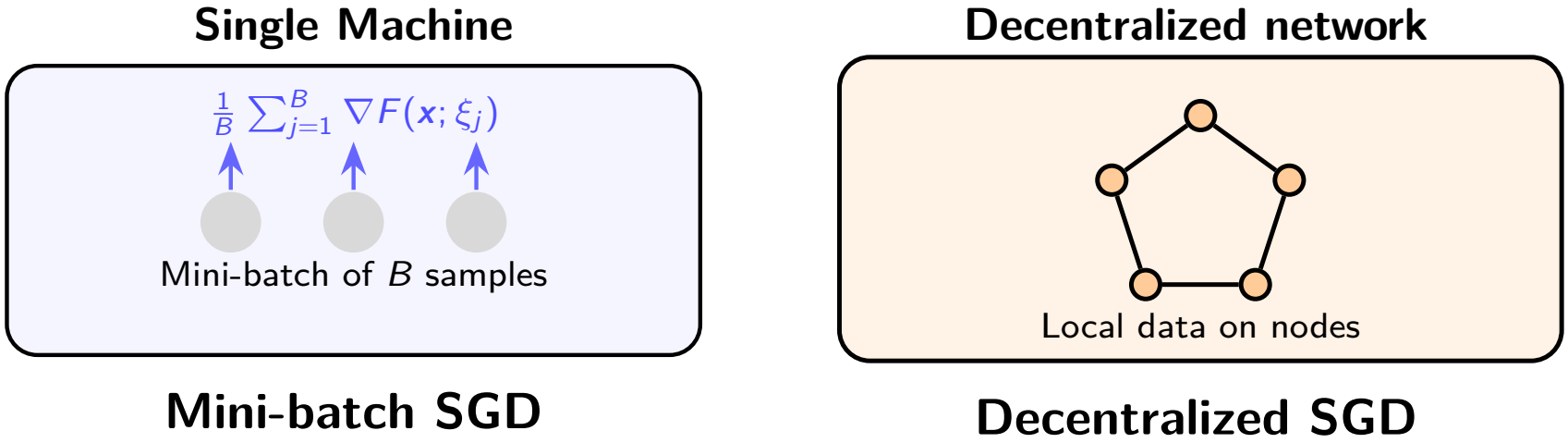
  suggesting the stationary point is not stable.

- Cause divergence when $b^2 \triangleq \frac{1}{n} \sum_{i=1}^n \left\| \nabla f_i(x) - \nabla f(x) \right\|^2$ is large.

Q: How to alleviate this issue?

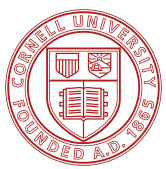


Insights from mini-batch SGD



- Gradient $\nabla f_i(\boldsymbol{x})$ for each node can be viewed as a "mini-batch" gradient with respect to the the global gradient $\nabla f(\boldsymbol{x})$.
- Data heterogeneity $b^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2$ is in the same spirit of "variance", but is for each worker.

A: Applying 'variance reduction' technique to local workers.



Recall: how SVRG reduces the variance

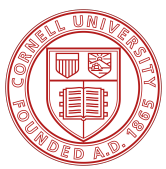
Key idea: SVRG replaces noisy gradients with a corrected version that re-centers them around the full gradient at a *snapshot point*.

- Periodically calculate the **full gradient** at a "snapshot" point $\tilde{\boldsymbol{x}}$.
- Use this full gradient as a "low-variance anchor" to correct

$$\mathbf{v}_{\text{SVRG}}^t = \nabla f_{i_t}(\boldsymbol{x}^t) - \nabla f_{i_t}(\tilde{\boldsymbol{x}}) + \nabla f(\tilde{\boldsymbol{x}})$$

- f_{i_t}: gradient for the current random sample i_t.
- $\tilde{\boldsymbol{x}}$: snapshot point (updated every epoch).
- $\nabla f(\tilde{\boldsymbol{x}})$: full gradient at $\tilde{\boldsymbol{x}}$.

How to track the global gradient in decentralized setting?



Track the global gradient

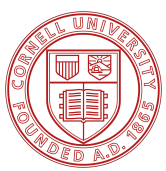
Key idea: introduce additional variable y which converges to the global gradient when $\nabla f_i(\boldsymbol{x}_i^k)$ achieves stationary point.

- Initialize the tracking variable for each node as $\boldsymbol{y}_i^0 = \nabla f_i(\boldsymbol{x}_i^{(0)})$.
- Synchronize the smoothed version with neighbors

$$\boldsymbol{y}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \left(\boldsymbol{y}_j^k + \nabla f_j(\boldsymbol{x}_j^{k+1}) - \nabla f_j(\boldsymbol{x}_j^k) \right)$$

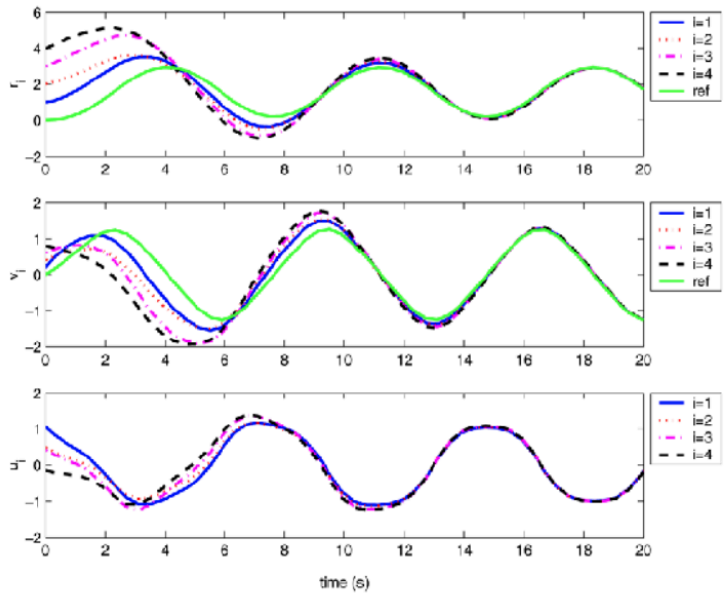
- When each node reaches stationary, the gradient difference of two steps vanishes and

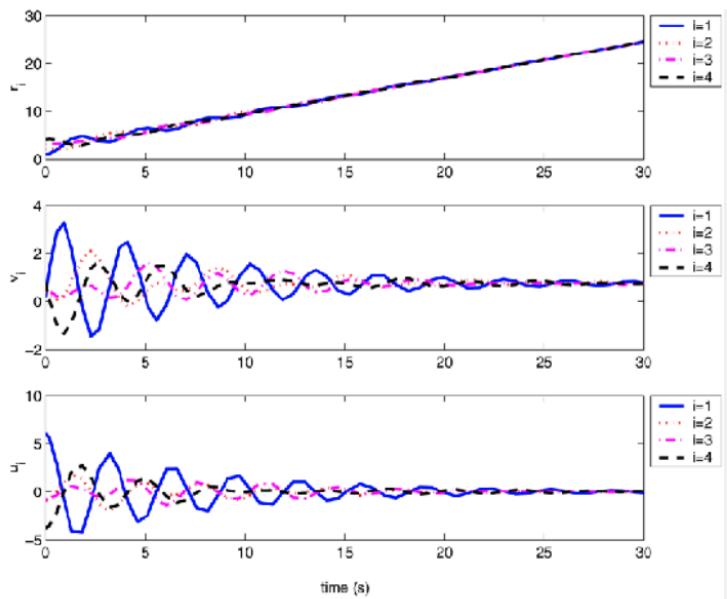
$$\boldsymbol{y}_i^{k+1} \to \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\boldsymbol{x}_i^k) \quad \leftarrow \text{ global gradient}$$



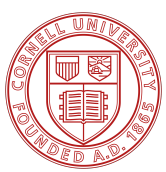
Tracking variable converges to global gradient

- Green: global gradient





- Tracking variable converges to the global gradient



Simulation results are from [Ren IEEE TAC 2007]

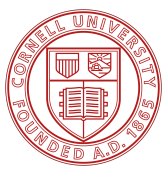
Gradient tracking algorithm

- Consider the general setting where we have gradient noise. Gradient tracking algorithm is defined as follows.

$$\boldsymbol{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \left(\boldsymbol{x}_i^k - \eta \boldsymbol{y}_i^k \right)$$

$$\boldsymbol{y}_i^{k+1} = \sum_{j \in \mathcal{N}_i} w_{ij} \left(\boldsymbol{y}_j^k + \nabla F(\boldsymbol{x}_j^{k+1}; \xi_j^{k+1}) - \nabla F(\boldsymbol{x}_j^k; \xi^k) \right)$$

- **Key difference**: introduce a tracking variable that can converge to the global averaged gradient asymptotically



Convergence rate of Gradient Tracking

Assumptions:

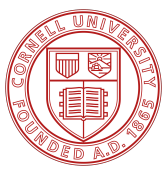
- f_i: L-smooth
- $\nabla F(\boldsymbol{x}; \xi_i)$ is an unbiased estimate of $\nabla f_i(\boldsymbol{x})$, with bounded variance

Theorem 2 (Convergence rate of GT)

Suppose above assumptions hold and with proper η. Then

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[||\nabla f(\bar{\boldsymbol{x}}^k)||_2^2] \leq \mathcal{O}\left(\frac{\sigma}{\sqrt{nK}} + \frac{\rho^{2/3}\sigma^{2/3}}{K^{2/3}(1-\rho)^{1/3}}\right)$$

- **Key feature:** remove the bounded data heterogeneity assumption!

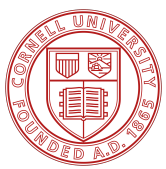


DSGD v.s. Gradient Tracking

DSGD: $\mathcal{O}\Big(\frac{\sigma}{\sqrt{nK}} + \underbrace{\boxed{\frac{\rho^{2/3}\sigma^{2/3}}{K^{2/3}(1-\rho)^{1/3}} + \frac{\rho^{2/3}b^{2/3}}{K^{2/3}(1-\rho)^{2/3}}}}_{\text{Extra overhead}}\Big)$

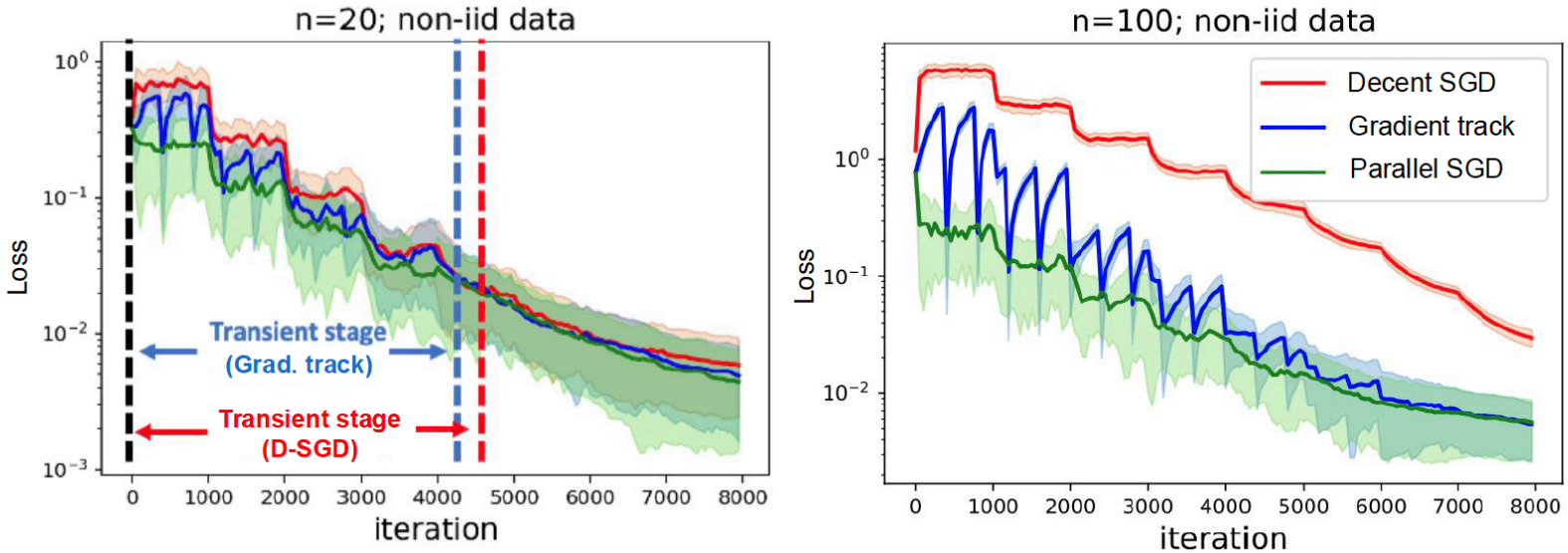
GT: $\mathcal{O}\Big(\frac{\sigma}{\sqrt{nK}} + \underbrace{\boxed{\frac{\rho^{2/3}\sigma^{2/3}}{K^{2/3}(1-\rho)^{1/3}}}}_{\text{Extra overhead}}\Big)$

- Gradient Tracking shorten the transient stage
- DSGD: $\mathcal{O}\left(\frac{\rho^4 n^3}{\sigma^2(1-\rho)^2} + \frac{\rho^4 n^3 b^4}{\sigma^6(1-\rho)^4}\right) \longrightarrow$ GT: $\mathcal{O}\left(\frac{\rho^4 n^3}{\sigma^2(1-\rho)^2}\right)$

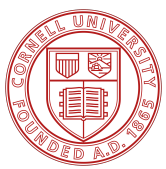


Empirical studies on heterogeneous data

- DNN training on ring graph $(1 - \rho = O\left(n^{-2}\right))$



- Gradient tracking has shorter transient period.
- Gradient tracking outperforms DSGD on heterogeneous data.



Simulation results are from [Ren IEEE TAC 2007]

Recap and fine-tuning

- What we have talked about today?

  $\Rightarrow$ **Decentralized SGD** reduces spatial communication reduction via partial averaging with neighbors.

  $\Rightarrow$ **Decentralized SGD** achieves the same convergence rate as Parallel SGD asymptotically.

  $\Rightarrow$ **Gradient Tracking** tackles the data hetergenoity issue by tracking the global averaged gradient.

Welcome anonymous survey!