北京大学
PEKING UNIVERSITY

Optimization for Deep Learning

Lecture 13-4: Decentralized Stochastic Optimization

Kun Yuan

北京大学
PEKING UNIVERSITY

Part 01

Linear speedup and transient stage

Distributed learning

北京大学 PEKING UNIVERSITY

- Training deep neural networks typically requires **massive** datasets; efficient and scalable distributed optimization algorithms are in urgent need

- A network of n nodes (devices such as GPUs) collaborate to solve the problem:

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

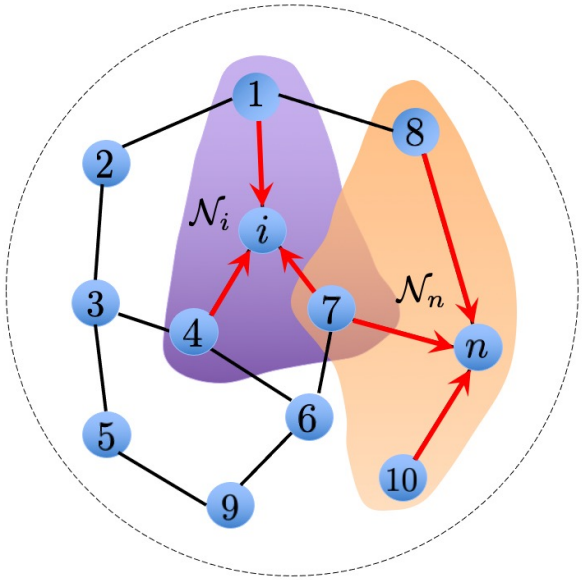
  - Each component $f_i : \mathbb{R}^d \to \mathbb{R}$ is local and private to node i
  - Random variable ξ_i denotes the local data that follows distribution D_i
  - Each local distribution D_i is different; data heterogeneity exists

Decentralized SGD (DSGD)

- To break $O(n)$ comm. overhead, we replace global average with partial average

$$x_i^{(k+\frac{1}{2})} = x_i^{(k)} - \gamma \nabla F(x_i^{(k)}; \xi_i^{(k)}) \quad \text{(Local update)}$$

$$x_i^{(k+1)} = \sum_{j \in \mathcal{N}_i} w_{ij} x_j^{(k+\frac{1}{2})} \quad \text{(Partial averaging)}$$



- DSGD = local SGD update + partial averaging [LS08]
- $\mathcal{N}_i$ is the set of neighbors at node i ; w_{ij} scales information from j to i and satisfies $\sum_{j \in \mathcal{N}_i} w_{ij} = 1$
- Incurs $O(d_{\max})$ comm. overhead per iteration where $d_{\max} = \max_i \{|\mathcal{N}_i|\}$ is the graph maximum degree

DSGD is more communication-efficient than PSGD

北京大学 PEKING UNIVERSITY

- A real experiment on a 256-GPUs cluster [CYZ+21]

Model	Ring-Allreduce	Partial average
ResNet-50 (25.5M)	278 ms	150 ms

Table. Comparison of per-iter comm. time in terms of runtime with 256 GPUs

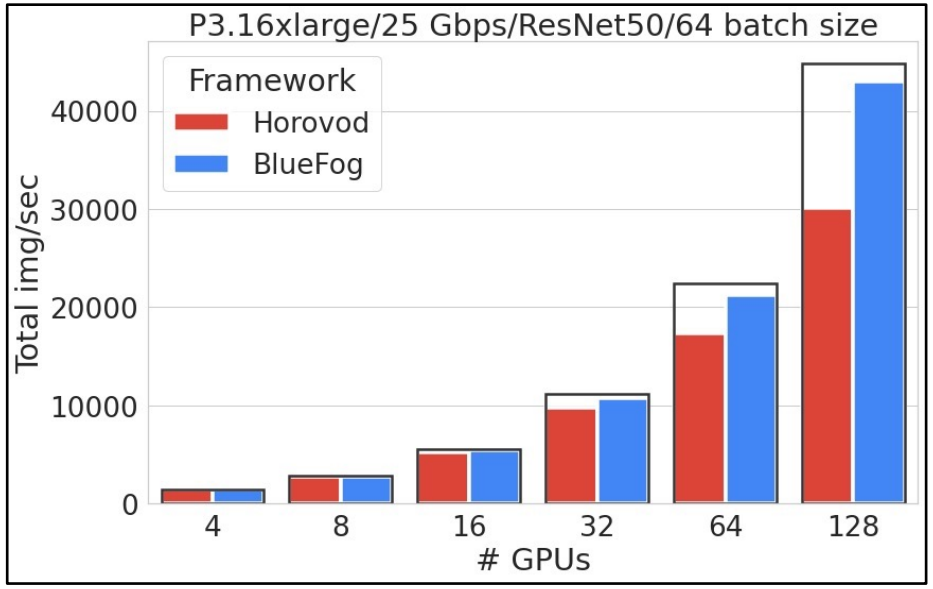
- DSGD saves more communications per iteration for larger models

[CYZ+21] Y. Chen*, K. Yuan*, Y. Zhang, P. Pan, Y. Xu, and W. Yin, ``Accelerating Gossip SGD with Periodic Global Averaging", ICML 2021

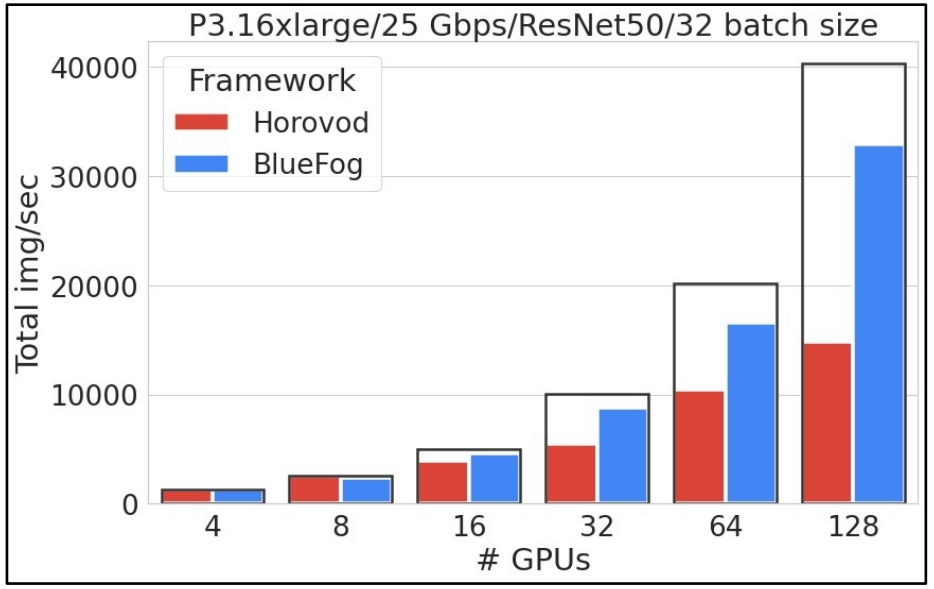
DSGD is more communication-efficient than PSGD

北京大学 PEKING UNIVERSITY

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



Small comm.-to-compt. ratio



Large comm.-to-compt. ratio

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

However, DSGD has slower convergence

- The efficient comm. comes with a cost: slower convergence
- Partial average $x_i^+ = \sum w_{ij} x_j$ is less effective to aggregate information than global average
- The average effectiveness can be evaluated by **graph spectral gap**:

$$\rho = \|W - \frac{1}{n}\mathbb{1}\mathbb{1}^T\|_2 \in (0,1) \text{ where } W = [w_{ij}] \in \mathbb{R}^{n\times n}$$

  - Well-connected topology has $\rho \to 0$, e.g. fully-connected topology
  - Sparsely-connected topology has $\rho \to 1$, e.g. ring has $\rho = O(1 - \frac{1}{n^2})$
  - ρ or $1-\rho$ essentially gauges the **graph connectivity**

DSGD convergence rate

- Convergence comparison (non-convex and **data-homogeneous** scenario) [KLB+20]:

$$\text{P-SGD}: \quad \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = O\left(\frac{\sigma}{\sqrt{nT}}\right)$$

$$\text{D-SGD}: \quad \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = O\left(\frac{\sigma}{\sqrt{nT}} + \underbrace{\frac{\rho^{2/3}\sigma^{2/3}}{T^{2/3}(1-\rho)^{1/3}}}_{\text{extra overhead}}\right)$$

where σ^2 is the gradient noise, and T is the number of iterations

- D-SGD can asymptotically converge as fast as P-SGD when $T \to \infty$; the first term dominates; reach **linear speedup** asymptotically

- But D-SGD **requires more iteration** to reach that stage due to the overhead caused by partial average

Transient iterations

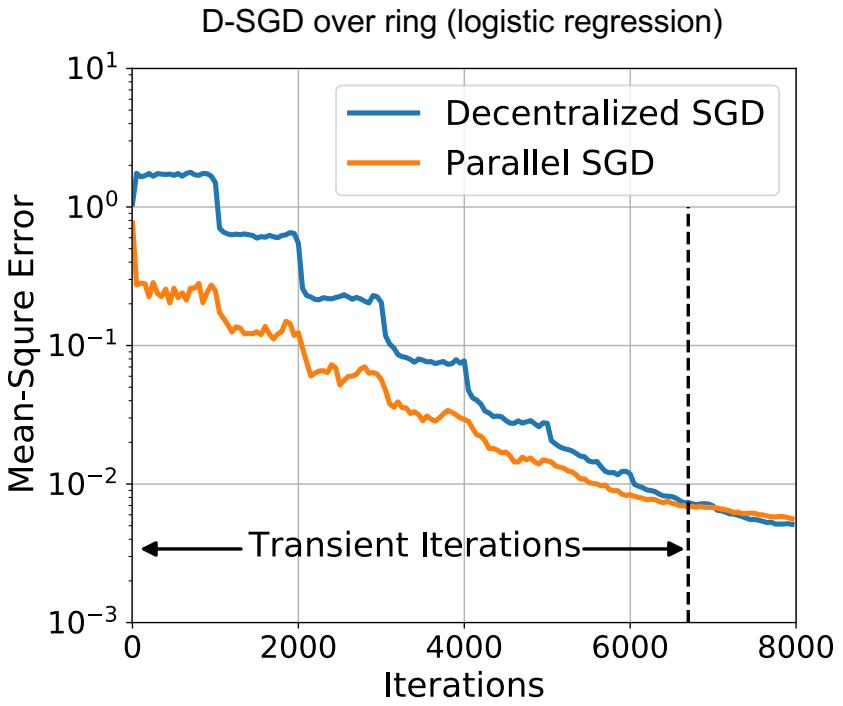
- Definition [POP21]: number of iterations before D-SGD achieves linear speedup

- D-SGD for non-convex and data-homogeneous scenario has $O(n^3(1-\rho)^{-2})$ transient iterations

$$\frac{\rho^{2/3}\sigma^{2/3}}{T^{2/3}(1-\rho)^{1/3}} \leq \frac{\sigma}{\sqrt{nT}} \quad \Longrightarrow \quad O(\frac{\rho^4 n^3}{(1-\rho)^2})$$

- Topology significantly influence the trans. stage.

- Sparse topology $\rho \to 1$ incurs longer tran. Iters.



北京大学
PEKING UNIVERSITY

PART 02

Data Heterogeneity leads to longer transient stage

Data heterogeneity causes even longer transient stage

- Recall the distributed stochastic optimization problem

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

- If local data distribution $\mathcal{D}_i$ is different from each other, we have

$$f_i(x)\neq f_j(x) \quad \text{when} \quad i\neq j$$

- We assume the gradient dissimilarity is upper bounded

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)-\nabla f(x)\|^2\leq b^2, \qquad \forall\, x\in\mathbb{R}^d$$

We term b^2 as **data heterogeneity**. When $b^2=0$, we have $\nabla f_i(x)=\nabla f(x)$ which reduces to homogeneity

Data heterogeneity causes even longer transient stage

- D-SGD convergence (non-convex and data-heterogeneous scenario) [KLB20+]

$$\text{P-SGD:} \quad \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = O\left(\frac{\sigma}{\sqrt{nT}}\right)$$

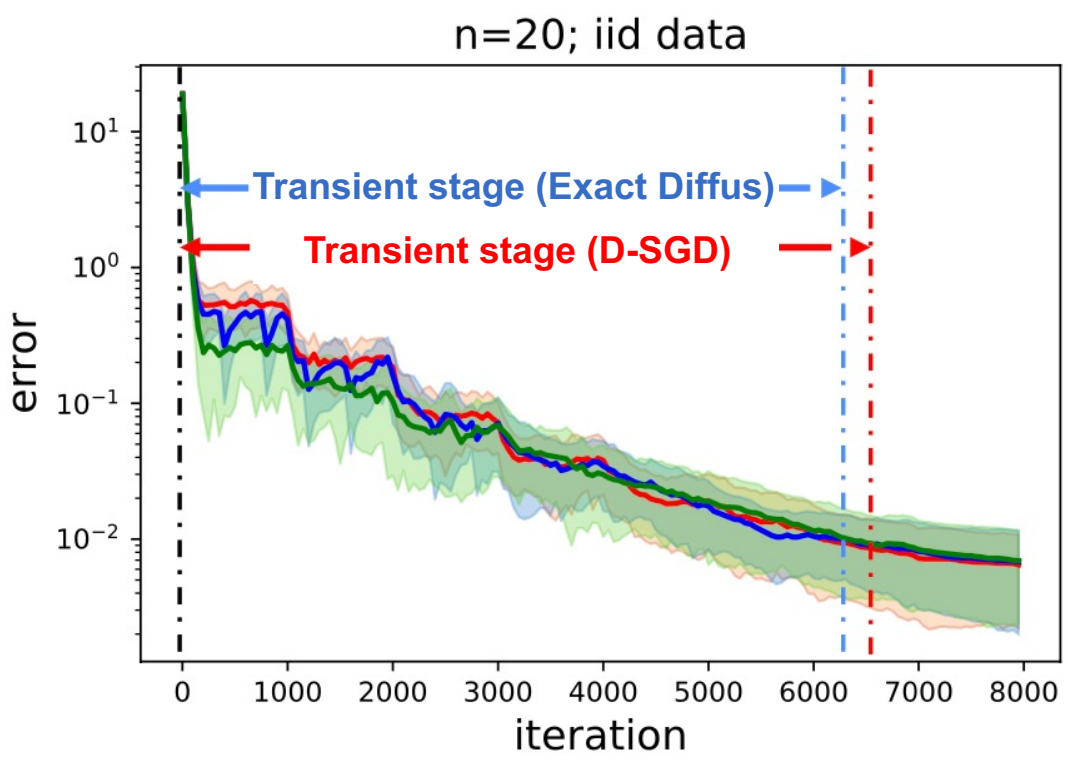
$$\text{D-SGD:} \quad \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = O\left(\frac{\sigma}{\sqrt{nT}} + \frac{\rho^{2/3}\sigma^{2/3}}{T^{2/3}(1-\rho)^{1/3}} + \frac{\rho^{2/3}b^{2/3}}{T^{2/3}(1-\rho)^{2/3}}\right)$$

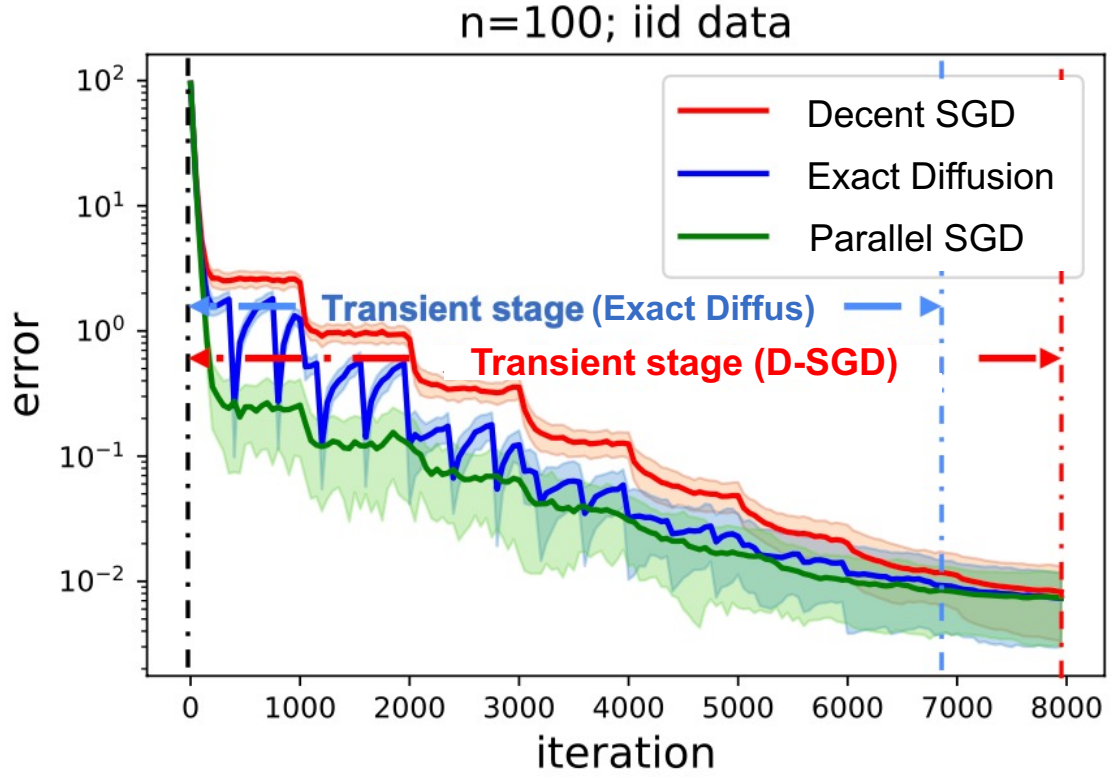
- Convergence gets slower due to additional data heterogeneity term
- For sparse topology in which $\rho \to 1$, transient stage gets significantly **worse**

Homogeneous: $O(\frac{n^3}{(1-\rho)^2})$ Heterogeneous: $O(\frac{n^3}{(1-\rho)^4})$

Data heterogeneity causes even longer transient stage

- Decentralized logistic regression with ring topology and **iid data**

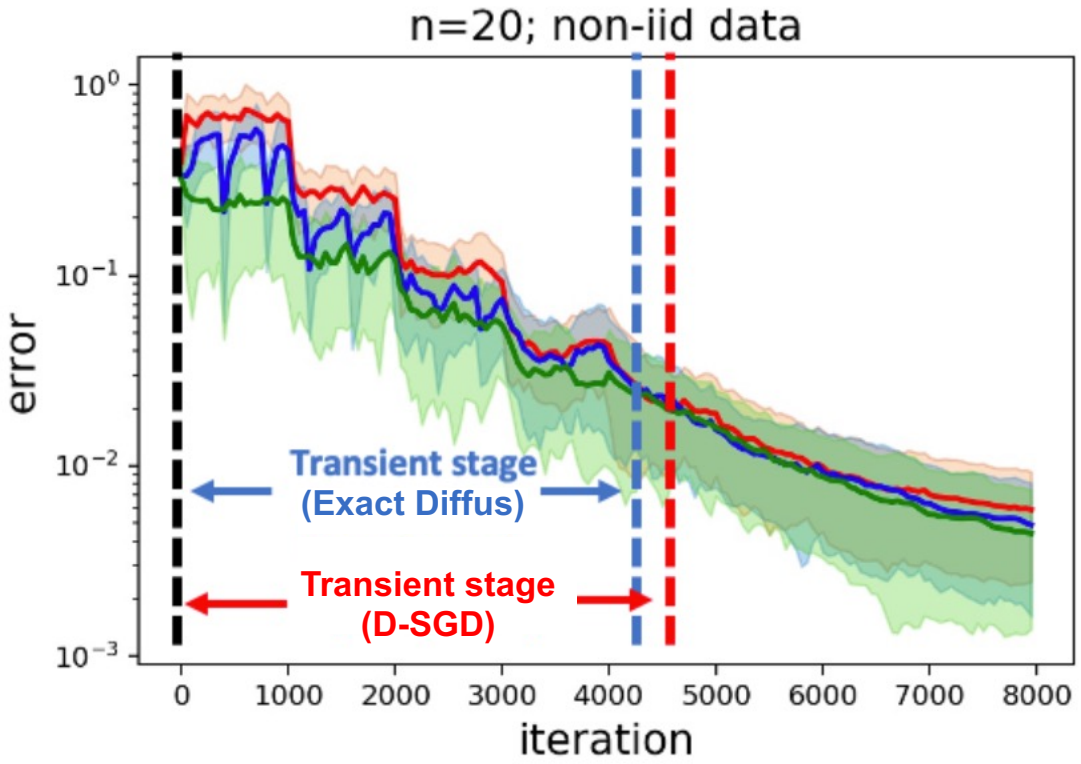


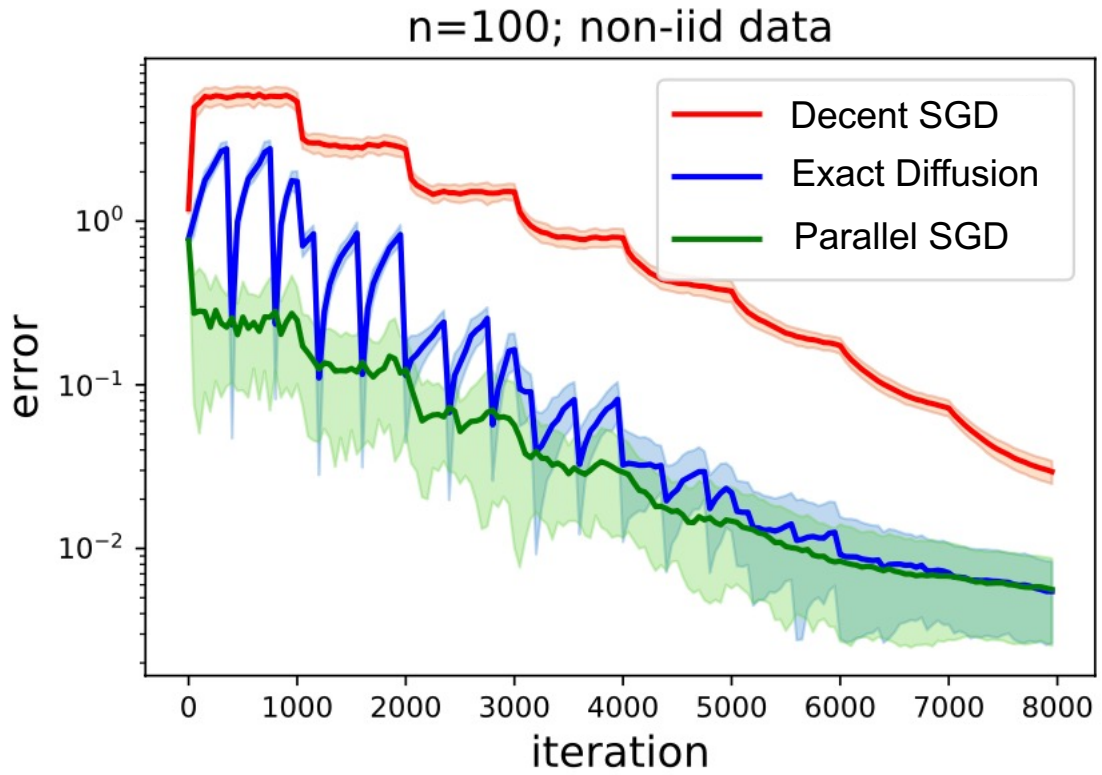


- When topology gets sparser, the transient stage of D-SGD gets longer due to $O(\frac{n^3}{(1-\rho)^2})$

Data heterogeneity causes even longer transient stage

- Decentralized logistic regression with ring topology with **non-iid data**





- When topology gets sparser, the transient stage of D-SGD gets **significantly longer** due to $O(\frac{n^3}{(1-\rho)^4})$

PART 03

Remove the Influence of Data Heterogeneity

Decentralized algorithms (adapt-then-combine version)

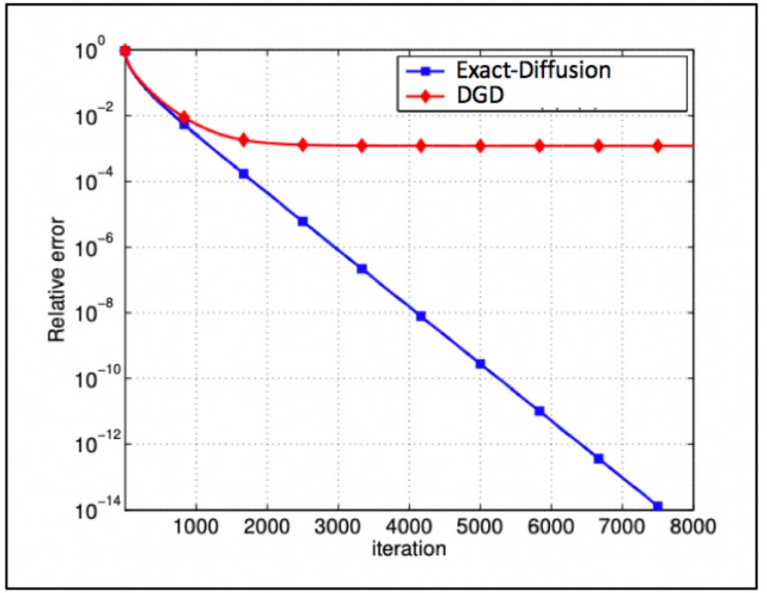
- In Part III, we have introduced several popular decentralized algorithms

(DGD) $$\mathbf{x}^{(k+1)} = W\left(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)})\right)$$

(Exact-Diffusion) $$\mathbf{x}^{(k+1)} = \left(\frac{W+I}{2}\right)\left[2\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} - \gamma\left(\nabla \mathbf{F}(\mathbf{x}^{(k)}) - \nabla \mathbf{F}(\mathbf{x}^{(k-1)})\right)\right]$$

(Gradient tracking)
$$\mathbf{x}^{(k+1)} = W\left(\mathbf{x}^{(k)} - \gamma \mathbf{y}^{(k)}\right)$$
$$\mathbf{y}^{(k+1)} = W\mathbf{y}^{(k)} + \nabla \mathbf{F}(\mathbf{x}^{(k+1)}) - \nabla \mathbf{F}(\mathbf{x}^{(k)})$$

- Exact-Diffusion and Gradient tracking remove the influence of data heterogeneity



Stochastic Exact-Diffusion

- Stochastic Exact-Diffusion

$$\mathbf{x}^{(k+1)} = \left(\frac{W+I}{2}\right)\left[2\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} - \gamma\left(\nabla\mathbf{F}(\mathbf{x}^{(k)};\xi^{(k)}) - \nabla\mathbf{F}(\mathbf{x}^{(k-1)};\xi^{(k)})\right)\right]$$

- We can rewrite it into a node-wise manner

$$\begin{aligned}
\psi_i^{(k)} &= x_i^{(k)} - \gamma\nabla F(x_i^{(k)};\xi_i^{(k)}) && \text{(Local update)}\\
\phi_i^{(k)} &= \psi_i^{(k)} + x_i^{(k)} - \psi_i^{(k-1)} && \text{(Correction)}\\
x_i^{(k+1)} &= \sum_{j\in\mathcal{N}_i} w_{ij}\phi_j^{(k)} && \text{(Partial averaging)}
\end{aligned}$$

- Exact-Diffusion with stochastic gradient is also called D2 which was studied in [TLYZ+18]

Exact-Diffusion has a shorter transient stage with non-iid data

Assumption 1. Each $f_i(x)$ is L-smooth

Assumption 2. The stochastic gradient satisfies

$$\mathbb{E}[\nabla F(x;\xi_i)] = \nabla f_i(x)$$
$$\mathbb{E}\|\nabla F(x;\xi_i) - \nabla f_i(x)\|^2 \le \sigma^2$$

Assumption 3. The weight matrix satisfies $W = W^T$, $W \succ 0$ and

$$\|W - \frac{1}{n}\mathbf{1}\mathbf{1}^T\| \le \rho$$

Exact-Diffusion has a shorter transient stage with non-iid data

Theorem 1. (Informal) Under Assumption 1-3, Exact-Diffusion with heterogenous data will converge as follows

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = O\left(\frac{\sigma}{\sqrt{nT}} + \frac{\rho^{2/3}\sigma^{2/3}}{T^{2/3}(1-\rho)^{1/3}}\right)$$

- Recall D-SGD with heterogeneous data will converge as follows

$$\text{D-SGD:}\quad \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 = O\left(\frac{\sigma}{\sqrt{nT}} + \frac{\rho^{2/3}\sigma^{2/3}}{T^{2/3}(1-\rho)^{1/3}} + \frac{\rho^{2/3}b^{2/3}}{T^{2/3}(1-\rho)^{2/3}}\right)$$

- Exact-Diffusion eliminates the influence of data heterogeneity and shorten the transient stage

D-SGD: $O(\frac{n^3}{(1-\rho)^4})$ ⟹ Exact-Diffusion: $O(\frac{n^3}{(1-\rho)^2})$

Exact-Diffusion has a shorter transient stage with non-iid data

北京大学 PEKING UNIVERSITY

- We establish similar results in both strongly-convex and generally-convex scenarios [AY22, YAH23]

	D-SGD	Exact-Diffusion
Non-convex	$O(\frac{n^3}{(1-\rho)^4})$	$O(\frac{n^3}{(1-\rho)^2})$
Convex	$O(\frac{n^3}{(1-\rho)^4})$	$O(\frac{n^3}{(1-\rho)^2})$
Strongly convex	$O(\frac{n^3}{(1-\rho)^2})$	$O(\frac{n^3}{1-\rho})$

- Exact-Diffusion always improves the transient stage and the dependence on network topology

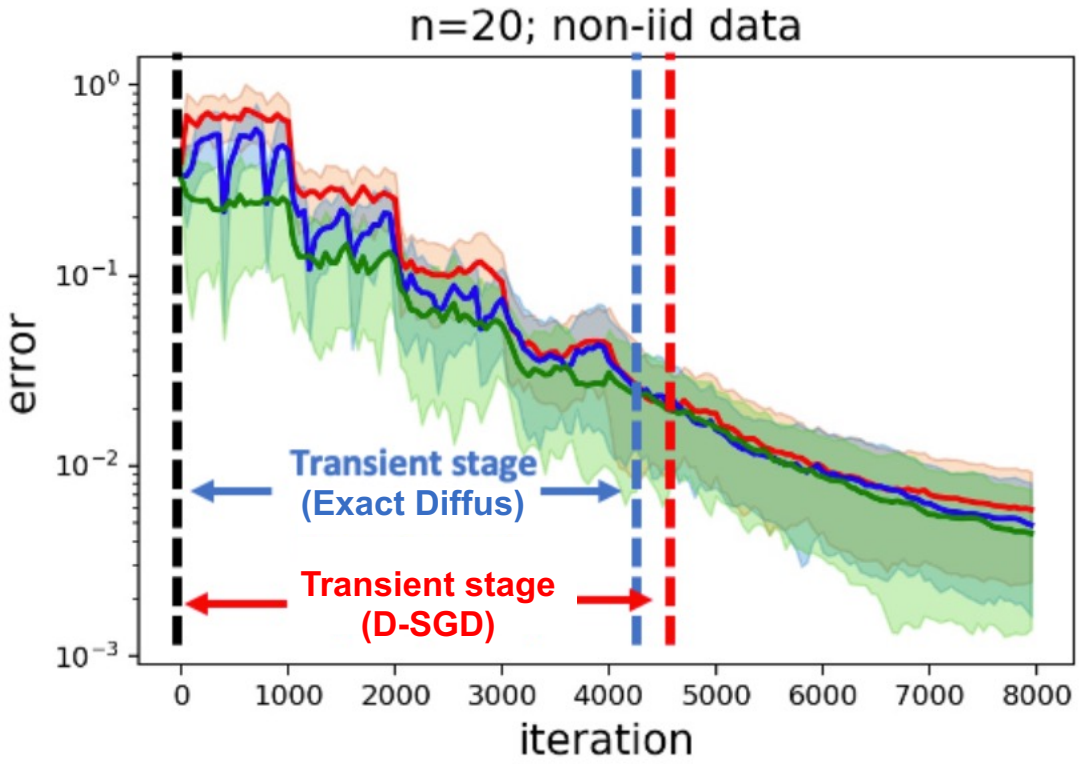
- Gradient tracking can also improve the transient stage by removing the influence of data heterogeneity

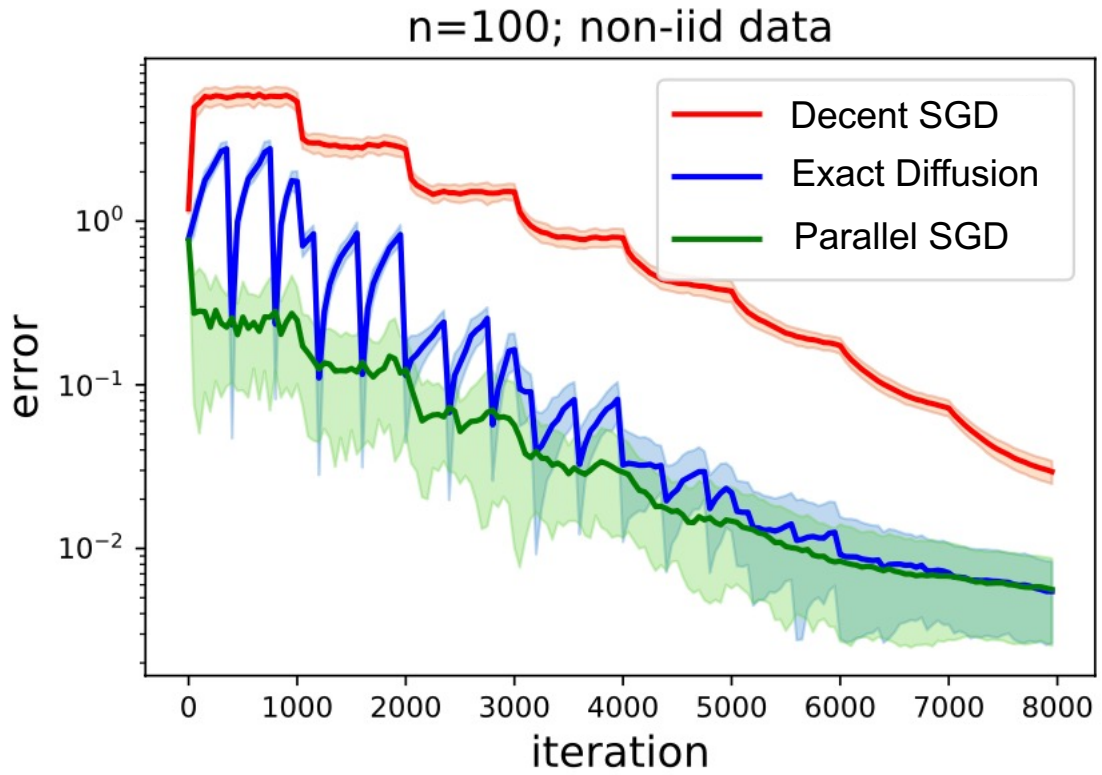
K. Yuan, S. Alghunaim, X. Huang, "Removing Data Heterogeneity Influence Enhances Network Topology Dependence of Decentralized SGD", JMLR 2023

S. Alghunaim and K. Yuan, "A Unified and Refined Convergence Analysis for Non-Convex Decentralized Learning", IEEE TSP 2022

Empirical studies

- Decentralized logistic regression with ring topology with **non-iid data**





- Exact-Diffusion has much shorter transient stage than D-SGD for sparse topology