On the Role of Data Homogeneity in Multi-Agent Non-convex Stochastic Optimization

Qiang Li, Hoi-To Wai

Dept of System Engineering and Engineering Management,
The Chinese University of Hong Kong

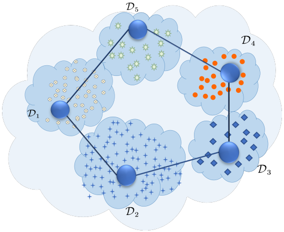
November 5, 2024
IEEE CDC 2022, Cancun, Mexico

Multi-agent Stochastic Optimization

- Consider tackling the optimization problem on a network with n agents:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \; f(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta}), \tag{1}$$

  - **Applications**: decentralized ML, control, etc.
  - $f_i(\boldsymbol{\theta}) = \mathbb{E}_{Z_i\sim \mathsf{B}_i}[\ell(\boldsymbol{\theta}; Z_i)]$ is a smooth (possibly non-convex) obj. function of agent i.
  - B_i is the data distribution at the ith agent.



- Algorithms: **decentralized stochastic gradient (DSGD)** [Sundhar Ram et al., 2010], GT-HSGD [Xin et al., 2021], D^2 [Tang et al., 2018], GNSD [Lu et al., 2019], many others ...

Decentralized SGD

Let $\boldsymbol{W}$ be a doubly stochastic matrix, the DSGD does

$$\boldsymbol{\theta}_i^{t+1} = \underbrace{\sum_{j=1}^{n} W_{ij}\boldsymbol{\theta}_j^t}_{\text{Consensus}} - \underbrace{\gamma_{t+1}\nabla\ell(\boldsymbol{\theta}_i^t; Z_i^{t+1})}_{\text{Local Update}}, \; i \in [n] \tag{2}$$

- Across the network, it uses n samples per iteration – $Z_i^{t+1} \sim \mathsf{B}_i$.

- [Lian et al., 2017] showed DSGD can achieve **linear speedup** – its performance approaches SGD with large batch, e.g.,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \gamma_{t+1}(1/n)\sum_{i=1}^{n} \nabla\ell(\boldsymbol{\theta}^t; Z_i^{t+1}) \longleftarrow \text{batch size } n$$

- This speedup only holds **asymptotically** when $t \to \infty$.

- **Transient time (informal)** $:=$ min. no. of iterations required such that DSGD can achieve comparable performance as CSGD.

Standard Assumptions

A1. Mixing matrix $\boldsymbol{W}$

Doubly stochastic, $\boldsymbol{W}\mathbf{1} = \boldsymbol{W}^\top \mathbf{1} = \mathbf{1}$. $\exists \rho \in (0, 1]$ and a projection matrix $\boldsymbol{U} \in \mathbb{R}^{n \times (n-1)}$ such that $\left\| \boldsymbol{U}^\top \boldsymbol{W} \boldsymbol{U} \right\|_2 \leq 1 - \rho$.

A2. L-Lipschitz continuous gradient

$$\|\nabla f_i(\boldsymbol{\theta}') - \nabla f_i(\boldsymbol{\theta})\| \leq L \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|, \; \forall \; \boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^d.$$

A3. Bounded variance σ

$$\mathbb{E}_{z_i \sim \mathsf{B}_i} [\|\nabla \ell(\boldsymbol{\theta}; z_i) - \nabla f_i(\boldsymbol{\theta})\|^2] \leq \sigma^2.$$

A4. Data Heterogeneity ς

$$\|\nabla f(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta})\| \leq \varsigma, \; \forall \; \boldsymbol{\theta} \in \mathbb{R}^d.$$

Convergence of Plain DSGD

Theorem 1 (Basic Result) [Lian et al., 2017]

Under A1–4, assume γ_t is sufficiently small, denote $\mathsf{D} := f(\overline{\boldsymbol{\theta}}^0) - f^\star$. For any $T \geq 1$, it holds

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_{t+1} \|\nabla f(\overline{\boldsymbol{\theta}}^t)\|^2\right] \lesssim \mathsf{D} + \frac{L\sigma^2}{n} \sum_{t=0}^{T-1} \gamma_{t+1}^2 + \frac{L^2(\varsigma^2+\sigma^2)}{\rho^2} \sum_{t=0}^{T-1} \gamma_{t+1}^3.$$

- For $\gamma_{t+1} = 1/\sqrt{T}$, let T be chosen uniformly from $\{0, \ldots, T-1\}$,

$$\mathbb{E}\left[\|\nabla f(\overline{\boldsymbol{\theta}}^{\mathsf{T}})\|^2\right] = \mathcal{O}\Big(\underbrace{(\mathsf{D} + L\sigma^2/n) T^{-1/2}}_{\propto \mathsf{D} + n^{-1} L\sigma^2 \ \text{CSGD term}} + \underbrace{\frac{L^2(\varsigma^2+\sigma^2)}{\rho^2} T^{-1}}_{\text{network depen.}} \Big)$$

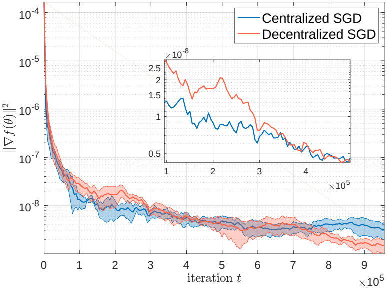
- *Transient time:* $T_{\text{trans}} = \Theta\left(n^2/\rho^4\right)$ – undesirable for large scale network and sparse network[1].

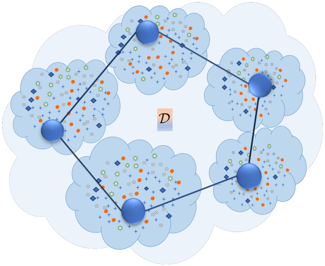
- *Remedy*: sophisticated algorithms, e.g., with gradient tracking, variance reduction, etc. [Lu et al., 2019, Huang and Pu, 2022] – is it necessary?

[1] E.g.: Ring graph: $\rho = \Theta(1/n^2)$, 2d-torus graph: $\rho = \Theta(1/n)$.

Observation

- DSGD sometimes performs almost as good as centralized SGD. Why?





- **Possible Reason**: homogeneous data (with $\mathrm{B}_i \approx \mathrm{B}_j$) are common in applications.
- Previous analysis (Theorem 1) does not take this into account.

Motivating Example

Question: Can DSGD (with homo. data) achieve fast convergence with a shorter transient time ?

- Consider a special case of (1),

$$f_i(\boldsymbol{\theta}) = (1/2)\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{b}, \tag{3}$$

  where $\boldsymbol{A}$ is PD, $\boldsymbol{b}$ is fixed vector (shared among agents).

- $\nabla f(\boldsymbol{\theta}) = \nabla f_i(\boldsymbol{\theta}) \Rightarrow \varsigma = 0 \longleftarrow$ Homogeneous data.
- Consider stochastic gradient map: $z_i \equiv \tilde{\boldsymbol{b}}_i \sim \mathsf{B}_i \equiv \mathsf{B}$ satisfies

$$\nabla \ell(\boldsymbol{\theta}; z_i) = \boldsymbol{A}\boldsymbol{\theta} + \widetilde{\boldsymbol{b}}_i, \quad \mathbb{E}[\widetilde{\boldsymbol{b}}_i] = \boldsymbol{b}, \quad \mathbb{E}[\|\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}\|^2] \leq \sigma^2 \tag{4}$$

$$\Rightarrow \mathbb{E}[\|\nabla \ell(\boldsymbol{\theta}; z_i) - \nabla f_i(\boldsymbol{\theta})\|^2] \leq \sigma^2 \Rightarrow \text{A3 } \checkmark$$

- **Note:** agents still draw independent and different samples.

Motivating Example

- Consider a special case of (1),

$$f_i(\boldsymbol{\theta}) = (1/2)\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{b}, \tag{3}$$

  where $\boldsymbol{A}$ is PD, $\boldsymbol{b}$ is fixed vector (shared among agents).

- The averaged iterate recursion of DSGD is:

$$\overline{\boldsymbol{\theta}}^{t+1} = \overline{\boldsymbol{\theta}}^{t} - \gamma_{t+1}\Big(\underbrace{\boldsymbol{A}\overline{\boldsymbol{\theta}}^{t} + \textstyle\sum_{i=1}^{n} \widetilde{\boldsymbol{b}}_i / n}_{\text{unbiased estimate of } \nabla f(\overline{\boldsymbol{\theta}}^{t})} \Big)$$

  variance: $\mathbb{E}[\|\boldsymbol{A}\overline{\boldsymbol{\theta}}^{t} + n^{-1}\sum_{i=1}^{n} \widetilde{\boldsymbol{b}}_i - \nabla f(\overline{\boldsymbol{\theta}}^{t})\|^2] \leq n^{-1}\sigma^2$.

- The above is identical to running CSGD with n samples per iter.
- **Transient time**: 0.

Does the observation generalize to nonlinear function?

Additional Assumptions

A5 Lipschitz continuous Hessian (High Order Smoothness)

$$\|\nabla^2 f_i(\boldsymbol{\theta}') - \nabla^2 f_i(\boldsymbol{\theta})\| \leq L_H \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|, \ \forall\ \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d.$$

A6 High-order heterogeneity ς_H

$$\|\nabla^2 f(\boldsymbol{\theta}) - \nabla^2 f_i(\boldsymbol{\theta})\| \leq \varsigma_H, \ \forall\ \boldsymbol{\theta} \in \mathbb{R}^d.$$

A7 Unbiased gradient & 4^{th}-order moment bound

$$\mathbb{E}_{z \sim \mathsf{B}_i}[\|\nabla \ell(\boldsymbol{\theta}; z) - \nabla f_i(\boldsymbol{\theta})\|^4] \leq \sigma^4.$$

- Note that $\varsigma = 0 \Longrightarrow \varsigma_H = 0$.
- Our notion of data homogeneity **only** requires $\varsigma_H \approx 0$ — quadratic (or higher order) terms of f_i, f to be similar.

Main Theorem

Theorem 2 (Our Bound)

Under A1–7. Assume $\{\gamma_t\}_{t\geq 1}$ is suff. small. For any $T \geq 1$, it holds

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_{t+1} \|\nabla f(\overline{\boldsymbol{\theta}}^t)\|^2\right] \lesssim \mathsf{D} + \frac{L\sigma^2}{n} \sum_{t=0}^{T-1} \gamma_{t+1}^2 + \frac{\varsigma_H^2(\varsigma^2+\sigma^2)}{\rho^2} \sum_{t=0}^{T-1} \gamma_{t+1}^3 + \frac{L_H^2}{\rho^4}(\sigma^4 + 4\varsigma^2) \sum_{t=0}^{T-1} \gamma_{t+1}^5 \quad (4)$$

- Set $\gamma_{t+1} = 1/\sqrt{T}$ and T be chosen uniformly in $\{0, \ldots, T-1\}$. Suppose that $\varsigma_H = 0$, it holds

$$\mathbb{E}\left[\|\nabla f(\overline{\boldsymbol{\theta}}^{\mathsf{T}})\|^2\right] = \mathcal{O}\Big(\underbrace{(\mathsf{D} + L\sigma^2/n)\, T^{-1/2}}_{\propto \mathsf{D} + n^{-1}L\sigma^2, \text{CSGD term}} + \underbrace{\frac{L_H^2(\sigma^4+\varsigma^4)}{\rho^4/n} T^{-2}}_{\text{network depen. term}} \Big)$$

- Transient time is now $T = \Theta\left(\frac{n^{4/3}}{\rho^{8/3}}\right)$. If $\varsigma_H \approx 0$, the above still holds approximately.

More on the Main Theorem

- **Improved transient time for** DSGD:

$$\Theta\left(\frac{n^2}{\rho^4}\right) \quad \longrightarrow \quad \Theta\left(\frac{n^{1.333}}{\rho^{2.667}}\right)$$

[Lian et al., 2017] Our analysis

Significant improvement when $n \gg 1, \rho \ll 1$.

- **Main technique:** Approximation error of gradient map ∇f_i is:

$$\mathcal{E}_i(\boldsymbol{\theta}'; \boldsymbol{\theta}) := \nabla f_i(\boldsymbol{\theta}') - \nabla f_i(\boldsymbol{\theta}) - \nabla^2 f_i(\boldsymbol{\theta})(\boldsymbol{\theta}' - \boldsymbol{\theta}).$$

Under A5, it holds that the following quadratic bound,

$$\|\mathcal{E}_i(\boldsymbol{\theta}'; \boldsymbol{\theta})\| \leq \frac{L_H}{2} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2, \ \forall \ \boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^d.$$

rather than applying Lip-gradient to obtain a linear bound.

Simulation Setup

- **Task**: binary classification using SVM.
- **Loss**: a non-convex sigmoid function on a 12-agents ring graph, where $\boldsymbol{W}_{ii} = 0.9$.

$$\ell(\boldsymbol{\theta}; z) = \frac{1}{1 + \exp(y\langle x \,|\, \boldsymbol{\theta}\rangle)} + \frac{\beta}{2} \|\boldsymbol{\theta}\|^2 ,$$

- **Synthetic Dataset**: with different ground truth $\boldsymbol{\theta}_{\circ,i}$, generate

$$x_j^i \sim \mathcal{U}[-1,1]^5, \quad y_j^i = \operatorname{sign}(\langle x_j^i \,|\, \boldsymbol{\theta}_{\circ,i} \rangle).$$

- **Benchmarks**: CSGD, DSGD with homogeneous data (Homo-DSGD) and heterogeneous data (Hete-DSGD).

Simulation Result

- **Observation:** DSGD always approach the same steady state convergence behavior as CSGD as $t \to \infty \implies$ Theorem 1 ✓.
- With **homogeneous data**, DSGD matches the performance of CSGD with a much smaller *transient time* than the case with heterogeneous data. $\implies$ Theorem 2 ✓

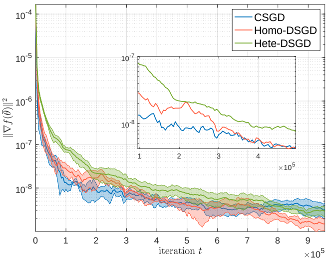


Figure 1: Compare the norm of gradient $\|\nabla f(\overline{\boldsymbol{\theta}}^t)\|^2$ against the number of iteration t. The shaded region indicate the 90% confidence interval.

Conclusion

- Plain DSGD algorithm still achieves fast convergence when the data distribution across agents are similar to each other.
- **Key Obs.:** Exploiting high-order smoothness gives tightened result.
- Our theoretical results are supported by numerical experiment.
- **Limitation/ongoing work**: the speedup happens only with $\overline{\boldsymbol{\theta}}^t$ instead of the local variables $\boldsymbol{\theta}_i^t$.

Questions & Comments?

References I

Huang, K. and Pu, S. (2022).
Improving the transient times for distributed stochastic gradient methods.
IEEE Transactions on Automatic Control.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017).
Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent.
In *NeurIPS*.

Lu, S., Zhang, X., Sun, H., and Hong, M. (2019).
Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization.
In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321. IEEE.

Sundhar Ram, S., Nedić, A., and Veeravalli, V. V. (2010).
Distributed stochastic subgradient projection algorithms for convex optimization.
JOTA, 147(3):516–545.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. (2018).
d^2: Decentralized training over decentralized data.
In *International Conference on Machine Learning*, pages 4848–4856. PMLR.

References II

Xin, R., Khan, U., and Kar, S. (2021).
A hybrid variance-reduced method for decentralized stochastic non-convex optimization.
In *ICML*.