Decentralized SGD and Average-direction SAM are Asymptotically Equivalent

Embracing Decentralization for Improved Communication Efficiency, Privacy, and Generalization

Tongtian Zhu[1], Fengxiang He[2, 3] ✉, Kaixuan Chen[1], Mingli Song[1], Dacheng Tao[4]

1 Zhejiang University, 2 JD Explore Academy, JD.com, Inc,
3 AIAI, School of Informatics, University of Edinburgh, 4 The University of Sydney

Paper

Home page

Contact Information

Fengxiang He:
F.He@ed.ac.uk
Tongtian Zhu:
raiden@zju.edu.cn

Is it possible to improve communication efficiency, privacy, and generalizability all at once 🤔?

Our paper shows that decentralized training might be the answer!

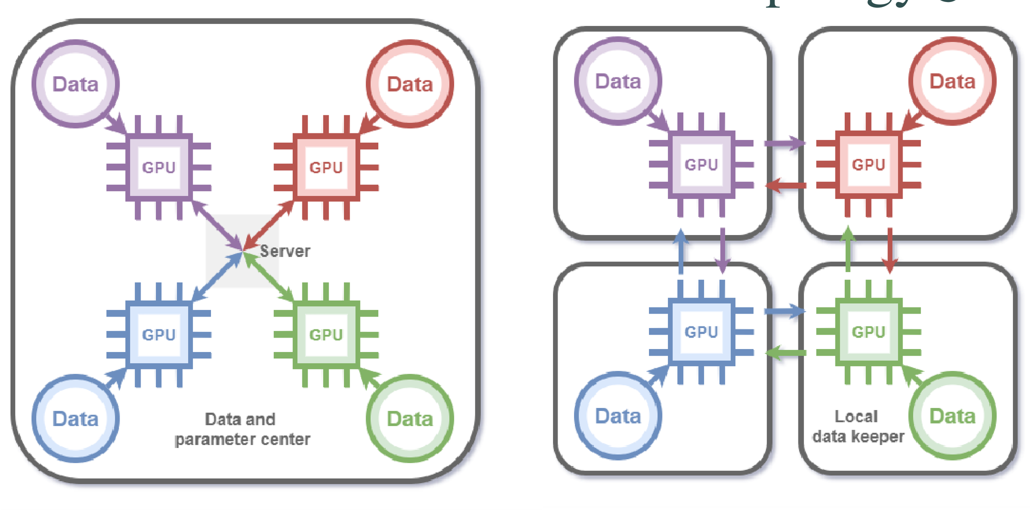
Problem

Training objective: $\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{z_j\sim\tilde{\mathcal{D}}_j}[L(\mathbf{w}; z_j)]$

$$\text{Centralized SGD: } \mathbf{w}_{a(t+1)} = \mathbf{w}_{a(t)} - \eta \underbrace{\frac{1}{m}\sum_{j=1}^{m} \cdot \overbrace{\nabla L^{\mu_j(t)}(\mathbf{w}_{a(t)})}^{\text{gradient computation}}}_{\text{average gradients on server}} .$$

$$\text{Decentralized SGD: } \mathbf{w}_{j(t+1)} = \overbrace{\sum_{k=1}^{m} \mathbf{P}_{j,k}\mathbf{w}_{k(t)}}^{\text{communication}} - \eta \cdot \overbrace{\nabla L^{\mu_j(t)}(\mathbf{w}_{j(t)})}^{\text{gradient computation}} ,$$

where $\mathbf{P}$ characterizes the communication topology $\mathcal{G}$.



Central training

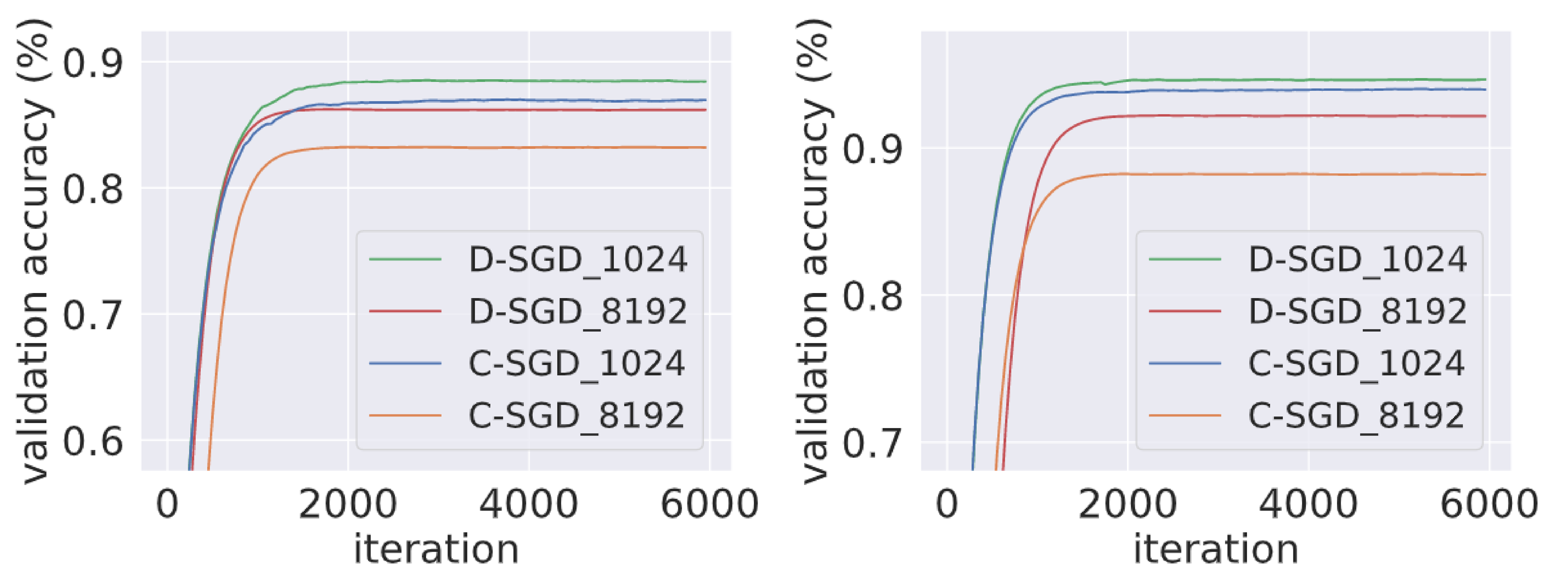
Decentralized training

Research gap

Bad news: Existing theories claim that decentralization invariably undermines generalization.

$$\text{Generalization error} \leq \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) + \text{extra error from decentralization.}$$

Some phenomena in decentralized learning are **not well explained**! D-SGD can **generalize better** than SGD in large-batch scenarios.



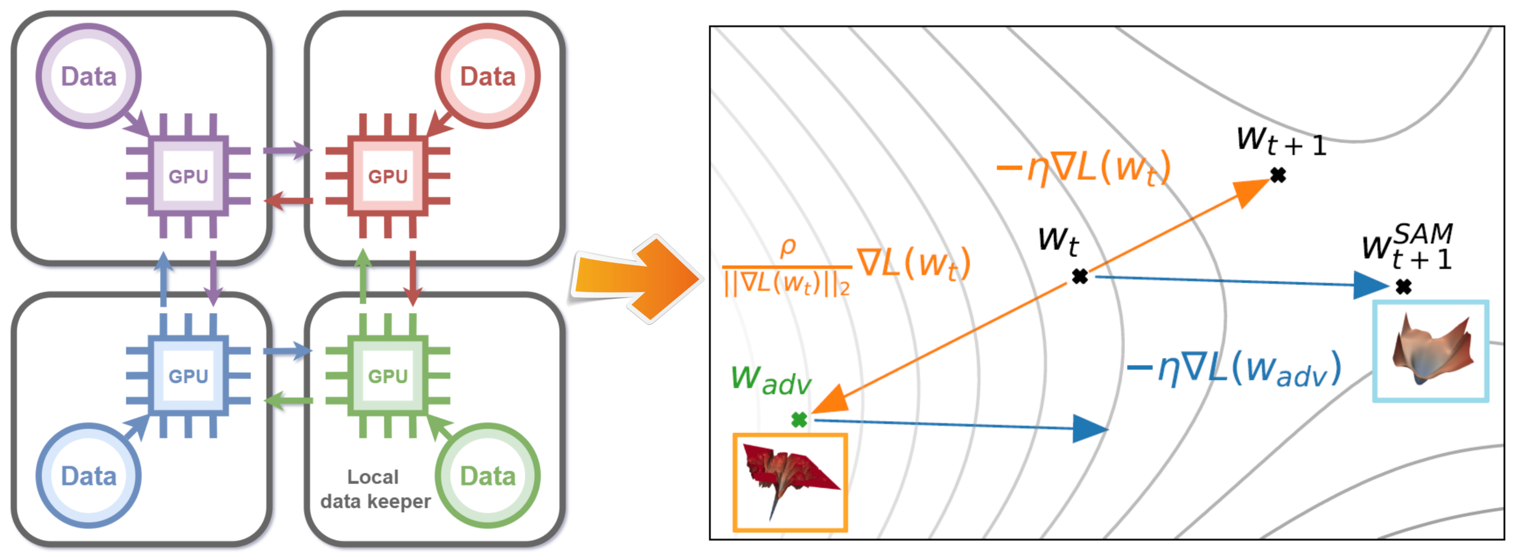
(a) AlexNet (b) ResNet-18

Non-negligible **gap** between theory and experiments exists. Important characteristics of decentralization might be underexamined!

Question: *what are the* ***inductive biases*** *of decentralization?*

🚨 Main Results 🚨

⭐ Decentralized SGD "magically" performs **sharpness-aware minimization** in an implicit way.



Decentralized training with D-SGD

Sharpness-aware Minimization

Main theorem. Given the objective L is continuous and has fourth-order partial derivatives. The mean iterate of the global averaged model of D-SGD can be written as follows:

$$\mathbb{E}_{\mu(t)}[\mathbf{w}_{a(t+1)}] = \mathbf{w}_{a(t)} - \eta \underbrace{\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Xi(t))}[\nabla L_{\mathbf{w}_{a(t)}+\epsilon}]}_{\text{asymptotic descent direction}}$$
$$+ \underbrace{\mathcal{O}\Big(\eta\, \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Xi(t))}\|\epsilon\|_2^3 + \frac{\eta}{m}\sum_{j=1}^{m}\|\mathbf{w}_{j(t)} - \mathbf{w}_{a(t)}\|_2^3\Big)}_{\text{higher-order residual terms}},$$

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{w}_{j(t)} - \mathbf{w}_{a(t)})(\mathbf{w}_{j(t)} - \mathbf{w}_{a(t)})^{\top}$.

- **Sharpness regularization**.

$$\mathbb{E}_{\mu(t)}[L_{\mathbf{w}}^{\text{D-SGD}}] \approx \underbrace{L_{\mathbf{w}}}_{\textit{original loss}} + \underbrace{\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Xi(t))}[L_{\mathbf{w}+\epsilon} - L_{\mathbf{w}}]}_{\textit{sharpness-aware regularizer}}$$

- **Regularization-optimization trade-off**.

consensus distance $\uparrow \Rightarrow$ sharpness regularization $\uparrow$ optimization $\downarrow$
consensus distance $\downarrow \Rightarrow$ sharpness regularization $\downarrow$ optimization $\uparrow$

- $\Xi(t)$, the empirical covariance matrix of $\mathbf{w}_{j(t)}$, implicitly estimate Σ_q, the intractable posterior covariance of weights,

$$\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{w}_{j(t)} - \mathbf{w}_{a(t)})(\mathbf{w}_{j(t)} - \mathbf{w}_{a(t)})^{\top} \approx \Sigma_q.$$

Proof idea

- **D-SGD iterate $\Rightarrow$ SGD iterate + noise**:

$$\underbrace{\mathbf{w}_{a(t+1)} = \mathbf{w}_{a(t)} - \eta\nabla L^{\mu(t)}_{\mathbf{w}_{a(t)}}}_{\textit{SGD iterate}} + \eta \underbrace{\frac{1}{m}\sum_{j=1}^{m}\left(\nabla L^{\mu_j(t)}_{\mathbf{w}_{j(t)}} - \nabla L^{\mu_j(t)}_{\mathbf{w}_{a(t)}}\right)}_{\textit{noise form decentralization}} .$$

- Characterize the unique noise in decentralization via a high order Taylor expansion on $\frac{1}{m}\sum_{j=1}^{m}\nabla L^{\mu_j(t)}_{\mathbf{w}_{j(t)}}$ around $\mathbf{w}_{a(t)}$.