# Decentralize to Generalize? 🤔

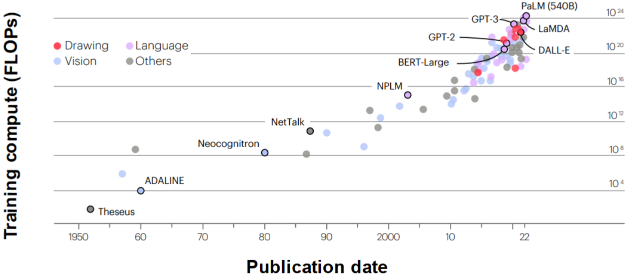
## On the Asymptotic Equivalence of Decentralized SGD and Average-direction SAM

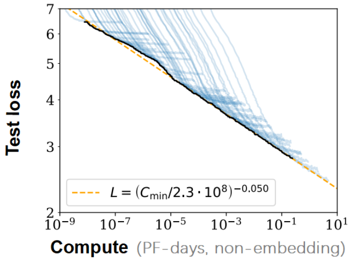
**Tongtian Zhu**[1], **Fengxiang He**[2,3], **Kaixuan Cheng**[1], **Mingli Song**[1], **Dacheng Tao**[4]

1 Zhejiang University, 2 University of Edinburgh, 3 JD Explore Academy, 4 The University of Sydney

# Overview

# Deep Learning is in Hunger

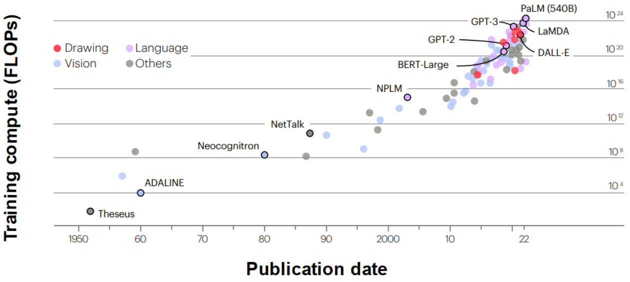


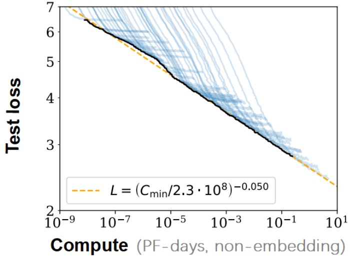




- Deep learning models are increasingly "hungry" for computing power.

"*Compute Trends Across Three Aras of Machine Learning*." Sevilla et al., arXiv, 2022.
"*Huge 'Foundation Models' Are Turbo-charging AI Progress*." The Economist, Jun 11th, 2022.
"*Scaling Laws for Neural Language Models*." Kaplan et al., arXiv, 2020.

# Deep Learning is in Hunger



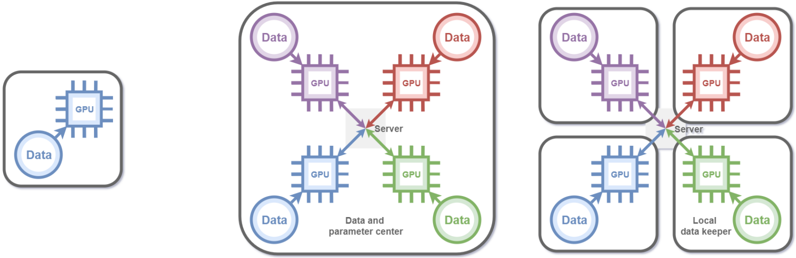




*Question: How to aggregate computing power?*

"*Compute Trends Across Three Aras of Machine Learning*." Sevilla et al., arXiv, 2022.
"*Huge 'Foundation Models' Are Turbo-charging AI Progress*." The Economist, Jun 11th, 2022.
"*Scaling Laws for Neural Language Models*." Kaplan et al., arXiv, 2020.

# Aggregate Computing Power

## Distributed training




**Single-device training**

Training without communication

**Central training**

Data centralization + Model parallelism

**Federated training**

Data decentralization

**Server-based distributed training**

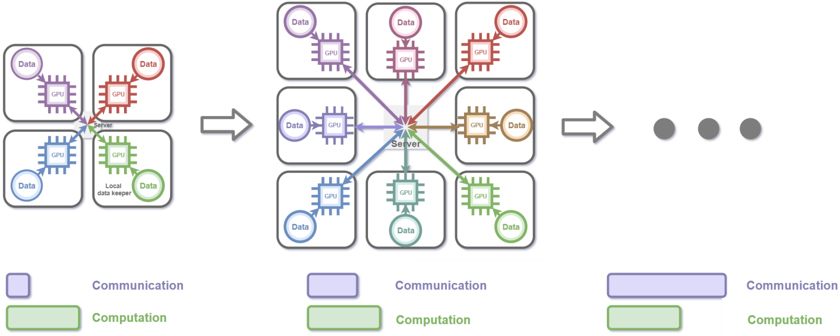
# Aggregate Computing Power

**Distributed training**

*Question: Are there any limitations to server-based distributed training?*

# Limitations of Server-based Training

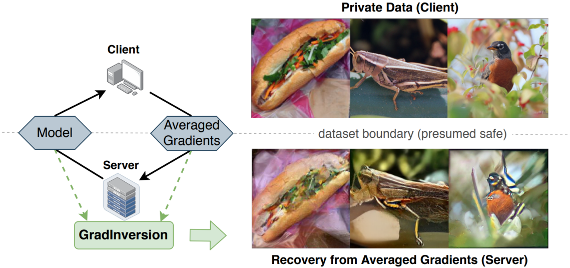
## 1 Communication bottleneck




- As the number of workers increases, communication time gradually dominates total training time.

# Limitations of Server-based Training

## 2 Privacy and security issues




(a) Inverting averaged gradients to recover original image batches

- Inverting averaged gradients on server can recover original image batches.

"*See through Gradients: Image Batch Recovery via Gradient Inversion*." Yin et al., CVPR, 2021.
"*Reconstructing Training Data from Model Gradient, Provably*." Wang et al., NeurIPS, 2022.

# Limitations of Server-based Training

**Distributed training**

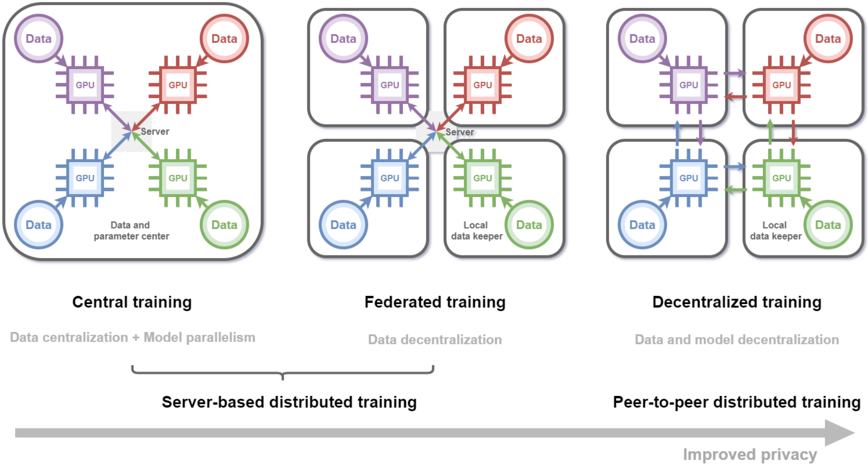
*Question: Is it possible to mitigate these limitations simultaneously?*

# Limitations of Server-based Training

## Distributed training

*Server-based communication* $\overset{?}{\Rightarrow}$ *Peer-to-peer communication*

# Possible Solution: Decentralized Training



# Possible Solution: Decentralized Training

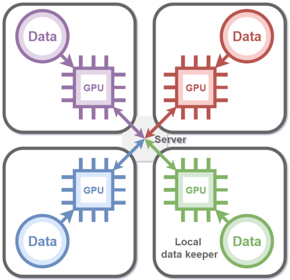

Compared with centralized training, training in a fully decentralized fashion

- avoids the requirements of a costly central server with heavy communication burdens;
- mitigates the risk of local information leakage;
- support more flexible and dynamic participation of workers;
- ...

"*Swarm Learning for Decentralized and Confidential Clinical Machine Learning*." Warnat-Herresthal et al., Nature, 2021.

# Notations

## Server-based distributed training



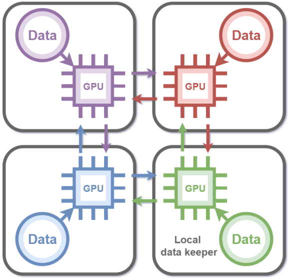

- Training objective: $\min_{w\in\mathbb{R}^d} \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{z_j\sim\tilde{\mathcal{D}}_j}[L(w;z_j)]$.
- Server-based distributed training with centralized SGD (C-SGD):

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

"*Large Scale Distributed Deep Networks.*" Dean et al., NeurIPS, 2012.
"*Communication Efficient Distributed Machine Learning with the Parameter Server.*" Li et al., NeurIPS, 2014.

# Notations

## Decentralized training



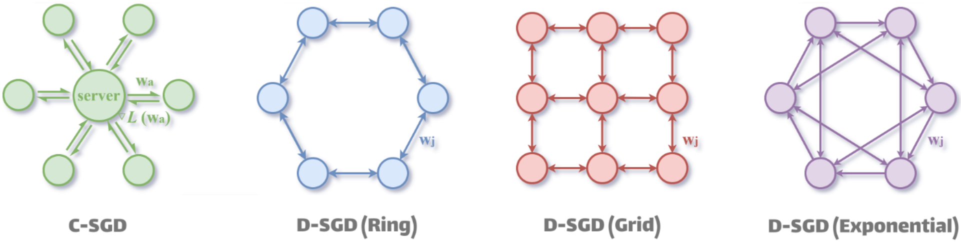

- Training objective: $\min_{w \in \mathbb{R}^d} \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}_{z_j \sim \tilde{\mathcal{D}}_j}[L(w; z_j)]$.
- Peer-to-peer distributed training with decentralized SGD (D-SGD):

$$w_{j(t+1)} = \overbrace{\sum_{j=1}^{m} P_{j,k} w_{k(t)}}^{\text{Communication}} - \eta \cdot \overbrace{\nabla L^{\mu_j(t)}(w_{j(t)})}^{\text{gradient computation}},$$

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

"*Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent.*" Lian et al., NeurIPS, 2017.

# Communication Topology in D-SGD




- Collaborations are flexible and dynamic.
- Information is exchanged only among (trusted) neighbors.

"*Topology-aware Generalization of Decentralized SGD*." Zhu et al., ICML, 2022.

# Recap


Compared with centralized training, training in a fully decentralized fashion

- avoids the requirements of a costly central server with heavy communication burdens;
- mitigates the risk of local information leakage;
- support more flexible and dynamic participation of workers.
- ...

# Motivation

**No free lunch? Are there trade-offs to the benefits of decentralization?**

# Motivation

**No free lunch? Are there trade-offs to the benefits of decentralization?**

- Bad news: Despite the aforementioned merits, it is regrettable that the existing theories claim decentralization to invariably undermines generalization. 😥

### Existing Generalization Bounds of D-SGD

$$\text{Generalization error of D-SGD} \leq \mathcal{O}\left(\frac{1}{\sqrt{\text{sample size}}}\right) + \text{additional error from decentralization.}$$

"*Stability and Generalization of Decentralized Stochastic Gradient Descent*." Sun et al., AAAI, 2021.
"*Topology-aware Generalization of Decentralized SGD*." Zhu et al., ICML, 2022.
"*Stability-Based Generalization Analysis of the Asynchronous Decentralized SGD*." Deng et al., AAAI, 2023.

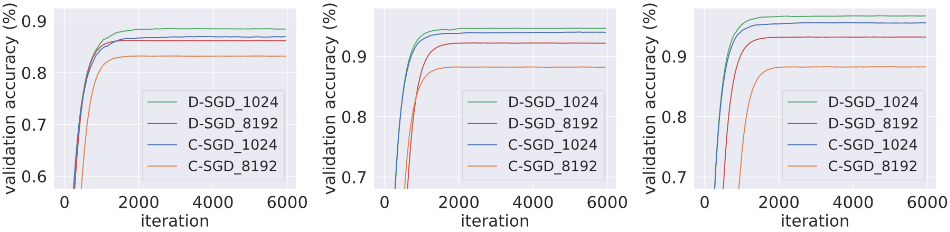
# Motivation

**Really? Some phenomena in decentralized deep learning are not well explained!**

# Motivation

**Really? Some phenomena in decentralized deep learning are not well explained!**

- D-SGD can outperform C-SGD in large-batch settings, achieving higher validation accuracy and smaller validation-training accuracy gap, despite both being fine-tuned (Zhang et al., 2021).




(a) AlexNet (b) ResNet-18 (c) DenseNet-121

"*Loss Landscape Dependent Self-Adjusting Learning Rates in Decentralized Stochastic Gradient Descent.*." Zhang et al., arXiv, 2021.

# Motivation

**Really? Some phenomena in decentralized deep learning are not well explained!** 🧐

- A non-negligible consensus distance (i.e., a measure of discrepancy between workers) at middle phases of decentralized training can improve generalization over centralized training (Kong et al., 2021).

Table 2: **The impact of consensus distance of different phases on generalization performance** (test top-1 accuracy) of training ResNet-20 on CIFAR-10 on ring. The All-Reduce performance for $n = 32$ and $n = 64$ are $92.82 \pm 0.27$ and $92.71 \pm 0.11$ respectively. The fine-tuned normal (w/o control) decentralized training performance for $n = 32$ and $n = 64$ are $91.74 \pm 0.15$ and $89.87 \pm 0.12$ respectively.

| target $\Xi$ / # nodes | dec-phase-1 | | | dec-phase-2 | | | dec-phase-3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\Xi_{\max}$ | 1/2 $\Xi_{\max}$ | 1/4 $\Xi_{\max}$ | $\Xi_{\max}$ | 1/2 $\Xi_{\max}$ | 1/4 $\Xi_{\max}$ | $\Xi_{\max}$ | 1/2 $\Xi_{\max}$ | 1/4 $\Xi_{\max}$ |
| n=32 | $91.78 \pm 0.35$ | $92.36 \pm 0.21$ | $92.74 \pm 0.10$ | $93.04 \pm 0.01$ | $92.99 \pm 0.30$ | $92.87 \pm 0.11$ | $92.60 \pm 0.00$ | $92.82 \pm 0.21$ | $92.85 \pm 0.24$ |
| n=64 | $90.31 \pm 0.12$ | $92.18 \pm 0.07$ | $92.45 \pm 0.17$ | $93.14 \pm 0.04$ | $92.94 \pm 0.10$ | $92.79 \pm 0.07$ | $92.23 \pm 0.12$ | $92.50 \pm 0.09$ | $92.60 \pm 0.10$ |

"*Consensus Control for Decentralized Deep Learning*." Kong et al., ICML, 2021.

# Motivation

**Really? Some phenomena in decentralized deep learning are not well explained!**

- A non-negligible consensus distance (i.e., a measure of discrepancy between workers) at middle phases of decentralized training can improve generalization over centralized training (Kong et al., 2021).

Table 2: **The impact of consensus distance of different phases on generalization performance** (test top-1 accuracy) of training ResNet-20 on CIFAR-10 on ring. The All-Reduce performance for $n = 32$ and $n = 64$ are $92.82 \pm 0.27$ and $92.71 \pm 0.11$ respectively. The fine-tuned normal (w/o control) decentralized training performance for $n = 32$ and $n = 64$ are $91.74 \pm 0.15$ and $89.87 \pm 0.12$ respectively.

| target $\Xi$ / # nodes | dec-phase-1 | | | dec-phase-2 | | | dec-phase-3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\Xi_{\max}$ | 1/2 $\Xi_{\max}$ | 1/4 $\Xi_{\max}$ | $\Xi_{\max}$ | 1/2 $\Xi_{\max}$ | 1/4 $\Xi_{\max}$ | $\Xi_{\max}$ | 1/2 $\Xi_{\max}$ | 1/4 $\Xi_{\max}$ |
| n=32 | $91.78 \pm 0.35$ | $92.36 \pm 0.21$ | $92.74 \pm 0.10$ | $93.04 \pm 0.01$ | $92.99 \pm 0.30$ | $92.87 \pm 0.11$ | $92.60 \pm 0.00$ | $92.82 \pm 0.21$ | $92.85 \pm 0.24$ |
| n=64 | $90.31 \pm 0.12$ | $92.18 \pm 0.07$ | $92.45 \pm 0.17$ | $93.14 \pm 0.04$ | $92.94 \pm 0.10$ | $92.79 \pm 0.07$ | $92.23 \pm 0.12$ | $92.50 \pm 0.09$ | $92.60 \pm 0.10$ |

*Takeaway: Global coherence is not always optimal.*

"*Consensus Control for Decentralized Deep Learning*." Kong et al., ICML, 2021.

# Motivation

**Really? Some phenomena in decentralized deep learning are not well explained!** 🧐

- D-SGD can outperform C-SGD in large-batch settings, achieving higher validation accuracy and smaller validation-training accuracy gap, despite both being fine-tuned (Zhang et al., 2021).
- A non-negligible consensus distance (i.e., a measure of discrepancy between workers) at middle phases of decentralized training can improve generalization over centralized training (Kong et al., 2021).

*Non-negligible gap between theory and experiments exists!*

"*Loss Landscape Dependent Self-Adjusting Learning Rates in Decentralized Stochastic Gradient Descent.*." Zhang et al., arXiv, 2021.
"*Consensus Control for Decentralized Deep Learning.*" Kong et al., ICML, 2021.

# Motivation

**Really? Some phenomena in decentralized deep learning are not well explained!** 🧐

- D-SGD can outperform C-SGD in large-batch settings, achieving higher validation accuracy and smaller validation-training accuracy gap, despite both being fine-tuned (Zhang et al., 2021).
- A non-negligible consensus distance (i.e., a measure of discrepancy between workers) at middle phases of decentralized training can improve generalization over centralized training (Kong et al., 2021).

*Non-negligible gap between theory and experiments exists!*

**Our Goal: Bridge the Gap**

*Thoroughly examine the unique, underexamined characteristics of decentralized training.*

"*Loss Landscape Dependent Self-Adjusting Learning Rates in Decentralized Stochastic Gradient Descent.*." Zhang et al., arXiv, 2021.
"*Consensus Control for Decentralized Deep Learning.*" Kong et al., ICML, 2021.

# From Gap to Solution: Our Journey

## How to Bridge the Gap?

*Understanding decentralization requires thinking its inductive bias.*

# From Gap to Solution: Our Journey

## How to Bridge the Gap?

*Understanding decentralization requires thinking its inductive bias.*

*Let us do some simple math!*

# From Gap to Solution: Our Journey

## How to Bridge the Gap?

*Understanding decentralization requires thinking its inductive bias.*

**Recall the iterate of D-SGD**: $w_{j(t+1)} = \overbrace{\sum_{j=1}^{m} P_{j,k} w_{k(t)}}^{\text{Communication}} - \eta \cdot \overbrace{\nabla L^{\mu_j(t)}(w_{j(t)})}^{\text{gradient computation}}$ .

# From Gap to Solution: Our Journey

## How to Bridge the Gap?

*Understanding decentralization requires thinking its inductive bias.*

**Recall the iterate of D-SGD**: $w_{j(t+1)} = \overbrace{\sum_{j=1}^{m} P_{j,k} w_{k(t)}}^{\text{Communication}} - \eta \cdot \overbrace{\nabla L^{\mu_j(t)}(w_{j(t)})}^{\text{gradient computation}}$ .

**What about its global average**: $w_{a(t+1)} = w_{a(t)} - \eta \cdot \dfrac{1}{m} \sum_{j=1}^{m} \nabla L^{\mu_j(t)}(w_{j(t)})$.

# From Gap to Solution: Our Journey

## How to Bridge the Gap?

*Understanding decentralization requires thinking its inductive bias.*

**Recall the iterate of D-SGD**: $w_{j(t+1)} = \overbrace{\sum_{j=1}^{m} P_{j,k} w_{k(t)}}^{\text{Communication}} - \eta \cdot \overbrace{\nabla L^{\mu_j(t)}(w_{j(t)})}^{\text{gradient computation}}$ .

**What about its global average**: $w_{a(t+1)} = w_{a(t)} - \eta \cdot \frac{1}{m} \sum_{j=1}^{m} \nabla L^{\mu_j(t)}(w_{j(t)})$.

**Rearrange**: $w_{a(t+1)} = w_{a(t)} - \eta \big[ \nabla L^{\mu(t)}_{w_a(t)} + \frac{1}{m} \sum_{j=1}^{m} (\nabla L^{\mu_j(t)}_{w_j(t)} - \nabla L^{\mu_j(t)}_{w_a(t)}) \big]$.

# From Gap to Solution: Our Journey

## How to Bridge the Gap?

*Understanding decentralization requires thinking its* ***inductive bias***.

**We obtain**: $$w_{a(t+1)} = w_{a(t)} - \eta\Big[ \underbrace{\nabla L^{\mu(t)}_{w_a(t)}}_{\text{gradient at } w_a(t)} + \underbrace{\frac{1}{m}\sum_{j=1}^{m}\big(\nabla L^{\mu_j(t)}_{w_j(t)} - \nabla L^{\mu_j(t)}_{w_a(t)}\big)}_{\text{gradient diversity among local workers}} \Big].$$

# From Gap to Solution: Our Journey

## How to Bridge the Gap?

*Understanding decentralization requires thinking its inductive bias.*

**We obtain**: $$w_{a(t+1)} = w_{a(t)} - \eta\Big[\underbrace{\nabla L^{\mu(t)}_{w_a(t)}}_{\text{gradient at } w_a(t)} + \underbrace{\frac{1}{m}\sum_{j=1}^{m}\big(\nabla L^{\mu_j(t)}_{w_j(t)} - \nabla L^{\mu_j(t)}_{w_a(t)}\big)}_{\text{gradient diversity among local workers}}\Big].$$

## Lightbulb Moment

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

# From Gap to Solution: Our Journey

## Question

*What are the inductive bias of the unique noise?*

**We obtain**: $$\underbrace{w_{a(t+1)} = w_{a(t)} - \eta \nabla L^{\mu(t)}_{w_a(t)}}_{\textit{SGD iterate}} \underbrace{+\eta \frac{1}{m}\sum_{j=1}^{m}\left(\nabla L^{\mu_j(t)}_{w_j(t)} - \nabla L^{\mu_j(t)}_{w_a(t)}\right)}_{\textit{noise form decentralization}}.$$

# From Gap to Solution: Our Journey

## Question

*What are the inductive bias of the unique noise?*

**We obtain**: $$\underbrace{w_{a(t+1)} = w_{a(t)} - \eta \nabla L^{\mu(t)}_{w_a(t)}}_{\textit{SGD iterate}} \underbrace{+\eta \frac{1}{m}\sum_{j=1}^{m}\left(\nabla L^{\mu_j(t)}_{w_j(t)} - \nabla L^{\mu_j(t)}_{w_a(t)}\right)}_{\textit{noise form decentralization}}.$$

Inspired by Gu et al. (2023), which shows that local steps in local SGD inject extra noise and drive the iterate to converge faster towards flatter minima, a natural question arises:

## A Possible Direction

*Would the noise from decentralization induces flatness bias?*

"*Why (and When) does Local SGD Generalize Better than SGD?*." Gu et al., ICLR, 2023.

# From Gap to Solution: Our Journey

**What are flat minima?**

# From Gap to Solution: Our Journey

**The flat minima hypothesis**

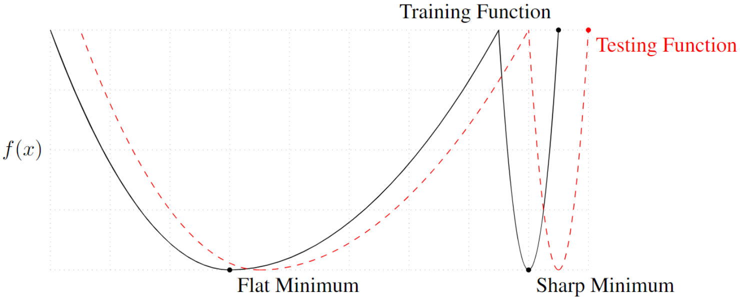
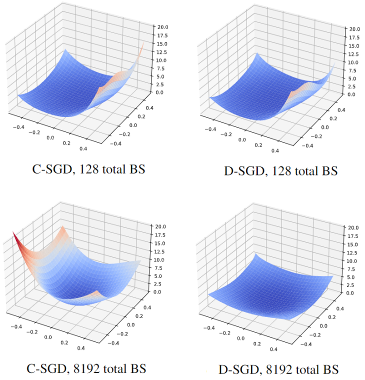



Figure: A Conceptual Sketch of Flat and Sharp Minima.

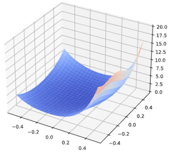
"*On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*." Keskar et al., ICLR, 2017.
"*Label Noise SGD Provably Prefers Flat Global Minimizers*." Damian et al., NeurIPS, 2021.
"*Why (and When) does Local SGD Generalize Better than SGD?*." Gu et al., ICLR, 2023.

# From Gap to Solution: Preliminary Experiments

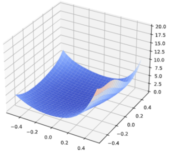



Loss landscape visualization of ResNet-18 trained on CIFAR-10 using C-SGD and D-SGD, respectively.

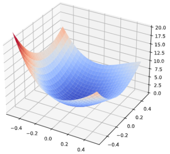
# From Gap to Solution: Preliminary Experiments



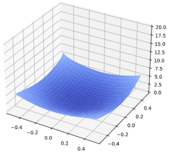
C-SGD, 128 total BS



D-SGD, 128 total BS



C-SGD, 8192 total BS

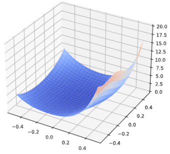


D-SGD, 8192 total BS

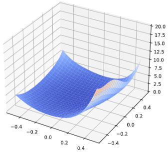
Loss landscape visualization of ResNet-18 trained on CIFAR-10 using C-SGD and D-SGD, respectively.

- The minima of D-SGD are flatter than those of C-SGD, especially in large-batch scenarios;
- The observation holds true across common topologies.

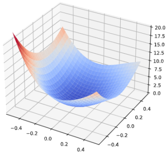
# From Gap to Solution: Preliminary Experiments



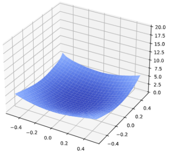
C-SGD, 128 total BS



D-SGD, 128 total BS



C-SGD, 8192 total BS



D-SGD, 8192 total BS

Loss landscape visualization of ResNet-18 trained on CIFAR-10 using C-SGD and D-SGD, respectively.

- The minima of D-SGD are flatter than those of C-SGD, especially in large-batch scenarios;
- The observation holds true across common topologies.

**Question**

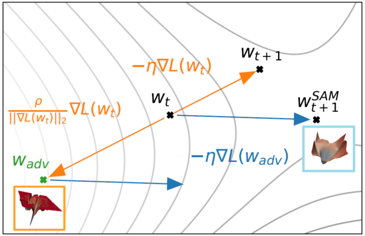
*How does decentralization (or gossip averaging) improve flatness?*

# Main Contribution

### What we find

- D-SGD and average-direction Sharpness-aware minimization (SAM) are asymptotically equivalent.


**Decentralized training with D-SGD**

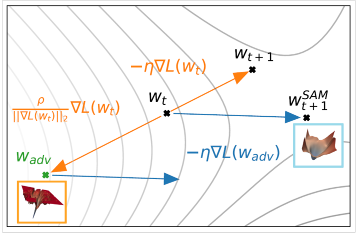



**Sharpness-aware Minimization**

---

"*Sharpness-aware Minimization for Efficiently Improving Generalization*." Foret et al., ICLR, 2021.

# Sharpness-aware Minimization




**Sharpness-aware Minimization**

Training objective: $\min_{w\in\mathbb{R}^d} \max_{\|\epsilon\|_p\leq\rho} \mathbb{E}_{z\sim\tilde{\mathcal{D}}} L(w+\epsilon; z)$.

- Foret et al. (2021) propose to use a first-order approximation to simplify the max step:

$$L^{\text{SAM}}(w) \approx \max_{\|\epsilon\|_p\leq\rho} [L(w) + \epsilon^\top \nabla L(w)].$$

- The gradient update of vanilla SAM becomes

$$\nabla L^{\text{SAM}}(w) \approx \nabla L(w+\epsilon^*) = \nabla L(w + \rho\frac{\nabla L(w)}{\|\nabla L(w)\|_2}).$$

"*Sharpness-aware Minimization for Efficiently Improving Generalization*." Foret et al., ICLR, 2021.

# Decentralized SGD as Average-direction SAM

## Main theorem (Decentralized SGD as SAM)

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:

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

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t) - w_a(t))(w_j(t) - w_a(t))^\top \in \mathbb{R}^{d\times d}$ denotes the weight diversity matrix.

# Decentralized SGD as Average-direction SAM

## Main theorem (Decentralized SGD as SAM)

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:

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

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t)-w_a(t))(w_j(t)-w_a(t))^\top \in \mathbb{R}^{d\times d}$ denotes the weight diversity matrix.

- **Asymptotic equivalence.** Note $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Xi(t))}[\nabla L_{w_a(t)+\epsilon}]$ is of the order $L_{w_a(t)} + \mathcal{O}(\frac{1}{m}\sum_{j=1}^{m}\|w_j(t)-w_a(t)\|_2^2)$ while the residuals are of the higher-order $\mathcal{O}(\frac{1}{m}\sum_{j=1}^{m}\|w_j(t)-w_a(t)\|_2^3)$. Therefore, $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Xi(t))}[\nabla L_{w_a(t)+\epsilon}]$ gradually dominates the optimization direction as the local models are near consensus (i.e., $w_j(t) \to w_a(t), \forall j$).

# Decentralized SGD as Average-direction SAM

## Main theorem (Decentralized SGD as SAM)

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:

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

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t) - w_a(t))(w_j(t) - w_a(t))^\top \in \mathbb{R}^{d\times d}$ denotes the weight diversity matrix.

- **Universality**. The theory is applicable to arbitrary communication topologies and general **non-convex and non-$\beta$-smooth** problems (e.g., deep neural networks training).

# Decentralized SGD as Average-direction SAM

## Main theorem (Decentralized SGD as SAM)

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:

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

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t) - w_a(t))(w_j(t) - w_a(t))^\top \in \mathbb{R}^{d\times d}$ denotes the weight diversity matrix.

- **Sharpness regularization.** D-SGD asymptotically optimizes $\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Xi(t))}[L_{w+\epsilon}]$, an averaged perturbed loss in a "basin" around $w$, rather than the original point-loss.

# Decentralized SGD as Average-direction SAM

## Main theorem (Decentralized SGD as SAM)

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:

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

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t) - w_a(t))(w_j(t) - w_a(t))^\top \in \mathbb{R}^{d\times d}$ denotes the weight diversity matrix.

- **Sharpness regularization.** Split "true objective" of D-SGD near consensus into the original loss plus an average-direction sharpness:

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

.

# Decentralized SGD as Average-direction SAM

## Main theorem (Decentralized SGD as SAM)

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:

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

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t) - w_a(t))(w_j(t) - w_a(t))^\top \in \mathbb{R}^{d\times d}$ denotes the weight diversity matrix.

- **Regularization-optimization trade-off.** Increasing the consensus distance enhances the sharpness of the regularization effect, but it also complicates the optimization of $\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Xi(t))}[\nabla L_{w_a(t)+\epsilon}]$, and can cause higher-order residual terms to dominate the whole optimization direction.

# Decentralized SGD as Average-direction SAM

## Main theorem (Decentralized SGD as SAM)

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:

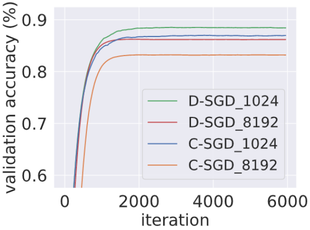
$$\mathbb{E}_{\mu(t)}[w_a(t+1)] = w_a(t) - \eta \underbrace{\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Xi(t))}[\nabla L_{w_a(t)+\epsilon}]}_{\text{asymptotic descent direction}} + \underbrace{\mathcal{O}\Big(\eta\, \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Xi(t))}\|\epsilon\|_2^3 + \frac{\eta}{m}\sum_{j=1}^{m}\|w_j(t)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

where $\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t) - w_a(t))(w_j(t) - w_a(t))^\top \in \mathbb{R}^{d\times d}$ denotes the weight diversity matrix.

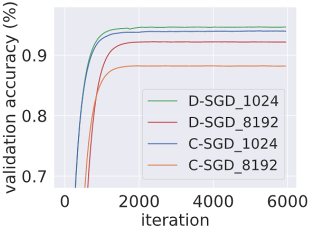
- **Variational interpretation.** D-SGD estimates uncertainty for free: The weight diversity matrix $\Xi(t)$ (i.e., the empirical covariance matrix of $w_j(t)$) implicitly estimate $\Sigma_q$, the intractable posterior covariance,

$$\Xi(t) = \frac{1}{m}\sum_{j=1}^{m}(w_j(t) - w_a(t))(w_j(t) - w_a(t))^\top \approx \Sigma_q.$$

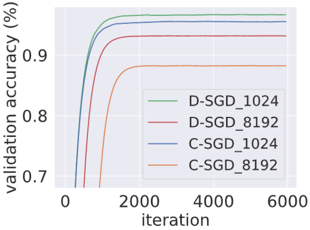
# Recap




(a) AlexNet

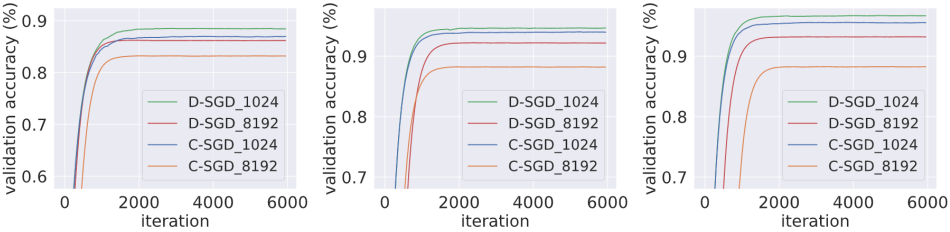



(b) ResNet-18




(c) DenseNet-121

# Recap




(a) AlexNet (b) ResNet-18 (c) DenseNet-121

## Question

Why is the generalization benefit of decentralization more significant in large-batch settings?

# Generalization Benefit in Large-batch Scenarios

## Corollary

Recall that N denotes the total training sample size and let $B = |\mu|$ denote the total batch size. With a probability greater than $1 - \mathcal{O}(\frac{B}{(N-B)\eta^2})$, D-SGD implicit minimizes

$$
\begin{aligned}
L_w^{\text{D-SGD}} = & L_w^\mu + \underbrace{\operatorname{Tr}(H_w^\mu \Xi_{(t)}) + \frac{\eta}{4}\operatorname{Tr}((H_w^\mu)^2 \Xi_{(t)})}_{\text{batch size independent sharpness regularizer}} \\
& + \underbrace{\boxed{\kappa} \cdot \frac{1}{N}\sum_{j=1}^{N}\left[\|\nabla L_w^j - \nabla L_w^\mu\|_2^2 + \operatorname{Tr}((H_w^j - H_w^\mu)^2 \Xi_{(t)})\right]}_{\text{batch size dependent variance regularizer}} + \frac{\eta}{4}\|\nabla L_w^\mu\|_2^2 + \mathcal{R}^A + \mathcal{O}(\eta^2),
\end{aligned}
$$

where $\kappa = \frac{\eta}{B} \cdot \frac{N-B}{(N-1)}$, and $\mathcal{R}^A$ absorbs all higher-order residuals.

- Compared with C-SGD, D-SGD exhibits additional batch size-independent sharpness regularization.

# The Best of All Worlds? 🤔


Compared with centralized training, training in a fully decentralized fashion

- avoids the requirements of a costly central server with heavy communication burdens;
- mitigates the risk of local information leakage;
- support more flexible and dynamic participation of workers;
- can potentially improve generalization (**this paper**).

# Discussion and Broader Impact

## Improve Convergence and Generalization Analyses

Can we utilize the connection between decentralized training and centralized training to improve the existing convergence and generalization bounds of decentralized algorithms?

# Discussion and Broader Impact

### Improve Convergence and Generalization Analyses

Can we utilize the connection between decentralized training and centralized training to improve the existing convergence and generalization bounds of decentralized algorithms?

### Bridge Decentralized Training and SAM

Does D-SGD share the properties of SAM, beyond generalizablity, including better interpretability (Andriushchenko et al., 2023) and transferability (Chen et al., 2022)?

"*When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations*." Chen et al., ICLR, 2022.
"*Sharpness-Aware Minimization Leads to Low-Rank Features*." Andriushchenko et al., HiLD Workshop, ICML, 2023.

# Summary

## Research gap

- Existing theories: Decentralization invariably undermines generalization;
- Experiments: D-SGD can generalizes better than its centralized counterpart in some scenarios.

## Main results

- D-SGD asymptotically performs sharpness-aware minimization.

## Implications

- Regularization-optimization trade-off;
- Free uncertainty evaluation mechanism;
- The sharpness regularization is batch size-independent.

## Furture work

- Bridge decentralized training and SAM.

# Reference

Andriushchenko, M., Bahri, D., Mobahi, H., and Flammarion, N. (2023). Sharpness-aware minimization leads to low-rank features. *High-dimensional Learning Dynamics Workshop in International Conference on Machine Learning*.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. (2019). Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pages 344–353. PMLR.

Chen, X., Hsieh, C.-J., and Gong, B. (2022). When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representations*.

Damian, A., Ma, T., and Lee, J. D. (2021). Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, volume 34, pages 27449–27461. Curran Associates, Inc.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M. a., Senior, A., Tucker, P., Yang, K., Le, Q., and Ng, A. (2012). Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Deng, X., Sun, T., Li, S., and Li, D. (2023). Stability-based generalization analysis of the asynchronous decentralized sgd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7340–7348.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2021). Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Gu, X., Lyu, K., Huang, L., and Arora, S. (2023). Why (and when) does local SGD generalize better than SGD? In *International Conference on Learning Representations*.

Hochreiter, S. and Schmidhuber, J. (1997). Flat minima. *Neural computation*, 9(1):1–42.

Jiang*, Y., Neyshabur*, B., Mobahi, H., Krishnan, D., and Bengio, S. (2020). Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

# Reference


Kong, L., Lin, T., Koloskova, A., Jaggi, M., and Stich, S. (2021). Consensus control for decentralized deep learning. In *International Conference on Machine Learning*, pages 5686–5696. PMLR.

Li, M., Andersen, D. G., Smola, A. J., and Yu, K. (2014). Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. (2022). Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Sun, T., Li, D., and Wang, B. (2021). Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9756–9764.

Wang, Z., Lee, J., and Lei, Q. (2023). Reconstructing training data from model gradient, provably. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6595–6612. PMLR.

Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N. A., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270.

Wei, Z., Zhu, J., and Zhang, Y. (2023). Sharpness-aware minimization alone can improve adversarial robustness. *New Frontiers in Adversarial Machine Learning Workshop in International Conference on Machine Learning*.

Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., and Molchanov, P. (2021). See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16337–16346.

Zhang, W., Liu, M., Feng, Y., Cui, X., Kingsbury, B., and Tu, Y. (2021). Loss landscape dependent self-adjusting learning rates in decentralized stochastic gradient descent. *arXiv preprint arXiv:2112.01433*.

Zhu, T., He, F., Zhang, L., Niu, Z., Song, M., and Tao, D. (2022). Topology-aware generalization of decentralized SGD. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27479–27503. PMLR.

# Thank You!

Contact: raiden@zju.edu.cn (Tongtian Zhu)

https://arxiv.org/abs/2306.02913

Paper