ICLR Oral

On the Surprising Effectiveness of Single Global Merging in Decentralized Learning

The "**Grokking Moment**" in Decentralized Learning

Tongtian Zhu*, Tianyu Zhang*, Mingze Wang, Zhanpeng Zhou✉, Can Wang

Zhejiang University, Mila / Universite de Montreal, Peking University, Shanghai Jiao Tong University

Authors and Presenter

Tongtian Zhu
Zhejiang University

Tianyu Zhang
Mila / Université de Montréal

Mingze Wang
Peking University

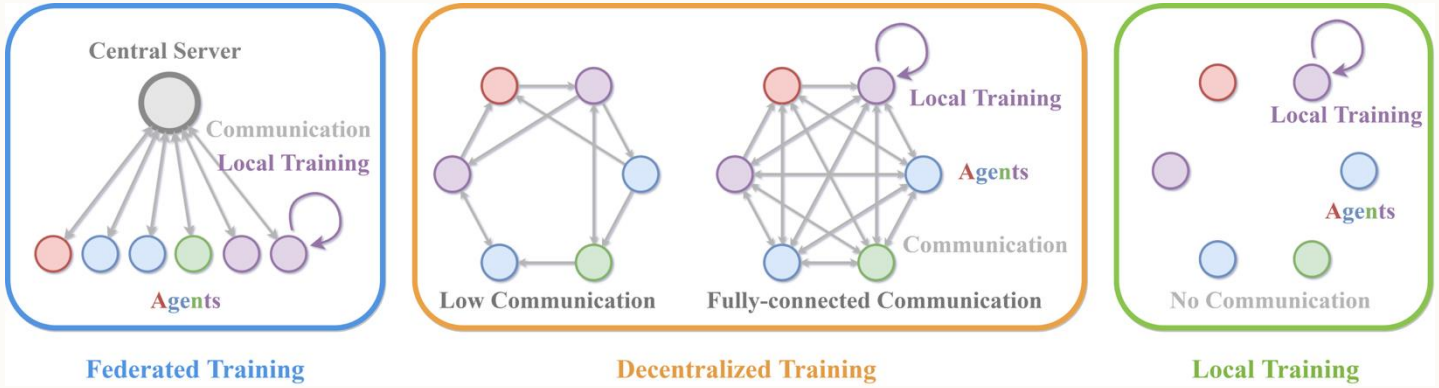
Zhanpeng Zhou
Shanghai Jiao Tong University

Can Wang
Zhejiang University

Background of Decentralized Learning

Key setup

Gossip communication: no central server
Data distribution: **Non-IID**
Evaluation metric: **global test accuracy**



Global Objective. Each node k maintains a local model θ_k. *The goal is to minimize the average objective across nodes:*

$$\min_{\theta} \mathcal{L}(\theta) = \frac{1}{m} \sum_{k \in \mathcal{V}} L_k(\theta) .$$

Local Optimization. Each node first performs a local update using its own sample $\xi_k^{(t)}$:

$$\theta_k^{(t+1/2)} := \theta_k^{(t)} - \eta g_k^{(t)}.$$

Communication means weighted averaging with neighbors through the mixing matrix $W^{(t)}$:

$$\theta_k^{(t+1)} = \sum_{l=1}^{m} W_{k,l}^{(t)} \theta_l^{(t+1/2)}.$$

Communication is determined by W: A sparse W matrix → Sparse communication

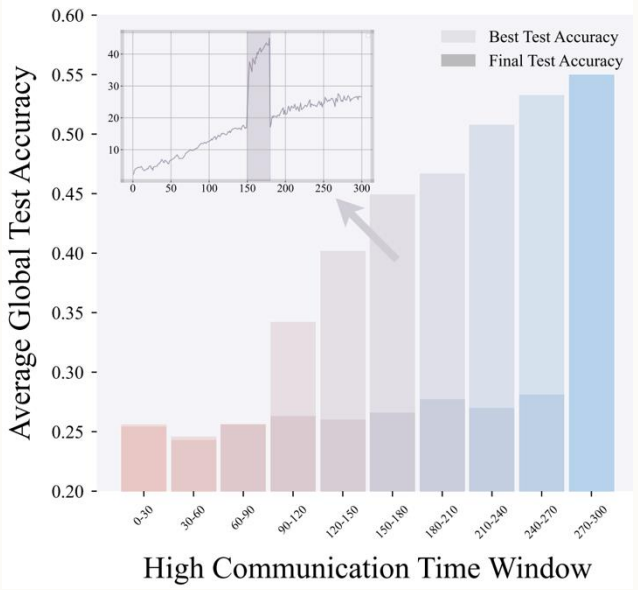
Motivation: Communication Timing

Prior work focuses on fixed W.

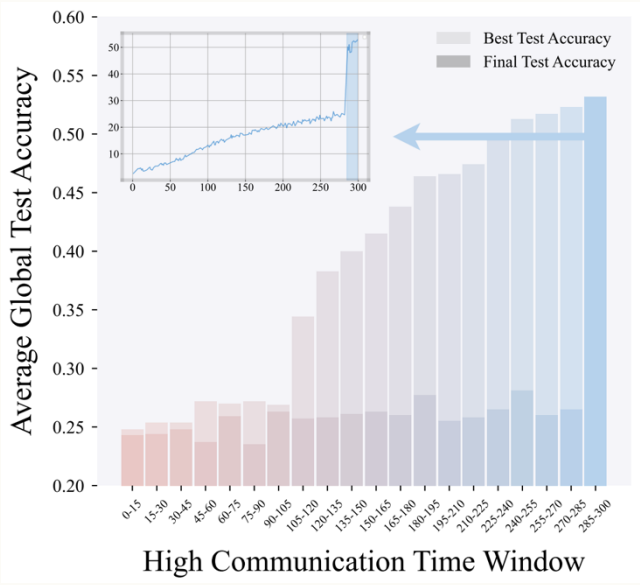
$$\theta_k^{(t+1)} = \sum_{l=1}^{m} W_{k,l}^{(t)} \, \theta_l^{(t+1/2)}$$

A sparse W matrix
→ Sparse communication

Our question: How to allocate communication over time → vary $W^{(t)}$ across training rounds.



High communication in 10% of rounds



High communication in 5% of rounds

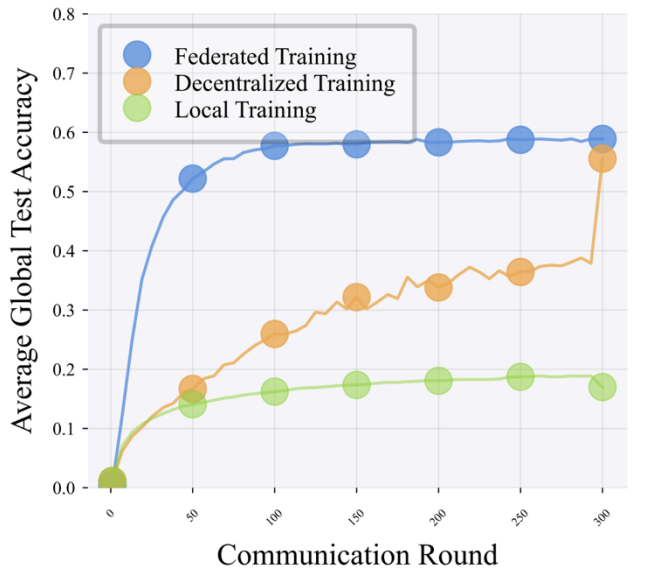
ResNet-18 on CIFAR-100

Setup: We fix the total communication budget and shift the high-communication window across training.

Key finding: Late communication consistently improves final test accuracy under the same total budget.

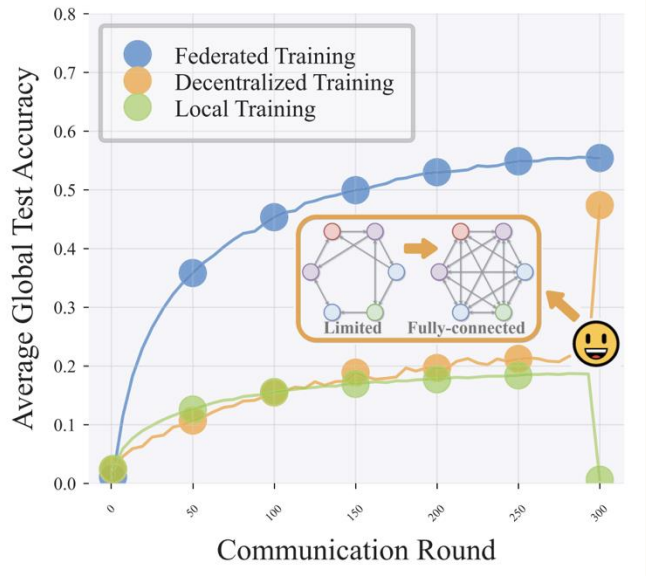
The Surprising Power of a Single Final Merging

Question: What happens if all fully connected communication is concentrated into a single final round?



CLIP ViT-B/32 on Tiny ImageNet

a sparse W matrix



ResNet-18 on Tiny ImageNet

a dense W matrix with no zero entries

Similar to grokking

$$\theta_k^{(t+1)} = \sum_{l=1}^{m} W_{k,l}^{(t)} \, \theta_l^{(t+1/2)}$$

Setup: Sparse communication is used throughout training, with one final global merging.

Key finding: Even under severe communication constraints, a single final global merging substantially improves test accuracy.

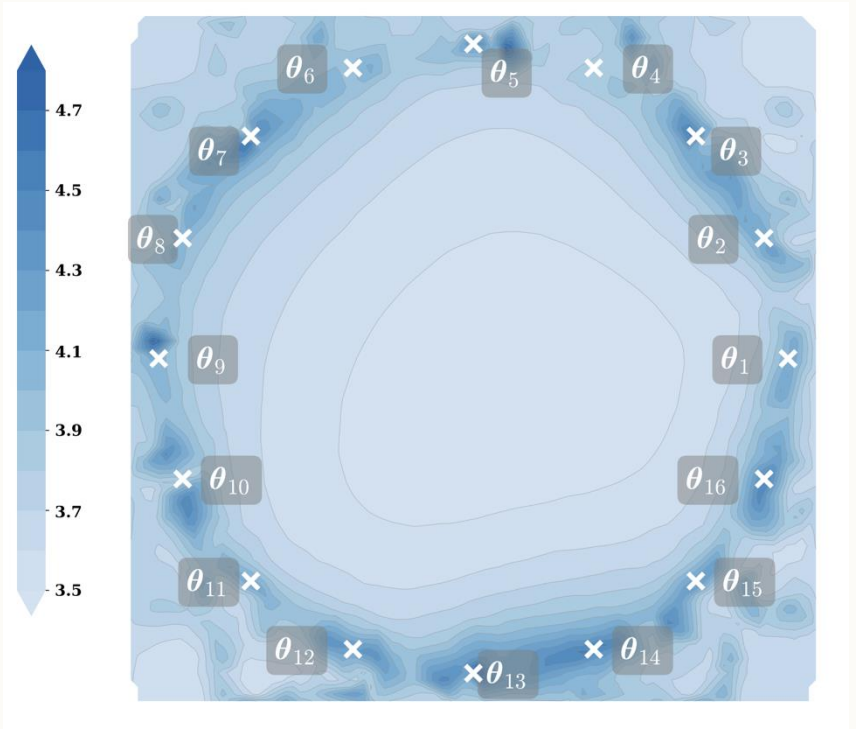
Practical implication: The total communication cost can be reduced far more aggressively than expected!

A Loss Landscape View

Diverse yet mergeable: Before the final merge, local models form a ring around a shared low-loss basin.

Implication: Mergeability maintains under sparse communication.

Mergeablity



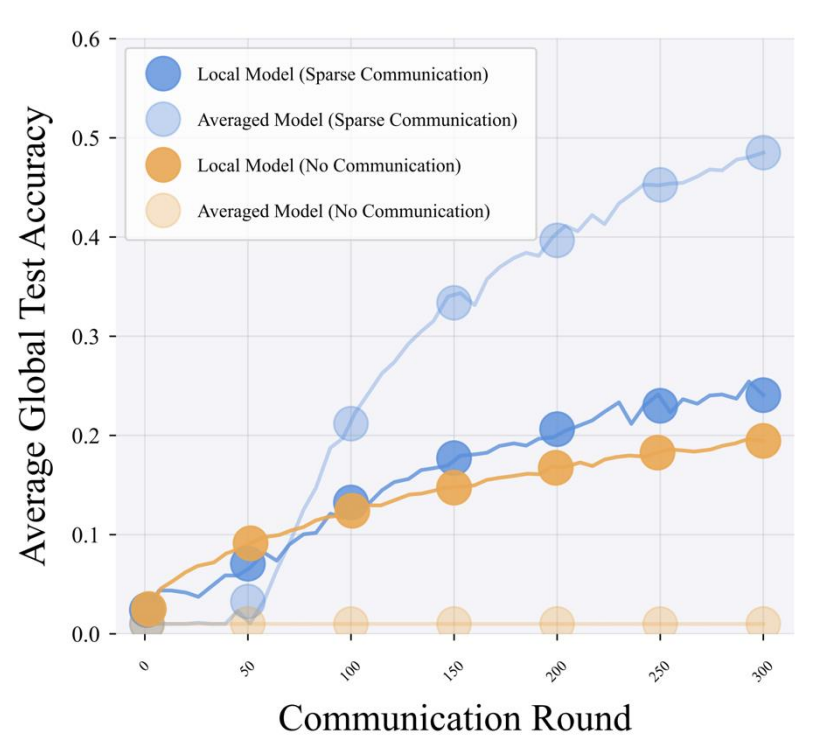
Before the final merge, local models form a ring around a central low-loss basin.

Communication is the Key to Mergeability

Question: Where does pre-merge mergeability come from?

Key findings:

- Under sparse gossip (blue), the merged model stays strong throughout training; with no communication (orange), it remains weak.
- Before the final merge, local-model accuracy under sparse gossip is close to that under no communication, yet merged performance differs dramatically



Setup: Highly non-IID decentralized training of ResNet-18 on CIFAR-100 with 13 agents.

Proposed metric (new): Counterfactual merged model: the hypothetical globally averaged model evaluated at each round for analysis only; no actual global merge is performed.

Understanding The Dynamics of Merged-Model

Question: Why Can Merged Model Work Well

Theorem 1 (Non-convex Convergence Rate of DSGD). *Suppose Assumption 2 and Assumption 3 hold. Consider decentralized SGD (DSGD) with initializations* $\theta_k^{(0)} = \theta^{(0)}$ *for all* $k \in \mathcal{V}$*, and a constant learning rate satisfying* $\eta < \frac{2}{L_2}$*. Let* $\bar{\theta}^{(t)} = \frac{1}{m}\sum_{k=1}^{m}\theta_k^{(t)}$ *denote the averaged model at the* t*-th step. To achieve an* ε*-stationary point such that* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\|_2^2\right] \le \varepsilon$*, the total number of steps* T *satisfies:*

$$T = \mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2} + \frac{1}{\varepsilon} + \sum_{t=0}^{T-1} U^{(t)}\right) \cdot \left(\mathcal{L}(\theta^{(0)}) - \mathcal{L}^\star\right),$$

where $U^{(t)} = (\eta L_2 - 1)(\nabla\mathcal{L}(\bar{\theta}^{(t)}))^\top \nabla \operatorname{Tr}\left(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\,\Gamma^{(t)}\right) + \Theta(\Xi_t^3)$*, with* $\Gamma^{(t)} = \frac{1}{m}\sum_{k=1}^{m}(\theta_k^{(t)} - \bar{\theta}^{(t)})(\theta_k^{(t)} - \bar{\theta}^{(t)})^\top$ *and the consensus distance* $\Xi_t^2 = \operatorname{Tr}(\Gamma^{(t)})$.

Algorithm	Parallel SGD	DSGD (prior work)	DSGD (ours)
Rate	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2} + \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2} + \frac{1}{p\varepsilon} + \frac{\sqrt{p}\,\sigma+\zeta}{p\,\varepsilon^{3/2}}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2} + \frac{1}{\varepsilon} + \frac{1}{\varepsilon}\sum_{t=0}^{T-1} U^{(t)}\right)$

Important Notations: m denotes the number of nodes, p the consensus rate (a spectral-gap quantity related to $(1-\lambda_2(W))$, $\overline{\theta^{(t)}}$ the merged model, $\Gamma^{(t)}$ the covariance of $\{\theta_k^{(t)}\}_{k=1}^m$, $\Xi_t^2 = \operatorname{Tr}(\Gamma^{(t)})$ the consensus error, σ the level of gradient noise, and ζ the level of data heterogeneity.

Understanding The Dynamics of The Merged-Model

Question: Why Can Merged Model Work Well

Proposition 2. *Suppose Assumption 2 and Assumption 4 hold. Assume η satisfies $\eta > 1/L_2$, and assume $\|\nabla \mathcal{L}(\bar{\theta}^{(t)})\| \geq \mu_t > 0$ for all t. Consider the matrix $\Gamma^{(t)} = \frac{1}{m}\sum_{k=1}^{m}(\theta_k^{(t)} - \bar{\theta}^{(t)})(\theta_k^{(t)} - \bar{\theta}^{(t)})^\top$ and its trace $\Xi_t^2 = \mathrm{Tr}(\Gamma^{(t)})$. Then, for any fixed $m > 0$, there exists $\Xi_t^2 > 0$ such that*

$$U^{(t)} \triangleq \frac{1}{2}(\eta L_2 - 1)\nabla \mathcal{L}(\bar{\theta}^{(t)})^\top \nabla \mathrm{Tr}\left(\nabla^2 \mathcal{L}(\bar{\theta}^{(t)})\,\Gamma^{(t)}\right) + O(\Xi_t^3) < 0. \tag{8}$$

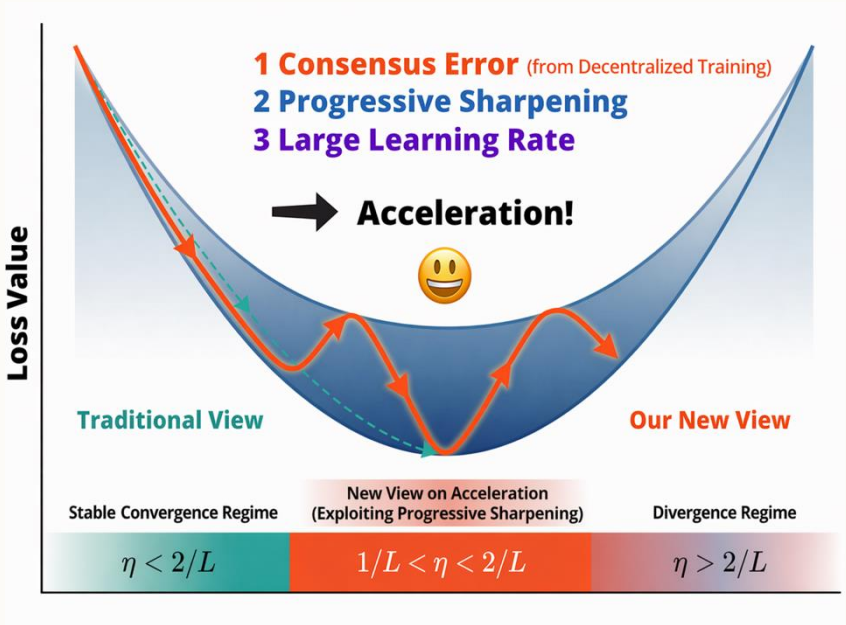
Algorithm	Parallel SGD	DSGD (prior work)	DSGD (ours)
Rate	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2} + \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2} + \frac{1}{p\varepsilon} + \frac{\sqrt{p}\,\sigma+\zeta}{p\,\varepsilon^{3/2}}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2} + \frac{1}{\varepsilon} + \frac{1}{\varepsilon}\sum_{t=0}^{T-1} U^{(t)}\right)$
Interpretation	**Baseline**	Consensus error is treated as purely harmful	Controlled consensus error can be constructive

Key finding: Full consensus is not always necessary for fast convergence; controlled consensus error can be constructive.

Understanding The Dynamics of The Merged-Model

Proof Idea and Intuition

Algorithm	Parallel SGD	DSGD (prior work)	DSGD (ours)
Rate	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2}+\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2}+\frac{1}{p\varepsilon}+\frac{\sqrt{p}\,\sigma+\zeta}{p\,\varepsilon^{3/2}}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2}+\frac{1}{\varepsilon}+\frac{1}{\varepsilon}\sum_{t=0}^{T-1}U^{(t)}\right)$
Interpretation	**Baseline**	Consensus error is treated as purely harmful	Controlled consensus error can be constructive



Descent Lemma

$$\Delta\mathcal{L}(\bar{\theta}^{(t)})$$
$$\leq -\underbrace{\left(\eta-\frac{\eta^2 L_2}{2}\right)}_{>0}\underbrace{\left\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\right\|^2}_{\text{Standard Descent}}+\frac{\eta^2 L_2\sigma^2}{2m}+\mathcal{O}(\Xi_t^3)$$
$$+\underbrace{\left(\eta^2 L_2-\eta\right)}_{>0}\underbrace{\nabla\mathcal{L}(\bar{\theta}^{(t)})^\top\nabla\mathrm{Tr}\left(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)}\right)}_{<0,\ \text{progressive sharpening}}.$$

Takeaways & Resources

Main takeaways

In Non-IID decentralized learning,

- Late communication matters: even a single final global merging can be surprisingly effective.
- Strong performance does not require full consensus.

Broader implications

- Decentralized/gossip learning is not merely a communication-constrained technique; it is a distinct optimization regime with rich phenomena of its own.
- Our new perspective opens new avenues for model merging, and communication-efficient large-scale distributed learning.

Paper

Blog

Reaching out please!

🚨 **Tongtian** and **Zhanpeng** are currently on the **job market.** Feel free to **reach out us!**

Tongtian Zhu First author | ZJU

Decentralized Learning & Optimization
raiden@zju.edu.cn

Zhanpeng Zhou Corresponding author | SJTU

Science of LLM pre-training & Deep Learning Theory
zzp1012@sjtu.edu.cn

Thank you very much for listening.