ON THE SURPRISING EFFECTIVENESS OF A SINGLE GLOBAL MERGING IN DECENTRALIZED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized learning provides a scalable alternative to parameter-server-based training, yet its performance is often hindered by limited peer-to-peer communication. In this paper, we study how communication should be scheduled over time to improve global generalization, including determining when and how frequently devices synchronize. Counterintuitive empirical results show that concentrating communication budgets in the later stages of decentralized training remarkably improves global generalization. Surprisingly, we uncover that fully connected communication at the final step, implemented by a single global merging, can significant improve the generalization performance of decentralized learning under serve high data heterogeneity. Our theoretical contributions, which explains these phenomena, are first to establish that the globally merged model of decentralized SGD can match the convergence rate of parallel SGD. Technically, we reinterpret part of the discrepancy among local models, which were previously considered as detrimental noise, as constructive components essential for matching this rate. This work provides promising results that decentralized learning is able to generalize under high data heterogeneity and limited communication, while offering broad new avenues for model merging research. The code will be made publicly available.

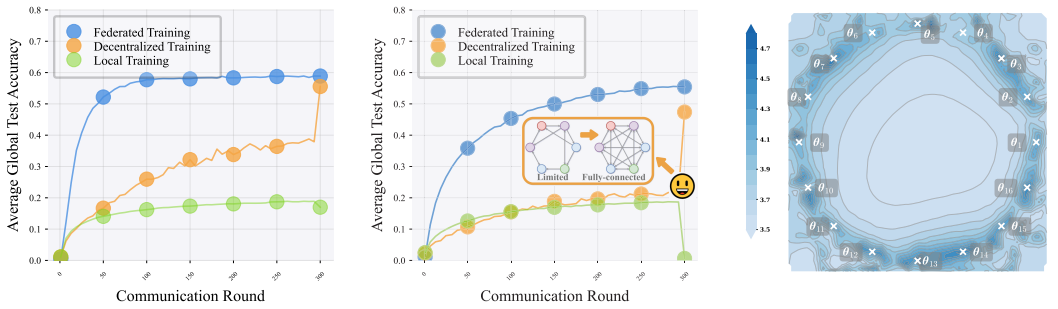
1 INTRODUCTION

Decentralized learning offers a promising approach to crowdsource computational workloads across geographically distributed compute (Yuan et al., 2022; Borzunov et al., 2023b; Jaghouar et al., 2024). A defining characteristic of this setting is the reliance on peer-to-peer communication during training, involving the peer-level exchange of model parameters or gradients during training. However, such communication is often constrained in practice due to limited bandwidth between geographically distant nodes, making it a scarce resource. These constraints can significantly degrade the performance of decentralized learning, both theoretically and empirically (Lian et al., 2017; Koloskova et al., 2020; Vogels et al., 2021). As a result, efficiently allocating limited communication resources becomes a fundamental challenge in decentralized learning, especially in heterogeneous environments where varying local data distributions intensify communication demands (Martínez Beltrán et al., 2023).

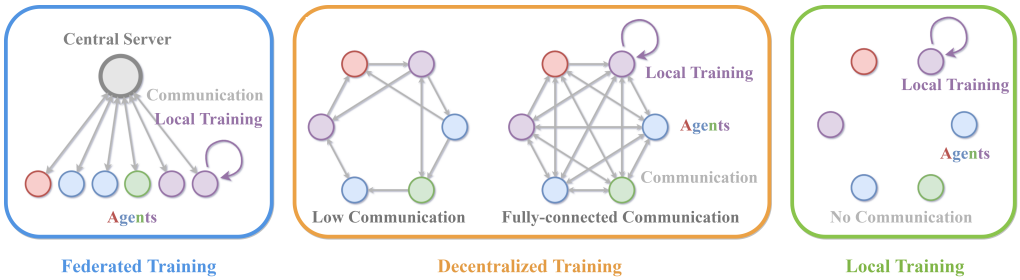
To date, most efforts addressing this challenge have focused on optimizing communication allocation at the *spatial level*, particularly through the design of communication graphs (Ying et al., 2021; Li et al., 2022b; Takezawa et al., 2023; Kharrat et al., 2024). In contrast, the *temporal* allocation of communication, i.e., deciding when and how frequently agents synchronize with others, remains a significant yet underexplored direction for improving decentralized learning. Although temporal communication allocation has been studied in federated learning (FL) (Tang et al., 2020), this problem remains largely untouched in the fully decentralized setting, which is fundamentally different due to the lack of a central server for global aggregation (see discussions in Section 2 and Remark 1).

Question: *How to allocate communication budget in decentralized learning over temporal levels?*

To answer this question, we design a series of experiments that allocate communication budgets across different time windows during training (see Figure 2). Specifically, we divide the training process into consecutive windows, each consisting of a fixed number of communication rounds. We assign higher communication budgets to selected windows using global synchronization via `AllReduce` (Sergeev



(a) CLIP ViT-B/32 (b) ResNet-18 (w/o pretraining) (c) Landscape before final merging



(d) A comparative illustration of federated, decentralized, and local training.

Figure 1: **(a, b)**: Global test accuracy (see Definition 1) of CLIP ViT-B/32 (a) and ResNet-18 (b) trained on Tiny ImageNet using FedAvg (blue), decentralized SGD (orange), and one-shot FedAvg (green), distributed across 32 agents with high data heterogeneity (Dirichlet $\alpha = 0.1$). Decentralized training involves each agent syncing model parameters with a random peer per round with a probability of 0.2, followed by a global merging at the final round (see details in Appendix C.1). **(c)** Loss landscape for 16-agent training with decentralized SGD, prior to the final merging (see details in Appendix C.1).[3] **(d)**: An illustration comparing federated, decentralized, and local training.

& Del Balso, 2018), while keeping communication low otherwise by infrequent synchronization with random peer agents.[1] This enables us to gain insights into how the temporal communication allocation affect the generalization performance under constrained budgets. We observe that allocating higher communication budgets toward the later stages of training consistently leads to improved final global test performance (see Definition 1). More surprisingly, we observe the remarkable effect of a single round of fully-connected communication.[2]

Surprising Phenomenon: *A single global merging of decentralized models, even under severely constrained communication and high data heterogeneity, can significantly improve global generalization.*

Our Contributions are summarized below.

- **Empirical Observations**. **(1)**: We highlight the critical role of a single global merging in decentralized training, showing that it can achieve performance close to federated learning, even under severe communication constraints and data heterogeneity (see Figure 1a, Figure 1b). The results remain consistent across different hyperparameter setups, datasets, degree of data heterogeneity, model architectures, optimizers, initialization schemes, and communication topologies (see additional results in Appendix C.3). **(2)**: We observe that limited but non-zero communication preserves the "mergeability" of local models throughout training (see Definition 2, Figure 1c, and the blue curve in Figure 2c), which does not hold under complete local training (green curve in Figure 1a, Figure 1b).

[1] Agents refer to participants in decentralized learning. "Communication" and "synchronization" are used interchangeably.

[2] Fully-connected communication refers to global synchronization via `AllReduce`. In this paper, fully-connected communication is realized through parameter averaging over the models on all agents, namely *global merging*.

[3] We use 16 agents for loss landscape visualization to ensure visual clarity.

Notably, our work takes the first step towards a systematic study of the global merging strategy in decentralized learning, revealing its standalone effectiveness in generalization improvement.

- **Theoretical Contributions**. We investigate the underlying mechanism that enables the *mergeability* of local models in decentralized learning. Specifically, we provide the first convergence analysis showing that the globally merged model of decentralized SGD can match the rate of parallel SGD (Theorem 1 and Proposition 2). Furthermore, we offer a theoretical explanation for why limited but nonzero communication can ensure mergeability, and why communication should be concentrated in the later stages of training (see Proposition 3).

We anticipate that this work will pave the way for principled decentralized training algorithms capable of generalizing under severe communication constraints and data heterogeneity, while also advancing model merging research (see discussions in Section 6). We also provide additional insights and address potential limitations in a **Q&A Section** (see Appendix A).

2 Related Work

Temporal Communication Allocation in Parallel, Federated, and Decentralized Learning. Communication allocation is well-studied in both data-centric parallel learning (Li et al., 2014), and Federated Learning (FL) (McMahan et al., 2017). In parallel learning settings, Gu et al. (2024) proposed a novel strategy for scheduling local steps through analyzing the implicit bias of Local SGD (Gu et al., 2023b). FL extends this server-based paradigm to handle not identically and independently distributed (non-IID) data, but it critically retains a global model. This reliance on a global model has shaped a broad consensus in the FL literature: frequent, early-stage communication is considered essential for aligning local models (Wang et al., 2019; Tang et al., 2020).

In contrast, our work addresses fully decentralized learning, a fundamentally different setting that lacks a central server. Instead of optimizing a generic global model, the goal is to make local models generalize to the global distribution. Despite extensive work focusing on communication allocation at the spatial level in decentralized learning (e.g., designing communication topologies) (Ying et al., 2021; Li et al., 2022b; Takezawa et al., 2023; Kharrat et al., 2024), few studies have examined the communication allocation problem over temporal levels. Pioneering work by Kong et al. (2021) demonstrated that in IID scenarios, aligning local models more closely with their global average early in training modestly improves generalization. However, these findings do not directly translate to non-IID scenarios, as they are based on the IID assumption where the global population risk $\mathcal{L}(\cdot)$ reduces to the local population risk $\mathcal{L}_k(\cdot)$ (see Equation (1) and Definition C.2). Therefore, their results primarily address local generalization, as opposed to the global generalization (see Definition 1) in our work. Due to space constraints, we refer readers to Appendix B.2 and Appendix B.3 for related work on the implicit bias of decentralized learning, and on the topic of model merging.

3 Notations and Preliminaries

3.1 Non-IID Decentralized Learning

Decentralized learning formalizes distributed learning as an optimization problem over a connected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ contains m agents and $\mathcal{E}$ denotes the communication links. Each agent $k \in \mathcal{V}$ samples data from a local distribution $\mathcal{D}_k$ and maintains a local model $\theta_k \in \mathbb{R}^d$. The objective is to learn a consensus model θ that minimize the global population risk (Koloskova et al., 2020):

$$\min_{\theta \in \mathbb{R}^d} \left[\mathcal{L}(\theta) \triangleq \frac{1}{m} \sum_{k \in \mathcal{V}} \mathbb{E}_{\xi_k \sim \mathcal{D}_k} \mathcal{L}(\theta; \xi_k) \right], \tag{1}$$

where $\mathbb{E}_{\xi_k \sim \mathcal{D}_k} \mathcal{L}(\theta; \xi_k) \triangleq \mathcal{L}_k(\theta)$ denotes the local population risk of θ on unseen instance $\xi_k \sim \mathcal{D}_k$.

In practice, the optimization of Equation (1) is performed under the empirical risk minimization framework, leveraging m local datasets $S \triangleq \bigcup_{k=1}^{m} S_k$, where $S_k = \{\xi_{k,1}, \ldots, \xi_{k,\zeta}\}$ denotes the dataset of agent k sampled from $\mathcal{D}_k$. The resulting optimization problem is given by:

$$\min_{\theta \in \mathbb{R}^d} \left[\mathcal{L}_{\mathcal{S}}(\theta) \triangleq \frac{1}{m} \sum_{k \in \mathcal{V}} \sum_{\zeta=1}^{n_k} \mathcal{L}(\theta; \xi_{k,\zeta}) \right]. \tag{2}$$

To solve the optimization problem in Equation (2), decentralized algorithms minimize the global empirical risk with only local computations and peer-to-peer communication (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009). The communication graph is governed by a weighted adjacency matrix $W^{(t)} \in [0,1]^{m\times m}$, sampled from a distribution $\mathcal{W}^{(t)}$, where each entry $W_{k,l}^{(t)} \geq 0$ reflects the influence of agent l on agent k.[4] Decentralized learning algorithms operate by alternating between local updates and model aggregation through communication with neighbors, as outlined in Algorithm 1.

Algorithm 1 Decentralized Learning

input Initialize values $\theta_k^{(0)} \in \mathbb{R}^d$ on each agent $k \in \mathcal{V}$, number of steps T, mixing matrix W

1: **in parallel on all agent** $k \in \mathcal{V}$**, for** $t = 0, \ldots, T-1$ **do**
2: Sample training data $\xi_k^{(t)}$ from $\mathcal{D}_k$, $\theta_k^{(t+1)} \leftarrow \textit{Optimizer}(\theta_k^{(t)}, \xi_k^{(t)})$ ▷ Local update
3: Send $\theta_k^{(t)}$ to out-neighbor(s) and receive $\{\theta_l^{(t)}\}_{l\in\mathcal{N}_{\text{in}}(k)}$ from in-neighbor(s) ▷ Communication
4: Sample mixing matrix $W^{(t)} \sim \mathcal{W}^{(t)}$, $\theta_k^{(t+1)} \leftarrow \sum_{l\in\mathcal{N}_{\text{in}}(k)} W_{k,l}^{(t)}\theta_l^{(t)}$ ▷ Gossip averaging
5: **end parallel for**

Practical Evaluation Metrics. In decentralized learning, models are often evaluated in the absence of a full consensus model θ due to data heterogeneity and limited training time. In this paper, we adopt the *average global test accuracy*, a proxy of average global population risk, as the primary evaluation metric, which quantifies how well local models generalize to the global data distribution.

Definition 1 (Average Global Test Accuracy)**.** *The average accuracy of agents $k \in \mathcal{V}$ is defined as:*

$$\underbrace{\overline{\text{Acc}}(\{\theta_k^{(t)}\}_{k\in\mathcal{V}}) = \frac{1}{m}\sum_{k\in\mathcal{V}} \text{Acc}(\theta_k^{(t)})}_{\textit{Average Accuracy across agents}}, \quad \textit{where } \text{Acc}(\cdot) \triangleq \underbrace{\frac{1}{m}\sum_{l\in\mathcal{V}} \mathbb{E}_{\xi_l\sim\mathcal{D}_l}\,\text{Acc}(\cdot;\xi_l)}_{\textit{Test accuracy on the global distribution}} .$$

Remark 1 (Metric Justification)**.** This metric is specifically designed to address a core question in fully decentralized learning: *how well do local models $\{\theta_k^{(t)}\}_{k\in\mathcal{V}}$, trained with limited peer-to-peer synchronization, generalize to the global data distribution $\mathcal{D}$?* This metric offers a more realistic evaluation for decentralized settings without a global model. See discussions in Appendix C.2.

3.2 Mergeability

Definition 2 (Mergeability under Global Population Risk)**.** *A set of local models $\{\theta_k\}_{k\in V}$ is globally mergeable if there exist combination weights $\{w_k\}_{k\in V} \in [0,1]$ such that:*

$$\mathcal{L}\left(\sum_{k\in V} w_k\theta_k\right) \leq \sum_{k\in V} w_k\mathcal{L}(\theta_k), \tag{3}$$

where $\mathcal{L}(\cdot)$ denotes the global population risk.

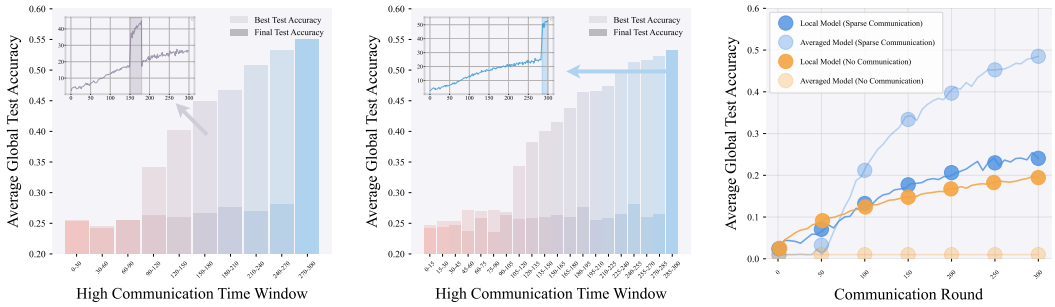
Definition 2 formalizes the intuition that a linearly interpolated model perform no worse than the original local models. The Definition is inherently non-trivial due to the *non-convexity* of $\mathcal{L}$.

4 Empirical Observations

4.1 Increasing Impact of Communication in the Later Stages of Training

The primary objective of this paper is to design a temporal communication strategy for decentralized learning that enables local models to generalize effectively to the global data distribution (see Remark 1). To investigate potential solutions, we explore a direct strategy: Concentrate communication

[4] Our framework incorporates randomized decentralized learning setting where the weighted adjacency matrix $W^{(t)}$ can change during training (Boyd et al., 2006; Koloskova et al., 2020; Vos et al., 2023).



(a) (b) (c)

Figure 2: **(a, b)**: Comparisons of global test accuracy (see Definition 1) in decentralized training of ResNet-18 on CIFAR-100 with AdamW, distributed across 16 agents with Dirichlet $\alpha = 0.1$ (see details in Appendix C.1). Fully-connected communication (i.e., AllReduce) is activated only in specific windows, while low communication with one random peer with a probability of 0.2 is used elsewhere. **(a)**: Fully-connected communication in $1/10$ of total rounds. **(b)**: Fully-connected communication in $1/20$ of total rounds. In both, lighter bars show peak accuracy, darker bars show final accuracy. **(c)**: Global test accuracy curves for local models and the globally averaged model (counterfactual) under persistent low communication (blue) and no communication (orange).[6]

in a small subset of communication rounds. To this end, we divide the training process into consecutive windows, each consisting of a fixed length of communication rounds. Specifically, the communication scheme is as follows: (1) fully-connected communication (see Figure 1d (b)) is activated only within specific communication windows (i.e., global synchronization via `AllReduce` (Sergeev & Del Balso, 2018)[5]); (2) while in all other rounds, each agent communicates only with *one* random peer with a probability of 0.2 (see "Communication Graph" in Appendix C.1).

As shown in Figure 2, training is divided into 10 (a) and 20 (b) communication windows, respectively. The bars in Figure 2 show both the best global test accuracy achieved during training (lighter-colored bars) and the final test accuracy at the end of training (darker-colored bars). Each bar corresponds to one communication window, where fully connected communication is applied *only* to the rounds within that window, while random peer communication is used in all other rounds. For instance, the inset in Figure 2a presents the complete test accuracy trajectory when fully-connected communication is applied during rounds 150 to 180. A consistent trend emerges: *allocating communication budgets toward the later stages of training yields substantial improvements, particularly in final test accuracy*.

4.2 A Single Global Merging Significantly Improves Global Generalization

In Figure 2b, we reduce the fully-connected communication window length to 10 rounds, yet still observe substantial improvements in global generalization. This observation naturally raises the question: *What happens if the fully-connected window is reduced to a single round?*

To investigate this, we conduct experiments where fully-connected communication is applied only once, implemented by a single global merging. As shown in Figure 1a and Figure 1b, a single global merging is sufficient to significantly improve global generalization. Consistent gains by a single global merging are observed across a wide range of settings, including different datasets, model architectures, optimizers, initialization schemes, and communication topologies (see additional experimental results in Appendix C.3). The significant increase in performance suggests that the global generalization potential of decentralized learning might be considerably underestimated.

Comparisons. D-PSGD (Lian et al., 2017) introduced the idea of final global merging under IID settings, yet the performance gain before and after merging was not analyzed. In contrast, we provide the first systematic study of this performance recovery in challenging non-IID scenarios. Further, Chen et al. (2021) demonstrated the benefits of periodic global averaging. However, their method

[5]We note that `AllReduce` can be efficiently realized in a decentralized manner such as `Ring-ALLReduce`.

[6]The term "counterfactual" refers to the fact that no global merging occurs during decentralized training. Instead, we manually compute the test accuracy of the hypothetical globally averaged model to quantify the "mergeability" of local models.

requires frequent global communication every $H = 48$ steps; in contrast, we achieve recovery with only *a single* merging. We also note that Aketi et al. (2021) proposed Skew-Compensated Sparse Push (SCSP), an effective strategy to improve the communication efficiency of decentralized learning, which also includes a final global merging step. While both works share the goal of reducing communication, our approaches differ in methodology and experimental setting: (1) *Methodology.* SCSP proposes a *gradient sparsification* algorithm (top-k gradients) over a fixed topology. In contrast, we investigate the phenomenon of mergeability under *topological sparsification* (i.e., sparse gossip). (2) *Experimental setting.* Their analysis focuses on settings with a single local step ($H = 1$). In contrast, we demonstrate that mergeability is remarkably robust even with a large number of local update steps (e.g., $H = 100$) and high data heterogeneity. While these works share the broader goal of improving communication efficiency, our work offers a new perspective by investigating the *mergeability* itself: Why local models retain this property despite extremely limited communication and high data heterogeneity.

Cost Comparison and the Practical Feasibility of Global Merging. Let P be the model size, m the number of agents, and T the number of training rounds. A standard `AllReduce`-based protocol incurs a total communication cost of $\mathcal{O}(m^2PT)$ throughout training. In contrast, our decentralized setup has a cost of $\mathcal{O}(mRPT + m^2P)$, where $R \ll m$ denotes the expected number of peers per round, and the $\mathcal{O}(m^2P)$ term arises from final merging. We also note that while a global merging may appear impractical in some decentralized settings due to the lack of `AllReduce` communication, it can be effectively approximated via multiple rounds of local synchronization (i.e, gossip).

4.3 Mergeability Persists under Limited but Nonzero Communication

A follow-up question is whether the effectiveness of the global merging is specific to the end of training. To investigate this, we assess the *counterfactual* performance of the globally averaged model at each training round, as depicted by the light-blue curve in Figure 2c. The experiments are conducted under a lower-communication setting, where each agent communicates with one random peer at each round with probability 0.2 (see "Communication Graph" in Subsection C.1). A consistent superiority of the merged model (light-blue curve) over the local models (dark-blue curve) is observed throughout training, suggesting that local models remain mergeable at all stages (see Definition 2).

As an ablation, we conduct an experiment in which all models are trained entirely locally without any communication (see Figure 2c). In this case, the counterfactual test performance of the globally averaged model remains close to zero (light-orange curve), indicating that without communication, local models are not mergeable. This suggests that mergeability does *not* arise inherently from the local models themselves. Interestingly, under the low-communication setting, the test performance of local models before merging (dark-blue curve) remains similar to that in the no-communication case (dark-orange curve). However, after global merging, the resulting model shows shows significant generalization improvement. This clear contrast implies that extremely limited but nonzero communication plays a pivotal role in enabling mergeability.

Mergeability without Consensus. Prior work on gossip algorithms has suggested that local models may converge to a similar state even in minimal communication regimes (Jelasity et al., 2005). In contrast, our work addresses a more challenging heterogeneous data setting where we find that local models do not reach a single consensus point, yet remain mergeable. Specifically, we identify an emergent geometric structure where decentralized training guides local models to a ring-like high-loss region surrounding a central low-loss basin (see Figure 1c).

5 Theoretical Analysis

In this section, we examine the underlying mechanisms that enable the mergeability of local models in decentralized learning. As an initial step, we conduct a fine-grained convergence analysis of the globally merged models trained by Decentralized SGD (DSGD).[7] To substantiate the mergeability of local models, we compare the convergence rate of the merged model of DSGD model to that of parallel SGD. Remarkably, we prove that the merged model in decentralized learning can match the rate of parallel SGD (Dekel et al., 2012; Li et al., 2014). This supports the empirical findings that the merged model can preserve the performance of individual local models (see Definition 2).

[7] DSGD refers to standard decentralized SGD where the optimizer in Algorithm 1 is replaced with SGD.

5.1 Assumptions

We start by introducing the commonly used assumptions (Kong et al., 2021; Koloskova et al., 2020).

Assumption 1 (Mixing matrix). *Each sample of the (randomized) mixing matrix $W \in \mathbb{R}^{m\times m}$ is doubly stochastic. Moreover, there exists $p > 0$ such that*

$$\mathbb{E}_W \|\Theta W - \bar{\Theta}\|_F^2 \le (1-p)\|\Theta - \bar{\Theta}\|_F^2, \ \forall\, \Theta \in \mathbb{R}^{d\times m}. \tag{4}$$

Here $\Theta = [\theta_1, \dots, \theta_m]$, $\bar{\Theta} = [\bar{\theta}, \dots, \bar{\theta}] \equiv \Theta \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ where $\bar{\theta} = \frac{1}{m}\sum_{k=1}^m \theta_k$.

Assumption 2 (Regularity). *The objective function $\mathcal{L}$ is four-times continuously differentiable (i.e., $\mathcal{L} \in \mathcal{C}^4$) and there exist constants $L_q \ge 0$ for $q \in \{1, \dots, 4\}$ such that:*

$$\|\nabla^q \mathcal{L}(\theta)\| \le L_q, \quad \forall\, \theta \in \mathbb{R}^d. \tag{5}$$

We note that given $\mathcal{L} \in \mathcal{C}^2$, the boundedness of the Hessian norm (i.e., the case $q = 2$) implies that $\mathcal{L}$ is L_2-smooth, thereby recovering Assumption D.1 with ($L = L_2$).

Assumption 3 (Bounded noise and diversity). *There exist $\sigma^2, \zeta^2 \ge 0$ such that for any $\theta_k \in \{\theta_k\}_{k=1}^m$:*

$$\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\xi_k}\|\nabla \mathcal{L}_k(\theta_k;\xi_k) - \nabla \mathcal{L}_k(\theta_k)\|_2^2 \le \sigma^2, \quad \frac{1}{m}\sum_{k=1}^m \|\nabla \mathcal{L}_k(\theta_k) - \nabla \mathcal{L}(\theta_k)\|_2^2 \le \zeta^2, \tag{6}$$

where $\mathcal{L}(\theta) = \frac{1}{m}\sum_{k=1}^m \mathcal{L}_k(\theta)$.

Here σ measures the local noise level and ζ is measure of the heterogeneity among agents.

5.2 Convergence Analysis

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 $\eta \le \frac{1}{L_2}$. 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 $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla \mathcal{L}(\bar{\theta}^{(t)})\|_2^2\big] \le \varepsilon$, the total number of steps T satisfies:*

$$T = \mathcal{O}\Big(\tfrac{\sigma^2}{m\varepsilon^2} + \tfrac{1}{\varepsilon} + \tfrac{1}{\varepsilon}\big([\textstyle\sum_{t=0}^{T-1} A^{(t)}]^+\big)^{1/2}\Big) \cdot L_2\big(\mathcal{L}(\theta^{(0)}) - \mathcal{L}^\star\big),$$

where $[\cdot]^+ \triangleq \max(0, \cdot)$ and $A^{(t)}$ is defined as:

$$A^{(t)} \triangleq \eta L_2 \left(2T_2 + L_3^2 \Xi_t^4 + \left(2L_1 + 2L_3\Xi_t^2 + \frac{mL_4^2}{24^2}\right)\sqrt{m}\Xi_t^3\right),$$

with T_2 and the consensus distance Ξ_t^2 given by:

$$T_2 \triangleq (\nabla \mathcal{L}(\bar{\theta}^{(t)}))^\top \nabla \operatorname{Tr}\big(\nabla^2 \mathcal{L}(\bar{\theta}^{(t)})\, \Gamma^{(t)}\big), \quad \Xi_t^2 \triangleq \frac{1}{m}\sum_{k=1}^m (\theta_k^{(t)} - \bar{\theta}^{(t)})^\top (\theta_k^{(t)} - \bar{\theta}^{(t)}).$$

Remark 2. We note that Theorem 1 gives an implicit bound depending on $A^{(t)}, t \in \{1, 2, \dots, T-1\}$, rather than a closed-form expression. It primarily serves to bridge convergence with the per-iteration dynamics of $A^{(t)}$, facilitating the subsequent derivation of the conditions on consensus and communication required to recover the parallel SGD rate (see Proposition 2 and Proposition 3).

Comparison. As summarized in Table 1, unified analysis by Koloskova et al. (2020) showed that DSGD suffers from additional terms of order $\mathcal{O}\left(\frac{1-p}{p\varepsilon} + \frac{\sqrt{p}\,\sigma+\zeta}{p\,\varepsilon^{3/2}}\right)$ in the convergence rate compared to parallel SGD. The core idea behind their analysis is to separate the effects of three key factors: the descent force (i.e., the squared gradient norm), gradient noise, and parameter discrepancy among agents. Each of these components is then analyzed and controlled separately. Among them, both the gradient noise and the model discrepancy are treated as detrimental to convergence. In

Table 1: Comparison of non-convex convergence rates for parallel SGD and DSGD, both run with m agents under non-iid data.

Algorithm	Parallel SGD	DSGD (Koloskova et al., 2020)	DSGD (ours)
Rate	$\mathcal{O}\Big(\frac{\sigma^2}{m\varepsilon^2}+\frac{1}{\varepsilon}\Big)$	$\mathcal{O}\Big(\frac{\sigma^2}{m\varepsilon^2}+\frac{1}{p\varepsilon}+\frac{\sqrt{p}\,\sigma+\zeta}{p\,\varepsilon^{3/2}}\Big)$	$\mathcal{O}\Big(\frac{\sigma^2}{m\varepsilon^2}+\frac{1}{\varepsilon}+\frac{1}{\varepsilon}\big([\sum_{t=0}^{T-1}A^{(t)}]^+\big)^{1/2}\Big)$

contrast, we adopt a new proof framework that leverages the implicit bias of decentralized learning (see Proposition D.3 (Zhu et al., 2023b) and Appendix B.2). Rather than treating the discrepancy among agents purely as noise, we partially incorporate it as a constructive component essential for matching the rate of paralle SGD. This intuition is formalized through the convergence guarantee provided in Theorem 1, which introduces an additional term of $\mathcal{O}\left(\frac{1}{\varepsilon^2}\big([\sum_{t=0}^{T-1}A^{(t)}]^+\big)^{1/2}\right)$, where $[\cdot]^+ \triangleq \max(0,\cdot)$. In what follows, we conduct a fine-grained analysis on the sign of $A^{(t)}$.

Remark 3 (Reduction to Standard Rates)**.** We consider two special cases where the term $A^{(t)}$ vanishes because the consensus error is identically zero ($\Xi_t \equiv 0$):

- The single-agent case ($m = 1$);
- The fully synchronous Parallel SGD case, where perfect synchronization ensures identical local models ($\theta_k^{(t)} \equiv \bar{\theta}^{(t)}$ for all k).

In both settings, the auxiliary term $A^{(t)}$ in Theorem 1 strictly equals zero. Consequently, Theorem 1 naturally recovers the convergence rate of standard (Parallel) SGD, which is of the order $\mathcal{O}\left(\frac{\sigma^2}{m\varepsilon^2}+\frac{1}{\varepsilon}\right)$. This confirms that the comparison in Table 1 is fair, as our unified bound applies to both settings without requiring any additional assumptions for the decentralized setting.

To better characterize how the high-order loss landscape affects the dynamics of $A^{(t)}$, we introduce a new assumption that is theoretically novel yet empirically supported by prior literature.

Assumption 4 (Progressive sharpening)**.** *For any positive semi-definite matrix Σ, the gradient of population risk negatively aligns with the gradient of sharpness. Formally, $\forall\theta \in \mathbb{R}^d$,*

$$\nabla\mathcal{L}(\theta)^\top \nabla \operatorname{Tr}(\nabla^2\mathcal{L}(\theta)\Sigma) < 0. \tag{7}$$

Remark 4. Intuitively, $\operatorname{Tr}(\nabla^2\mathcal{L}(\theta)\Sigma)$ can be interpreted as an "average sharpness" around θ; see similar metrics in (Gu et al., 2023a; Zhu et al., 2023b). Assumption 4 reflects a widely observed phenomenon in deep learning: The loss gradient exhibits a negative correlation with the gradient of sharpness (Wang et al., 2022; Damian et al., 2023; Cohen et al., 2025).

Assumption 4 ensures that T_2 in $A^{(t)}$ remains negative. In the following, we formally establish that T_2 can dominate the other terms in $A^{(t)}$, thereby ensuring that $A^{(t)}$ remains non-positive.

Proposition 2. *Suppose Assumption 2 and Assumption 4 hold, 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 $\Xi_t^2 = \operatorname{Tr}(\Gamma^{(t)})$. Then, for any fixed $m > 0$, there exists a $\Xi_t^2 > 0$ such that*

$$A^{(t)} \triangleq \eta L\Big(2T_2 + L_3^2\Xi_t^4 + \big(2L_1 + 2L_3\Xi_t^2 + \tfrac{mL_4^2}{24^2}\big)\sqrt{m}\,\Xi_t^3\Big) \leq 0, \tag{8}$$

where $T_2 = (\nabla\mathcal{L}(\bar{\theta}^{(t)}))^\top \nabla \operatorname{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\,\Gamma^{(t)})$.

Explanations for Assumptions.

- We assume a lower bound on the global gradient norm evaluated at the averaged parameters $\bar{\theta}^{(t)}$, i.e., $\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\| \geq \mu_t > 0$. We note that this applies to the gradient on the global data set, which can remain significant even if individual local gradients vanish. The assumption is motivated by the Polyak-Lojasiewicz (PL) condition (Polyak, 1963), $\frac{1}{2}\|\nabla\mathcal{L}(\theta)\|^2 \geq \mu(\mathcal{L}(\theta) - \mathcal{L}^*)$, which ensures the gradient is bounded from zero before reaching the optimum. Our new assumption formalizes this property for the pre-convergence phase by denoting this lower bound at iteration t as $\frac{1}{2}\mu_t^2$.

- We note that Assumption 2 requires that the norm of the loss derivatives are bounded up to the fourth order, $\|\nabla^q \mathcal{L}(\theta)\| \leq L_q$ for $q = 1, 2, 3, 4$. These higher-order bounds are necessary to analyze the the interaction between the consensus error Ξ_t and higher-order landscape geometry.

Proposition 2 highlights the critical role of the consensus violation term $\Xi_t = \sqrt{\mathrm{Tr}(\Gamma^{(t)})}$. In conjunction with Theorem 1, Proposition 2 implies that DSGD can match the parallel SGD rate if Ξ_t $(\forall t \in [T])$ is properly controlled. According to Corollary D.2, $\mathbb{E}\left[\Xi_t^2\right]$ is bounded by

$$\mathbb{E}\left[\Xi_t^2\right] \leq \mathcal{O}\left(\frac{(1-p)\,\eta^2}{p^2}\right). \tag{9}$$

The parameter $p \in (0, 1]$ reflects the level of connectivity in the communication graph (see Assumption 1). A larger p indicates better connectivity and faster consensus, while a smaller p implies a sparse communication graph (i.e., lower communication) and slower information propagation. For example, $p = 1$ corresponds to a fully connected topology, enabling perfect communication, whereas $p = 0$ represents the extreme case of complete local training with no communication.

Remark 5. For the fully-connected case where $p = 1$, we observe that $A^{(t)} \equiv 0$ as $\Xi_t \equiv 0$. In this case, Theorem 1 recovers the rate of standard SGD.

Why Limited but Nonzero Communication Enables Mergeability. Notably, random communication graphs can achieve $p = \Theta(1)$, striking a favorable trade-off: they require relatively low communication overhead while still maintaining efficient information mixing due to randomized edge sampling, which ensures a rapid decrease of Ξ_t (Vos et al., 2023). This is why we adopt random topologies as the primary setup in our experiments: They can satisfy the condition in Proposition 2 even under extremely limited communication, thereby ensuring that mergeability (see Figure 1).

However, in the case of full local training where $p = 0$ (see Figure 1d), the right-hand side of Equation (9) increases to infinity, indicating that Ξ_t may diverge. As a consequence, the condition of Ξ_t in Proposition 2 can no longer be satisfied, which explains why local models after complete local training may not be reliably merged (see the green curve in Figure 1b).

5.3 A Theoretical Explanation for Communication Allocation

Recall that Proposition 2 shows there exists a threshold of consensus violation Ξ_t^2 for which Inequality (8) holds. This motivates the question of how small Ξ_t^2 (or how large p) should be, which we answer by providing the following sufficient condition.

Proposition 3 (Critical Consensus Edge)**.** *Suppose Assumption 1 and Assumption 2 hold. Assume the averaged squared gradient norm is bounded by* $\frac{1}{m}\sum_{k=1}^{m} \|\nabla \mathcal{L}_k(\theta_k^{(t)})\|^2 \leq \phi^2$ *for all* t*. Then the following condition ensures that the critical Inequality (8) is satisfied:*

$$\frac{12(1-p)\eta^2}{p^2}\left(\phi^2 + \sigma^2\right) < \min\left\{\sqrt{\frac{\gamma\mu_t}{2L_3^2}},\ \frac{\gamma\mu_t}{\left(2L_1 + \frac{\gamma\mu_t}{L_3} + \frac{mL_4^2}{24^2}\right)\sqrt{m}}\right\}, \tag{10}$$

where γ *denotes the degree of progressive sharpening (see Assumption 4), and* μ_t *is the lower bound on the gradient norm (i.e.,* $\|\nabla \mathcal{L}(\bar{\theta}^{(t)})\| \geq \mu_t > 0$ *for all* t*).*

Practical Guidance. Proposition 3 provides a guide for allocating communication to ensure $A^{(t)} \leq 0$, contributing to the non-positiveness of the cumulative sum $\sum_{t=0}^{T-1} A^{(t)}$ in Theorem 1. To derive a practical strategy from Equation (10), we observe that parameters ϕ, σ^2, γ, m, and $L_q (q = 1, 3, 4)$ are time-independent constants, the only quantity that vary with the iteration t is the gradient norm lower bound μ_t. The condition therefore simplifies to how p should be adjusted over time in response to the changing μ_t. Crucially, the left-hand side of Equation (10) is a decreasing function of p, while its right-hand side is an increasing function of μ_t. This means more communication (i.e., a larger p) makes the condition easier to satisfy, whereas a smaller μ_t tightens the bound. Specifically,

- Early, High-Gradient Regime: In the starting phase of training, when models are far from a minimum, the lower bound on gradient norm μ_t is large. This corresponds to a relaxed consensus

requirement in Equation (10), which permits low-frequency communication (i.e., smaller p) without significantly impacting the performance of the globally merged model.

- Late, Low-Gradient Regime: As models approach a solution and training enters a convergence phase, the gradient norm μ_t decreases. This tightens the constraint in Equation (10). In this regime, frequent communication (i.e., larger p) becomes critical.

We note that this theoretically motivated guidance aligns well with our empirical findings in Section 4 that more communication should be concentrated in the later stages of training.

6 Implications and Discussions

Model Merging. The success of a single merging of decentralized models has significant implications for the broader field of model merging. A recent work showed that pre-trained models occupy a large, flat "basic capability basin", within which fine-tuning creates smaller "specific capability basins" (Chen et al., 2025). The observed "mergeability" of local models in our paper implies that decentralized learning can "guide" each agent into specific capability basins that are inherently connected. This allows simple merging without permutation to effectively create a new model that successfully integrates the specialized knowledge. The insight opens a promising new avenue: introducing lightweight synchronization during local training may promote the connectivity between specialized models, thus simplifying their subsequent merging into a more capable model.

Decentralized Learning. Our work provides promising empirical and theoretical evidence that decentralized learning can generalize under high data heterogeneity and limited communication. More importantly, our findings could directly motivate a new class of adaptive, communication-efficient decentralized algorithms, which dynamically allocate their communication budget by monitoring training dynamics to satisfy the critical consensus edge condition in Equation (10).

References

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.

Sai Aparna Aketi, Amandeep Singh, and Jan Rabaey. Sparse-push: Communication-& energy-efficient decentralized distributed learning over directed & time-varying graphs with non-iid datasets. *arXiv preprint arXiv:2102.05715*, 2021.

Youssef Allouah, Anastasia Koloskova, Aymane El Firdoussi, Martin Jaggi, and Rachid Guerraoui. The privacy power of correlated noise in decentralized learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 1115–1143, 2024.

Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, 1999.

Marco Bornstein, Tahseen Rabbani, Evan Z Wang, Amrit Bedi, and Furong Huang. SWIFT: Rapid decentralized federated learning via wait-free model communication. In *The Eleventh International Conference on Learning Representations*, 2023.

Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Maksim Riabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. Petals: Collaborative inference and fine-tuning of large models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 558–568. Association for Computational Linguistics, 2023a.

Alexander Borzunov, Max Ryabinin, Artem Chumachenko, Dmitry Baranchuk, Tim Dettmers, Younes Belkada, Pavel Samygin, and Colin A Raffel. Distributed inference and fine-tuning of large language models over the internet. In *Advances in Neural Information Processing Systems*, 2023b.

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

Ying Cao, Zhaoxian Wu, Kun Yuan, and Ali H Sayed. On the trade-off between flatness and optimization in distributed learning. *arXiv preprint arXiv:2406.20006*, 2024.

CCAF. Cambridge bitcoin electricity consumption index (CBECI). https://ccaf.io/cbnsi/cbeci, 2023.

Huanran Chen, Yinpeng Dong, Zeming Wei, Yao Huang, Yichi Zhang, Hang Su, and Jun Zhu. Understanding pre-training and fine-tuning from loss landscape perspectives. *arXiv preprint arXiv:2505.17646*, 2025.

Lesi Chen, Haishan Ye, and Luo Luo. An efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization. *International Conference on Artificial Intelligence and Statistics*, 2024.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishna Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 4641–4671. PMLR, 2023.

Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Accelerating gossip sgd with periodic global averaging. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 1791–1802. PMLR, 2021.

Jeremy M Cohen, Alex Damian, Ameet Talwalkar, Zico Kolter, and Jason D Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025.

Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodolà. c^2m^3: Cycle-consistent multi-model merging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Edwige Cyffers, Aurélien Bellet, and Jalaj Upadhyay. Differentially private decentralized learning with random walks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 9762–9783, 2024.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *the Eleventh International Conference on Learning Representations*, 2023.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(6):165–202, 2012.

Arthur Douillard, Qixuan Feng, Andrei A Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. Diloco: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Mathieu Even, Anastasia Koloskova, and Laurent Massoulie. Asynchronous SGD on graphs: a unified framework for asynchronous decentralized and federated optimization. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.

C. Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. In *International Conference on Learning Representations*, 2017.

Hongchang Gao and Heng Huang. Fast training method for stochastic compositional optimization problems. *Advances in Neural Information Processing Systems*, 34:25334–25345, 2021.

Hongchang Gao, Bin Gu, and My T. Thai. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 9238–9281. PMLR, 2023.

Anissa Gardizy and Amir Efrati. Microsoft and OpenAI plot $100 billion stargate AI supercomputer. *The Information*, 2024. URL https://www.theinformation.com/articles/microsoft-and-openai-plot-100-billion-stargate-ai-supercomputer.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Grand View Research. Ai infrastructure market size, share & growth report, 2030, 2024. URL https://www.grandviewresearch.com/industry-analysis/ai-infrastructure-market-report.

Xinran Gu, Kaifeng Lyu, Longbo Huang, and Sanjeev Arora. Why (and when) does local SGD generalize better than SGD? In *The Eleventh International Conference on Learning Representations*, 2023a.

Xinran Gu, Kaifeng Lyu, Longbo Huang, and Sanjeev Arora. Why (and when) does local SGD generalize better than SGD? In *International Conference on Learning Representations*, 2023b.

Xinran Gu, Kaifeng Lyu, Sanjeev Arora, Jingzhao Zhang, and Longbo Huang. A quadratic synchronization rule for distributed deep learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Mert Gurbuzbalaban, Yuanhan Hu, Umut Simsekli, Kun Yuan, and Lingjiong Zhu. Heavy-tail phenomenon in decentralized sgd. *arXiv preprint arXiv:2205.06689*, 2022.

F. He, L. Nan, and T. Zhu. Imagining a democratic, affordable future of foundation models: A decentralised avenue. In *Handbook of Blockchain Analytics*. Springer, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, 2016.

Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via clippedgossip. *arXiv preprint arXiv:2202.01545*, 2022.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pp. 876–885. AUAI Press, 2018.

Sami Jaghouar, Jack Min Ong, Manveer Basra, Fares Obeid, Jannik Straube, Michael Keiblinger, Elie Bakouch, Lucas Atkins, Maziyar Panahi, Charles Goddard, et al. Intellect-1 technical report. *arXiv preprint arXiv:2412.01152*, 2024.

Márk Jelasity, Alberto Montresor, and Ozalp Babaoglu. Gossip-based aggregation in large dynamic networks. *ACM Trans. Comput. Syst.*, 23(3):219–252, August 2005. ISSN 0734-2071.

Salma Kharrat, Marco Canini, and Samuel Horvath. Decentralized personalized federated learning. *arXiv preprint arXiv:2406.06520*, 2024.

Jari Kolehmainen, Nikolay Blagoev, John Donaghy, Oğuzhan Ersoy, and Christopher Nies. Noloco: No-all-reduce low communication training method for large models. *arXiv preprint arXiv:2506.10911*, 2025.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, 2020.

Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, and Sebastian Stich. Consensus control for decentralized deep learning. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Alex Krizhevsky, G Hinton, et al. Learning multiple layers of features from tiny images (tech. rep.). *University of Toronto*, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Erick Lavoie, and Anne-Marie Kermarrec. Refined convergence and topology learning for decentralized sgd with heterogeneous data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Kevin Scaman, and Giovanni Neglia. Improved stability and generalization guarantees of the decentralized SGD algorithm. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. In *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*, 2022a.

Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. *Advances in Neural Information Processing Systems*, 2014.

Shuangtong Li, Tianyi Zhou, Xinmei Tian, and Dacheng Tao. Learning to collaborate in decentralized learning of personalized models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9766–9775, 2022b.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022c.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, 2018.

Tao Lin, Sai Praneeth Karimireddy, Sebastian Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Kaifeng Lyu. *Implicit Bias of Deep Learning Optimization: A Mathematical Examination*. PhD thesis, Princeton University, 2024.

Enrique Tomás Martínez Beltrán, Mario Quiles Pérez, Pedro Miguel Sánchez Sánchez, Sergio López Bernal, Gérôme Bovet, Manuel Gil Pérez, Gregorio Martínez Pérez, and Alberto Huertas Celdrán. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 25(4):2983–3013, 2023.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Michalis Mavrovouniotis, Changhe Li, and Shengxiang Yang. A survey of swarm intelligence for dynamic optimization: Algorithms and applications. *Swarm and Evolutionary Computation*, 33: 1–17, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1273–1282. PMLR, 2017.

Abdellah El Mrini, Edwige Cyffers, and Aurélien Bellet. Privacy attacks in decentralized learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Giorgi Nadiradze, Amirmojtaba Sabour, Peter Davies, Shigang Li, and Dan Alistarh. Asynchronous decentralized sgd with quantized and local updates. *Advances in Neural Information Processing Systems*, 2021.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Angelia Nedi'c and Alex Olshevsky. Distributed optimization over time-varying directed graphs. volume 60, pp. 601–615. IEEE, 2014.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

OpenAI. Announcing the stargate project. https://openai.com/index/announcing-the-stargate-project/, 2025.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Konstantin F Pilz, James Sanders, Robi Rahman, and Lennart Heim. Trends in ai supercomputers. *arXiv preprint arXiv:2504.16026*, 2025.

B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Sameera Ramasinghe, Thalaiyasingam Ajanthan, Gil Avraham, Yan Zuo, and Alexander Long. Protocol models: Scaling decentralized training with communication-efficient model parallelism. *arXiv preprint arXiv:2506.01260*, 2025.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, patrick gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Alexandre Rame, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Leon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28656–28679. PMLR, 23–29 Jul 2023.

Dominic Richards et al. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 2020.

Max Ryabinin, Tim Dettmers, Michael Diskin, and Alexander Borzunov. SWARM parallelism: Training large models can be surprisingly communication-efficient. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29416–29440. PMLR, 2023.

Ali H. Sayed. *Adaptation, Learning, and Optimization over Networks*. Now Publishers, 2014.

Alexander Sergeev and Mike Del Balso. Horovod: fast and easy distributed deep learning in tensorflow. *arXiv preprint arXiv:1802.05799*, 2018.

Li Shen, Yan Sun, Zhiyuan Yu, Liang Ding, Xinmei Tian, and Dacheng Tao. On efficient training of large-scale deep learning models. *ACM Computing Surveys*, 57(3), 2024.

Tao Shen, Didi Zhu, Ziyu Zhao, Chao Wu, and Fei Wu. Will llms scaling hit the wall? breaking barriers via distributed resources on massive edge devices. *arXiv preprint arXiv:2503.08223*, 2025.

Abhishek Singha, Charles Lua, Gauri Guptaa, Ayush Chopraa, Jonas Blanca, Tzofi Klinghoffera, Kushagra Tiwarya, and Ramesh Raskara. A perspective on decentralizing ai. 2024.

Ankit Sonthalia, Alexander Rubinstein, Ehsan Abbasnejad, and Seong Joon Oh. Do deep neural network solutions form a star domain? In *The Thirteenth International Conference on Learning Representations*, 2025.

Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Yuki Takezawa, Ryoma Sato, Han Bao, Kenta Niwa, and Makoto Yamada. Beyond exponential graph: Communication-efficient topologies for decentralized learning via finite-time convergence. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D2: Decentralized training over decentralized data. In *International Conference on Machine Learning*. PMLR, 2018.

Zhenheng Tang, Shaohuai Shi, Wei Wang, Bo Li, and Xiaowen Chu. Communication-efficient distributed deep learning: A comprehensive survey. *arXiv preprint arXiv:2003.06307*, 2020.

J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Commun. ACM*, 66(6):86–93, 2023.

Thijs Vogels, Lie He, Anastasia Koloskova, Sai Praneeth Karimireddy, Tao Lin, Sebastian U Stich, and Martin Jaggi. Relaysum for decentralized deep learning on heterogeneous data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Thijs Vogels, Hadrien Hendrikx, and Martin Jaggi. Beyond spectral gap: The role of the topology in decentralized learning. *Journal of Machine Learning Research*, 24(355):1–31, 2023.

Martijn De Vos, Sadegh Farhadkhani, Rachid Guerraoui, Anne marie Kermarrec, Rafael Pires, and Rishi Sharma. Epidemic learning: Boosting decentralized learning with randomized communication. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jue Wang, Yucheng Lu, Binhang Yuan, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Re, and Ce Zhang. CocktailSGD: Fine-tuning foundation models over 500Mbps networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36058–36076. PMLR, 2023.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022a.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, June 2022b.

Tongle Wu and Ying Sun. Implicit regularization of decentralized gradient descent for sparse regression. In *the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Wenhan Xian, Feihu Huang, Yanfu Zhang, and Heng Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34: 25865–25877, 2021.

Jie Xu, Wei Zhang, and Fei Wang. A(dp)^{2}sgd: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shuoguang Yang, Xuezhou Zhang, and Mengdi Wang. Decentralized gossip-based stochastic bilevel optimization over communication networks. *Advances in Neural Information Processing Systems*, 35:238–252, 2022.

Haoxiang Ye and Qing Ling. Generalization error matters in decentralized learning under Byzantine attacks. *IEEE Transactions on Signal Processing*, 2025.

Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems*, 2021.

Binhang Yuan, Yongjun He, Jared Quincy Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy Liang, Christopher Re, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 2022.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Liangqi Yuan, Ziran Wang, Lichao Sun, Philip S. Yu, and Christopher G. Brinton. Decentralized federated learning: A survey and perspective. *IEEE Internet of Things Journal*, pp. 1–1, 2024.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher G Brinton. Decentralized sporadic federated learning: A unified algorithmic framework with convergence guarantees. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wei Zhang, Mingrui Liu, Yu Feng, Xiaodong Cui, Brian Kingsbury, and Yuhai Tu. Loss landscape dependent self-adjusting learning rates in decentralized stochastic gradient descent. *arXiv preprint arXiv:2112.01433*, 2021.

Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *The Thirteenth International Conference on Learning Representations*, 2025.

Miaoxi Zhu, Li Shen, Bo Du, and Dacheng Tao. Stability and generalization of the decentralized stochastic gradient descent ascent algorithm. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized SGD. In *International Conference on Machine Learning*. PMLR, 2022.

Tongtian Zhu, Fengxiang He, Kaixuan Chen, Mingli Song, and Dacheng Tao. Decentralized SGD and average-direction SAM are asymptotically equivalent. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Tongtian Zhu, Wenhao Li, Can Wang, and Fengxiang He. DICE: Data influence cascade in decentralized learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

LLM USAGE STATEMENT

We use large language models (LLMs) as writing-assistance tools. Their role is confined to proof-reading and language polishing.

IMPACT STATEMENT

This paper studies the problem of temporal communication allocation in decentralized distributed learning, a topic of very high significance in the era of communication-intensive large model training. Specifically, we aim to contribute to the development of communication-efficient decentralized learning without compromising performance. The potential positive social impact are twofold:

- **Democratizing Access.** For individuals and organizations with constrained infrastructure, our work contributes to the democratization of access to large-scale collaborative training. By reducing communication requirements, we lower the barrier to entry for participating in advanced model development. Such inclusivity can extend the applicability of distributed learning systems to edge environments, thereby promoting more equitable contributions to models trained at scale.
- **Reducing Training Costs.** In data center environments, our approach can alleviate communication bottlenecks of distributed training. This reduction directly translates to shorter total wall-clock training time, thereby lowering the overall costs and energy consumption associated with large-scale distributed training.

No negative societal impacts are identified.

ETHICS STATEMENT

Our research strictly adheres to the ICLR Code of Ethics. The work is foundational, focusing on the algorithmic and theoretical properties of decentralized learning, and does not involve human subjects or the collection of new sensitive data. All experiments were conducted on publicly available, standard academic datasets. We foresee no direct negative societal impacts; on the contrary, by reducing communication overhead, our findings may contribute positively by democratizing access to large-scale distributed training and lowering the associated resource footprint.

REPRODUCIBILITY STATEMENT

We are committed to the reproducibility of our research. Our theoretical claims, including all assumptions and their justifications, are presented in Section 5 with complete, step-by-step proofs provided in Appendix D. Comprehensive details for reproducing our empirical results, including model architectures, data processing, hyperparameter settings, and communication configurations, are well documented in Appendix C.1.

A LIMITATIONS AND POTENTIAL QUESTIONS

***Q**: Why use decentralized **AdamW** in some experiments when the theory is on decentralized **SGD**?*

A:We use decentralized AdamW in some of our experiments for its superior performance in Non-IID settings. Crucially, we note that all reported empirical observations are fully consistent when using decentralized SGD, which directly align with our theoretical analysis (see Figure 1 and Subsection C.3).

***Q**: How does theory part explain "local models in decentralized learning are globally mergeable"?*

A: The theoretical explanation of the "mergeability" of local models in decentralized learning is supported by the result that a globally merged model converges faster to the optimum than the individual local models. Specifically, we provide a fine-grained convergence analysis showing

that the merged model trained via Decentralized SGD (DSGD) can match the convergence rate to optimum of parallel SGD, despite using limited communication. Since the rate of m-agent parallel SGD is superior to that of single local model, the result transitively justifies the merged model's superior performance to that of any individual model, thereby providing theoretical support for their mergeability.

***Q (Hyperparameter tunning)**: How the baselines were tuned in terms of hyperparameter?*

A: All hyperparameters were tuned via grid search based on global generalization performance, with the batch size searched over $\{64, 128\}$. For ResNet-18 trained from scratch on Tiny ImageNet, we searched the learning rate over $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$ for AdamW and $\{1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$ for SGD. For CLIP ViT-B/32 on Tiny ImageNet, we searched the learning rate over $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$ for AdamW and $\{5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$ for SGD. For the optimal hyperparameters selected for our main experiments, please refer to the **Implementation Details** in Appendix C.1 and the additional empirical results in Subsection C.3).

***Q (Comparison with Model Soup)**: How does different or same initialization affect results? The performance gain from merging has been observed in Model Soup (Wortsman et al., 2022a).*

A: We use different initialization schemes and observe consistent performance gains from global merging whether models start from different random initializations or a pretrained state. The majority of our experiments use different initializations, demonstrating that local models in decentralized learning can be effectively merged regardless of their starting points. This is quite surprising as it contrasts with methods like **Model Soup**, which require models to be fine-tuned from an identical pretrained state. Furthermore, our experiments with a shared pretrained state confirm that the performance gains hold in that setting as well (see Figure 1a and Subsection C.3).

***Q (Methodology for Landscape Visualization)**: Please clarify the methodology for visualizing the loss landscape in Figure 1c, including the basis for the visualization grid.*

A: We adopt the visualization tool from (Crisostomi et al., 2024), positioning 16 trained models at the vertices of a regular hexadecagon. Any point within this polygon is an interpolated model whose parameters are determined by Wachspress barycentric coordinates; we then evaluate its cross-entropy loss to generate the contour map. Unlike methods that use random directions, our visualization grid is **deterministically** defined by the models themselves, allowing a direct investigation of their geometric connectivity. The full implementation is available in their official code repository https://github.com/crisostomi/cycle-consistent-model-merging/blob/master/notebooks/plots/plot_loss_contours_n_models.ipynb

***Q (Experimental Scope)**: The empirical findings are restricted to visual tasks.*

A: Our empirical findings primarily focus on tasks within the vision domain. We note that this is consistent with most existing decentralized learning literature (Lin et al., 2021; Kong et al., 2021; Ying et al., 2021; Vogels et al., 2021; Li et al., 2022b; Zehtabi et al., 2025). Extending the experimental setup to broader tasks is a meaningful direction for future research.

***Q**: The findings in Figure 2 (c) that local models eventually converge to a similar state even with limited communication was observed by prior work on gossip algorithms (Jelasity et al., 2005).*

A: In our setting, the local models do not, in fact, converge to a similar state or a single consensus point. This is because our work addresses a more challenging heterogeneous data regime, which differs from the setting in the cited prior work. Instead, we identify an emergent geometric structure where decentralized training guides local models to a shared "high-loss ring" surrounding a central low-loss basin (see Figure 1c). Although the models do not reach a consensus, they remain surprisingly mergeable within this region. This geometric arrangement allows their average, i.e., the globally merged model, to fall directly into the low-loss basin. To the best of our knowledge, we are the first to identify this emergent phenomenon in decentralized learning.

B Additional Background and Related Work

B.1 Decentralized Learning

Modern large-scale model training and inference are predominantly conducted within centralized, high-cost data centers. Driven by mounting constraints on computational resources and power availability (Pilz et al., 2025), both academia and industry are increasingly exploring decentralized training approaches (OpenAI, 2025; Grand View Research, 2024). This paradigm, drawing inspiration from swarm intelligence systems (Bonabeau et al., 1999; Mavrovouniotis et al., 2017), offers a more economical and scalable approach by distributing computational tasks across globally distributed nodes, rather than relying solely on a single central server (Yuan et al., 2022; Borzunov et al., 2023b; Jaghouar et al., 2024; Ramasinghe et al., 2025). A notable illustration of the computational potential through decentralization is the Bitcoin system, which sustains workloads equivalent to a 16 GW power draw (CCAF, 2023), surpassing by a factor of three the estimated 5 GW consumption of the largest AI supercluster under development (Gardizy & Efrati, 2024; OpenAI, 2025).

To provide context, we summarize key algorithmic and theoretical advances in decentralized learning. While our discussion highlights several notable contributions, it is not exhaustive; readers are referred to recent advances and surveys (Zhu et al., 2025; Martínez Beltrán et al., 2023; Singha et al., 2024; Yuan et al., 2024; He et al., 2025; Ramasinghe et al., 2025; Kolehmainen et al., 2025).

Algorithmic Progress in Decentralized Learning. The advancement of decentralized learning algorithms has been primarily driven by the need for communication-efficiency in practical distributed learning. Decentralized algorithms have been refined to handle a variety of realistic scenarios, including time-varying communication topologies (Nedi'c & Olshevsky, 2014; Koloskova et al., 2020; Ying et al., 2021; Takezawa et al., 2023), asynchronous updates (Lian et al., 2018; Xu et al., 2021; Nadiradze et al., 2021; Bornstein et al., 2023; Even et al., 2024), statistical heterogeneity (Tang et al., 2018; Vogels et al., 2021; Le Bars et al., 2023), and robustness to Byzantine failures (He et al., 2022; Ye & Ling, 2025). Moreover, recent works extended beyond standard empirical risk minimization to more structured problem classes, such as compositional (Gao & Huang, 2021), minimax (Xian et al., 2021; Zhu et al., 2023a; Chen et al., 2024), and bi-level optimization (Yang et al., 2022; Gao et al., 2023; Chen et al., 2023). Additionally, privacy concerns in decentralized learning are also critical, with efforts focusing on differentially privacy (Cyffers et al., 2024; Allouah et al., 2024) and data reconstruction attacks (Mrini et al., 2024).

Theoretical Progress in Decentralized Learning. Foundational work on decentralized optimization (Nedic & Ozdaglar, 2009; Sayed, 2014; Yuan et al., 2016; Lian et al., 2017) laid the groundwork for understanding convergence. Building on this, Lu & De Sa (2021) proposed a hierarchical abstraction of decentralization, distinguishing it into three layers, providing a unified view across federated and decentralized paradigms. Koloskova et al. (2020) consolidated synchronous decentralized SGD algorithms with changing communication topologies and local updates, and Even et al. (2024) extended the unifying perspective to asynchronous protocols. More recently, Zehtabi et al. (2025) developed these frameworks further by considering the sporadicity of both communication and computations. On the generalization front, Richards et al. (2020) derived stability-based bounds for decentralized SGD in convex settings, while Sun et al. (2021) extended these to non-convex objectives, revealing a dependency on the spectral gap of the communication graph. This dependency was subsequently refined by Zhu et al. (2022), who introduced a Gaussian weight difference assumption to tight the bound. Complementary results showed that in convex regimes, the generalization of decentralized SGD matches that of centralized SGD (Le Bars et al., 2024), while in non-convex landscapes, decentralization primarily impacts worst-case generalization behavior. To account for unexplained generalization behaviors in decentralized training (Kong et al., 2021; Gurbuzbalaban et al., 2022; Vogels et al., 2023), Zhu et al. (2023b) linked decentralized SGD to random sharpness-aware minimization (SAM), revealing a bias toward flatter minima. Notably, akin to our finding that decentralized learning generalizes when allocated high communication late in training, Zhou et al. (2025) showed that SAM efficiently selects flatter minima when applied in the later stage of training.

Towards Decentralized Training of Foundation Models. Recent advances have shown the feasibility of training large-scale foundation models in decentralized environments. DT-FM (Yuan et al., 2022) introduced tasklet-based scheduling for Transformer training under bandwidth-constrained settings, enabling efficient resource allocation. SWARM Parallelism (Ryabinin et al., 2023) scaled decentralized training through resilient pipeline design and adaptive load balancing. CocktailSGD

(Wang et al., 2023) further improved efficiency via a combination of decentralization, gradient sparsification, and quantization for LLM fine-tuning. On the inference side, Petal (Borzunov et al., 2023a) exploited peer-to-peer networks to amortize computational costs across heterogeneous nodes. Most recently, Intellect (Jaghouar et al., 2024), building on Diloco (Douillard et al., 2023), leveraged hybrid parallelism, i.e., both data and model parallelism, to collaboratively train models with billions of parameters. NoLoCo (Kolehmainen et al., 2025) further extended Diloco to gossip-type decentralized settings. For a broad survey of large-scale deep learning practice, see Shen et al. (2024; 2025).

B.2 Implicit Bias of Decentralized Learning

The concept of implicit bias, i.e., the intrinsic preference of learning algorithms for solutions with certain properties, has emerged as a key concept in explaining the empirical success of modern deep learning (Li et al., 2022c; Vardi, 2023; Lyu, 2024). Recent studies have highlighted intriguing distinctions between decentralized stochastic gradient descent (DSGD) and its centralized counterpart (CSGD). Gurbuzbalaban et al. (2022) demonstrated that under certain conditions, DSGD operating on large, sparse topologies exhibits heavier-tailed parameter distributions compared to CSGD. Zhang et al. (2021) showed that decentralization introduces a landscape-dependent noise, which can improve tolerance to larger learning rates. This observation aligns with findings by Vogels et al. (2023), who revealed that collaboration in decentralized settings permits the use of larger learning rates. Zhu et al. (2023b) first explicitly characterized the implicit bias of decentralized SGD by establishing its connection with random sharpness-aware minimization, proving the existence of flatne bias in decentralized training. Complementing this, Cao et al. (2024) offered a detailed analysis of the interplay between flatness and optimization in DSGD, particularly its ability to escape local minima. More recently, Wu & Sun (2024) investigated the implicit regularization properties of decentralized optimization in non-convex sparse regression problems, recovering the convergence rates achieved by gradient descent in centralized settings.

Comparison with Zhu et al. (2023b). We note that Zhu et al. (2023b) has highlighted the generalization benefits of decentralized learning, but key differences exist in terms of the experimental setup and the insights derived. While Zhu et al. (2023b) focused on IID scenarios and specific cases involving exceptionally large batch sizes, we consider the more realistic non-IID setting using standard batch sizes. This shift in focus allows us to uncover phenomena not observed by Zhu et al. (2023b), including insights into communication allocation strategies.

B.3 Model Merging

Mode Connectivity and Model Merging Techniques. Recent works on *(Linear) Mode Connectivity* have advanced our understanding of the complex loss landscape in neural networks. Freeman & Bruna (2017); Draxler et al. (2018); Garipov et al. (2018); Nagarajan & Kolter (2019); Frankle et al. (2020) discovered that different solutions of deep neural networks can be merged together by simply averaging their parameters. Sonthalia et al. (2025) further showed that the solutions may form a star domain. We note that these phenomenon are observed in the following scenarios:

- *Shared initialization (Frankle et al., 2020; Fort et al., 2020; Zhou et al., 2023).* Models are initialized from a pretrained checkpoint.
- *Homogeneous data distribution (Wortsman et al., 2022a).* Models are trained on homogeneous data distribution.
- *Permutation (Ainsworth et al., 2023; Entezari et al., 2022).* Models are independently trained. The neurons of one model are permuted to match the neurons of the other while maintaining a functionally equivalent network.

These findings have inspired a range of model merging techniques for various applications. Izmailov et al. (2018); Matena & Raffel (2022); Rame et al. (2022; 2023); Wortsman et al. (2022a;b) found that merging the parameters of models that start from the same pretrained model and finetune over the same task leads to improved generalization and robustness. Furthermore, Ilharco et al. (2022); Li et al. (2022a); Ilharco et al. (2023); Ortiz-Jimenez et al. (2023); Yadav et al. (2023) showed that merging models that finetune over different tasks enables multi-task abilities.

Comparisons with Model Merging Literature. Our results show that mode connectivity, or mergeability, can still emerge in decentralized learning, even when the local models are initialized *differently*, trained on highly *heterogeneous* data, and merged *without* any permutation. Our findings offer new insights into both model merging techniques and the geometry of the neural network loss landscape, which we anticipate will motivate further advances in both areas.

C Additional Experiments

C.1 Experimental Setups

Computational Resources. The experiments were conducted on a computing facility equipped with 80 GB NVIDIA® A100™ GPUs. All implementations are based on PyTorch, and computations are distributed across multiple GPUs for efficiency.

Dataset. We use three widely adopted image classification datasets: CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and Tiny ImageNet (Le & Yang, 2015). CIFAR-10 consists of 60,000 RGB images across 10 classes, while CIFAR-100 contains 60,000 RGB images across 100 classes. The images in both datasets have a spatial resolution of 32×32 pixels. Tiny ImageNet is a subset of the ImageNet dataset, comprising 100,000 images drawn from 200 classes, with each image resized to 64×64 pixels. It provides a mid-scale benchmark that is more challenging than CIFAR datasets but less computationally demanding than training full ImageNet. To incorporate data augmentation, we employ a combination of RandomCrop with 4-pixel padding, RandomHorizontalFlip, and RandAugment with `num_ops`=2 and `magnitude`=9.

Details of Decentralized Learning. We simulate a heterogeneous decentralized learning environment. For our main experiments (Figure 1a and Figure 1b), we use $m = 32$ agents, while for other experiments, including the sliding window experiments (Figure 2) and the loss landscape visualizations (Figure 1c), we use $m = 16$ agents. The number of agents for the visualization was chosen as 16 for clarity, as a plot with 32 models would be visually crowded. In all configurations, we employ a Dirichlet distribution characterized by $\alpha = 0.1$ to partition the data among agents. The Dirichlet distribution is commonly used to partition data in federated learning scenarios, as it allows for the control of label distribution skew among agents (Yurochkin et al., 2019; Hsu et al., 2019). A smaller α results in more imbalanced data distributions, where some agents predominantly receive data from a limited number of classes, while a larger α results in more uniform label distributions across agents. This configuration effectively captures the realistic non-IID nature of decentralized learning, where different agents may have access to personalized data reflective of their local environments.

- **Communication Graph.** We evaluate three decentralized communication topologies: random graph, ring graph, and exponential graph. In the random graph setting, during each communication round, each agent selects a random subset of its neighbors for gossip averaging. For "R 1", each agent selects exactly one random neighbor in each round. For "R 0.2", each agent selects one neighbor with a probability of 0.2 and continues local training without communication with a probability of 0.8. The ring graph enforces a fixed cyclic communication structure, while the exponential graph ensures connectivity by allowing agents to communicate with exponentially increasing distances in the ring graph.
- **Communication Rounds and Local Steps.** The decentralized learning process is conducted over $T = 300$ communication rounds. We use a local training step size of $H = 100$ batches per communication round to balance communication and computation costs.
- **Local Data per Agent.** Each agent is assigned a subset of the dataset with a fixed size of 4096 samples, drawn according to a Dirichlet distribution to simulate realistic non-IID scenarios.

Model Architecture. To ensure a representative comparison across different model families, we adopt ResNet-18 (He et al., 2016) and CLIP ViT-B/32 (Radford et al., 2021) as backbone architectures in our experiments. ResNet-18 is a widely used lightweight convolutional neural network that serves as a canonical example of traditional CNN-based architectures. In contrast, CLIP ViT-B/32 is a transformer-based vision model pre-trained on large-scale image–text pairs. For experiments on Tiny ImageNet, where images are resized to 64×64 pixels, we adjust the CLIP visual encoder to handle the lower resolution. With a patch size of 32, each image yields 4 visual tokens arranged in a 2×2 grid, plus a [CLS] token, resulting in a 5-token input sequence.

Implementation Details. All hyperparameters are tuned through grid search based on global generalization performance (see Definition 1). For experiments using decentralized SGD, the optimal learning rates were found to be 1×10^{-2} for ResNet-18 (trained from scratch) and 1×10^{-3} for CLIP ViT-B/32. When using decentralized AdamW, the optimal learning rate is 5×10^{-4} for ResNet-18 (both when trained from scratch and fine-tuned from ImageNet-pretrained weights) and 1×10^{-5} for the pretrained CLIP ViT-B/32 on Tiny ImageNet. For all experiments, weight decay is set to 5×10^{-4} and the batch size is selected as 128. The key empirical results remain consistent across these optimizer and hyperparameter choices, indicating that our conclusions are stable and not sensitive to specific hyperparameter configurations.

Details of Loss Landscape Visualization in Figure 1c. To analyze the geometric connections among models after decentralized training, we visualize the loss landscape spanning their parameter spaces. We adopt the visualization tool from (Crisostomi et al., 2024), which is specifically designed to analyze the interpolation space within the convex hull formed by a given set of models. In our implementation, we position the 16 trained models at the vertices of a regular hexadecagon. Any point within this polygon represents an interpolated model, whose parameters are a weighted sum of the parameters of the 16 vertex models; the weights are determined by the point's Wachspress barycentric coordinates. We then evaluate the cross-entropy loss of each interpolated model on the entire test set to generate the final loss contour map, as shown in Figure 1c. The implementation is available in their notebook https://github.com/crisostomi/cycle-consistent-model-merging/blob/master/notebooks/plots/plot_loss_contours_n_models.ipynb within the official code repository for (Crisostomi et al., 2024). We note two key aspects of this visualization approach:

- **Focus on Convex Combinations.** For points outside the polygon, one or more of their barycentric coordinates become negative, corresponding to an extrapolation, which is often unstable. This visualization approach is consistent with Definition 2, focusing on the space of convex combinations among the models.
- **Deterministic Grid vs. Random Directions.** Notably, the visualization method differs from approaches that use random directions to probe the landscape of a single model, as our visualization grid is defined directly by the 16 models themselves. This allows us to directly investigate the geometric connectivity and interpolation properties among this predefined set of models.

Computational Resource Requirements and Runtime. To enhance accessibility for researchers working with diverse computational environments, our code includes a centralized simulation of decentralized training. This enables the reproduction and extension of our decentralized learning experiments using fewer GPUs. A single decentralized AdamW training experiment with 16 agents using ResNet-18 on the Tiny ImageNet dataset requires approximately 15 GB of GPU memory and can be conducted on a single GPU with sufficient memory, such as an NVIDIA V100, RTX 3090, RTX 4090, or A100. On an A100 GPU, the typical runtime is approximately 8 hours for 300 communication rounds, each comprising 100 local steps. For the CLIP ViT-B/32 model, the memory demand rises to about 30 GB, yet it remains feasible on a single A100 GPU, with a runtime of approximately 12 hours under the same configuration of 300 communication rounds and 100 local steps per round.

C.2 Practical Evaluation Metrics

The standard evaluation metric of parallel and federated learning is the accuracy of the global model.

Definition C.1 (Test Accuracy of Global Model)**.** *The accuracy of the global model θ is defined as:*

$$\mathrm{Acc}(\theta) \triangleq \frac{1}{m}\sum_{k\in\mathcal{V}} \mathbb{E}_{\xi_k\sim\mathcal{D}_k}\,\mathrm{Acc}(\theta;\xi_k) \overset{\text{if IID}}{=} \mathbb{E}_{\xi\sim\mathcal{D}}\,\mathrm{Acc}(\theta;\xi).$$

In decentralized learning, models are often evaluated in the absence of a full consensus model θ due to data heterogeneity and limited training time. Two major metrics are adopted in this scenario.

Definition C.2 (Average Local Test Accuracy)**.** *The average accuracy of agents $k \in \mathcal{V}$ is defined as:*

$$\underbrace{\overline{\mathrm{Acc}}(\{\theta_k\}_{k\in\mathcal{V}}) \triangleq \frac{1}{m}\sum_{k\in\mathcal{V}} \mathbb{E}_{\xi_k\sim\mathcal{D}_k}\,\mathrm{Acc}(\theta_k;\xi_k)}_{\textit{Average Test accuracy on the local distribution across agents}} \overset{\text{if IID}}{=} \frac{1}{m}\sum_{k\in\mathcal{V}} \mathbb{E}_{\xi\sim\mathcal{D}}\,\mathrm{Acc}(\theta_k;\xi).$$

Remark C.1 (Local Generalization)**.** This metric aims to address the following question in decentralized learning: *how well do local models* $\{\theta_k\}_{k\in\mathcal{V}}$*, with the aid of peer-to-peer communication, generalize to their local (personalized) data distribution* $\mathcal{D}_l$*?* This is the standard evaluation metric in personalized decentralized settings, where the goals are to optimize local objectives.

However, in real-world scenarios, local data distributions are often heterogeneous and not guaranteed to be IID across agents. In such settings, an important goal is to understand how well local models, trained on limited local data, generalize to the global data distribution. To account for this, we adopt the following *average global test accuracy*, a proxy of average global population risk, as the primary evaluation metric, which quantifies how well local models generalize to the global distribution.

Definition C.3 (Average Global Test Accuracy)**.** *The average accuracy of agents* $k \in \mathcal{V}$ *is defined as:*

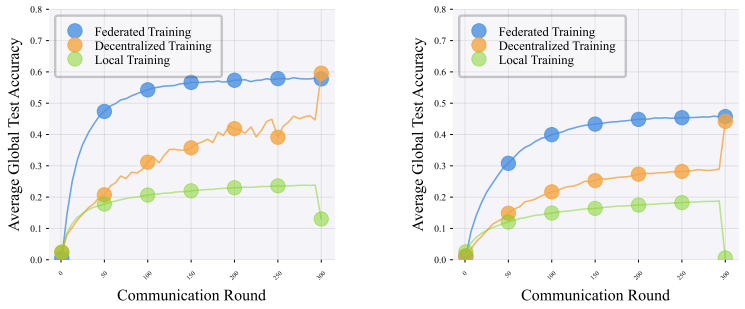
$$\underbrace{\overline{\mathrm{Acc}}(\{\theta_k\}_{k\in\mathcal{V}}) = \frac{1}{m}\sum_{k\in\mathcal{V}} \mathrm{Acc}(\theta_k),}_{\textit{Average Accuracy across agents}} \quad \textit{where } \mathrm{Acc}(\cdot) \triangleq \underbrace{\frac{1}{m}\sum_{l\in\mathcal{V}} \mathbb{E}_{\xi_l\sim\mathcal{D}_l}\, \mathrm{Acc}(\cdot;\xi_l)}_{\textit{Test accuracy on the global distribution}} .$$

Remark C.2 (Global Generalization)**.** This metric is specifically designed to address a core research question in fully decentralized learning with non-IID data: *how well do local models* $\{\theta_k\}_{k\in\mathcal{V}}$*, trained with limited peer-to-peer synchronization, generalize to the global data distribution* $\mathcal{D}$*?* We note that this objective is particularly critical in the highly non-IID scenarios we study, where local models drift significantly apart. Unlike federated learning that measures the performance of a global model, this metric offers a more realistic evaluation for decentralized settings where no central server is present.

C.3 Additional Experiments

C.3.1 Different number of agents and optimizers

We conduct additional experiments by varying the number of agents (from 16 to 32) and comparing different optimizers (SGD to AdamW). The effect of single merging remains consistent.



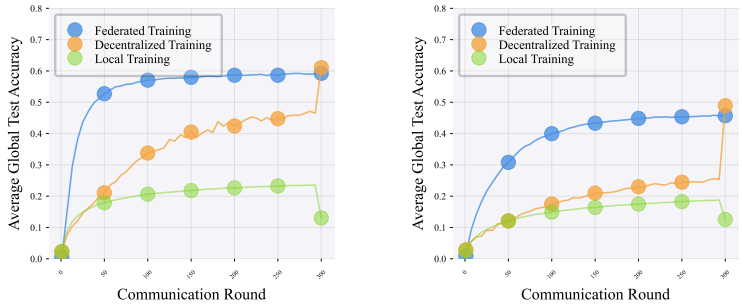
(a) CLIP ViT-B/32 (b) ResNet-18 (w/o pretraining)

Figure C.1: **(a, b)**: Global test accuracy (see Definition 1) of CLIP ViT-B/32 (a) and ResNet-18 (b) trained on Tiny ImageNet using FedAdamW (blue), decentralized AdamW (orange), and one-shot FedAdamW (green), distributed across **16** agents with high data heterogeneity (Dirichlet $\alpha = 0.1$). Decentralized training involves each agent syncing model parameters with a random peer per round with a probability of 0.2, with a single global merging at the final round (see details in Appendix C.1).

C.3.2 Different communication topologies

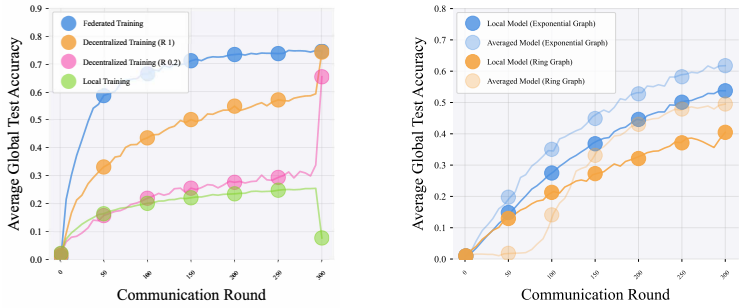
We also conduct additional experiments with different communication topologies to examine whether the empirical results remain consistent. New observations are summarized below.

• **Models remain mergeable under different number of peers.** We evaluate two settings (random topology with $R = 0.2$ and $R = 1$; see "Communication Graph" in Appendix C.1). As shown in



(a) CLIP ViT-B/32 (b) ResNet-18 (w/o pretraining)

Figure C.2: **(a, b)**: Global test accuracy (see Definition 1) of CLIP ViT-B/32 (a) and ResNet-18 (b) trained on Tiny ImageNet using FedAdamW (blue), decentralized AdamW (orange), and one-shot FedAdamW (green), distributed across **32** agents with high data heterogeneity (Dirichlet $\alpha = 0.1$). Decentralized training involves each agent syncing model parameters with a random peer per round with a probability of 0.2, with a single global merging at the final round (see details in Appendix C.1).



(a) Different Number of Peers R (b) Different Topologies

Figure C.3: Global test accuracy (see Definition 1) of training ResNet-18 on Tiny ImageNet, distributed across 16 agents with high heterogeneity (Dirichlet $\alpha = 0.1$; see details in Appendix C.1). We evaluate the effects of different **(a)** number of peers R, and **(b)** communication topologies. Pretrained weights are used only in (a).

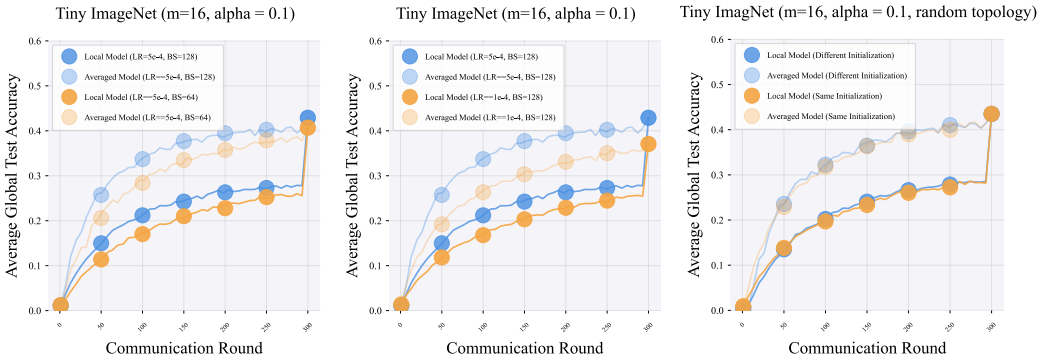
Figure C.3a, performance improvements are consistently observed, with more significant gains in the $R = 0.2$ case.

• **Models remain mergeable across different communication topologies.** We evaluate two topologies: exponential and ring graphs. As shown in Figure C.3b, both topologies preserve the mergeability of local models, with exponential graphs yielding slightly better generalization for both local and merged models. The trend of mergeability persists across topologies throughout training, though performance may vary.

C.3.3 Different hyperparameters, dataset, and heterogeneity level

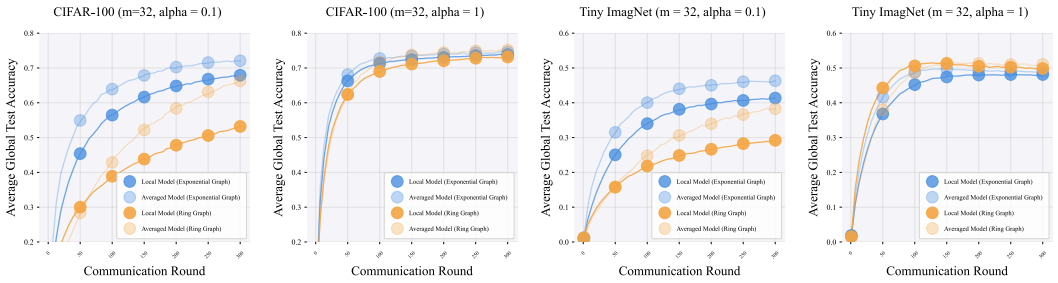
We further experiments with different hyperparameters (e.g., learning rate and batch size), dataset, and degree of data heterogeneity to examine whether the empirical observations remain consistent.

Summary. Consistent generalization improvement of a single global merging across a wide range of settings are observed, including different hyperparameter setups, datasets, degree of data heterogeneity, model architectures, optimizers, initialization schemes, and communication topologies.



(a) Different Batch Sizes (b) Different Learning Rates (c) Different Initialization Schemes

Figure C.4: Global test accuracy (see Definition 1) of training ResNet-18 on Tiny ImageNet with decentralized AdamW, distributed across 16 agents with high heterogeneity (Dirichlet $\alpha = 0.1$; see details in Appendix C.1). We evaluate the effects of different **(a)** batch sizes (64 vs. 128), **(b)** learning rates (5×10^{-4} vs. 1×10^{-4}), and **(c)** different initialization schemes.



(a) (b) (c) (d)

Figure C.5: Global test accuracy (see Definition 1) of training ResNet-18 with decentralized AdamW across 32 agents under different levels of data heterogeneity (Dirichlet $\alpha = 0.1$ (**a**, **c**) vs. $\alpha = 1.0$ (**b**, **d**); see Appendix C.1). Results are on both CIFAR-100 (**a**, **b**) and Tiny ImageNet (**c**, **d**).

D Theory

This section provides the proofs of the main theoretical results presented in this paper. For simplicity, and following the setup in the existing literature, we assume that the sample size of local agents is $n_k = n$ for all $k \in \mathcal{V}$.

Lemma D.1 (Consensus Distance Recursion under Local Updates (Kong et al., 2021)). *Suppose Assumption 1–Assumption 3 hold. Let $\theta_k^{(t)}$ be the local parameter on client k at t-th step, and denote their average by $\bar{\theta}^{(t)} = \frac{1}{m}\sum_{k=1}^m \theta_k^{(t)}$. Define the* consensus distance *and the* average gradient norm *at round t by $\Xi_t^2 = \frac{1}{m}\sum_{k=1}^m \|\theta_k^{(t)} - \bar{\theta}^{(t)}\|^2$ and $\phi_t^2 = \frac{1}{m}\sum_{k=1}^m \|\nabla \mathcal{L}_k(\theta_k^{(t)})\|^2$, where $\mathcal{L}_k(\theta) = \mathbb{E}_{\xi_k \sim \mathcal{D}_k}[\mathcal{L}(\theta; \xi_k)]$. Let $\eta > 0$ the learning rate, and σ^2 the variance bound from Assumption 3. Then there exists a constant $p > 0$ (see Assumption 1) such that for all $t \geq 0$, the following inequality holds:*

$$\mathbb{E}\left[\Xi_{t+1}^2\right] \leq \left(1 - \frac{p}{2}\right)\Xi_t^2 + \frac{12(1-p)}{p}\eta^2\left(\phi_t^2 + \sigma^2\right), \tag{D.1}$$

where the expectation is taken over the stochastic gradients in the t-th update phase.

Proof. For completeness, we provide the proof of Lemma D.1, with minor corrections and additional details. In decentralized SGD (Algorithm 1 with SGD as the local optimizer), each agent $k \in \mathcal{V}$

performs at each iteration

$$\theta_k^{(t+1)} = \sum_{l=1}^{m} W_{k,l}\big(\theta_l^{(t)} - \eta\,\nabla\mathcal{L}_l(\theta_l^{(t)};\,\xi_l^{(t)})\big).$$

In matrix form, letting

$$\Theta^{(t)} = [\theta_1^{(t)},\dots,\theta_m^{(t)}] \in \mathbb{R}^{d\times m}, \quad \nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)}) = [\nabla\mathcal{L}_1(\theta_1^{(t)};\xi_1^{(t)}),\dots,\nabla\mathcal{L}_m(\theta_m^{(t)};\xi_m^{(t)})],$$

we have

$$\Theta^{(t+1)} = \big(\Theta^{(t)} - \eta\,\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})\big)\,W.$$

The consensus matrix after mixing is

$$\bar{\Theta}^{(t+1)} = \Theta^{(t+1)}\,\tfrac{1}{m}\mathbf{1}\mathbf{1}^\top = \big(\Theta^{(t)} - \eta\,\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})\big)\,\tfrac{1}{m}\mathbf{1}\mathbf{1}^\top,$$

since $\mathbf{1}^\top W = \mathbf{1}^\top$.

Thus the consensus distance satisfies

$$m\,\Xi_{t+1}^2 = \big\|\Theta^{(t+1)} - \bar{\Theta}^{(t+1)}\big\|_F^2 = \big\|\big(\Theta^{(t)} - \eta\,\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})\big)\,(W - \tfrac{1}{m}\mathbf{1}\mathbf{1}^\top)\big\|_F^2.$$

By Assumption 1, for any $\Theta \in \mathbb{R}^{d\times m}$,

$$\mathbb{E}_W\big\|\Theta W - \bar{\Theta}\big\|_F^2 \le (1-\rho)\big\|\Theta - \bar{\Theta}\big\|_F^2,$$

we obtain,

$$m\,\Xi_{t+1}^2 \le (1-p)\Big\|\Theta^{(t)}(I - \tfrac{1}{m}\mathbf{1}\mathbf{1}^\top) - \eta\,\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})(I - \tfrac{1}{m}\mathbf{1}\mathbf{1}^\top)\Big\|_F^2.$$

Applying the inequality $\|A+B\|_F^2 \le (1+\alpha)\|A\|_F^2 + (1+1/\alpha)\|B\|_F^2$ with $\alpha = \frac{p}{2}$ gives

$$\begin{aligned} m\,\Xi_{t+1}^2 &\le (1-p)\Big[(1+\tfrac{p}{2})\big\|\Theta^{(t)}(I - \tfrac{1}{m}\mathbf{1}\mathbf{1}^\top)\big\|_F^2 + (1+\tfrac{2}{p})\,\eta^2\,\big\|\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})\big\|_F^2\Big] \\ &\le \Big(1-\tfrac{p}{2}\Big)\,m\,\Xi_t^2 \;+\; \frac{6(1-p)}{p}\,\eta^2\,\big\|\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})\big\|_F^2, \end{aligned}$$

where we used $(1+p/2) \le 1+p$ and $(1+2/p) \le 6/p$ for $p \in (0,1)$.

We now decompose the stochastic gradient as

$$\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)}) = \nabla\mathcal{L}(\Theta^{(t)}) + \big[\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)}) - \nabla\mathcal{L}(\Theta^{(t)})\big],$$

so by Young's Inequality, we have

$$\big\|\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})\big\|_F^2 \le 2\big\|\nabla\mathcal{L}(\Theta^{(t)})\big\|_F^2 + 2\big\|\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)}) - \nabla\mathcal{L}(\Theta^{(t)})\big\|_F^2.$$

Taking expectation over $\xi^{(t)}$ and invoking Assumption 3, we get

$$\mathbb{E}\big[\|\nabla\mathcal{L}(\Theta^{(t)};\xi^{(t)})\|_F^2\big] \le 2\|\nabla\mathcal{L}(\Theta^{(t)})\|_F^2 + 2\,\sigma^2 m.$$

Substituting back and dividing by m yields

$$\mathbb{E}\big[\Xi_{t+1}^2\big] \le \Big(1-\tfrac{p}{2}\Big)\,\Xi_t^2 \;+\; \frac{12(1-p)}{p}\,\eta^2\big(\phi_t^2 + \sigma^2\big),$$

which completes the proof. □

Corollary D.2 ((Kong et al., 2021)). *Define the* consensus distance *and the* average gradient norm *at round t by $\Xi_t^2 = \frac{1}{m}\sum_{k=1}^m \|\theta_k^{(t)} - \bar{\theta}^{(t)}\|^2$ and $\phi_t^2 = \frac{1}{m}\sum_{k=1}^m \|\nabla\mathcal{L}_k(\theta_k^{(t)})\|^2$, where $\mathcal{L}_k(\theta) = \mathbb{E}_{\xi_k\sim\mathcal{D}_k}[\mathcal{L}(\theta;\xi_k)]$. Under the conditions of Lemma D.1, suppose that for all iterations t, the gradient norms are uniformly bounded by a constant ϕ, i.e. $\phi_t^2 \le \phi^2$, $\forall t \in \{1,\dots,T\}$. Then the expected consensus distance satisfies*

$$\mathbb{E}\big[\Xi_t^2\big] \;\le\; \frac{24\,(1-p)\,\eta^2}{p^2}\Big(\phi^2 + \sigma^2\Big).$$

In the general case where the gradient-norms change slowly, i.e., $\phi_t^2 \le (1+\frac{p}{4})\,\phi_{t+1}^2$, we have

$$\mathbb{E}\big[\Xi_t^2\big] \;\le\; \frac{48\,(1-p)\eta^2}{p^2}\left(\phi_{t-1}^2 + \sigma^2\right).$$

The expectation here is taken over the stochastic gradients in the t-th update phase.

Proof. Consider the key recursion from Lemma D.1:

$$\mathbb{E}\big[\Xi_{t+1}^2\big] \;\le\; \Big(1-\tfrac{p}{2}\Big)\,\Xi_t^2 \;+\; \frac{12\,(1-p)}{p}\,\eta^2\big(\phi_t^2+\sigma^2\big).$$

(1) Special Case: uniformly bounded gradient norms.

Assume $\phi_t^2 \le \phi^2$. Unrolling the above gives

$$\mathbb{E}\big[\Xi_{t+1}^2\big] \;\le\; \sum_{i=0}^{t-1}\Big(1-\tfrac{p}{2}\Big)^i\,\frac{12\,(1-p)}{p}\,\eta^2\,(\phi^2+\sigma^2).$$

Since $\sum_{i=0}^{t-1}(1-\frac{p}{2})^i \le \frac{2}{p}$, we can bound the consensus distance as

$$\mathbb{E}\big[\Xi_{t+1}^2\big] \;\le\; \frac{12\,(1-p)}{p}\,\eta^2(\phi^2+\sigma^2) \;\times\; \frac{2}{p} \;=\; \frac{24\,(1-p)\,\eta^2}{p^2}\,(\phi^2+\sigma^2),$$

which yields the first claim.

(2) Special Case: slowly changing gradient norms.

If $\phi_t^2 \le (1+\frac{p}{4})\,\phi_{t+1}^2$, and since

$$\Big(1-\tfrac{p}{2}\Big)^i\Big(1+\tfrac{p}{4}\Big)^i \;\le\; \Big(1-\tfrac{p}{4}\Big)^i,$$

the consensus distance satisfies

$$\begin{aligned}\mathbb{E}\left[\Xi_{t+1}^2\right] &\le \sum_{i=0}^{t-i-1}\left(1-\frac{p}{2}\right)^i\frac{12(1-p)\eta^2\left(\phi_{t-1}^2+\sigma^2\right)}{p}\\ &\le \sum_{i=0}^{t-1}\left(1-\frac{p}{4}\right)^i\frac{12(1-p)\eta^2(\phi_{t-1}^2+\sigma^2)}{p} \le \frac{48(1-p)\eta^2}{p^2}\left(\phi_{t-1}^2+\sigma^2\right). \qquad \text{(D.2)}\end{aligned}$$

□

Proposition D.3 (Implicit Bias of Decentralized SGD (Zhu et al., 2023b)). *Assume $\mathcal{L} \in C^4\left(\mathbb{R}^d\right)$, the globally averaged model of decentralized SGD (DSGD), defined by $\bar{\theta}^{(t)} = \frac{1}{m}\sum_{k=1}^m \theta_k^{(t)}$, follows the following gradient descent direction:*

$$\mathbb{E}_{\xi^{(t)}}[\bar{\theta}^{(t+1)}] = \bar{\theta}^{(t)} - \eta\cdot\mathbb{E}_{\epsilon^{(t)}\sim\mathcal{N}(0,\Gamma^{(t)})}\left[\nabla\mathcal{L}(\bar{\theta}^{(t)}+\epsilon^{(t)})\right] + \delta^{(t)},$$

where $\Gamma^{(t)} = \frac{1}{m}\sum_{k=1}^m(\theta_k^{(t)}-\bar{\theta}^{(t)})(\theta_k^{(t)}-\bar{\theta}^{(t)})^\top \in \mathbb{R}^{m\times m}$ denotes the consensus distance matrix, and $\delta^{(t)} = \Theta\left(\frac{\eta}{m}\sum_{k=1}^m\|\theta_k^{(t)}-\bar{\theta}^{(t)}\|_2^3\right)$ denotes the high-order terms. The first expectation eliminates the randomness from sampled data $\xi^{(t)} = \{\xi_k^{(t)}\}_{k\in\mathcal{V}}$ at step (t).

We can then control the expected squared distance between two consecutive steps of the globally averaged model with Corollary D.4.

Corollary D.4. *Under the assumptions in Proposition D.3, the expected squared distance between two consecutive iterates of decentralized SGD can be bounded as follows:*

$$\mathbb{E}_{\xi^{(t)}}\left\|\bar{\theta}^{(t+1)}-\bar{\theta}^{(t)}\right\|^2 \le \frac{\sigma^2}{m} + \eta^2\left\|\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big) + \nabla\,\mathrm{Tr}\left(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)}\right) + \delta^{(t)}\right\|^2. \qquad \text{(D.3)}$$

Proof. Denote $\gamma^{(t+1)} = \mathbb{E}_{\xi^{(t)}}\bar{\theta}^{(t+1)} - \bar{\theta}^{(t)}$. We can expand the expected distance as follows:

$$\mathbb{E}_{\xi^{(t)}}\left\|\bar{\theta}^{(t+1)}-\bar{\theta}^{(t)}\right\|^2 = \mathbb{E}_{\xi^{(t)}}\left\|\bar{\theta}^{(t+1)}\right\|^2 - \left\|\bar{\theta}^{(t)}\right\|^2 - 2(\bar{\theta}^{(t)})^\top\gamma^{(t+1)}$$

$$
\begin{aligned}
&= \operatorname{Tr}\left(\operatorname{Cov}(\bar{\theta}^{(t+1)})\right) + \left\|\mathbb{E}_{\xi^{(t)}}\,\bar{\theta}^{(t+1)}\right\|^2 - \left\|\bar{\theta}^{(t)}\right\|^2 - 2(\bar{\theta}^{(t)})^\top \gamma^{(t+1)} \\
&= \operatorname{Tr}\left(\operatorname{Cov}(\bar{\theta}^{(t+1)})\right) + \left\|\mathbb{E}_{\xi^{(t)}}[\bar{\theta}^{(t+1)} - \bar{\theta}^{(t)}]\right\|^2 \\
&= \operatorname{Tr}\left(\operatorname{Cov}(\frac{1}{m}\sum_{k=1}^{m}\nabla\mathcal{L}(\theta_k^{(t)};\xi_k^{(t)}))\right) + \left\|\mathbb{E}_{\xi^{(t)}}[\bar{\theta}^{(t+1)} - \bar{\theta}^{(t)}]\right\|^2, \qquad \text{(D.4)}
\end{aligned}
$$

where in the second equality we substitute $\mathbb{E}_{\xi^{(t)}}\bar{\theta}^{(t+1)}$ with $\gamma^{(t+1)} + \bar{\theta}^{(t)}$. The final equality is derived from the update rule:

$$
\bar{\theta}^{(t+1)} = \bar{\theta}^{(t)} - \frac{1}{m}\sum_{k=1}^{m}\nabla\mathcal{L}(\theta_k^{(t)};\xi_k^{(t)}).
$$

According to the convexity of the vector norm and the fact that

$$
\operatorname{Tr}\left(\operatorname{Cov}(\frac{1}{m}\sum_{k=1}^{m}\nabla\mathcal{L}(\theta_k^{(t)};\xi_k^{(t)}))\right) = \mathbb{E}_{\xi^{(t)}}\left\|\frac{1}{m}\sum_{k=1}^{m}\nabla\mathcal{L}(\theta_k^{(t)};\xi_k^{(t)}) - \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\xi_k^{(t)}}\nabla\mathcal{L}(\theta_k^{(t)};\xi_k^{(t)})\right\|^2, \qquad \text{(D.5)}
$$

we then complete the proof by applying Proposition D.3 and the bounded noise assumption in Assumption 3. □

Corollary D.5. *Let* $\Gamma^{(t)} = \frac{1}{m}\sum_{k=1}^{m}(\theta_k^{(t)} - \bar{\theta}^{(t)})(\theta_k^{(t)} - \bar{\theta}^{(t)})^\top \in \mathbb{R}^{d\times d}$, *where* $\bar{\theta}^{(t)} = \frac{1}{m}\sum_{k=1}^{m}\theta_k^{(t)} \in \mathbb{R}^d$ *denotes the globally averaged model across* m *agents. Assume the loss function* $\mathcal{L} \in C^4(\mathbb{R}^d)$, *with its fourth derivative* $\nabla^4\mathcal{L}(\cdot)$ *uniformly bounded by a constant* $L_4 > 0$, *i.e.,* $\|\nabla^4\mathcal{L}(\cdot)\| \le L_4$. *Then, for* $\epsilon^{(t)} \sim \mathcal{N}(0, \Gamma^{(t)})$, *the expected gradient perturbation satisfies:*

$$
\begin{aligned}
&\mathbb{E}_{\epsilon^{(t)}\sim\mathcal{N}(0,\Gamma^{(t)})}\left[\nabla\mathcal{L}(\bar{\theta}^{(t)} + \epsilon^{(t)})\right] - \nabla\mathcal{L}(\bar{\theta}^{(t)}) \\
&\qquad = \nabla\operatorname{Tr}\left(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)}\right) + \mathbb{E}_{\epsilon^{(t)}\sim\mathcal{N}(0,\Gamma^{(t)})}\left[R_3(\epsilon^{(t)})\right], \qquad \text{(D.6)}
\end{aligned}
$$

where $\|R_3(\epsilon^{(t)})\|$ *is bounded by* $\frac{L_4}{24}\|\epsilon^{(t)}\|^3$.

Proof. We apply the third-order Taylor expansion to $\nabla\mathcal{L}$ around $\bar{\theta}^{(t)}$:

$$
\nabla\mathcal{L}(\bar{\theta}^{(t)} + \epsilon^{(t)}) = \nabla\mathcal{L}(\bar{\theta}^{(t)}) + \nabla^2\mathcal{L}(\bar{\theta}^{(t)})\epsilon^{(t)} + \frac{1}{2}\nabla^3\mathcal{L}(\bar{\theta}^{(t)})[\epsilon^{(t)}, \epsilon^{(t)}] + R_3(\epsilon^{(t)}),
$$

with the remainder:

$$
R_3(\epsilon^{(t)}) = \int_0^1 \frac{(1-\tau)^3}{6}\nabla^4\mathcal{L}(\bar{\theta}^{(t)} + \tau\epsilon^{(t)})[\epsilon^{(t)}, \epsilon^{(t)}, \epsilon^{(t)}]d\tau.
$$

Taking expectations over $\epsilon^{(t)} \sim \mathcal{N}(0, \Gamma^{(t)})$, since $\mathbb{E}[\epsilon^{(t)}] = 0$, the linear term vanishes. The quadratic term $\mathbb{E}\left[\nabla^3\mathcal{L}(\bar{\theta}^{(t)})[\epsilon^{(t)}, \epsilon^{(t)}]\right]$ simplifies to $\nabla\operatorname{Tr}\left(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)}\right)$ due to properties of the Gaussian distribution. The remainder bound can be bounded as

$$
\|R_3(\epsilon^{(t)})\| \le \int_0^1 \frac{(1-\tau)^3}{6}L_4\|\epsilon^{(t)}\|^3 d\tau = L_4\|\epsilon^{(t)}\|^3 \cdot \frac{1}{6}\int_0^1 (1-\tau)^3 d\tau.
$$

Since $\int_0^1 (1-\tau)^3 d\tau = \frac{1}{4}$, we have:

$$
\|R_3(\epsilon^{(t)})\| \le L_4\|\epsilon^{(t)}\|^3 \cdot \frac{1}{6}\cdot\frac{1}{4} = \frac{L_4}{24}\|\epsilon^{(t)}\|^3.
$$

□

For comparison, we restate the convergence rate of DSGD by Koloskova et al. (2020).

Assumption D.1 (L-smoothness). *Each population risk $\mathcal{L}_k = \mathbb{E}_{\xi_k \sim \mathcal{D}_k} \mathcal{L}(\theta; \xi_k)$ for $k \in \{1, \dots, m\}$ is continuously differentiable, and there is a constant $L \geq 0$ such that:*

$$\|\nabla \mathcal{L}_k(\theta) - \nabla \mathcal{L}_k(\vartheta)\| \leq L\|\theta - \vartheta\|, \quad \forall\, \theta, \vartheta \in \mathbb{R}^d. \tag{D.7}$$

Theorem D.6 (Non-convex Convergence Rate of DSGD (Koloskova et al., 2020)). *Under Assumption 1, Assumption D.1 and Assumption 3, let the learning rate η satisfy $\eta \leq \eta_{\max} = \mathcal{O}\big(\frac{p}{L}\big)$ 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 $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla \mathcal{L}(\bar{\theta}^{(t)})\|_2^2\big] \leq \varepsilon$, the total number of steps T satisfies:*

$$T = \mathcal{O}\Big(\frac{\sigma^2}{m\,\varepsilon^2} + \frac{\sqrt{p}\,\sigma + \zeta}{p\,\varepsilon^{3/2}} + \frac{1}{p\varepsilon}\Big) \cdot L\Big(\mathcal{L}(\theta_0) - \mathcal{L}^\star\Big).$$

We then provide our main theoretical results as follows.

Theorem D.7 (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 $\eta \leq \frac{1}{L_2}$. 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 $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla \mathcal{L}(\bar{\theta}^{(t)})\|_2^2\big] \leq \varepsilon$, the total number of steps T satisfies:*

$$T = \mathcal{O}\Big(\tfrac{\sigma^2}{m\varepsilon^2} + \tfrac{1}{\varepsilon} + \tfrac{1}{\varepsilon}\big([\textstyle\sum_{t=0}^{T-1} A^{(t)}]^+\big)^{1/2}\Big) \cdot L_2\big(\mathcal{L}(\theta^{(0)}) - \mathcal{L}^\star\big),$$

where $[\cdot]^+ \triangleq \max(0, \cdot)$ and $A^{(t)}$ is defined as:

$$A^{(t)} \triangleq \eta L_2 \left(2T_2 + L_3^2 \Xi_t^4 + \left(2L_1 + 2L_3\Xi_t^2 + \frac{mL_4^2}{24^2}\right)\sqrt{m}\Xi_t^3\right),$$

with T_2 and the consensus distance Ξ_t^2 given by:

$$T_2 \triangleq (\nabla \mathcal{L}(\bar{\theta}^{(t)}))^\top \nabla\, \mathrm{Tr}\big(\nabla^2 \mathcal{L}(\bar{\theta}^{(t)})\, \Gamma^{(t)}\big), \quad \Xi_t^2 \triangleq \tfrac{1}{m}\sum_{k=1}^m (\theta_k^{(t)} - \bar{\theta}^{(t)})^\top (\theta_k^{(t)} - \bar{\theta}^{(t)}).$$

Proof. We structure the proof into several key steps.

Step (A): Descent Force Decomposition.

Based on the L_2-smoothness (Assumption D.1) of the loss function $\mathcal{L}$ (as implied by Assumption 2)), we can apply the first-order Taylor expansion around $\bar{\theta}^{(t)}$ to establish an upper bound for $\mathcal{L}(\bar{\theta}^{(t+1)})$:

$$\mathcal{L}\big(\bar{\theta}^{(t+1)}\big) \leq \mathcal{L}\big(\bar{\theta}^{(t)}\big) + \nabla \mathcal{L}\big(\bar{\theta}^{(t)}\big)^\top \big(\bar{\theta}^{(t+1)} - \bar{\theta}^{(t)}\big) + \tfrac{L_2}{2}\big\|\bar{\theta}^{(t+1)} - \bar{\theta}^{(t)}\big\|^2.$$

According Proposition D.3, we have

$$\mathbb{E}_{\xi^{(t)}}[\bar{\theta}^{(t+1)}] = \bar{\theta}^{(t)} - \eta\,\big(\nabla \mathcal{L}\big(\bar{\theta}^{(t)}\big) + \nabla\ \mathrm{Tr}(\nabla^2 \mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})\big) + \delta^{(t)},$$

where $\bar{\theta}^{(t+\frac{1}{2})} = \bar{\theta}^{(t)} + \epsilon^{(t)}$ and $\epsilon^{(t)} \sim \mathcal{N}(0, \Gamma^{(t)})$.

Substituting this into the previous bound and taking the expectation with respect to random data sampling yields:

$$\begin{aligned}
&\mathbb{E}_{\xi^{(t)}}[\mathcal{L}\big(\bar{\theta}^{(t+1)}\big)] \\
&\leq \mathcal{L}\big(\bar{\theta}^{(t)}\big) - \eta\, \nabla \mathcal{L}\big(\bar{\theta}^{(t)}\big)^\top \Big(\nabla \mathcal{L}\big(\bar{\theta}^{(t)}\big) + \nabla\ \mathrm{Tr}(\nabla^2 \mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)}) - \delta^{(t)}\Big) + \mathbb{E}_{\xi^{(t)}} \tfrac{\eta^2 L_2}{2}\big\|\bar{\theta}^{(t+1)} - \bar{\theta}^{(t)}\big\|^2.
\end{aligned}$$

According to Corollary D.4, we obtain

$$\mathbb{E}_{\xi^{(t)}}\big\|\bar{\theta}^{(t+1)} - \bar{\theta}^{(t)}\big\|^2 \leq \frac{\sigma^2}{m} + \eta^2\big\|\nabla \mathcal{L}\big(\bar{\theta}^{(t)}\big) + \nabla\ \mathrm{Tr}(\nabla^2 \mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)}) + \delta^{(t)}\big\|^2.$$

To further refine the analysis, we decompose the squared norm:

$$\begin{aligned}&\left\|\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)+\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})\right\|^2\\&=\left\|\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})\right\|^2+\left\|\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)\right\|^2+2\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})^\top\mathcal{L}\big(\bar{\theta}^{(t)}\big).\end{aligned}$$

Combining the previous steps, we obtain:

$$\begin{aligned}\mathbb{E}_{\xi^{(t)}}\,\mathcal{L}\big(\bar{\theta}^{(t+1)}\big)\le\mathcal{L}\big(\bar{\theta}^{(t)}\big)-(\eta-\tfrac{\eta^2L_2}{2})\left\|\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)\right\|^2+\frac{\eta^2L_2}{2}\underbrace{\left\|\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})\right\|^2}_{T_1}\\\eta^2L_2\underbrace{\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)^\top\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})}_{T_2}+\frac{\sigma^2}{m}\cdot\frac{\eta^2L_2}{2}\\+\eta^2L_2\underbrace{\big(\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)+\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})\big)^\top\delta^{(t)}}_{T_3}+\frac{\eta^2L_2}{2}\underbrace{\left\|\delta^{(t)}\right\|^2}_{T_4}.\end{aligned}\tag{D.8}$$

We subsequently control terms related to $\mathbb{E}_{\epsilon^{(t)}\sim\mathcal{N}(0,\Gamma^{(t)})}\nabla\mathcal{L}\big(\bar{\theta}^{(t+\frac{1}{2})}\big)\big)-\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)$ in Equation (D.8).

Step (B): Control Consensus-related Terms

Applying the logic in Corollary D.5 to bound residuals, we can derive

$$\left\|\delta^{(t)}\right\|\le\frac{L_4}{24}\cdot\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^3\ \le\frac{L_4}{24}\cdot\sqrt{m}\big(\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2\big)^{\frac{3}{2}},$$

and thus by the convexity of square operation,

$$T_4=\left\|\delta^{(t)}\right\|^2\le\frac{mL_4^2}{24^2}\cdot\big(\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2\big)^3.$$

Given $\left\|\nabla^3\mathcal{L}(\cdot)\right\|\le L_3$ we can upper-bound T_1 as

$$T_1=\left\|\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})\right\|^2\le L_3^2\cdot\big(\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2\big)^2=L_3^2\cdot\big(\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2\big)^4.$$

We can also bound T_3 as follows:

$$\begin{aligned}T_3&\le\left\|\nabla\ \mathrm{Tr}(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\Gamma^{(t)})+\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)\right\|\cdot\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^3\\&\le\big(L_3\,\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2+L_1\big)\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^3\\&\le\big(L_3\,\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2+L_1\big)\sqrt{m}\big(\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2\big)^{\frac{3}{2}}.\end{aligned}$$

Recall the notation $\Xi_t^2=\big(\frac{1}{m}\sum_{k=1}^m\left\|\theta_k^{(t)}-\bar{\theta}^{(t)}\right\|^2\big)$. Therefore,

$$A^{(t)}\triangleq\ \eta L_2\,\big(2T_2\ +L_3^2\,\Xi_t^4+\big(2L_1+2L_3\,\Xi_t^2+\frac{mL_4^2}{24^2}\big)\sqrt{m}\Xi_t^3\big).$$

Then Equation (D.8) becomes

$$\mathbb{E}_{\xi^{(t)}}\,\mathcal{L}\big(\bar{\theta}^{(t+1)}\big)\le\mathcal{L}\big(\bar{\theta}^{(t)}\big)-\Big(\eta-\tfrac{\eta^2L_2}{2}\Big)\left\|\nabla\mathcal{L}\big(\bar{\theta}^{(t)}\big)\right\|^2+\eta^2L_2A^{(t)}+\tfrac{\sigma^2}{m}\cdot\tfrac{\eta^2L_2}{2},\tag{D.9}$$

where

$$A^{(t)}\ =\eta L_2\big(2T_2+T_1+2T_3+T_4\big)\ \le2T_2+L_3^2\,\Xi_t^4\ +\Big(2L_1+2L_3\,\Xi_t^2+\tfrac{m\,L_4^2}{24^2}\Big)\sqrt{m}\,\Xi_t^3.\tag{D.10}$$

Step (C): Derive the Rate

Starting from the descent inequality equation D.9:

$$\mathbb{E}_{\xi^{(t)}}\left[\mathcal{L}(\bar{\theta}^{(t+1)})\right] \leq \mathcal{L}(\bar{\theta}^{(t)}) - \left(\eta - \frac{\eta^2 L_2}{2}\right)\left\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\right\|^2 + \eta^2 L_2\, A^{(t)} + \frac{\sigma^2}{m}\,\frac{\eta^2 L_2}{2}.$$

Taking full expectation and summing over $t = 0, \ldots, T-1$, we obtain

$$\sum_{t=0}^{T-1}\left(\eta - \frac{\eta^2 L_2}{2}\right)\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\right\|^2 \leq \mathcal{L}(\theta^{(0)}) - \mathbb{E}\left[\mathcal{L}(\bar{\theta}^{(T)})\right] + \eta^2 L_2 \sum_{t=0}^{T-1} A^{(t)} + \frac{\sigma^2\,\eta^2 L_2\, T}{2m}.$$

Since $\eta \leq 1/L_2$ implies $\eta - \frac{\eta^2 L_2}{2} \geq \eta/2$, and denoting $\Delta = \mathcal{L}(\bar{\theta}^{(0)}) - \mathcal{L}^*$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\right\|^2 \leq \frac{2\Delta}{\eta\, T} + \frac{2\eta L_2}{T}\sum_{t=0}^{T-1} A^{(t)} + \frac{\sigma^2\,\eta L_2}{m}. \tag{D.11}$$

To ensure this is at most ε, it suffices to enforce

$$\frac{\sigma^2\,\eta L_2}{m} \leq \frac{\varepsilon}{3}, \quad \frac{2\eta L_2}{T}\sum_{t=0}^{T-1} A^{(t)} \leq \frac{\varepsilon}{3}, \quad \text{and} \quad \frac{2\Delta}{\eta\, T} \leq \frac{\varepsilon}{3}.$$

Regarding the second inequality, we consider two cases for the sign of $\sum_{t=0}^{T-1} A^{(t)}$. If this sum is non-positive, the inequality is trivially satisfied. Otherwise, if the sum is positive, we must choose η to satisfy the resulting bound. To satisfy all three conditions simultaneously, along with a stability condition like $\eta \leq 1/L_2$, we must select η from the minimum of all applicable upper bounds. This logic suggests the following choices:

$$\eta \;\leq\; \min\left\{\frac{1}{L_2},\; \frac{m\varepsilon}{3\sigma^2 L_2},\; \frac{T\varepsilon}{6L_2\sum_{t=0}^{T-1} A^{(t)}}\right\}, \quad \text{and} \quad T \geq \frac{6\,\Delta}{\eta\,\varepsilon}.$$

To ensure a valid step-size η exists, we substitute these three upper bounds into the condition for T. This yields three distinct lower bounds on the total number of iterations T that must be satisfied. By rearranging the inequality $T\eta \geq \frac{6\Delta}{\varepsilon}$, we require:

$$T \;\geq\; \max\left\{\frac{6\Delta L_2}{\varepsilon},\; \frac{18\Delta\sigma^2 L_2}{m\varepsilon^2},\; \frac{6}{\varepsilon}\sqrt{\Delta L_2 \sum_{t=0}^{T-1} A^{(t)}}\right\}.$$

The first two bounds are derived directly by substituting the first two terms from the $\min\{\cdot\}$ expression for η. The third bound arises specifically in the case where $\sum_{t=0}^{T-1} A^{(t)} > 0$.

Therefore, the total number of iterations T should be large enough to satisfy all applicable lower bounds. This leads to the sufficient condition:

$$T = \mathcal{O}\left(\frac{L_2\,\Delta}{\varepsilon} + \frac{L_2\,\Delta\,\sigma^2}{m\,\varepsilon^2} + \frac{\sqrt{L_2\,\Delta}}{\varepsilon}\sqrt{\left[\sum_{t=0}^{T-1} A^{(t)}\right]^+}\right),$$

where $[\cdot]^+ \triangleq \max(0, \cdot)$ is the positive part function. This complexity bound elegantly and directly reflects the impact of the higher-order term: its contribution to the iteration count only materializes when its cumulative sum is positive,. This condition is sufficient to guarantee

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\right\|_2^2 \leq \varepsilon.$$

The proof is now complete.

□

Proposition D.8. *Suppose Assumption 4 holds and assume that the gradient norm satisfies* $\|\nabla\mathcal{L}(\bar{\theta}^{(t)})\| \geq \mu_t > 0$ *for positive constant* μ_t*. Then, for any fixed* $m > 0$*, there exists a sufficiently small* $\Xi_t^2 > 0$*, where* $\Xi_t^2 = \mathrm{Tr}(\Gamma^{(t)})$ *with* $\Gamma^{(t)} = \frac{1}{m}\sum_{k=1}^{m}(\theta_k^{(t)} - \bar{\theta}^{(t)})(\theta_k^{(t)} - \bar{\theta}^{(t)})^\top$*, such that the inequality*

$$A^{(t)} \triangleq \eta L\Big(2T_2 + L_3^2\Xi_t^4 + \big(2L_1 + 2L_3\Xi_t^2 + \tfrac{mL_4^2}{24^2}\big)\sqrt{m}\,\Xi_t^3\Big) \leq 0 \tag{D.12}$$

holds. Here $T_2 = (\nabla\mathcal{L}(\bar{\theta}^{(t)}))^\top\nabla\,\mathrm{Tr}\big(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\,\Gamma^{(t)}\big)$ *and* L_1, L_3, L_4 *are the upper bounds for the first, third, and fourth derivatives of* $\mathcal{L}$*, respectively.*

Remark D.1. We note that Proposition D.8 does not contradict Equation (D.11) when both Δ and σ are zero. The condition $\Delta = \mathcal{L}(\bar{\theta}^{(0)}) - \mathcal{L}^* = 0$ implies that the models are initialized at an optimal point. In Theorem D.7, we assume that all initializations are identical ($\theta_k^{(0)} = \theta^{(0)}, \forall k \in \mathcal{V}$), so it follows that all models begin at the same optimum. Consequently, the consensus error remains zero throughout all iterations, meaning the model covariance matrix $\Gamma^{(t)}$ is the zero matrix and its trace Ξ_t is also zero. Since every component of the term $A^{(t)}$, defined as

$$A^{(t)} \triangleq \eta L\Big(2(\nabla\mathcal{L}(\bar{\theta}^{(t)}))^\top\nabla\,\mathrm{Tr}\big(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\,\Gamma^{(t)}\big) + L_3^2\Xi_t^4 + \big(2L_1 + 2L_3\Xi_t^2 + \tfrac{mL_4^2}{24^2}\big)\sqrt{m}\,\Xi_t^3\Big),$$

is a function of either $\Gamma^{(t)}$ or Ξ_t, the entire expression becomes $A^{(t)} = 0$. This causes the inequality in Equation (D.11) to hold trivially as both sides are zero, thus resolving any apparent inconsistency.

Proof Idea. Assuming a positive lower bound μ_t on the gradient norm, the term $-T_2$ can be lower bounded by a positive term of the form $\gamma\,\mu_t\,\Xi_t^2$, where $\gamma > 0$. The terms

$$L_3^2\Xi_t^4 + \big(2L_1 + 2L_3\Xi_t^2 + \tfrac{mL_4^2}{24^2}\big)\sqrt{m}\,\Xi_t^3$$

is a polynomial in Ξ_t with leading terms of order Ξ_t^4 and Ξ_t^3, both exceeding the quadratic order of $-2T_2 \geq 2\gamma\mu_t\Xi_t^2$. As $\Xi_t \to 0^+$, the higher-order terms approaches zero faster than the right side remains positive, so by continuity and the intermediate value theorem there exists a sufficiently small $\Xi_t > 0$ satisfying the inequality.

Proof.

Step (A): Derivation of γ. Denote $g^{(t)} = \nabla\mathcal{L}(\bar{\theta}^{(t)})$. Define

$$F(\delta) = \frac{\nabla^3\mathcal{L}(\bar{\theta}^{(t)})[\delta, \delta, g^{(t)}]}{\|g^{(t)}\|\|\delta\|^2}, \quad \delta \neq 0.$$

Since $\nabla^3\mathcal{L}(\bar{\theta}^{(t)})[\delta, \delta, g^{(t)}] < 0$, we have $F(\delta) < 0$ by Assumption 4. On the compact unit sphere $S = \{\delta : \|\delta\| = 1\}$, F attains its maximum $M < 0$. We can then set $\gamma = -M > 0$. Then for all δ,

$$\nabla^3\mathcal{L}(\bar{\theta}^{(t)})[\delta, \delta, g^{(t)}] \leq -\gamma\,\|g^{(t)}\|\|\delta\|^2,$$

and in particular for each k,

$$\nabla^3\mathcal{L}(\bar{\theta}^{(t)})[\delta_k^{(t)}, \delta_k^{(t)}, g^{(t)}] \leq -\gamma\,\|g^{(t)}\|\|\delta_k^{(t)}\|^2. \tag{D.13}$$

The intuition behind the parameter γ is that it reflects the *relative degree of progressive sharpening* (Assumption 4) during training.

Step (B): Bound on T_2. Denote

$$\Gamma^{(t)} = \frac{1}{m}\sum_{k=1}^{m}\delta_k^{(t)}(\delta_k^{(t)})^\top, \quad \Xi_t^2 = \mathrm{Tr}(\Gamma^{(t)}) = \frac{1}{m}\sum_{k=1}^{m}\|\delta_k^{(t)}\|^2.$$

Then

$$T_2 = (g^{(t)})^\top\nabla\,\mathrm{Tr}\big(\nabla^2\mathcal{L}(\bar{\theta}^{(t)})\,\Gamma^{(t)}\big) = \frac{1}{m}\sum_{k=1}^{m}\nabla^3\mathcal{L}(\bar{\theta}^{(t)})[\delta_k^{(t)}, \delta_k^{(t)}, g^{(t)}].$$

Using the bound in Equation (D.13),

$$T_2 \leq \frac{1}{m}\sum_{k=1}^{m}\left(-\gamma\,\|g^{(t)}\|\|\delta_k^{(t)}\|^2\right) = -\gamma\,\|g^{(t)}\|\Xi_t^2.$$

Therefore, we have

$$-T_2 \geq \gamma\,\|g^{(t)}\|\Xi_t^2 \geq \gamma\,\mu_t\,\Xi_t^2.$$

Step (C): Backward proof. We present a proof by working backwards from the desired result. The goal is to show that there exists $\Xi_t^2 > 0$ with

$$L_3^2(\Xi_t^2)^2 + \left(2L_1 + 2L_3\Xi_t^2 + \tfrac{mL_4^2}{24^2}\right)\sqrt{m}\,(\Xi_t^2)^{3/2} \leq -2T_2.$$

Since $-2T_2 \geq 2\gamma\mu_t\Xi_t^2$, it suffices that

$$L_3^2(\Xi_t^2)^2 + \left(2L_1 + 2L_3\Xi_t^2 + \tfrac{mL_4^2}{24^2}\right)\sqrt{m}\,(\Xi_t^2)^{3/2} \leq 2\gamma\mu_t\,\Xi_t^2,$$

where dividing both sides with $\Xi_t^2 > 0$ yields an equivalent form:

$$L_3^2\Xi_t^2 + \left(2L_1 + 2L_3\Xi_t^2 + \tfrac{mL_4^2}{24^2}\right)\sqrt{m}\,(\Xi_t^2)^{1/2} \leq 2\gamma\mu_t.$$

We can set

$$h(u) = L_3^2 u + \left(2L_1 + 2L_3 u + \tfrac{mL_4^2}{24^2}\right)\sqrt{m}\,\sqrt{u}, \quad u > 0.$$

Then $\lim_{u\to 0^+} h(u) = 0 < 2\gamma\mu_t$. By continuity, there is $\delta > 0$ such that for $0 < u < \delta$, $h(u) < 2\gamma\mu_t$. Hence for sufficiently small Ξ_t^2, the desired inequality holds. The proof is now complete. □