When Can Federated Learning Match Centralized Learning? A PAC-Bayesian Generalization Gap Analysis

Xuanyu Chen[1] **Shuai Wang**[2] **Nan Yang**[1*] **Dong Yuan**[1*]
The University of Sydney[1] Northwestern Polytechnical University[2]

Abstract

The growing focus on distributed data and privacy has spurred the rise of Federated Learning (FL). Empirical studies show that, under equal resources, FL often underperforms centralized training, but the reasons behind this gap remain theoretically unclear. This lack of understanding leaves open whether FL is inherently inferior in generalization and how the gap might be closed. We address this by formulating FL as a server-based SGD optimization problem over distributed data and analyzing the generalization gap within the PAC-Bayesian framework. Our analysis derives non-vacuous bounds on this gap, showing that such a gap necessarily exists under equal resources and depends on training parameters. We further prove that the gap can be fully eliminated only by introducing new clients or adding new data to existing clients, with the latter being more efficient. In contrast, allowing FL to have advantages in other resources, such as larger models or more communication rounds, cannot close the gap. As a complementary analysis, we also confirm from a stability perspective that centralized FL holds a generalization advantage over decentralized FL, justifying our FL formulation choice. Extensive experiments across different model architectures and datasets validate our theory.

* Corresponding authors: Nan Yang and Dong Yuan. (n.yang@sydney.edu.au and dong.yuan@sydney.edu.au)

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Classical deep learning algorithms are typically performed in centralized settings (He et al., 2016; Vaswani et al., 2017), where deep neural networks are trained with massive amounts of data on servers equipped with strong computational power. This setup has been proven effective in practice, including the training of Large Language Models (LLMs) (Brown et al., 2020; Black et al., 2022; Thoppilan et al., 2022), which have recently received significant attention due to their impressive performance on various tasks. However, an inherent limitation of this approach is the imperative centralization of training data (Chen et al., 2023). In reality, the majority of data is generated and stored in a distributed manner. If data containing sensitive information is centralized, the privacy of participating parties will likely be compromised. The challenge of expanding data size while protecting data privacy has led to the emergence of a new type of learning methods that exploit training with distributed data, among which Federated Learning (FL) (McMahan et al., 2017; Zhuang et al., 2021; Karimireddy et al., 2021) is the most popular. In FL, training data remains on local clients, and multiple clients collaborate with a central server to train a model without sharing data (Li et al., 2021).

The introduction of FL effectively alleviates the privacy problem, but there is no perfect solution (Abdul-Rahman et al., 2020). By comparing the two types of training setups, many studies have found that under equal training resources, the models trained in a federated scenario do not perform as well as models trained in a centralized scenario in test datasets or downstream tasks (Elnakib et al., 2023; Zhao et al., 2018), which drastically hinders the broad application of FL. Notably, this conclusion has mainly been established through empirical evidence, and the theoretical aspect has yet to be fully explored (Garst et al., 2023; Mar'i et al., 2023). Previous theoretical studies mainly study the generalization behavior of stochastic optimization within a single training scenario (He

et al., 2019; Zhao et al., 2024). The lack of theoretical understanding has resulted in long-term arguments on the existence of the performance gap (Drainakis et al., 2023) and also prevented the identification of feasible ways to close this gap.

In this paper, we revisit the fundamental question: ***Given the same training resource (including model, training data, total training compute, etc), can federated learning catch up with or surpass centralized learning in terms of generalization performance?*** To advance the theoretical underpinnings, we model two types of learning as server-based Stochastic Gradient Descent (SGD) (Bottou, 1998; Sutskever et al., 2013) optimization problems on centralized and decentralized data, respectively, and establish PAC-Bayesian (Probably Approximately Correct Bayesian) bounds (McAllester, 1998, 1999) on the generalization error of the models. Since the generalization bound is usually considered an essential index of the generalization ability of the learning algorithm, the performance gap is formulated as the distance between two generalization bounds. Our analysis shows that such a gap necessarily exists under equivalent training conditions and is affected by training settings. Therefore, completely bridging this gap requires FL to be provided with more training resources. We prove that only incorporating new clients or adding data to existing clients can fully close the gap, while scaling model size or increasing communication rounds is not feasible. Furthermore, as a complementary analysis, we also provide a uniform stability perspective (Hardt et al., 2016) to justify our formulation of FL when quantifying the FL-to-centralized gap. Finally, extensive experiments on two model architectures (ResNet (He et al., 2016) and Vision Transformer (Dosovitskiy et al., 2020)) and two datasets (CIFAR-10 (Krizhevsky, 2009) and Mini-ImageNet (Vinyals et al., 2016; Deng et al., 2009)) confirm that the empirical results are closely aligned with our theory, showing the practical applicability of our findings.

In summary, the key contributions of our paper are shown below:

1. We theoretically characterize the generalization gap between centralized and federated learning, defining this gap as the distance between their PAC-Bayesian generalization error bounds. As a complementary analysis, we also adopt a stability perspective, which suggests the correct modeling of FL for a rigorous gap characterization.

2. We establish non-vacuous lower and upper bounds on the generalization gap by proving the monotonicity between the gap and the number of clients. With these bounds, we discover that when two training scenarios are provided with equivalent training resources, the gap cannot be fully closed even in the best case, and we also find the influence of training settings on this gap.

3. We further explore whether the gap can be completely bridged if FL is given with training advantages compared to centralized learning. Our study shows interesting results that some common approaches in improving the performance, such as scaling up model size or increasing communication rounds, cannot fully close this gap. Only introducing clients with new data or adding data to existing clients is possible. Among these two feasible strategies, the latter one is found to be more efficient in closing the gap.

4. Through extensive experiments on different model architectures and datasets, the correctness of our theoretical results is validated.

2 RELATED WORKS

2.1 Federated Learning

Federated learning (FL) is a class of distributed methods for collaborative model training without directly sharing local data, thereby preserving privacy (Abdul-Rahman et al., 2020; Li et al., 2021). The benchmark approach is Federated Averaging (FedAvg) (McMahan et al., 2017), where a central server broadcasts a global model, clients update it using local data, and the server aggregates their updates. Beyond this centralized FL design, an alternative line of work adopts decentralized communication, referred to as decentralized FL (Yuan et al., 2024; Beltrán et al., 2023). Here, clients periodically average parameters or gradients with their neighbors in a peer-to-peer network, avoiding a central coordinator. This improves robustness to single-point failures and can reduce communication bottlenecks, but performance then depends heavily on network topology and connectivity (Sun et al., 2024a). In recent years, as people have become more concerned about privacy and data security, many research works related to FL have emerged (Zhuang et al., 2021; Zhao et al., 2018; Tran et al., 2019). These studies generally hold the impression that centralized learning must perform better than FL, and many of them focus on proposing advanced FL algorithms to catch up with the centralized baseline (Zhuang et al., 2021). However, the correctness of this impression has not been fully explored from a theoretical aspect. Our work fills this gap and identifies generic strategies that can bridge the gap between the two training setups.

2.2 Studies that Compare Federated Learning with Centralized Learning

Since FL was proposed, there have been studies focusing on the comparison between federated and centralized training. Some works aim to compare the performance of the models trained in each training scenario. These comparative evaluations report that models trained in a centralized setup generally outperform models trained in a federated setup across a variety of tasks and datasets, such as MNIST (Peng et al., 2022; Mar'i et al., 2023), CIFAR-10 (Zhao et al., 2018), and CICIDS2017 (Elnakib et al., 2023). Similar experimental results are also found in the federated studies that adopt the centralized training results as one of the baselines (Zhuang et al., 2021). In addition to performance comparison, there are comparisons on the convergence rate. Unlike the above studies, these studies show that federated algorithms can attain the same order or faster convergence rate than centralized algorithms (Karimireddy et al., 2021). Furthermore, a recent study by Drainakis et al. explores the differences between federated and centralized training from the perspectives of energy cost and bandwidth cost (Drainakis et al., 2023). However, these existing comparisons mainly fall into two categories. First, empirical studies provide useful observations but do not offer theoretical explanations for why the gap arises. Second, some works attempted to analyze the gap from the theoretical perspective, which they focused on optimization efficiency rather than generalization behavior. Consequently, the fundamental question of whether a generalization gap necessarily exists between federated and centralized training remains theoretically unclear. In this work, we address this question by providing a PAC-Bayesian characterization of the generalization gap and deriving analytic conditions under which the gap can be closed.

2.3 Generalization Bound for Stochastic Gradient Descent Algorithms

Stochastic Gradient Descent (SGD) (Bottou, 1998; Sutskever et al., 2013) is a foundational optimization method in machine learning (Goodfellow et al., 2014; McMahan et al., 2017). The generalization of stochastic algorithms has been primarily studied through two complementary frameworks: uniform stability and PAC-Bayesian analysis. Uniform stability measures how sensitive a learning algorithm is to the replacement of training samples (Hardt et al., 2016). Specifically, an ϵ-uniformly stable algorithm is guaranteed to have generalization error at most ϵ, linking training stability to its generalization ability. This perspective is algorithm-dependent and well-suited for comparing the generalization behaviors of different algorithms under the same scenario (Sun et al., 2024b). The PAC-Bayesian framework, in contrast, characterizes generalization through the divergence between the prior and posterior distributions over hypotheses (He et al., 2019; Mou et al., 2018; London, 2017). This divergence describes the expected gap between training and test error, providing a principled way to quantify generalization performance. Both frameworks have recently been extended to FL settings. Stability-based analyses study how communication protocols (Sun et al., 2024a) and data heterogeneity affect generalization (Zhu et al., 2024), while PAC-Bayesian studies investigate algorithm design for non-IID data (Zhao et al., 2024) and the role of training parameters in generalization (Sefidgaran et al., 2024). However, most existing work analyzes generalization within a single training scenario, either centralized or federated, leaving the generalization difference between the two regimes insufficiently understood. In this work, we adopt a combined perspective to study the gap between federated and centralized learning. The PAC-Bayesian framework is used to rigorously characterize the generalization gap and to analyze feasible ways to bridge it, while stability analysis provides a complementary justification for modeling FL in a server-based setting. A detailed comparison between our work and prior generalization studies is summarized in Appendix A.

3 PRELIMINARIES

3.1 Generalization Error

In machine learning, let the hypothesis class of a model be denoted as $\Theta \subset \mathbb{R}^d$. The primary goal of learning algorithms is to identify a parameter vector $\theta \in \Theta$ that minimizes the expected risk, expressed as $\mathcal{R}(\theta) = \mathbb{E}_{\xi\sim\mathcal{D}}F(\theta;\xi)$. Here, d represents the dimension of Θ, F is the loss function, and $\mathcal{D}$ is the unknown distribution of the test data. When the parameter θ is treated as a random variable following a distribution Q, the expected risk with respect to Q can be written as

$$\mathcal{R}(Q) = \mathbb{E}_{\theta\sim Q}\mathbb{E}_{\xi\sim\mathcal{D}}F(\theta;\xi). \tag{1}$$

Since the true data distribution $\mathcal{D}$ is typically unknown, based on the training data's distribution $\hat{\mathcal{D}}$, the expected risk $\mathcal{R}$ is approximated by the empirical risk $\hat{\mathcal{R}}$ as

$$\hat{\mathcal{R}}(Q) = \mathbb{E}_{\theta\sim Q}\mathbb{E}_{\zeta\sim\hat{\mathcal{D}}}F(\theta;\zeta), \tag{2}$$

which is equivalent to the sample-average form $\hat{\mathcal{R}}(Q) = \frac{1}{|\hat{\mathcal{D}}|}\sum_{\zeta\in\hat{\mathcal{D}}}\mathbb{E}_{\theta\sim Q}F(\theta;\zeta)$. The discrepancy $\mathcal{R}(Q) - \hat{\mathcal{R}}(Q)$ between the expected risk $\mathcal{R}$ and the empirical risk $\hat{\mathcal{R}}$ defines the generalization error.

3.2 PAC-Bayesian Upper Bound for Generalization Error

Within the PAC-Bayesian (Probably Approximately Correct Bayesian) framework (McAllester, 1998, 1999), hypothesis functions learned by stochastic algorithms are viewed as randomly sampled functions from a hypothesis class. The generalization ability of an algorithm is measured by the distance between the posterior distribution of the output hypothesis Q and the prior distribution P, which is typically assumed to be Gaussian or Uniform. This leads to a classic result that provides a uniform bound on the expected risk $\mathcal{R}(Q)$, presented as follows:

Lemma 1. *[From (McAllester, 1998, 1999)] For any positive real number $\delta \in (0,1)$, and for all distributions Q, the following inequality holds with probability at least $1-\delta$ over a sample of size N:*

$$R(Q) \leq \hat{R}(Q) + \sqrt{\frac{\mathcal{D}(Q||P) + \log(\frac{1}{\delta}) + \log(N) + 2}{2N-1}}. \tag{3}$$

where $\mathcal{D}(Q||P)$ denotes the KL divergence between Q and P, defined as:

$$\mathcal{D}(Q||P) = \mathbb{E}_{\theta\sim Q} \log(\frac{Q(\theta)}{P(\theta)}). \tag{4}$$

4 THEORETICAL ANALYSIS

In this section, we develop theoretical foundations for the performance gap between federated and centralized settings and identify theoretically feasible approaches to close this gap. The main ingredient of our theory is the expression of this gap in the view of the PAC-Bayesian framework. We derive non-vacuous bounds for this theoretical expression, showing that the performance gap necessarily exists under equal training resources and how this gap varies with the parameters. Further analysis suggests that only the strategy of introducing new clients or adding data to existing clients is possible to close this gap fully. The definition of the notations used in this analysis can be found in Appendix B, and the full proof of the PAC-Bayesian gap analysis is provided in Appendix C.

4.1 Problem Setup

We compare FL with centralized learning under equal training conditions. Specifically, the same dataset and model are used for both sides, and with equal training compute, which we define following previous scaling law studies (Kaplan et al., 2020; Muennighoff et al., 2024) as the total number of samples processed through training (dataset size times the number of training rounds). We assume that they are equal so that the comparison focuses on the effect of the learning paradigm itself rather than differences caused by unequal optimization budgets. Within FL, we focus on the classical server-client design and model the federated training by the FedAvg framework (McMahan et al., 2017), rather than decentralized FL. This modeling choice is motivated by our complementary stability analysis provided in Appendix D, which proves that centralized FL holds stronger generalization guarantees than decentralized FL. Hence, to precisely quantify the gap between federated and centralized training, it suffices to be based on centralized FL. Concretely, in such a federated scenario, there are n clients and a central server that coordinates them. Each client $i \in \{1,\dots,n\}$ possesses a local dataset $\mathcal{D}_i$, with the average size denoted as $m = \frac{1}{n}\sum_{i=1}^{n}|\mathcal{D}_i|$. Thus, the total amount of data across all clients is nm. Assuming FL iterates T communication rounds, its formal update in round $j \in \{1,\dots,T\}$ can be written as:

$$\begin{aligned}\bar{\theta}(j+1) &= \frac{1}{n}\sum_{i=1}^{n}\theta_i(j+1) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\bar{\theta}(j) - \eta\nabla_{\bar{\theta}(j)}\mathbb{E}_{\zeta_i\sim\mathcal{D}_i}F(\bar{\theta}(j);\zeta_i)),\end{aligned} \tag{5}$$

where η is the learning rate. The first line describes the model aggregation on the central server, and the second line shows local training of $\bar{\theta}(j)$ on client i. Note that Eq.(5) summarizes the outcome of the t local SGD steps performed on each client during round j. The gradients are evaluated at intermediate iterates $\theta_i(j,\tau)$ for $\tau = 0,\dots,t-1$, starting from $\theta_i(j,0) = \bar{\theta}(j)$. Since local training is carried out with SGD, we define the local batch size as $S_{Fed} = k_{Fed}m$, where k_{Fed} denotes the batch-size ratio (i.e., $\frac{1}{m} \leq k_{Fed} \leq 1$). In contrast, the centralized scenario works with a dataset $\mathcal{D} = \bigcup_{i=1}^{n}\mathcal{D}_i$ of total size $D = nm$, and the initial model weights are identical to those used in the federated scenario, expressed as $\{\theta(0) = \theta_i(0)|i \in n\}$. The centralized training of θ is denoted as:

$$\theta(j+1) = \theta(j) - \eta\nabla_{\theta(j)}\mathbb{E}_{\zeta\sim\mathcal{D}}F(\theta(j);\zeta), \tag{6}$$

and is iterated for $\frac{T}{n}$ rounds to match the total training compute. In each round, following SGD optimization, the model θ is trained using data from $\mathcal{D}$ for t steps with the batch size $S_{Cen} = k_{Cen}D \in \{1,\dots,nm\}$, where k_{Cen} is the batch size ratio for centralized SGD, satisfying $\frac{1}{nm} \leq k_{Cen} \leq 1$. Moreover, we have $k_{Fed}m \leq k_{Cen}D$ due to more training data and stronger computation allocated to centralized settings in practice, leading to a generally larger batch size.

4.2 PAC-Bayesian Generalization Gap

To derive the PAC-Bayesian view of the generalization gap between FL and centralized learning, we first need to establish the PAC-Bayesian upper bounds for the generalization error of models trained in each scenario. Similar to the previous studies (Stephan et al., 2017; He et al., 2019), we make some assumptions on SGD to help our proof.

Assumption 1. *Considering that the stochastic gradient $\hat{g}_s(\theta) = \nabla_{\theta(t)}\hat{\mathcal{R}}(\theta(t))$ is computed as the sum of S independent gradients uniformly sampled from the training dataset, we assume that the gradient noise is Gaussian with covariance $\frac{1}{S}C(\theta)$, so $\hat{g}_s(\theta)$ can be approximated as*

$$\hat{g}_s(\theta) \approx g(\theta) + \frac{1}{\sqrt{S}}\Delta g(\theta), \quad \Delta g(\theta) \sim \mathcal{N}(0, C(\theta)), \tag{7}$$

where $g(\theta)$ denotes the full gradient of the expected loss. We further assume that $C(\theta)$ remains approximately constant with respect to θ and can be factorized into $C(\theta) \approx C = BB^\top$, where $C \in \mathbb{R}^{d\times d}$ is symmetric and (semi) positive semi-definite.

We justify Assumption 1 by the central limit theorem when the training data size is substantially larger than the batch size. Since deep neural networks are typically trained on large-scale datasets in real-world applications, the Gaussian assumption about gradient noise is generally valid (Weinan, 2017; Stephan et al., 2017). Also, the constant matrix C can be justified when the iterates of SGD are confined to a small enough region around a local optimum of the loss, where the noise covariance does not vary significantly in that region. Recent empirical studies have also observed that gradient noise may exhibit heavy-tailed behavior in certain deep learning settings (Zhang et al., 2020). In this work, we adopt the Gaussian approximation that are more commonly adopted in general SGD optimization and PAC-Bayesian analyses to maintain analytical tractability.

Assumption 2. *Assuming the loss function $F(\theta)$ is smooth, when the stationary distribution of the iterates is confined to a local region near a minimum θ^*, the loss gradient satisfies $\nabla F(\theta) \approx A(\theta - \theta^*)$, where $A \in \mathbb{R}^{d\times d}$ is a constant (semi) positive-definite matrix representing the local Jacobian of the gradient field.*

Assumption 2 makes sense when SGD converges to a low-variance quasi-stationary distribution near a deep local minimum, where the gradient noise is small compared to the average gradient. According to the fact that the exit time of a stochastic process is typically exponential in the height of the barriers between minima (Stephan et al., 2017), local optima are very stable even in the presence of noise. Thus SGD follows a relatively directed path toward the optimum. This assumption is also supported by empirical evidence (see p.1, Figures 1(a) and 1(b) and p.6, Figures 4(a) and 4(b) in (Li et al., 2018)). Moreover, this assumption can be extended to general cases through translation operations, which would not modify the geometry of the objective and its associated generalization ability.

Based on the above assumptions and Lemma 1, we derive a generalization bound for the models trained by federated SGD optimization.

Theorem 2. *For any positive real number $\delta \in (0,1)$, with probability at least $1-\delta$ over a decentralized training dataset of total size nm across n clients, the following inequality holds for the distribution Q_{Fed} of the output hypothesis learned by federated SGD:*

$$R(Q_{Fed}) - \hat{R}(Q_{Fed}) \leq \sqrt{\frac{H_F + \frac{T\eta}{2k_{Fed}m}tr(\bar{C}\bar{A}^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm - 2}}. \tag{8}$$

where $H_F = -\log(\det(\Sigma_{Fed}))$, Σ_{Fed} denotes the covariance matrix for the stationary distribution of FL, C_i is the covariance of the loss gradients and A_i is the Jacobian matrix around the minimum of the loss function for local training on client i, $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$, $\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i$, d is the dimension of the model parameter θ (parameter size), η is the learning rate and $tr(\bar{C}\bar{A}^{-1})$ is the trace of the product matrix $\bar{C}\bar{A}^{-1}$.

By a similar approach, the generalization bound for centralized training under equal training resources can also be proved as follows.

Corollary 3. *For any positive real number $\delta \in (0,1)$, with probability at least $1-\delta$ over a centralized training dataset of total size D on server, the following inequality holds for the distribution Q_{Cen} of the output hypothesis learned by centralized SGD:*

$$R(Q_{Cen}) - \hat{R}(Q_{Cen}) \leq \sqrt{\frac{H_C + \frac{T\eta}{2nk_{Cen}D}tr(CA^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(D) + 4}{4D - 2}}. \tag{9}$$

where $H_C = -\log(\det(\Sigma_{Cen}))$, Σ_{Cen} is the covariance matrix for the stationary distribution of centralized training, C and A are the covariance and Jacobian matrix around the minima for training with the centralized dataset.

Since the gradient covariance matrix C, the Jacobian matrix A, and the constant matrix Σ are from the stationary distribution of the SGD optimization, it is easy to notice that the comparison of two bounds becomes intractable without the knowledge of how these matrices vary with changes in the training setup. Therefore, we further present two assumptions and study a special case of the generalization bound.

Assumption 3. *We assume that A and Σ are symmetric matrices satisfying $A\Sigma = \Sigma A$.*

Assumption 3 implies that the local geometry around the global minimum and the stationary distribution are homogeneous across all dimensions of the parameter space. Similar assumptions were also used in previous papers (He et al., 2019; Jastrzkebski et al., 2017).

Assumption 4. *Under the fair comparison condition that the same training dataset is used for both training scenarios, we assume that the local data distributions $\mathcal{D}_1, \ldots, \mathcal{D}_n$ across n clients of size m form a heterogeneous partition of the global dataset $\mathcal{D}$ of size $D = nm$. Hence, we have the following approximate relationships:*

$$\bar{A} \approx A + \Delta_A, \quad \bar{C} \approx \frac{1}{n^\gamma}(C + \Delta_C), \tag{10}$$

where $\gamma > 1$, and Δ_A, Δ_C are deviation terms introduced by data heterogeneity. These deviations are assumed to be bounded in norm: $\|\Delta_A\| \leq \epsilon_A$, $\|\Delta_C\| \leq \epsilon_C$, where ϵ_A, ϵ_C grow with the non-IID degree.

Assumption 4 reflects the realistic cases where client datasets are drawn from heterogeneous (non-IID) distributions and could be justified by the central limit theorem when the average data size m across clients and the size of the global dataset D are both large enough. While the centralized quantities A and C characterize curvature and noise under the full dataset, their decentralized counterparts $\bar{A}$ and $\bar{C}$ may deviate due to non-IID local sampling. The inclusion of bounded deviation terms Δ_A and Δ_C, whose magnitudes reflect the degree of heterogeneity across clients, and scaling variable γ captures this variability while retaining analytical traceability for us to quantify the impact of non-IID distributions. Under the two new assumptions, we can characterize the difference between decentralized and centralized generalization behavior and formally establish the theorem that follows.

Theorem 4. *Under the above assumptions and when training resources for federated and centralized learning are equal, the generalization gap between federated and centralized SGD optimization has the following analytic solution:*

$$\mathcal{G}_{Fed} - \mathcal{G}_{Cen} = \frac{(\frac{T\eta}{2n^\gamma k_{Fed} m} - \frac{T\eta}{2nk_{Cen}D}) tr(CA^{-1})}{4D-2} + \frac{d\log(\frac{n^{\gamma-1}k_{Fed}m}{k_{Cen}D}) + \frac{T\eta}{2n^\gamma k_{Fed}m} tr(\Delta_1) + \log(\det(\Delta_2)^{-1})}{4D-2} \tag{11}$$

where $\Delta_1 = (CA^{-1}\Delta_A + \Delta_C(I + A^{-1}\Delta_A))A^{-1}$, $\Delta_2 = (I + C^{-1}\Delta_C)(I + \Delta_A A^{-1})$, and $\mathcal{G}$ is the generalization bound of a learning algorithm.

Theorem 4 shows the analytic solution of the generalization gap in the PAC-Bayesian framework.

4.3 Non-Vacuous Generalization Gap Bounds

In this subsection, we continue to explore this theoretical expression to gain a deeper understanding of the gap. As pointed out at the beginning of the paper, our interest lies in these questions: 1) Does the generalization gap always exist with equal training resources? 2) How is this gap affected by the environmental variables in the federated scenario? We answer these questions using the following theorem.

Theorem 5. *Under the above assumptions, and assuming equal training resources between federated and centralized scenarios, the generalization gap between federated and centralized SGD optimization satisfies the following inequalities:*

$$\frac{d\log(3^{\gamma-1}) + T\left(\frac{\eta tr(CA^{-1})}{2*3^\gamma k_{Fed}m} - \frac{\eta tr(CA^{-1})}{6k_{Cen}D} + \frac{\eta tr(\tilde{\Delta}_1)}{2^{\gamma+1}k_{Fed}m}\right)}{4D-2} + \frac{\log(\det(\Delta_2)^{-1})}{4D-2} \leq O\left(\mathcal{G}_{Fed} - \mathcal{G}_{Cen}\right) \leq \frac{d\log(\frac{D^{\gamma-1}k_{Fed}m}{k_{Cen}D})}{4D-2} + \frac{\left(\frac{\eta(tr(CA^{-1}))}{2D^\gamma k_{Fed}m} - \frac{\eta(tr(CA^{-1})}{2D^2 k_{Cen}}\right) + \frac{T\eta tr(\tilde{\Delta}_1)}{2^{\gamma+1}k_{Fed}m} + \log(\det(\Delta_2)^{-1})}{4D-2}, \tag{12}$$

for $3 \leq n \leq D$, where $\tilde{\Delta}_1$ satisfies $(\tilde{\Delta}_1)_{i,j} = |(\Delta_1)_{i,j}|$, n represents the number of clients and $D = nm$ is the total data size across clients. Additionally, when $n = 2$, for any constant $\gamma \geq 2$, the generalization gap satisfies the following inequality:

$$O\left(\mathcal{G}_{Fed} - \mathcal{G}_{Cen}\right) \geq \frac{\log(\det(\Delta_2)^{-1})}{4D-2} + \frac{d\log(2^{\gamma-1}) + T\left(\frac{\eta(tr(CA^{-1}) + tr(\tilde{\Delta}_1))}{2^{\gamma+1}k_{Fed}m} - \frac{\eta tr(CA^{-1})}{4k_{Cen}D}\right)}{4D-2}. \tag{13}$$

Remark 1. *Theorem 5 establishes non-vacuous upper and lower bounds for the generalization gap between federated and centralized training. These bounds allow us to analyze if the gap necessarily exists and how the gap is affected by various parameters:*

- ***Gap Existence:*** *Since C and A are (semi) positive-definite matrices, we can observe from Eqs.(12) and (13) that $\mathcal{G}_{Fed} - \mathcal{G}_{Cen} > 0$ requires satisfying $d > \frac{\frac{T\eta tr(CA^{-1})}{2nk_{Cen}D} + \log(\det(\Delta_2))}{\log(n^{\gamma-1})}$. Considering that deep learning typically involves over-parameterized neural networks to perform well (Kaplan et al., 2020; Hoffmann et al., 2022) and federated scenarios often scale to a significant number of devices (i.e., leads to large n and D), this condition is readily satisfied in practice.*

- ***Number of clients*** *n: As shown through the proved monotonicity, the gap increases with n.*

- ***Model dimensionality*** d*: Both lower bounds scale with the term* $d\log(n^{\gamma-1})$*. Since* $\gamma > 1$ *and* $n \geq 2$*, the gap increases with* d*.*
- ***Communication rounds*** T*: The impact of* T *appears in the form of a difference between two trace terms, which makes its sign unclear in general. In the special case where client data is i.i.d. and both training scenarios use identical batch size, the term can become positive, and the gap grows linearly with* T*. However, this assumption is rarely satisfied in realistic federated setups.*
- ***Non-IID degree:*** *The gap is explicitly affected by* $\tilde{\Delta}_1$ *and* Δ_2*, which quantify client heterogeneity. As the term* $tr(\tilde{\Delta}_1)$ *increases with* T*, and the term* $\log(\det(\Delta_2)^{-1})$ *remains fixed, the gap grows with the non-IID level across clients.*
- ***Total dataset size*** D*: Both lower bounds are inversely proportional to* D*, so increasing* D *consistently reduces the gap.*

With the above analysis, we can further summarize the below important insight:

- *Under equivalent training resources, a generalization gap necessarily exists for deep learning between federated and centralized settings. This gap is small if the training in a federated scenario satisfies a small number of clients, mild data heterogeneity, and a small model size. Additionally, this gap is also mitigated if the total data size across clients is sufficiently large.*

4.4 Strategies for Fully Closing the Gap

The above theoretical results demonstrate that the gap cannot be eliminated completely as long as training resources are equal between the two scenarios. Therefore, if we still look forward to federated training catching up with centralized training, the federated scenario has to be allowed with an advantage in some training resources. To understand how different resources influence the gap when federated learning is given progressively stronger advantages, we analyze the asymptotic behavior of the gap with respect to several key parameters. Generally, increasing the data size and model size can result in an improvement in performance. For example, researchers have concluded scaling laws indicating that the performance of large language models is related to these two parameters (Kaplan et al., 2020; Hoffmann et al., 2022). Besides, previous federated studies have also empirically shown that increasing the number of communication rounds or the number of clients also leads to improved performance (McMahan et al., 2017; Zhuang et al., 2021). So, we study the related parameters n, m, d, and T in federated settings and derive the following theorems.

Theorem 6. *Under the above assumptions and assuming that the federated scenario is provided with an advantage in training conditions, the following inequalities hold for the generalization gap between federated and centralized SGD optimization:*

$$\lim_{n\to\infty}(\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}) = \lim_{n\to\infty} \left(O(\frac{(\gamma d+2)\log(n)}{n}) + O(\frac{1}{n^{\gamma+1}}) + O(\frac{1}{n}) - O(1)\right) < 0; \tag{14}$$

$$\lim_{m\to\infty}(\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}) = \lim_{m\to\infty} \left(O(\frac{(d+2)\log(m)}{m}) + O(\frac{1}{m^2}) + O(\frac{1}{m}) - O(1)\right) < 0. \tag{15}$$

Here, $\gamma > 1$*,* $\tilde{\mathcal{G}}_{Fed}$ *is the generalization bound for federated scenarios having an advantage, and* $\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen} \leq 0$ *implies that federated training catches up with or outperform centralized training.*

Theorem 7. *Under the above assumptions and assuming that the federated scenario is provided with an advantage in training conditions, the following inequality hold for the generalization gap between federated and centralized SGD optimization:*

$$\lim_{T\to\infty}(\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}) = \infty. \tag{16}$$

Besides, if the federated scenario contains a large number of clients satisfying $n > \sqrt[\gamma]{\frac{T\eta e}{2k_{Fed}m}}$ *for any* $\gamma > 1$*, we also have:*

$$\lim_{d\to\infty}(\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}) = \infty. \tag{17}$$

Remark 2. *Theorems 6 and 7 show how the gap between federated and centralized training behaves as key parameters approach infinity, representing a continually growing advantage of federated training in these parameters. The condition in Theorem 7 basically holds in practice, considering that realistic federated scenarios generally scale to a sufficiently large number of clients (Kairouz et al., 2021) (e.g., phones with user data, edge sensors, etc.) According to Eqs.(14), (15), (16) and (17), we find that simply increasing the number of communication rounds (*T*) or the model size (*d*) cannot close the generalization gap unless more data is introduced. The two feasible approaches to do so are: (1) increasing the number of clients, or (2) increasing the average data per client. Among these, the latter is more efficient, as the gap decreases at a faster rate with respect to* m *than with respect to* n *(i.e.* $O(\frac{(d+2)\log(m)}{m})$ *vs* $O(\frac{(\gamma d+2)\log(n)}{n})$*). This suggests that, in reality, focusing on growing the local dataset in existing clients would be a more efficient way to make FL catch up with centralized training than introducing new clients.*

5 EMPIRICAL VALIDATION

5.1 Experiment Setup

To empirically validate our theoretical findings across different scenarios, we conduct extensive experiments on two representative model architectures, ResNet-18 (He et al., 2016) and Vision Transformer (ViT) (Dosovitskiy et al., 2020), which exemplify Convolutional Neural Networks (CNNs) (LeCun et al., 1998) and Transformers (Vaswani et al., 2017). For each architecture, we built 10 models of varying sizes to study the effect of model scale. We evaluate them on two standard datasets: CIFAR-10 (Krizhevsky, 2009) with 50,000 training and 10,000 validation images, and Mini-ImageNet (Vinyals et al., 2016) with 60,000 images in 100 classes extracted from ImageNet (Deng et al., 2009). Since Mini-ImageNet does not provide a complete train/validation split, we randomly divide it into 48,000 training and 12,000 validation images. The full training sets are used for centralized training, while for FL we partition them into n client datasets using Dirichlet sampling (Hsu et al., 2019) with parameter $\alpha = 0.1$, where a smaller α yields more heterogeneous splits. Detailed settings and server configuration are provided in Appendix E for reproducibility. All experiments were repeated with three random seeds (i.e., 0, 10, and 100), and the reported results correspond to the average performance across these runs.

5.2 Empirical Evidence

5.2.1 Generalization Gap under Equal Resources

We verify our non-vacuous bounds about the generalization gap by first constructing federated and centralized scenarios with equivalent training resources based on our problem setup. According to Eqs.(12) and (13), we observe that the gap is affected by the number of clients n, the model dimensionality d, and the data heterogeneity across clients. Hence, we conduct three sets of experiments to validate their respective impact. Figure 1 shows that the testing accuracy of models decreases with the number of clients. Since the centralized scenario can be considered as containing only one client (which is the server), the impact of n on the performance gap is justified. Next, the light blue area in Figure 2 demonstrates that enlarging the model size widens the gap for both ViT and ResNet architectures, which aligns with our theory. Finally, we also observe that the increasing non-IID level contributes to the enlargement of the gap. Figure 3 illustrates the impact of the non-IID degree, where a smaller α implies stronger heterogeneity across clients. As the heterogeneity level increases (moving right along the x-axis),

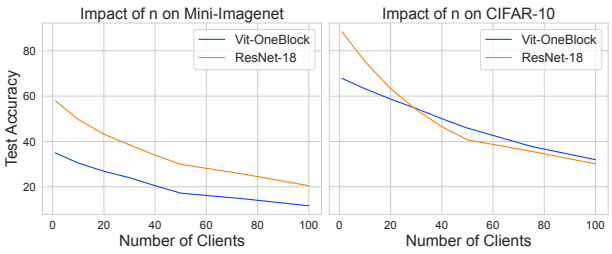

Figure 1: **Impact of the client number n on generalization. (Left)** Curves of Mini-ImageNet accuracy (%) to n. **(Right)** Curve of CIFAR-10 accuracy (%). The centralized scenario is considered as $n = 1$.

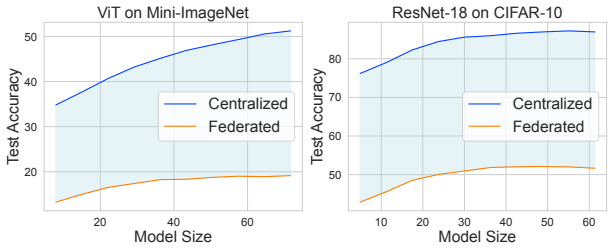

Figure 2: **Impact of the model size d (measured in M (millions parameters)) on generalization.** The gap between federated and centralized training is demonstrated by the light-blue area.

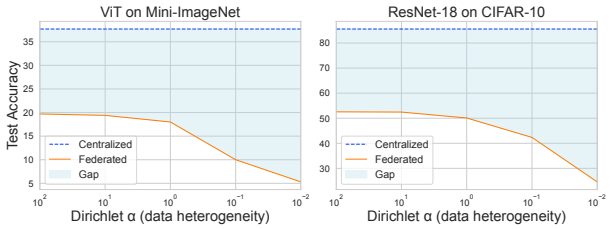

Figure 3: **Impact of the non-IID degree on the performance.** Smaller α implies greater heterogeneity across clients (i.e., α decreases from left to right on the x-axis).

the light-blue shaded area becomes larger, indicating a wider performance gap between federated and centralized training. This observation is consistent with our theoretical analysis, which predicts that stronger data heterogeneity enlarges the gap.

5.2.2 Bridge Gap by Increasing Resources

To empirically investigate our theoretical insights about the complete elimination of the performance gap, we designed four sets of experiments for the four parameters involved in Theorems 6 and 7. In each experiment, centralized training of ViT is compared with

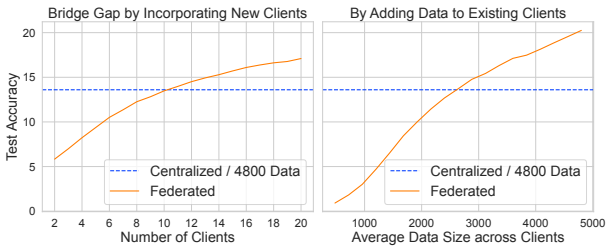

Figure 4: **Empirical evidence for strategies that are feasible in closing the gap. (Left)** By incorporating new clients (increasing n). **(Right)** By adding data to existing clients (increasing m).

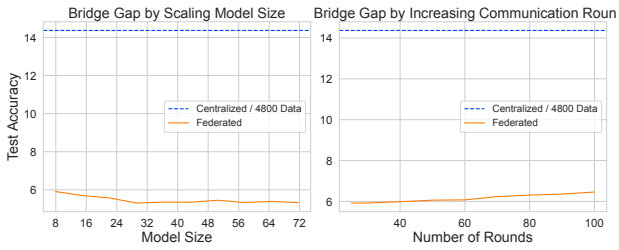

Figure 5: **Additional evidence for strategies that are unable to close the gap.** The baseline centralized scenario contains 4800 data, aligned with the settings in Figure 4. **(Left)** The strategy of scaling model sizes (increasing d). **(Right)** The strategy of increasing communication rounds (increasing T).

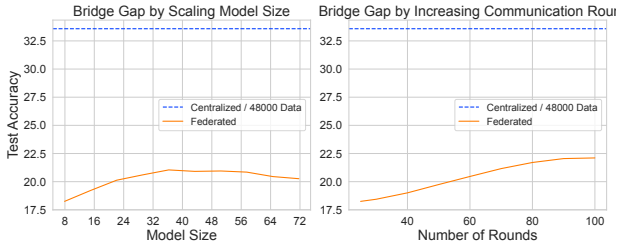

Figure 6: **Further evidence for strategies that are unable to close the gap.** The baseline centralized scenario holds the complete training set, which contains 48000 data. **(Left)** The strategy of scaling model sizes (increasing d). **(Right)** The strategy of increasing communication rounds (increasing T).

FL on Mini-ImageNet, where FL holds an advantage in one kind of training resource. We start from the setting where this training resource is equal and gradually amplify the focused parameter in the federated scenario to examine whether the performance gap can be progressively closed.

The results in Figures 4, 5, and 6 validate our theoretical findings. Figure 4 corresponds to Theorem 6. Specifically, the generalization performance of models trained in federated setups catches up with or surpasses that of centralized training by either incorporating new clients or adding data to existing clients. Moreover, Figure 4 shows that the latter approach is more efficient in closing the gap, as indicated by the steeper improvement curve. For example, scaling up the average data size by ten times (i.e., from $m = 480$ to $m = 4800$) results in a larger generalization improvement than scaling up the number of clients by ten times (i.e., from $n = 2$ to $n = 20$) for the same amount of increased data.

Figures 5 and 6 provide complementary evidence for Theorem 7. In these experiments, we examine whether giving federated learning an advantage in model size (d) or communication rounds (T) can eliminate the gap. The results consistently show that simply increasing d or T fails to close the gap with centralized training. This observation holds under both a partial dataset setting (4800 samples) and a full dataset setting (48000 samples), further confirming that scaling model size or communication rounds alone cannot eliminate the performance gap.

6 CONCLUSION

This paper re-examines why models trained in FL often underperform compared to centralized ones, focusing on the theoretical exploration of the generalization gap and strategies to bridge it. By formulating the gap as the distance between the PAC-Bayesian generalization bounds of server-based FL and centralized learning, we derive non-vacuous bounds on this gap, showing that it necessarily exists under equal training resources and is shaped by training settings. We further prove that the gap can only be closed by introducing new clients or adding data to existing clients, with the latter being more efficient, while common strategies such as scaling model size or increasing communication rounds are ineffective. As a complementary analysis, we also provide a stability view, confirming that centralized FL holds a generalization advantage over decentralized FL and justifying our FL formulation for a rigorous gap analysis. Finally, our empirical studies across different architectures and datasets corroborate the theory and highlight its practical relevance.

Acknowledgments

This work is partly supported by the Australian Research Council Linkage Project (Grant No. LP220200893).

References

Sawsan AbdulRahman, Hanine Tout, Hakima Ould-Slimane, Azzam Mourad, Chamseddine Talhi, and Mohsen Guizani. A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. *IEEE Internet of Things Journal*, 8(7):5476–5497, 2020.

Ghadir Ayache and Salim El Rouayheb. Random walk gradient descent for decentralized learning on graphs. In *2019 IEEE International Parallel and Distributed Processing Symposium Workshops*, pages 926–931. IEEE, 2019.

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.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

Léon Bottou. Online learning and stochastic approximations. In David Saad, editor, *Online Learning in Neural Networks*, pages 9–42. Cambridge University Press, 1998.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chaochao Chen, Xiaohua Feng, Jun Zhou, Jianwei Yin, and Xiaolin Zheng. Federated large language model: A position paper. *arXiv preprint arXiv:2307.08925*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Georgios Drainakis, Panagiotis Pantazopoulos, Konstantinos V Katsaros, Vasilis Sourlas, Angelos Amditis, and Dimitra I Kaklamani. From centralized to federated learning: Exploring performance and end-to-end resource consumption. *Computer Networks*, 225:109657, 2023.

Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

Omar Elnakib, Eman Shaaban, Mohamed Mahmoud, and Karim Emara. Evaluation of centralized, distributed and federated learning for iot intrusion detection systems. In *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 315–320. IEEE, 2023.

Paul Erdős and Alfréd Rényi. On random graphs. i. *Publicationes Mathematicae*, 6:290–297, 1959.

Swier Garst, Julian Dekker, and Marcel Reinders. A comprehensive experimental comparison between federated and centralized learning. *bioRxiv*, pages 2023–07, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, volume 48, pages 1225–1234. PMLR, 2016.

Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in neural information processing systems*, 32, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 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.

Stanisław Jastrzkebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, Toronto, Canada, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4): 3347–3366, 2021.

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Farhanna Mar'i, Ahmad Afif Supianto, and Fitra Abdurrachman Bachtiar. Comparation of federated and centralized learning for image classification. *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic*, 11(2):393–400, 2023.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, volume 54, pages 1273–1282. PMLR, 2017.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Sony Peng, Yixuan Yang, Makara Mao, and Doo-Soon Park. Centralized machine learning versus federated averaging: A comparison using mnist dataset. *KSII Transactions on Internet and Information Systems (TIIS)*, 16(2):742–756, 2022.

Milad Sefidgaran, Romain Chor, Abdellatif Zaidi, and Yijun Wan. Lessons from generalization error analysis of federated learning: You may communicate less often! In *International Conference on Machine Learning*, volume 235. PMLR, 2024.

Mandt Stephan, Matthew D Hoffman, David M Blei, et al. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Tao Sun, Dongsheng Li, and Bao Wang. Adaptive random walk gradient descent for decentralized optimization. In *International Conference on Machine Learning*, pages 20790–20809. PMLR, 2022.

Yan Sun, Li Shen, and Dacheng Tao. Towards understanding generalization and stability gaps between centralized and decentralized federated learning, 2024a. URL `https://arxiv.org/abs/2310.03461`.

Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, volume 238, pages 676–684. PMLR, 2024b.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Nguyen H Tran, Wei Bao, Albert Zomaya, Minh NH Nguyen, and Choong Seon Hong. Federated learning over wireless networks: Optimization model design and analysis. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, pages 1387–1395. IEEE, 2019.

George E Uhlenbeck and Leonard S Ornstein. On the theory of the brownian motion. *Physical review*, 36 (5):823, 1930.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Stability and generalization for markov chain stochastic gradient methods. *Advances in Neural Information Processing Systems*, 35:37735–37748, 2022.

Ee Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Honglin Yuan, Warren Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? In *International Conference on Learning Representations*, 2022.

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*, 11(21):34617–34638, 2024.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zihao Zhao, Yang Liu, Wenbo Ding, and Xiao-Ping Zhang. Federated pac-bayesian learning on non-iid data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5945–5949. IEEE, 2024.

Miaoxi Zhu, Li Shen, Bo Du, and Dacheng Tao. Stability and generalization of the decentralized stochastic gradient descent ascent algorithm. *Advances in Neural Information Processing Systems*, 36, 2024.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4912–4921, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, it is provided in Sections 3 and 4.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, it is provided in supplementary material.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, as shown in 4.]
   (b) Complete proofs of all theoretical results. [Yes, as shown in Appendix C.]
   (c) Clear explanations of any assumptions. [Yes, as shown in Section 4.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, as shown in Appendix E.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, run with three random seeds.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or

cloud provider). [Yes, as shown in Appendix E.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes, citations can be found in ReadME file in the supplemental material]
   (b) The license information of the assets, if applicable. [Yes, it is provided in ReadMe file]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Comparison of Existing Works Analyzing Generalization in FL

This section presents the detailed related work comparison omitted from the main manuscript, highlighting how our analysis differs from existing studies. As shown in Table 1, prior studies mainly focus on generalization within a single training scenario, either federated or centralized, and many recent works adopt only the stability perspective. While a few early efforts compare both training settings, they often rely solely on empirical observations or lack rigorous theoretical guarantees. As a result, a combined theoretical treatment using both the PAC-Bayesian theory and the stability tool for comparing federated and centralized training remains largely unexplored. Our work addresses this gap by applying PAC-Bayesian analysis to rigorously quantify the generalization gap and derive theoretical insights on how it can be bridged, and employing stability analysis to support our problem setup.

Table 1: **Generalization Analysis Comparison between Our Paper and Related Works.**

Paper	**Theoretical Analysis**	**Analysis Framework**	**Scenario Setup**	**Gap Study**	**Gap-bridging Insights**
London (2017)	✓	PAC-Bayes	Centralized	×	×
Mou et al. (2018)	✓	PAC-Bayes / Stability	Centralized	×	×
He et al. (2019)	✓	PAC-Bayes	Centralized	×	×
Yuan et al. (2022)	✓	Independent Analysis	×	✓	✓
Peng et al. (2022)	×	×	×	✓	×
Mar'i et al. (2023)	×	×	×	✓	×
Zhao et al. (2024)	✓	Pac-Bayes	Federated	×	×
Sefidgaran et al. (2024)	✓	Pac-Bayes	Federated	×	✓
Zhu et al. (2024)	✓	Stability	Federated	×	×
Sun et al. (2024b)	✓	Stability	Federated	×	×
Sun et al. (2024a)	✓	Stability	Federated	✓	×
Ours	✓	PAC-Bayes / Stability	Federated/Centralized	✓	✓

B Notations and Definitions

n	Number of clients in the network
m	Average number of local data across clients
θ, Θ	Model parameters
d	Dimension of model parameters / Model Size
f, F	Loss function
$\mathcal{D}$	Dataset
z	Data sample
η	Learning rate
T	Communication rounds
i, j, s	Indexes
t	Round Index (Stability) / Number of Local Epochs (PAC-Bayes)
$\mathcal{A}$	Algorithm
ϵ	Small constant
$L, G, \mathcal{B}$	Constants related to assumptions
K	Number of participating clients in each round
$\mathcal{P}$	Transition matrix of communication walk
$c_{\mathcal{P}}$	Spectral contraction constant depending on $\mathcal{P}$
$\mathbb{W}$	Communication weight matrix for decentralized mixing in decentralized FL
λ_2	Second largest eigenvalue
deg	Node degree in network
E	Number of edges in network
β	Training perturbation term
δ	Probability
Q	Distribution of the output hypothesis
P	Prior distribution
B, C	Covariance matrix about gradient
A	Jacobian matrix of the gradient field
k	Batch ratio
N, D	Data size
γ	Value depending on data heterogeneity
$\mathcal{G}$	Generalization error bound
$\mathcal{R}, \hat{\mathcal{R}}$	Expected risk and empirical risk
$\hat{g}_s, g$	Stochastic gradient and full gradient
$\Sigma_{Fed}, \Sigma_{Cen}$	Stationary covariance matrices of federated and centralized SGD
S_{Fed}, S_{Cen}	Batch sizes in federated and centralized training
Δ_A, Δ_C	Deviation terms induced by data heterogeneity
$H_F, H_C, \Delta_1, \Delta_2$	Composite deviation terms in the gap analysis

C Full Proofs for PAC-Bayesian Analysis on Generalization Gap

This section shows the full proof for our theoretical analysis shown in Section 4.

When Assumptions 1 and 2 hold, we can introduce some necessary lemmas for our proof.

Lemma 8. *Under the above assumptions, if learning rate η and batch size $S = k_{Fed}m$ are fixed, we can derive the following analytic solution for the output parameter $\theta_{Fed}(T)$ of federated SGD:*

$$\theta_{Fed}(T) = \frac{1}{n}\sum_{i=1}^{n}\theta_i(T) = \theta_i(0)e^{-T\bar{A}t} + T\sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-T\bar{A}(t-t')}\bar{B}dW(t'). \tag{18}$$

where A_i is the Jacobian matrix and B_i is the covariance matrix for local training on client i, respectively. Besides, we have $\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i$ and $\bar{B} = \frac{1}{n}\sum_{i=1}^{n} B_i$.

Proof. From the result of the Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930), the analytical solution for the local SGD training on client i in the first round $j = 1$ is expressed as follows:

$$\theta_i(1) = \theta_i(0)e^{-A_i t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t'), \tag{19}$$

where $W(t')$ is a white noise and follows $\mathcal{N}(0, I)$. Then based on the update rule of FedAvg defined in Eq. (5), the analytic solution for local training on client i in the round $j = 2$ should be:

$$\theta_i(2) = \frac{1}{n}\sum_{i=1}^{n}\theta_i(1)e^{-A_i t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t'). \tag{20}$$

Substituting Eq.(19) into Eq.(20), we have

$$\begin{aligned}
\theta_i(2) &= \frac{1}{n}\sum_{i=1}^{n}\left(\theta_i(0)e^{-A_i t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t')\right)e^{-A_i t} \\
&\quad + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t') \\
&= \theta_i(0)e^{-2\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{-t}^0 e^{-\bar{A}(t-t')}\bar{B}dW(t') + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t').
\end{aligned} \tag{21}$$

In the same way, we formulate the analytic solution in the round $j = 3$ as follows:

$$\begin{aligned}
\theta_i(3) &= \frac{1}{n}\sum_{i=1}^{n}(\theta_i(0)e^{-2\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{-t}^0 e^{-\bar{A}(t-t')}\bar{B}dW(t') \\
&\quad + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t'))e^{-A_i t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t') \\
&= \theta_i(0)e^{-2\bar{A}t}\frac{1}{n}\sum_{i=1}^{n}e^{-A_i t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{-t}^0 e^{-\bar{A}(t-t')}\bar{B}dW(t')\frac{1}{n}\sum_{i=1}^{n}e^{-A_i t} \\
&\quad + \sqrt{\frac{\eta}{k_{Fed}m}}\frac{1}{n}\sum_{i=1}^{n}\int_0^t e^{-A_i(t-t')}e^{-A_i t}B_i dW(t') + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t') \\
&= \theta_i(0)e^{-3\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\left(\int_{-2t}^{-t} e^{-\bar{A}(t-t')}\bar{B}dW(t') + \int_{-t}^0 e^{-\bar{A}(t-t')}\bar{B}dW(t')\right) \\
&\quad + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t') \\
&= \theta_i(0)e^{-3\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{-2t}^0 e^{-\bar{A}(t-t')}\bar{B}dW(t') + \sqrt{\frac{\eta}{k_{Fed}m}}\int_0^t e^{-A_i(t-t')}B_i dW(t').
\end{aligned} \tag{22}$$

Similarly, the analytic solution after T rounds of federated training can be derived as the following equation:

$$\begin{aligned}
\theta_{Fed}(T) &= \frac{1}{n}\sum_{i=1}^{n}\theta_i(T) \\
&= \theta_i(0)e^{-T\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{(1-T)t}^{0} e^{-\bar{A}(t-t')}\bar{B}dW(t') + \sqrt{\frac{\eta}{k_{Fed}m}}\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{t} e^{-A_i(t-t')}B_i dW(t') \\
&= \theta_i(0)e^{-T\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{(1-T)t}^{0} e^{-\bar{A}(t-t')}\bar{B}dW(t') + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{0}^{t} e^{-\bar{A}(t-t')}\bar{B}dW(t') \\
&= \theta_i(0)e^{-T\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\int_{(1-T)t}^{t} e^{-\bar{A}(t-t')}\bar{B}dW(t') \\
&= \theta_i(0)e^{-T\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\frac{1-e^{-T\bar{A}t}}{\bar{A}}\bar{B} \\
&= \theta_0 e^{-T\bar{A}t} + \sqrt{\frac{\eta}{k_{Fed}m}}\frac{T(1-e^{-T\bar{A}t})}{T\bar{A}}\bar{B} \\
&= \theta_0 e^{-T\bar{A}t} + T\sqrt{\frac{\eta}{k_{Fed}m}}\int_{0}^{t} e^{-T\bar{A}(t-t')}\bar{B}dW(t'),
\end{aligned} \tag{23}$$

which completes the proof. □

Lemma 9. *Under the Assumption 2, the stationary distribution of the Ornstein-Uhlenbeck process for the federated SGD,*

$$q(\theta_{Fed}) = M\exp\left\{-\frac{1}{2}\theta_{Fed}^{\intercal}\Sigma_{Fed}^{-1}\theta\right\}, \tag{24}$$

has the following property,

$$T\bar{A}\Sigma_{Fed} + \Sigma_{Fed}T\bar{A} = \frac{T^2\eta}{k_{Fed}m}\bar{C}. \tag{25}$$

where M is the normalizer and Σ_{Fed} is the covariance matrix of the stationary distribution.

Proof. From Eq.(24), we know that

$$\Sigma_{Fed} = \mathbb{E}_{\theta\sim Q}[\theta_{Fed}\theta_{Fed}^{\intercal}]. \tag{26}$$

Then, according to Eq.(23), we can derive the following equation:

$$\begin{aligned}
T\bar{A}\Sigma_{Fed} + \Sigma_{Fed}T\bar{A} &= \frac{T^2\eta}{k_{Fed}m}\int_{-\infty}^{t} T\bar{A}e^{-T\bar{A}(t-t')}\bar{C}e^{-T\bar{A}(t-t')}dt' \\
&\quad + \frac{T^2\eta}{k_{Fed}m}\int_{-\infty}^{t} e^{-T\bar{A}(t-t')}\bar{C}e^{-T\bar{A}(t-t')}dt'T\bar{A} \\
&= \frac{T^2\eta}{k_{Fed}m}\int_{-\infty}^{t}\frac{d}{dt'}(e^{-T\bar{A}(t-t')}\bar{C}e^{-T\bar{A}(t-t')}) \\
&= \frac{T^2\eta}{k_{Fed}m}\bar{C},
\end{aligned} \tag{27}$$

which completes the proof. □

C.1 Proof of Theorem 2

Next, we start to prove Theorem 2.

Proof. Following the classical Pac-Bayesian framework, we suppose the prior distribution over the parameter space θ is P, and the distribution of the learned hypothesis from the federated SGD algorithm is Q. Then

according to Eq.(24), the densities of the stationary distribution Q and the prior distribution P are respectively $q(\theta)$ and $p(\theta)$ in terms of the parameter θ and can be expressed as the following equations:

$$\begin{aligned} q(\theta) &= \frac{1}{\sqrt{2\pi \det(\Sigma_{Fed})}} \exp\left\{-\frac{1}{2}\theta^\intercal \Sigma_{Fed}^{-1}\theta\right\}, \\ p(\theta) &= \frac{1}{\sqrt{2\pi \det(I)}} \exp\left\{-\frac{1}{2}\theta^\intercal I\theta\right\}. \end{aligned} \tag{28}$$

Thus we have

$$\begin{aligned} \log\left(\frac{q(\theta)}{p(\theta)}\right) &= \log\left(\frac{\sqrt{2\pi \det(I)}}{\sqrt{2\pi \det(\Sigma_{Fed})}} \exp\left\{\frac{1}{2}\theta^\intercal I\theta - \frac{1}{2}\theta^\intercal \Sigma_{Fed}^{-1}\theta\right\}\right) \\ &= \frac{1}{2}\log\left(\frac{1}{\det(\Sigma_{Fed})}\right) + \frac{1}{2}\left(\theta^\intercal I\theta - \theta^\intercal \Sigma_{Fed}^{-1}\theta\right). \end{aligned} \tag{29}$$

Here, we can calculate the KL divergence between the distribution Q and P by applying Eq.(4) in Lemma 1:

$$\begin{aligned} D(Q||P) &= \mathbb{E}_{\theta\sim Q}\left(\log\frac{Q(\theta)}{P(\theta)}\right) \\ &= \int_{\theta\in\Theta} \log\left(\frac{q(\theta)}{p(\theta)}\right) q(\theta)d\theta \\ &= \int_{\theta\in\Theta} \left[\frac{1}{2}\log\left(\frac{1}{\det(\Sigma_{Fed})}\right) + \frac{1}{2}\left(\theta^\intercal I\theta - \theta^\intercal \Sigma_{Fed}^{-1}\theta\right)\right] q(\theta)d\theta \\ &= \frac{1}{2}\log\left(\frac{1}{\sqrt{\det(\Sigma_{Fed})}}\right) + \frac{1}{2}\int_{\theta\in\Theta} \theta^\intercal I\theta q(\theta)d\theta - \frac{1}{2}\int_{\mathbb{R}^{|\mathcal{S}|}} \theta^\intercal \Sigma_{Fed}^{-1} q(\theta)d\theta \\ &= \frac{1}{2}\log\left(\frac{1}{\sqrt{\det(\Sigma_{Fed})}}\right) + \frac{1}{2}\mathbb{E}_{\theta\sim\mathcal{N}(0,\Sigma_{Fed})}\theta^\intercal I\theta - \frac{1}{2}\mathbb{E}_{\theta\sim\mathcal{N}(0,\Sigma_{Fed})}\theta^\intercal \Sigma_{Fed}^{-1}\theta \\ &= \frac{1}{2}\log\left(\frac{1}{\sqrt{\det(\Sigma_{Fed})}}\right) + \frac{1}{2}\mathrm{tr}(\Sigma_{Fed} - I). \end{aligned} \tag{30}$$

Since we have proved from Lemma 9 that $T\bar{A}\Sigma_{Fed} + \Sigma_{Fed}T\bar{A} = \frac{T^2\eta}{k_{Fed}m}\bar{C}$, we have

$$\begin{aligned} \bar{A}\Sigma_{Fed}\bar{A}^{-1} + \Sigma_{Fed} &= \frac{T^2\eta}{Tk_{Fed}m}\bar{C}\bar{A}^{-1} \\ \mathrm{tr}(\bar{A}\Sigma_{Fed}\bar{A}^{-1} + \Sigma_{Fed}) &= \mathrm{tr}(\frac{T\eta}{k_{Fed}m}\bar{C}\bar{A}^{-1}). \end{aligned} \tag{31}$$

For the left-hand side, we can change it to the following equation:

$$\begin{aligned} \mathrm{LHS} &= \mathrm{tr}(\bar{A}\Sigma_{Fed}\bar{A}^{-1} + \Sigma_{Fed}) \\ &= \mathrm{tr}(\bar{A}\Sigma_{Fed}\bar{A}^{-1}) + \mathrm{tr}(\Sigma_{Fed}) \\ &= \mathrm{tr}(\bar{A}\bar{A}^{-1}\Sigma_{Fed}) + \mathrm{tr}(\Sigma_{Fed}) \\ &= \mathrm{tr}(\Sigma_{Fed}) + \mathrm{tr}(\Sigma_{Fed}) \\ &= 2\mathrm{tr}(\Sigma_{Fed}). \end{aligned} \tag{32}$$

Therefore,

$$\mathrm{tr}(\Sigma_{Fed}) = \frac{1}{2}\mathrm{tr}(\frac{T\eta}{k_{Fed}m}\bar{C}\bar{A}^{-1}) = \frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}). \tag{33}$$

On the other side, we can simply calculate that $\mathrm{tr}(I) = d$, because $I \in \mathbb{R}^{d\times d}$, where d is the dimension of the parameter θ. Then we can have

$$\begin{aligned} D(Q_{Fed}||P) &= -\frac{1}{2}\log(\det(\Sigma_{Fed})) + \frac{1}{2}\mathrm{tr}(\Sigma_{Fed}) - \frac{1}{2}\mathrm{tr}(I) \\ &= -\frac{1}{2}\log(\det(\Sigma_{Fed})) + \frac{T\eta}{4k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - \frac{1}{2}d. \end{aligned} \tag{34}$$

By inserting the Eq.(34) into Eq.(3), we can drive the following inequality for the global training sample set of size nm:

$$R(Q_{Fed}) - \hat{R}(Q_{Fed}) \leq \sqrt{\frac{-\log(\det(\Sigma_{Fed})) + \frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm - 2}}, \tag{35}$$

which has completed the proof. □

C.2 Proof of Corollary 3

By a similar approach, we can prove the generalization bound for centralized SGD. We start by proving the required lemmas.

Lemma 10. *Under all assumptions of Lemma 8, if learning rate η and batch size $S = k_{Cen}D$ are fixed, we can derive the following analytic solution for the output parameter of centralized SGD trained on the same amount of training data:*

$$\theta_{Cen}(T) = \theta(0)e^{-\frac{T}{n}At} + \frac{T}{n}\sqrt{\frac{\eta}{k_{Cen}D}}\int_0^t e^{-\frac{T}{n}A(t-t')}BdW(t')). \tag{36}$$

where A is the Jacobian matrix and B is the covariance matrix for training on the centralized dataset of size D.

Proof. Based on Eq.(6) and the result of the Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930), we can simply derive the following analytic solution for the baseline centralized SGD:

$$\theta_{Cen}(T) = \theta(0)e^{-\frac{T}{n}At} + \frac{T}{n}\sqrt{\frac{\eta}{k_{Cen}D}}\int_0^t e^{-\frac{T}{n}A(t-t')}BdW(t')). \tag{37}$$

Thus completing the proof. □

Lemma 11. *When Assumption 2 holds, the Ornstein-Uhlenbeck process's stationary distribution for the baseline centralized SGD,*

$$q(\theta_{Cen}) = M\exp\left\{-\frac{1}{2}\theta^{\intercal}\Sigma_{Cen}^{-1}\theta\right\}, \tag{38}$$

has the following property,

$$\frac{T}{n}A\Sigma_{Cen} + \Sigma_{Cen}\frac{T}{n}A = \frac{T^2\eta}{n^2k_{Cen}D}C. \tag{39}$$

Proof. Based on Eq.(38), we know that

$$\Sigma_{Cen} = \mathbb{E}_{\theta\sim Q}[\theta_{Cen}\theta_{Cen}^{\intercal}]. \tag{40}$$

Then, by combining Eq.(36) and Eq.(40), we can derive the following equation:

$$\begin{aligned}
\frac{T}{n}A\Sigma_{Cen} + \Sigma_{Cen}\frac{T}{n}A &= \frac{T^2\eta}{n^2k_{Cen}D}\int_{-\infty}^t \frac{T}{n}Ae^{-\frac{T}{n}A(t-t')}Ce^{-\frac{T}{n}A(t-t')}dt' \\
&\quad + \frac{T^2\eta}{n^2k_{Cen}D}\int_{-\infty}^t e^{-\frac{T}{n}A(t-t')}Ce^{-\frac{T}{n}A(t-t')}dt'\frac{T}{n}A \\
&= \frac{T^2\eta}{n^2k_{Cen}D}\int_{-\infty}^t \frac{d}{dt'}(e^{-\frac{T}{n}A(t-t')}Ce^{-\frac{T}{n}A(t-t')}) \\
&= \frac{T^2\eta}{n^2k_{Cen}D}C,
\end{aligned} \tag{41}$$

which completes the proof. □

Based on the above lemmas, the proof of Corollary 3 is shown below.

Proof. Since we have proved from Lemma 11 that $\frac{T}{n}A\Sigma_{Cen} + \Sigma_{Cen}\frac{T}{n}A = \frac{T^2\eta}{n^2 k_{Cen} D}C$, we have

$$\begin{aligned} A\Sigma_{Cen} + \Sigma_{Cen}A &= \frac{T\eta}{nk_{Cen}D}C \\ A\Sigma_{Cen}A^{-1} + \Sigma_{Cen} &= \frac{T\eta}{nk_{Cen}D}CA^{-1} \\ \mathrm{tr}(A\Sigma_{Cen}A^{-1} + \Sigma_{Cen}) &= \mathrm{tr}(\frac{T\eta}{nk_{Cen}D}CA^{-1}) \\ 2\mathrm{tr}(\Sigma_{Cen}) &= \mathrm{tr}(\frac{T\eta}{nk_{Cen}D}CA^{-1}) \\ \mathrm{tr}(\Sigma_{Cen}) &= \frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1}). \end{aligned} \tag{42}$$

Like the proof of Theorem 2, by substituting the Eq.(42) into Eq.(30), we can compute the KL divergence between the distribution of the output hypothesis and the prior distribution as below:

$$\begin{aligned} D(Q_{Cen}||P) &= -\frac{1}{2}\log(\det(\Sigma_{Cen})) + \frac{1}{2}\mathrm{tr}(\Sigma_{Cen}) - \frac{1}{2}\mathrm{tr}(I) \\ &= -\frac{1}{2}\log(\det(\Sigma_{Cen})) + \frac{T\eta}{4nk_{Cen}D}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - \frac{1}{2}d. \end{aligned} \tag{43}$$

According to Lemma 1, then we can derive the following inequality to bound the generalization error of the baseline centralized SGD:

$$\begin{aligned} &R(Q_{Cen}) - \hat{R}(Q_{Cen}) \\ &\leq \sqrt{\frac{-\log(\det(\Sigma_{Cen})) + \frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(D) + 4}{4D-2}}. \end{aligned} \tag{44}$$

The proof has been completed. □

C.3 Proof of Theorem 4

Under the new Assumptions 3 and 4, we can characterize the difference between federated and centralized generalization behavior and formally establish Theorem 4 that follows.

Proof. Based on Assumption 3, we can re-formulate Eq.(25) in Lemma 9 to

$$\begin{aligned} T\bar{A}\Sigma_{Fed} + \Sigma_{Fed}T\bar{A} &= \frac{T^2\eta}{k_{Fed}m}\bar{C} \\ 2T\Sigma_{Fed}\bar{A} &= \frac{T^2\eta}{k_{Fed}m}\bar{C} \\ \Sigma_{Fed} &= \frac{T\eta}{2k_{Fed}m}\bar{C}\bar{A}^{-1}. \end{aligned} \tag{45}$$

By substituting Eq.(45) into Eq.(35), we have

$$\begin{aligned} R(Q_{Fed}) - \hat{R}(Q_{Fed}) &\leq \sqrt{\frac{-\log(\det(\frac{T\eta}{2k_{Fed}m}\bar{C}\bar{A}^{-1})) + \frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2}} \\ &\leq \sqrt{\frac{-\log((\frac{T\eta}{2k_{Fed}m})^d\det(\bar{C}\bar{A}^{-1})) + \frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2}} \\ &\leq \sqrt{\frac{d\log(\frac{2k_{Fed}m}{T\eta}) - \log(\det(\bar{C}\bar{A}^{-1})) + \frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2}} \\ &\leq \frac{d\log(\frac{2k_{Fed}m}{T\eta}) - \log(\det(\bar{C}\bar{A}^{-1})) + \frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2}. \end{aligned} \tag{46}$$

Similarly, according to Assumption 3, we can re-formulate Eq.(39) to:

$$\begin{aligned}\frac{T}{n}A\Sigma_{Cen}+\Sigma_{Cen}\frac{T}{n}A&=\frac{T^2\eta}{n^2k_{Cen}D}C\\2\Sigma_{Cen}A&=\frac{T\eta}{nk_{Cen}D}C\\\Sigma_{Cen}&=\frac{T\eta}{2nk_{Cen}D}CA^{-1}.\end{aligned}\tag{47}$$

By inserting Eq.(47) into Eq.(44) and re-arranging the equation, we have

$$\begin{aligned}R(Q_{Cen})-\hat{R}(Q_{Cen})&\leq\sqrt{\frac{-\log(\det(\frac{T\eta}{2nk_{Cen}D}CA^{-1}))+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(\frac{1}{\delta})+2\log(D)+4}{4D-2}}\\&\leq\sqrt{\frac{d\log(\frac{2nk_{Cen}D}{T\eta})-\log(\det(CA^{-1}))+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(\frac{1}{\delta})+2\log(D)+4}{4D-2}}\\&\leq\frac{d\log(\frac{2nk_{Cen}D}{T\eta})-\log(\det(CA^{-1}))+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(\frac{1}{\delta})+2\log(D)+4}{4D-2}\end{aligned}\tag{48}$$

For Eqs.(46) and (48), we define

$$\begin{aligned}\mathcal{G}_{Fed}&=\frac{d\log\left(\frac{2k_{Fed}m}{T\eta}\right)-\log\left(\det\left(\bar{C}\bar{A}^{-1}\right)\right)+\frac{T\eta}{2k_{Fed}m}\mathrm{tr}\left(\bar{C}\bar{A}^{-1}\right)-d+2\log\left(\frac{1}{\delta}\right)+2\log\left(nm\right)+4}{4nm-2},\\\mathcal{G}_{Cen}&=\frac{d\log\left(\frac{2nk_{Cen}D}{T\eta}\right)-\log\left(\det\left(CA^{-1}\right)\right)+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}\left(CA^{-1}\right)-d+2\log\left(\frac{1}{\delta}\right)+2\log\left(D\right)+4}{4D-2}.\end{aligned}\tag{49}$$

The difference between $\mathcal{G}_{Fed}$ and $\mathcal{G}_{Cen}$, which is considered as the gap in the generalization performance, can be derived with the following form:

$$\begin{aligned}\mathcal{G}_{Fed}-\mathcal{G}_{Cen}=&\frac{d\log(\frac{2k_{Fed}m}{T\eta})-\log(\det(\bar{C}\bar{A}^{-1}))+\frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1})-d+2\log(\frac{1}{\delta})+2\log(nm)+4}{4nm-2}\\&-\frac{d\log(\frac{2nk_{Cen}D}{T\eta})-\log(\det(CA^{-1}))+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(\frac{1}{\delta})+2\log(D)+4}{4D-2}.\end{aligned}\tag{50}$$

When Assumption 4 hold, we have:

$$\begin{aligned}\bar{C}\bar{A}^{-1}&=\frac{1}{n^\gamma}(C+\Delta_C)(A+\Delta_A)^{-1}\\&\approx\frac{1}{n^\gamma}(C+\Delta_C)(A^{-1}+A^{-1}\Delta_AA^{-1}).\end{aligned}\tag{51}$$

Hence,

$$\begin{aligned}\mathrm{tr}(\bar{C}\bar{A}^{-1})&=\mathrm{tr}(\frac{1}{n^\gamma}(C+\Delta_C)(A+\Delta_A)^{-1})\\&=\mathrm{tr}(\frac{1}{n^\gamma}(CA^{-1}+\underbrace{CA^{-1}\Delta_AA^{-1}+\Delta_C\left(A^{-1}+A^{-1}\Delta_AA^{-1}\right)}_{\Delta_1}))\\&=\frac{1}{n^\gamma}(\mathrm{tr}(CA^{-1})+\mathrm{tr}(\Delta_1));\end{aligned}\tag{52}$$

$$
\begin{aligned}
-\log(\det(\bar{C}\bar{A}^{-1})) &= -\log(\det(\frac{1}{n^{\gamma}}(C+\Delta_C)(A+\Delta_A)^{-1})) \\
&= -\log(\det(\frac{1}{n^{\gamma}}CA^{-1}\underbrace{(I+C^{-1}\Delta_C)(I+\Delta_A A^{-1})}_{\Delta_2})) \\
&= -\log(\frac{1}{n^{\gamma d}}\det(CA^{-1}\Delta_2)) \\
&= -\log(\frac{1}{n^{\gamma d}}\det(CA^{-1})\det(\Delta_2)) \\
&= \gamma d\log(n) - \log(\det(CA^{-1})) + \log(\det(\Delta_2)^{-1}).
\end{aligned} \tag{53}
$$

Substituting Eqs.(52) and (53) into Eq.(50) derives:

$$
\begin{aligned}
\mathcal{G}_{Fed} - \mathcal{G}_{Cen} &= \frac{d\log(\frac{2k_{Fed}m}{T\eta}) - \log(\det(\bar{C}\bar{A}^{-1})) + \frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2} \\
&\quad - \frac{d\log(\frac{2nk_{Cen}D}{T\eta}) - \log(\det(CA^{-1})) + \frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(D) + 4}{4D-2} \\
&= \frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta}) - \log(\det(CA^{-1})) + \frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1))}{4D-2} \\
&\quad + \frac{\log(\det(\Delta_2)^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4D-2} \\
&\quad - \frac{d\log(\frac{2nk_{Cen}D}{T\eta}) - \log(\det(CA^{-1})) + \frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1}) - d + 2\log(\frac{1}{\delta}) + 2\log(D) + 4}{4D-2} \\
&= \frac{d\log(\frac{n^{\gamma-1}k_{Fed}m}{k_{Cen}D}) + (\frac{T\eta}{2n^{\gamma}k_{Fed}m} - \frac{T\eta}{2nk_{Cen}D})\mathrm{tr}(CA^{-1}) + \frac{T\eta}{2n^{\gamma}k_{Fed}m}\mathrm{tr}(\Delta_1) + \log(\det(\Delta_2)^{-1})}{4D-2}.
\end{aligned} \tag{54}
$$

The proof has been completed.

□

C.4 Proof of Theorem 5

We begin to bound the established form of the generalization gap. This section shows the proof of Theorem 5.

Proof. At the beginning, we construct the following helper function:

$$
f(n) = d\log(\frac{n^{\gamma-1}k_{Fed}m}{k_{Cen}D}) + (\frac{T\eta}{2n^{\gamma}k_{Fed}m} - \frac{T\eta}{2nk_{Cen}D})\mathrm{tr}(CA^{-1}) + \frac{T\eta}{2n^{\gamma}k_{Fed}m}\mathrm{tr}(\Delta_1) + \log(\det(\Delta_2)^{-1}). \tag{55}
$$

Let $S_{Fed} = k_{Fed}m$ and $S_{Cen} = k_{Cen}D$. By the fact that $n \geq 2$, we further derive

$$
\begin{aligned}
f(n) &= d\log(\frac{n^{\gamma-1}S_{Fed}}{S_{Cen}}) + (\frac{T\eta}{2n^{\gamma}S_{Fed}} - \frac{T\eta}{2nS_{Cen}})\mathrm{tr}(CA^{-1}) + \frac{T\eta}{2n^{\gamma}S_{Fed}}\mathrm{tr}(\Delta_1) + \log(\det(\Delta_2)^{-1}) \\
&\leq d\log(\frac{n^{\gamma-1}S_{Fed}}{S_{Cen}}) + (\frac{T\eta}{2n^{\gamma}S_{Fed}} - \frac{T\eta}{2nS_{Cen}})\mathrm{tr}(CA^{-1}) + \frac{T\eta}{2^{\gamma+1}S_{Fed}}\mathrm{tr}(\tilde{\Delta}_1) + \log(\det(\Delta_2)^{-1}),
\end{aligned} \tag{56}
$$

where $\tilde{\Delta}_1$ satisfies $(\tilde{\Delta}_1)_{i,j} = |(\Delta_1)_{i,j}|$. Then, we define

$$
g(n) = d\log(\frac{n^{\gamma-1}S_{Fed}}{S_{Cen}}) + (\frac{T\eta}{2n^{\gamma}S_{Fed}} - \frac{T\eta}{2nS_{Cen}})\mathrm{tr}(CA^{-1}) + \frac{T\eta}{2^{\gamma+1}S_{Fed}}\mathrm{tr}(\tilde{\Delta}_1) + \log(\det(\Delta_2)^{-1}). \tag{57}
$$

The derivative of $g(n)$ is:

$$
g'(n) = \frac{(\gamma-1)d}{n} + \frac{T\eta}{2n^{\gamma+1}S_{Fed}S_{Cen}}(n^{\gamma-1}S_{Fed} - \gamma S_{Cen})\mathrm{tr}(CA^{-1}). \tag{58}
$$

Since $\gamma > 1$, we know that $g'(n) > 0$ requires $n^{\gamma-1}S_{Fed} - \gamma S_{Cen} > 0$, this implies:

$$\begin{aligned} n^{\gamma-1}S_{Fed} &> \gamma S_{Cen} \\ n^{\gamma-1} &> \gamma \frac{S_{Cen}}{S_{Fed}} \\ n^{\gamma-1} &\geq \gamma \\ n &\geq \sqrt[\gamma-1]{\gamma}, \end{aligned} \tag{59}$$

where the third inequality adopts the fact that $S_{Cen} \geq S_{Fed}$. Since the constant γ satisfies $\gamma > 1$, we can prove $g'(n) > 0$ when $n \geq \sqrt[\gamma-1]{\gamma}$. Then, we construct another helper function and the derivative of this new helper function as follows:

$$\begin{aligned} h(x) &= x^{\frac{1}{x-1}} = e^{\frac{1}{x-1}\log(x)} \\ h'(x) &= e^{\frac{1}{x-1}\log(x)} \frac{1 - \frac{1}{x} - \log(x)}{(x-1)^2}. \end{aligned} \tag{60}$$

From Eq.(60), since $1 - \frac{1}{x} - \log(x) < 0$, it is clear that $h'(x) < 0$. Thus, we have $h(x) < h(1) = e$ and $\sqrt[\gamma-1]{\gamma} < e$. According to Eq.(55), the analytic solution of $O(\mathcal{G}_{Fed} - \mathcal{G}_{Cen})$ is monotonically increasing with n when $n \geq e$. Because of $n \in \mathbb{Z}^+$, substituting $n = 3$ and $n = D$ into Eq.(56) will derive the following inequalities for $3 \leq n \leq D$:

$$\begin{aligned} &\frac{d\log(\frac{3^{\gamma-1}k_{Fed}m}{k_{Cen}D}) + T\left(\frac{\eta \mathrm{tr}(CA^{-1})}{2*3^{\gamma}k_{Fed}m} - \frac{\eta \mathrm{tr}(CA^{-1})}{6k_{Cen}D}\right) + \frac{T\eta \mathrm{tr}(\tilde{\Delta}_1)}{2^{\gamma+1}k_{Fed}m} + \log(\det(\Delta_2)^{-1})}{4D-2} \\ &\leq O(\mathcal{G}_{Fed} - \mathcal{G}_{Cen}) \leq \frac{d\log(\frac{D^{\gamma-1}k_{Fed}m}{k_{Cen}D}) + T\left(\frac{\eta(\mathrm{tr}(CA^{-1}))}{2D^{\gamma}k_{Fed}m} - \frac{\eta(\mathrm{tr}(CA^{-1}))}{2D^2k_{Cen}}\right) + \frac{T\eta \mathrm{tr}(\tilde{\Delta}_1)}{2^{\gamma+1}k_{Fed}m} + \log(\det(\Delta_2)^{-1})}{4D-2}. \end{aligned} \tag{61}$$

By again applying the condition $\frac{S_{Fed}}{S_{Cen}} \leq 1$, we can get a tighter lower bound as below:

$$O(\mathcal{G}_{Fed} - \mathcal{G}_{Cen}) \geq \frac{d\log(3^{\gamma-1}) + T\left(\frac{\eta \mathrm{tr}(CA^{-1})}{2*3^{\gamma}k_{Fed}m} - \frac{\eta \mathrm{tr}(CA^{-1})}{6k_{Cen}D}\right) + \frac{T\eta \mathrm{tr}(\tilde{\Delta}_1)}{2^{\gamma+1}k_{Fed}m} + \log(\det(\Delta_2)^{-1})}{4D-2}. \tag{62}$$

However, the lower bound of n is actually $n = 2$. To find the bound of $O(\mathcal{G}_{Fed} - \mathcal{G}_{Cen})$ covering the entire range $\{2 \leq n \leq D | n \in Z\}$, we solve:

$$\begin{aligned} \sqrt[\gamma-1]{\gamma} &= 2 \\ \gamma &= 2. \end{aligned} \tag{63}$$

Since $\sqrt[\gamma-1]{\gamma} < 2$ for any $\gamma > 2$, we know that when $\gamma \geq 2$ is satisfied, the following inequality holds for the case of $n = 2$:

$$O(\mathcal{G}_{Fed} - \mathcal{G}_{Cen}) \geq \frac{d\log(2^{\gamma-1}) + T\left(\frac{\eta(\mathrm{tr}(CA^{-1}) - \eta \mathrm{tr}(\tilde{\Delta}_1))}{2^{\gamma+1}k_{Fed}m} - \frac{\eta \mathrm{tr}(CA^{-1})}{4k_{Cen}D}\right) + \log(\det(\Delta_2)^{-1})}{4D-2}, \tag{64}$$

which is derived by substituting $n = 2$ into the lower bound of Eq.(61). The proof has been completed. □

C.5 Proof of Theorems 6 and 7

Finally, we study which kind of training advantage is possible to close this gap from a theoretical view. We first focus on the validity of n and m and show the following proof for Theorem 6.

Proof. We define $\tilde{\mathcal{G}}_{Fed}$ for the generalization bound of federated scenarios having an advantage in training resources and start with the case of n tends to infinity. The generalization performance gap $\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen}$ for this

case is formulated as the below form:

$$\begin{aligned}\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen} &= \frac{d\log(\frac{2k_{Fed}m}{T\eta})-\log(\det(\bar{C}\bar{A}^{-1}))+\frac{T\eta}{2k_{Fed}m}\mathrm{tr}(\bar{C}\bar{A}^{-1})-d+2\log(\frac{1}{\delta})+2\log(nm)+4}{4nm-2}\\ &\quad-\frac{d\log(\frac{2nk_{Cen}D}{T\eta})-\log(\det(CA^{-1}))+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(\frac{1}{\delta})+2\log(D)+4}{4D-2}\\ &= \frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta})-\log(\det(CA^{-1}))+\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1})+\mathrm{tr}(\Delta_1))}{4nm-2}\\ &\quad+\frac{\log(\det(\Delta_2)^{-1})-d+2\log(\frac{1}{\delta})+2\log(nm)+4}{4nm-2}\\ &\quad-\frac{d\log(\frac{2nk_{Cen}D}{T\eta})-\log(\det(CA^{-1}))+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(\frac{1}{\delta})+2\log(D)+4}{4D-2}\\ &= \frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta})+\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1})+\mathrm{tr}(\Delta_1))+\log(\det(\Delta_2)^{-1})-d+2\log(nm)}{4nm-2}\\ &\quad-\frac{d\log(\frac{2nk_{Cen}D}{T\eta})+\frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(D)}{4D-2}.\end{aligned} \tag{65}$$

According to the definition of PAC-Bayesian bound in Lemma 1, we have $\mathcal{G}_{Cen}>0$. Considering increasing n leads to $nm\geq D$, Eq.(65) turns to

$$\begin{aligned}\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen} &= \frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta})+\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1})+\mathrm{tr}(\Delta_1))+\log(\det(\Delta_2)^{-1})-d+2\log(nm)}{4nm-2}\\ &\quad-\frac{d\log(\frac{2\tilde{n}k_{Cen}D}{T\eta})+\frac{T\eta}{2\tilde{n}k_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(D)}{4D-2},\end{aligned} \tag{66}$$

where $n\geq\tilde{n}$. Then, we derive the limit of $\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}$ when n approaches infinity as follows:

$$\begin{aligned}\lim_{n\to\infty}\left(\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}\right) &= \lim_{n\to\infty}(\frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta})+\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1})+\mathrm{tr}(\Delta_1))+\log(\det(\Delta_2)^{-1})-d+2\log(nm)}{4nm-2}\\ &\quad-\frac{d\log(\frac{2\tilde{n}k_{Cen}D}{T\eta})+\frac{T\eta}{2\tilde{n}k_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(D)}{4D-2})\\ &= \lim_{n\to\infty}\left(O(\frac{(\gamma d+2)\log(n)}{n})+O(\frac{1}{n^{\gamma+1}})+O(\frac{1}{n})-O(1)\right)<0.\end{aligned} \tag{67}$$

Similarly, the limit of $\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}$ when m approaches infinity is established below:

$$\begin{aligned}\lim_{m\to\infty}\left(\tilde{\mathcal{G}}_{Fed}-\mathcal{G}_{Cen}\right) &= \lim_{m\to\infty}(\frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta})+\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1})+\mathrm{tr}(\Delta_1))+\log(\det(\Delta_2)^{-1})-d+2\log(nm)}{4nm-2}\\ &\quad-\frac{d\log(\frac{2\tilde{n}k_{Cen}D}{T\eta})+\frac{T\eta}{2\tilde{n}k_{Cen}D}\mathrm{tr}(CA^{-1})-d+2\log(D)}{4D-2})\\ &= \lim_{m\to\infty}\left(O(\frac{(d+2)\log(m)}{m})+O(\frac{1}{m^2})+O(\frac{1}{m})-O(1)\right)<0,\end{aligned} \tag{68}$$

which completes the proof.

□

Next, we shift our focus to study T and d. The proof of Theorem 7 is demonstrated below.

Proof. Similar to the proof of Theorem 6, we study the case when T tends to positive infinity. Here, we represent the number of iterations for the centralized scenario as T_{Cen}. Increasing the number of communication rounds

T in the federated scenario results in $T \geq T_{Cen}$. Thus, the performance gap $\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen}$ denoted in Eq.(65) can be expressed as follows:

$$\begin{aligned}\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen} = &\frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta}) + \frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1)) + \log(\det(\Delta_2)^{-1}) - d + 2\log(nm)}{4nm - 2} \\ &- \frac{d\log(\frac{2nk_{Cen}D}{T_{Cen}\eta}) + \frac{T_{Cen}\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1}) - d + 2\log(D)}{4D - 2}.\end{aligned} \tag{69}$$

It is easy to recognize that the value of $\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen}$ depends on the first term in the right of Eq.(69) when T tends to infinity. To understand how this term changes as T increases, we need to compare the impact of $d\log\left(\frac{2n^{\gamma}k_{Fed}m}{T\eta}\right)$ and $\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1))$, which is expressed as follows:

$$\begin{aligned}&\lim_{T\to\infty} \frac{d\log\left(\frac{2n^{\gamma}k_{Fed}m}{T\eta}\right)}{\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1))} \\ &= \lim_{T\to\infty} \frac{\frac{d}{dT}\left(d\log\left(\frac{2n^{\gamma}k_{Fed}m}{T\eta}\right)\right)}{\frac{d}{dT}\left(\frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1))\right)} \\ &= \lim_{T\to\infty} \frac{-\frac{d}{T}}{\frac{\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1))} = 0.\end{aligned} \tag{70}$$

From Eq.(70), we know that

$$\lim_{T\to\infty}\left(d\log\left(\frac{2n^{\gamma}k_{Fed}m}{T\eta}\right) + \frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1))\right) = \infty. \tag{71}$$

Hence, we have

$$\lim_{T\to\infty}\left(\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen}\right) = \infty. \tag{72}$$

Then, we consider the case when d tends to positive infinity. Like above, we denote the model size in the centralized scenario as d_{Cen}. Since we attempt to increase the model size d in the federated scenario, we have $d \geq d_{Cen}$. With this condition, we reformulate Eq.(65) with the following form:

$$\begin{aligned}\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen} = &\frac{d\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta}) + \frac{T\eta}{2n^{\gamma}k_{Fed}m}(\mathrm{tr}(CA^{-1}) + \mathrm{tr}(\Delta_1)) + \log(\det(\Delta_2)^{-1}) - d + 2\log(nm)}{4nm - 2} \\ &- \frac{d_{Cen}\log(\frac{2nk_{Cen}D}{T\eta}) + \frac{T\eta}{2nk_{Cen}D}\mathrm{tr}(CA^{-1}) - d_{Cen} + 2\log(D)}{4D - 2}.\end{aligned} \tag{73}$$

When the number of clients is large enough to satisfy $n > \sqrt[\gamma]{\frac{T\eta e}{2k_{Fed}m}}$, we have

$$\begin{aligned}&n^{\gamma} > \frac{T\eta e}{2k_{Fed}m} \\ &\frac{2n^{\gamma}k_{Fed}m}{T\eta} > e \\ &\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta}) > \log(e) \\ &\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta}) - 1 > 0.\end{aligned} \tag{74}$$

Therefore,

$$\begin{aligned}&\lim_{d\to\infty}\left(\tilde{\mathcal{G}}_{Fed} - \mathcal{G}_{Cen}\right) \\ &= \lim_{d\to\infty}\left(\frac{d(\log(\frac{2n^{\gamma}k_{Fed}m}{T\eta}) - 1)}{4nm - 2}\right) = \infty.\end{aligned} \tag{75}$$

The proof has been completed with Eqs.(72) and (75). □

D Complementary Study: Stability and Generalization Superiority between Centralized and Decentralized FL

The gap analysis shown in the main manuscript introduces uniform stability as a complementary theoretical tool. Uniform stability is a standard technique for analyzing the generalization behavior of stochastic optimization algorithms, defined as follows:

Definition 1. ***(Uniform Stability (Sun et al., 2024b))*** *Consider a dataset $\mathcal{D}$ consisting of local datasets across all clients. Let $\tilde{\mathcal{D}}$ be a neighboring dataset that differs from $\mathcal{D}$ in at most one data point within some client's local dataset $\mathcal{D}_i$. A learning algorithm $\mathcal{A}$ is said to be ϵ-uniformly stable if*

$$\sup z \sim \mathcal{D}_i, \mathbb{E}\Big[f(\mathcal{A}(\mathcal{D}), z) - f(\mathcal{A}(\tilde{\mathcal{D}}), z)\Big] \leq \epsilon, \tag{76}$$

where $f(\cdot, z)$ is the loss evaluated at sample z, and the expectation is taken over the randomness of $\mathcal{A}$.

Lemma 12. *((Elisseeff et al., 2005; Hardt et al., 2016)) If a stochastic learning algorithm $\mathcal{A}$ is ϵ-uniformly stable, then its generalization error satisfies $\epsilon_G \leq \epsilon$.*

The above definition and lemma show that bounding the stability ϵ of training algorithms directly provides a bound on the generalization error. In this section, we follow the above concept to establish which of the two prevalent FL paradigms (i.e., centralized and decentralized FL) has the advantage in terms of generalization performance, and use the results to support our problem setup in comparing the generalization between centralized and federated training. To derive the results, we introduce three new assumptions below:

Let $G, L > 0$ and $\|\cdot\|$ denote the Euclidean form.

Assumption 5. *For all $\theta, \tilde{\theta} \in \Theta$ and any data $z \sim \mathcal{D}_i$, the loss $f(\cdot; z)$ is L-smooth:*

$$f(\theta; z) - f(\tilde{\theta}; z) \leq \langle \partial f(\tilde{\theta}; z), \theta - \tilde{\theta} \rangle + \frac{L}{2}\|\theta - \tilde{\theta}\|^2. \tag{77}$$

Assumption 6. *For all $\theta, \tilde{\theta} \in \Theta$ and any data $z \sim \mathcal{D}_i$, the loss $f(\cdot; z)$ is G-Lipschitz continuous:*

$$|f(\theta; z) - f(\tilde{\theta}; z)| \leq G\,\|\theta - \tilde{\theta}\|. \tag{78}$$

Assumption 7. *For all $\theta \in \Theta$, clients $i \in \{1, \ldots, n\}$, and any data $z \sim \mathcal{D}_i$, the stochastic gradient is bounded:*

$$\|\nabla f_i(\theta; z)\| \leq \mathcal{B}. \tag{79}$$

These assumptions are standard in prior uniform stability analyses (Hardt et al., 2016; Sun et al., 2024a; Wang et al., 2022) and enable us to establish the following stability bounds for FL. Noticeably, decentralized FL typically requires full client participation in each round (Yuan et al., 2024), whereas centralized FL can naturally support both partial and full participation. To enable a consistent comparison, we employ the random walk view of decentralized communication (Ayache and El Rouayheb, 2019; Sun et al., 2022), which provides a standard mathematical tool to represent partial client participation. This modeling choice is purely for analytical consistency and does not change the practical interpretation of decentralized FL. It allows us to align the participation assumptions across paradigms and then extend the analysis to the more common full-participation decentralized FL setting.

Theorem 13. ***(Stability and Generalization of Decentralized FL)*** *Let the learning rate be constant $\eta_t \equiv \eta$ and define $\rho = 1 + \eta L$. Consider a decentralized FL process on a network of n clients with the maximum degree $deg_{\max}$ and the number of edges E. For the case with partial client participation, let the number of clients participating in each communication round be $K \in \{1, \ldots, n\}$. Then, under the above assumptions, the expected stability and generalization for this training can be formulated as:*

$$\mathbb{E}\big[\,\big|f(w^T; z) - f(\tilde{w}^T; z)\big|\,\big] \leq\ \eta G\mathcal{B}\left[\, n c_{\mathcal{P}}\, \rho^{T-1}\, \frac{1 - \big(\lambda_2(\mathcal{P})/\rho\big)^T}{1 - \big(\lambda_2(\mathcal{P})/\rho\big)}\ +\ \frac{\rho^T - 1}{\rho - 1} \cdot \frac{2(1 - \big(1 - \frac{deg_{\max}}{2E}\big)^K)}{K}\,\right], \tag{80}$$

where $\mathcal{P}$ is the transition matrix of the communication walk, $c_{\mathcal{P}}$ is a contraction constant depending on $\mathcal{P}$, and $\lambda_2(\mathcal{P})$ is the second largest eigenvalue of $\mathcal{P}$. While for the classical decentralized FL with full client participation,

the expected stability and generalization are bounded by:

$$\mathbb{E}\big[\,\big|f(w^T;z)-f(\tilde{w}^T;z)\big|\,\big] \le\ \eta G\mathcal{B}\left[\,\rho^{T-1}\,\frac{1-\big(\lambda_2(\mathbb{W})/\rho\big)^T}{1-\big(\lambda_2(\mathbb{W})/\rho\big)}\ +\ \frac{\rho^T-1}{\rho-1}\cdot\frac{2}{n}\,\right]. \tag{81}$$

where $\mathbb{W}\in\mathbb{R}^{n\times n}$ is the communication weight matrix employed for decentralized mixing across all n clients and $\lambda_2(\mathbb{W})$ is the second largest eigenvalue of $\mathbb{W}$.

Proof. To prove this theorem, we couple two executions of decentralized FL on training datasets that differ in exactly one example and start with the form of partial client participation. The two runs produce parameter sequences $\{\theta^t\}_{t=0}^T$ and $\{\tilde{\theta}^t\}_{t=0}^T$. We define the deviation between them by $\Delta_t := \|\theta^t-\tilde{\theta}^t\|$. At iteration t, let the two coupled runs visit clients i_t and $\tilde{i}_t$, and let the corresponding local samples be z_{i_t} and $z_{\tilde{i}_t}$. We consider the following two cases.

Firstly, if the same sample is selected for both training, we have ***Case A:***

$$\begin{aligned}\Delta_{t+1} &= \big\|\theta^t-\tilde{\theta}^t-\eta\big(\nabla f_{i_t}(\theta^t;z_{i_t})-\nabla f_{i_t}(\tilde{\theta}^t;z_{i_t})\big)\big\|\\ &\le \big\|\theta^t-\tilde{\theta}^t\big\|+\eta\big\|\nabla f_{i_t}(\theta^t;z_{i_t})-\nabla f_{i_t}(\tilde{\theta}^t;z_{i_t})\big\|\\ &\le (1+\eta L)\,\Delta_t.\end{aligned} \tag{82}$$

Otherwise, we have ***Case B:***

$$\begin{aligned}\Delta_{t+1} &= \big\|\theta^t-\tilde{\theta}^t-\eta\big(\nabla f_{i_t}(\theta^t;z_{i_t})-\nabla f_{\tilde{i}_t}(\tilde{\theta}^t;z_{\tilde{i}_t})\big)\big\|\\ &\le \big\|\theta^t-\tilde{\theta}^t\big\|+\eta\big\|\nabla f_{i_t}(w^t;z_{i_t})-\nabla f_{\tilde{i}_t}(\tilde{\theta}^t;z_{i_t})\big\|+\eta\big\|\nabla f_{i_t}(\tilde{\theta}^t;z_{i_t})-\nabla f_{\tilde{i}_t}(\tilde{\theta}^t;z_{\tilde{i}_t})\big\|\\ &\le (1+\eta L)\,\Delta_t+\eta\,\beta_t,\end{aligned} \tag{83}$$

where we define $\beta_t = \big\|\nabla f_{i_t}(\tilde{\theta}^t;z_{i_t})-\nabla f_{\tilde{i}_t}(\tilde{\theta}^t;z_{\tilde{i}_t})\big\|$. Combining the two cases yields

$$\Delta_{t+1}\le(1+\eta L)\Delta_t+\eta\,\beta_t. \tag{84}$$

Extending this into the recursion of T steps further produces

$$\Delta_T\le\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)\eta_t\beta_t. \tag{85}$$

According to Assumption 7, the following statement holds depending on the circumstance if the same client i is selected for both training:

$$\delta_t\le\begin{cases}2\mathcal{B}, & i_t\neq\tilde{i}_t,\\ 2\mathcal{B}, & i_t=\tilde{i}_t=i,\\ 0, & \text{otherwise.}\end{cases} \tag{86}$$

Then, by taking expectations in (85) and using Assumption 5, we establish

$$\begin{aligned}E[\Delta_T] &\le E[\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)\eta_t\beta_t]\\ &=\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)\eta_t E[\beta_t]\\ &\le\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)2\eta_t\mathcal{B}(Pr(i_t\neq\tilde{i}_t)+Pr(i_t=\tilde{i}_t=i)).\end{aligned} \tag{87}$$

Let μ,ν be the initial distributions of the two coupled communication walks, and let P be the transition kernel of the communication process. Considering that each walk is determined by the selected training client in the

last round, it thus follows a Markovian chain (Wang et al., 2022). Standard mixing estimates give

$$\begin{aligned}\Pr(i_t \neq \tilde{i}_t) &\leq \|\mu\mathcal{P}^t - \nu\mathcal{P}^t\|_{\mathrm{TV}} \\ &= \frac{1}{2}\|\mu\mathcal{P}^t - \nu\mathcal{P}^t\|_1 \\ &\leq \frac{n}{2}\|\mu\mathcal{P}^t - \nu\mathcal{P}^t\|_\infty \\ &\leq \frac{n}{2}c_{\mathcal{P}}\lambda_2^t,\end{aligned} \tag{88}$$

where $||\cdot||_{TV}$ denotes the total variation. For the collision probability $\Pr(i_t = \tilde{i}_t = i)$, using the stationary distribution $\pi(i) = \deg_i/(2E)$ and the upper bound $\sum_i \pi(i)^2 \leq \max_i \pi(i) \leq \deg_{\max}/(2E)$, we have

$$\Pr(i_t = \tilde{i}_t = i) \leq \frac{\deg_{\max}}{2E}. \tag{89}$$

Substituting (88) and (89) into (87) gives the single-walk estimate

$$E[\Delta_T] \leq 2\mathcal{B}\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)\eta_t(\frac{n}{2}c_{\mathcal{P}}\lambda_2^t + \frac{\deg_{max}}{2E}). \tag{90}$$

By the Assumption 6,

$$\begin{aligned}\mathbb{E}\left[|f\left(\theta^T;z\right) - f\left(\tilde{\theta}^T;z\right)|\right] &\leq G\mathbb{E}\left[||\theta^T - \tilde{\theta}^T||\right] \\ &\leq G\mathcal{B}\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)\eta_t(nc_{\mathcal{P}}\lambda_2^t + \frac{\deg_{max}}{E})\end{aligned} \tag{91}$$

Next, we attempt to generalize the above result to k parallel communication walks. Considering that each work adopts the same initial learning rate and learning rate decay (i.e., $\forall_j : \eta_s^j \equiv \eta_s, \eta_t^j \equiv \eta_t$), we have

$$\begin{aligned}&\mathbb{E}\left[|f\left(\theta^T;z\right) - f\left(\tilde{\theta}^T;z\right)|\right] \\ &\leq GE\left[||\theta^T - \tilde{\theta}^T||\right] \\ &= GE\left[||\frac{1}{K}\sum_{j=1}^K \theta_j^T - \frac{1}{K}\sum_{j=1}^K \tilde{\theta}_j^T||\right] \\ &= GE\left[||\frac{1}{K}\sum_{j=1}^K (\theta_j^T - \tilde{\theta}_j^T)||\right] \\ &= \frac{G}{K}E\left[||\sum_{j=1}^K (\theta_j^T - \tilde{\theta_j^T})||\right] \\ &\leq \frac{G}{K}\sum_{j=1}^K\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)2\eta_t\mathcal{B}(Pr(i_t^j \neq \tilde{i}_t^j) + Pr(i_t^j = \tilde{i}_t^j = i)) \\ &\leq \frac{G}{K}\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_s L)\right)2\eta_t\mathcal{B}(Pr(\exists_j : i_t^j \neq \tilde{i}_t^j) + Pr(\exists_j : i_t^j = \tilde{i}_t^j = i)).\end{aligned} \tag{92}$$

Based on Eqs.(88) and (89) and by applying a union bound, we find

$$\Pr\left(\exists j :\, i_t^j \neq \tilde{i}_t^j\right) \leq \sum_{j=1}^K \frac{n}{2}c_{\mathcal{P}}\lambda_2^t = K\cdot\frac{n}{2}c_{\mathcal{P}}\lambda_2^t, \tag{93}$$

and

$$\Pr\left(\exists j :\, i_t^j = \tilde{i}_t^j = i\right) \leq 1 - \left(1 - \frac{\deg_{\max}}{2E}\right)^K. \tag{94}$$

Plugging Eqs.(93) and (94) into Eq.(92) produces

$$\begin{aligned}&\mathbb{E}\Big[\big|f(\theta^T;z)-f(\tilde{\theta}^T;z)\big|\Big]\\&\leq\frac{G}{K}\mathcal{B}\sum_{t=0}^{T-1}\left(\prod_{s=t+1}^{T-1}(1+\eta_sL)\right)\eta_t(Knc_{\mathcal{P}}\lambda_2^t+2(1-\left(1-\frac{\deg_{\max}}{2E}\right)^K)).\end{aligned}\tag{95}$$

Let $\eta_t\equiv\eta$, and $\rho=1+\eta L$. We can simplify the above equation into

$$\mathbb{E}\left[|f\left(\theta^T;z\right)-f\left(\tilde{\theta}^T;z\right)|\right]\leq G\eta\mathcal{B}\sum_{t=0}^{T-1}(1+\eta L)^{T-1-t}(nc_{\mathcal{P}}\lambda_2^t+\frac{2}{K}(1-(1-\frac{\deg_{max}}{2E})^K)).\tag{96}$$

This further simplifies into the following closed form:

$$\mathbb{E}\left[|f\left(\theta^T;z\right)-f\left(\tilde{\theta}^T;z\right)|\right]\leq\eta G\mathcal{B}\left[nc_{\mathcal{P}}\,\rho^{T-1}\,\frac{1-\left(\lambda_2/\rho\right)^T}{1-\left(\lambda_2/\rho\right)}\;+\;\frac{\rho^T-1}{\rho-1}\cdot\frac{2(1-\left(1-\frac{\deg_{\max}}{E}\right)^K)}{K}\right],\tag{97}$$

by the fact that

$$\begin{aligned}&\sum_{t=0}^{T-1}\rho^{T-1-t}\lambda_2^t=\rho^{T-1}\sum_{t=0}^{T-1}(\frac{\lambda_2}{\rho})^t=\rho^{T-1}\cdot\frac{1-(\lambda_2/\rho)^T}{1-(\lambda_2/\rho)},\\&\sum_{t=0}^{T-1}\rho^{T-1-t}=\sum_{s=0}^{T-1}\rho^s=\frac{\rho^T-1}{\rho-1}.\end{aligned}\tag{98}$$

This completes the proof for decentralized FL with partial client participation.

We next extend this to establish the stability and generalization of decentralized FL with full client participation. In this setting, every client is activated in each round, and the communication step is represented not by a random walk with transition kernel $\mathcal{P}$, but by a deterministic mixing through the communication weight matrix $\mathbb{W}\in\mathbb{R}^{n\times n}$. As a result, the divergence term $\|\mu\mathcal{P}^t-\nu\mathcal{P}^t\|$ in the partial-participation proof is replaced by the spectral contraction of $\mathbb{W}$, which decays at rate $\lambda_2(\mathbb{W})^t$ with $\lambda_2(\mathbb{W})$ being the second-largest eigenvalue of $\mathbb{W}$. Moreover, since all clients participate, the collision probability no longer needs to be bounded by degree-based arguments; instead, the effect of a single-sample perturbation is deterministically diluted by the factor $1/n$ across all clients. Following the same recursion as before, we obtain the stability bound

$$\mathbb{E}\Big[\big|f(\theta^T;z)-f(\tilde{\theta}^T;z)\big|\Big]\;\leq\;\eta G\mathcal{B}\left[\rho^{T-1}\frac{1-(\lambda_2(\mathbb{W})/\rho)^T}{1-\lambda_2(\mathbb{W})/\rho}+\frac{\rho^T-1}{\rho-1}\cdot\frac{2}{n}\right],\tag{99}$$

where $1-\lambda_2(W)$ denotes the spectral gap of the communication graph. This completes the proof for decentralized FL with full client participation. With Eqs.(97) and (99), the full proof for this theorem is completed. □

Corollary 14. ***(Stability and Generalization of Centralized FL)*** *Let the learning rate be constant $\eta_t\equiv\eta$ and define $\rho=1+\eta L$. Consider a centralized FL process on a network of n clients and let the number of clients participating in each communication round be $K\in\{1,\ldots,n\}$. Then, under the above assumptions, the expected stability and generalization for this training can be formulated as:*

$$\mathbb{E}\big[\,\big|f(w^T;z)-f(\tilde{w}^T;z)\big|\,\big]\;\leq\;\eta G\mathcal{B}\frac{\rho^T-1}{\rho-1}\cdot\frac{2}{n}.\tag{100}$$

Proof. The proof follows directly from the decentralized FL analysis. In decentralized FL with full participation, the stability bound contains a term involving $\lambda_2(\mathbb{W})$ due to the spectral gap of the network topology. In centralized FL, however, aggregation is performed by the server and corresponds to IID sampling of clients each round, so there is no network mixing effect and the spectral term vanishes. The second term, which captures the effect of a single-sample perturbation, remains $2/n$ because under partial participation each client is selected with probability K/n while its contribution is weighted by $1/K$, so the two factors cancel and the expected influence of any client is $1/n$. Substituting these into the recursive bound of Eq.(84) and applying the assumptions then gives

$$\mathbb{E}\Big[\,\big|f(\theta^T;z)-f(\tilde{\theta}^T;z)\big|\,\Big]\;\leq\;\eta G\mathcal{B}\frac{\rho^T-1}{\rho-1}\cdot\frac{2}{n},\tag{101}$$

which is exactly Eq.(100), completing the proof. □

Remark 3. *The bounds derived in Theorem 13 reveal how the stability of decentralized FL depends jointly on the client participation size K and the network connectivity. In the partial-participation case, the first term in Eq.*(80) *grows with the contraction constant $c_{\mathcal{P}}$ and the spectral factor $\lambda_2(\mathcal{P})$, indicating that poor connectivity and slower mixing enlarge the bound. The second term is inversely proportional to K, so increasing the number of participating clients per round improves stability by reducing the perturbation contribution. In the case of full participation (i.e., in the classical form of decentralized FL), Eq.*(99) *shows that the network eigenvalue $\lambda_2(\mathbb{W})$ still affects generalization, but the generalization performance is better than in the partial participation case due to the smaller diluted factor of $1/n$. Comparing these with Corollary 14, we see that centralized FL holds a strictly tighter bound under equal conditions. This is because the central server aggregates globally, which corresponds to setting the spectral term to zero (no dependency on λ_2 or $c_{\mathcal{P}}$), and the perturbation term always scales as $2/n$ regardless of K. Notably, this superiority holds for both participation regimes: centralized FL with either partial or full participation yields a smaller generalization bound than decentralized FL under the same setup. Therefore, centralized FL is guaranteed to generalize at least as well as decentralized FL, justifying our problem setup choice on centralized FL as the canonical representative of FL when analyzing the precise gap against centralized learning. We also provide an empirical sanity check in Section F.1 to support this claim.*

E Detailed Experiment Setup

In this section, we present the details of our experiment setup through two tables. Table 2 details the experiment system, covering the specific settings for model architecture, dataset, federated scenario, and training. Table 3 outlines the running environment, including the configuration of the executed codes and the test server.

Table 2: **Experiment System Settings.**

System	Value
Model Architecture	Vision Transformer (ViT) (Dosovitskiy et al., 2020) ResNet (He et al., 2016)
Dataset	Mini-ImageNet (Vinyals et al., 2016) CIFAR-10 (Krizhevsky, 2009)
Range on Communication Rounds	$25 \leq T \leq 100$
Range on Number of Clients	$2 \leq n \leq 100$
Data Distribution on Clients	Non-IID ($\alpha = 0.1$ (default))
Decentralized Network Connectivity	0.5 (default)
ViT Model Size Options (Millions)	$\{7.91, 15.00, 22.08, 29.17, 36.26,$ $43.35, 50.44, 57.52, 64.61, 71.70\}$
ResNet Model Size Options (Millions)	$\{4.91, 11.18, 17.45, 23.72, 29.99,$ $36.26, 42.54, 48.81, 55.08, 61.35\}$
Local Training Epochs	$t = 2$
Batch Size	256
Base Learning Rate	1.5e-4

Table 3: **Running Environment Settings.**

Config	Details
Server GPU Count	8
Server GPU Type	RTX A5000 (24GB)
Server CPU Type	AMD EPYC 7513 32-core
Programming Language	Python
CUDA	11.3
Framework	PyTorch

F Additional Experiments

F.1 Generalization Comparison between Centralized and Decentralized FL

To further substantiate our stability-based analysis in Appendix D, we present an additional experiment comparing centralized FL with decentralized FL under equal training resources (i.e., same model size, identical training data, and equal training compute). The results are reported in Table 4. Several clear observations emerge. First, centralized FL consistently outperforms decentralized FL across both model architectures (Tiny-ViT and ResNet-18) and datasets (CIFAR-10 and Mini-ImageNet). This empirically confirms our stability-based conclusion that centralized FL enjoys a generalization advantage over decentralized FL. Furthermore, since centralized FL already represents the best generalization performance achievable under the federated setting, our PAC-Bayesian findings, such as the insight that the gap can be eliminated only by introducing new clients or adding new data, naturally extend to decentralized FL as well, where the gap to centralized training is even larger.

Second, within the decentralized FL results, we observe a strong dependence on network connectivity. A denser peer-to-peer network initialized using the Erdős-Rényi model (Erdős and Rényi, 1959) with 0.5 connectivity produces substantially higher accuracy than a sparse ring topology, confirming our results in Theorem 13 that the stability and generalization of decentralized FL are shaped by the communication graph structure. In other words, better-connected networks facilitate more effective information mixing and reduce the degradation caused by data heterogeneity, while sparse networks amplify instability and yield poorer generalization.

Table 4: **Top-1 accuracy (%) on CIFAR-10 and Mini-ImageNet.** Models are trained by different learning paradigms with equal training resources. The table reports the mean of three trials with random seeds.

	CIFAR-10 (%)		**Mini-ImageNet (%)**	
Method	Tiny-ViT	ResNet-18	Tiny-ViT	ResNet-18
Centralized Learning	55.97	66.85	29.51	37.63
Centralized FL	38.81	46.31	22.28	22.34
Decentralized FL (ER graph with 0.5 connectivity)	32.02	37.35	14.86	15.30
Decentralized FL (Ring graph)	28.93	30.07	12.96	14.43