# Expediting Federated Learning on Non-IID Data by Maximizing Communication Channel Utilization

Qi Tan, *Member, IEEE*, Yi Zhao, *Member, IEEE*, Qi Li, *Senior Member, IEEE*, and Ke Xu, *Fellow, IEEE*

***Abstract*—Federated learning (FL) is at the core of intelligent Internet architecture. It allows clients to jointly train a model without direct data sharing. In such a process, clients and the central server share information through communication channels formed by parameters. However, the non-iid training data in clients significantly impacts global model convergence and brings difficulties for the evaluation of local contributions. Most of existing studies try to expand the communication channel by improving consistency with variance reduction or regularization, but such methods neglect an important factor, i.e., channel utilization, hence their capability for sharing information is under-utilized. Moreover, the issue of contribution evaluation is still unsolved. In this paper, we simultaneously solve the former two challenges (i.e., model convergence and contribution evaluation) by modeling the indirect data sharing of FL as a problem of information communication. We prove that FL with non-iid data forms noisy communication channels, which have limited capability for information transmission, i.e., limited channel capacity. The main factor in deciding the channel capacity is the Gradient Signal to Noise Ratio (GSNR). Through analyzing GSNR, we further prove that channel capacity can be reached by optimal local updates and propose a method FedGSNR to calculate it, which allows us to maximize channel utilization in FL, leading to faster model convergence. Moreover, as the contribution of the local dataset depends on the amount of provided information, the derived GSNR allows the server to accurately evaluate the contributions of different clients (i.e., the quality of local datasets).**

***Index Terms*—Federated learning, information communication, gradient signal to noise ratio, channel utilization.**

## I. Introduction

FEDERATED learning (FL) [25] is at the core of intelligent Internet architecture [37], [40]. It focuses on the practical scenario with multiple clients to collaboratively train a model without direct data sharing. Typically, FL sends the global model to different clients, then all clients optimize the global model with their own datasets, which always follow non-iid distributions in reality. Finally, optimization results bring back the information of local datasets for server aggregation. In a nutshell, the server and all clients share information through transmitting gradients, i.e., the model gradients form a logical communication channel. Particularly, FL algorithms such as FedAvg [25] typically accelerate global model convergence through multiple local updates, which load more information from local datasets onto the shared gradients.

Although it achieves great performance in various practical applications, FL on non-iid data still has unknown territories. Previous studies [6], [9], [16], [17], [22], [38] devote to analyze the convergence of FL with different conditions. Meanwhile, [1], [15], [27], [34], [35] utilize variance reduction or regularization to improve the consistency among all clients, which alleviates the impact of non-iid distributed data. Most of these investigations deal with identical local updates, and the improving consistency makes the communication channel wider, which can hold more information. Whereas, as an important factor for FL, the issue of channel utilization has not been well studied. For instance, if the channel utilization is sufficiently low, we cannot get the expected information even if we have a wider channel.

To bridge the gap, we model the indirect data sharing of FL as a problem of information communication. Based on this insight, multiple local updates increase the amount of information loaded on the gradients, which means we can get more information within fewer communication rounds and it possibly leads to the faster convergence. Moreover, as illustrated in Fig. 1, if data are non-iid distributed among

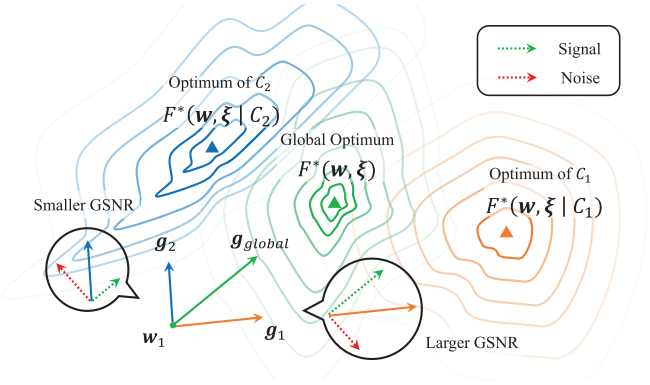



Fig. 1. Overview of GSNR. Different local gradients ($\boldsymbol{g}_1$ and $\boldsymbol{g}_2$) can be decomposed into two components: signal and noise, the former parallels to the global gradient ($\boldsymbol{g}_{\text{global}}$), and the latter orthogonal to it. GSNR is defined as the ratio between the norm of these components.

Received 24 October 2023; revised 6 August 2024; accepted 7 April 2025; approved by IEEE TRANSACTIONS ON NETWORKING Editor G. Joshi. Date of publication 20 May 2025; date of current version 17 October 2025. This work was supported in part by the National Science Foundation for Distinguished Young Scholars of China under Grant 62425201; in part by the National Natural Science Foundation of China under Grant 62472036, Grant 62202258, Grant 62132011, and Grant U22B2031; and in part by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under Grant 62221003. *(Corresponding authors: Yi Zhao; Ke Xu.)*

Qi Tan is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: tanqi@szu.edu.cn).

Yi Zhao is with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: zhaoyi@bit.edu.cn).

Qi Li is with the Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China (e-mail: qli01@tsinghua.edu.cn).

Ke Xu is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Zhongguancun Laboratory, Beijing 100094, China (e-mail: xuke@tsinghua.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TON.2025.3565822, provided by the authors.

Digital Object Identifier 10.1109/TON.2025.3565822

all clients, the local gradients are different from the global gradient (the ideal gradients calculated by global data distribution), which means the communication channels are noisy. Inspired by information theory, the transmission capability of a noisy channel is limited, and the achievable channel capacity (the maximum capability for information transmission) is decided by the Signal to Noise Ratio (SNR), i.e., Shannon's formula: $C = W \cdot \log(1 + \text{SNR})$ [5], where $C$ and $W$ represent the channel capacity and the bandwidth of the channel respectively, SNR is the signal-to-noise ratio. Hence, both superabundant local updates and insufficient local updates lead to low channel utilization, which under-utilizes existing channels. The optimal strategy is to calculate the optimal number of local updates to reach the channel capacity.

To formalize SNR in FL scenario, i.e., the Gradient Signal to Noise Ratio (GSNR), we first define the signal components and the noise components. As illustrated in Fig. 1, in the non-iid scenario, we can decompose the local gradient ($\boldsymbol{g}_1$ or $\boldsymbol{g}_2$ in Fig. 1) into mutually orthogonal signal component and noise component, the signal component is parallel to the global gradient, while the noise component is orthogonal to it. GSNR is defined as the ratio between the norm of the signal component and the noise component. Based on these ideas, we prove that GSNR can be calculated by the optimal number of local updates, which is consistent with the intuition that utilizing optimal local updates reaches the channel capacity, and the maximal channel utilization accelerates the convergence of FL. Using this insight, we propose FedGSNR, a practical method to calculate the optimal number of local updates for different clients. Moreover, we also develop a specific method to calculate GSNR for each client, which allows the server to evaluate the contribution of each client.

Particularly, the newly proposed GSNR strategy FedGSNR is orthogonal to existing methods, which mostly devote to improving the consistency in order to expand channels, it can help these methods to further improve the performance by maximizing channel utilization.

In summary, our contributions in this paper are as follows:

- We first propose to model FL as a problem of information communication and prove that the communication channel formed by the gradients converges to the Gaussian channel.
- We prove GSNR can be calculated by the optimal local updates, which decides the maximum capability of the communication channel for information transmission, and the optimal local updates maximizes channel utilization, leading to faster convergence.
- We propose FedGSNR, a practical algorithm that can be combined with various FL algorithms (e.g., FedAvg, FedProx, Scaffold, etc) to calculate their optimal local updates, which maximizes their channel utilization to further accelerate convergence.
- We derive a function $r(\boldsymbol{w})$ to calculate GSNR, which can be used to evaluate the local contributions of different clients.
- We validate our theoretical results on CIFAR-10 and CIFAR-100 datasets, and the experiments indicate that FedGSNR with different FL algorithms achieve on average a $1.69\times$ speedup over their original version, and $r(\boldsymbol{w})$ is an accurate metric for local contributions.

The rest of this paper is organized as follows: Section II presents the related work. Section III provides the preliminaries for the following analysis. Section IV explores the correlations between GSNR and optimal local updates, derives the method to calculate the optimal local updates, and provides the method to calculate corresponding GSNR. Section V provides a practical algorithm FedGSNR to maximize the channel utilization. Then we provide the convergence analysis of FedGSNR in Section VI and analyze the concepts of GSNR in detail in Section VII. Section VIII validates our theoretical results with real-world datasets. Finally, Section IX concludes the paper.

## II. Related Work

There has been a lot of literatures devoted to improving FL, including the convergence [15], [22], [31], [34], the robustness [7], [20], [28], and the data privacy [2], [4], [26], [41]. Regarding GSNR, [23] and [30] try to analyze the generalization and variational bounds with such a concept. In this work, we focus on the relationship between GSNR and the optimal local updates in FL scenarios. To control the noise component (client drift), SCAFFOLD [15] proposes a specific gradient calculating method based on variance reduction. FedProx [21] indicates that under non-iid FL conditions, a large number of local updates lead to divergence or instability. Moreover, FedNova [35] tries to stabilize the training process with a new average strategy. Additionally, [36] proposes a practical optimization problem with the resource constraints, and it determines the number of local updates for each client according to the resource constraints.

A similar work [17] derives an upper bound of local updates by the total iterations $T$ and the number of clients $M$, which provides a theoretical analysis of local updates. However, they treat each client equally, and fail to propose a method to calculate the optimal number of local updates directly from the heterogeneous data.

So far, we have explained the difference between the correlated literature and our work. To understand the concepts of GSNR more comprehensively, we will discuss more researches in FL, which are correlated to GSNR.

### A. Gradient Diversity

Gradient diversity is a key ingredient of FL, which captures the differences between the datasets possessed by different clients. Reference [38] employs gradient diversity to investigate the relationship between batch size and the convergence rate in parallel SGD. Reference [39] analyzes why periodical model averaging is suitable for deep learning, and provides a deep understanding of model averaging. Reference [8] tries to mitigate the gradient diversity through sharing a small batch of data among all clients, but it also introduces higher privacy risks. Reference [1] introduces a dynamic regularization term to resolve the problem of gradient divergence. In summary, most of the previous investigations try to solve the issue of gradient diversity through the gradient calculating, such as gradient prediction, regularization, personalized target function, and so on. However, the influence of local updates gains less attention. In this paper, we propose a new perspective to analyze the optimization process by Gradient Signal to Noise Ratio, it reduces the required communication rounds via an elaborate configuration of local updates and proposes a method to evaluate the contributions of different clients.

### B. Personalization in FL

Another important problem in federated learning is personalization. Formally, personalization transforms the optimization issue from the global distribution to a specific local

distribution on client $C_k$, it scarifies the global performance in order to gain more benefit in local scenario. Reference [19] reviews the investigations of personalization, and the situations are divided into three categories: device heterogeneity, data heterogeneity, and model heterogeneity, the last one is the motivation of personalization. Reference [24] proposes three methods to achieve personalization, these methods try to balance the model performance on global data distribution and the local data distribution. Reference [32] proposes a method to optimize the global model and local model separately in order to make the local model more personalized. Reference [13] proposes three objectives to make personalization easily. However, personalization is an important topic in FL since different clients confront different issues, but if we greedily utilize global information for personalization, there is likely to appear Prisoner's Dilemma, the collective benefit for all clients is not optimal. Therefore, the profit for each client can probably be further improved. Hence, cooperation is also an important problem, and the better goal of personalization is to search for optimum on conditional data distribution combined with cooperation.

## III. Preliminary

### A. Federated Averaging (FedAvg)

In FL, we have a set of clients $\boldsymbol{C} = \{C_1, \cdots, C_K\}$ and the problem can be formalized as:

$$\begin{aligned}\min_{\boldsymbol{w}} \overline{F(\boldsymbol{w})} &:= \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w},\ \boldsymbol{\xi})] = \mathbb{E}_C[\mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w},\ \boldsymbol{\xi})] \mid C] \\ &:= \sum_{k=1}^{K} P(C = C_k) \cdot \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w},\ \boldsymbol{\xi}|C_k)], \end{aligned} \tag{1}$$

where $F(\boldsymbol{w},\ \cdot\ )$ is a specified loss function with model $\boldsymbol{w}$, $K$ is the number of clients, and $P(C)$ is a discrete probability distribution correlated to the importance of different clients. Usually, $P(C)$ is a uniform distribution or proportional to local data quantity. $\boldsymbol{\xi}|C_k$ is a random sample drawn from the dataset of $C_k$, i.e., $\boldsymbol{\xi}|C_k \sim p(\boldsymbol{x}|C_k)$.

Regarding traditional machine learning, the global dataset is gathered from all clients, and the goal is to minimize

$$F(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{\xi}}[F(\boldsymbol{w},\ \boldsymbol{\xi})], \tag{2}$$

where $\boldsymbol{\xi}$ is a random sample of global dataset, i.e., $\boldsymbol{\xi} \sim p(\boldsymbol{x}) = \sum_{k=1}^{K} P(C = C_k) p(\boldsymbol{x}|C_k)$. However, we cannot gather data from different clients in most cases due to privacy concerns. Thus, we separate the target function as Eq. (1), and send the initial model $\boldsymbol{w}_1$ to all clients. Then the clients do the optimization locally and send back the corresponding results. Finally, the results are obtained by Eq. (1).

If each client does only one-step optimization, according to the property of conditional expectation, minimizing Eq. (1) is equivalent to minimizing Eq. (2). However, this procedure puts much pressure on communication, so researchers propose to do more local updates for efficiency. Hence,for client $C_k$, the optimization procedure of a typical round can be formalized as

$$\boldsymbol{w}_{i+1}^k \leftarrow \boldsymbol{w}_i^k - \eta \mathbb{E}_{\boldsymbol{\xi}|C_k}[\nabla_{\boldsymbol{w}} F(\boldsymbol{w}_i^k, \boldsymbol{\xi}|C_k)],\ i = 1, \cdots, n.$$

Then the server aggregates local models $\boldsymbol{w}_{n+1}^1, \cdots, \boldsymbol{w}_{n+1}^K$ to update the global model by

$$\overline{\boldsymbol{w}} = \sum\nolimits_{k=1}^{K} p_k \boldsymbol{w}_{n+1}^k, \tag{3}$$

where we denote $p_k$ for $P(C = C_k)$ for convenience.

### B. Wasserstein Distance

Wasserstein distance [33] is a metric in probabilistic space inspired by the problem of optimal transport. It is a distance between probability distributions that takes geometric information into account. The Wasserstein distance is defined as

$$\boldsymbol{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \inf_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}(\boldsymbol{\mu}, \boldsymbol{\nu})} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \boldsymbol{\gamma}}[\|\boldsymbol{x} - \boldsymbol{y}\|_p],$$

which is difficult to find a closed-form solution. However, if we choose 2-norm as the geometric measure and simplify the issue to Gaussian distribution, the distance becomes

$$d^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + tr\left(\left(\boldsymbol{\Sigma}_1^{\frac{1}{2}} - \boldsymbol{\Sigma}_2^{\frac{1}{2}}\right)^2\right), \tag{4}$$

where we define $d := \boldsymbol{W}_2(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))$. In Section IV, we use it to transform the issue of calculating GSNR to the issue of minimizing the distance between the global and local gradients.

## IV. Optimal Local Updates Leads to Maximal Channel Utilization

In this section, we establish a quantitative correlation between GSNR and the optimal local updates, which achieves the channel capacity for information communication. Then we derive a method to estimate the optimal local updates with initial information, which expedites FL convergence by maximizing the channel utilization. Finally, based on optimal local updates, we derive a method to calculate GSNR, which is a metric for local contributions.

### A. Gradient Signal to Noise Ratio

Regarding FL, we have a set of clients $\boldsymbol{C}$, and each client $C_k \in \boldsymbol{C}$ has a local data distribution $p(\boldsymbol{x}|C_k)$. Ideally, our target is gathering data from all clients to minimize Eq. (2), which means we optimize the model with the global gradient

$$\boldsymbol{g}_{\text{global}} = \mathbb{E}_{\boldsymbol{\xi}}[\nabla_{\boldsymbol{w}} F(\boldsymbol{w}, \boldsymbol{\xi})],$$

where $\boldsymbol{\xi} \sim p(\boldsymbol{x}) = \sum_{k=1}^{K} P(C = C_k) p(\boldsymbol{x}|C_k)$ is the sample drawn from global distribution. Due to limited resources and privacy concerns, we cannot gather data to minimize Eq. (2). In practice, the optimization processes are distributed to different clients. Specifically, client $C_k$ optimizes Eq. (2) with the local distribution $p(\boldsymbol{x}|C_k)$, which means $C_k$ utilizes local gradient for optimization, i.e.,

$$\boldsymbol{g}_l = \mathbb{E}_{\boldsymbol{\xi}|C_k}[\nabla_{\boldsymbol{w}} F(\boldsymbol{w}, \boldsymbol{\xi}|C_k)],\ \boldsymbol{\xi}|C_k \sim p(\boldsymbol{x}|C_k),$$

Usually, in non-iid scenario, $\boldsymbol{g}_l$ is different from $\boldsymbol{g}_{\text{global}}$ (as illustrated in Fig. 1), hence we can decompose $\boldsymbol{g}_l$ orthogonally according to $\boldsymbol{g}_{\text{global}}$, and define GSNR with model $\boldsymbol{w}$ as:

$$\begin{aligned} &r(\boldsymbol{w}) = \frac{\|\boldsymbol{g}_l^{\parallel}\|}{\|\boldsymbol{g}_l^{\perp}\|}, \\ &s.t.\ \boldsymbol{g}_l = \boldsymbol{g}_l^{\parallel} + \boldsymbol{g}_l^{\perp},\ \boldsymbol{g}_l^{\parallel} = \lambda \cdot \boldsymbol{g}_{\text{global}},\ \langle \boldsymbol{g}_l^{\parallel}, \boldsymbol{g}_l^{\perp} \rangle = 0. \end{aligned} \tag{5}$$

Note that $r(\boldsymbol{w})$ is decided by four elements: the model $\boldsymbol{w}$, the loss function $F(\ \cdot\ ,\ \cdot\ )$, the global distribution $p(\boldsymbol{x})$, and the local distribution $p(\boldsymbol{x}|C_k)$. Among them, $\boldsymbol{w}$ is variable during training, while others are pre-decided by various training techniques.

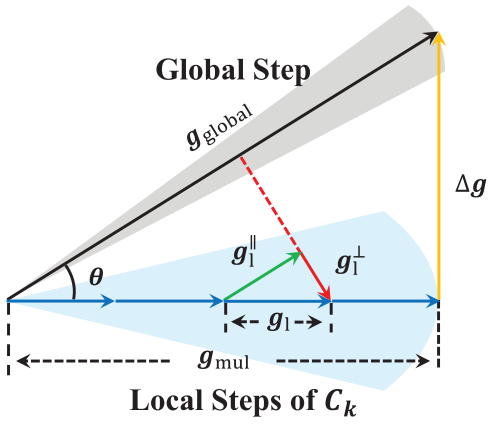



Fig. 2. The relation between GSNR and optimal local updates: when $C_k$ utilizes optimal local updates, $C_k$ reaches the channel capacity decided by GSNR.

Moreover, at a specific point in parameter space, $r(\boldsymbol{w})$, i.e., the GSNR, measures the channel's quality for information sharing in FL. Fig. 1 intuitively displays that if GSNR is large, then the directions of $\boldsymbol{g}_l$ and $\boldsymbol{g}_{\text{global}}$ are similar to each other. In this case, a large number of local updates will not lead to too much deviation from $\boldsymbol{g}_{\text{global}}$, which means the channel has better capability for information transmission. Moreover, if GSNR is small, the opposite is true.

*1) The Relation Between GSNR and Optimal Local Updates:* Intuitively, as illustrated in Fig. 2, angle $\theta$ can be viewed as the similarity between $\boldsymbol{g}_{\text{global}}$ and $\boldsymbol{g}_l$, and GSNR represents the $\cot\theta$. Moreover, when we utilize optimal local updates $\boldsymbol{g}_{\text{mul}}$, where $\boldsymbol{g}_{\text{mul}} = n \cdot \boldsymbol{g}_l$ and $n$ is the number of local updates, to approximate $\boldsymbol{g}_{\text{global}}$, which means the distance between $\boldsymbol{g}_{\text{mul}}$ and $\boldsymbol{g}_{\text{global}}$ reaches the minimal value $\Delta$, where $\Delta = \|\Delta \boldsymbol{g}\|$ and $\Delta \boldsymbol{g} = \boldsymbol{g}_{\text{global}} - \boldsymbol{g}_{\text{mul}}$, we have $\Delta \boldsymbol{g} \perp \boldsymbol{g}_{\text{mul}}$. Hence, we can get $\csc\theta = \|\boldsymbol{g}_{\text{global}}\|/\Delta$. Then according to trigonometric transformation, i.e., $\cot^2\theta = \csc^2\theta - 1$, we conclude that utilizing optimal local updates reaches the channel capacity, which maximizes the channel utilization. We will formalize these intuitions under certain assumptions.

### B. Maximize Channel Utilization With Optimal Local Updates

In practice, we utilize mini-batch SGD for optimization, which means $\boldsymbol{g}_{\text{global}}$ and $\boldsymbol{g}_l$ are random variables, hence their distributions are important factors for the analysis. Meanwhile, another important factor is the relationship between $\boldsymbol{g}_l$ and $\boldsymbol{g}_{\text{mul}}$. Here we prove that under bounded in-client variance and smoothness assumptions, the distributions converge to Gaussian and the relationship is linear.

To get the distributions of $\boldsymbol{g}_{\text{global}}$ and $\boldsymbol{g}_l$, we make the following assumption:

*Assumption 1 (Bounded in-Client Variance):* The variance of stochastic gradients are uniformly bounded, i.e., $\forall C_k \in \boldsymbol{C}$, $\forall \boldsymbol{w}$, $\mathbb{E}_{\boldsymbol{\xi}|C_k}\|\nabla_{\boldsymbol{w}} F(\boldsymbol{w}, \boldsymbol{\xi}|C_k) - \boldsymbol{\mu}_k\|^2 \leq \sigma^2$, where $\boldsymbol{\mu}_k := \mathbb{E}_{\boldsymbol{\xi}|C_k}[\nabla_{\boldsymbol{w}} F(\boldsymbol{w}, \boldsymbol{\xi}|C_k)]$.

With this assumption, we have Lemma 1, which implies the distributions of $\boldsymbol{g}_{\text{global}}$ and $\boldsymbol{g}_l$ for mini-batch stochastic gradient descent converges to Gaussian distributions.

*Lemma 1:* With Assumption 1, let $\{\boldsymbol{\xi}_{i,b} \mid 1 \leq i \leq n; 1 \leq b \leq B\}$ be a set of iid samples of a specific dataset, $\boldsymbol{g} = (\boldsymbol{g}_1, \cdots, \boldsymbol{g}_n)$ be a finite-dimensional gradient vector, where $\boldsymbol{g}_i = \frac{1}{B}\sum_{b=1}^{B} \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_i, \boldsymbol{\xi}_{i,b})$, $i \in \{1, \cdots, n\}$, then $\sqrt{B}(\boldsymbol{g} - \mathbb{E}[\boldsymbol{g}])$ converges to multivariate Gaussian distribution.

*Proof:* For any constant $n$, $\boldsymbol{g}$ can be rewritten as

$$\boldsymbol{g} = \frac{1}{B}\sum_{b=1}^{B}(\nabla_{\boldsymbol{w}} F(\boldsymbol{w}_1, \boldsymbol{\xi}_{1,b}), \cdots, \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_n, \boldsymbol{\xi}_{n,b})),$$

let $\tilde{\boldsymbol{g}}_b = (\nabla_{\boldsymbol{w}} F(\boldsymbol{w}_1, \boldsymbol{\xi}_{1,b})), \cdots, \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_n, \boldsymbol{\xi}_{n,b})))$, we have

$$\boldsymbol{g} = \frac{1}{B}\sum_{b=1}^{B} \tilde{\boldsymbol{g}}_b,$$

then with Assumption 1 and $n$ is a constant, $\tilde{\boldsymbol{g}}_b$ is subject to some complex distribution with bounded covariance matrix. As $\boldsymbol{\xi}_{i,b}$ is iid sampled from a specific dataset, $\boldsymbol{g}$ is the mean vector of $\tilde{\boldsymbol{g}}_1, \cdots, \tilde{\boldsymbol{g}}_B$, which are iid random vectors.

Therefore, based on the classical Central Limit Theory, with $B$ growing large, $\sqrt{B}(\boldsymbol{g} - \mathbb{E}[\boldsymbol{g}])$ converges to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ in distribution, where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{g}$.■

Moreover, Lemma 1 also implies that with mini-batch stochastic gradient descent, multiple local updates, i.e., $\bar{\boldsymbol{g}} = \boldsymbol{w}_{n+1} - \boldsymbol{w}_1 = \sum_{i=1}^{n} \boldsymbol{g}_i = \mathbb{1}^T \boldsymbol{g}$, converges to a Gaussian distribution since $\bar{\boldsymbol{g}}$ is a linear transformation of a joint Gaussian vector, which can be used to describe $\boldsymbol{g}_{\text{mul}}$.

Moreover, to analyze the relationship between $\boldsymbol{g}_l$ (the initial local gradient for each round) and $\boldsymbol{g}_{\text{mul}}$, we have Assumption 2 of smoothness.

*Assumption 2 (Smoothness):* The target function $F(\boldsymbol{w}, \cdot)$: $R^m \to R$ is twice differentiable, and the expected matrix norm of hessian matrix $H(F(\boldsymbol{w}, \cdot))$ is bounded, i.e., $\mathbb{E}_{\boldsymbol{\xi}}\|H(F(\boldsymbol{w}, \boldsymbol{\xi}))\|^2 \leq L^2$, where $\boldsymbol{\xi}$ is randomly sampled from a specific dataset.

Note that Assumption 2 is weaker than L-smooth Assumption, since if a function $F(\boldsymbol{w}, \cdot)$ is L-smooth, it conforms to Assumption 2, but not vice versa. Moreover, Assumption 2 always holds for typical machine learning tasks, e.g., logistic regression, soft-max classification, and so on. With these assumptions, we have the following lemma.

*Lemma 2:* If Assumption 1 and Assumption 2 hold, let $\{\eta_r\}_{r=1}^{+\infty}$ be a sequence of real number such that $\lim_{r\to+\infty} \eta_r = 0$, and $\{\boldsymbol{\varepsilon}_r\}_{r=1}^{+\infty}$ be a sequence of random vectors, where $r$ denotes for the communication round index, $\boldsymbol{\varepsilon}_r = \hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}$, $\hat{\boldsymbol{g}} = \sum_{i=1}^{n} \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_1, \boldsymbol{\xi})$, and $\bar{\boldsymbol{g}} = \sum_{i=1}^{n} \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_i, \boldsymbol{\xi})$ with $\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \eta_r \boldsymbol{g}_{i-1}, i \in \{2, \cdots, n\}$, then we have $\boldsymbol{\varepsilon}_r \xrightarrow{L} 0$, which implies $\hat{\boldsymbol{g}} \xrightarrow{L} \bar{\boldsymbol{g}}$.

*Proof:* First, we prove $\lim_{r\to+\infty} \mathbb{E}\|\boldsymbol{\varepsilon}_r\| = 0$.

Regarding the gradient $\boldsymbol{g}_i$, $i \in \{1, \cdots, n\}$, due to the smoothness of $\nabla_{\boldsymbol{w}} F(\boldsymbol{w}_i, \boldsymbol{\xi})$, we can expand $\boldsymbol{g}_i$ based on Lagrange's mean value theorem as

$$\begin{aligned}\boldsymbol{g}_i &= \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_i, \boldsymbol{\xi}) = \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_1, \boldsymbol{\xi}) + H(F(\tilde{\boldsymbol{w}}_i, \boldsymbol{\xi}))(\boldsymbol{w}_i - \boldsymbol{w}_1) \\ &= \boldsymbol{g}_1 + H(F(\tilde{\boldsymbol{w}}_i, \boldsymbol{\xi}))(\boldsymbol{w}_i - \boldsymbol{w}_1), \end{aligned} \tag{6}$$

where $\tilde{\boldsymbol{w}} := \lambda \boldsymbol{w}_i + (1-\lambda)\boldsymbol{w}_1$, $\lambda \in [0, 1]$. Then

$$\begin{aligned}\mathbb{E}\|\boldsymbol{g}_i - \boldsymbol{g}_1\| &= \mathbb{E}\|H(F(\tilde{\boldsymbol{w}}_i, \boldsymbol{\xi}))(\boldsymbol{w}_i - \boldsymbol{w}_1)\| \\ &\overset{(a)}{\leq} \mathbb{E}\|H(F(\tilde{\boldsymbol{w}}_i, \boldsymbol{\xi}))\|\|(\boldsymbol{w}_i - \boldsymbol{w}_1)\| \\ &\overset{(b)}{\leq} \sqrt{\mathbb{E}\|H(F(\tilde{\boldsymbol{w}}_i, \boldsymbol{\xi}))\|^2 \mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{w}_1\|^2} \\ &\overset{(c)}{\leq} L \cdot \sqrt{\mathbb{E}\|\eta_r \sum_{j=1}^{i-1} \boldsymbol{g}_j\|^2}\end{aligned}$$

$$\overset{(d)}{\leq} L^2 \cdot \eta_r \sqrt{(i-1)\sum_{j=1}^{i-1}\mathbb{E}\|\boldsymbol{g}_j\|^2} \overset{(e)}{\leq} (i-1)\eta_r LG, \tag{7}$$

where $(a)$ follows from sub-multiplicative property of matrix norm, $(b)$ is based on Cauchy-Schwarz inequality, $(c)$ is an immediate consequence of Assumption 2 and the local optimization process, $(d)$ comes from the fact $\|\sum_{i=1}^n \boldsymbol{a}_i\|^2 \leq n\sum_{i=1}^n \|\boldsymbol{a}_i\|^2$, and $(e)$ is based on Assumption 1, where $G^2 := \sigma^2 + \mu^2$, $\mu = \max(\{\|\boldsymbol{\mu}_i\|\}_{i\in\{1,\cdots,n\}})$. Hence,

$$\mathbb{E}\|\boldsymbol{\varepsilon}_r\| = \mathbb{E}\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}\| = \mathbb{E}\|\sum_{i=1}^n(\boldsymbol{g}_i - \boldsymbol{g}_1)\| \leq \sum_1^n \mathbb{E}\|\boldsymbol{g}_i - \boldsymbol{g}_1\|$$
$$\leq \left(\sum_{i=1}^n (i-1)\right)\eta_r LG = \frac{n(n-1)}{2}\eta_r LG$$

where the first inequality follows from the triangle inequality, and the second inequality is based on Eq. (7).

As $n$ represents the number of local steps, which is a constant, $\mathbb{E}\|\boldsymbol{\varepsilon}_r\|$ is upper bounded by $\eta_r \cdot M$, where $M$ is a bounded value. Therefore,

$$0 \leq \lim_{r\to+\infty}\mathbb{E}\|\boldsymbol{\varepsilon}_r\| \leq \lim_{r\to+\infty}\eta_r \cdot M = 0. \tag{8}$$

Eq. (8) implies $\boldsymbol{\varepsilon}_r$ converge to 0 in mean, i.e., $\boldsymbol{\varepsilon}_r \xrightarrow{L} 0$, which immediately completes the proof.■

*Remark 1:* Based on the proof of Lemma 2, the estimation error is $\mathbb{E}\|\bar{\boldsymbol{g}} - \hat{\boldsymbol{g}}\| \leq n(n-1)\eta_r LG$, which implies that if we consider learning rate decay,[1] then

$$\forall\epsilon,\ \lim_{r\to+\infty} Pr(\|\bar{\boldsymbol{g}} - \hat{\boldsymbol{g}}\| > \epsilon) = 0. \tag{9}$$

Moreover, in a typical communication round $r$, the optimization process implies that $\boldsymbol{w}_{n+1} - \boldsymbol{w}_1 = \eta_r\bar{\boldsymbol{g}}$. While based on Eq. (9), we can use $\eta_r\hat{\boldsymbol{g}} = \eta_r\sum_{i=1}^n \nabla_{\boldsymbol{w}}F(\boldsymbol{w}_1, \boldsymbol{\xi})$ to approximate $\eta_r\bar{\boldsymbol{g}}$. With multiplying Eq. (9) by $\eta_r$, the estimation error becomes $\mathcal{O}((n\eta_r)^2 LG)$.

Particularly, in local optimization, we have $\boldsymbol{g}_l = \boldsymbol{g}_1 = \nabla_{\boldsymbol{w}}F(\boldsymbol{w}_1, \boldsymbol{\xi})$ and $\boldsymbol{g}_{\text{mul}} = \bar{\boldsymbol{g}}$. Then based on Lemma 2, we can estimate $\boldsymbol{g}_{\text{mul}}$ by $n \cdot \boldsymbol{g}_l$. Hence, if we denote the mean values and the covariance matrices of $\boldsymbol{g}_l$ and $\boldsymbol{g}_{\text{mul}}$ as $\boldsymbol{\mu}_{\text{mul}}$, $\boldsymbol{\mu}_l$, $\boldsymbol{\Sigma}_{\text{mul}}$, and $\boldsymbol{\Sigma}_l$, respectively, we have

$$\boldsymbol{\mu}_{\text{mul}} = n \cdot \boldsymbol{\mu}_l,\ \boldsymbol{\Sigma}_{\text{mul}} = n^2 \cdot \boldsymbol{\Sigma}_l.$$

For convenience, in the rest of our work, we use $\boldsymbol{\mu}_*$ and $\boldsymbol{\Sigma}_*$ to denote the corresponding mean vector and covariance matrix of gradients estimated by a specific dataset. Moreover, if we utilize mini-batch SGD, the covariance matrix is scaled by the batch size $B$. Then, with the initial parameters, the distribution of $n$ updates can be estimated by $\mathcal{N}\left(n\eta_r\boldsymbol{\mu}_*, n^2\eta_r^2\frac{\boldsymbol{\Sigma}_*}{B}\right)$.

According to Lemma 1 and Lemma 2, we can approximate distributions of $\eta_r\boldsymbol{g}_{\text{global}}$ and $\eta_r\boldsymbol{g}_{\text{mul}}$ with $n$ local steps by $\mathcal{N}\left(\eta_r\boldsymbol{\mu}_g, \eta_r^2\frac{\boldsymbol{\Sigma}_g}{B}\right)$ and $\mathcal{N}\left(n\eta_r\boldsymbol{\mu}_l, n^2\eta_r^2\frac{\boldsymbol{\Sigma}_l}{B}\right)$ respectively. Then the optimal number of local updates to reach channel capacity is implied by the following theorem.

*Theorem 1:* The minimal Wasserstein distance between two multivariate Gaussian distributions denoted by $\mathcal{N}\left(\eta_r\boldsymbol{\mu}_g, \eta_r^2\frac{\boldsymbol{\Sigma}_g}{B}\right)$ and $\mathcal{N}\left(n\eta_r\boldsymbol{\mu}_l, n^2\eta_r^2\frac{\boldsymbol{\Sigma}_l}{B}\right)$ with variable $n$ is achieved when $n$ is

$$n_1^{opt} = \max\left(0,\ \frac{\boldsymbol{\mu}_l^T\boldsymbol{\mu}_g + \frac{tr\left((\boldsymbol{\Sigma}_l\boldsymbol{\Sigma}_g)^{\frac{1}{2}}\right)}{B}}{\|\boldsymbol{\mu}_l\|^2 + \frac{tr(\boldsymbol{\Sigma}_l)}{B}}\right),$$

and the minimum distance is $(\Delta_1^{opt})^2 = \eta_r^2\Delta^2$,

$$\Delta^2 = \|\boldsymbol{\mu}_g\|^2 + \frac{tr(\boldsymbol{\Sigma}_g)}{B} - \frac{\left(\boldsymbol{\mu}_l^T\boldsymbol{\mu}_g + \frac{tr\left((\boldsymbol{\Sigma}_l\boldsymbol{\Sigma}_g)^{\frac{1}{2}}\right)}{B}\right)^2}{\|\boldsymbol{\mu}_l\|^2 + \frac{tr(\boldsymbol{\Sigma}_l)}{B}}.$$

*Proof:* According to Eq. (4), to minimize the distance between $\mathcal{N}\left(\eta_r\boldsymbol{\mu}_g, \eta_r^2\frac{\boldsymbol{\Sigma}_g}{B}\right)$ and $\mathcal{N}\left(n\eta_r\boldsymbol{\mu}_l, n^2\eta_r^2\frac{\boldsymbol{\Sigma}_l}{B}\right)$, we can build an optimization problem as

$$\begin{aligned}\min_n\quad & d^2 = \|\eta_r\boldsymbol{\mu}_g - n\eta_r\boldsymbol{\mu}_l\|^2 + tr(\boldsymbol{M}^2)\\ \text{s.t.}\quad & \boldsymbol{M} = \left(\frac{\eta_r^2\boldsymbol{\Sigma}_g}{B}\right)^{\frac{1}{2}} - \left(\frac{n^2\eta_r^2\boldsymbol{\Sigma}_l}{B}\right)^{\frac{1}{2}}\\ & n \geq 0.\end{aligned} \tag{10}$$

Note that Eq. (10) is a quadratic function of $n$, which immediately completes the proof.■

*Corollary 1:* For two distributions $\mathcal{N}(m\eta_r\boldsymbol{\mu}_g, m^2\eta_r^2\boldsymbol{\Sigma}_g)$ and $\mathcal{N}(n\eta_r\boldsymbol{\mu}_l, n^2\eta_r^2\boldsymbol{\Sigma}_l)$, where $m$ is a constant, the optimal $n$ to minimize the Wasserstein distance is $n_m^{opt} = m \cdot n_1^{opt}$ and the minimal distance is $(\Delta_m^{opt})^2 = m^2 \cdot (\Delta_1^{opt})^2$.

*Proof:* In this case, we change the distribution $\mathcal{N}\left(\eta_r\boldsymbol{\mu}_g, \eta_r^2\frac{\boldsymbol{\Sigma}_g}{B}\right)$ to $\mathcal{N}\left(m\eta_r\boldsymbol{\mu}_g, m\eta_r^2\frac{\boldsymbol{\Sigma}_g}{B}\right)$, and reformulate problem (10) as

$$\begin{aligned}\min_n\quad & d^2 = m^2\left(\|\eta_r\boldsymbol{\mu}_g - \frac{n}{m}\eta_r\boldsymbol{\mu}_l\|^2 + tr(\boldsymbol{M}^2)\right)\\ \text{s.t.}\quad & \boldsymbol{M} = \left(\frac{\eta_r^2\boldsymbol{\Sigma}_g}{B}\right)^{\frac{1}{2}} - \left(\frac{\left(\frac{n}{m}\right)^2\eta_r^2\boldsymbol{\Sigma}_l}{B}\right)^{\frac{1}{2}}\\ & n \geq 0,\end{aligned}$$

let $\tilde{d} = \frac{d}{m}$ and $\tilde{n} = \frac{n}{m}$, then the new problem reduces to problem (10), which concludes the proof immediately.■

Based on Corollary 1, if $m$ is a constant, the minimum Wasserstein distance is achieved when $n = m * n_1^{opt}$, which is the optimal number of local updates for maximizing channel utilization, leading to the maximal capability for information communication.

### C. Contribution Evaluation With GSNR

As aforementioned, GSNR decides the channel capacity of information communication, thus it can be used to evaluate the contribution of different clients. Here we derive a method to calculate GSNR.

First, for convenience, we define a matrix as follows:

$$\boldsymbol{R}_* = \begin{pmatrix} u_*^1 & & & \\ & u_*^2 & & \\ & & \ddots & \\ & & & u_*^d \\ \hline & & & \left(\frac{1}{B}\boldsymbol{\Sigma}_*\right)^{\frac{1}{2}} \end{pmatrix}$$

[1]For example, $\eta_r = \eta_0\alpha^r$ refers to a widely used learning rate decay method with a decay rate $\alpha < 1$.

where $\mu_*^i$ is the component of $\boldsymbol{\mu}_* = (\mu_*^1, \cdots, \mu_*^d)$. Then we have $\|\boldsymbol{g}_{\text{global}}\| = \|\boldsymbol{R}_g\|_F^2$,[2] and $\|\boldsymbol{g}_l\| = \|\boldsymbol{R}_l\|_F^2$.

From Theorem 1, we obtain the optimal distance between $\boldsymbol{g}_{\text{global}}$ and $\boldsymbol{g}_{\text{mul}}$, i.e., $\Delta$, hence we can get the $\csc\theta = \|\boldsymbol{g}_{\text{global}}\|/\Delta$. Moreover, as mentioned in Section IV-A, $r(\boldsymbol{w})$ is the $\cot\theta$, which can be calculated by trigonometric transformation, i.e., $\cot^2\theta = \csc^2\theta - 1$, then we can derive Proposition 1.

*Proposition 1* (Gradient Signal to Noise Ratio (GSNR)): For a local dataset $\boldsymbol{D}_l$ and a global dataset $\boldsymbol{D}_g$, with a loss function $F(\boldsymbol{w}, \cdot)$, the GSNR is a function of $\boldsymbol{w}$ as

$$r(\boldsymbol{w}) = \max\left(0, \frac{\langle \boldsymbol{R}_l, \boldsymbol{R}_g\rangle_F}{\sqrt{\|\boldsymbol{R}_l\|_F^2\|\boldsymbol{R}_g\|_F^2 - \langle \boldsymbol{R}_l, \boldsymbol{R}_g\rangle_F^2}}\right).$$

Proposition 1 can be used to calculate GSNR directly from local datasets.

**Algorithm 1** FedGSNR in Conjunction With FedAvg

**Input:** initial model $\boldsymbol{w}_1$, learning rate $\eta_0$, sample size $B$, and chosen global steps $E_{const}$
**for** $r = 1$ **to** $R$ **do**
Sample clients $\boldsymbol{S} \subseteq \boldsymbol{C}$
**Server**: send $\boldsymbol{w}_1$ and $\eta_r$ to each client $C_k \in \boldsymbol{S}$
**On each active client** $C_k$ **in parallel**: initialize local model $\boldsymbol{w}^k \leftarrow \boldsymbol{w}_1$, compute $\tilde{\boldsymbol{g}}^k$ and $diag\left(\tilde{\boldsymbol{\Sigma}}^k\right)$, and send them to the server
**Server**: compute $n_{1,k}^{opt}$ according to Theorem 1 for each client $C_k$, and send it to client $C_k$
**On each active client** $C_k$ **in parallel**:
**for** $t = 1$ **to** $E_{const} \cdot n_{1,k}^{opt}$ **do**
$\boldsymbol{w}_t \leftarrow \boldsymbol{w}_t - \eta_r \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_t, \boldsymbol{\xi}|C_k)$
**end for**
**Server**: $\boldsymbol{w}_1 \leftarrow \sum_{k=1}^{|\tilde{\boldsymbol{S}}|} \tilde{p}_k \boldsymbol{w}^k$, where $\tilde{\boldsymbol{S}} = \left\{C_k \mid C_k \in S, n_{1,k}^{\text{opt}} > 0\right\}$, and $\tilde{p}_k$ is the corresponding probability ratio, i.e., $p_k / \sum_{s\in\tilde{S}} p_s$
**end for**

## V. FedGSNR: Proposed Algorithm for Maximizing Channel Utilization

Theorem 1 describes the method to calculate the optimal local updates, and the mean vectors and the covariance matrices for both local distribution and global distribution are the key parameters. In practice, we can use sample mean vector and sample covariance matrix to estimate the parameters of local distribution. Specifically, for client $C_k$ with model $\boldsymbol{w}_1$, the corresponding statistics are $\tilde{\boldsymbol{g}}_k = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{g}_{k,b}$ and $\tilde{\boldsymbol{\Sigma}}_k = \frac{1}{B}\sum_{b=1}^{B}(\boldsymbol{g}_{k,b} - \tilde{\boldsymbol{g}}_k)(\boldsymbol{g}_{k,b} - \tilde{\boldsymbol{g}}_k)^T$, where $\boldsymbol{g}_{b,k} = \nabla_{\boldsymbol{w}} F(\boldsymbol{w}_1, \boldsymbol{\xi}_b|C_k)$. For the server, based on the theory of conditional expectation, the corresponding global statistics are

$$\tilde{\boldsymbol{g}} = \mathbb{E}_C[\tilde{\boldsymbol{g}}|C] = \sum_{k=1}^{K} p_k \tilde{\boldsymbol{g}}_k, \tag{11}$$

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}} &= \mathbb{E}_C[\tilde{\boldsymbol{\Sigma}}|C] + Cov_C(\tilde{\boldsymbol{g}}|C) \\ &= \sum_{k=1}^{K} p_k \tilde{\boldsymbol{\Sigma}}_k + \sum_{k=1}^{K} p_k[(\tilde{\boldsymbol{g}}_k - \tilde{\boldsymbol{g}})(\tilde{\boldsymbol{g}}_k - \tilde{\boldsymbol{g}})^T].\end{aligned} \tag{12}$$

[2] $\|\cdot\|_F$ and $\langle\cdot,\cdot\rangle_F$ are Frobenius inner product and Frobenius norm respectively.

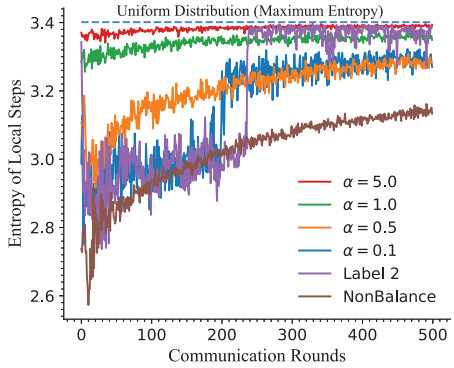



Fig. 3. The entropy of optimal local updates for different partition methods. Note that the larger entropy means the channel capacities of different clients are similar to each other.

In practice, as the covariance matrix increases the communication traffic and the calculation of the matrix introduces lots of computation, we use trace to simplify the procedure. Specifically, in Theorem 1, we mainly use the trace of the covariance matrix. Meanwhile, according to [3], the covariance matrix of gradient is a sparse matrix and the estimated error can be scaled down by the batch size $B$. Therefore, we instead utilize the principal diagonal element of $\tilde{\boldsymbol{\Sigma}}_k$ for efficiency. Based on former analysis, we propose an algorithm FedGSNR to calculate the optimal number of local updates, and Algorithm 1 is a typical example of FedGSNR in conjunction with FedAvg.[3]

### A. Fairness Analysis

An important concern of FedGSNR is the fairness of different clients. For this purpose, we experimentally analyze the entropy of optimal local updates, i.e., $H = \sum_{k=1}^{K} p(C_k)\log p(C_k)$, where $p(C_k) = \frac{n_{1,k}^{opt}}{\sum_{i=1}^{K} n_{1,i}^{opt}}$, and the results are illustrated in Fig. 3 (the blue dashed line on the top represents the uniform distribution, i.e., identical local updates, which is the maximum entropy distribution). In these experiments, we distributed the data according to different techniques, and the details of distributing methods can be found in Section VIII. The key observations of Fig. 3 include two aspects. First, the entropy becomes smaller when the degree of non-iid is increased, which means with the increased degree of non-iid, the channel capacity varies dramatically according to different clients, indicating that the optimal local updates becomes more important. Second, for all distribution methods, the entropy is an increasing function according to communication rounds, which means FedGSNR naturally seeks the fair path (the entropy gets closer to uniform distribution) for optimization.

## VI. Convergence Analysis

In this section, we aim to analyze the convergence of FedGSNR and demonstrate that it is a convergent algorithm. Table I lists major notations used in this section, and we will describe the remaining notations when they are utilized.

[3] Note that our proposed FedGSNR is a compatible method, and the referred FedAvg can also be replaced by other methods (e.g., FedProx, Scaffold, etc.).

TABLE I
MAJOR NOTATION EXPLANATION

| Notations | Explanation |
|---|---|
| $t$, $k$ | The iteration number and the client number, respectively |
| $F_k(\cdot)$, $F(\cdot)$ | Loss function with sample $\boldsymbol{\xi}\|C_k$ and $\boldsymbol{\xi}$, respectively |
| $\boldsymbol{w}^{k*}$, $\boldsymbol{w}^*$ | The optimum of $F_k(\cdot)$ and $F(\cdot)$ |
| $\boldsymbol{w}_t^k$ | The model at time $t$ on client $k$ |
| $\bar{\boldsymbol{w}}_t$ | The average model at time $t$, i.e. $\bar{\boldsymbol{w}}_t = \sum_{k=1}^K p_k \boldsymbol{w}_t^k$ |
| $\sigma_{opt}^2$ | The variance of the stochastic gradients at the optimum |
| $B$ | Batch size for optimization |
| $D_F(\boldsymbol{w}, \boldsymbol{v})$ | The Bregman divergence between $\boldsymbol{w}$ and $\boldsymbol{v}$ (Lemma 5) |
| $V_t$ | $V_t = \sum_{k=1}^K p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2$, i.e., the model divergence at time $t$ |
| $R_{g,t}$ | The matrix $R_*$ with $\boldsymbol{\xi}$ at time $t$ |
| $R_{l,t}^k$ | The matrix $R_*$ with $\boldsymbol{\xi}\|C_k$ at time $t$ on client $k$ |
| $E_{const}$ | The hyper-parameter for calculating local steps |
| $n_{1,k,r}^{opt}$ | The optimal $n$ at round $r$ on client $k$ (Theorem 1) |
| $E_r^k$ | The optimal local steps at round $r$ on client $k$ |
| $\|\zeta_t\|^2$ | $\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^*\|^2$, i.e., the deviation of the average model from the optimum at time $t$ |

As defined in Section IV-C, we have

$$\boldsymbol{R}_* = \begin{pmatrix} u_*^1 & & & \\ & u_*^2 & & \\ & & \ddots & \\ & & & u_*^d \\ \hdashline & & & \left(\frac{1}{B}\boldsymbol{\Sigma}_*\right)^{\frac{1}{2}} \end{pmatrix} \tag{13}$$

where $\boldsymbol{\mu}_*$ and $\boldsymbol{\Sigma}_*$ are the corresponding mean vector and covariance matrix calculated by the samples sampled from different datasets, $\mu_*^i$ is the component of $\boldsymbol{\mu}_* = (\mu_*^1, \cdots, \mu_*^d)$. Then we have the following lemmas.

*Lemma 3:* $\|\boldsymbol{R}_*\|^2 \geq \frac{1}{B}\mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{g}_*\|^2]$, where $B \geq 1$.

*Proof:* First, we have

$$\boldsymbol{\mu}_*^T \boldsymbol{\mu}_* = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*]^T \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*] = tr(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*]\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*]^T),$$

Note that $B \geq 1$, hence

$$\begin{aligned} \|\boldsymbol{R}_*\|^2 &= \boldsymbol{\mu}_*^T \boldsymbol{\mu}_* + \frac{tr(\boldsymbol{\Sigma}_*)}{B} \geq \frac{\boldsymbol{\mu}_*^T \boldsymbol{\mu}_* + tr(\boldsymbol{\Sigma}_*)}{B} \\ &= \frac{tr(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*]\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*]^T + \boldsymbol{\Sigma}_*)}{B} \\ &= \frac{\mathbb{E}_{\boldsymbol{\xi}}[tr(\boldsymbol{g}_*\boldsymbol{g}_*^T)]}{B} = \frac{1}{B}\mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{g}_*\|^2], \end{aligned}$$

which concludes the proof.■

*Lemma 4:* $\|\boldsymbol{R}_*\|^2 \leq \mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{g}_*\|^2]$, where $B \geq 1$.

*Proof:* Similarly,

$$\begin{aligned} \|\boldsymbol{R}_*\|^2 &= \boldsymbol{\mu}_*^T \boldsymbol{\mu}_* + \frac{tr(\boldsymbol{\Sigma}_*)}{B} \leq \boldsymbol{\mu}_*^T \boldsymbol{\mu}_* + tr(\boldsymbol{\Sigma}_*) \\ &= tr(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*]\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_*]^T + \boldsymbol{\Sigma}_*) \\ &= \mathbb{E}_{\boldsymbol{\xi}}[tr(\boldsymbol{g}_*\boldsymbol{g}_*^T)] = \mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{g}_*\|^2], \end{aligned}$$

which concludes the proof.■

With these lemmas, we can analyze the convergence of FedGSNR with FedAvg according to existing technique [17]. For convenience, we denote $F(\boldsymbol{w}; \boldsymbol{\xi}|C_k)$ and $F(\boldsymbol{w}; \boldsymbol{\xi})$ as $F_k(\boldsymbol{w})$ and $F(\boldsymbol{w})$, respectively. It is worth noting that $F(\boldsymbol{w}) = \sum_{i=1}^K p_k F_k(\boldsymbol{w})$. Then we make additional assumptions.

*Assumption 3:* The functions $F_k(\cdot)$, $k \in \{1, \cdots, K\}$, are all L-smooth: for all $\boldsymbol{w}$ and $\boldsymbol{v}$, $F_k(\boldsymbol{v}) \leq F_k(\boldsymbol{w}) + \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_k(\boldsymbol{w})\rangle + \frac{L}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2$.

*Assumption 4:* The functions $F_k(\cdot)$, $k \in \{1, \cdots, K\}$, are all $\gamma$-strongly convex: for all $\boldsymbol{w}$ and $\boldsymbol{v}$, $F_k(\boldsymbol{v}) \geq F_k(\boldsymbol{w}) + \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F_k(\boldsymbol{w})\rangle + \frac{\gamma}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2$.

*Assumption 5:* The local optimization variance at the optimum is bounded: $\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^K p_k \|\nabla F_k(\boldsymbol{w}^*)\|^2\right] \leq \sigma_{opt}^2$, where $\boldsymbol{w}^* := \min_{\boldsymbol{w}} F(\boldsymbol{w})$.

As $\boldsymbol{w}^*$ is a fixed point for $F(\cdot)$, Assumption 5 always holds with finite $\sigma_{opt}^2$ for all non-degenerate sampling distributions, which is a much more meaningful quantity for the convergence analysis [17]. Then we have the following lemmas.

*Lemma 5:* If Assumption 3 holds, we have

$$\mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\boldsymbol{w}) - \nabla F_k(\boldsymbol{v})\|^2 \leq 2LD_{F_k}(\boldsymbol{w}, \boldsymbol{v}),$$

where $D_{F_k}(\boldsymbol{w}, \boldsymbol{v}) := \mathbb{E}_{\boldsymbol{\xi}}[F_k(\boldsymbol{w}) - F_k(\boldsymbol{v}) - \langle \boldsymbol{w} - \boldsymbol{v}, \nabla F_k(\boldsymbol{v})\rangle]$ is the Bregman divergence associated with function $F_k(\cdot)$ and arbitrary $\boldsymbol{w}$, $\boldsymbol{v}$.

Lemma 5 is an immediate consequence of Proposition 2 in [17], which can be proved by the L-smoothness of $F_k(\cdot)$.

*Lemma 6:* Suppose Assumption 3, 4, and 5 hold, we have

$$\mathbb{E}_{\boldsymbol{\xi}}\|\sum_{k=1}^K p_k \nabla F_k(\boldsymbol{w}_t^k)\|^2 \leq 2L^2 V_t + 8LD_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) + 4\sigma_{opt}^2,$$

where we denote $V_t = \sum_{k=1}^K p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2$.

*Proof:* With the fact that $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}}\|\sum_{k=1}^K p_k \nabla F_k(\boldsymbol{w}_t^k)\|^2 \\ &\leq \underbrace{2\mathbb{E}_{\boldsymbol{\xi}}\|\sum_{k=1}^K p_k \nabla F_k(\boldsymbol{w}_t^k) - \sum_{k=1}^K p_k \nabla F_k(\bar{\boldsymbol{w}}_t)\|^2}_{A_1} \\ &\quad + \underbrace{2\mathbb{E}_{\boldsymbol{\xi}}\|\sum_{k=1}^K p_k \nabla F_k(\bar{\boldsymbol{w}}_t)\|^2}_{A_2}. \end{aligned} \tag{14}$$

For the first term $A_1$, we have

$$\begin{aligned} A_1 &\overset{(a)}{\leq} 2\sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\boldsymbol{w}_t^k) - \nabla F_k(\bar{\boldsymbol{w}}_t)\|^2 \\ &\overset{(b)}{\leq} 2L^2 \sum_{k=1}^K p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2, \end{aligned}$$

where inequality $(a)$ depends on the Jensen's inequality and inequality $(b)$ is a consequence of L-Smoothness of $F_k(\cdot)$.

For the second term $A_2$, we have

$$\begin{aligned} A_2 &= 2\mathbb{E}_{\boldsymbol{\xi}}\|\nabla F(\bar{\boldsymbol{w}}_t)\|^2 \\ &\overset{(a)}{\leq} 4\mathbb{E}_{\boldsymbol{\xi}}\|\nabla F(\bar{\boldsymbol{w}}_t) - \nabla F(\boldsymbol{w}^*)\|^2 + 4\mathbb{E}_{\boldsymbol{\xi}}\|\nabla F(\boldsymbol{w}^*)\|^2 \\ &\overset{(b)}{\leq} 8LD_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) + 4\mathbb{E}_{\boldsymbol{\xi}}\|\sum_{k=1}^K p_k \nabla F_k(\boldsymbol{w}^*)\|^2 \\ &\overset{(c)}{\leq} 8LD_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) + 4\sigma_{opt}^2, \end{aligned}$$

where inequality $(a)$ depends on the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, inequality $(b)$ is based on Lemma 5, and inequality $(c)$ is the consequence of Jensen's inequality and Assumption 5.

Using $A_1$ and $A_2$ in Eq. (14) concludes the proof.■

*Lemma 7:* If Assumption 3 and 4 hold, we have

$$-2\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^{K} p_k \langle \bar{\boldsymbol{w}}_t - \boldsymbol{w}^*, \nabla F_k(\boldsymbol{w}_t^k)\rangle\right] \leq -2D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) - \gamma\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^*\|^2 + L\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2,$$

*Proof:* We split the left-hand side as

$$-2\langle \bar{\boldsymbol{w}}_t - \boldsymbol{w}^*, \nabla F_k(\boldsymbol{w}_t^k)\rangle = A_1 + A_2. \tag{15}$$

where $A_1 = -2\langle \bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k, \nabla F_k(\boldsymbol{w}_t^k)\rangle$ and $A_2 = -2\langle \boldsymbol{w}_t^k - \boldsymbol{w}^*, \nabla F_k(\boldsymbol{w}_t^k)\rangle$. Using the $\gamma$-strongly convexity and L-smoothness, we have

$$A_1 \leq 2\left(F_k(\boldsymbol{w}^*) - F_k(\boldsymbol{w}_t^k) - \frac{\gamma}{2}\|\boldsymbol{w}_t^k - \boldsymbol{w}^*\|^2\right)$$
$$A_2 \leq 2\left(F_k(\boldsymbol{w}_t^k) - F_k(\bar{\boldsymbol{w}}_t) + \frac{L}{2}\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2\right).$$

Averaging Eq. (15) over $k$, we have

$$\begin{aligned}
&-2\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^{K} p_k \langle \bar{\boldsymbol{w}}_t - \boldsymbol{w}^*, \nabla F_k(\boldsymbol{w}_t^k)\rangle\right] \\
&\leq 2\mathbb{E}_{\boldsymbol{\xi}}\sum_{k=1}^{K} p_k [F_k(\boldsymbol{w}^*) - F_k(\bar{\boldsymbol{w}}_t)] - \gamma\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \boldsymbol{w}^*\|^2 \\
&\quad + L\sum_{k=1}^{K} p_k \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2 \\
&\overset{(a)}{\leq} -2D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) - \gamma\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^*\|^2 + LV_t,
\end{aligned}$$

where $V_t = \sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2$ and inequality (a) depends on the definition of $D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*)$ and the convexity of $\|\cdot\|^2$, which concludes the proof.■

*Lemma 8 (Bounding the Divergence of $\boldsymbol{w}_t^k$ in FedGSNR):* Suppose Assumption 3, 4, and 5 hold, if $\eta \leq \frac{1}{4E_{const}L}\sqrt{\frac{\gamma}{2BL}}$, we have

$$\mathbb{E}_{\boldsymbol{\xi}}[V_t] \leq \frac{16\eta^2 E_{const}^2 BL^2}{\gamma} D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) + \frac{8\eta^2 E_{const}^2 BL}{\gamma}\sigma_{opt}^2,$$

where $V_t = \sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2$.

*Proof:* Based on the strategy of FedGSNR, the number of local steps is individually decided by $E_r^k = n_{1,k,r}^{opt} * E_{const}$, where $E_{const}$ is a constant. Hence we define the local optimization process as

$$\boldsymbol{w}_{t+1}^k = \begin{cases} \boldsymbol{w}_t^k - \eta\nabla_{\boldsymbol{w}} F(\boldsymbol{w}_t^k), & 0 \leq t - t_0 < E_r^k \\ \boldsymbol{w}_t^k, & E_r^k \leq t - t_0 < E_{max}, \end{cases} \tag{16}$$

where $\boldsymbol{w}_{t_0}^k = \bar{\boldsymbol{w}}_{t_0}$ represents the aggregation step, and without loss of generality, $t_0$ is the initial time of communication round $r$, $E_{max} = \max\{E_r^k \,|\, 1 \leq k \leq K, 1 \leq r \leq R\}$. Then the situation becomes FL with identical local steps, and we will prove that the divergence is independent of $E_{max}$. We use the fact that $t - t_0 \leq E_{max}$, where $t_0$ represents the latest aggregation step before $t$, $\eta$ is the learning rate. Then we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^{K} p_k \|(\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{t_0}) - (\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t_0})\|^2\right] \\
&\overset{(a)}{=} \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{t_0}\|^2\right] - \|\bar{\boldsymbol{w}}_t - \bar{\boldsymbol{w}}_{t_0}\|^2 \\
&\leq \underbrace{\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^{K} p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_{t_0}\|^2\right]}_{A_1}.
\end{aligned} \tag{17}$$

In equality $(a)$, we expand the quadratic equation and use the fact that $\bar{\boldsymbol{w}}_t = \sum_{k=1}^{K} p_k \boldsymbol{w}_t^k$. Then we further have

$$\begin{aligned}
A_1 &\overset{(a)}{=} \sum_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{\xi}}\left[\Big\|\sum_{s=t_0}^{\min\{t, t_0+E_r^k-1\}} \eta\nabla F_k(\boldsymbol{w}_s^k)\Big\|^2\right] \\
&\overset{(b)}{\leq} \sum_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{s=t_0}^{t_0+E_r^k-1} E_r^k \eta^2 \|\nabla F_k(\boldsymbol{w}_s^k)\|^2\right] \\
&\overset{(c)}{\leq} 2L\eta^2 \sum_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{s=t_0}^{t_0+E_r^k-1} E_r^k (F_k(\boldsymbol{w}_s^k) - F_k(\boldsymbol{w}^{k^*}))\right] \\
&\overset{(d)}{\leq} 2L\eta^2 \sum_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{s=t_0}^{t_0+E_r^k-1} E_r^k (F_k(\boldsymbol{w}_{t_0}^k) - F_k(\boldsymbol{w}^{k^*}))\right] \\
&= 2L\eta^2 \sum_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{\xi}}[E_r^{k^2}(F_k(\bar{\boldsymbol{w}}_{t_0}) - F_k(\boldsymbol{w}^{k^*}))],
\end{aligned} \tag{18}$$

where equality $(a)$ is based on the local optimization process, i.e., Eq. (16). Inequality $(b)$ is a consequence of $\|\sum_{i=1}^{n} a_i\|^2 \leq n\sum_{i=1}^{n}\|a_i\|^2$. Inequality $(c)$ depends on the L-Smoothness of $F_k(\cdot)$, i.e., $\|\nabla F_k(\boldsymbol{w}_s^k)\|^2 \leq 2L(F_k(\boldsymbol{w}_s^k) - F_k(\boldsymbol{w}^{k^*}))$. Inequality $(d)$ depends on the fact that $F_k(\boldsymbol{w}_{t+1}^k) \leq F(\boldsymbol{w}_t^k)$, $\forall t \in \{t_0, \cdots, t_0 + E_r^k - 1\}$ (i.e., the local optimization process is a non-increasing sequence [14]), where $\boldsymbol{w}^{k^*} := \min_{\boldsymbol{w}} F_k(\boldsymbol{w})$. Finally, as $t_0$ is the latest aggregation step, we have $\bar{\boldsymbol{w}}_{t_0} = \boldsymbol{w}_{t_0}^k$.

For $E_r^k$, we have

$$E_r^k = E_{const} * n_{1,k,r}^{opt} \leq E_{const} * \frac{\|R_{g,t_0}\|}{\|R_{l,t_0}^k\|}, \tag{19}$$

where the inequality depends on the Cauchy-Schwarz inequality, i.e., $n_1^{opt} \leq \frac{\|R_g\|\|R_l\|}{\|R_l\|^2} = \frac{\|R_g\|}{\|R_l\|}$. Using Eq. (19) in Eq. (18)

$$A_1 \leq 2L\eta^2 E_{const}^2 \underbrace{\sum_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\|R_{g,t_0}\|^2}{\|R_{l,t_0}^k\|^2}(F_k(\bar{\boldsymbol{w}}_{t_0}) - F_k(\boldsymbol{w}^{k^*})\right]}_{A_2}$$

$$A_2 \overset{(a)}{\leq} B\sum_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\|\nabla F(\bar{\boldsymbol{w}}_{t_0})\|^{2\,2}}{\|\nabla F_k(\bar{\boldsymbol{w}}_{t_0}\|}(F_k(\bar{\boldsymbol{w}}_{t_0}) - F_k(\boldsymbol{w}^{k^*})\right], \tag{20}$$

where inequality $(a)$ is a consequence of Lemma 3 and 4. According to the $\gamma$-strong convexity, we have

$$\|\nabla F(\boldsymbol{w}_s^k)\|^2 \geq 2\gamma(F(\boldsymbol{w}_s^k) - F(\boldsymbol{w}^{k^*})).$$

Using it in Eq. (20)

$$\begin{aligned} A_1 &\leq \frac{LB\eta^2 E_{const}^2}{\gamma}\mathbb{E}_{\boldsymbol{\xi}}\|\nabla F(\bar{\boldsymbol{w}}_{t_0})\|^2 \\ &\overset{(a)}{\leq} \frac{LB\eta^2 E_{const}^2}{\gamma}\sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\bar{\boldsymbol{w}}_{t_0})\|^2 \\ &\overset{(b)}{\leq} \frac{2LB\eta^2 E_{const}^2}{\gamma}\Bigg[\underbrace{\sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\bar{\boldsymbol{w}}_{t_0}) - \nabla F_k(\boldsymbol{w}_t^k)\|^2}_{A_3} \\ &\quad + \underbrace{\sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\boldsymbol{w}_t^k)\|^2}_{A_4}\Bigg], \end{aligned} \tag{21}$$

where inequality $(a)$ depends on the convexity of $\|\cdot\|^2$, inequality $(b)$ is a consequence of $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. With the L-Smoothness of $F_k(\cdot)$, we have

$$A_3 \leq L^2 \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^K p_k \|\bar{\boldsymbol{w}}_{t_0} - \boldsymbol{w}_t^k\|^2\right] = L^2 A_1.$$

Using $\|\sum_{i=1}^n a_i\|^2 \leq n\sum_{i=1}^n \|a_i\|^2$ to bound $A_4$, we have

$$\begin{aligned} A_4 \leq\ & \underbrace{3\sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\boldsymbol{w}_t^k) - \nabla F_k(\bar{\boldsymbol{w}}_t)\|^2}_{A_5} \\ & + \underbrace{3\sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\bar{\boldsymbol{w}}_t) - \nabla F_k(\boldsymbol{w}^*)\|^2}_{A_6} \\ & + \underbrace{3\sum_{k=1}^K p_k \mathbb{E}_{\boldsymbol{\xi}}\|\nabla F_k(\boldsymbol{w}^*)\|^2}_{A_7}. \end{aligned} \tag{22}$$

As Assumption 5 hold, we have $A_7 \leq 3\sigma_{opt}^2$. Then with Lemma 5 and the fact that $F(\boldsymbol{w}) = \sum_{i=1}^K p_k F_k(\boldsymbol{w})$, we have $A_6 \leq 6LD_F(\bar{\boldsymbol{w}}, \boldsymbol{w}^*)$. Moreover, with the L-Smoothness of $F_k(\cdot)$, we have

$$\begin{aligned} A_5 &\leq 3L^2 \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^K p_k \|\bar{\boldsymbol{w}}_t - \boldsymbol{w}_t^k\|^2\right] \\ &\leq 3L^2 \mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^K p_k \|\bar{\boldsymbol{w}}_{t_0} - \boldsymbol{w}_t^k\|^2\right] = 3L^2 A_1, \end{aligned}$$

where the last inequality depends on Eq. (17). Using $A_3$ and $A_4$ in Eq. (21), we have

$$A_1 \leq \frac{2LB\eta^2 E_{const}^2}{\gamma}(4L^2 A_1 + \underbrace{6LD_F(\bar{\boldsymbol{w}}, \boldsymbol{w}^*) + 3\sigma_{opt}^2}_{A_8}),$$

rearranging the inequality, we have

$$\left(1 - \frac{8L^3 B\eta^2 E_{const}^2}{\gamma}\right) A_1 \leq \frac{2LB\eta^2 E_{const}^2}{\gamma} A_8.$$

Since $\eta \leq \frac{1}{4E_{const}L}\sqrt{\frac{\gamma}{2BL}}$, we have $1 - \frac{8L^3 B\eta^2 E_{const}^2}{\gamma} \geq \frac{3}{4}$, then we have

$$A_1 \leq \frac{16L^2 B\eta^2 E_{const}^2}{\gamma} D_F(\bar{\boldsymbol{w}}, \boldsymbol{w}^*) + \frac{8LB\eta^2 E_{const}^2}{\gamma}\sigma_{opt}^2,$$

which immediately concludes the proof.■

*Lemma 9:* (Bound for One step recursion) Assume Assumption 3, 4, and 5 hold, if $\eta \leq \frac{1}{8L}$, we have

$$\begin{aligned} \|\zeta_{t+1}\|^2 \leq\ & (1 - \eta\gamma)\|\zeta_t\|^2 + \frac{5\eta L}{4} V_t - \eta D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) \\ & + 4\eta^2 \sigma_{opt}^2, \end{aligned}$$

where $\|\zeta_t\|^2 = \mathbb{E}_{\boldsymbol{\xi}}\|\bar{\boldsymbol{w}}_t - \boldsymbol{w}^*\|^2$, $V_t = \sum_{k=1}^K p_k \|\boldsymbol{w}_t^k - \bar{\boldsymbol{w}}_t\|^2$.

*Proof:* According to the definition of $\bar{\boldsymbol{w}}_t$ and the local optimization process, we have

$$\begin{aligned} \|\zeta_{t+1}\|^2 &= \mathbb{E}_{\boldsymbol{\xi}}\|\bar{\boldsymbol{w}}_t - \eta\sum_{k=1}^K p_k \nabla F_k(\boldsymbol{w}_t^k) - \boldsymbol{w}^*\|^2 \\ &= \|\zeta_t\|^2 - 2\eta\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{k=1}^K p_k \langle \bar{\boldsymbol{w}}_t - \boldsymbol{w}^*, \nabla F_k(\boldsymbol{w}_t^k)\rangle\right] \\ &\quad + \eta^2 \mathbb{E}_{\boldsymbol{\xi}}\|\sum_{k=1}^K p_k \nabla F_k(\boldsymbol{w}_t^k)\|^2 \\ &\leq (1 - \eta\gamma)\|\zeta_t\|^2 + \eta L(1 + 2\eta L)V_t \\ &\quad - 2\eta(1 - 4\eta L)D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) + 4\eta^2\sigma_{opt}^2, \end{aligned}$$

where the last inequality depends on Lemma 6 and Lemma 7. Since $\eta \leq \frac{1}{8L}$, we have $1 - 4\eta L \geq \frac{1}{2}$ and $1 + 2\eta L \leq \frac{5}{4}$, hence

$$\begin{aligned} \|\zeta_{t+1}\|^2 \leq\ & (1 - \eta\gamma)\|\zeta_t\|^2 + \frac{5\eta L}{4} V_t - \eta D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) \\ & + 4\eta^2 \sigma_{opt}^2, \end{aligned}$$

which concludes the proof.■

*Theorem 2:* Suppose Assumption 3, 4, and 5 hold, then for a learning rate $\eta > 0$ such that $\eta \leq \min\left\{\frac{1}{8L}, \frac{1}{4E_{const}L}\sqrt{\frac{\gamma}{2BL}}\right\}$, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\xi}}[F(\hat{\boldsymbol{w}}_T) - F(\boldsymbol{w}^*)] \leq\ & \frac{8\|\zeta_0\|^2}{3\eta T} + \frac{80\eta^2 E_{const}^2 BL}{3\gamma}\sigma_{opt}^2 \\ & + \frac{32\eta}{3}\sigma_{opt}^2, \end{aligned}$$

where $\hat{\boldsymbol{w}}_T := \frac{1}{T}\sum_{t=0}^{T-1} \bar{\boldsymbol{w}}_t$ and $\|\zeta_0\|^2 = \|\bar{\boldsymbol{w}}_0 - \boldsymbol{w}^*\|^2$.

*Proof:* We start with Lemma 9. As $\eta\gamma \geq 0$, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\xi}}\|\zeta_{t+1}\|^2 \leq\ & \mathbb{E}_{\boldsymbol{\xi}}\|\zeta_t\|^2 + \frac{5\eta L}{4} V_t - \eta D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*) \\ & + 4\eta^2 \sigma_{opt}^2. \end{aligned}$$

Summing up over $t$, we have

$$\begin{aligned} &\sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\xi}}[\|\zeta_{t+1}\|^2 - \|\zeta_t\|^2] \\ &\leq \eta\left[\frac{5L}{4}\sum_{t=0}^{T-1} V_t - \sum_{t=0}^{T-1} D_F(\bar{\boldsymbol{w}}_t, \boldsymbol{w}^*)\right] + 4T\eta^2\sigma_{opt}^2 \end{aligned}$$

$$\leq \eta\left(\frac{20\eta^2 E_{const}^2 BL^3}{\gamma}-1\right)\sum_{t=0}^{T-1} D_F(\bar{\boldsymbol{w}}_t,\boldsymbol{w}^*)+4T\eta^2\sigma_{opt}^2 + \frac{10T\eta^3 E_{const}^2 BL}{\gamma}\sigma_{opt}^2, \tag{23}$$

where the last inequality depends on Lemma 8. As $\eta \leq \frac{1}{4E_{const}L}\sqrt{\frac{\gamma}{2BL}}$, we have $\frac{20\eta^2 E_{const}^2 BL^3}{\gamma}-1 \leq -\frac{3}{8}$. Then we rearrange Eq. (23)

$$\begin{aligned}&\frac{3\eta}{8}\sum_{t=0}^{T-1} D_F(\bar{\boldsymbol{w}}_t,\boldsymbol{w}^*)\\ &\leq \|\zeta_0\|^2-\|\zeta_T\|^2+4T\eta^2\sigma_{opt}^2+\frac{10T\eta^2 E_{const}^2 BL}{\gamma}\sigma_{opt}^2\\ &\leq \|\zeta_0\|^2+4T\eta^2\sigma_{opt}^2+\frac{10T\eta^3 E_{const}^2 BL}{\gamma}\sigma_{opt}^2.\end{aligned}$$

Hence, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} D_F(\bar{\boldsymbol{w}}_t,\boldsymbol{w}^*) \leq \frac{8\|\zeta_0\|^2}{3\eta T}+\frac{80\eta^2 E_{const}^2 BL}{3\gamma}\sigma_{opt}^2 + \frac{32\eta}{3}\sigma_{opt}^2$$

Finally, according to the convexity of $F(\cdot)$, we have $\mathbb{E}_{\boldsymbol{\xi}}[F(\hat{\boldsymbol{w}}_T)-F(\boldsymbol{w}^*)] \leq \frac{1}{T}\sum_{t=0}^{T-1} D_F(\bar{\boldsymbol{w}}_t,\boldsymbol{w}^*)$, which immediately concludes the proof.■

*Corollary 2:* Choose $E_{const} \leq \sqrt{T}$ and $B \geq \frac{\gamma}{2L}$, then $\eta = \frac{1}{8\sqrt{T}L} \leq \min\left\{\frac{1}{8L}, \frac{1}{4E_{const}L}\sqrt{\frac{\gamma}{2BL}}\right\}$, we have

$$\mathbb{E}_{\boldsymbol{\xi}}[F(\hat{\boldsymbol{w}}_T)-F(\boldsymbol{w}^*)] \leq \frac{64L\|\bar{\boldsymbol{w}}_0-\boldsymbol{w}^*\|^2}{3\sqrt{T}}+\frac{4\sigma_{opt}^2}{3L\sqrt{T}} + \frac{5E_{const}^2 B\sigma_{opt}^2}{12TL\gamma},$$

which gives $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ convergence rate.

## VII. Detailed Analysis of GSNR

### A. The Necessity of FedGSNR

As FedGSNR calculates the optimal local updates for different clients according to their local datasets, it employs the heterogeneous local steps in FL when data is non-iid distributed among clients. Previous work indicates that heterogeneous local steps of FL cause objective inconsistency issue [35], but the elaborately calculated local steps is necessary when the data is non-iid distributed. It minimizes the gap between the ideal optimization process with the global dataset and the practical optimization process in FL to extract more information from the local datasets, thus expediting the convergence of FL.

To objectively illustrate this issue, we provide the following theorem to theoretically explain that when data is non-iid distributed, there is a gap between the ideal optimization process with the global dataset and the FL optimization process.

*Theorem 3 (Gap Between the Global and Local Gradients):* If we denote the distribution of local gradient calculated by the data of $C_k$ as $\phi_k(\boldsymbol{\xi})$, i.e., $\nabla F(\boldsymbol{w};\boldsymbol{\xi}|C_k) \sim \phi_k(\boldsymbol{\xi})$, which implies that the distribution of global gradient calculated by the global dataset is $\phi(\boldsymbol{\xi}) = \sum_{k=1}^{K} p_k\phi_k(\boldsymbol{\xi})$, we have

$$0 \leq Gap_1 \leq \sum_{k=1}^{K}\sum_{i=1}^{K} p_k p_i D_{KL}(\phi_k(\boldsymbol{\xi})\|\phi_i(\boldsymbol{\xi})),$$

where $Gap_1 := \sum_{k=1}^{K} p_k D_{KL}(\phi_k(\boldsymbol{\xi})\|\phi(\boldsymbol{\xi}))$, which is the average KL-divergence between the distributions of global gradient and local gradients. The first equality holds if and only if $\phi_1(\boldsymbol{\xi}) = \cdots = \phi_K(\boldsymbol{\xi})$, which implies that the data is iid distributed.

*Proof:* Based on the definition of $Gap_1$ and the fact that $D_{KL} \geq 0$, we have $0 \leq Gap_1$, which is the first inequality, and the equality holds if and only if $\phi_1(\boldsymbol{\xi}) = \cdots = \phi_K(\boldsymbol{\xi})$.

Then we focus on the second inequality

$$\begin{aligned}&\sum_{k=1}^{K} p_k D_{KL}(\phi_k(\boldsymbol{\xi})\|\phi(\boldsymbol{\xi}))\\ &=\sum_{k=1}^{K} p_k\int_{\boldsymbol{\xi}}\phi_k(\boldsymbol{\xi})\log\frac{\phi_k(\boldsymbol{\xi})}{\sum_{i=1}^{K} p_i\phi_i(\boldsymbol{\xi})}d\boldsymbol{\xi}\\ &=\sum_{k=1}^{K} p_k\int_{\boldsymbol{\xi}}\left[\phi_k(\boldsymbol{\xi})\log\phi_k(\boldsymbol{\xi})-\phi_k(\boldsymbol{\xi})\log\sum_{i=1}^{K} p_i\phi_i(\boldsymbol{\xi})\right]d\boldsymbol{\xi}\\ &\leq\sum_{k=1}^{K}\sum_{i=1}^{K} p_k p_i\int_{\boldsymbol{\xi}}\phi_k(\boldsymbol{\xi})\log\phi_k(\boldsymbol{\xi})d\boldsymbol{\xi}\\ &\quad-\sum_{k=1}^{K}\sum_{i=1}^{K} p_k p_i\int_{\boldsymbol{\xi}}\phi_k(\boldsymbol{\xi})\log\phi_i(\boldsymbol{\xi})d\boldsymbol{\xi}\\ &=\sum_{k=1}^{K}\sum_{i=1}^{K} p_k p_i D_{KL}(\phi_k(\boldsymbol{\xi})\|\phi_i(\boldsymbol{\xi})),\end{aligned}$$

where the inequality depends on the Jensen's inequality and the fact that $\sum_{i=1}^{K} p_i = 1$.■

*Corollary 3 (More Information in Global Gradient):* For Shannon's entropy $H(\nabla F(\boldsymbol{w};\boldsymbol{\xi}))$ and $H(\nabla F(\boldsymbol{w};\boldsymbol{\xi}|C_k))$, $k \in \{1,\cdots,K\}$, we have the following result

$$H(\nabla F(\boldsymbol{w};\boldsymbol{\xi})) - \sum_{k=1}^{K} p_k H(\nabla F(\boldsymbol{w};\boldsymbol{\xi}|C_k)) = Gap_1 \geq 0.$$

Corollary 3 is an immediate consequence according to the definition of Shannon's entropy [5].

Theorem 3 proves that there is a gap, i.e., $Gap_1$, between the ideal optimization process and FL optimization process with identical local steps, and $Gap_1$ vanishes if and only if the data is iid distributed among clients. Moreover, as Shannon's entropy quantifies the information contained in a random variable, Corollary 3 implies that approximating the ideal optimization process can extract more information from the local datasets, thus expediting the convergence of FL. Hence, $Gap_1$ may be the reason that FL with non-iid data suffers from unstable and slow convergence issues [12], [15], [31]. Combining Theorem 3 and Corollary 3, we conclude that minimizing the gap between ideal optimization process and FL optimization process can expedite the convergence of FL, which is the purpose of FedGSNR.

To clarify the necessity of FedGSNR, we go deeper into the gap between the ideal optimization process with the global dataset and the FL optimization process with heterogeneous local steps. As displayed in Fig. 4, the gap is composed of two parts: $Gap_1$, which is the gap between the ideal optimization process and FL with identical local steps (Theorem 3); and $Gap_2$, the gap between FL with identical local steps and the one with heterogeneous local steps (the objective inconsistency [35]). It is worth noting that Theorem 3 demonstrates that $Gap_1$ is decided by the degree of non-iid, which cannot

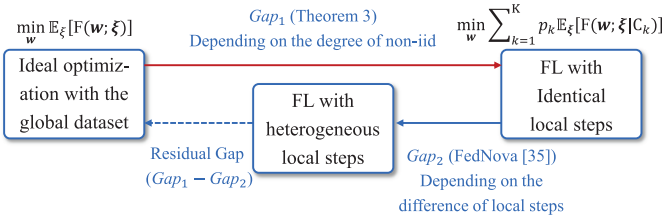



Fig. 4. FedGSNR minimizes the residual gap by solving for the optimal local updates (results in a crafted $Gap_2$) according to the non-iid data (results in $Gap_1$), which elaborately calculates the local steps to make the FL optimization process closer to the ideal optimization process with the global dataset. This strategy extracts more information from the local datasets, thus expediting the convergence of FL.

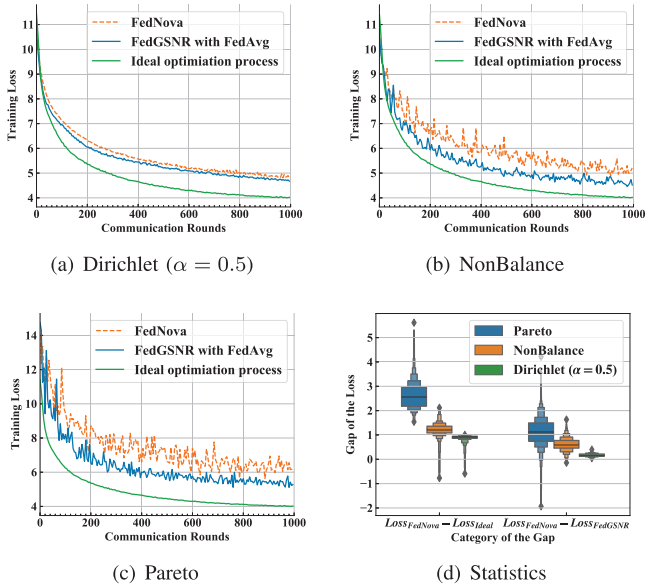
be changed in FL due to the pre-determined data and loss function. Moreover, Wang et al. [35] indicate that $Gap_2$ depends on the difference in local steps, which can be easily changed during training. If the residual gap in Fig. 4 gets smaller, Corollary 3 indicates that FL optimization process can extract more information from the local datasets, thus expediting the convergence of FL. Hence, our target is to minimize the residual gap between the ideal optimization process and the FL optimization process.

Moreover, $Gap_2$ caused by randomly decided local steps may enlarge the residual gap to keep the FL optimization process away from the ideal optimization process (e.g., to make residual gap equal $Gap_1+Gap_2$), this may be the reason that FedNova tries to eliminate $Gap_2$. However, an elaborately calculated local steps can be used to offset the influence of $Gap_1$ and make the residual gap equal $Gap_1 - Gap_2$, which leads the FL optimization process much closer to the ideal optimization process, thus expediting the convergence of FL.

To this end, FedGSNR employs the optimal strategy to set local steps to minimize the residual gap between the FL optimization process and the ideal optimization process. As explained in Theorem 1, FedGSNR decides the local steps of FL by directly solving for the optimal steps (i.e., $n_1^{opt}$), which employs the information of $Gap_1$ described by $\boldsymbol{\mu}_g$, $\boldsymbol{\mu}_l$, $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\Sigma}_l$ to minimize the Wasserstein distance, i.e., the residual gap, between the FL optimization process and the ideal optimization process to expedite the convergence of FL.

Specifically, the situations in FedGSNR are two-fold: if the data is iid distributed ($Gap_1 = 0$), the optimal local steps, i.e., $n_{1,k}^{opt}$ ($k$ represents the index of client), are identical for all clients $k \in \{1, \cdots, K\}$, which implies the local steps are identical ($Gap_2 = 0$). Our proposed strategy is equivalent to the ideal optimization process. Moreover, if the data is non-iid distributed ($Gap_1 > 0$), $n_{1,k}^{opt}$, $k \in \{1, \cdots, K\}$, are usually different for different clients, which results in optimal $Gap_2 > 0$ to offset the influence of $Gap_1$. And Theorem 1 guarantees the local steps minimizes the residual gap between the FL optimization process and the ideal optimization process in this case.

To validate this analysis, we compare FedGSNR with FedNova [35] under different non-iid conditions (the details of non-iid condition will be explained in Sec. VIII). In these experiments, we distributed the data among 30 clients with different partition methods and set $E_{const} = 20$. Fig. 5(a), 5(b), and 5(c) indicate that FedGSNR expedites the convergence of FL when the data is non-iid distributed. Moreover, as displayed in Fig. 5(d), we gather statistics of the gaps between different loss curves under different




(a) Dirichlet ($\alpha = 0.5$) (b) NonBalance

(c) Pareto (d) Statistics

Fig. 5. We compare FedGSNR with FedNova [35] under different settings and demonstrate that FedGSNR uses elaborately calculated local steps to expedite the convergence of FL. It is worth noting that Fig. 5(d) displays the statistics of the loss gap in all communication rounds between different loss curves.

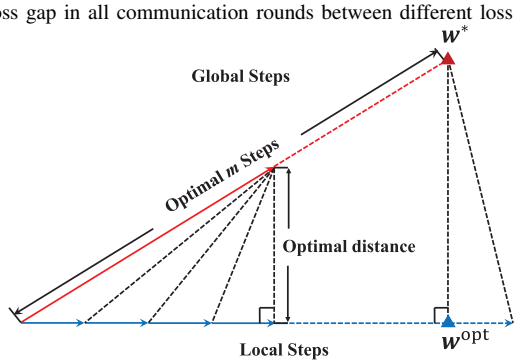



Fig. 6. A case for choosing the global steps: there is an optimal distance between one-step global update and multiple local updates. Furthermore, when global updates converge to the optimum $\boldsymbol{w}^*$, the optimal local updates converge to $\boldsymbol{w}^{\text{opt}}$, which achieves the minimal distance between multiple global updates and multiple local updates.

non-iid conditions. Specifically, as FedNova solves the objective inconsistency issue, the loss gap between it and the ideal optimization process represents $Gap_1$. Meanwhile, the loss gap between FedNova and FedGSNR reflects $Gap_2$. In Fig. 5(d), the boxes on the left represent $Loss_{FedNova} - Loss_{Ideal}$, i.e., $Gap_1$, which indicates that the degree of non-iid decreases from Pareto to Dirichlet ($\alpha = 0.5$) partition. Furthermore, the boxes on the right represent $Loss_{FedNova} - Loss_{FedGSNR}$, i.e., $Gap_2$ caused by FedGSNR, these results indicate that the optimal $Gap_2$ increases according to the degree of non-iid. The increased $Gap_2$ leads the optimization process of FedGSNR much closer to the ideal optimization process, thus expediting the convergence of FL more significantly, which is consistent with our analysis.

### B. Parameter Analysis of FedGSNR

According to Fig. 6, the stochastic gradient descent algorithm converges to the $\epsilon$-neighborhood of optimum after a constant steps (usually more than $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ [29]). For convenience, we denote this constant as $m_{opt}$. In practice, as Fig. 6 indicates, with determined differences between global

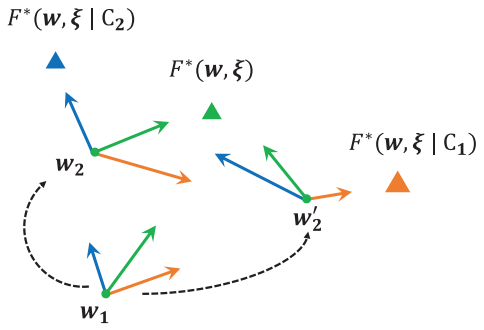



Fig. 7. A representative scenario of GSNR: $r(\boldsymbol{w})$ is a random variable with regard to $\boldsymbol{w}$. If we get closer to the optimum of $C_2$, the GSNR of $C_1$ will increase and the GSNR of $C_2$ will decrease. If we get closer to the optimum of $C_1$, the situation changes oppositely.

distribution and local distribution, i.e., maximal GSNR is constant during a specific round, set the target global steps as $m_{opt}$ is the optimal choice. However, a large number of $m$ leads to a large error of gradient estimation, which is determined by $O(\eta^2 n^2)$. Hence, we need to trade off between the estimation error and the corresponding convergence rate in practice to decide $m$.

Then we focus on Proposition 1, the function to calculate GSNR. Specifically, Cauchy-Schwarz inequality implies that $\Delta^2 \geq 0$, and the equality holds when the local distribution is the same as the global distribution, i.e., the data is iid distributed among all clients. With $\Delta^2$ decreases, which implies the local data distribution approaches the global data distribution, $n_1^{opt}$ gradually increases. Moreover, when $\Delta^2$ reaches its minimum 0, $n_1^{opt}$ attains its maximum value 1. Based on the analysis, we can conclude that the more similarity between the local dataset and the global dataset, i.e., the local dataset achieves the larger GSNR, the more local updates we need for optimization process, which is heuristically experimented in [22].

As described in Proposition 1, $r(\boldsymbol{w})$ is positive related to $n_1^{opt}$. On the one hand, when $n_1^{opt} = 0$, we know that $\langle \boldsymbol{R}_l, \boldsymbol{R}_g \rangle_F = 0$ from its definition, hence the optimal distance $\Delta^2$ achieves its maximum $\|\boldsymbol{R}_g\|_F^2$, and $r(\boldsymbol{w})$ attains its minimum 0. On the other hand, when the local distribution is the same as the global distribution, i.e., $\Delta^2 = 0$, $r(\boldsymbol{w}) \to +\infty$, we denote this scenario as a noiseless optimization process, and the data is iid distributed among all clients. Therefore, we have $r(\boldsymbol{w}) \in (0, +\infty)$.

Based on the former analysis, we know that $r(\boldsymbol{w}) \in (0, +\infty)$. On the one hand, as the data is iid distributed among all clients, i.e., GSNR goes to $+\infty$, the distributed optimization is a noiseless procedure, which means the local updates is unbiased. On the other hand, when GSNR is 0, which means $\langle \boldsymbol{R}_l, \boldsymbol{R}_g \rangle_F \leq 0$, the angle between the local gradient and the global gradient is greater than 90°. In other words, for current optimization, local data distribution is independent of global data distribution, thus for global optimization, it is no better than a random guess, then its signal component will be set to 0, which leads GSNR to be 0.

As for the relationship between GSNR and the parameter $\boldsymbol{w}$, Fig. 7 displays the representative scenario. Due to the randomness of SGD, the new parameter after $\boldsymbol{w}_1$ with another aggregation can be either $\boldsymbol{w}_2$ or $\boldsymbol{w}_2'$. If the parameter is $\boldsymbol{w}_2$, which means we get closer to the optimum of client $C_2$, there are different changes of the GSNR for different clients: for $C_1$, the GSNR is increased; while for $C_2$, the vector is almost orthogonal to global optimization vector, which implies its GSNR is closer to 0. If the parameter is $\boldsymbol{w}_2'$, which is closer to the optimum of $C_1$, the phenomenon is different: the GSNR of $C_1$ is decreased, and the GSNR of $C_2$ is increased. Hence, during the training process, $r(\boldsymbol{w})$ is a random variable correlated to the random variable, i.e., $\boldsymbol{w}$, and we can use the mean or median of GSNR to evaluate the contributions of different clients.

## VIII. Experiment

We run the experiments on the well-known real-world datasets CIFAR-10 and CIFAR-100 mentioned in [18] to validate our design. For all experiments, we use LeNet for CIFAR-10 and VGG-16 for CIFAR-100.

### A. Different Methods of Data Partition

For non-iid settings, we utilize 3 methods for the data partition. First, we follow the settings in [11] to generate non-iid data across different clients by Dirichlet distribution, where $\alpha$ is a parameter representing the non-iid level. Second, we propose NonBalance and Pareto for imbalanced partition, which simulate the imbalanced distributed information in practice.

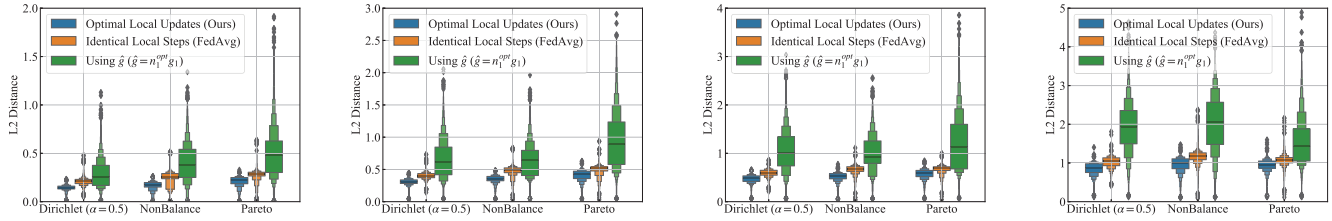
*1) Dirichlet Partition:* Specifically, the prior distribution of the Dirichlet partition is set to be uniform, and the parameter $\alpha$ represents the concentration level. With $\alpha \to +\infty$, the data distributions of all clients tend to be identical, hence the data is iid distributed among all clients. While $\alpha \to 0$, each client only possesses data chosen from just one class, i.e., one label for each client. As for **Label 2.**, it is a specific partition method in [11], and each client owns the data sampled from 2 classes.

*2) NonBalance Partition:* In this method, we simulate the practical scenario of imbalanced information distribution. Specifically, we divide all clients into three categories: abundant information, medium information, and less information, which means the clients' data is chosen from different numbers of labels. First, for clients with abundant information, we randomly choose the data from all the classes, and they account for 10% of the total clients. Second, for the clients with medium information, we randomly choose the data from 50% of the classes, and the ratio of them is 40%. Finally, for the clients with less information, we randomly choose the data from 20% of the classes, and the ratio of them raises to 50%.

*3) Pareto Partition:* In practice, Pareto distribution is a common scenario. It represents the long-tailed distribution of practical scenarios such as the degree of nodes in the complex network, the distribution of social wealth, the distribution of followers in the social network, etc. Hence, we design the Pareto partition to simulate the so-called Two-Eight distribution in practice. In this method, we first sample $N$ points from the Pareto distribution,

$$p(x) = \begin{cases} \dfrac{k \cdot x_{min}^k}{x^{k+1}}, & \text{if } x \geq x_{min} \\ 0, & \text{otherwise,} \end{cases}$$

where $N$ represents the number of clients. Then we denote the corresponding samples as $\boldsymbol{X} = \{x_i\}_{i=1}^N$ and normalize $x_i$ as $\tilde{x}_i = x_i / \max(\boldsymbol{X})$ to guarantee all data distributed in [0, 1]. Finally, we use $\tilde{x}_i$ as the ratio of classes possessed by different clients for random sampling, and set the minimum number of labels among all clients to be 1.




(a) Learning Rate=0.01 (b) Learning Rate=0.03 (c) Learning Rate=0.05 (d) Learning Rate=0.1

Fig. 8. We calculate the L2 distance between the ideal optimization updates with the global dataset ($E_{const} = 20$) and the local updates of different strategies, including optimal local updates, identical local steps, and using $\hat{\boldsymbol{g}}$ ($\hat{\boldsymbol{g}} = n_1^{opt}\boldsymbol{g}_1$). The results demonstrate that L2 distances increase as the learning rate increases, but the strategy of optimal local updates achieves the minimal distance from the ideal optimization process.

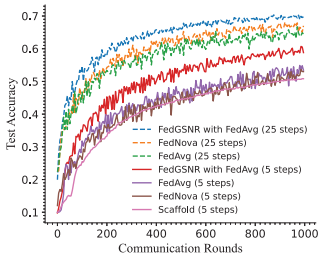



Fig. 9. Test accuracy of different algorithms with different local steps.

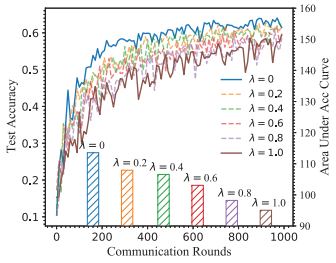
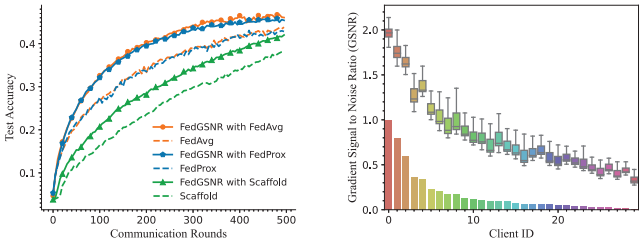



Fig. 10. Test accuracy and the area under acc curve by interpolation.




(a) Test accuracy (b) GSNR

Fig. 11. An imbalanced scenario with Pareto partition on CIFAR-100 dataset.

### B. Experiment Results

*1) The Gap Between the Ideal Updates and the Practical Updates:* We first investigate the distance between the ideal optimization updates and the local updates of different strategies, including optimal local updates (FedGSNR), identical local steps (FedAvg), and using $\hat{\boldsymbol{g}}$ (i.e., $\hat{\boldsymbol{g}} = n_1^{opt}\boldsymbol{g}_1$ in Lemma 2). In these experiments, we employ the global dataset, which gathers all clients' data to calculate the gradient, to update the model for $E_{const} = 20$ steps to calculate the ideal updates. Then we use different strategies to update the model parameters from the identical initial model and calculate the L2 distance between different local strategies and the ideal updates. As illustrated in Fig. 8, the distance between ideal updates and different strategies consistently increases as the learning rate increases, the reason is that a large learning rate will enlarge the gap between the ideal updates and the practical updates regardless of the strategies. However, the distance between optimal local updates and the ideal updates is the minimum among the three strategies, which validates our analysis of FedGSNR in Sec. VII, indicating that FedGSNR can minimize the gap to the ideal optimization process. Moreover, the strategy of using $\hat{\boldsymbol{g}}$ ($\hat{\boldsymbol{g}} = n_1^{opt}\boldsymbol{g}_1$) results in the maximal distance because compared to the strategy of FedGSNR, which only uses one approximation for calculating the optimal local steps, the strategy of using $\hat{\boldsymbol{g}}$ brings additional approximation errors of gradient calculation into the optimization process, thus enlarging the approximation errors.

Then we compare FedGSNR with other methods. To ensure all methods are comparable, we set the total computation (i.e., local updates) to be equal. Therefore, we set the local updates for different clients to be $E_k = NE_{const}\frac{n_{1,k}^{opt}}{\sum_{i=1}^{K} n_{1,i}^{opt}}$ in FedGSNR, where $N$ and $E_{const}$ represent the active clients and the local updates of baseline algorithms, respectively. Note that $E_k$ is a redistribution of local steps.

*2) Model Convergence of FedGSNR:* In these experiments, we set $\mu = 0.01$ for FedProx, and compare the performance of different algorithms to its combination with FedGSNR. According to Table II, FedGSNR with different algorithms achieve faster convergence versus its original, it reaches a $1.69\times$ speedup on average with comparable accuracy. Besides, Fig. 9 displays the accuracy when we use different $E_{const}$, and FedGSNR with FedAvg converges faster and achieves better accuracy when we use different local steps (Scaffold fails to work when we set $E_{const} = 25$). Moreover, in practice, Pareto's Law is a common principle, which means a small number of clients possess a large number of information. Fig. 11(a) indicates that FedGSNR with different algorithms converge faster and reaches comparable accuracy. These results demonstrate the importance of maximizing channel utilization. Meanwhile, the GSNRs of different clients resemble their label distribution (the histogram at the bottom of Fig. 11(b)), which demonstrates that GSNR can distinguish the information quality between different local datasets. Furthermore, Table IV indicates that the growth of active clients speeds up the convergence of different algorithms. Particularly, FedGSNR gains more benefit from global information as its speedup is increased from $1.4\times$ to $1.8\times$ when active clients grow.

Moreover, with the Pareto partition on CIFAR10, we use the interpolation to further investigate the performance of deviating from optimal local updates, i.e., $\boldsymbol{E} = (1-\lambda)(E_1, \cdots, E_K) + \lambda(E_{const}, \cdots, E_{const})$ (when $\lambda = 1$, the algorithm reduces to FedAvg). The results in Fig. 10 demon-

TABLE II

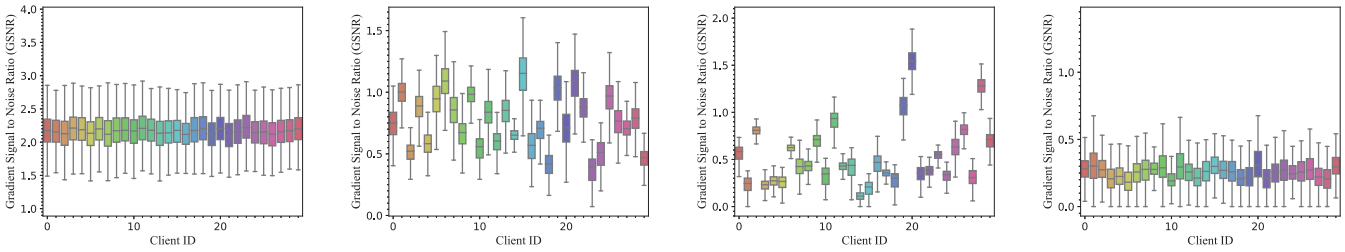
COMMUNICATION ROUNDS TO REACH 0.5 ACCURACY AND CORRESPONDING SPEEDUP[4] OF FEDGSNR ON CIFAR10. WE DISTRIBUTED THE DATA AMONG 30 CLIENTS, UTILIZE THE BATCH SIZE OF 64, AND SET $E_{const} = 20$

| Algorithms | $\alpha = 0.5$ | $\alpha = 0.1$ | Label 2 | NonBalance | Pareto |
|---|---|---|---|---|---|
| FedAvg | 170 (1.0×) | 355 (1.0×) | 470 (1.0×) | 210 (1.0×) | 550 (1.0×) |
| FedGSNR with FedAvg | **115 (1.5×)** | **285 (1.3×)** | **340 (1.4×)** | **155 (1.4×)** | **170 (3.2×)** |
| FedProx | 185 (1.0×) | 430 (1.0×) | 500 (1.0×) | 220 (1.0×) | 385 (1.0×) |
| FedGSNR with FedProx | **140 (1.3×)** | **300 (1.4×)** | **405 (1.2×)** | **180 (1.2×)** | **210 (1.8×)** |
| Scaffold | 385 (1.0×) | 770 (1.0×) | 870 (1.0 ×) | 390 (1.0 ×) | >1K |
| FedGSNR with Scaffold | **140 (2.7×)** | **425 (1.8×)** | **770 (1.1×)** | **170 (2.3×)** | >1K |

TABLE III

BEST TEST ACCURACY ON CIFAR-10. WE DISTRIBUTED THE DATA AMONG 30 CLIENTS, UTILIZE BATCH SIZE OF 64, AND SET LOCAL STEPS $E_{const} = 20$ FOR DIFFERENT ALGORITHMS

| Algorithms | $\alpha = 0.5$ | $\alpha = 0.1$ | Label 2 | NonBalance | Pareto |
|---|---|---|---|---|---|
| FedAvg | **67.06%** | 57.44% | **58.25%** | 65.48% | 57.23% |
| FedGSNR with FedAvg | 66.23% | **61.41%** | 57.69% | **66.42%** | **63.38%** |
| FedProx | 63.14% | 57.80% | **58.84%** | 63.87% | 56.96% |
| FedGSNR with FedProx | **63.5%** | **59.39%** | 58.39% | **68.17%** | **63.39%** |
| Scaffold | 62.05% | 53.98% | 50.38% | 62.79% | **40.73%** |
| FedGSNR with Scaffold | **63.53%** | **58.95%** | **55.48%** | **64.96%** | 39.83% |




(a) IID (b) Dirichlet(1.0) (c) Dirichlet(0.5) (d) Label 2

Fig. 12. The GSNR of different clients. We observe that GSNR is larger and almost the same among all clients when data is iid distributed, then it gets smaller and heterogeneous as the non-iid level grows. Finally, when the data partition method is Label 2, GSNR is small but similar to each other again, which indicates that data is distributed with symmetries regarding the information.

TABLE IV

COMMUNICATION ROUNDS TO REACH 0.5 TEST ACCURACY FOR CLASSIFICATION ON NONBALANCE CIFAR-10 OF 100 CLIENTS AS WE VARY THE NUMBER OF ACTIVE CLIENTS

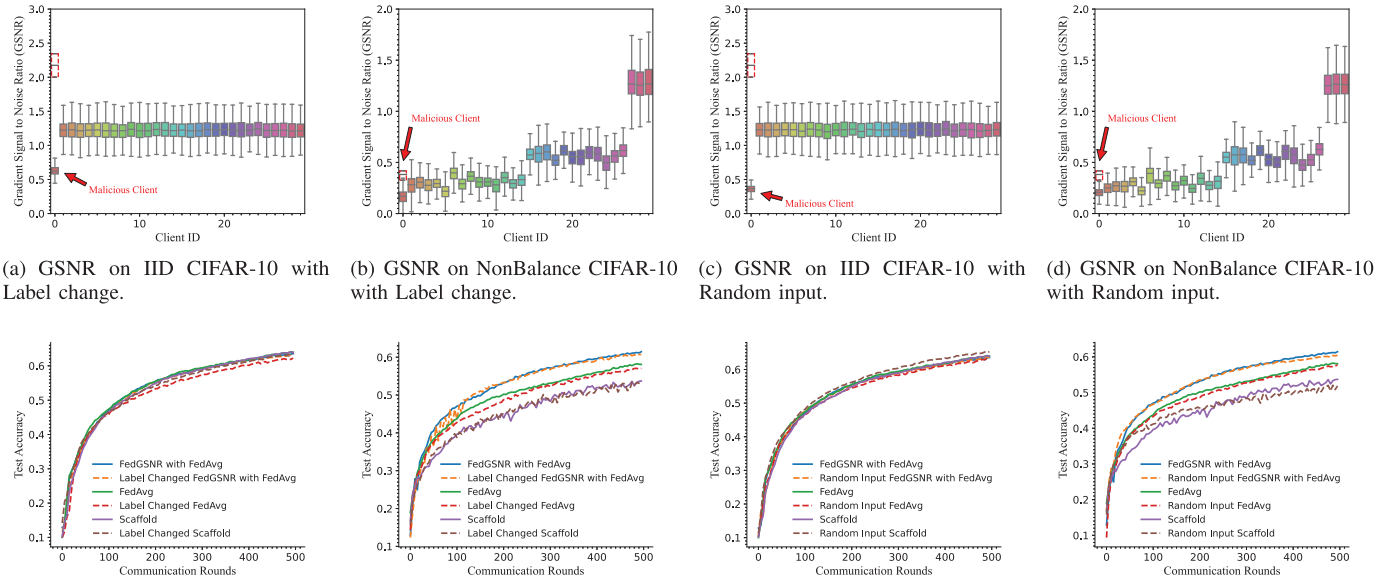
| | 10% | 20% | 100% |
|---|---|---|---|
| FedAvg | 210 (1.0×) | 140 (1.0×) | 100 (1.0×) |
| FedNova | 235 (0.9×) | 140 (1.0×) | 80 (1.2×) |
| Scaffold | 230 (0.9×) | - | - |
| FedGSNR with FedAvg | **150 (1.4×)** | **80 (1.7×)** | **55 (1.8×)** |

strate that the convergence gets faster when we get closer to optimal local updates, which indicates maximizing channel utilization can accelerate the convergence of FL.

*3) The Impact on Test Accuracy:* Versus their original version, FedGSNR with different FL algorithms achieve comparable test accuracy and even outperform their original version when the non-iid degree is increased. For example, in the Pareto scenario, the accuracy of FedGSNR with FedProx achieves an increase of 6.43%. Table III displays the corresponding test accuracy of different algorithms aforementioned in Table II, and it indicates that FedGSNR not only converges faster but also achieves comparable accuracy to its opponents. As for the accuracy drop in Table III, it is because that FedGSNR is a gradient-based algorithm, if the basic method introduces the gradient estimation (i.e., Scaffold), the performance of FedGSNR will be correlated to the precision of such an estimation, and a relatively low precision leads to the corresponding accuracy drop. Meanwhile, as illustrated in Fig. 12, the method of Label 2 distributes the data with symmetries regarding the information, i.e., the GSNRs of all clients are similar to each other, hence it is naturally compatible to identical local updates, and the test accuracy between FedGSNR and its opponents are close to each other.

*4) Evaluate Local Contributions With GSNR:* Fig. 12 displays the variation of GSNR when we utilize the Dirichlet method with different $\alpha$ for the data partition. The results demonstrate that when the non-iid level is increased, the GSNRs of different clients vary dramatically, which indicate the contributions of different clients are different. Moreover, combined GSNR with the results in Table II, the model convergence gets faster than its opponents when the channel utilization is maximized. To further investigate the characteristics of data evaluation, we change the labels $l$ on client 0 to be $(l + 5)\, mod\, 10$, so that client 0 conducts the label

[4]Speedup [15], i.e., $S = \frac{T_{old}}{T_{new}}$, measures the relative performance of two methods.




(a) GSNR on IID CIFAR-10 with Label change.

(b) GSNR on NonBalance CIFAR-10 with Label change.

(c) GSNR on IID CIFAR-10 with Random input.

(d) GSNR on NonBalance CIFAR-10 with Random input.

(e) Test accuracy on IID CIFAR-10 with Label change.

(f) Test accuracy on NonBalance CIFAR-10 with Label change.

(g) Test accuracy on IID CIFAR-10 with Random input.

(h) Test accuracy on NonBalance CIFAR-10 with Random input.

Fig. 13. We make different malicious changes to client 0. In (a) and (b), we change the labels $l$ of client 0 to $l+3$. While in (c) and (d), the labels remain unchanged, and we change the input data to be a uniform distribution $U(-1,1)$. (e)-(h) are the corresponding test accuracy in different scenarios.

flipping attack [10]. Fig. 13(a) and 13(b) illustrate the changes of GSNR when the labels are changed, and the red dashed line box represents the original GSNR when the labels are unchanged. Specifically, we observe that the GSNR is dramatically decreased when we make the malicious change to the labels. Additionally, we instead change the data to be sampled from a uniform distribution and observe a similar phenomenon. For both of the malicious changes, we observe that FedGSNR is more robust.

Fig. 13 displays the change of GSNR and corresponding test accuracy when we apply different malicious changes to the client. Interestingly, comparing Fig. 13(a) and 13(b) with Fig. 13(c) and 13(d), we discover that the decrease of GSNR for random input is larger than label changing. This phenomenon is consistent with our intuition since the malicious attack of label changing still provides more information than the attack of random input. Specifically, the model can still get the information that the data belongs to the same class after label changing. For example, if we change all labels of 'cat' to 'dog', we still know the data of 'cat' belongs to the same class, though they are called 'dog' now. On the contrary, the change of random input cannot provide this information.

## IX. Conclusion

In this paper, we innovatively investigated the FL issue via the perspective of information communication. Under non-iid scenarios, we maximize the channel utilization with the optimal local updates. Then we propose a practical algorithm FedGSNR to calculate the optimal local updates for different FL algorithms, which leads to faster model convergence. Additionally, we derive a method to calculate GSNR directly from the local datasets, which can be utilized to evaluate the local contributions of different clients. Finally, extensive experiments demonstrate the beneficial effect of optimizing FL from the new perspective of information communication, and the perspective also opens up a promising new direction for follow-up research.

## References

[1] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *Proc. ICLR*, Nov. 2021, pp. 2748–2783.

[2] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 2938–2948.

[3] D. Balduzzi, M. Frean, L. Leary, J. Lewis, W.-D. Kurt, and B. McWilliams, "The shattered gradients problem: If resnets are the answer, then what is the question?," in *Proc. ICML*, Aug. 2017, pp. 342–350.

[4] J. Chen, Y. Zhao, Q. Li, X. Feng, and K. Xu, "FedDef: Defense against gradient leakage in federated learning-based network intrusion detection systems," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4561–4576, 2023.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Hoboken, NJ, USA: Wiley, 2006.

[6] C. T. Dinh et al., "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 398–409, Feb. 2021.

[7] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 1605–1622.

[8] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe, "Trading redundancy for communication: Speeding up distributed SGD for non-convex optimization," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2545–2554.

[9] F. Haddadpour and M. Mahdavi, "On the convergence of local descent methods in federated learning," 2019, *arXiv:1910.14425*.

[10] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 603–618.

[11] T.-M. Harry Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," 2019, *arXiv:1909.06335*.

[12] D. Jhunjhunwala, S. Wang, and G. Joshi, "FedExP: Speeding up federated averaging via extrapolation," in *Proc. ICLR*, Jan. 2023, pp. 10924–10958.

[13] Y. Jiang, J. Konečný, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," 2019, *arXiv:1909.12488*.

[14] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Proc. ECML/PKDD*, in Lecture Notes in Computer Science, vol. 9851, Cham, Switzerland. Springer, 2016, pp. 795–811.

[15] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. 37th Int. Conf. Mach. Learn.*, Jul. 2020, pp. 5132–5143.
[16] A. Khaled, K. Mishchenko, and P. Richtárik, "First analysis of local GD on heterogeneous data," 2019, *arXiv:1909.04715*.
[17] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local SGD on identical and heterogeneous data," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2020, pp. 4519–4529.
[18] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 0, 2009.
[19] V. Kulkarni, M. Kulkarni, and A. Pant, "Survey of personalization techniques for federated learning," 2020, *arXiv:2003.08673*.
[20] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6357–6368.
[21] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, Jan. 2018, pp. 429–450.
[22] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. ICLR*, 2020, pp. 1421–1446.
[23] J. Liu, G. Jiang, Y. Bai, T. Chen, and H. Wang, "Understanding why neural networks generalize well through GSNR of parameters," in *Proc. ICLR*, Jan. 2020, pp. 2064–2078.
[24] Y. Mansour, M. Mohri, J. Ro, and A. Theertha Suresh, "Three approaches for personalization with applications to federated learning," 2020, *arXiv:2002.10619*.
[25] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.
[26] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 691–706.
[27] K. Mishchenko, G. Malinovsky, S. U. Stich, and P. Richtárik, "ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally!," in *Proc. ICML*, vol. 162, Jan. 2022, pp. 15750–15769.
[28] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4615–4625.
[29] Y. Nesterov, "Introductory lectures on convex programming volume I: Basic course," *Lect. Notes*, vol. 3, no. 4, p. 5, 1998.
[30] T. Rainforth et al., "Tighter variational bounds are not necessarily better," in *Proc. ICML*, Jul. 2018, pp. 4277–4285.
[31] S. J. Reddi et al., "Adaptive federated optimization," in *Proc. ICLR*, Jan. 2020, pp. 2691–2728.
[32] K. C. Sim, P. Zadrazil, and F. Beaufays, "An investigation into on-device personalization of end-to-end automatic speech recognition models," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 774–778.
[33] C. Villani, *Optimal Transport: Old New*, vol. 338. Berlin, Germany: Springer, 2009.
[34] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *Proc. ICLR*, Jan. 2020, pp. 1739–1754.
[35] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. NeurIPS*, Jan. 2020, pp. 7611–7623.
[36] S. Wang et al., "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.
[37] H. Wei, Y. Zhao, and K. Xu, "G-routing: Graph neural networks-based flexible online routing," *IEEE Netw.*, vol. 37, no. 4, pp. 90–96, Jul. 2023.
[38] D. Yin, A. Pananjady, M. Lam, D. Papailiopoulos, K. Ramchandran, and P. Bartlett, "Gradient diversity: A key ingredient for scalable distributed learning," in *Proc. AISTATS*, 2018, pp. 1998–2007.
[39] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proc. Conf. Artif. Intell. (AAAI)*, 2019, pp. 5693–5700.
[40] Y. Zhao, K. Xu, J. Chen, and Q. Tan, "Collaboration-enabled intelligent internet architecture: Opportunities and challenges," *IEEE Netw.*, vol. 36, no. 5, pp. 98–105, Sep. 2022.
[41] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 14747–14756.

**Qi Tan** (Member, IEEE) received the master's degree from Tsinghua University, Beijing, China, in 2019, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, in 2024. He is currently an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University. His research interests include machine learning, network security, and data privacy.

**Yi Zhao** (Member, IEEE) received the B.Eng. degree from the School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an, China, in 2016, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2021. From 2021 to 2023, he was an Assistant Researcher and a Post-Doctoral Fellow with the Department of Computer Science and Technology, Tsinghua University, where he was a recipient of the Shuimu Tsinghua Scholar Program. He is currently an Associate Researcher with the School of Cyberspace Science and Technology, Beijing Institute of Technology. His research interests include next-generation internet, network security, machine learning, and game theory. He is a member of ACM.

**Qi Li** (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University. He is currently an Associate Professor with the Institute for Network Sciences and Cyberspace, Tsinghua University. His research interests include internet and cloud security, mobile security, and big data security. He is an Editorial Board Member of IEEE TRANSACTIONS ON DEPENDABLE AND SECURITY COMPUTING and ACM DTRAP.

**Ke Xu** (Fellow, IEEE) received the Ph.D. degree from Tsinghua University, Beijing, China. He is currently a Full Professor with the Department of Computer Science and Technology, Tsinghua University. He has published more than 200 technical articles and holds 11 U.S. patents in the research areas of next-generation internet, blockchain systems, the Internet of Things, and network security. He serves as the Steering Committee Chair for IEEE/ACM IWQoS. He has guest-edited several special issues for IEEE and Springer journals. He is an Editor of IEEE INTERNET OF THINGS JOURNAL.