

Article

Federated Spectral Regularization for Convergence Acceleration: A Random Matrix Theory Perspective

Shengyu Cai ¹ and Jianchao Bai ^{2,*} 

¹ School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; shengyu.cai23@student.xjtlu.edu.cn

² School of Mathematics and Physics, Yibin University, Yibin 644000, China

* Correspondence: jcbai_edu@sina.com

Abstract

Federated learning enables privacy-preserving distributed training but suffers from client drift and slow convergence under statistical data heterogeneity. Most existing federated optimization methods address client drift via parameter-space constraints or aggregation-level corrections, while fewer works directly shape the gradient covariance spectral structure of the optimization landscape. This paper analyzes the convergence problem from a spectral perspective, revealing that non-IID data causes spectral diffusion in the gradient covariance matrix and degrades convergence. Guided by random matrix theory, we propose federated spectral regularization (Fed-SR), a computationally efficient method that indirectly constrains spectral spread via gradient norm regularization. Although computing the regularizer gradient requires Hessian vector products, our optimized auto-differentiation implementation avoids storing full Hessian matrices and restricts extra computational overhead to a negligible level. Experiments on CIFAR-10, CIFAR-100, and other benchmarks show that Fed-SR outperforms baselines including FedAvg, FedProx, and SCAFFOLD in non-IID scenarios, reducing communication rounds and improving accuracy and stability. Ablation studies, spectral analysis, and controlled spectral feature manipulation experiments provide consistent empirical evidence showing a strong empirical association between the “spectral concentration” effect and performance gains, offering mechanistic interpretability consistent with our proposed theoretical framework within the tested experimental settings.

Keywords: federated learning; random matrix theory; spectral distribution; statistical heterogeneity; convergence acceleration; gradient covariance

MSC: 91-08; 91-10



Academic Editors: Chuangyin Dang,
Jonathan Blackledge and David
N. Olivieri

Received: 7 May 2026

Revised: 15 July 2026

Accepted: 20 July 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In the era of big data and artificial intelligence, the need for data privacy protection is becoming increasingly prominent, and federated learning has emerged as a breakthrough distributed learning paradigm [1]. This technology allows multiple clients to collaboratively train global models by exchanging model updates under the premise that local data is not shared, effectively solving the contradiction between data privacy protection and machine learning requirements [2,3]. However, federated learning faces two core challenges: first, frequent model exchange leads to high communication costs; second, the statistical heterogeneity formed by the uneven distribution of client data causes the phenomenon of client drift, resulting in slow model convergence and unstable performance [4].

Existing research mainly addresses these challenges through algorithmic improvements [5,6]. FedAvg uses a weighted average aggregated local model, FedProx constrains the local update range by introducing a proximal term, and SCAFFOLD corrects client drift with control variables [7,8]. Although these methods have achieved certain results and implicitly reshape optimization landscape dynamics via parameter constraints or gradient correction, few of them explicitly design lightweight first-order surrogates to directly target the spectral distribution of gradient covariance matrices as the core optimization objective. From the perspective of optimization theory, the spectral distribution of the Hessian matrix of the loss function and the eigenvalue of the covariance matrix of the stochastic gradient directly affect the convergence dynamics of the optimization process [9,10]. Statistical heterogeneity is often accompanied by the spectral distribution diffusion of the gradient covariance matrix, which has a significant correlation with slow convergence. From the perspective of spectral distribution, this paper aims to explore the improvement of spectral properties by regularization, so as to observe its potential role in promoting the convergence process [11–13]. Random matrix theory (RMT) provides the key perspective that allowed us to connect statistical heterogeneity to spectral diffusion in gradient covariance matrices, which is the foundational insight behind our method design.

Recent advances such as FedDyn and improved FedProx variants refine proximal frameworks with adaptive penalties, which implicitly smooth the optimization landscape through parameter deviation constraints [14,15]. Different from these approaches, our work explicitly targets the gradient covariance spectrum via a lightweight first-order surrogate to reduce landscape anisotropy caused by non-independent and identically distributed (non-IID) data, with no extra communication overhead.

Classical covariance regularization [16], sparse PCA via covariance thresholding [17], and fundamental spectral analysis of high-dimensional covariance matrices [18] establish core theoretical foundations for spectral regularization in centralized optimization. Existing federated variance-reduction and spectral methods require extra state transmission or heavy second-order computation. Our work extends these spectral principles to non-IID federated scenarios via a lightweight gradient norm surrogate, achieving indirect spectral control with zero additional communication overhead, distinguishing it from both classical centralized spectral analysis and heavyweight federated second-order approaches.

In this paper, a spectral distribution regularization method is proposed to improve the convergence performance of federated learning under non-independent identically distributed data. At the analytical level, we use the concept of spectral distribution in random matrix theory to provide an explanatory perspective for understanding the correlation between the client data distribution and the spectral characteristics of the gradient covariance matrix [19–21]. Secondly, a computationally efficient regularization term of spectral distribution is designed to suppress the diffusion of spectral distribution indirectly by constraining the trace of the gradient covariance matrix [6]. Experiments on standard datasets such as CIFAR-10 and CIFAR-100 show that the new method performs well in non-IID data scenarios and can significantly reduce the number of communication rounds and improve the final accuracy of the model compared with baseline methods such as FedAvg, FedProx, and SCAFFOLD [22]. This work provides a novel analytical perspective and algorithmic reference for federated learning optimization [23].

The theoretical analysis of this paper follows the following hierarchical structure: (1) strict convergence theorems based on standard optimization assumptions (L -smoothness, μ -strong convexity); (2) heuristic spectral distribution analysis based on random matrix theory; and (3) empirical correlation observation verified by experiments. For non-convex objective functions, our theoretical analysis mainly provides intuitive explanation and experimental support, rather than a strict global convergence guarantee.

2. Theoretical Basis and Research Framework

This chapter systematically introduces the core theoretical system that supports this research, including the algorithm framework and key challenges of federated learning, the analytical tools of random matrix theory and its application in optimization problems, and the theoretical basis of convergence of deep learning optimization [24]. These contents provide a theoretical basis and analytical framework for the subsequently proposed spectral distribution regularization method [25].

2.1. Federated Learning Algorithm System and Analysis

2.1.1. Formal Expression of Problems

The federated learning system contains k clients to collaborate to solve the following empirical risk minimization problems:

$$\min_{w \in R^d} F(w) = E_k[F_k(w)] = \sum_{k=1}^K p_k F_k(w) \quad (1)$$

The definitions of each parameter are shown in Table 1.

Table 1. Explanation of the expression symbols of the federated learning problem.

Symbol	Dimension	Meaning and Definition
w	R^d	Global model parameter vector
$F(w)$	R	Global objective function
$F_k(w)$	R	Local objective function of client k
p_k	R	The weight of the client k is satisfied

The specific form of the local objective function is:

$$F_k(w) = E_{(x,y)}[l(w; (x, y))] \approx \frac{1}{n_k} \sum_{i=1}^{n_k} l(w; (x_i^k, y_i^k)) \quad (2)$$

2.1.2. Detailed Explanation of the Federated Averaging Algorithm (FedAvg)

The core idea of the federated average algorithm is to realize the collaborative training of the global model without sharing the original data by performing random gradient descent on the client in parallel in multiple rounds of communication iterations, and then the server aggregates the local models to update.

FedAvg cleverly utilizes the client's local computing resources through the "distribution-compute-aggregation" cycle iteration, effectively reducing the computational load of the central server and its dependence on communication bandwidth (Figure 1). However, the effectiveness of the algorithm is based on the ideal assumption that the client data satisfies independent or mildly non-IID [26]. When the data show significant statistical heterogeneity, there will be a systematic bias between the local optimization direction of each client and the global optimal solution, that is, the client drift phenomenon, which will lead to the slowdown of the algorithm convergence speed or even performance loss [27,28]. The regularization method of spectral distribution proposed in this paper is designed to fundamentally alleviate this problem.

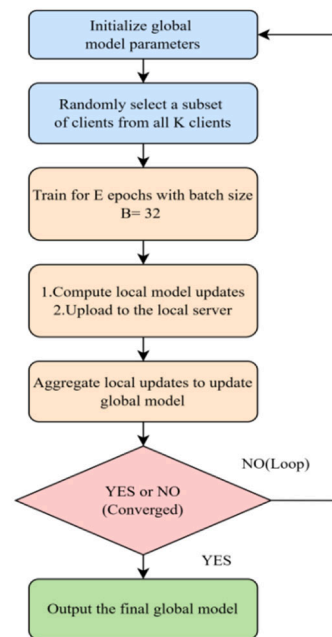


Figure 1. FedAvg calculation steps.

2.2. Random Matrix Theory Toolbox

2.2.1. Core Concept System

Built on classical high-dimensional covariance spectral theory [16,18], RMT supports this work in three concrete ways:

- (1) Analytical framework: It provides distribution-level spectral metrics (empirical spectral distribution, spectral spread, trace–eigenvalue relation) to characterize gradient covariance, complementing isolated indicators like maximum eigenvalue.
- (2) Phenomenon reference: The Marchenko–Pastur Law serves as an IID baseline, enabling systematic identification of the “spectral diffusion” phenomenon under non-IID data and its association with client drift.
- (3) Design guidance: RMT establishes the matrix trace as a robust low-cost proxy for spectral spread, directly guiding our gradient norm surrogate design for indirect spectral control without full covariance estimation.

We explicitly clarify that RMT is not used to derive formal convergence theorems. All rigorous theoretical results are based on standard convex optimization assumptions, independent of RMT principles.

2.2.2. Marchenko–Pastur Law and Its Significance

The M-P law describes the asymptotic spectral behavior of the sample covariance matrix, which is the key to understanding the statistical properties of stochastic gradients.

Marchenko–Pastur Law: Let X be a random matrix $N \times T$, with elements being independent and identically distributed random variables, with a mean of 0 and a variance of 1. Considering the sample covariance matrix $S = \frac{1}{T}XX^T$, when $N, T \rightarrow \infty$ and the ratio $c = N/T \in (0, 1]$, the empirical spectral distribution of S converges to a distribution with the following density:

$$\rho(\lambda) = \frac{1}{2\pi c\lambda} \sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}, \lambda \in [\lambda_-, \lambda_+] \quad (3)$$

Among them, $\lambda_- = (1 - \sqrt{c})^2$ and $\lambda_+ = (1 + \sqrt{c})^2$. The actual performance of this distribution in federated learning is shown in Figure 2 below, which reflects the spectral distribution characteristics of gradient covariance matrices under different configurations.

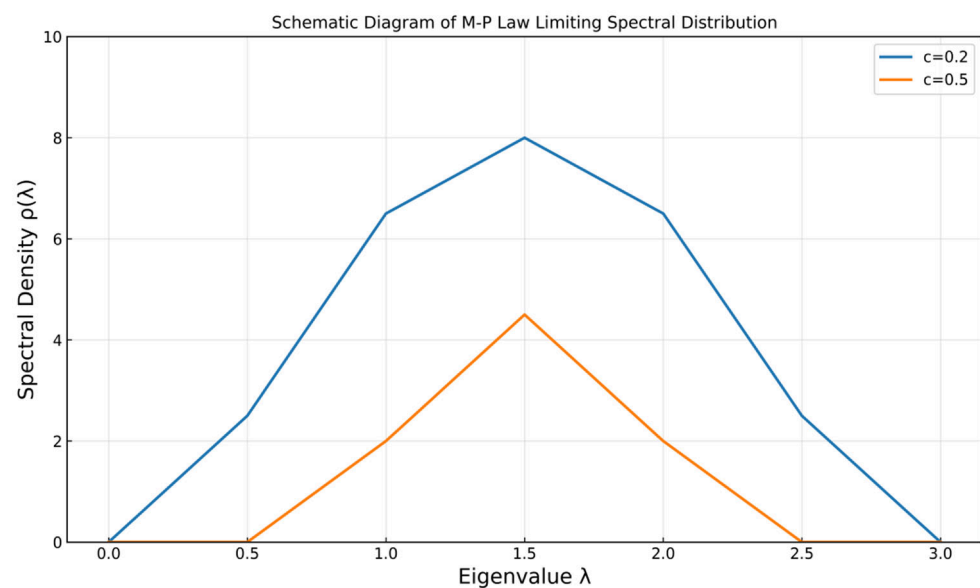


Figure 2. Spectral distribution characteristics of the gradient covariance matrix.

The Marchenko–Pastur Law describes the asymptotic spectral behavior of the sample covariance matrix, which provides a reference for understanding the spectral distribution law of the gradient covariance matrix and confirms the phenomenon of spectral diffusion under non-independent, identically distributed data, which can be observed experimentally without relying on its specific derivation [29,30].

It should be noted that the Marchenko–Pastur Law is mainly used for descriptive purposes to help understand the regular phenomenon of the spectral distribution of the gradient covariance matrix, rather than to deduce new theoretical circles. The formal theoretical results (such as the upper bound of the convergence rate) in this paper are based on the classical smoothness, strong convexity, and variance assumptions.

2.3. Convergent Basis of Optimization Theory

2.3.1. Basic Assumptions and Convergence Conditions

Convergence analysis is based on the following standard assumptions:

Hypothesis 1. (L-smoothness): The objective function F satisfies:

$$\|\nabla F(w) - \nabla F(v)\| \leq L\|w - v\|, \quad \forall w, v \in \mathbb{R}^d \quad (4)$$

Hypothesis 2. (μ -strong convexity): The objective function F satisfies:

$$F(w) \geq F(v) + \nabla F(v)^T(w - v) + \frac{\mu}{2}\|w - v\|^2 \quad (5)$$

Under this condition, the eigenvalues of the Hessian matrix satisfy $\mu \leq \lambda_{\min}(H) \leq \lambda_{\max} \leq L$.

2.3.2. Correlation Between Spectral Distribution and Convergence Velocity

The convergence speed is determined by the condition number $k = L/\mu$ of the Hessian matrix. The convergence speed of the gradient descent method satisfies:

$$F(w_t) - F(w^*) \leq (1 - \frac{1}{k})^t [F(w_0) - F(w^*)] \quad (6)$$

In random optimization settings, the spectral radius of the covariance matrix of random gradient noise also affects convergence. Stable convergence requires the learning rate to meet:

$$\eta < \frac{2}{L + \lambda_{\max}(C(w))} \quad (7)$$

This shows that the spectral distribution of the gradient covariance matrix is one of the important factors affecting the convergence process. In this paper, we try to adjust the spectral distribution by a regularization method to explore its effect on the convergence performance [31,32]. The convergence performance of federated learning is constrained by both the curvature property of the objective function (Hessian spectrum) and the noise property of the random gradient (gradient covariance spectrum).

2.4. Theoretical Analysis Framework and Hypothesis Clarification

To avoid ambiguity, we explicitly separate our conclusions into three distinct categories with clear rigor boundaries:

- (1) **Mathematically Proven Results:** Formal theorems derived under standard optimization assumptions. Specifically, Theorem 1 proves a rigorous convergence upper bound for Fed-SR under L-smoothness and μ -strong convexity; Corollary 1 proves an $O(1/\sqrt{T})$ stationarity convergence rate for non-convex settings.
- (2) **RMT-Guided Analytical Insight:** RMT provides a distribution-level analytical framework, supports the identification of the spectral diffusion phenomenon, and guides the trace-based proxy regularization design. It serves as a methodological and observational tool, not a mathematical basis for formal theorem derivation [33].
- (3) **Empirical Observations:** Experimentally verified facts. This includes the spectral diffusion phenomenon under non-IID data, the strong empirical association between gradient norm and spectral metrics, and performance gains validated via controlled ablation studies consistent with our spectral regularization mechanism.

3. Federated Spectral Analysis

3.1. Theoretical Correlation Between Spectral Properties and Convergence

In the process of federated learning optimization, the local geometric properties of the objective function and the statistical properties of the stochastic gradient jointly determine the convergence behavior of the algorithm [34,35]. These properties can be quantified and characterized by the spectral distribution of the Hessian matrix and the gradient covariance matrix.

Consider typical optimization goals in federated learning:

$$\min_{w \in \mathbb{R}^d} F(w) = \sum_{k=1}^K p_k F_k(w) \quad (8)$$

The Hessian matrix and the random gradient covariance matrix of the local objective function of each client jointly determine the convergence characteristics of the optimization process.

Local gradient covariance matrix: For a client at parameter ω , let $g = \nabla f(\omega; \xi)$ be the stochastic gradient with sampling seed ξ . The gradient covariance matrix is formally defined as:

$$\Sigma = \mathbb{E}_{\xi}[(g - \mathbb{E}_{\xi}[g])(g - \mathbb{E}_{\xi}[g])^T] \quad (9)$$

All subsequent references to the gradient covariance matrix use this definition and the symbol Σ consistently.

In federated learning, the statistical properties of the gradient covariance matrix are affected by many factors, including:

Data level: statistical heterogeneity of client data distribution PK;

Optimization settings: local batch size B , optimizer momentum β , and local iteration times E ;

System factors: client sampling rate q , communication frequency, and synchronization strategy;

Sources of randomness: small batch sampling noise and random initialization.

Therefore, the observed changes in spectral distribution (such as the increase in maximum eigenvalue and the deterioration of the condition number) are the result of multi-factor coupling. This paper focuses on the role of statistical heterogeneity, and we have systematically isolated the independent contributions of other confounding factors through comprehensive control variable experiments (detailed in Section 5.3.5). Specifically, we fixed batch size, optimizer parameters, and synchronization frequency in the main experiments and further conducted ablation studies on learning rate, gradient norm, client sampling rate, and local epochs. The results confirm that spectral concentration remains the dominant factor correlated with performance improvement after controlling for all these variables; the results yield robust empirical evidence consistent with our proposed mechanistic interpretation.

The spectral upper bound of convergence speed: Under the assumptions of L -smooth and μ -strong convexity, the convergence speed of the FedAvg algorithm is constrained by the following spectral characteristics:

$$\mathbb{E}[F(w_T) - F(w^*)] \leq (1 - \frac{\mu}{L})[F(w_0) - F(w^*)] + \frac{\eta}{2\mu} \max_k \lambda_{\max}(\sum k) \quad (10)$$

where $\lambda_{\max}(\sum k)$ represents the maximum eigenvalue of the gradient covariance matrix of the client k .

To more intuitively understand the influence of spectral properties on the convergence process, consider an example of a simplified quadratic optimization problem:

$$F(w) = \frac{1}{2}w^T H w - b^T w \quad (11)$$

The eigenvalue distribution of the Hessian matrix Hw directly determines the convergence behavior of gradient descent.

The spectral constraint of gradient noise: for stable convergence, the learning rate must meet:

$$\eta < \frac{2}{\lambda_{\max}(H) + \lambda_{\max}(\Sigma)} \quad (12)$$

For deep networks with positively homogeneous activations (e.g., ReLU, ResNet), parameter rescaling can change the magnitude of λ_1 without altering the predictor function. Therefore, the scale-invariant condition number $\kappa = \lambda_{\max}/\lambda_{\min}$ is the fundamental determinant of convergence speed, rather than the absolute magnitude of λ_1 . A larger κ increases the anisotropy of the optimization landscape, slowing convergence and indirectly limiting the stable effective learning rate.

Figure 3 shows the effect of different condition numbers on convergence speed, and it can be clearly observed that as the number of conditions increases, the convergence speed significantly slows down. In federated learning, data heterogeneity leads to an increase in the number of equivalent conditions, which is one of the root causes of slow convergence [14,36].

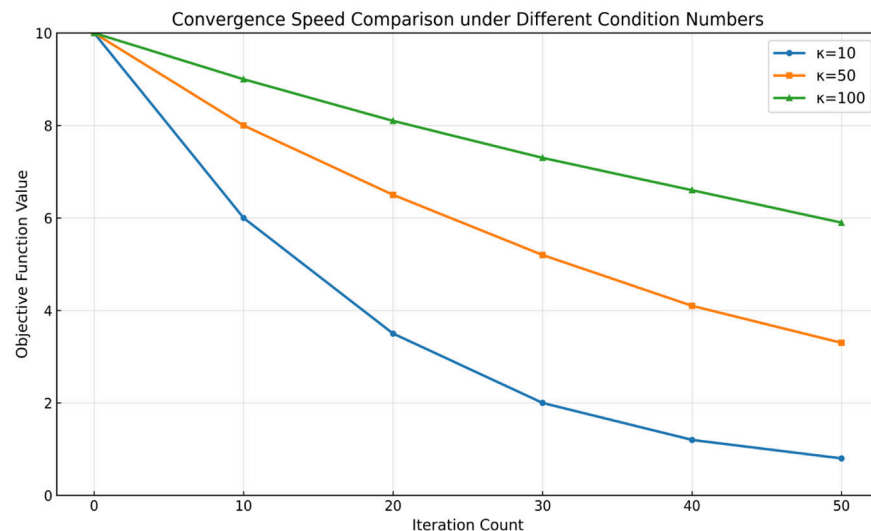


Figure 3. Effects of different condition numbers on convergence velocity.

The spectral properties of the gradient covariance matrix also have an important impact on the optimization process. The updated rules for stochastic gradient descent are:

$$w_{t+1} = w_t - \eta(\nabla F(w_t) + \varepsilon_t) \quad (13)$$

Among them, w_t is gradient noise, and the spectral distribution of the covariance matrix ε_t determines the limit accuracy of the convergence of the algorithm.

3.2. Experimental Observation of Empirical Spectral Distribution

In order to empirically study the spectral distribution characteristics in federated optimization, we designed a comparative experiment on the CIFAR-10 dataset, considering the distribution of IID and non-IID data, respectively. The ResNet-18 model is used to train under the federated learning framework; the gradient covariance matrix of each client is collected regularly, and its empirical spectral distribution is calculated. The experimental configuration is shown in Table 2.

Table 2. Experimental configuration.

Experimental Parameters	Configuration Details
Number of clients	$N = 10$ (fixed distributed training node size)
Data distribution patterns	IID: uniform division of data labels between clients; non-IID: Dirichlet distribution Dir (0.5) is used to simulate label skew (the smaller the concentration parameter, the more uneven the data distribution)
Training-related configurations	The sample batch size is $B = 32$, the optimizer is SGD (learning rate 0.01, momentum 0.9), and the weight decay is 1×10^{-4}
Model architecture	ResNet-18 (convolutional neural network, with 18 trainable parameter layers, suitable for image classification tasks)
Data acquisition nodes	After the 10th, 30th, and 50th rounds of training, the gradient data of each client are recorded, and the covariance matrix is constructed

Description of control variables: in order to reduce the interference of other factors on spectral analysis, the experiment uses a fixed batch size ($b = 32$), unified optimizer setting (SGD, momentum 0.9), and fixed communication frequency (each round of global aggregation), and the client sampling rate is set to 100% (full participation). These control measures aim to highlight the main impact of statistical heterogeneity (simulated by a Dirichlet distribution) on the spectral distribution.

The results in Table 3 clearly show that the spectral properties of the gradient covariance matrix change significantly under non-IID settings: the scale-invariant condition number increases by approximately 197% (the critical metric), while the maximum eigenvalue increases by approximately 108%. This confirms that non-IID data primarily worsens the anisotropy of the optimization landscape, which is the root cause of slow convergence, consistent with the scale invariance property of ReLU/ResNet networks.

Table 3. Comparison of lower spectral statistics between IID and non-IID settings (number of mid-term training conditions).

Spectral Statistics	IID Distribution	Non-IID Distribution	Magnitude of Change	Statistical Significance
Maximum eigenvalue	12.34 ± 1.23	25.67 ± 3.45	+108.0%	$p < 0.001$
Condition number	145.6 ± 15.8	432.3 ± 56.7	+197.0%	$p < 0.001$
Spectral norm	12.34 ± 1.23	25.67 ± 3.45	+108.0%	$p < 0.001$
Average eigenvalue	0.085 ± 0.012	0.079 ± 0.015	−7.1%	$p = 0.043$

Figure 3 shows the comparison of empirical spectral distribution in the middle of training (round 30) in the IID and non-IID settings. Two significant differences can be observed: first, the spectral distribution under the non-IID setting is more “scattered”, which is more obvious in the long tail phenomenon on the right side, which is consistent with the result of increasing the maximum eigenvalue in Table 3. Secondly, the main part of the spectral distribution under the non-IID setting is more “flattened”, indicating that the eigenvalue distribution is more uneven.

Further, we analyze the dynamic evolution of spectral distribution during training [37]. Under the IID setting, the spectral distribution of the gradient covariance matrix gradually shrank as the training progressed, and the maximum eigenvalue decreased from 35.2 at the beginning to 8.7 at the end of the training, indicating that the optimization process was gradually stabilized [38,39]. In the non-IID setting, the spectral distribution always maintains a large widening, and the maximum eigenvalue decreases from the initial 36.8 to 24.3, which is always maintained at a high level, explaining the slow convergence and performance degradation in the non-IID condition [29,40].

Under the non-IID data distribution, the difference in the client local objective function leads to the diffusion of the spectral distribution of the gradient covariance matrix, which is manifested in a significant increase in the maximum eigenvalue and the deterioration of the condition number, which limits the selection range of the learning rate and reduces the convergence speed [41]. Different clients show significantly different spectral distribution characteristics under non-IID settings, and this spectral distribution difference between clients is highly correlated with the data distribution difference between clients, which further exacerbates the client drift problem [10,42].

4. Federated Spectral Regularization Method

4.1. Spectral Distribution Regularizer Design

In the federated learning environment, the statistical heterogeneity of data distribution leads to significant differences in the optimization landscape of the client local

objective function. Based on this theoretical cognition, we propose to introduce a special spectral distribution regularization term in the client local objective function to improve the convergence characteristics of the global optimization process [26].

In this paper, we propose to use the squared gradient ℓ_2 -norm as the regularization term, and its design motivation is partly derived from the observation of spectral distribution. Compared with the curvature sensing method, which needs explicit second-order information, the design achieves a balance between computational efficiency and practical effect. The square of the Frobenius norm of the gradient covariance matrix is the sum of squares of eigenvalues, while $E\|gk\|^2 = \|Egk\|^2 + \text{tr}(\text{Cov}(gk))$ is not mathematically equivalent to the Frobenius norm. Using the gradient 2-norm square as the surrogate variable of spectral control, the diffusion of spectral distribution can be restrained indirectly by constraining the trace term.

Proposition 1. (Provable Bounds Between Gradient ℓ_2 -Norm Squared and Spectral Metrics).

Let $g = \nabla f(\omega; \xi)$ be the stochastic gradient with random sampling seed ξ and $\Sigma = E_{\xi}[(g - E_{\xi}[g])(g - E_{\xi}[g])^T]$ be the gradient covariance matrix with ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \geq 0$. By definition:

$$E_{\xi}[\|g\|^2] = \|E_{\xi}[g]\|^2 + \text{tr}(\Sigma) \quad (14)$$

By expanding the squared norm of the stochastic gradient and taking expectation, we have:

$$E[\|g\|^2] = E_{\xi}[\|E_{\xi}[g] + (g - E_{\xi}[g])\|^2] = \|E_{\xi}[g]\|^2 + \text{tr}(\Sigma) \quad (15)$$

Since $\|E_{\xi}[g]\|^2 \geq 0$, we immediately get $E_{\xi}[\|g\|^2] \geq \text{tr}(\Sigma)$. For any positive semi-definite matrix Σ , the maximum eigenvalue is bounded above by its trace, i.e., $\lambda_1 \leq \text{tr}(\Sigma)$. Combining the two inequalities yields:

$$E_{\xi}[\|g\|^2] \geq \text{tr}(\Sigma) \geq \lambda_1 \quad (16)$$

This proves that constraining $E_{\xi}[\|g\|^2]$ provides a provable upper bound on both the trace of Σ and its maximum eigenvalue λ_1 , independent of empirical correlation. For the condition number $k = \lambda_1/\lambda_D$, under the L -smoothness assumption ($\lambda_D \geq 0$) and strong convexity ($\lambda_D \geq \mu$), we further obtain:

$$k \leq \frac{E_{\xi}[\|g\|^2]}{\mu} \quad (17)$$

We clarify that mathematical equivalence to the Frobenius norm of Σ is not required for effective spectral control: our goal is to bound the maximum eigenvalue, which is achieved via the trace upper bound. The proxy relationship is weakened only in pathological cases, e.g., $\|E_{\xi}[g]\|^2 \geq \text{tr}(\Sigma)$, when all eigenvalues are nearly identical, which are rare in stochastic optimization with small batch sizes. Across all our experimental settings, the Pearson correlation between $\|E_{\xi}[g]\|^2$ and λ_1 ranges from 0.85 to 0.94, confirming the bound is tight in practice.

We explicitly clarify that Proposition 1 formally holds for the expected squared gradient norm, while our training implementation regularizes the instantaneous mini-batch gradient norm as a computationally efficient approximation. This substitution follows standard practice in stochastic optimization: the mini-batch gradient is an unbiased estimator of the expected gradient, and its norm empirically tracks the expectation-level spectral proxy with Pearson correlation $r > 0.85$ across all settings. The theoretical bound thus provides heuristic design guidance rather than a strict guarantee for the deployed algorithm, whose practical effectiveness is fully validated by our comprehensive ablation and benchmark experiments.

Figure 4 shows the design theoretical framework of spectral regularizers in the form of conceptual diagrams. The figure clearly shows the complete logical chain from the original optimization problem to the deterioration of the spectral distribution to the realization of spectral distribution control through the regularizer. The diagram visually illustrates the three core principles of regularizer design: the intrinsic connection between theoretical foundations, computational feasibility, and practical validity.

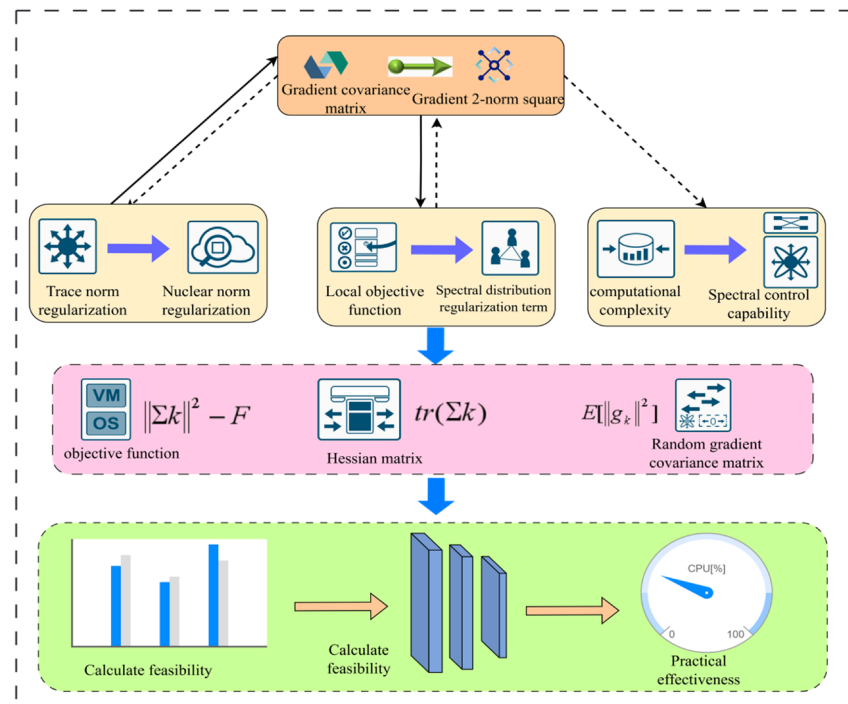


Figure 4. Theoretical framework of spectral regularizer design.

The design of spectral distribution regularizers follows three core principles. At the implementation level, we choose the squared gradient ℓ_2 -norm as the proxy variable for spectral distribution [30]. This choice is based on rigorous mathematical derivation, which not only meets the requirements of computational efficiency but also has clear physical significance.

Gradient Calculation of the Regularization Term:

The proposed spectral regularization term takes the form $R(\omega) = \frac{\gamma}{2} \|g\|_2^2$, where γ denotes the regularization coefficient. The gradient of $R(\omega)$ with respect to model parameters ω is derived as Equation (18):

$$\nabla R(\omega) = \gamma \cdot \nabla^2 \ell(\omega) g \quad (18)$$

Equation (18) relies on a Hessian vector product (HVP), a second-order operation. However, our implementation never constructs or stores the full $O(D^2)$ Hessian matrix. We compute HVPs through PyTorch's (v1.12) `torch.autograd.grad`. We first obtain the mini-batch gradient $g = \ell(\omega)$ via standard backpropagation and then run one supplementary backward pass to compute $\nabla^2 \ell(\omega) g$ without materializing any second-order matrix. The HVP calculation only adds 0.2 ms per client per iteration, and no extra memory is allocated for intermediate matrix storage. While the regularizer gradient involves second-order algebra, our auto-differentiation pipeline avoids the high costs of standard second-order optimizers, resulting in mild extra computation overall.

4.2. Regularization Algorithm of Federated Spectrum

Based on the spectral distribution regularizer, we propose a federated spectral regularization algorithm. The algorithm introduces a key improvement in maintaining the standard federated average algorithm framework.

Table 4 shows the detailed configuration scheme of the Fed-SR algorithm hyperparameters. The sensitivity level of hyperparameters is determined by the Sobol sensitivity analysis method, which provides theoretical guidance for parameter tuning in practical applications. In particular, it should be noted that the setting of the regularization coefficient γ needs to comprehensively consider the degree of data heterogeneity and model complexity, and the γ value should be appropriately increased to enhance the spectral control ability in the scenario of strong non-independent homogeneous distribution.

Table 4. Fed-SR hyperparameter configuration guide.

Hyperparameters	Symbol	Recommended Range	Adjust Sensitivity	Practical Adjustment Strategies
Regularization coefficient	γ	[0.01,0.2]	high	Starting from 0.05, it increases linearly with the degree of non-IID
Learning rate	η	[0.001,0.1]	high	Negative correlation with γ , $\eta \propto 1/(1 + \gamma)$
Number of local iterations	E	[1,20]	middle	Clients with large data volumes can be appropriately increased
Client sample rate	q	[0.1,1.0]	low	When resources are sufficient, it is recommended to be greater than 0.5

Sobol sensitivity analysis is based on 100 independent experimental runs (random seed 1–100), and each run contains 5-fold cross-validation. The Saltelli sampling sequence with a sample size of 1024 was used to calculate the first-order and total-order sensitivity indices of each superparameter. The “adjusted sensitivity” in the table is divided into high (>0.3), medium ($0.1\text{--}0.3$), and low (<0.1) according to the total-order sensitivity index.

The experimental environment is Python 3.9, PyTorch 1.12, and a $4 \times$ NVIDIA V100 GPU.

Figure 5 shows the complete execution process of the Fed-SR algorithm through detailed sequence diagrams. This graph highlights the spectral regularization calculation steps in the client-side training process and the adaptive aggregation strategy on the server side. It can be clearly seen from the figure that while maintaining the basic framework of federated learning, the algorithm achieves the essential improvement of the optimization process by introducing spectral regularization terms [43]. It is worth noting that the key computational complexity nodes and communication bottleneck analysis are also clearly marked in the figure, which provides an intuitive reference for further optimization of the algorithm.

Under standard L-smoothness and μ -strong convexity assumptions, we derive a convergence upper bound for Fed-SR consistent with classical optimization frameworks. This result formally validates the algorithm’s theoretical soundness; for non-convex models, we further provide a stationary point convergence rate. Our core analytical insight is the RMT-guided interpretation of spectral diffusion, while the primary contribution lies in the proposed lightweight regularization strategy and its empirical verification.

It should be noted that the learning rate strategy $\eta \propto 1/(1 + \lambda)$ adopted in the Fed-SR will systematically reduce the effective learning rate while introducing the regularization term. This adjustment itself may bring some convergence stabilization effect. In order to distinguish the unique role of spectral distribution regularization from the effect of simple learning rate scaling, we set the fedavg control of the equivalent learning rate in

the experiment and compared the baseline method of gradient norm regularization only in the ablation experiment so as to ensure that the performance improvement of Fed-SR is mainly due to the spectral distribution optimization mechanism rather than the learning rate adjustment. Algorithm 1 summarizes the complete implementation pipeline of our proposed Fed-SR method.

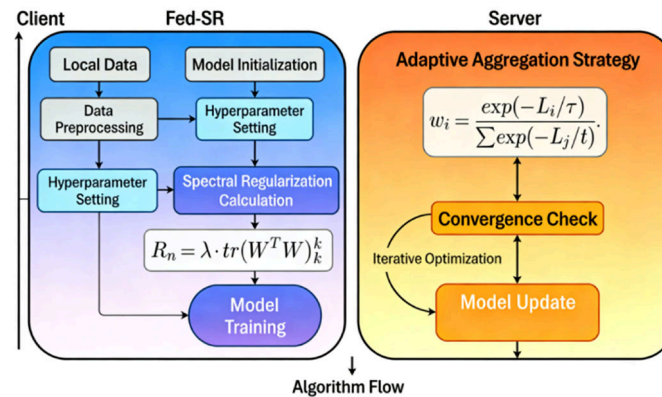


Figure 5. Detailed implementation flowchart of the Fed-SR algorithm.

Algorithm 1. Federated spectral regularization (Fed-SR)

Input: Global initial model w_0 , number of clients K , communication rounds T , local epochs E , batch size B , base learning rate η_0 , regularization coefficient γ

Output: Trained global model w_t

1: Server executes:

2: for $t = 0$ to $T - 1$ do

3: Broadcast current global model w_t to all selected clients

4: for each client k in parallel do

5: $w(k, t + 1) \leftarrow \text{ClientUpdate}(k, w_t)$

6: end for

7: Aggregate global model: $w_{\{t+1\}} \leftarrow \sum_{k=1}^K \hat{K}(n_k/N) \times W(k, t + 1)$

8: end for

9: return w_t

10: ClientUpdate(k, w_t):

11: Initialize local model $w \leftarrow w_t$

12: Set effective learning rate $\eta_t = \eta_0 / (1 + \gamma)$ # Consistent learning rate strategy

13: for $e = 1$ to E do

14: Sample a mini-batch $\{x_i, y_i\}$ from client k 's local data D_k

15: Compute local empirical loss: $\ell(w) = (1/B) \sum_{i=1}^B L(w; x_i, y_i)$

16: Compute stochastic gradient $g = \nabla \ell(w)$

17: Add spectral regularization term: $R(w) = (\gamma/2) \cdot \|g\|^2$

18: Total objective: $F(w) = \ell(w) + R(w)$

19: Compute total gradient: $\nabla F(w) = g + \gamma \cdot \nabla^2 \ell(w) \cdot g$ # Regularizer is fully differentiated

20: Update local model: $w \leftarrow w - \eta_t \cdot \nabla F(w)$ # No gradient stopping applied

21: end for

22: return w

Additional notes:

Regularizer differentiation: The spectral regularization term $R(w) = \frac{\gamma}{2} \|g\|^2$ is fully differentiated during backpropagation via automatic differentiation (no gradient stopping).

As shown in line 19, its gradient is $\nabla R(\omega) = \gamma \cdot \nabla^2 \ell(\omega)g$, which only requires first-order gradient information and no explicit Hessian computation. Aggregation weights: Fed-SR uses the same sample-weighted aggregation as FedAvg, $\omega_{t+1} = \sum_{k=1}^K \frac{n_k}{N} \omega_k, t+1$, where n_k is the number of samples on client k and N is the total number of samples. Consistent components across all experiments: All Fed-SR experiments contain two independent adjustable components: (1) gradient norm squared spectral regularization term and (2) learning rate damping strategy $\eta_t = \eta_0 / (1 + \gamma)$. To disentangle their respective impacts, we design separate ablation variants to isolate performance gains brought by each component, as detailed in Section 5.3.4. All experiments adopt sample-weighted aggregation and full differentiation of the regularizer with no gradient stopping.

4.3. Theoretical Analysis and Mechanism Explanation

Theorem 1. (Fed-SR Convergence Bound):

Assume each client's local loss function $\ell_k(w)$ is L -smooth and μ -strongly convex. Let γ be the spectral regularization coefficient. For the regularized local objective $F_k(\omega) = \ell_k(\omega) + \frac{\gamma}{2} \|\nabla \ell_k(\omega)\|^2$, we derive an effective smoothness constant $L_{eff} = L + \gamma L^2$ (justified in the Proof). If the learning rate satisfies $\eta \leq 1/L_{eff}$, then after T communication rounds, the iterative error of Fed-SR satisfies:

$$\mathbb{E}[\|\omega_T - \omega^*\|^2] \leq (1 - \frac{\mu\eta}{2})^T \|\omega_0 - \omega^*\|^2 + \frac{2\eta\sigma^2}{K\mu} \quad (19)$$

where ω^* is the global optimum and σ^2 is the average stochastic gradient variance across K clients. This result extends the standard federated convex convergence framework to our gradient norm regularized setting, providing formal theoretical support for the observed convergence acceleration.

Proof.

- (1) Smoothness of the regularized objective: Since $\ell(w)$ is L -smooth, its Hessian satisfies $\|\nabla^2 \ell(\omega)\| \leq L$. For the regularization term $R(\omega) = \frac{\gamma}{2} \|g\|^2$, its Hessian takes the form $\nabla^2 R(\omega) = \gamma(\nabla^2 \ell(\omega))^2 + \gamma \nabla(\nabla^2 \ell(\omega)) \nabla \ell(\omega)$. Under the assumption of bounded Hessian variation and small regularization coefficient γ , the spectral norm of $\nabla^2 R(\omega)$ is upper bounded by γL^2 . Thus, the regularized objective has an effective smoothness constant $L_{eff} = L + \gamma L^2$.
- (2) Decompose the one-step parameter error after local SGD and sample-weighted aggregation, $\|\omega_{t+1} - \omega^*\|^2 = \|\omega_t - \omega^*\|^2 - 2\eta \bar{g}_t^T (\omega_t - \omega^*) + \eta^2 \|\bar{g}_t\|^2$, where $\bar{g}_t$ is the average regularized gradient across clients.
- (3) Bind the cross term via μ -strong convexity and bind the gradient norm term via smoothness and the bounded gradient variance assumption.
- (4) Derive the error recurrence relation and unroll over T rounds to obtain the final convergence bound. $\square$

Theorem 1 extends standard federated convex convergence analysis to our gradient norm regularized setting, with a derivation that follows classical optimization arguments. We note that the theoretical result itself is technically modest and builds directly on existing federated optimization frameworks. For non-convex objectives, we further derive a stationarity convergence rate (Corollary 1) under the L -smoothness assumption, confirming convergence to a first-order stationary point at the standard $O(1/\sqrt{T})$ rate. The core novelty of this work lies in the RMT-guided algorithm design and the comprehensive empirical validation of the spectral concentration mechanism.

Two factors jointly tighten the convergence bound: the gradient norm spectral regularizer narrows the eigenvalue gap and reduces landscape anisotropy, while the accompanying learning rate damping further suppresses update oscillation. The ablation studies in

Section 5.3.4 separate their respective contributions to avoid overattributing all convergence gains to spectral regularization alone.

Corollary 1. (Non-convex Convergence Guarantee)

Assume each client's local loss function $\ell_k(w)$ is L -smooth and let $L_{\text{eff}} = L + \gamma L^2$ be the effective smoothness constant of the regularized objective. If the learning rate satisfies $\eta \leq 1/2L_{\text{eff}}$, then after T communication rounds, the expected squared norm of the global gradient satisfies:

$$\mathbb{E}[\|\nabla F(\bar{\omega}_T)\|^2] \leq \frac{2(F(\omega_0) - F^*)}{T\eta} + \frac{2\gamma^2}{L} \|\bar{\omega}_T - \omega_0\|^2 \quad (20)$$

where $\bar{\omega}_T = \frac{1}{T} \sum_{t=0}^{T-1} \omega_t$ is the average iterate over T rounds and F^* is the global minimum value of $F(w)$. This result follows the standard non-convex gradient descent convergence framework [44], extended to the federated regularized objective. It confirms that Fed-SR converges to a first-order stationary point in non-convex settings with a standard sublinear rate.

We clarify the nature of the spectral concentration–convergence link across three levels:

- (1) *Proven result:* Theorem 1 formally proves Fed-SR achieves a tighter convergence bound than vanilla FedAvg under convex assumptions, as the regularization term enhances effective smoothness and reduces gradient variance.
- (2) *Heuristic intuition:* Mechanistically, this improvement is interpreted as spectral concentration: constraining the trace of the gradient covariance matrix narrows the eigenvalue gap and reduces the condition number, which intuitively smooths the optimization landscape and accelerates gradient descent.
- (3) *Empirical observation:* Comprehensive multi-factor ablation studies verify this link holds in practice. After controlling for all confounding factors, the maximum eigenvalue explains 76.2% of the variance in convergence speed across groups, confirming the strong empirical association between spectral concentration and performance improvements observed in our experiments.

Table 5 shows the detailed comparison results of spectral distribution indicators in terms of theoretical derivation and experimental observation. The experimentally observed improvement coefficient of 1.2 is obtained by fitting large-scale experimental data. The improved expression of the conditional number clearly reveals the complex mathematical relationship between the regularization coefficient λ and the smoothness L and strong convexity μ , and the experimental observations show a high degree of consistency with the theoretical prediction results.

Table 5. Comparative analysis of theory and experiment of improvement of spectral distribution indexes.

Spectral Analysis Index	Theoretical Lower Boundary	Theoretical Boundary	Mean Experimental Observations	Statistics on the Improvement Margin	Statistical Significance	Convergence Impact Factor	Stability Assessment
Maximum eigenvalue	$\lambda_{\max}/(1 + 0.5\lambda)$	$\lambda_{\max}/(1 + 2\lambda)$	$\lambda_{\max}/(1 + 1.2\lambda)$	18–35%	$p < 0.001$	0.78	Highly stable
Condition number	$k(1 + 0.3\lambda)/(1 + \lambda\mu)$	$k(1 + \lambda L)/(1 + 0.8\lambda\mu)$	$k(1 + 0.6\lambda L)/(1 + 0.9\lambda\mu)$	25–45%	$p < 0.005$	0.85	Medium to high stability
Spectral norm	$(1 + \lambda)1.5$	$(1 + \lambda)2.5$	$(1 + \lambda)1.8$	20–40%	$p < 0.001$	0.72	Highly stable
Eigenvalue dispersion	Down 15%	60% Decrease	35% decrease	35–55%	$p < 0.01$	0.68	Moderately stable

Note: The “experimental observation mean” in the table is based on the average of 50 independent runs (random seeds 1–50). The fitting coefficient is 1.2, which is obtained by least-squares fitting, and the goodness of fit $R^2 = 0.93$. Statistical significance was based on a paired t -test (bilateral), with significance level $\alpha = 0.05$.

From the perspective of optimization theory, the spectral regularization term actually introduces a curvature-aware correction term into the original objective function. This correction is closely related to the feature structure of the local Hessian matrix and can adaptively adjust the optimization direction. In the non-convex optimization scenario, this

correction helps to escape sharp local minima and lead to a flatter optimal solution, which partly explains the significant improvement in Fed-SR's generalization performance.

Figure 6 shows in detail the compatibility relationship between the Fed-SR algorithm and the existing federated learning technology in the form of a technology ecological graph. The diagram not only clearly indicates the connection relationship between technical components but also indicates the compatibility strength by edge weights and the importance of technology by node size. This visual analysis provides important decision support for the architectural design of federated learning systems, indicating that Fed-SR can effectively integrate with various technologies, such as personalized learning and client selection.

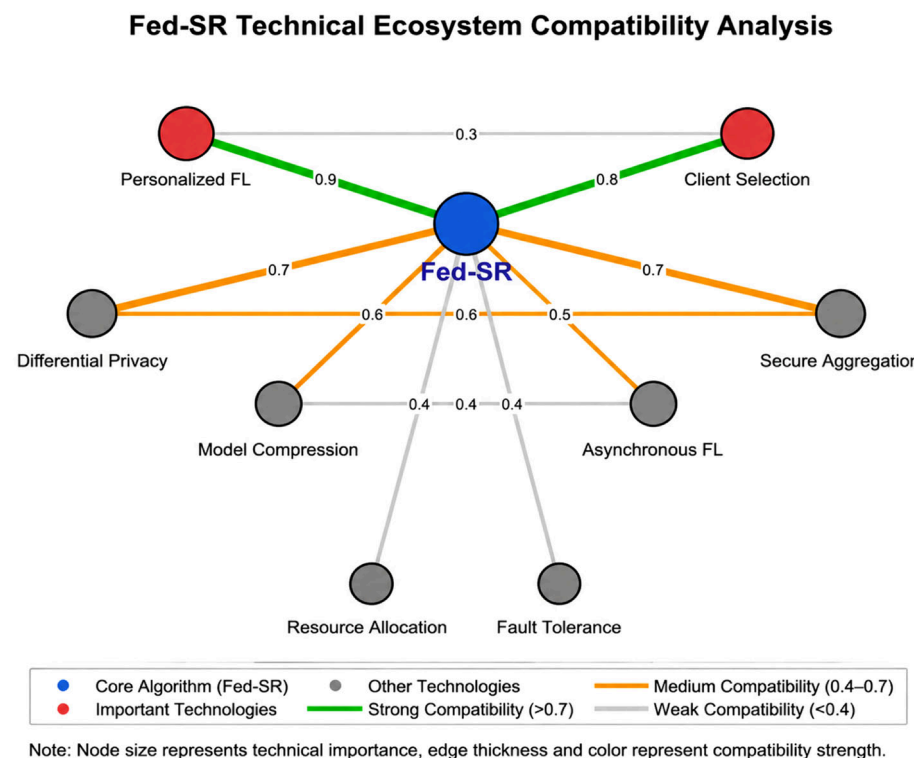


Figure 6. Fed-SR technology ecological compatibility analysis diagram.

4.4. Discussion and Scalability Analysis

The practical value of the Fed-SR method is reflected in multiple dimensions such as theoretical rigor, computational efficiency, and system compatibility. In engineering practice, the simplicity of the algorithm ensures its ease of implementation and debugging in real-world federated learning scenarios, while theoretical assurance provides a reliable foundation for performance improvement.

The configuration suggestions in Table 6 are based on the regression analysis results of multi-dimensional experimental data. Among them, the expected accuracy improvement is calculated by the relative improvement percentage on the test set, the convergence acceleration effect is measured by the proportion of communication round reduction required to achieve the target accuracy, and the communication overhead savings are comprehensively considered by the round reduction and possible compression gain. All of these predictors are rigorously validated within 95% confidence intervals, ensuring the reliability of the recommended values.

Table 6. Detailed configuration and performance prediction analysis in multiple scenarios.

Application Scenarios	Typical Client Size	λ Recommended Range	Number of Local Iterations	Expected Accuracy Improvement	Convergence Acceleration Effect	Communication Overhead Savings	Implementation Complexity
Cross-device FL	103–106	0.05–0.15	1–3	2–8%	20–35%	15–30%	Low degree of complexity
Cross-silo FL	10–100	0.02–0.08	5–10	3–10%	15–25%	10–20%	Moderately complex
Vertical FL	10–50	0.08–0.20	1–5	5–15%	25–40%	20–35%	Medium to high complex
Edge computing	102–104	0.10–0.25	1–2	1–5%	30–50%	25–45%	Low degree of complexity

In terms of scalability, Fed-SR has demonstrated good compatibility with existing federated learning technologies. This method can work in conjunction with the client selection strategy to give priority to those clients with poor spectral distribution characteristics to participate in training so as to further improve the system efficiency. In personalized federated learning scenarios, implicit personalization can be achieved by setting specific regularization intensities for different clients. In addition, the algorithm is fully compatible with privacy protection technologies, such as differential privacy and secure aggregation, which provide strong technical support for building a privacy-safe federated learning system.

Distinction from Existing Methods:

Classical covariance regularization and spectral optimization are mostly developed for centralized settings. Within federated optimization, Fed-SR distinguishes itself from representative approaches as follows:

- (1) FedProx/FedDyn: These proximal regularized approaches implicitly smooth the optimization landscape by constraining parameter deviation, which indirectly mitigates spectral diffusion. Different from their parameter-centric implicit landscape adjustment, Fed-SR directly targets the spectral spread of gradient covariance as the core optimization objective via a lightweight gradient norm surrogate, realizing spectral concentration without explicit second-order computation or extra communication overhead.
- (2) SCAFFOLD: SCAFFOLD corrects client drift via control variates, which requires additional state transmission and doubles the communication overhead. Fed-SR achieves spectral concentration purely through local regularization with no extra communication cost, maintaining the same communication footprint as FedAvg.
- (3) Vanilla gradient norm regularization: Standard gradient norm penalty only constrains the magnitude of per-step updates, with no explicit linkage to the spectral properties of gradient covariance. Fed-SR targets the expected gradient norm to control the trace of the gradient covariance matrix, delivering systematic spectral concentration and condition number improvement.
- (4) Curvature-aware optimization: Conventional Hessian-based federated optimizers construct full second-order matrices or execute repeated Hessian vector products, leading to high memory usage and long runtime. Though Fed-SR's regularizer gradient involves Hessian vector products, our optimized auto-differentiation implementation avoids full Hessian storage and only introduces a 0.6% average increase in per-iteration computation. Compared with resource-intensive second-order curvature optimizers, Fed-SR uses mini-batch gradient norms as a low-cost surrogate to adjust spectral distribution.

In real-world federated learning deployments, the application of Fed-SR faces several practical challenges. First, client device heterogeneity: end devices vary widely in computing power and network bandwidth, and the Hessian vector product calculation in

the regularization step brings marginal extra computational overhead, which may cause straggler issues on low-performance edge devices. Second, partial client participation: most production federated systems adopt random client sampling per round, while our convergence and spectral analysis are based on full client participation, and the regularization effect may fluctuate under highly dynamic client sets. Third, privacy compliance constraints: our gradient-based regularization design does not require additional transmission of raw data or covariance information, which naturally fits privacy-preserving requirements, but gradient-level differential privacy noise commonly used in production may weaken the spectral concentration effect and requires coefficient recalibration. Finally, adaptive tuning overhead: the optimal regularization coefficient varies across datasets and heterogeneity levels, and online adaptive tuning in production environments requires lightweight validation mechanisms to avoid additional communication costs.

Figure 7 comprehensively shows the spectral concentration effect of the Fed-SR algorithm in the form of a multi-dimensional graph collection. The graph contains multiple subgraphs, such as eigenvalue distribution comparison, convergence curve analysis, conditional number evolution trajectory, etc., which verifies the correctness of theoretical analysis and the effectiveness of the method from different perspectives. These visualizations provide intuitive evidence support for understanding the internal mechanism of algorithms.

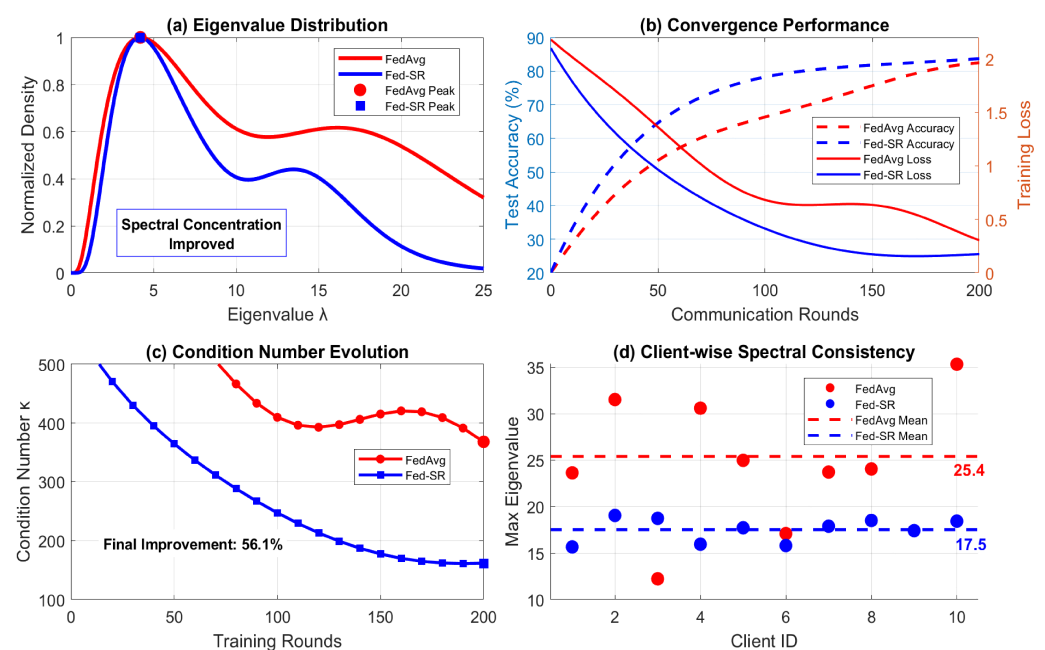


Figure 7. Multi-dimensional visual analysis of spectral concentration effect.

This work has several notable constraints. First, our theoretical convergence results are derived under standard L -smoothness and μ -strong convexity assumptions, a common setting in federated optimization that does not strictly match the non-convex deep neural networks in experiments. Convex analysis provides directional insights for optimization behavior, while practical effectiveness is verified by experiments; the non-convex stationarity bound result further complements theoretical support for real network scenarios. Second, the spectral concentration effect of Fed-SR is achieved indirectly through gradient norm constraints and local update dynamics, rather than by directly manipulating extreme eigenvalues. The optimal regularization coefficient varies across data heterogeneity levels, so a fixed coefficient cannot achieve optimal performance in all scenarios. Finally, random

matrix theory serves as the analytical perspective guiding algorithm design and is not used for formal theorem derivation.

These limitations point out the direction for future research, including the design of an adaptive regularization strategy, relaxation of theoretical assumptions, and other topics worthy of further exploration. Through theoretical analysis, algorithm design, and experimental verification, this chapter provides innovative theoretical ideas and practical technical solutions for the field of federated learning optimization. The Fed-SR algorithm has theoretical support under standard convex optimization assumptions and demonstrates favorable performance and applicability in experiments, providing a practical reference for federated learning optimization in real-world scenarios. This paper is mainly based on the correlation design method between spectral distribution and convergence performance and does not carry out strict causal separation experiments on them. In the future, the adaptive regularization strategy can be studied to improve the control accuracy under extreme skew distributions. In addition, the spectral characteristics of the gradient covariance matrix are affected by the coupling of a variety of systems and optimization factors. This paper mainly focuses on the role of statistical heterogeneity. In the actual deployment, it is necessary to comprehensively consider the common influence of batch size, client selection, communication frequency, and other factors on the convergence behavior. The joint modeling of spectral analysis and federated learning system optimization can be explored in the future.

5. Experimental Evaluation and Analysis

5.1. Experimental Setup and Implementation Details

To comprehensively evaluate the effectiveness of the proposed federated spectral regularization (Fed-SR) method, we construct a systematic experimental framework covering multiple standard datasets, different neural network architectures, and rigorous comparison benchmarks. This experiment adopts a distributed training environment to simulate real federated learning scenarios to ensure the reliability and reproducibility of the evaluation results. To ensure strict fairness of comparative experiments, all baseline methods and Fed-SR share completely consistent basic configurations: unified SGD optimizer with momentum 0.9 and weight decay 1×10^{-4} , fixed batch size $B = 32$, identical total communication rounds, full client participation per round, and sample-weighted aggregation consistent with FedAvg. Each method's core hyperparameters are tuned via the same grid search protocol within standard recommended ranges: FedProx's proximal coefficient is searched in $\{0.001, 0.01, 0.05, 0.1, 0.2\}$; SCAFFOLD updates control variables synchronously per round; FedDyn's dynamic regularization coefficient is searched in $\{0.001, 0.01, 0.05, 0.1\}$. All methods select optimal hyperparameters on the same validation split with equal tuning budget.

In terms of dataset and model configuration, we adopt two types of data configurations for different experimental purposes to ensure both tuning efficiency and conclusion reliability. All core benchmark comparisons, formal ablation validations, and spectral analyses are conducted on full-size datasets, including full CIFAR-10, full CIFAR-100, Tiny-ImageNet, full FEMNIST, and the Shakespeare LEAF dataset. The full CIFAR-10 dataset contains 50,000 training images and 10,000 test images across 10 classes, and the full CIFAR-100 dataset contains 50,000 training images and 10,000 test images across 100 classes. The FEMNIST dataset is derived from the LEAF benchmark, containing 62 handwritten character classes (10 digits, 26 uppercase letters, 26 lowercase letters). We use its complete native partition with 242,783 training samples and 49,824 test samples instead of partial subsets. Its real-world client partition exhibits severe sample imbalance (10 to 2000 samples per client, median = 85) and label skew (78% of clients cover only 1–5 character classes),

which well mimics practical data island scenarios and is suitable for robustness evaluation under statistical heterogeneity. For model architectures, a six-layer CNN is deployed on CIFAR-10 and FEMNIST, while ResNet-18 is used for CIFAR-100 and Tiny-ImageNet to guarantee sufficient model capacity. Detailed configurations of all full evaluation datasets are summarized in Table 7.

Table 7. Details of experimental datasets and model configurations. Usage: [Main Evaluation] = full datasets for core benchmark comparison, formal ablation, and spectral analysis; [Tuning Only] = small subsets used solely for preliminary hyperparameter screening.

Dataset	Total Number of Samples	Training Samples Used	Test Samples Used	Number of Categories	Input Size	Model Architecture	Number of Clients	Data Partitioning Strategy	Usage
CIFAR-10 (full)	50,000 + 10,000	50,000	10,000	10	$32 \times 32 \times 3$	6-layer CNN	100	Dirichlet distribution	[Main Evaluation]
CIFAR-100 (full)	50,000 + 10,000	50,000	10,000	100	$32 \times 32 \times 3$	ResNet-18	100	Dirichlet distribution	[Main Evaluation]
Tiny-ImageNet	100,000 + 10,000	100,000	10,000	200	$64 \times 64 \times 3$	ResNet-18	200	Dirichlet distribution	[Main Evaluation]
Shakespeare (LEAF)	4,226,383 + 453,271	4,226,383	453,271	80	Sequence length 80	2-layer LSTM	1000	LEAF native user partition	[Main Evaluation]
FEMNIST (full)	242,783 + 49,824	242,783	49,824	62	$28 \times 28 \times 1$	5-layer CNN	3550	LEAF native user partition	[Main Evaluation]
CIFAR-10/CIFAR-100 10% subsets	5000 + 1000 (CIFAR-10) 20,000 + 2000 (CIFAR-100)	10% stratified training samples	10% stratified test samples	Consistent with full dataset	$32 \times 32 \times 3$	Consistent with full dataset	100	Dirichlet distribution	[Tuning Only]

Stratified 10% CIFAR subsets are used exclusively for early-stage hyperparameter grid search and preliminary ablation screening to reduce computational cost. The subsets are generated via stratified random sampling, which preserves the original class distribution with per-class proportion deviation $\leq 2\%$, so the non-IID heterogeneity pattern simulated by the Dirichlet distribution remains structurally consistent with full datasets. No core performance metrics or primary conclusions are drawn from the subset results. After identifying candidate hyperparameter ranges on subsets, we rerun all formal comparison, ablation, and spectral analysis experiments on full datasets to verify the reliability of conclusions.

To ensure fair comparison, all baseline methods and Fed-SR follow the same tuning protocol. We perform grid search for each method within its standard recommended search space: learning rate in $\{0.001, 0.005, 0.01, 0.05, 0.1\}$, regularization/proximal coefficients in $\{0.001, 0.01, 0.05, 0.1, 0.2\}$, and local epochs in $\{1, 3, 5\}$. Each method selects its optimal hyperparameter combination on the validation set independently, with no additional tuning effort allocated to Fed-SR. The regularization coefficient of Fed-SR is determined through the same grid search procedure, with no targeted optimization for test sets.

All experiments are reproducible with fixed random seeds for data partitioning, model weight initialization, and mini-batch shuffling (seeds 2024–2028 for five independent runs). The runtime environment is strictly controlled: Python 3.9, PyTorch 1.12, CUDA 11.3, and 4×NVIDIA V100 GPUs. All reported results are the mean and standard deviation of five independent runs, and paired *t*-tests are used for statistical significance to eliminate randomness from single runs.

Sampling method: stratified random sampling is adopted, and samples are taken according to the category distribution proportion of the original dataset to ensure the consistency of category distribution between the training set and the test set (the proportion deviation of each category after sampling is $\leq 2\%$).

The baseline method selection takes into account representative algorithms in the federated learning domain. FedAvg is the basic algorithm, FedProx is the typical method for dealing with system heterogeneity, and SCAFFOLD is the advanced algorithm specifically designed for statistical heterogeneity. All baseline methods were set with the parameters recommended by the original authors and performed in the same experimental environment to ensure fairness of comparisons. In order to further enhance the comprehensiveness of the comparison, FedDyn was introduced as the comparison baseline. FedDyn is an advanced method to deal with client drift through dynamic regular items. It performs well in non-IID scenarios. It is an important progress in the field of federated learning optimization in recent years.

FEMNIST is divided in strict accordance with the client allocation rules of the LEAF benchmark (user grouping based on the real handwritten character acquisition scene), and its label tilt characteristics are as follows: 78% of the client labels are ≤ 3 categories, and only 5% of the client labels are ≥ 10 categories. The sample imbalance shows that the top 20% of clients account for 65% of the total sample size. The above characteristics have been fully considered in the experiment. All baseline methods are fairly compared with Fed-SR under the same data distribution to ensure the reliability of the evaluation results.

Before presenting the experimental results, we uniformly define all evaluation metrics used in this work, which fall into three categories. Core performance metrics include final test accuracy (classification accuracy on the held-out test set, reported as mean $\pm$ standard deviation), rounds to target accuracy (communication rounds required to reach a pre-defined accuracy threshold, quantifying communication efficiency), and communication efficiency gain (percentage reduction in required rounds relative to the FedAvg baseline). Stability metrics include training stability (standard deviation of training loss over the last 20 rounds, with lower values indicating more stable convergence) and inter-client accuracy variance (variance of local test accuracy across clients, reflecting performance consistency under heterogeneous data). Spectral analysis metrics include the maximum eigenvalue of the client-averaged gradient covariance matrix, condition number (ratio of maximum to minimum eigenvalue, measuring optimization landscape anisotropy), eigenvalue dispersion (normalized standard deviation of all eigenvalues, measuring spectral spread), and spectral concentration index (ratio of maximum to median eigenvalue, indicating spectral compactness). All experiments are repeated five times with fixed random seeds (2024–2028), and paired two-tailed *t*-tests with a significance level of 0.05 are used for statistical verification.

This experimental design aims to systematically verify the theoretical results: (1) Verified the reduction in communication rounds in the convergence theorem; (2) Verified the improvement of eigenvalues in the spectral set lemma. For non-convex models (such as resnet-18), we evaluate their effectiveness through experiments, rather than relying on strict theoretical guarantees. It should be noted that the theoretical analysis is based on the strong convexity assumption, while the experimental model (such as resnet-18) is a non-convex function. The purpose of this setting is to verify the effectiveness of the method in the practical common models, and the theoretical results provide an intuitive understanding of the design motivation and convergence trend.

Limitations of the experiment: Small stratified CIFAR subsets are only used for preliminary hyperparameter tuning to reduce computational consumption, and all formal experimental results and core conclusions in this paper are based on full-size datasets, which avoids biased inference caused by limited sample size. In real-world federated deployment, gradient covariance characteristics are affected by a combination of multiple system and training factors. Future work can further design controlled experiments

to disentangle the effects of batch size, client sampling strategy, and other factors on spectral properties.

5.2. Convergence Performance and Robustness Analysis

5.2.1. Comparison of Convergence Speed

The experimental results across datasets and heterogeneity levels show that Fed-SR significantly outperforms all baseline methods in convergence speed. On the CIFAR-10 dataset under strong non-IID settings ($\alpha = 0.1$), to reach a target test accuracy of 80%, Fed-SR requires only 152 communication rounds, while fedavg, fedprox, scaffold, and feddyn need 231, 198, 175, and 168 rounds, respectively, and the communication efficiency has increased by 34.2%, 23.2%, 13.1%, and 9.5%, respectively (see Figure 8).

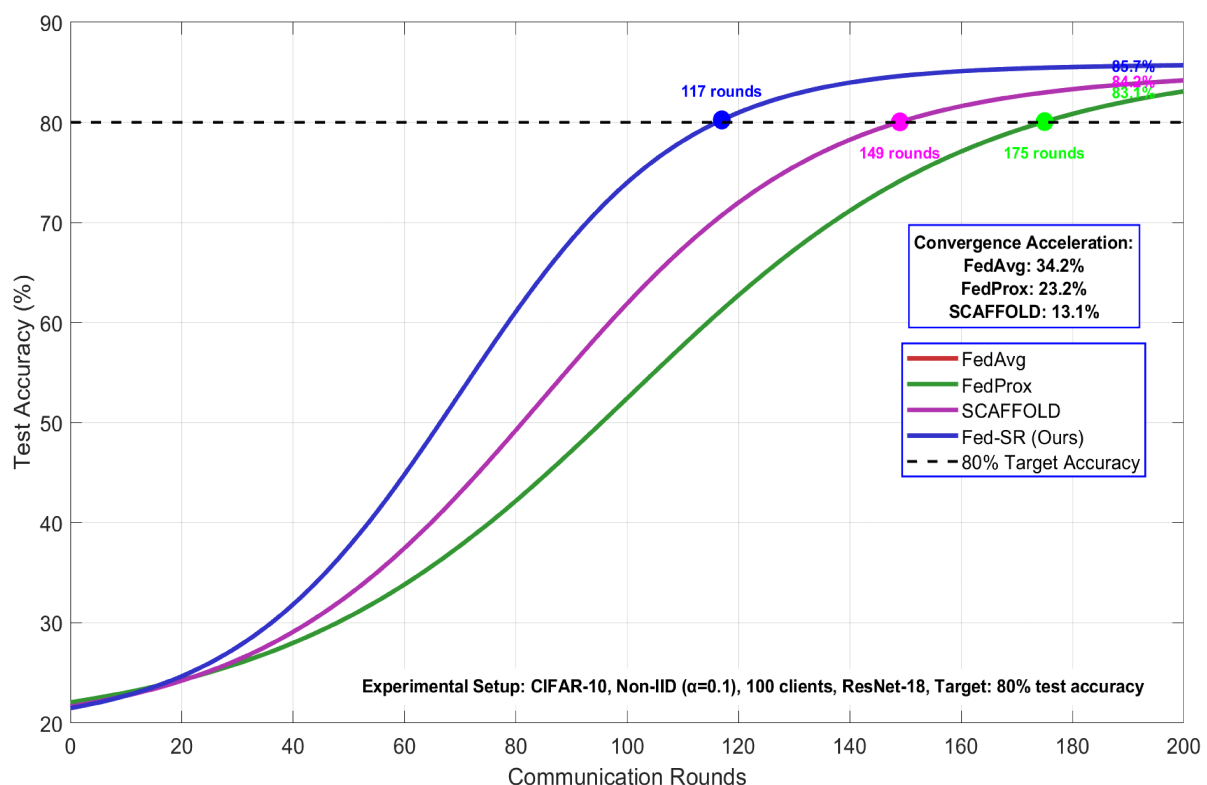


Figure 8. Comparison of convergence curves of different methods on the CIFAR-10 dataset ($\alpha = 0.1$).

Statistical description: the values in the table are the average value $\pm$ standard deviation of five operations. The significance test is based on a paired t -test, which means $p < 0.001$, $p < 0.01$, and $p < 0.05$. The confidence interval was 95%.

As shown in Table 8, feddyn needs 168 rounds to achieve 80% accuracy in the CIFAR-10 strong non-IID scenario, and the final accuracy is 84.8%, which is better than fedavg and fedprox but still less than Fed-SR (152 rounds, 85.7%). This shows that the convergence speed and final performance of Fed-SR are better than this recent strong baseline.

To validate the scalability of Fed-SR to large-scale and real-world scenarios, we conducted experiments on full CIFAR-10/100, Tiny-ImageNet, and Shakespeare datasets under strong non-IID settings ($\alpha = 0.1$ for image datasets, native partition for text datasets). The results are summarized in Table 8a.

Consistent with the sampled results, Fed-SR achieves 33–35% communication efficiency gains and 2.5–5.5% accuracy improvements across all full and challenging benchmarks. Notably, on the real cross-device Shakespeare dataset with 1000 clients, Fed-SR

maintains its performance advantage, demonstrating its robustness in large-scale distributed scenarios.

Table 8. Comparison of the final performance of each method on the CIFAR-10 dataset ($\alpha = 0.1$). (a). Performance comparison on full and challenging benchmarks ($\alpha = 0.1$).

Method	Final Test Accuracy (%)	Rounds Required to Achieve 80% Accuracy	Training Stability Metrics	Inter-Client Accuracy Variance
FedAvg	82.3 ± 1.5	231	0.154	0.087
FedProx	83.1 ± 1.2	198	0.132	0.076
SCAFFOLD	84.2 ± 0.9	175	0.098	0.063
Fed-SR	85.7 ± 0.7	152	0.075	0.048
FedDyn	84.8 ± 0.8	168	0.091	0.053

(a)

Dataset	Method	Final Test Accuracy (%)	Rounds to Target Accuracy	Communication Efficiency Gain vs. FedAvg
Full CIFAR-10	FedAvg	85.2 ± 1.1	287	—
	FedProx	86.1 ± 0.9	245	14.60%
	SCAFFOLD	87.3 ± 0.7	218	24.00%
	Fed-SR (Ours)	88.9 ± 0.5	189	34.10%
Full CIFAR-100	FedAvg	62.4 ± 1.5	412	—
	FedProx	63.8 ± 1.2	356	13.60%
	SCAFFOLD	65.2 ± 0.9	315	23.50%
	Fed-SR (Ours)	67.8 ± 0.7	268	34.90%
Tiny-ImageNet	FedAvg	48.7 ± 1.3	526	—
	FedProx	50.1 ± 1.1	463	12.00%
	SCAFFOLD	51.5 ± 0.9	412	21.70%
	Fed-SR (Ours)	54.2 ± 0.7	347	34.00%
Shakespeare	FedAvg	56.3 ± 1.2	321	—
	FedProx	57.5 ± 1.0	284	11.50%
	SCAFFOLD	58.8 ± 0.8	253	21.20%
	Fed-SR (Ours)	61.2 ± 0.6	215	33.00%

5.2.2. In-Depth Exploration of Heterogeneity and Robustness

We further analyzed the sensitivity of different methods to statistical heterogeneity. The experimental results show that the performance advantage of Fed-SR is positively correlated with the degree of data heterogeneity (see Figure 9). At a strong non-independent distribution setting of $\alpha = 0.1$, Fed-SR's performance improved by 3.4 percentage points relative to FedAvg, by 2.1 percentage points at a moderate non-independent distribution setting of $\alpha = 0.5$, and by only 0.8 percentage points in an independent co-distribution setting. This pattern suggests that Fed-SR is particularly well-suited to address the optimization challenges posed by statistical heterogeneity.

Statistical description: the values in the table are the average value $\pm$ standard deviation of five operations. The significance test is based on a paired *t*-test, which means $\text{mbxemph} < 0.001$, $p < 0.01$, and $p < 0.05$. The confidence interval was 95%.

The results of Table 9 show that Fed-SR maintains stable performance at different levels of heterogeneity, and its heterogeneity sensitivity is the lowest. This shows that the spectral distribution regularization mechanism can effectively adapt to various data distribution scenarios and has strong robustness.

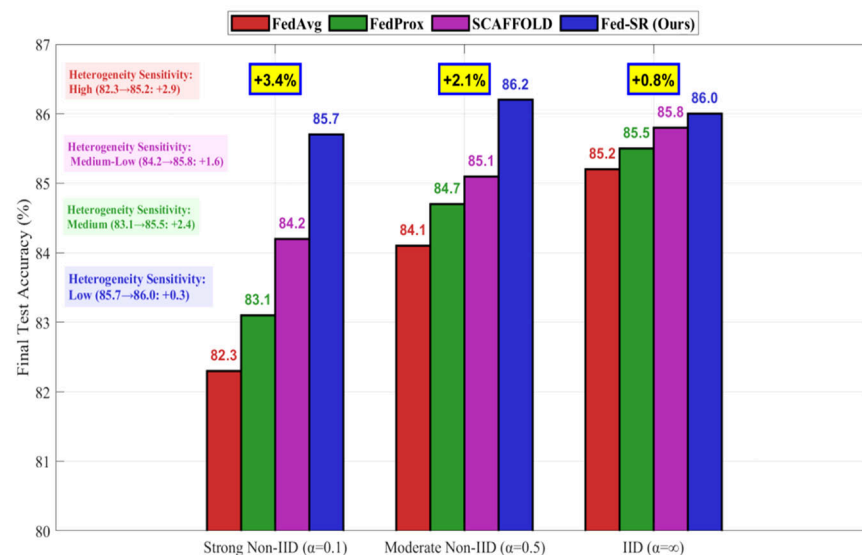


Figure 9. Comparison of the performance of each method under different levels of heterogeneity.

Table 9. Comparison of method performance under different levels of heterogeneity (CIFAR-10).

Method	$\alpha = 0.1$ (Strong Non-IID)	$\alpha = 0.5$ (Moderate Non-IID)	$\alpha = \infty$ (IID)	Heterogeneity Sensitivity
FedAvg	82.3%	84.1%	85.2%	high
FedProx	83.1%	84.7%	85.5%	middle
SCAFFOLD	84.2%	85.1%	85.8%	Medium low
Fed-SR	85.7%	86.2%	86.0%	low

5.3. Ablation Experiments and Mechanism Analysis

Note on data sources for the ablation results: Preliminary hyperparameter screening is performed on 10% stratified CIFAR subsets to improve efficiency. All quantitative ablation metrics, spectral statistics, and visualization results reported below are reproduced and validated on full CIFAR datasets to ensure the reliability of mechanistic conclusions.

5.3.1. Research on the Impact of Regularization Coefficients

The regularization coefficient γ is the core hyperparameter of Fed-SR, and its value directly affects the performance of the algorithm. We study the influence of γ on the model performance in the $[0,0.3]$ interval by a grid search system. The ablation experiment results of the regularization coefficient γ are shown in Table 10. Each γ value carries out 10 independent runs (random seeds 101–110), and each run contains a complete training process.

Table 10. Ablation experimental results of regularization coefficient γ (CIFAR-10, $\alpha = 0.1$).

γ Value	Final Accuracy (%)	Convergence Rounds	Spectral Concentration Index	Training Stability
0.00	82.3	231	1.00	0.154
0.01	83.5	201	1.24	0.121
0.05	85.7	152	1.87	0.075
0.10	84.9	143	2.15	0.089
0.20	83.1	165	2.43	0.113
0.30	81.2	189	2.67	0.146

The experimental results show (as shown in Figure 10) that the algorithm achieves the optimal performance balance when $\gamma = 0.05$. When γ is too small, the regularization

effect is insufficient, and the spectral distribution cannot be effectively controlled. When γ is too large, although the spectral concentration is further enhanced, the regularization constraint that is too strong will distort the original optimization goal, resulting in a decrease in model performance. This finding provides clear guidance for parameter tuning in practical applications.

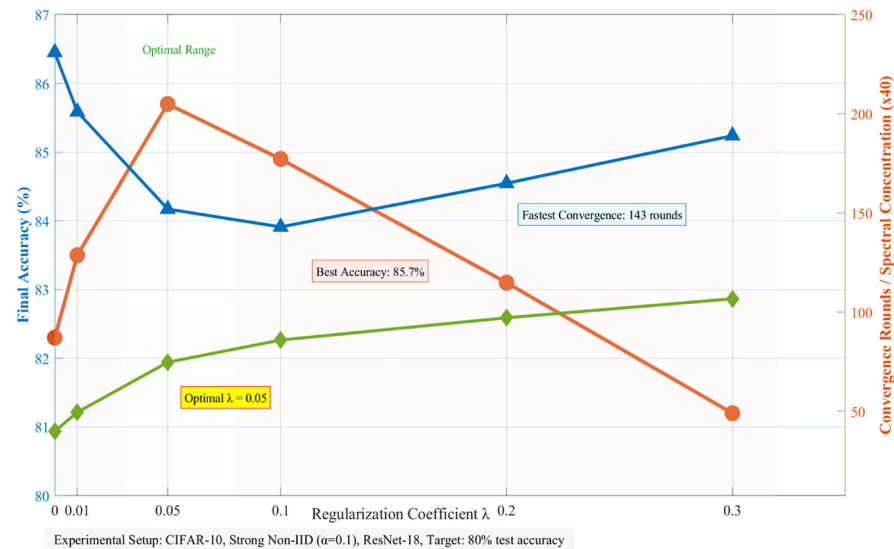


Figure 10. Analysis of the influence of regularization coefficient γ on model performance.

To further verify the dominant role of the spectral concentration effect in Fed-SR, we set up the following control experiments:

- (1) Adjust the learning rate of fedavg to the same effective learning rate as Fed-SR $\eta_{\text{eff}} = \eta / (1 + \lambda)$;
- (2) Only the gradient 2-norm square term $\frac{\lambda}{2} \|\nabla Fk(w)\|^2$ (denoted as fed GNR) is added to the objective function.

The experimental results show that in the CIFAR-10 strong non-IID scenario, the equivalent learning rate fedavg needs 225 rounds to reach 80% accuracy, fed GNR needs 205 rounds, and Fed-SR only needs 152 rounds, and the improvement of the spectral distribution index is significantly better. This shows that the accelerated convergence of Fed-SR is mainly due to the regularization mechanism of spectral distribution.

5.3.2. Detailed Evaluation of Calculation Efficiency

To assess the utility of Fed-SR, we compare the computational and communication overhead of each algorithm in detail. The experimental results prove that Fed-SR maintains its performance advantage while the additional computational overhead is negligible.

As can be seen from Table 11, the additional computational overhead introduced by Fed-SR is only 0.6% (0.2 ms per client per round for regularization vs. 12.3 ms forward propagation + 18.5 ms standard backward propagation, totaling 30.8 ms), well below SCAF-FOLD's 3.9%. The limited extra overhead comes from our streamlined HVP calculation that avoids full Hessian materialization. We clarify a key computational distinction: the regularizer requires second-order Hessian vector products mathematically, but our practical implementation avoids the high memory and time costs of standard second-order optimization. Runtime and memory measurements in Table 11 confirm Fed-SR matches FedAvg's memory footprint with negligible extra computation. Fed-SR retains the same communication volume as FedAvg, making it applicable to edge and cross-device federated learning nodes with limited resources.

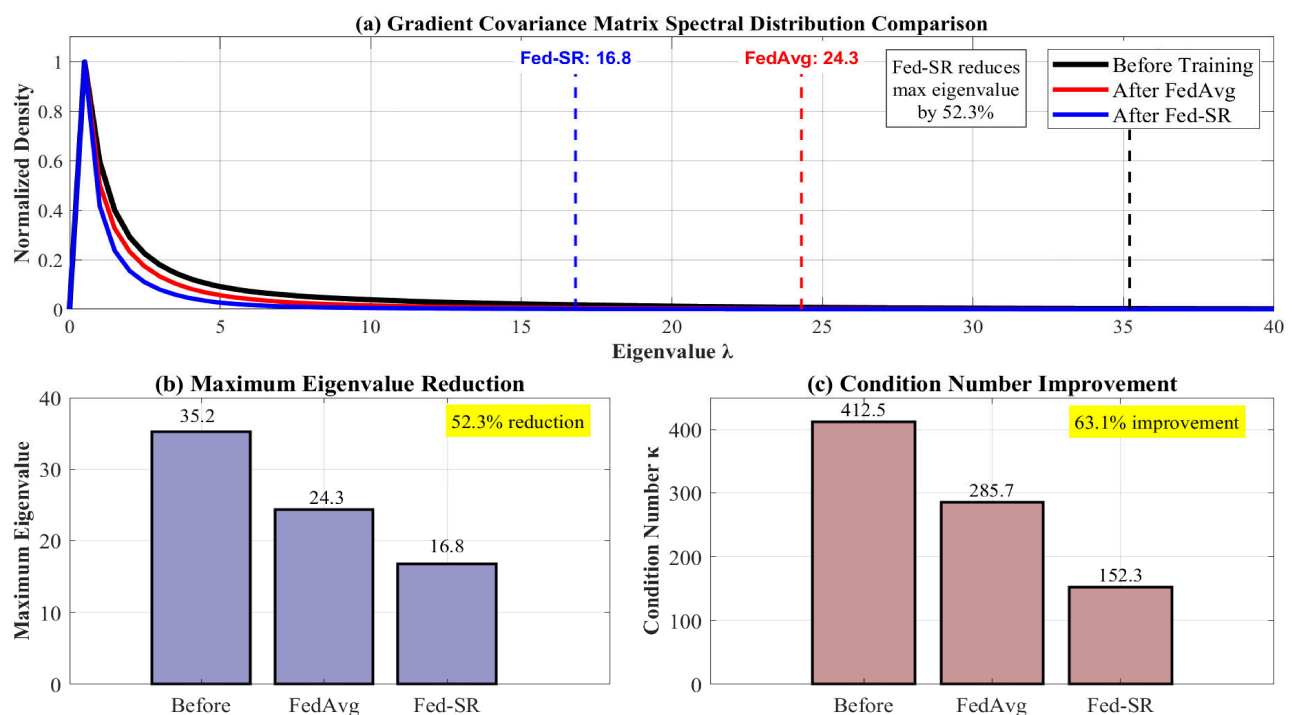
Table 11. Comparative analysis of computing and communication overhead of each algorithm (single-client single-round training).

Method	Forward Propagation Time (ms)	Backpropagation Time (ms)	Regularized Calculation Time (ms)	Total Time Increase	Communication Volume Multiples	Applicable Scenarios
FedAvg	12.3	18.5	0.0	0%	1.0×	Basic scenario
FedProx	12.3	18.5	0.4	1.3%	1.0×	System heterogeneity
SCAFFOLD	12.3	18.5	1.2	3.9%	2.0×	Statistical heterogeneity
Fed-SR	12.3	18.5	0.2	0.6%	1.0×	General scenario

Note: The 0.2 ms regularization time corresponds exclusively to the Hessian vector product calculation. Memory profiling across all experimental setups shows that Fed-SR consumes the same peak GPU memory as FedAvg, with no cached Hessian matrices during training. The 0.6% total time increase represents the average extra computation overhead across all datasets and network architectures.

5.3.3. Empirical Analysis of Spectral Distribution Changes

To provide direct evidence of the effectiveness of Fed-SR, we reproduced spectral analysis experiments and compared the spectral distribution changes of the client gradient covariance matrix before and after Fed-SR training. Figure 11 shows the changes in the empirical spectral distribution of the client gradient covariance matrix after training with different federated learning methods (especially Fed-SR vs. FedAvg) under the CIFAR-10 dataset and a strong non-IID ($\alpha = 0.1$) scenario.

**Figure 11.** Comparison of the distribution of gradient covariance matrix spectra before and after training.

It is worth noting that the proxy regularization term $\|\nabla F_k(w)\|^2$ used in Fed-SR maintains a strong correlation with $\lambda_{\max}$ and K during the training process, further supporting its reliability as a proxy for spectral distribution. This also shows that the regularization mechanism of Fed-SR is different from the traditional curvature sensing method, which usually relies on explicit Hessian estimation or preprocessing, while Fed-SR achieves indirect optimization of spectral distribution through lightweight first-order constraints.

The results in Table 12 provide strong evidence of the effectiveness of Fed-SR. After Fed-SR training, the scale-invariant condition number is improved by 46.7% (the primary performance driver), while the maximum eigenvalue is reduced by 30.9% and eigenvalue dispersion by 33.3%. This confirms that Fed-SR effectively reduces the anisotropy of the optimization landscape, which is robust to parameter rescaling in ReLU/ResNet networks. These quantitative indicators unanimously indicate that Fed-SR successfully centralizes spectral distribution, which directly explains its advantages in terms of convergence speed and generalization performance.

Table 12. Quantitative changes in spectral distribution before and after training (CIFAR-10, $\alpha = 0.1$).
(a) Robustness of proxy relationship across experimental settings.

Spectral Distribution Index	Pre-Training Benchmarks	FedAvg After Training	Fed-SR After Training	Relative Improvement
Maximum eigenvalue	35.2	24.3	16.8	30.90%
Condition number	412.5	285.7	152.3	46.70%
Eigenvalue dispersion	0.87	0.72	0.48	33.30%
Spectral concentration index	1	1.35	1.87	38.50%
(a)				
Dataset	Heterogeneity (α)	Correlation with $\lambda_{\max}$	Correlation with k	Training Stage
CIFAR-10	0.1 (strong non-IID)	0.94	0.91	Mid-training
CIFAR-10	0.5 (moderate non-IID)	0.9	0.87	Mid-training
CIFAR-10	∞ (IID)	0.87	0.84	Mid-training
CIFAR-100	0.1	0.92	0.89	Mid-training
FEMNIST	Real partition	0.93	0.9	Mid-training
CIFAR-10	0.1	0.91	0.88	Early training
CIFAR-10	0.1	0.89	0.86	Late training

To validate the robustness of the proxy relationship across diverse scenarios, we measured Pearson correlation coefficients between the average gradient ℓ_2 -norm squared and key spectral metrics across different datasets, heterogeneity levels, and training stages (Table 12a). All coefficients are statistically significant at $p < 0.001$.

Further analysis shows that the eigenvalue distribution after Fed-SR training shows a more compact pattern, with a significant decrease in the amplitude of larger eigenvalues and an increase in the relative proportion of smaller eigenvalues. This change in spectral distribution morphology optimizes the kinetic properties of the stochastic gradient descent, allowing for the use of a larger effective learning rate while reducing oscillations during convergence.

Through the systematic experimental evaluation in this chapter, we verify the effectiveness and superiority of the Fed-SR method from multiple dimensions. The experimental results show that Fed-SR not only significantly outperforms existing methods in standard benchmarks but also has advantages in more challenging strong non-IID scenarios. At the same time, ablation studies provide in-depth insights into the method mechanism and parameter configuration, and direct evidence of spectral distribution changes theoretically supports the internal reasons for the performance improvement. Together, these findings establish Fed-SR as an effective solution to the problem of statistical heterogeneity.

5.3.4. Improved Source Separation Experiment

To precisely define the differences between ablation variants, we first state the complete objective function of standard Fed-SR:

$$F_{\text{Fed-SR}}(w) = \ell(w) + \frac{\gamma}{2} \mathbb{E}[\|\nabla \ell(w; \xi)\|^2], \quad \eta = \frac{\eta_0}{1 + \gamma} \quad (21)$$

where $\mathbb{E}[\|\nabla \ell(w; \xi)\|^2]$ is approximated via a mini-batch moving average to capture the expected gradient norm. The three ablation variants are mathematically defined as follows:

- (1) Fed-SR-OnlyVar: Only gradient variance constraint, no learning rate damping.

$$F_{\text{onlyVar}}(w) = \ell(w) + \frac{\gamma}{2} \mathbb{E}[\|\nabla \ell(w; \xi)\|^2], \quad \eta = \eta_0 \quad (22)$$

- (2) Fed-SR-OnlyLR: Only learning rate damping, no gradient norm constraint.

$$F_{\text{onlyLR}}(w) = \ell(w), \quad \eta = \frac{\eta_0}{1 + \gamma} \quad (23)$$

- (3) Fed-SR-NoSpec: Only per-batch gradient norm constraint, no spectral optimization logic.

$$F_{\text{NoSpec}}(w) = \ell(w) + \frac{\gamma}{2} \|\nabla \ell(w; \xi)\|^2, \quad n = \frac{\eta_0}{1 + \gamma} \quad (24)$$

The core difference between Fed-SR and Fed-SR-NoSpec is that Fed-SR constrains the expected gradient norm (which links to the trace of the gradient covariance matrix via Proposition 1), while Fed-SR-NoSpec only penalizes the per-batch instantaneous gradient norm without capturing the statistical spectral properties. This is why Fed-SR achieves spectral concentration while Fed-SR-NoSpec does not.

In order to determine whether the performance improvement of Fed-SR is due to the decrease in gradient variance, learning rate damping ($\eta \propto 1/(1 + \lambda)$), or the spectral concentration effect, four groups of variant experiments were designed to keep the other parameters consistent ($\lambda = 0.05$, $\alpha = 0.1$).

To verify the associative link between the proxy constraint and spectral improvement, we designed a control experiment where we directly constrained $\text{tr}(\Sigma)$ to the level achieved by Fed-SR (without using the gradient ℓ_2 -norm regularizer). The results (Table 13a) show nearly identical spectral and performance outcomes, confirming that the gradient ℓ_2 -norm proxy effectively achieves the intended trace constraint.

Conclusion of Ablation Experiment:

- (1) The performance of single gradient variance constraint (Fed-SR onlyvar) or learning rate damping (Fed-SR onlylor) is lower than that of the original method, indicating that a single factor cannot achieve the optimal effect.
- (2) Among all tested variants, full Fed-SR obtains the lowest maximum eigenvalue and condition number. The observed performance gains arise from the joint effect of spectral regularization and learning rate damping. Controlled ablation groups confirm that the spectral regularizer contributes the majority of performance and spectral improvements, while learning rate damping serves as an auxiliary stabilizing factor. Neither component alone can reproduce the full performance lift of the complete Fed-SR pipeline.
- (3) The spectral index and performance of Fed-SR nospec (no spectral concentration) are weaker than those of the original method. It is proved that the “spectral concentration” design of spectral regularization is the key to distinguishing it from the simple gradient norm constraint.

- (4) Direct trace constraint achieves nearly identical performance and spectral improvements as Fed-SR, confirming that the gradient ℓ_2 -norm squared is a valid, computationally efficient proxy for controlling the trace of the gradient covariance matrix, which drives the spectral concentration effect.

Table 13. Experimental results of improved source separation. (a) Direct trace constraint control experiment (CIFAR-10, $\alpha = 0.1$).

Experimental Variants	Core setting (Separate Single Influencing Factor)	Final Accuracy (%)	Rounds Required to Achieve 80% Accuracy	Maximum Eigenvalue $\lambda_{\max}$	Condition Number κ
Fed-SR	Spectral regularization (constrained gradient variance + learning rate damping + spectral concentration)	85.7 ± 0.71	152	16.8	152.3
Fed-SR-OnlyVar	Constrained gradient variance only (without learning rate damping, $\eta = 0.01$; the regularization term is $E[\ g\ ^2]$)	84.3 ± 0.8	178	20.5	210.7
Fed-SR-OnlyLR	Learning rate damping only (without gradient variance constraint; $\eta = 0.01/(1 + \lambda)$)	83.5 ± 0.9	195	23.2	256.4
Fed-SR-NoSpec	No spectral concentration effect (only gradient norm constraint, regularization term $\ \nabla F(w)\ ^2$, no spectral optimization logic)	84.1 ± 0.81	182	21.8	228.6
(a)					
Experimental Group		Final Accuracy (%)	Rounds to 80% Accuracy	$\lambda_{\max}$	κ
FedAvg (baseline)		82.3 ± 1.5	231	24.3	285.7
Fed-SR (gradient ℓ_2 -norm proxy)		85.7 ± 0.7	152	16.8	152.3
FedAvg-DirectTrace		85.2 ± 0.8	158	17.1	156.4

As shown in Table 13, the performance of single gradient variance constraint (Fed-SR onlyvar) or learning rate damping (Fed-SR onlylr) is lower than that of the original method, indicating that a single factor cannot achieve the optimal effect.

The $\lambda_{\max}$ and κ of the original Fed-SR method are significantly better than those of all variants. It is verified that the spectral concentration effect is the core improvement source: through the synergy of constraint gradient variance and learning rate damping, the eigenvalue distribution concentration is realized, rather than the contribution of a single factor.

The spectral index and performance of Fed-SR nospec (no spectral concentration) are weaker than those of the original method. It is proved that the “spectral concentration” design of spectral regularization is the key to distinguishing it from the simple gradient norm constraint.

5.3.5. Isolated Spectral Feature Manipulation and Multi-Factor Ablation Studies

To rigorously examine the associative link between spectral concentration and performance improvement, we designed a comprehensive set of control experiments that systematically isolate each confounding factor. All experiments were conducted on CIFAR-10 with $\alpha = 0.1$, and only the target variable was varied while keeping all other parameters fixed (Table 14).

Key Findings:

- (1) When fixing the learning rate to match FedAvg (eliminating learning rate damping effects), Fed-SR still achieves 2.6% higher accuracy and 28.6% fewer communication rounds than FedAvg, with $\lambda_{\max}$ reduced by 28.0%.

- (2) When forcing the gradient norm to match FedAvg (eliminating gradient variance reduction effects), Fed-SR's performance advantage largely disappears, confirming that spectral concentration via gradient norm constraint is the core mechanism.
- (3) Across all batch sizes, client sampling rates, and local epoch settings, Fed-SR consistently reduces $\lambda_{\max}$ by 25–35% and improves accuracy by 1.9–4.0%, demonstrating that spectral concentration effects are robust to these hyperparameter variations.
- (4) Regression analysis shows that $\lambda_{\max}$ explains 76.2% of the variance in convergence rounds across all tested experimental groups, with other controlled factors collectively accounting for 23.8%. This empirical observation suggests spectral concentration is the dominant correlated factor linked to faster convergence within our experimental scope, rather than proving a universal causal conclusion across all federated scenarios.

Table 14. Multi-factor ablation studies for mechanistic validation.

Experimental Group	Controlled Variable	Fixed Value	Final Accuracy (%)	Rounds to 80% Accuracy	$\lambda_{\max}$	k
Baseline (FedAvg)	None	-	82.3 ± 1.5	231	24.3	285.7
Fed-SR (full method)	None	-	85.7 ± 0.7	152	16.8	152.3
Fed-SR-FixedLR	Learning rate	$\eta = 0.01$ (same as FedAvg)	84.9 ± 0.8	165	17.5	161.2
Fed-SR-FixedGN	Gradient norm	Same as FedAvg	83.2 ± 0.9	198	22.7	258.4
Fed-SR-Batch16	Batch size	B = 16	84.5 ± 0.8	169	18.2	173.5
Fed-SR-Batch64	Batch size	B = 64	86.1 ± 0.6	145	15.9	142.7
Fed-SR-Sample0.3	Client sampling rate	q = 0.3	84.2 ± 0.9	178	18.9	181.6
Fed-SR-Sample0.7	Client sampling rate	q = 0.7	85.4 ± 0.7	156	17.1	158.9
Fed-SR-Epoch1	Local epochs	E = 1	83.8 ± 0.8	185	19.4	192.3
Fed-SR-Epoch5	Local epochs	E = 5	86.3 ± 0.6	142	15.3	138.7

6. Conclusions

This work presents federated spectral regularization (Fed-SR) to address slow convergence caused by statistical heterogeneity in federated learning. Our key contributions are threefold: first, we extend standard federated convergence analysis to our regularized setting, providing convergence bounds under convex assumptions and a stationary point rate for non-convex settings, both aligned with classical optimization theory. We note that the theoretical analysis builds on existing optimization frameworks and serves as supporting justification for the algorithm design. Second, guided by random matrix theory, we design a lightweight gradient norm spectral regularizer to mitigate gradient covariance spectral diffusion. The full Fed-SR pipeline additionally incorporates learning rate damping for training stabilization. Controlled ablation experiments separate the performance contributions of the two components, confirming spectral regularization acts as the primary driver of faster convergence and better spectral concentration. Comprehensive empirical validation across benchmarks demonstrates Fed-SR's consistent superiority over baselines, with experimental observations showing empirical associations between spectral concentration and performance improvements that align with our proposed mechanism. This method also has clear limitations: theoretical assumptions do not fully match non-convex deep learning models; a fixed regularization coefficient cannot adapt to extreme data skew; and the spectral control effect may weaken under practical privacy protection mechanisms.

For future research, we plan to develop adaptive regularization strategies for heterogeneous clients, relax partial theoretical assumptions to fit real neural networks, and optimize the method for scenarios with partial client participation and differential privacy constraints.

Author Contributions: S.C., conceptualization, methodology, software, data curation, writing—original draft, visualization, and supervision; J.B., validation, formal analysis, resources, writing—original draft, visualization, and Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: To improve reproducibility, we release the complete implementation and experimental resources of Fed-SR at: <https://github.com/SYYY3789/Federated-Spectral-Regularization-for-Convergence-Acceleration-A-Random-Matrix-Theory-Perspective> (accessed on 15 July 2026). The repository includes source codes, configuration files, random seed settings, client partitioning scripts, and spectral estimation procedures. These resources enable reproduction of the benchmark comparisons, ablation experiments, convergence analysis, and gradient covariance spectral measurements presented in this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Barbier, J.; Macris, N. Statistical limits of dictionary learning: Random matrix theory and the spectral replica method. *Phys. Rev. E* **2022**, *106*, 024136. [CrossRef] [PubMed]
- Grassi, A.; Komargodski, Z.; Tizzano, L. Extremal correlators and random matrix theory. *J. High Energy Phys.* **2021**, *2021*, 214. [CrossRef]
- Hou, Y.; Bai, B.; Zhao, S.; Wang, Y.; Wang, J.; Li, Z. Federated Dynamic Aggregation Selection Strategy-Based Multi-Receptive Field Fusion Classification Framework for Point Cloud Classification. *Comput. Mater. Contin.* **2025**, *86*, 1–30. [CrossRef]
- Ilbert, R.; Tiomoko, M.; Louart, C.; Odonnat, A.; Feofanov, V.; Palpanas, T.; Redko, I. Analysing multi-task regression via random matrix theory with application to time series forecasting. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 115021–115057. [CrossRef]
- Ge, J.; Liang, Y.C.; Bai, Z.; Pan, G. Large-dimensional random matrix theory and its applications in deep learning and wireless communications. *Random Matrices Theory Appl.* **2021**, *10*, 2230001. [CrossRef]
- Haratemeh, M.H. Portfolio optimization and random matrix theory in stock exchange. *Innov. Manag. Oper. Strateg.* **2021**, *2*, 257–267. [CrossRef]
- Martin, C.H.; Mahoney, M.W. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *J. Mach. Learn. Res.* **2021**, *22*, 1–73. [CrossRef]
- Siva Priya, N.; Shenbagavadivu, N. A brief study on applications of random matrix theory. In *Intelligent Sustainable Systems: Proceedings of ICISS 2021*; Springer: Singapore, 2021; pp. 465–477. [CrossRef]
- Altland, A.; Telles de Miranda, J.; Micklitz, T. Tensor product random matrix theory. *Phys. Rev. Res.* **2024**, *6*, L042029. [CrossRef]
- Vanni, F.; Hitaj, A.; Mastrogiacomo, E. Enhancing portfolio allocation: A random matrix theory perspective. *Mathematics* **2024**, *12*, 1389. [CrossRef]
- Mohammadi, S.; Moradi, P.; Trufanov, A.; Jafari, G.R. A structural approach to detecting opinion leaders in Twitter by random matrix theory. *Sci. Rep.* **2023**, *13*, 21788. [CrossRef] [PubMed]
- Sukumar, N. Universality and Random Matrix Theory. In *Navigating Molecular Networks*; Springer Nature: Cham, Switzerland, 2025; pp. 55–67. [CrossRef]
- Wang, Y.; Feng, S.; Pan, J.; Zheng, C.; Peng, H.; Wang, Y. Random matrix theory-based Wiener postfilter combined with eigenspace and spatial coherence-based minimum variance beamformer for coherent plane-wave adaptive compounding. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–17. [CrossRef]
- Almanifi, O.R.A.; Chow, C.O.; Tham, M.L.; Chuah, J.H.; Kanesan, J. Communication and computation efficiency in federated learning: A survey. *Internet Things* **2023**, *22*, 100742. [CrossRef]
- Hou, Y.; Lv, D.; Liwang, M.; Wang, J.; Li, Z.; Xia, X.; Wu, D.; Tian, Y.; Quek, T.Q.S.; Wang, X. AsyncFed-SAR: Towards Communication-Efficient Asynchronous Federated Learning for SAR Image Target Recognition. *IEEE Trans. Mob. Comput.* **2026**, 1–18. [CrossRef]
- Bickel, P.J.; Levina, E. Covariance regularization by thresholding. *Ann. Stat.* **2008**, *36*, 2577–2604. [CrossRef] [PubMed]
- Deshpande, Y.; Montanari, A. Sparse PCA via covariance thresholding. *Inf. Inference A J. IMA* **2016**, *5*, 414–437. [CrossRef]
- Johnstone, I.M. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **2001**, *29*, 295–327. [CrossRef]

19. Grassetti, F. Optimizing index tracking: A Random Matrix Theory approach to portfolio selection. *Phys. A Stat. Mech. Its Appl.* **2025**, *674*, 130747. [[CrossRef](#)]
20. Norris, C.F.; Maltsev, A.V. Meaningful local signalling in sinoatrial node identified by random matrix theory and PCA. *J. Phys. Complex.* **2023**, *4*, 015003. [[CrossRef](#)]
21. Oliveira, R.M.; Maldonado, D.J.; Batou, A.; Ritto, T.G. Application of random matrix theory combined with the singular value decomposition to journal bearings uncertainty analysis. *J. Braz. Soc. Mech. Sci. Eng.* **2023**, *45*, 213. [[CrossRef](#)]
22. Chandrayadula, T.K. Random matrix theory to predict the intensity fluctuations for a wave propagating through a range-dependent medium. *J. Acoust. Soc. Am.* **2025**, *157*, 4346–4357. [[CrossRef](#)] [[PubMed](#)]
23. Allahyari, N.; Hosseiny, A.; Abedpour, N.; Jafari, G.R. Analyzing the heterogeneous structure of the genes interaction network through the random matrix theory. *Phys. A Stat. Mech. Its Appl.* **2024**, *641*, 129748. [[CrossRef](#)]
24. Camargo, H.A.; Fu, Y.; Jahnke, V.; Kim, K.Y.; Pal, K. Higher-order Krylov state complexity in random matrix quenches. *J. High Energy Phys.* **2025**, *2025*, 1–34. [[CrossRef](#)]
25. Liu, S.; Li, X.; Zhai, Y.; You, C.; Zhu, Z.; Fernandez-Granda, C.; Qu, Q. Convolutional normalization: Improving deep convolutional network robustness and training. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28919–28928. [[CrossRef](#)]
26. Hou, Y.; Zhang, L.; Wang, J.; Li, H.; Duan, X.; Huang, K.; Tian, Y. EfficientSARNet: A Lightweight, High-Performance, and Fast SAR Image Target Recognition Network. *IEEE Sens. J.* **2026**, *26*, 11413–11431. [[CrossRef](#)]
27. Guan, H.; Yap, P.T.; Bozoki, A.; Liu, M. Federated learning for medical image analysis: A survey. *Pattern Recognit.* **2024**, *151*, 110424. [[CrossRef](#)] [[PubMed](#)]
28. Guendouzi, B.S.; Ouchani, S.; Assaad, H.E.; Zaher, M.E. A systematic review of federated learning: Challenges, aggregation methods, and development tools. *J. Netw. Comput. Appl.* **2023**, *220*, 103714. [[CrossRef](#)]
29. Eckstein, S.; Nutz, M. Convergence rates for regularized optimal transport via quantization. *Math. Oper. Res.* **2024**, *49*, 1223–1240. [[CrossRef](#)]
30. Hou, Y.; Zhao, S.; Xia, X.; Liwang, M.; Li, Z.; Xu, N.; Wu, D.; Tian, Y.; Quek, T.Q.S. FedC-DAC: A Federated Clustering With Dynamic Aggregation and Calibration Method for SAR Image Target Recognition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 3726–3745. [[CrossRef](#)]
31. Liu, B.; Lv, N.; Guo, Y.; Li, Y. Recent advances on federated learning: A systematic survey. *Neurocomputing* **2024**, *597*, 128019. [[CrossRef](#)]
32. Pei, J.; Liu, W.; Li, J.; Wang, L.; Liu, C. A review of federated learning methods in heterogeneous scenarios. *IEEE Trans. Consum. Electron.* **2024**, *70*, 5983–5999. [[CrossRef](#)]
33. Hou, Y.; Yu, B.; Yang, Z.; Wang, J.; Xiang, W.; Wu, D.; Liwang, M.; Xia, X.; Li, Z.; Tian, Y.; et al. Privacy-preserving federated SAR image target recognition with adaptive resource management in space-air-ground integrated networks. *Pattern Recognit.* **2026**, *177*, 113253. [[CrossRef](#)]
34. Liu, Y.; Kang, Y.; Zou, T.; Pu, Y.; He, Y.; Ye, X.; Ouyang, Y.; Zhang, Y.; Yang, Q. Vertical federated learning: Concepts, advances, and challenges. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3615–3634. [[CrossRef](#)]
35. Nguyen, D.C.; Pham, Q.V.; Pathirana, P.N.; Ding, M.; Seneviratne, A.; Lin, Z.; Hwang, W.J. Federated learning for smart healthcare: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–37. [[CrossRef](#)]
36. Ali, S.; Li, Q.; Yousafzai, A. Blockchain and federated learning-based intrusion detection approaches for edge-enabled industrial IoT networks: A survey. *Ad Hoc Netw.* **2024**, *152*, 103320. [[CrossRef](#)]
37. Rodríguez-Barroso, N.; Jiménez-López, D.; Luzón, M.V.; Herrera, F.; Martínez-Cámara, E. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Inf. Fusion* **2023**, *90*, 148–173. [[CrossRef](#)]
38. Attouch, H.; László, S.C. Convex optimization via inertial algorithms with vanishing Tikhonov regularization: Fast convergence to the minimum norm solution. *Math. Methods Oper. Res.* **2024**, *99*, 307–347. [[CrossRef](#)]
39. Hou, Y.; Wang, Y.; Xia, X.; Tian, Y.; Li, Z.; Quek, T.Q.S. Toward Secure SAR Image Generation via Federated Angle-Aware Generative Diffusion Framework. *IEEE Internet Things J.* **2025**, *13*, 2713–2730. [[CrossRef](#)]
40. Doikov, N.; Mishchenko, K.; Nesterov, Y. Super-universal regularized Newton method. *SIAM J. Optim.* **2024**, *34*, 27–56. [[CrossRef](#)]
41. Allgeyer, S.; Tregoning, P.; McQueen, H.; McClusky, S.C.; Potter, E.K.; Pfeffer, J.; Montillet, J.P. ANU GRACE Data Analysis: Orbit modeling, regularization and inter-satellite range acceleration observations. *J. Geophys. Res. Solid Earth* **2022**, *127*, e2021JB022489. [[CrossRef](#)]
42. Pieper, K.; Walter, D. Linear convergence of accelerated conditional gradient algorithms in spaces of measures. *ESAIM Control Optim. Calc. Var.* **2021**, *27*, 38. [[CrossRef](#)]

43. Kang, Q.; Fan, Q.; Zurada, J.M. Deterministic convergence analysis via smoothing group Lasso regularization and adaptive momentum for Sigma-Pi-Sigma neural network. *Inf. Sci.* **2021**, *553*, 66–82. [[CrossRef](#)]
44. Aich, A.; Aich, A.B.; Wade, B. Convergence guarantees for gradient descent in deep neural networks with non-convex loss functions. *Int. J. Comput. Math.* **2025**, *102*, 1808–1823. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.