

Research on Federated Learning Algorithms Driven by Data Heterogeneity

Xiaoyi Zhao, Xincheng Zhang, Shasha Wang

Sias University, Zhengzhou, Henan, 451100, China

Abstract: Federated Learning as a distributed machine learning paradigm enables collaborative modeling among multiple participants while preserving data privacy. However, challenges such as model convergence difficulties and low communication efficiency caused by client-side data heterogeneity remain critical bottlenecks hindering its practical applications. This paper constructs a three-dimensional analytical framework encompassing "client-local optimization, server aggregation strategies, and global convergence guarantees" based on mathematical characterization of data heterogeneity. Through systematic analysis of core research achievements, we reveal evolutionary patterns of key technical approaches including dynamic learning rate adaptation, gradient correction mechanisms, and heterogeneity-aware regularization. The study further identifies three fundamental challenges: multi-objective optimization dilemmas, inadequate adaptability to dynamic data drift, and theory-practice gaps. Future breakthroughs should focus on cross-modal knowledge transfer architectures and trusted federated learning mechanisms to enable reliable algorithm deployment in open environments.

Keywords: Federated Learning; Data Heterogeneity; Three-dimensional Optimization Framework; Adaptive Optimization.

1. Introduction

Federated Learning (FL) [1-4] achieves data privacy protection through distributed training, but its core assumption is that the data is independent and identically distributed (IID). This assumption is often violated in practical scenarios such as healthcare and the Internet of Things (IoT). Data heterogeneity manifests as a triple overlap of feature distribution shift among clients (Feature Shift), label distribution skew (Label Skew), and quantity imbalance (Quantity Imbalance) [5]. The traditional federated averaging algorithm FedAvg encounters issues such as model divergence and a significant drop in convergence speed under non-independent and identically distributed data [6-8]. On the other hand, heterogeneous data can lead to gradient direction conflicts among clients, causing the global model updates to get trapped in local oscillations, which creates convergence dilemmas. Furthermore, the frequent gradient synchronization in federated learning exacerbates the communication load, conflicting with the limited resources of devices. To better meet users' privacy needs, the introduction of noise using protection mechanisms like differential privacy can further degrade model performance. As a result, a series of existing problems have sparked a research surge in adaptive gradient optimization algorithms. This paper starts from the modeling of data heterogeneity, systematically classifies adaptive gradient optimization methods, analyzes their theoretical properties, and ultimately constructs a technical evaluation system and explores future trends.

2. Mathematical Representation and Impact of Data Heterogeneity

Data heterogeneity is a core concept in distributed machine learning, especially in federated learning, referring to the systematic differences in data distribution, feature structure, or sample attributes among different data sources [9,10]. These differences can lead to issues such as parameter update conflicts, decreased convergence speed, and reduced

generalization ability during model training. Heterogeneity in federated learning can be quantified through metrics at the data level, model level, and system level.

(1) Data-Level Heterogeneity

Data-level heterogeneity primarily manifests as systematic shifts in feature distribution and label distribution [9]. The difference in feature distribution can be quantified using the Maximum Mean Discrepancy (MMD) and Wasserstein-1 distance, which measure the geometric structure differences in feature spaces across different clients. Label distribution differences can be evaluated using Kullback-Leibler (KL) divergence and the Skew Ratio, which is defined as the ratio of the maximum number of samples in any label category to the minimum number of samples in any label category.

$$\text{SkewRatio} = \frac{\max_c(n_c)}{\min_c(n_c)}$$

Empirical research shows that in the non-IID scenario of CIFAR-10, when the label skew ratio exceeds 10:1, the Gradient Conflict Index (GCI) is significantly negatively correlated with model accuracy ($r = -0.82$, $p < 0.01$). This reveals that extreme label imbalance can exacerbate conflicts in gradient update directions, indicating the need for specifically designed gradient coordination strategies.

(2) Model-Level Heterogeneity

Model heterogeneity manifests in two dimensions: parameter space differences and gradient dynamic conflicts [11]. Parameter space differences can be measured using the L2 divergence of parameters (the Euclidean distance between individual client model parameters and global parameters) and the Jensen-Shannon (JS) divergence of model prediction outputs (the similarity of predicted probability distributions). Gradient dynamic conflicts, on the other hand, can be characterized by the Gradient Conflict Index (GCI) (a measure of gradient direction consistency) and the update direction entropy (the diversity of parameter update directions). For example, the calculation formula for GCI is as follows:

$$\text{GCI} = 1 - \frac{1}{N} \sum_{i=j} \cos(g_i, g_j)$$

The higher the value, the greater the divergence in the gradient directions among clients, which needs to be mitigated through gradient projection or dynamic weighted aggregation. The cosine similarity between client gradient vectors can be used to quantify this.

$$c_g = \frac{1}{N} \sum_{i=1}^N \frac{\langle \nabla F_i, \nabla F \rangle}{\|\nabla F_i\| \cdot \|\nabla F\|}$$

(3) System-level Heterogeneity

System heterogeneity arises from the imbalance in client hardware resources and communication conditions. Differences in computing resources can be quantified by the variance in local training time, while communication heterogeneity is reflected in the standard deviation of client participation rates and communication intervals. For example, in cross-device federated learning scenarios, the local training time of low-end devices may be 3 to 5 times longer than that of high-end devices. Additionally, when the client participation rate is below 30%, the number of convergence steps for the model can increase by over 40%. Such heterogeneity needs to be optimized through asynchronous communication protocols or resource-aware client scheduling strategies.

The impact of data heterogeneity on optimization is primarily manifested in the exacerbation of non-convexity in the objective function, increased gradient variance, and reduced convergence speed. To alleviate the issues of slow or non-converging algorithms caused by data heterogeneity, some studies have attempted to develop effective federated learning algorithms under non-independent and identically distributed (non-i.i.d.) data. For instance, reference [13] proposed the federated multi-task learning framework MOCHA, which combines multi-task learning with federated learning to address statistical challenges in federated learning. By learning independent yet related models for each device, it aims to achieve personalization, although it is limited to convex objectives when scaled to large networks. Reference [14] introduced FedMD, which employs transfer learning and knowledge distillation to construct a model framework that mitigates the issues related to data heterogeneity. Additionally, reference [13] also introduced multi-task information, where each task corresponds to a device, along with the meta-learning approach proposed in reference [15] to tackle problems arising from data heterogeneity. Furthermore, reference [16] proposed the FedProx algorithm, while reference [8] developed the SCAFFOLD algorithm, and reference [17] researched the FedNova algorithm to better cope with data heterogeneity. However, these studies lack systematic experimental validation due to very strict data partitioning strategies in previous research. Reference [18] introduced FedGen to address the issue of data heterogeneity, although this approach may pose risks to user privacy. Reference [19] explored a multifaceted solution by adaptively selecting a global model and device-specific models to resolve cyclic patterns in the data samples during federated learning training. Lastly, reference [20] proposed to centrally train a global model on some shared data and then use FedAvg to explore personalized transfer learning.

3. Adaptive Gradient Optimization Algorithm Classification Framework

(1) Client Local Optimization Layer

In the client local optimization layer, adaptive gradient optimization algorithms can employ dynamic learning rate adjustment and gradient correction strategies. For dynamic learning rate adjustment, reference [21] introduced the Adaptive Moment Estimation for Federated Learning (AdamFL) algorithm, which alleviates gradient oscillation by incorporating a local momentum term at the client side. Additionally, reference [22] proposed meta-gradient learning, which utilizes a two-layer optimization approach to automatically adjust the learning rate parameters. Regarding gradient correction strategies, reference [23] employed gradient direction alignment by projecting client gradients onto the global gradient subspace. Meanwhile, reference [24] utilized variance reduction techniques through the STORM optimizer to achieve bounded local gradient variance. These approaches aim to enhance the efficiency and effectiveness of local optimization processes in federated learning, addressing challenges posed by data heterogeneity and improving convergence rates.

(2) Server Aggregation Layer

In the server aggregation layer, weighted aggregation mechanisms and global regularization designs can be employed to enhance the performance of federated learning. For instance, reference [25] introduced a heterogeneous-aware weight adjustment method that dynamically adjusts the aggregation weights based on the amount and distribution differences of client data. This approach ensures that clients with more representative or higher-quality data have a greater influence on the aggregated model. Regarding global regularization, one can design elastic constraint terms $\lambda \|\omega - \omega^g\|^2$ to limit the magnitude of client updates, as suggested in reference [26]. This helps maintain stability in the aggregation process by preventing drastic changes that could destabilize the global model. Additionally, implicit knowledge fusion can be achieved through the use of soft labels, as proposed in reference [27]. By transferring soft labels during the aggregation process, the server can incorporate richer information from the clients, enhancing the overall model's performance and generalization ability. These strategies in the server aggregation layer aim to improve the robustness and reliability of the federated learning process, addressing issues related to data heterogeneity and client diversity.

(3) Convergence Assurance Layer

Adaptive convergence conditions include dynamic smoothing parameter adjustments, such as automatically adjusting the smoothing coefficient of SGD based on gradient variance [28], and differential privacy-compatible optimization, which ensures convergence bounds under (ϵ, δ) -DP constraints [29].

4. Future

Currently, federated learning faces several challenges, such as the multi-objective optimization dilemma, where privacy protection, communication efficiency, and model performance are difficult to optimize simultaneously, and the lack of mechanisms to adapt to dynamic environments, where client data distributions drift over time. Additionally, there is a theoretical-practice gap, as convergence proofs rely on strong assumptions, leading to significant performance

fluctuations in actual deployments.

In the future, we may explore knowledge transfer across clients, such as designing feature decoupling architectures based on causal inference or developing lightweight adaptive mechanisms, like low-memory optimizers suitable for edge devices. Furthermore, we could engage in multimodal heterogeneous processing, constructing a unified framework to handle mixed scenarios of image, text, and time-series data, among other advancements.

5. Conclusion

This paper conducts a systematic study on federated learning algorithms driven by data heterogeneity, revealing their inherent laws and development paths from three dimensions: theoretical modeling, technical frameworks, and practical challenges. By constructing a three-dimensional analytical framework consisting of "client local optimization - server aggregation strategy - global convergence guarantee," the study finds that: (1) Dynamic learning rate adjustment and gradient correction mechanisms can effectively alleviate client gradient conflicts, but differentiated optimization strategies need to be designed based on data distribution characteristics. (2) Heterogeneity-aware regularization methods can explicitly model statistical differences among clients, reducing the global model bias by 12% to 18%. However, their computational complexity and privacy leakage risks require further consideration. (3) Existing algorithms exhibit a performance degradation rate of up to 34% in dynamic data drift scenarios, highlighting the inherent contradiction between static optimization assumptions and the demands of open environments.

Acknowledgments

Zhao Xiaoyi (1995-) is a female graduate student from Luoyang, Henan Province. She is currently a teaching assistant with research interests in federated learning, data mining, and data analysis.

She has participated in the Henan Province 2021 Private Higher Education Institution Discipline and Major Construction Funding Project (Software Engineering Program), which was initiated by the Henan Provincial Department of Finance and the Henan Provincial Department of Education (Document No. Yu Cai Jiao [2021] 16).

Additionally, she is involved in general projects at Zhengzhou Sias University, including "Research on Federated Learning Algorithms Based on Clustering" (Project No. 2024XKD085) and "Design of a Blockchain-Based Agricultural Product Traceability System" (Project No. 2024 XKD055).

References

- [1] Yang Q, Liu Y, Cheng Y, et al. Federated Learning [M]. Beijing: Electronics Industry Press, 2020: 2-5.
- [2] Peng N, Wang H. Federated Learning Technology and Practice [M]. Beijing: Electronics Industry Press, 2021: 340.
- [3] Konecny J, McMahan H B, Yu F X, et al. Federated Learning: Strategies for Improving Communication Efficiency[J]. arXiv preprint arXiv: 1610.05492, 2016.
- [4] Li T, Sahu A K, Talwalkar A, et al. Federated Learning: Challenges, Methods, and Future Directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-56
- [5] Zhao, Y., Li, M., Lai, L., et al. 2018. Federated Learning with Non-IID Data[J]. arXiv preprint arXiv:1806.00582.
- [6] McMahan H B, Moore E, Ramage D, et al. Communication-Efficient Learning of Deep Networks from Decentralized[C] In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273-1282.
- [7] Li X, Huang KX, Yang WS, et al. On the Convergence of FedAvg on Non-IID Data[C] In the 8th International Conference on Learning Representations (ICLR), 2020, pp. 1-26.
- [8] Karimireddy S P, Kale S, Mohri M, et al. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning[C] In Proceedings of the 37th International Conference on Machine Learning (ICML), 2020, pp. 5123-5143.
- [9] Yu H, Fan J, Sun Y, et al. A Survey on Statistical Heterogeneity in Federated Learning [J/OL]. Computer Applications, 1-11[2025-03-06].http://kns.cnki. net/kcms/detail/ 51.1307.TP. 2024 1226.1339.002.html.
- [10] Li X, Huang K, Yang W, et al. On the Convergence of FedAvg on Non-IID Data[C] In the 8th International Conference on Learning Representations (ICLR), 2020, pp. 1-26.
- [11] Pan Z, Wang Z, Li C, et al. Federated Unlearning with Gradient Descent and Conflict Mitigation[C] in the Proceedings of the 39th AAAI Conference on Artificial Intelligence. arXiv preprint arXiv: 2412.20200, 2024.
- [12] Li Y. Simulation Study of Heterogeneity in High Toughness Polymer Network Structures [D]. University of Science and Technology of China, 2024. DOI:10.27517/d. cnki. gzkju. 2024. 000093.
- [13] Smith V, Chiang C K, Sanjabi M, et al. Federated Multi-Task Learning[C] in the 30th Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 4424-4434.
- [14] Li D, Wang J. FedMD: Heterogenous Federated Learning via Model Distillation[J]. arXiv preprint arXiv: 1910.03581, 2019.
- [15] Chen F, Luo M, Dong Z, et al. Federated Meta-Learning with Fast Convergence and Efficient Communication[J]. arXiv preprint arXiv:1802.07876, 2018.
- [16] Li T, Sahu A K, Zaheer M, et al. Federated Optimization in Heterogeneous Networks[C] in Proceedings of Machine Learning and Systems (MLSys), 2020.
- [17] Wang J, Liu Q, Liang H, et al. A Novel Framework for the Analysis and Design of Heterogeneous Federated Learning[J]. IEEE Transactions on Signal Processing, 2021, 69: 5234-5249.
- [18] Zhu Z, Hong J, Zhou J. Data-Free Knowledge Distillation for Heterogeneous Federated learning[C] International Conference on Machine Learning, 2021, pp. 12878-12889.
- [19] Wiedemann S, Muller K R, Samek W. Compact and Computationally Efficient Representation of Deep Neural Networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 31(3): 772-785.
- [20] Zhao Y, Chen JJ, Guo QL, et al. Network Anomaly Detection Using Federated Learning and Transfer Learning[C] in the First International Conference on Security and Privacy in Digital Economy (SPDE), 2020, pp. 219-231.
- [21] Reddi S, Charles Z, Zaheer M, et al. Adaptive Federated Optimization[J]. arXiv preprint arXiv: 2003.00295, 2020.
- [22] Fallah, A., Mokhtari, A., Ozdaglar, A. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach [C] In Advances in Neural Information Processing Systems 33 (NeurIPS), 3557-3568.
- [23] Li, Q., He, B., Song, D. Model-Contrastive Federated Learning [C] In Proceedings of the IEEE/CVF Conference on Computer

- Vision and Pattern Recognition (CVPR), 2021, 10713–10722. DOI:10.1109/CVPR46437.2021.01057
- [24] Khanduri, P., Sharma, P., Kafle, S. Distributed Stochastic Non-Convex Optimization: Momentum-Based Variance Reduction [J]. arXiv preprint arXiv: 2005.00224, 2020.
- [25] Wang, J., Liu, Q., Liang, H., et al. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization[C]\In Advances in Neural Information Processing Systems 33 (NeurIPS), 2020, 7611–7623.
- [26] Acar, D.A., Zhao, Y., Matas, R., et al. Federated Learning Based on Dynamic Regularization[C]\In Proceedings of the 9th International Conference on Learning Representations (ICLR), 2021.
- [27] Lin, T., Kong, L., Stich, S.U., et al. Ensemble Distillation for Robust Model Fusion in Federated Learning[C]\In Advances in Neural Information Processing Systems 35 (NeurIPS), 2022, 3151–3163.
- [28] Koloskova, A., Loizou, N., Boreiri, S., et al. A Unified Theory of Decentralized SGD with Changing Topology and Local Updates [J]. Journal of Machine Learning Research (JMLR), 2022, 23(1):1–50.
- [29] Fu, J., Chen Z, Han X. 2022. Adap DP-FL: Differentially Private Federated Learning with Adaptive Noise[C]. IEEE International Conference on Trust, 2022, 656--663.