NIRo: A Metric to capture non-IID robustness for Federated Learning Algorithms

Anupam Gupta
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
anupamguptacal@kgpian.iitkgp.ac.in

Pabitra Mitra
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
pabitra@cse.iitkgp.ac.in

Abstract

Federated Learning (FL) is a collaborative machine learning framework for decentralized nodes with non-IID(non- independent, identically, distributed) distributed private data, to create a globally un-biased, high performing central model. A majority of the proposed FL protocols report performances in varied non-IID settings. Heterogeneity in non-IID descriptions in each protocol paper makes it very hard to compare the robustness of approaches to other studied approaches in differing settings. In this paper, we define a metric, NIRo, to capture data-quantity and data label-skewness and use it to propose a cumulative area-under-the-curve metric that can be used to quantify the robustness of FL protocols in varied non-IID settings.

1 Introduction

Federated learning (FL) in non-IID setting has been widely studied in literature both theoretically (Wang et al., 2022; Ye et al., 2023) and empirically (Li et al., 2022). Various data benchmarks and metrics (Caldas et al., 2018; Yin et al., 2023; Haller et al., 2023) have been proposed for evaluation, each generating specific set of label, feature, and quantity skews. We propose a quantifiable method for exploring the entire skew spectrum and also an area-under-curve performance metric.

Consider the setting of a K - class classification problem $(C_1, \ldots, C_K)$ and a group of N decentralized nodes $(N_1, \ldots, N_N)$. Each node, N_i, holds a private dataset, D_i, comprising of individual class-labelled data points C_k^i s.t. $\sum_{j=1}^{j=K} C_j^i = D_i$. FL is usually studied in the following two non-IID settings: (1) **Label skewness** - Each of the C_k^i are distributed in a non-IID manner, within each D_i. (2) **Data quantity skewness** - $|D_1|, .., |D_N|$ are distributed in a non-IID manner. Commonly, real-world non-IID conditions are a combination of both of these non-IID settings. In most FL works, protocol robustness to non-IID settings is generally studied in *specific, simulated* non-IID data distributions(Gao et al. (2022); Zhang et al. (2021); Li et al. (2021b), which leads to significant difficulty in comparing the non-IID robustness of the protocols with each other, without complete experiment repetition. We attempt to suggest a comparison metric across various non-IID data distributions to facilitate easy comparison of non-IID robustness between protocols.

2 Comparison Metric (Non-IID Robustness) - NIRo

We define the metric NIRobust (NIRo), to 'measure' the degree of non-IID-ness in data distribution across the nodes:

$$NIRo(D_1, \ldots, D_N) = \frac{Cvar}{Cvar + DSvar}\dot{C}var + \frac{DSvar}{Cvar + DSvar}\dot{D}Svar,$$

where $DSvar = \frac{\sigma^2(|D_1|,\ldots,|D_N|)}{\sigma^2(NI_{max}^{Count}(|D_1|,\ldots,|D_N|))}$ and $Cvar = \Sigma_{i=1}^{i=N} \frac{|D_i|}{|D|} \times \frac{\sigma^2(|C_1^i|,\ldots,|C_K^i|)}{\sigma^2(NI_{max}^{Class}(|D_i|))}$.

Here, NI_{max}^{Count} is a vector of all $N-1$ ones in a single $|D|-|N|$ term. Similarly, NI_{max}^{Class} is a vector with $K-1$ zeros and a single $|D_i|$ term. They represent the highest variance permutation of each data-skew.

The NIRo metric ($\in [0, 1]$) provides a single estimate of the level of non-IID-ness in any given data distributed. A NIRo value of 0 corresponds to a IID setting i.e. equitable data-count distribution and within each node, an equitable label distribution and a NIRo metric of 1 corresponds to an extreme non-IID setting - i.e., a single data point in all nodes and a large dataset in one of the nodes or an equitable distribution of data points across nodes but a completely skewed label distribution within each node. The combination of both these extreme cases also yields a NIRo value of 1.

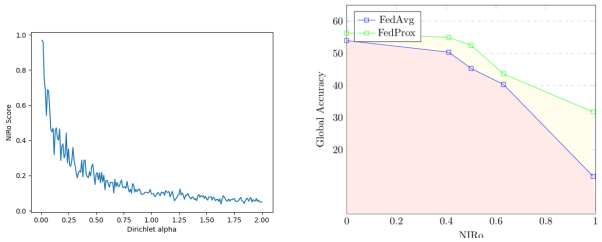
2.1 NIRo Area Under Curve (AuC)

In order to get a holistic perspective of algorithmic non-iid robustness, we quantify the performance of protocols across multiple non-IID levels and combine them to measure robustness through the proposed Area under the curve (NIRo-AuC) metric. We plot the global test accuracy and corresponding NIRo metrics for various non-IID settings in multiple runs. The area under this curve provides us with a single measure of unified non-IID performance for the specific protocol.

2.2 Experiment and Discussion

To generate non-IID partitions, use the Dirichlet distribution (over label, and volume distributions) with parameter $\alpha \in [0, 2.0]$ (higher is more IID). We rely on the numpy implementation of Dirichlet distribution and use NIID Bench (Li et al., 2022) in order to generate data partitions. We plot the computed NIRo against the α parameter in Figure 1(a) and find that as the parameter α draws higher (closer to IID), NIRo converges to a value close to 0, signifying minimal non-IID ness. When α is closer to 0 (high non-IIDness), we note a peak towards 1 in NIRo.

The NIRo-AuC is studied in Figure 1(b) for the FedAvg (McMahan et al., 2017) and the FedProx (Li et al., 2020) protocols on the CIFAR-10 dataset (12 nodes, 0.3 random client participation, $\mu_{FedProx} = 3$, Global Model Accuracy). As is known from literature, the FedProx protocol is more non-IID robust i.e. has a better NIRo-AuC, than the FedAvg protocol. We see the same trend with our NIRo-AuC as well. In addition, we can now also quantifiably infer that FedProx (area = 54.49) is 11% better at performing on non-IID distributions than FedAvg (area = 48.48) on average.



(a) (b)

Figure 1: (a) NIRo variation with different α parameters, (b) NIRo@5epoch vs global accuracy AuC for various non-IID data distribution (FedAvg vs FedProx) under Dirichlet($\alpha = 500.0, 0.055, 0.05, 0.045, 0.00001$). The @epoch suffix denotes the number of FL epochs.

3 Conclusion

In this paper, we address the issue of heterogenous non-IID settings that are explored for federated learning use-cases. We present the Non-IID Robustness (NIRo) metric for comparing the non-IID robustness of different FL protocols with a quantifiable measure (Area Under Curve).

URM Statement

The authors acknowledge that the first author of the paper meets the URM criteria for ICLR Tiny Track 2024 as they are non-white, of age 26, not a resident of North America, Western Europe, UK, or East Asian countries and are a first time submitter to the conference.

REFERENCES

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction, 2022.

Marc Haller, Christian Lenz, Robin Nachtigall, Feras M. Awayshehl, and Sadi Alawadi. Handling non-iid data in federated learning: An experimental evaluation towards unified metrics. In *IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 0762–0770, 2023. doi: 10.1109/DASC/PiCom/CBDCom/Cy59711.2023.10361408.

Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. 2019. URL https://arxiv.org/abs/1909.06335.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning, 2021.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study, 2021a.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 965–978, 2022. doi: 10.1109/ICDE53745.2022.00077.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization, 2021b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282, Fort Lauderdale, FL, 2017. PMLR, PMLR.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization, 2020.

Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022.

Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.*, 56(3), oct 2023. ISSN 0360-0300.

Kangning Yin, Zhen Ding, Zhihua Dong, Dongsheng Chen, Jie Fu, Xinhui Ji, Guangqiang Yin, and Zhiguo Wang. Nipd: A federated learning person detection benchmark based on real-world non-iid data. *arXiv preprint arXiv:2306.15932*, 2023.

Lin Zhang, Yong Luo, Yan Bai, Bo Du, and Ling-Yu Duan. Federated learning for non-iid data via unified feature learning and optimization objective alignment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4400–4408, 2021. doi: 10.1109/ICCV48922.2021.00438.

A TECHNICAL BACKGROUND AND INTRODUCTION

Federated Learning(FL) is a decentralized privacy-preserving machine learning framework. Introduced in 2016 by Google (McMahan et al., 2017), FL provides a framework to generate a high-performing global model by combining learnings from individually distributed local data-nodes. The traditionally proposed FL framework relied on using a weighted average of individual local gradients approach, termed as FedAvg, to generate global gradients and support iterations of global models along with globally synchronized local models. An active area of research in FL has been the performance of protocols in different data-distribution settings - specifically in IID (Independent, Identically distributed) and non-IID settings, since this is of particular importance in decentralized settings. In recent years, a thorough understanding of FL has resulted in the development of many non-IID distribution robust algorithms. Some of these include FedProx (Li et al., 2020), SCAFFOLD (Karimireddy et al., 2021) and FedNova (Wang et al., 2020), to name a few.

Each FL algorithm tackles the issue of non-IID robustness with a novel approach. In each such approach, a baseline set of protocols are defined and used in the introductory paper to measure the comparative robustness of the novel approach against previously introduced protocols (Wang et al., 2020). Traditionally, non-IID data-distributions (for testing) are generated by splitting the original dataset based on samples drawn from a Dirichlet distribution (Li et al., 2021a; Hsu et al., 2019; Gao et al., 2022), that can be fine tuned via an input alpha parameter (closer to 1 is more IID). However, most papers study and report robustness on specific non-IID distributions, that the authors explore and evaluate against. Given the variability in testing datasets (and distributions) per paper, a direct translation of FL protocol performance, without complete experiment reproduction, is infeasible. Contrary to metrics like accuracy and convergence, that can be directly compared in IID settings (since a model will learn almost similarly if trained across multiple IID samples of a distributions), the same is not true in non-IID cases, since the level of non-IID ness is an essential determinant of performance in a distributed learning case. In order to accurately compare performance of protocols, the underlying testing conditions of the protocols must be the same, similar or translatable across experiments.

On a similar thread, in order to aid comparative performance analysis, a better metric would be one that could capture the wholistic performance of protocols across multiple non-IID settings. In most real-world conditions, a spectrum-wide analysis is more informative and indicative of non-IID robustness than single-point analysis.

In this paper, we present a step in this direction by introducing the NIRo metric, which allows us to gather the level of non-IID-ness of any given data-distribution in terms of data-count heterogeneity as well as label heterogeneity and represent it with a single numeric metric. Given our representation of non-IID-ness with a single variable with a set, defined range (between 0 and 1), we can then evaluate the performance of protocols across the range of the variable and be able to gather an overall robustness analysis of any given protocol as well. Across papers, we can further use this to compare and contrast with other protocols that report their results on the same scale as well, regardless of whether the datapoint distribution across both the experimental settings is exactly the same. We propose the Area Under NIRo Curve (AuC) metric allows us this freedom and provides us with a comparable metric of wholistic non-IID robustness performance across multiple protocols and test settings.