Meituan

中国科学院软件研究所
Institute of Software Chinese Academy of Sciences

Elastic Aggregation for Federated Optimization

Dengsheng Chen, Jie Hu, Vince Junkai Tan, Xiaoming Wei, Enhua Wu

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

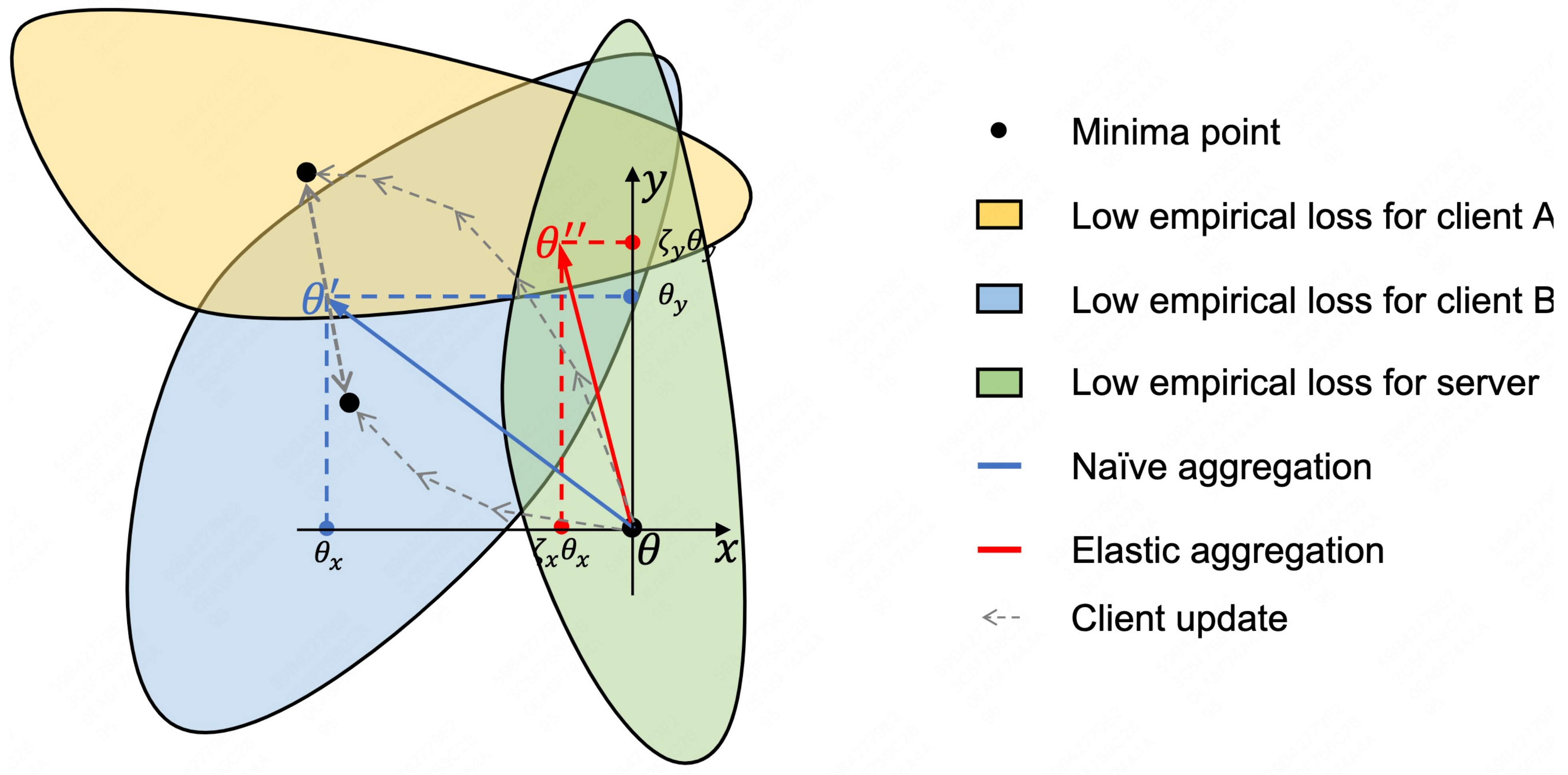
Introduction

- Federated learning enables the privacy-preserving training of neural network models using real-world data across distributed clients. FedAvg has become the preferred optimizer for federated learning because of its simplicity and effectiveness. FedAvg uses naïve aggregation to update the server model, interpolating client models based on the number of instances used in their training. However, naïve aggregation suffers from client drift when the data is heterogenous (non-IID), leading to unstable and slow convergence.
- We propose a novel aggregation approach, **elastic aggregation**, to overcome these issues. Elastic aggregation interpolates client models **adaptively according to parameter sensitivity**, which is measured by computing how much the overall prediction function output changes when each parameter is changed. This measurement is performed in an unsupervised and online manner.
- Elastic aggregation **reduces** the magnitudes of updates to the **more sensitive parameters** so as to prevent the server model from drifting to any one client distribution, and conversely **boosts** updates to the **less sensitive parameters** to better explore different client distributions.

Algorithm 1: Elastic aggregation within a single layer

A variable with a superscript i indicates the i^{th} element of the variable. A variable with a subscript k indicates the variable from k^{th} client. η, η' are learning rates of server and clients respectively. μ, τ are the hyper-parameters. $\theta, \theta_k \in \mathbb{R}^n$ are the server's and the k^{th} client's parameters respectively. $\Omega \in \mathbb{R}^n$ is the aggregated parameter sensitivity. $\Omega_k \in \mathbb{R}^n$ is the parameter sensitivity on the k^{th} client.

```
Initialize θ
B_k ← Sample a subset of training data D_k.
D_k ← Drop the samples of B_k from D_k.
for each round do
    for each activated client k do
        Initialize Ω_k as zeros.
        for each batch data x ∈ B_k do
            g = ∇||F(θ; x)||_2^2
            for i ∈ [1, ··· , n] do
                Ω_k^i ← μΩ_k^i + (1 − μ)|g^i|
        θ_k ← θ
        for each epoch do
            for each batch data x ∈ D_k do
                θ_k ← θ_k − η'∇ℓ_k(F(θ_k; x))
        Δ_k = θ_k − θ
    w_k ← |D_k| / Σ_k |D_k|; Ω = Σ_k (w_k · Ω_k);
      Ω' = max(Ω)
    for i ∈ [1, ··· , n] do
        ζ^i = 1 + τ − Ω^i/Ω'
        Δ^i = ζ^i · Σ_k (w_k · Δ_k^i)
        θ^i ← θ^i − η · Δ^i
```

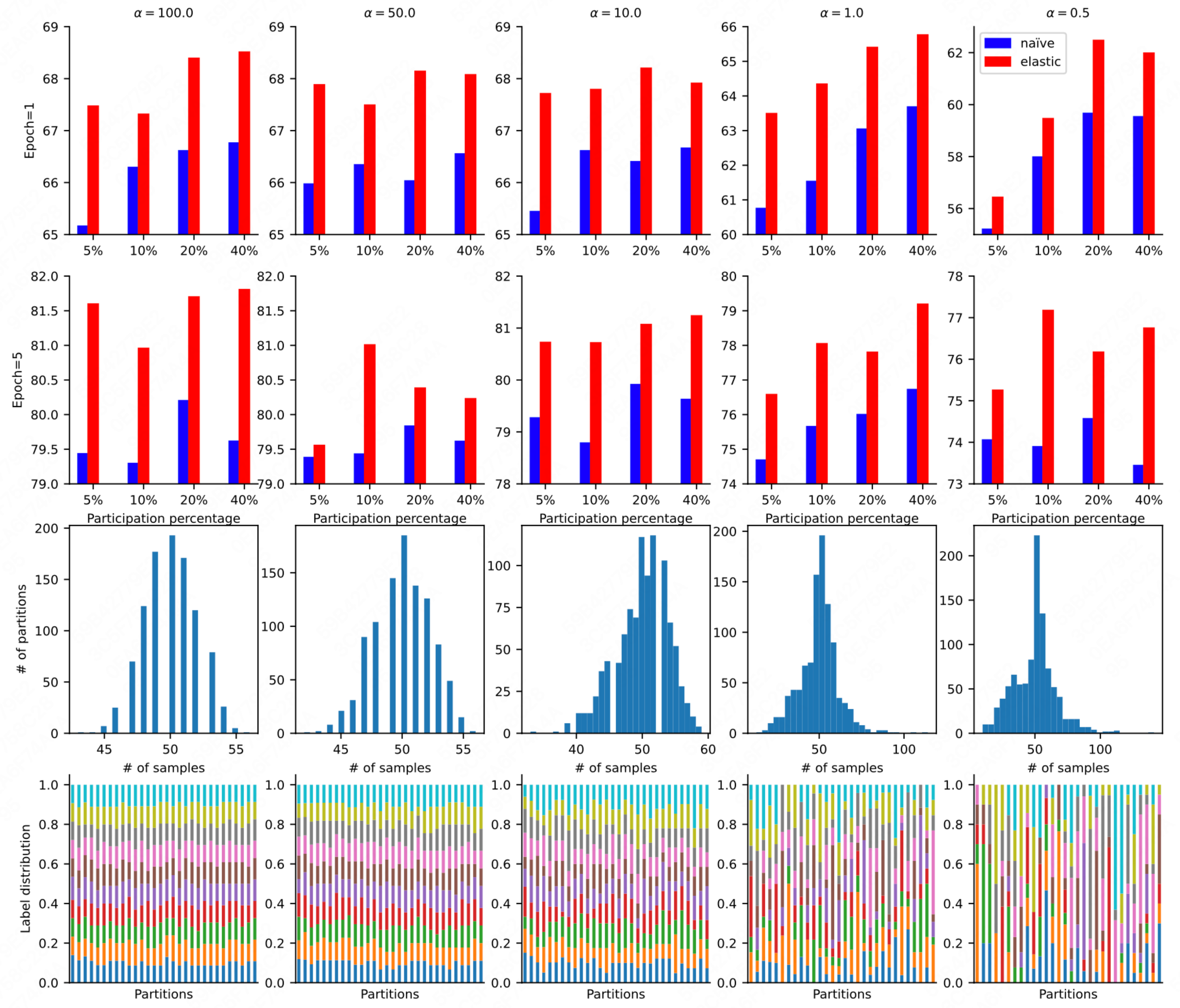


Elastic Aggregation

- The local updates of client A and client B drive the server model θ towards their individual minima (black dots in plot).
- Naïve aggregation simply averages the received model from clients A and B, yielding θ' as the new server model.
- Although θ' minimizes the local empirical loss of clients A and B, θ' drifts from ideal distribution for the server model.
- Elastic aggregation adjusts gradient with respect to parameter sensitivity which results in a better update θ''.
  - Parameter θ_x is more sensitive (has a larger gradient norm), and is restricted with $\zeta_x < 1$ to reduce the magnitude of its update.
  - Parameter θ_y is less sensitive (has a smaller gradient norm), and is correspondingly boosted with $\zeta_y > 1$ to better explore the parameter space.
- Elastic aggregation minimizes the loss for clients A and B, while not causing the server model to drift from its ideal distribution.

Experiments

Dataset	Rounds(Epochs)	Total	Sampled	Batch	Init. LR	Model	Naïve(%)	Elastic(%)
Balanced data across clients								
CIFAR-100	4000(~80)	500	10	10	0.05	ResNet-20	32.31	**56.64**
Unbalanced data across clients								
MNIST	20(~2)	1000	100	100	0.1	Logistic Regression	70.14	**73.64**
EMNIST	1000(~2.5)	3400	10	100	0.1	2Conv+2Linear	88.71	**89.82**
CIFAR-10	4000(~40)	1000	10	10	0.05	ResNet-20	66.84	**68.74**
CINIC-10	200(~20)	1000	100	10	0.05	ResNet-20	35.81	**36.29**
CINIC-10	4000(~40)	1000	10	10	0.05	ResNet-20	68.68	**69.25**



- (Upper) Table shows test accuracies for various datasets. elastic aggregation achieves superior performances on different datasets and settings.
- (Left) Figure illustrates that elastic aggregation achieves significant improvements across different partitioning distributions, participation rates and numbers of local epochs.

Conclusion

- We proposed a novel aggregation method, elastic aggregation, which utilizes parameter sensitivity to overcome gradient dissimilarity.
- We are the pioneers in utilizing unlabeled client data to enhance federated learning performances.
- Empirical evaluations across various federated datasets validate the theoretical analysis and reveal that elastic aggregation can significantly enhance the convergence behavior of federated learning in realistic heterogeneous scenarios.