Understanding Federated Learning through Loss Landscape Visualizations: A Pilot Study

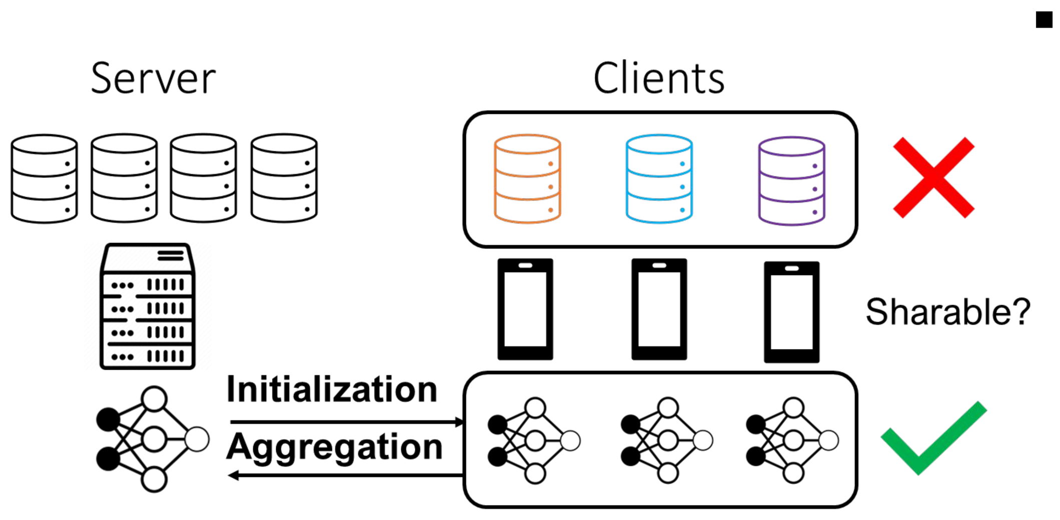
Ziwei Li, Hong-You Chen, Han-Wei Shen, Wei-Lun Chao

The Ohio State University

Highlights

- We extend the visualization techniques of 2D loss surface and optimization trajectories to understand federated learning for both global and local scope.
- We visually demonstrate the phenomenon of model drifting, the effect of data heterogeneity and model initialization.
- With proper initialization, the trajectories under different non-IID degrees would enter the same loss basin, which provides an explanation of why pre-training could largely improve FL.

Background



- **Federated Learning (FL)**
  - M clients, each client holds a data set $\mathcal{D}_m = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{|\mathcal{D}_m|}$
  - The optimization problem formulated as:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \mathcal{L}_m(\boldsymbol{\theta}),$$

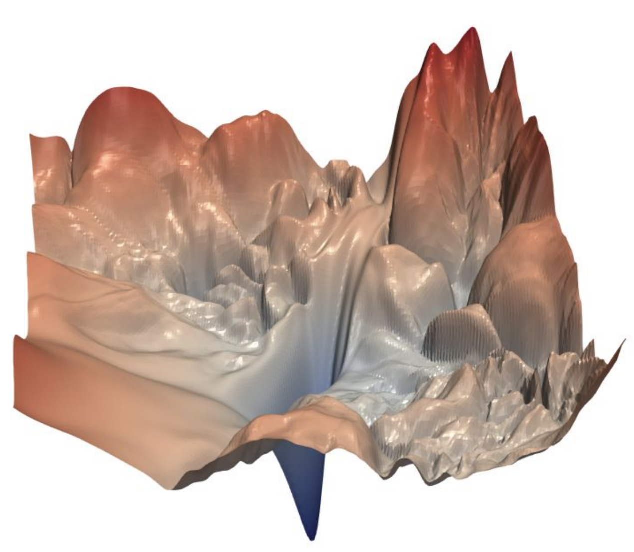
where $\mathcal{L}_m(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_m|} \sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{\theta})$

- **Federated averaging (FedAvg)** [1]
  - Local training and global aggregation, for multiple rounds of communication (indexed by t)

Local: $\boldsymbol{\theta}_m^{(t)} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_m(\boldsymbol{\theta})$, initialized with $\bar{\boldsymbol{\theta}}^{(t-1)}$; **Global:** $\bar{\boldsymbol{\theta}}^{(t)} \leftarrow \sum_{m=1}^{M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \boldsymbol{\theta}_m^{(t)}$

- **Visualization of loss landscape**
  - A powerful technique to analyze the high-dimensional learning behavior and generalization of neural networks [2]
  - Visualizing the loss in a 2D subspace at a a center point θ^* with two direction vectors, δ and η

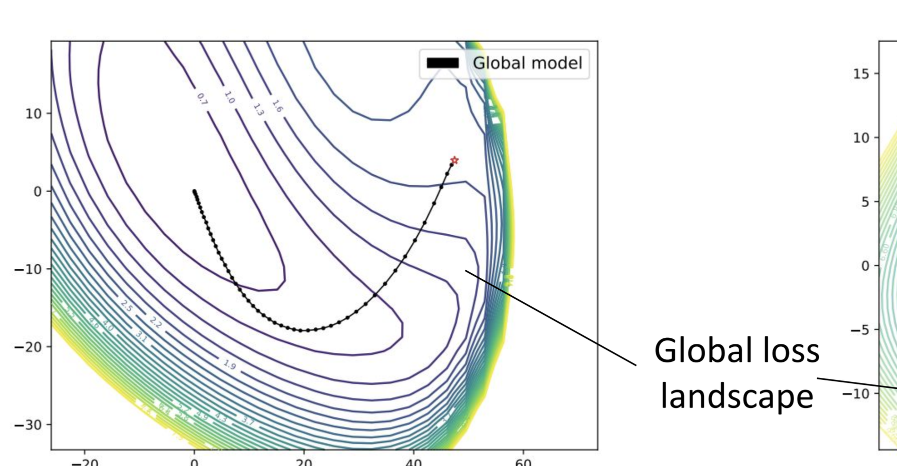
$$f(\alpha, \beta) = L(\theta^* + \alpha\delta + \beta\eta)$$



Loss Landscape Visualization for FL

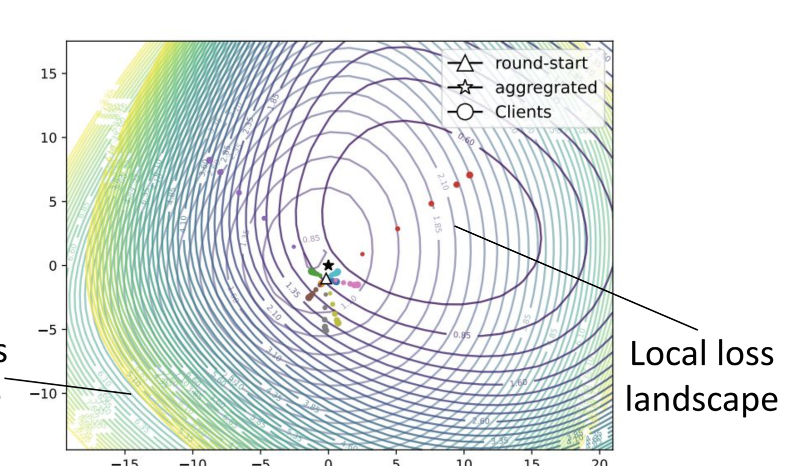
- ***Data preparation***
  - CIFAR-10 image classification
  - Number of clients: M = 10
  - Coefficient α indicates the non-IID degree
    - α=0.3 (middle non-IID)
- ***Training details***
  - Types of neural networks:
    - ConvNet and ResNet20
  - FedAvg with 100 rounds
  - 5 epochs for local training
  - Random initialization
  - Full client participation

- ***Global scope***



  - Trajectory of the aggregated global models over 100 rounds
  - Global model initialized at ☆
  - Loss values are computed using global training data

- ***Local scope***



  - Trajectories of 10 local models at round 9
  - Local training starts from △, and the final aggregated global model ☆ moves closer towards the minima

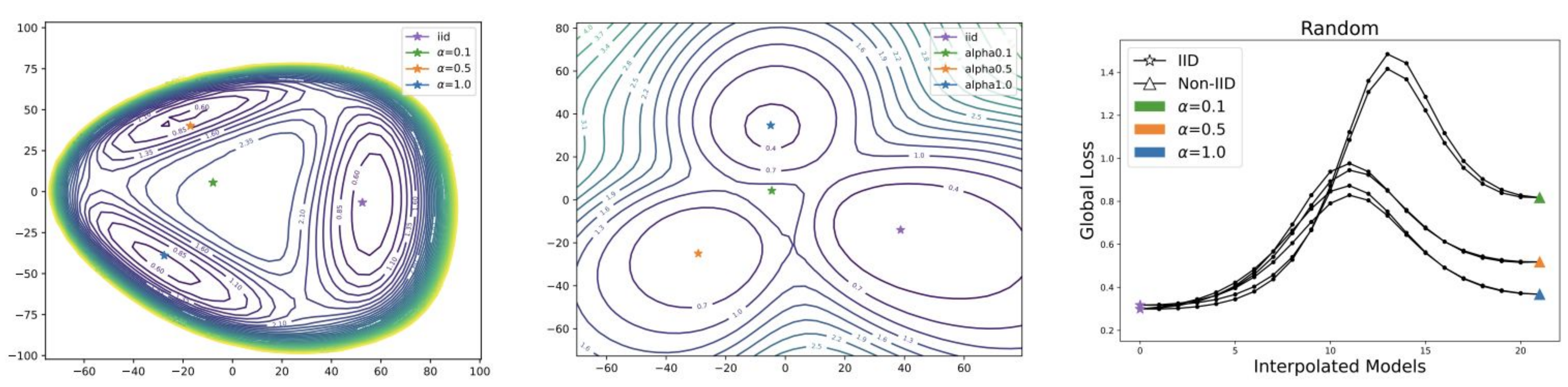
[1] McMahan et al., Communication-Efficient Learning of Deep Networks from Decentralized Data, 2017

[2] Li et al., Visualizing the Loss Landscape of Neural Nets, 2018

Detailed Study

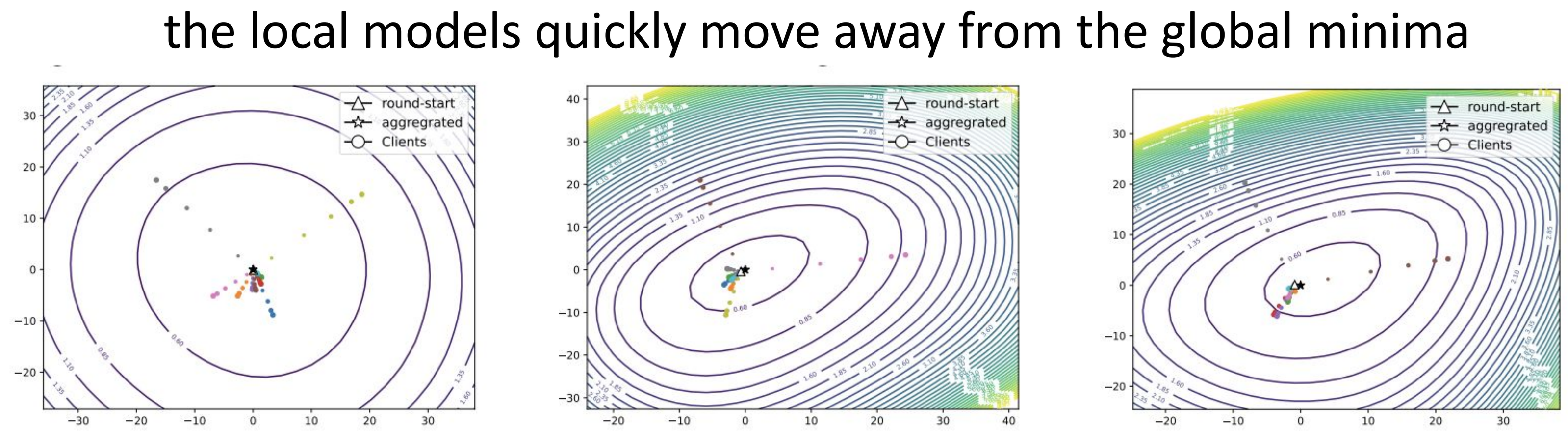
Effect of Data Heterogeneity

- **Random** weight initialization
  - Global scope
    - Visualizations of the final global models
    - With the same initialization, global models under different non-IID soon diverge into **different loss basins** (fig: a,b)
    - Global loss along the interpolation between the IID global model and each of the non-IID global model (fig: c)
      - a higher intermediate loss indicates existence of a larger barrier between two models



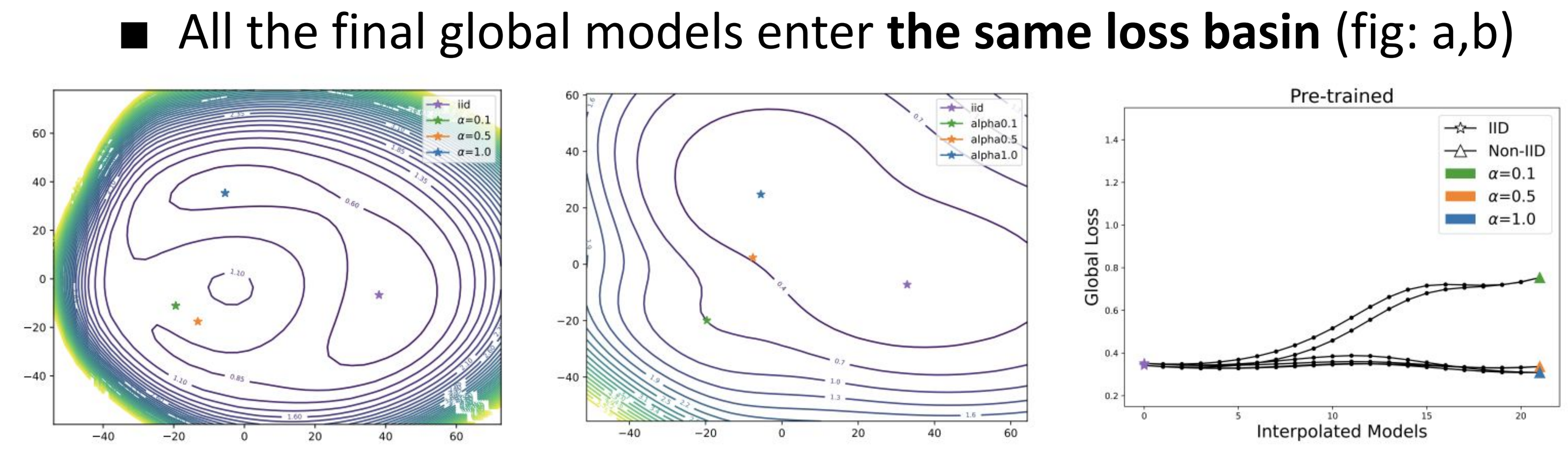
(a) ConvNet (b) ResNet-20 (c) Interpolation

  - Local scope
    - Different non-IID conditions, at round 30
    - Severer non-IID cases have denser loss contour, meaning that the local models quickly move away from the global minima



(a) IID (b) $\alpha = 1.0$ (c) $\alpha = 0.5$

- **Pre-trained** weight initialization
  - Global scope
    - Initialization with a well pre-trained model leads to a much smoother loss landscape
    - All the final global models enter **the same loss basin** (fig: a,b)



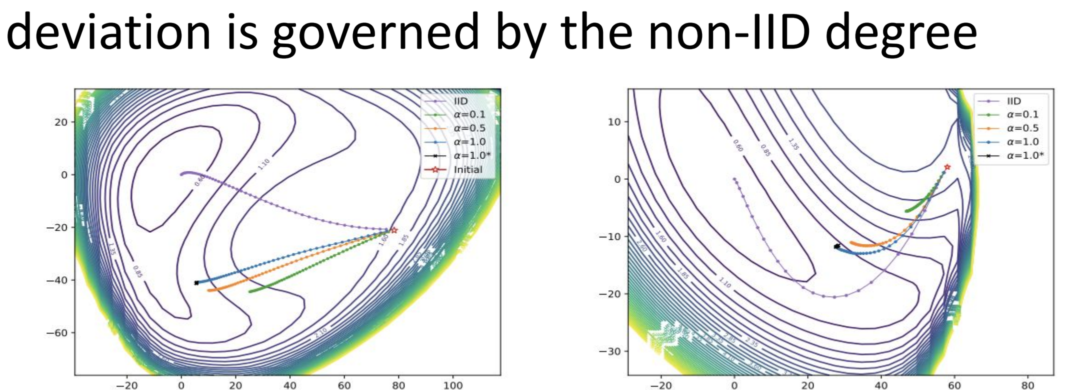
(a) ConvNet (b) ResNet-20 (c) Interpolation

Visualizations on training trajectories

(a) FedAvg initialized with the **ImageNet pre-trained weights**
(b) FedAvg initialized with the **weights after one round of IID training**

- When the data become more non-IID, the trajectory gradually deviates away from the ideal trajectory (purple one, under IID data) and ends at a higher loss value (fig: a,b)
- The degree of deviation is governed by the non-IID degree

- Four levels of data heterogeneity: IID to α = 1.0, α = 0.5, and α = 0.1

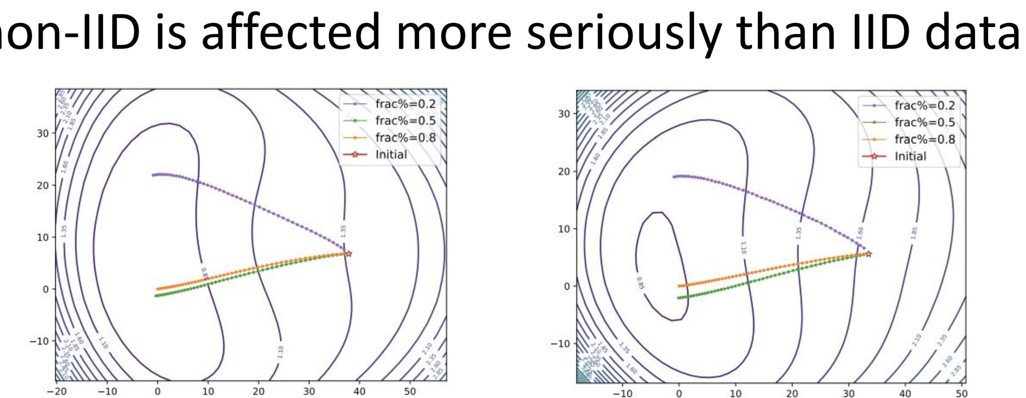


(a) Pre-trained weights. (b) Random weights, followed by one IID round.

Partial client participation

- Global scope
  - Under non-IID condition, if we only use a small portion of clients, the overall performance of FedAvg drops significantly
  - Performance of non-IID is affected more seriously than IID data

- At each round, randomly sample a portion of 50 clients
- Three client sampling rates: 0.2, 0.5, 0.8



(a) IID. (b) Non-IID with $\alpha = 0.5$.

This research is supported in part by grants from the National Science Foundation (IIS-2107077, OAC-2118240, and OAC-2112606) and Cisco Systems, Inc.