Tackling Distribution Shifts in Federated Learning

Krishna Pillutla
August 6, 2022 @ DRDS Workshop
Data is decentralized and private
Federated Learning

[Diagram showing a cloud labeled "Server" connected to multiple phones with data flowing into the cloud and training data flowing back out.]

Percentage of world population with a smartphone

- Year
  - 2000
  - 2005
  - 2010
  - 2015
  - 2020
  - 2025

- Percentage
  - 10
  - 20
  - 30
  - 40
  - 50
  - 60

Data Credit: Business Wire
Federated Learning

Percentage of world population with a smartphone

Data Credit: Business Wire
Federated Learning

Percentage of world population with a smartphone

Data Credit: Business Wire

Year


Percentage

60
50
40
30
20
10
0
Federated Learning

Percentage of world population with a smartphone

![Graph showing percentage of world population with a smartphone from 2000 to 2025.]

Data Credit: Business Wire
Federated Learning

Challenges:
- Communication efficiency
- Statistical heterogeneity
- Privacy of user data
Federated Learning

Challenges:
- Communication efficiency
- Statistical heterogeneity
- Privacy of user data
THE ACCENT GAP

We tested Amazon’s Alexa and Google’s Home to see how people with accents are getting left behind in the smart-speaker revolution.
Tackling distribution shifts in federated learning

- Improving tail performance with a single model
- Improving overall performance with local adaptation
Problem Setup

**Usual Learning Objective**

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} F_i(w)
\]

where

\[
F_i(w) = \mathbb{E}_{z \sim p_i} [f(w; z)]
\]

loss on client \(i\)

[McMahan et al. AISTATS (2017), Kairouz et al. (2021)]
$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} F_i(w)$$

Global model is trained on *average distribution* across clients
Global model is deployed on *individual* clients.
Global model is deployed on individual clients.
Global model is deployed on *individual* clients
Our goal: improve performance on “tail clients”
Simplicial federated learning

Our Approach: minimize the tail error directly!

Simplicial-FL Objective:

\[
\min_w \mathcal{S}_\theta\left( (F_1(w), \ldots, F_n(w)) \right)
\]

Superquantile | Conditional Value at Risk

[Rockafellar & Uryasev (2000; 2002)]
Distributional robustness in federated learning:

Assuming a new test client with mixture distribution $p_\pi = \sum \pi_i p_{ii}$, Simplicial-FL objective is equivalent to:

$$\min_w \max_{\pi : \pi_i \leq (n\theta)^{-1}} \mathbb{E}_{z \sim p_\pi}[f(w; z)]$$

**Dual expression** $\equiv$ continuous knapsack problem

$$\mathbb{S}_\theta(x_1, \cdots, x_n) = \max \left\{ \sum_i \pi_i x_i : \pi_i \geq 0, \sum_i \pi_i = 1, \pi_i \leq (n\theta)^{-1} \right\}$$

[Dantzig (1957), Ben-Tal & Teboulle (1987), Föllmer & Schied (2002)]
Challenge:

The superquantile is non-smooth

plot of $h(u_1, u_2) = S_{1/2}(u_1, u_2, 0, 0)$
**Nonsmooth:** The subdifferential has a tractable form

\[ \partial F_\theta(w) \ni \sum_{i=1}^{n} \pi_i^* \nabla F_i(w) \quad \text{where} \quad \pi_i^* \propto \mathbb{1}\left(F_i(w) \geq Q_\theta(F_1(w), \ldots, F_n(w)) \right) \]

assuming \( \theta_n \) is an integer
Nonsmooth: The subdifferential has a tractable form

$$\partial F_\theta(w) \supseteq \sum_{i=1}^{n} \pi^*_i \nabla F_i(w) \quad \text{where} \quad \pi^*_i \propto \mathbb{1}\left(F_i(w) \geq Q_\theta(F_1(w), \ldots, F_n(w))\right)$$

assuming $\theta n$ is an integer

Proof Chain rule $\implies$ subdifferential holds with

$$\pi^* \in \arg \max_{\pi \in \mathcal{P}_\theta} \sum_{i} \pi_i F_i(w)$$

Alternate form of $\pi^*$ comes from the continuous knapsack problem

[Dantzig. ORIJ (1957)]
Algorithm

In each communication round:

- Estimate the quantile
- Aggregate over the tail
Convergence rates

Non-convex case: $O(1/\sqrt{t}) + \text{lower order terms}$

Strongly convex case: $\tilde{O} \left( \kappa^{3/2} + \frac{1}{\lambda \epsilon} \right)$

$\kappa$: condition number
$\lambda$: strong convexity
Experiments: EMNIST
Histogram of per-client errors
Tackling distribution shifts in federated learning

- Improving tail performance with a single model

- Improving overall performance with local adaptation
The need for local adaptation a.k.a. personalization

Objective

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} F_i(w)
\]

where

\[
F_i(w) = \mathbb{E}_{z \sim p_i} [f(w; z)]
\]

loss on client \(i\)
Personalization: Each model has a global component and a per-client component

Shared Params $u$ + Personal Params $v_i$ = Full model $w_i = (u, v_i)$

Objective: \[
\min_{u, v_1, \ldots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)
\]

Example: \[
F_i(u, v_i) = \mathbb{E}_{(X,Y) \sim p_i} \left( \phi_g(X; u) + \phi_l(X; v_i) - Y \right)^2
\]
Personalization architectures

\[ F(u, v_i) = E_{(X,Y) \sim \rho_i} \left( \phi_g(X; u) + \phi_l(X; v_i) - Y \right)^2 \]

**Multi-task learning:** Caruana (1997), Baxter (2000), Evgeniou & Pontil (2004), Collobert & Weston (2005), Argyriou et al. (2008), …
Personalization architectures

![Diagram showing personalization architectures]

Multitask learning: Caruana (1997), Baxter (2000), Evgeniou & Pontil (2004), Collobert & Weston (2005), Argyriou et al. (2008), ...
Personalization architectures

**Multi-task learning:** Caruana (1997), Baxter (2000), Evgeniou & Pontil (2004), Collobert & Weston (2005), Argyriou et al. (2008), ...
Best personalization architecture depends on task heterogeneity

Next word prediction

Input
Output
Adapter

Speech recognition

y-axis shows error: lower is better
Best personalization architecture depends on task heterogeneity

Next word prediction

Speech recognition

Landmark detection

_y-axis shows error: lower is better_
Best personalization architecture depends on task heterogeneity

Next word prediction

Speech recognition

Landmark detection

$y$-axis shows error: lower is better
Open problems: Deeper understanding of shifts

Many negative results: optimization can slow down, makes robustness harder, ...

Yet, federated learning is used widely in practice
Open problems: Deeper understanding of shifts

Many negative results: optimization can slow down, makes robustness harder, ...
Yet, federated learning is used widely in practice

Quantify heterogeneity:

Measure gaps between distributions: MAUVE

Liu, Prabhat, Welleck, Oh, Choi, Harchaoui. NeurIPS (2021)]
Open problems: Deeper understanding of shifts

Many negative results: optimization can slow down, makes robustness harder, ...
Yet, federated learning is used widely in practice

Quantify heterogeneity:

Measure gaps between distributions: **MAUVE**

Best algorithms for different types of shifts (subject to federated constraints)

Statistical assumptions under which heterogeneity is benign?

What measures of heterogeneity impact optimization?

Liu, P., Welleck, Oh, Choi, Harchaoui. NeurIPS (2021)]
Federated Learning with Partial Model Personalization.
Krishna Pillutla, Kshitiz Malick, Abdulrehman Mohamed, Mike Rabbat, Maziar Sanjabi, Lin Xiao
ICML (2022).

Federated Learning with Heterogeneous Devices: A Superquantile Optimization Approach.
Under Review (arXiv 2112.09429)

A Superquantile Approach to Federated Learning with Heterogeneous Devices.

Superquantiles at Work : Machine Learning Applications and Efficient (Sub)gradient Computation.
Yassine Laguel, Krishna Pillutla, Jérôme Malick, Zaid Harchaoui.
Set-Valued and Variational Analysis (2021).