Federated Learning: Robustness, Heterogeneity and Optimization

Krishna Pillutla
June 24, 2022

Committee: Zaid Harchaoui (Chair), Sham Kakade (Chair), Yejin Choi, Lalit Jain (GSR), Kevin Jamieson, Jamie Morgenstern
Data is decentralized and private
Datacenter
Datacenter

Non-collaborative

Model 1

Model n

Training data

Common model

Training data
Peer-to-peer
Federated Learning

Percentage of world population with a smartphone

Year


Percentage

60
50
40
30
20
10
0

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
Federated Learning

Percentage of world population with a smartphone

Year


Percentage

0  10  20  30  40  50  60

Data Credit: Business Wire
Federated Learning

Percentage of world population with a smartphone

Communication cost > computation cost!

Data Credit: Business Wire
Challenges

• models are deployed on clients with heterogeneous 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.
Challenges

• models are deployed on clients with heterogeneous data

• training is not robust to potentially malicious clients
Alexa and Siri Can Hear This Hidden Command. You Can’t.

Researchers can now send secret audio instructions undetectable to the human ear to Apple’s Siri, Amazon’s Alexa and Google’s Assistant.
Clean Accuracy 64.3% -4.2pp
10% Corrupt 60.1%
Challenges

- models are deployed on clients with **heterogeneous data**
- training is **not robust** to potentially **malicious** clients
- solutions to both these problems are **conflicting**
1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]
This talk

1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]

Our approach

- Directly minimize the tail error
- Communication efficiency
- Privacy

![Histogram showing error distribution with low and high error categories.](Image)
This talk

1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]

2. **Robust aggregation** for federated learning
   [TSP ’22]
Clean 64.3% - 4.2pp
10% Corrupt 60.1%

Usual

Robust 62.9% - 0.6pp

10% Corrupt 62.3%
Clean 10% Corrupt

Usual

Robust

+64.3% -4.2pp 60.1% -1.4pp

+62.9% -0.6pp 62.3%
This talk

1. **Heterogeneity-aware objectives** for federated learning
   [CISS ’21, SVAA ’21, Under Review ’21]

2. **Robust aggregation** for federated learning
   [TSP ’22]

3. **Model personalization** for federated learning
   [ICML ’22, TSP ’22]
<table>
<thead>
<tr>
<th>Clean</th>
<th>25% Corrupt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usual</td>
<td>Robust</td>
</tr>
<tr>
<td>70.1%</td>
<td>69.8%</td>
</tr>
<tr>
<td>-45.5pp</td>
<td>-3.8pp</td>
</tr>
<tr>
<td>24.6%</td>
<td>66.0%</td>
</tr>
</tbody>
</table>
This talk

1. **Heterogeneity-aware objectives** for federated learning  
   [CISS ’21, SVAA ’21, Under Review ’21]

2. **Robust aggregation** for federated learning  
   [TSP ’22]

3. **Model personalization** for federated learning  
   [ICML ’22, TSP ’22]
Part 1: Heterogeneity-aware objectives for federated learning

[CISS ’21, SVAA ’21, Under Review ’21]
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 $i$

Data heterogeneity

[McMahan et al. AISTATS (2017), Kairouz et al. (2021)]
Our goal

Reduce tail error

![Histograms showing error distribution before and after reduction. The goal is to reduce the high error count.](image-url)
Our goal

Reduce tail error without sacrificing the mean error
Simplicial federated learning

Our Approach: minimize the tail error directly!

Simplicial-FL Objective:
\[
\min_w \mathbb{S}_\theta \left( (F_1(w), \ldots, F_n(w)) \right)
\]

Superquantile | Conditional Value at Risk

\[\mathbb{E}[Z]
\]

\[Q_\theta(Z)
\]

\[S_\theta(Z) = \mathbb{E}[Z \mid Z > Q_\theta(Z)]
\]

[Rockafellar & Uryasev (2002)]
**Dual expression** $\equiv$ continuous knapsack problem

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

[Dantzig (1957), Ben-Tal & Teboulle (1987), Föllmer & Schied (2002)]
**Dual expression** \(\equiv\) continuous knapsack problem

\[
\mathcal{S}_\theta(x_1, \ldots, x_n) = \max \left\{ \sum_i \pi_i x_i : \pi_i \geq 0, \sum_i \pi_i = 1, \pi_i \leq (n\theta)^{-1} \right\}
\]

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

Assuming a new test client with mixture distribution \(p_\pi = \sum_i \pi_i p_{ir}\), the Simplicial-FL objective is equivalent to:

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

\(\Rightarrow\) Distributionally robust learning
Optimization
Communication primitive: secure sum

Only reveal $x_1 + x_2$ to the server without revealing $x_1$ or $x_2$

Client 1

Client 2

[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]
Perform all operations modulo $M$

[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]
Client 1

$x'_1 = x_1 + \xi$

Client 2

$x'_2 = x_2 - \xi$

$\xi \sim \text{Unif}\left(\bigcirc\right)$
Server only sees $x'_1, x'_2 \sim \text{Unif}(\bigcirc)$ but calculates the correct sum $x'_1 + x'_2 = x_1 + x_2$.

[Bonawitz et al. CCS (2017), Bell et al. CCS (2020)]
Server only sees $x'_1, x'_2 \sim \text{Unif}(\bigcirc)$ but calculates the correct sum $x'_1 + x'_2 = x_1 + x_2$

Client 1

$x'_1 = x_1 + \xi$

$\xi \sim \text{Unif}(\bigcirc)$

Client 2

$x'_2 = x_2 - \xi$

Server

Total communication for $m$ vectors in $\mathbb{R}^d = O(m \log m + md)$ numbers
Real-world communication constraint:
All client-to-server communication must go through secure summation
ERM Algorithm (FedAvg):

$$\min_w \frac{1}{n} \sum_{i=1}^n F_i(w)$$

Simplicial-FL Algorithm:

$$\min_w \mathbb{S}_\theta\left( (F_1(w), \ldots, F_n(w)) \right)$$

FedAvg [MacMahan et al. AISTATS (2017)]

Parallel Gradient Distribution [Mangasarian. SICON (1995)]
Iterative Parameter Mixing [McDonald et al. ACL (2009)]
BMUF [Chen & Huo. ICASSP (2016)]
Local SGD [Stich. ICLR (2019)]
ERM Algorithm (FedAvg):

$$\min_w \frac{1}{n} \sum_{i=1}^{n} F_i(w)$$

Simplicial-FL Algorithm:

$$\min_w \mathbb{S}_\theta\left( (F_1(w), \ldots, F_n(w)) \right)$$

Step 1 of 3: Server samples $m$ clients and broadcasts global model
ERM Algorithm (FedAvg):

$$\min_w \frac{1}{n} \sum_{i=1}^{n} F_i(w)$$

Simplicial-FL Algorithm:

$$\min_w \mathbb{S}_\theta \left( (F_1(w), \ldots, F_n(w)) \right)$$

*Step 2 of 3: Clients perform $\tau$ local SGD steps on their local data*
ERM Algorithm (FedAvg):

$$\min_w \frac{1}{n} \sum_{i=1}^{n} F_i(w)$$

Step 3 of 3: Aggregate updates contributed by all clients

Simplicial-FL Algorithm:

$$\min_w \mathbb{S}_\theta \left( (F_1(w), \ldots, F_n(w)) \right)$$

Step 3 of 3: Aggregate updates contributed by tail clients only
Loss

Count

(1 - \(\theta\))-Quantile

(1 - \(\theta\))-Quantile

Loss

Count

Tail


\textbf{Loss - Count - Tail - Quantile} \quad \text{(1 - \theta)-Quantile}

\begin{itemize}
  \item \textbf{Loss}: Horizontal axis showing the range of losses.
  \item \textbf{Count}: Vertical axis showing the frequency of losses.
  \item \textbf{Tail}: The right end of the distribution, indicating high losses.
  \item \textbf{(1 - \theta)-Quantile}: The quantile that represents the upper tail of the distribution.
\end{itemize}
Per-client loss

\[ \sum \]

Histogram

(1 - \theta)-Quantile
$h'_i = h_i + \mathcal{N}(0, \sigma^2 I_b)$
$h_i' = h_i + \mathcal{N}(0,\sigma^2 I_b)$
Per-client loss

$\sum$ Histogram

(1 − $\theta$)-Quantile

Distributed discrete Gaussian mechanism

$\approx$ (1 − $\theta$)-Quantile

Noisy histogram

$\approx$ Tail

[Kairouz, Liu, Steinke. ICML (2021)]
**Proposition** [P., Laguel, Malick, Harchaoui]

Fix parameters $\epsilon, \delta > 0$ and $M \gtrsim m^{3/2}$. If we choose noise scale

$$\sigma \approx \frac{1}{\epsilon \sqrt{m} \sqrt{\log \frac{1}{\delta}}}$$

$\sigma$ #clients per round
$M$ modular ring size
$b$ #bins in the histogram
$(\epsilon, \delta)$ differential privacy parameters

then

- the noisy histogram (and hence all quantiles) are $(\epsilon, \delta)$-differentially private
- w.h.p., the estimated $(1 - \theta)$-quantile is actually the $(1 - \theta')$-quantile, with

$$|\theta' - \theta| \lesssim \sqrt{\frac{b\sigma^2}{m}} \approx \frac{1}{\epsilon m} \sqrt{b \log \frac{1}{\delta}}$$

Total communication cost $\approx bm \log^2 m$
Proposition [P., Laguel, Malick, Harchaoui]

Fix parameters $\epsilon, \delta > 0$ and $M \geq m^{3/2}$. If we choose noise scale

$$\sigma \approx \frac{1}{\epsilon\sqrt{m}} \sqrt{\log \frac{1}{\delta}}$$

then

- the noisy histogram (and hence all quantiles) are $(\epsilon, \delta)$-differentially private
- w.h.p., the estimated $(1 - \theta)$-quantile is actually the $(1 - \theta')$-quantile, with

$$|\theta' - \theta| \lesssim \sqrt{\frac{b\sigma^2}{m}} \approx \frac{1}{\epsilon m} \sqrt{b \log \frac{1}{\delta}}$$

Total communication cost $\approx bm \log^2 m$
**Proposition** [P., Laguel, Malick, Harchaoui]

Fix parameters $\epsilon, \delta > 0$ and $M \geq m^{3/2}$. If we choose noise scale

$$\sigma \approx \frac{1}{\epsilon \sqrt{m}} \sqrt{\log \frac{1}{\delta}}$$

then

- the noisy histogram (and hence all quantiles) are $(\epsilon, \delta)$-differentially private

- w.h.p., the estimated $(1 - \theta)$-quantile is actually the $(1 - \theta')$-quantile, with

$$|\theta' - \theta| \lesssim \sqrt{\frac{b\sigma^2}{m}} \approx \frac{1}{\epsilon m} \sqrt{b \log \frac{1}{\delta}}$$

Total communication cost $\approx bm \log^2 m$
Summary:

Simplicial-FL algorithm requires 2 secure summations per update
Convergence analysis (non-convex)
Challenge #1:

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

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

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

\[ \pi^* \in \text{arg max} \sum_{i} \pi_i F_i(w) \]

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

[Dantzig. ORIJ (1957)]
**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

**Other option**: Use smoothing

[Nesterov. Math. Prog. (2005),
Beck & Teboulle. SIAM J. Optim. (2012),
Laguel, P., Malick, Harchaoui. SVAA (2021)]
Challenge #2

The superquantile is *nonlinear*

\[ \implies \text{unbiased stochastic gradients not possible} \]

For i.i.d. copies \( Z_1, \ldots, Z_m \) of \( Z \), we have

\[
\mathbb{E} \left[ \frac{1}{m} \sum_{i=1}^{m} Z_i \right] = \mathbb{E}[Z] \quad \text{but} \quad \mathbb{E} \left[ S_\theta(Z_1, \ldots, Z_m) \right] \neq S_\theta(Z)
\]
Nonlinear: We minimize a close surrogate

\[ \bar{F}_\theta(w) = \mathbb{E}_{S: |S|=m} \left[ S_\theta \left( \left( F_i(w) : i \in S \right) \right) \right] \]
**Nonlinear**: We minimize a close surrogate

\[
\bar{F}_\theta(w) = \mathbb{E}_{S:|S|=m} \left[ S_\theta\left( (F_i(w) : i \in S) \right) \right]
\]

The surrogate is uniformly close for bounded losses:

For i.i.d. copies \(Z_1, \ldots, Z_m\) of \(Z\) with \(|Z| \leq B\) a.s., we have

\[
\left| \mathbb{E}\left[ S_\theta(Z_1, \ldots, Z_m) \right] - S_\theta(Z) \right| \leq \frac{B}{\sqrt{\theta m}} \\
\text{Var}\left[ S_\theta(Z_1, \ldots, Z_m) \right] \leq \frac{B^2}{\theta m}
\]

[Levy et al. NeurIPS (2020)]
Theorem [P., Laguel, Malick, Harchaoui]

Suppose each $F_i$ is $L$-smooth and $G$-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$
\mathbb{E}\left\| \nabla \Phi_\theta^{2L}(w_t) \right\|^2 \leq \sqrt{\frac{\Delta_0 LG^2}{t}} + (1 - \tau)^{1/3} \left( \frac{\Delta_0 LG}{t} \right)^{2/3} + \frac{\Delta_0 L}{t}
$$

$t$: #comm. rounds
$\tau$: #local update steps
$\Delta_0$: initial error

$$
\Phi_\theta^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2} \|z - w\|^2 \right\} \quad \text{← Moreau envelope of } F_\theta \text{ well defined for } \mu > L
$$
**Theorem** [P., Laguel, Malick, Harchaoui]

Suppose each $F_i$ is $L$-smooth and $G$-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$
\mathbb{E}\left\| \nabla \Phi^2_L(w_t) \right\|^2 
\leq \sqrt{\frac{\Delta_0 LG^2}{t}} + (1 - \tau)^{1/3} \left( \frac{\Delta_0 LG}{t} \right)^{2/3} + \frac{\Delta_0 L}{t}
$$

$t$: #comm. rounds

$\tau$: #local update steps

$\Delta_0$: initial error

\[
\Phi^\mu_\theta(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2} \|z - w\|^2 \right\} \quad \leftarrow \quad \text{Moreau envelope of } F_\theta \mid \text{well defined for } \mu > L
\]
**Theorem** [P., Laguel, Malick, Harchaoui]

Suppose each $F_i$ is $L$-smooth and $G$-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

$$
\mathbb{E} \left\| \nabla \Phi^{2L}_\theta(w_t) \right\|^2 \leq \sqrt{\frac{\Delta_0 L G^2}{t}} + (1 - \tau)^{1/3} \left( \frac{\Delta_0 L G}{t} \right)^{2/3} + \frac{\Delta_0 L}{t}
$$

$t$: #comm. rounds  
$\tau$: #local update steps  
$\Delta_0$: initial error

$$
\Phi^\mu_\theta(w) = \inf_z \left\{ F_\theta(z) + \frac{\mu}{2} \| z - w \|^2 \right\} \quad \text{Moreau envelope of } F_\theta \text{, well defined for } \mu > L
$$
**Theorem** [P., Laguel, Malick, Harchaoui]

Suppose each $F_i$ is $L$-smooth and $G$-Lipschitz.

Then, Simplicial-FL satisfies the convergence guarantee:

\[
\mathbb{E} \left\| \nabla \Phi^2L(w_t) \right\|^2 \leq \sqrt{\frac{\Delta_0 LG^2}{t} + (1 - \tau)^{1/3} \left( \frac{\Delta_0 LG}{t} \right)^{2/3} + \frac{\Delta_0 L}{t}}
\]

$t$: #comm. rounds
$\tau$: #local update steps
$\Delta_0$: initial error

\[\Phi^\mu(w) = \inf_z \left\{ \overline{F}_\theta(z) + \frac{\mu}{2} \|z - w\|^2 \right\} \quad \text{Moreau envelope of } F_\theta \text{ well defined for } \mu > L\]
\[ \Phi(w_{t+1}) = \min_z \left\{ F_\theta(z) + L\|z - w_{t+1}\|^2 \right\} \]

\[ \leq F_\theta(z_t) + L\|z_t - w_{t+1}\|^2 \]

Plug in a particular choice of \( z_t \) to be determined later

Definition of the Moreau envelope
\[ \Phi(w_{t+1}) = \min_z \left\{ \overline{F}_\theta(z) + L\|z - w_{t+1}\|^2 \right\} \]

\[ \leq \overline{F}_\theta(z_t) + L\|z_t - w_{t+1}\|^2 \]

\[ \leq \overline{F}_\theta(z_t) + L\|z_t - w_t\|^2 - \gamma L(w_t - z_t)^T g_t + O(\gamma^2) \]

Definition of the Moreau envelope

Plug in a particular choice of \( z_t \) to be determined later

Expand update
\[ w_{t+1} = w_t - \gamma g_t \]
\[ \Phi(w_{t+1}) = \min_z \left\{ \bar{F}_\theta(z) + L \| z - w_{t+1} \|^2 \right\} \]

\[ \leq \bar{F}_\theta(z_t) + L \| z_t - w_{t+1} \|^2 \]

\[ \leq \bar{F}_\theta(z_t) + L \| z_t - w_t \|^2 - \gamma L (w_t - z_t)^T g_t + O(\gamma^2) \]

\[ = \Phi(w_t) - \gamma \nabla \Phi(w_t)^T g_t + O(\gamma^2) \]

Definition of the Moreau envelope

Plug in a particular choice of \( z_t \) to be determined later

Expand update \( w_{t+1} = w_t - \gamma g_t \)

Choose \( z_t = \arg \min_z \left\{ \bar{F}_\theta(z) + L \| z - w_t \|^2 \right\} \)

so that \( \nabla \Phi(w_t) = L(w_t - z_t) \)
\[
\Phi(w_{t+1}) = \min_z \left\{ \overline{F}_{\theta}(z) + L\|z - w_{t+1}\|^2 \right\}
\]

\[
\leq \overline{F}_{\theta}(z_t) + L\|z_t - w_{t+1}\|^2
\]

\[
\leq \overline{F}_{\theta}(z_t) + L\|z_t - w_t\|^2 - \gamma L(w_t - z_t)^T g_t + O(\gamma^2)
\]

\[
= \Phi(w_t) - \gamma \nabla \Phi(w_t)^T g_t + O(\gamma^2)
\]

If \( E_i[g_i] \approx \nabla \Phi(w_i) \), proof is complete

Plug in a particular choice of \( z_t \) to be determined later

Expand update

Choose

so that \( \nabla \Phi(w_t) = L(w_t - z_t) \)

Definition of the Moreau envelope
\[ \Phi(w_{t+1}) \leq \Phi(w_t) - \gamma \nabla \Phi(w_t)^T g_t + O(\gamma^2) \]

- \( g_t \) comes from \( \tau \) local gradient steps of step size \( \gamma \)
- \( g_t' \) comes from one local gradient steps of step size \( \tau \gamma \)
\[ \Phi(w_{t+1}) \leq \Phi(w_t) - \gamma \nabla \Phi(w_t)^T g_t + O(\gamma^2) \]

\[ g_t \text{ comes from } \tau \text{ local gradient steps of step size } \gamma \]

\[ \Phi(w_{t+1}) \leq \Phi(w_t) - \tau \gamma \nabla \Phi(w_t)^T g'_t + O(\gamma^2) \]

\[ g'_t \text{ comes from one local gradient steps of step size } \tau \gamma \]

\[ \mathbb{E}_t[g'_t] \in \partial \bar{F}_\theta(w_t) \]
\[
\Phi(w_{t+1}) \leq \Phi(w_t) - \gamma \nabla \Phi(w_t)^T g_t + O(\gamma^2)
\]

\[
\Phi(w_{t+1}) \leq \Phi(w_t) - \tau \gamma \nabla \Phi(w_t)^T g'_t + O(\gamma^2)
\]

\[
\nabla \Phi(w_t)^T \mathbb{E}[g'_t] \geq \frac{1}{2} \| \nabla \Phi(w_t) \|^2
\]

\[g_t\] comes from \(\tau\) local gradient steps of step size \(\gamma\)

\[g'_t\] comes from one local gradient steps of step size \(\tau \gamma\)

\[\mathbb{E}[g'_t] \in \partial \bar{F}_\theta(w_t)\]

Prox-gradient and subgradient are closely aligned

[Davis & Drusvyatskiy. SIAM J. Optim. (2019)]
Experiments: EMNIST
Histogram of per-client errors

Misclassification Error
• Simplicial-FL has the smallest 90th percentile error

• Simplicial-FL is competitive on the mean error
Distributionally robust learning with 1 additional line of code

```python
import torch.nn.functional as F
from sqwash import reduce_superquantile

for x, y in dataloader:
    y_hat = model(x)
    batch_losses = F.cross_entropy(y_hat, y, reduction='none')  # must set `reduction='none'`
    loss = reduce_superquantile(batch_losses, superquantile_tail_fraction=0.5)  # Additional line
    loss.backward()  # Proceed as usual from here
```

Install: `pip install sqwash`

Documentation: [krishnap25.github.io/sqwash/](https://krishnap25.github.io/sqwash/)
Part 2: Robust aggregation for federated learning

[TSP ’22]
Arithmetic mean aggregation is *not robust* to corruptions ➞ Poor predictions!
Our goal

Design a robust aggregation algorithm for federated learning which is

Communication efficient

$O(1)$ times the communication cost as non-robust aggregation

Secure aggregation

Client-server communication via secure summation only

Note: not DP
Consider mean estimation in Huber’s contamination model:

\[ w_1, \ldots, w_n \sim (1 - \rho) \mathcal{N}(\mu, \sigma^2 I) + \rho Q \]

Any mean estimate \( \bar{w}_n \) must satisfy

\[ \| \bar{w}_n - \mu \|^2 \geq \sigma^2 \left( \rho^2 + \frac{d}{n} \right) \]

For general federated learning, fix a set $D$ of “inlier” distributions

Convergence only possible up to size of inlier set $D$ due to

$$\|\overline{w}_n - \mu\|^2 \geq \sigma^2 \left(\rho^2 + \frac{d}{n}\right)$$

Algorithm is agnostic to $D$ — only appears in analysis
Federated learning with robust aggregation
Fermat & Torricelli (~1600s), Weber (1909)
Fermat & Torricelli (~1600s), Weber (1909)

Geometric Median

$$GM(w_1, \ldots, w_m) = \arg \min_z \left\{ \sum_{i=1}^{m} \|z - w_i\|_2 \right\}$$
**Robustness**: Breakdown point of GM = 1/2

Weiszfeld’s Algorithm

Start with initial guess $z_0$ and iterate:

$$\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2, \nu\}}$$

$$z_{t+1} = \frac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$$

[Weiszfeld 1937]
Weiszfeld’s Algorithm

Start with initial guess $z_0$ and iterate:

$$\beta_{i,t} = \frac{1}{\max\{||z_t - w_i||_2, \nu\}}$$

$$z_{t+1} = \frac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}}$$

[Weiszfeld 1937]
Proposition \([P., \text{Kakade, Harchaoui}]

Assume that
\[
\min_i \|z^* - w_i\| \geq \nu.
\]
Then, we get an \(\varepsilon\)-approximate GM in
\[
O\left(\frac{1}{\nu \varepsilon}\right)
\]
iterations

\[z^* = \text{GM}(w_1, \ldots, w_m)\]

\(\nu\): smoothing in \(\beta\)-update

\[
\beta_{i,t} = \frac{1}{\max\{\|z_t - w_i\|_2, \nu\}}
\]

Communication efficient!

Empirically, \(3-5\) iterations suffice: rapid convergence

Even \(1\) iteration gives robustness
**RFA** = FedAvg + GM aggregation

Secure aggregation

Only client-server communication is via **secure summation** in

\[ z_{t+1} = \frac{\sum_i \beta_{i,t} w_i}{\sum_i \beta_{i,t}} \]
Step 1 of 3: Server broadcasts global model to sampled clients

Step 2 of 3: Clients perform some local SGD steps on their local data

So far, same as FedAvg
Step 3 of 3: Aggregate with multiple rounds of secure average
(weights $\beta_i$ from the Weiszfeld Algorithm)
Convergence analysis (least squares)
Data on client $i$: $(X_i, Y_i) \sim p_i$ satisfies

$$Y_i = X_i^T w_i^* + \xi_i \quad \text{where} \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

$X_1 \sim \mathcal{N}(0, H_1)$

$X_2 \sim \mathcal{N}(0, H_2)$
Data on client $i$: $(X_i, Y_i) \sim p_i$ satisfies

$$Y_i = X_i^T w_i^* + \xi_i \quad \text{where} \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

Measure of heterogeneity

$$\Omega_X = \max_{i \in \text{inlier}} \lambda_{\max}(H_i^{-1/2}H_iH_i^{-1/2}) \geq 1$$

$$H_i = \mathbb{E}[X_iX_i^T] \quad \text{marginal covariance of } i$$

$$H = \frac{1}{n} \sum_i H_i \quad \text{marginal covariance of mixture}$$

$X_1 \sim \mathcal{N}(0, H_1)$

$X_2 \sim \mathcal{N}(0, H_2)$
Data on client $i$: $(X_i, Y_i) \sim p_i$ satisfies

$$Y_i = X_i^T w_i^* + \xi_i \quad \text{where} \quad \xi_i \sim \mathcal{N}(0, \sigma^2)$$

Measure of heterogeneity

$$\Omega_Y = \max_{i,j \text{ inlier}} \|w_i^* - w_j^*\|_2 \geq 0$$
Fraction of non-corrupted clients = $\frac{1}{2} + c$

Number of clients per round = $m$

\( \mathcal{E} \) holds w.h.p. if $m \geq \frac{1}{c^2}$

**Theorem** [P., Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \leq 1$ and number of local steps $\propto 2^t$. Let $\mathcal{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with $\varepsilon$-approximate GM satisfies

\[
\mathbb{E} \left[ \|w_t - w^*\|^2 \right| \mathcal{E} \right] \leq \frac{\|w_0 - w^*\|^2}{2^t} + \frac{1}{c^2} \left( d\sigma^2 \frac{t}{2^t} + \frac{\varepsilon^2}{m^2} + \Omega^2_x \Omega^2_y \right)
\]

Optimization error $\rightarrow$ Heterogeneity Error

Statistical error $\rightarrow$ GM Approx. Error
Fraction of non-corrupted clients = \( \frac{1}{2} + c \)

Number of clients per round = \( m \)

\( \mathcal{E} \) holds w.h.p. if \( m \gtrsim \frac{1}{c^2} \)

**Theorem** [P., Kakade, Harchaoui]

Assume that \( F(w) \) is strongly convex, \( \|X_i\| \leq 1 \) and number of local steps \( \propto 2^t \). Let \( \mathcal{E} \) denote the event that at least \( 1/2 + c/2 \) non-corrupted devices are chosen in each round.

Then, RFA with \( \epsilon \)-approximate GM satisfies

\[
\mathbb{E} \left[ \|w_t - w^*\|^2 \mathbb{1}_{\mathcal{E}} \right] \leq \frac{\|w_0 - w^*\|^2}{2^t} + \frac{1}{c^2} \left( \frac{d \sigma^2 t}{2^t} + \frac{\epsilon^2}{m^2} + \Omega_X^2 \Omega_Y^2 \right)
\]

Optimization error

Statistical error

Heterogeneity Error

GM Approx. Error
Fraction of non-corrupted clients = $\frac{1}{2} + c$

Number of clients per round = $m$

$\mathcal{E}$ holds w.h.p. if $m \gtrsim \frac{1}{c^2}$

**Theorem** [P., Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \leq 1$ and number of local steps $\propto 2^t$. Let $\mathcal{E}$ denote the event that **at least** $1/2 + c/2$ **non-corrupted devices** are chosen in each round.

Then, RFA with $\varepsilon$-approximate GM satisfies

$$
\mathbb{E} [\|w_t - w^*\|^2] \lesssim \frac{\|w_0 - w^*\|^2}{2^t} + \frac{1}{c^2} \left( d\sigma^2 t \frac{t}{2^t} + \frac{\varepsilon^2}{m^2} + \Omega_X^2 \Omega_Y^2 \right)
$$

Optimization error

Statistical error

Heterogeneity Error

GM Approx. Error
Fraction of non-corrupted clients = $\frac{1}{2} + c$

Number of clients per round = $m$

$\mathcal{E}$ holds w.h.p. if $m \gtrsim \frac{1}{c^2}$

**Theorem** [P., Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \leq 1$ and number of local steps $\propto 2^t$. Let $\mathcal{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with $\epsilon$-approximate GM satisfies

$$
\mathbb{E} \left[ \|w_t - w^*\|^2 \mid \mathcal{E} \right] \lesssim \frac{\|w_0 - w^*\|^2}{2^t} + \frac{1}{c^2} \left( d\sigma^2 \frac{t}{2^t} + \frac{\varepsilon^2}{m^2} + \Omega_X^2 \Omega_Y^2 \right)
$$

**Optimization error**  
**Statistical error**  
**Heterogeneity Error**  
**GM Approx. Error**
Fraction of non-corrupted clients $= \frac{1}{2} + c$

Number of clients per round $= m$

$\mathcal{E}$ holds w.h.p. if $m \geq \frac{1}{c^2}$

**Theorem** [P., Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \leq 1$ and number of local steps $\propto 2^t$. Let $\mathcal{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with $\varepsilon$-approximate GM satisfies

$$
\mathbb{E} \left[ \|w_t - w^*\|^2 \mid \mathcal{E} \right] \leq \frac{\|w_0 - w^*\|^2}{2^t} + \frac{1}{c^2} \left( d\sigma^2 t + \frac{\varepsilon^2}{m^2} + \Omega_X^2 \Omega_Y^2 \right)
$$

**Optimization error**

**Statistical error**

**Heterogeneity Error**

**GM Approx. Error**
Fraction of non-corrupted clients = $\frac{1}{2} + c$

Number of clients per round = $m$

\[ \mathcal{E} \text{ holds w.h.p. if } m \geq \frac{1}{c^2} \]

**Theorem** [P., Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \leq 1$ and number of local steps \( \propto 2^t \). Let \( \mathcal{E} \) denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with $\epsilon$-approximate GM satisfies

\[
\mathbb{E} \left[ \|w_t - w^*\|^2 \mid \mathcal{E} \right] \leq \frac{\|w_0 - w^*\|^2}{2^t} + \frac{1}{c^2} \left( d\sigma^2 \frac{t}{2^t} + \frac{\epsilon^2}{m^2} + \Omega_X^2 + \Omega_Y^2 \right)
\]

Optimization error

Statistical error

Heterogeneity Error

GM Approx. Error
Fraction of non-corrupted clients $= \frac{1}{2} + c$

Number of clients per round $= m$

$\mathcal{E}$ holds w.h.p. if $m \geq \frac{1}{c^2}$

**Theorem** [P., Kakade, Harchaoui]

Assume that $F(w)$ is strongly convex, $\|X_i\| \leq 1$ and number of local steps $\propto 2^t$. Let $\mathcal{E}$ denote the event that at least $1/2 + c/2$ non-corrupted devices are chosen in each round.

Then, RFA with $\epsilon$-approximate GM satisfies

$$
\mathbb{E} \left[ \|w_t - w^*\|^2 \mid \mathcal{E} \right] \leq \frac{\|w_0 - w^*\|^2}{2^t} + \frac{1}{c^2} \left( d\sigma^2 \frac{t}{2^t} + \frac{\epsilon^2}{m^2} + \Omega^2_X \Omega^2_Y \right)
$$

- Optimization error
- Statistical error
- Heterogeneity Error
- GM Approx. Error
Experiments

Data Corruption

EMNIST Linear

EMNIST ConvNet

Shakespeare LSTM

Sent140 Linear

Update Corruption

EMNIST Linear

EMNIST ConvNet

Shakespeare LSTM

Sent140 Linear
Experiments

1.4pp gap at zero corruption
Reducing RFA’s communication

One round of RFA $\implies$ 3-5 rounds of communication (Weiszfeld)

Does 1 round of communication give robustness?

$$\beta_i = \frac{1}{\max\{\|w_i\|_2, \nu\}}$$

$$z = \frac{\sum_i \beta_i w_i}{\sum_i \beta_{i,t}}$$

Convergence of Weiszfeld’s algorithm
Does 1 round of communication give robustness? **Yes!**

![EMNIST Linear (Data)]

Test Accuracy vs Corruption Level for RFA, RFA-1, and FedAvg.
import torch
from geom_median.torch import compute_geometric_median  # PyTorch API
# from geom_median.numpy import compute_geometric_median  # NumPy API

points = [torch.rand(d) for _ in range(n)]  # list of n tensors of shape (d,)
# The shape of each tensor is the same and can be arbitrary (not necessarily 1-dimensional)
weights = torch.rand(n)  # non-negative weights of shape (n,)
out = compute_geometric_median(points, weights)
# Access the median via `out.median`, which has the same shape as the points, i.e., (d,)

Install: pip install geom-median

Documentation: github.com/krishnap25/geom-median
Part 3: Model personalization for federated learning

[TSP ’22, ICML ’22]
Two regimes

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 \(i\)
Option 1: Train a separate model per client (no collaboration)

**Objective:** \[\min_{w_i} F_i(w_i)\]

<table>
<thead>
<tr>
<th>Privacy</th>
<th>Best possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>Poor</td>
</tr>
</tbody>
</table>
**Option 2:** The 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: $F_i(u, v_i) = \mathbb{E}_{(X,Y) \sim p_i} \left( \phi_g(X; u) + \phi_l(X; v_i) - Y \right)^2$
**Option 2:** The model has a global component and a per-client component

\[
\text{Shared Params } u + \text{ Personal Params } v_i = \text{ Full model } w_i = (u, v_i)
\]

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

<table>
<thead>
<tr>
<th>Privacy</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>data and personal params on client</td>
</tr>
</tbody>
</table>
Option 2: The 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: $$\min_{u, v_1, \ldots, v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)$$

Privacy  

<table>
<thead>
<tr>
<th>Privacy</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>Yes</td>
</tr>
</tbody>
</table>

data and personal params on client
Multi-task learning: Caruana (1997), Baxter (2000), Evgeniou & Pontil (2004), Collobert & Weston (2005), Argyriou et al. (2008), ...
**Personalization Architectures**

### Multi-task learning:
- Caruana (1997)
- Baxter (2000)
- Collobert & Weston (2005)
- Argyriou et al. (2008)

Arivazhagan et al. (2019)  
Collins et al. (2021)  
Liang et al. (2019)
Personalization Architectures

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

- Server samples $m$ clients and broadcast global model $u$

- **Local updates** on client $i$: $(u_i^+, v_i^+) = \text{LocalUpdate}_i(u, v_i)$

- Aggregate updates to global part of the model:
  $$u^+ = \frac{1}{m} \sum_i u_i^+$$

**Alternating update**

- Simultaneous update:
  $$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$
  $$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$

**Simultaneous update**

- Alternating update:
  $$v_i^+ = v_i - \gamma \nabla_v F_i(u, v_i)$$
  $$u_i^+ = u - \gamma \nabla_u F_i(u, v_i)$$
For smooth, nonconvex functions, we have the rates:

**Theorem** [P., Malik, Mohamed, Rabbat, Sanjabi, Xiao]

Alternating update: \( \frac{\sigma_1^2}{\sqrt{t}} \)

Simultaneous update: \( \frac{\sigma_2^2}{\sqrt{t}} \)

where \( \sigma_1^2 < \sigma_2^2 \) under typical scenarios

**Alternating update**

\[
\begin{align*}
    v_i^+ &= v_i - \gamma \nabla_v F_i(u, v_i) \\
    u_i^+ &= u - \gamma \nabla_u F_i(u, v_i)
\end{align*}
\]

**Simultaneous update**

\[
\begin{align*}
    v_i^+ &= v_i - \gamma \nabla_v F_i(u, v_i) \\
    u_i^+ &= u - \gamma \nabla_u F_i(u, v_i)
\end{align*}
\]

Experimentally, small but consistent trend of alternating > simultaneous
Experiments

Next word prediction

Speech recognition

Landmark detection

$y$-axis shows error: lower is better
Recall: robust aggregation experiments

1.4pp gap at zero corruption
Improving robust aggregation with personalization

Shared + Personal

EMNIST Linear

0.3pp gap at zero corruption

Test Accuracy

Corruption Level

Robust + pers.

Robust + no pers.

Non-robust + pers. (baseline)
Summary
Part 1: Heterogeneity-aware objectives for federated learning

Heterogeneity $\rightarrow$
large tail errors
Part 1: Heterogeneity-aware objectives for federated learning

Heterogeneity $\rightarrow$ large tail errors

$$\min_w S_\theta\left( (F_1(w), \ldots, F_n(w)) \right)$$

$S_\theta(Z) = \mathbb{E}[Z \mid Z > Q_\theta(Z)]$
Part 1: Heterogeneity-aware objectives for federated learning

Heterogeneity →
large tail errors

\[
\min_w S_\theta \left( (F_1(w), \ldots, F_n(w)) \right)
\]

Our approach reduces tail error

![Histogram showing low and high error counts](Image)

- Usual
- Ours

![Misclassification Error plot](Image)
Part 2: Robust aggregation for federated learning

Arithmetic mean $\rightarrow$
not robust to
poisoned updates

<table>
<thead>
<tr>
<th>Corruption Level</th>
<th>Ours</th>
<th>Usual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>0.1</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>0.2</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>0.3</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td>0.4</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>0.5</td>
<td>0.65</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Part 2: Robust aggregation for federated learning

\[ \text{GM} = \arg \min_z \sum_{i=1}^{m} \|z - w_i\|_2 \]

Arithmetic mean → not robust to poisoned updates

Data poisoning → Model poisoning

Test Accuracy

- Usual
- Ours

Corruption Level

- 0
- 0.1
- 0.2
- 0.3
- 0.4
- 0.5
- 0.6
Part 2: Robust aggregation for federated learning

Arithmetic mean ➞ not robust to poisoned updates

GM = \arg \min_z \sum_{i=1}^{m} \|z - w_i\|_2

Our approach gives greater robustness

![Graph showing comparison between Ours and Usual approaches](image)
Part 3: Model personalization for federated learning

Heterogeneity & robustness at odds

![Graph showing accuracy vs corruption level for robust and usual models.](image-url)
Part 3: Model personalization for federated learning

\[
\min_{u,v_1,\ldots,v_n} \frac{1}{n} \sum_{i=1}^{n} F_i(u, v_i)
\]
Part 3: Model personalization for federated learning

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

Heterogeneity & robustness at odds

Can tailor to heterogeneity & retain robustness

<table>
<thead>
<tr>
<th>Corruption Level</th>
<th>Robust</th>
<th>Usual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.67</td>
<td>0.68</td>
</tr>
<tr>
<td>0.05</td>
<td>0.69</td>
<td>0.7</td>
</tr>
<tr>
<td>0.1</td>
<td>0.71</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Perspectives and conclusion
The small data problem

- Per-client evaluations are not reliable
- Personalization can overfit

![Sent140 dataset statistics](image)

Does personalization improve per-client accuracy?

![Graph showing data per client vs. Δ Accuracy](image)
Understanding heterogeneity

Many negative results: optimization can slow down, makes robustness harder, ...
Yet, federated learning works.
Understanding heterogeneity

Many negative results: optimization can slow down, makes robustness harder, ...
Yet, federated learning works.

Quantify heterogeneity:

Measure gaps between distributions: MAUVE

Liu, P., Welleck, Oh, Choi, Harchaoui. NeurIPS (2021)]
Understanding heterogeneity

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

Yet, federated learning works.

Quantify heterogeneity:

Measure gaps between distributions: **MAUVE**

Statistical assumptions under which heterogeneity is benign?

What measures of heterogeneity impact optimization?

Tension between heterogeneity and privacy

Thank you!