Dubhe: Towards Data Unbiasedness with Homomorphic Encryption in Federated Learning Client Selection

Shulai Zhang, Zirui Li, Quan Chen, Wenli Zheng, Jingwen Leng, Minyi Guo
Shanghai Jiao Tong University
Background

Federated Learning

- **Demand:** Security promise
  - Homomorphic Encryption

- **Problem:** Statistical heterogeneity
  - Local data skewness
  - Data discrepancy among clients
  - Global data skewness

- **Method:** Client selection
  - Random selection
  - Greedy selection

The framework of Federated Learning
Background

Federated Learning

**Demand:** Security promise
- Homomorphic Encryption

**Problem:** Statistical heterogeneity
- Local data skewness
- Data discrepancy among clients
- Global data skewness

**Method:** Client selection
- Random selection
- Greedy selection

The framework of Federated Learning
Motivation

Federated Learning

- **Demand: Security promise**
  - Homomorphic Encryption

- **Problem: Statistical heterogeneity**
  - Local data skewness
  - Data discrepancy among clients
  - Global data skewness

- **Method: Client selection**
  - Random selection
  - Greedy selection

Discrepancy of clients in CIFAR10 classification
Motivation

Federated Learning

- **Demand**: Security promise
  - Homomorphic Encryption

- **Problem**: Statistical heterogeneity
  - Local data skewness
  - Data discrepancy among clients
  - **Global data skewness**

- **Method**: Client selection
  - Random selection
  - Greedy selection

Global data skewness in CIFAR10 classification
Federated Learning

- **Demand: Security promise**
  - Homomorphic Encryption

- **Problem: Statistical heterogeneity**
  - Local data skewness
  - Data discrepancy among clients
  - Global data skewness

- **Method: Client selection**
  - Random selection
  - Greedy selection

\[
||\omega^f_{mT} - \omega^*_m|| \leq ||\omega^f_{mT} - \omega^c_{mT} + \omega^c_{mT} - \omega^*_m||
\]

\[
\leq \frac{1}{K} \sum_{i=1}^{K} \left[ ((1 + \eta \lambda)^T ||\omega^k_{(m-1)T} - \omega^c_{(m-1)T}||
\right.
\]

\[
+ \eta ||p_i^k - p_o||_1 \left( \sum_{j=2}^{T} g(\omega^c_{mT-j})(1+\eta \lambda)^{j-1} \right)
\]

\[
+ (1 + \eta \lambda)^T ||\omega^c_{(m-1)T} - \omega^*_m||
\]

\[
+ \eta ||p_o - p_u||_1 \left( \sum_{j=1}^{T} g(\omega^*_m)(1+\eta \lambda)^{j-1} \right)
\]

Target: \( \min ||p_o - p_u||_1 \)
The Design of Dubhe

The framework of Dubhe
The Design of Dubhe

*Registration and Probability Calculation*

- The *registry* encodes the client’s data distribution information in a one-hot manner.
- In classification problems, the registry encodes each client’s data distribution by its *dominating classes*.

**Samples of a 10-class classification**

**Data distribution**

**Registry**
The Design of Dubhe

*Registration and Probability Calculation*

**Dubhe’s Client Selection**

- **Client α**
- **Client β**
- **Agent c**
- **Server**

**Registration and Probability Calculation**

- **0/1**

**Multi-time selection**

- Start training
- Drop out
- Start training

**Start training**

**Client Determination**
The Design of Dubhe

*Registration* and *Probability Calculation*

**Dubhe’s Client Selection**

Encrypt the *Registry* with Public Key

\[
\begin{align*}
R_0^{(t,a)} & \rightarrow 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\
R_1^{(t,a)} & \rightarrow 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\
R_2^{(t,a)} & \rightarrow 1 & 0 & \cdots & 0 & 0 & \cdots & 0
\end{align*}
\]

Decrypt with Private Key & *Probability Calculation*

\[
\begin{align*}
f(R^{(t,a)}, R^t_A) & \rightarrow f(R^{(t,b)}, R^t_A) \\
f(R^{(t,c)}, R^t_A)
\end{align*}
\]

0 is flipped to 1

\[
\begin{align*}
Enc(R^t_A) & \rightarrow 12 & 53 & 8 & \cdots & 17 & 26 & 11 & \cdots & 0
\end{align*}
\]

**Multi-time selection**

Client Determination

Client Determination

Drop out

training

training
Evaluation of Dubhe

Test accuracy curves on MNIST and CIFAR10

Average accuracy
Evaluation of Dubhe

The data balancing performance of Dubhe is approaching the performance of the greedy selection.

Results on FEMNIST
The Design of Dubhe

Multi-time selection

Dubhe’s Client Selection

Client $a$

Client $b$

Agent $c$

Server

Registration and Probability Calculation

Client Determination

Multi-time selection

Start training

Drop out

Start training

Multi-time selection

Client Determination
The Design of Dubhe

Multi-time selection

Dubhe’s Client Selection

Client a

Client b

Agent c

Server

Multi-time selection
Repeat $H$ times

$p_i^{(t,a)}$

$p_o$

Client Determination

$\text{Enc}(p_o)$

Registration and Probability Calculation

0/1

0/1

0/1

training

Drop out

training

$\%$ 

$+$, $-$
Evaluation of Dubhe

*Multi-time selection*

Results with multi-time selection. $M$ refers to MNIST and $C$ refers to CIFAR10.

<table>
<thead>
<tr>
<th>H</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>opt</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{EMD}^*$</td>
<td>0.2946</td>
<td>0.2588</td>
<td>0.2176</td>
<td>0.1971</td>
<td><strong>0.1750</strong></td>
<td>0.0144</td>
</tr>
<tr>
<td>$\text{Acc}^M$</td>
<td>0.9662</td>
<td>0.9668</td>
<td>0.9665</td>
<td><strong>0.9684</strong></td>
<td>0.9678</td>
<td>0.9694</td>
</tr>
<tr>
<td>$\beta^M$</td>
<td>0.0%</td>
<td>17.6%</td>
<td>10.5%</td>
<td><strong>69.5%</strong></td>
<td>51.5%</td>
<td>100%</td>
</tr>
<tr>
<td>$\text{Acc}^C$</td>
<td>0.4300</td>
<td>0.4518</td>
<td>0.4486</td>
<td>0.4441</td>
<td><strong>0.4577</strong></td>
<td>0.5295</td>
</tr>
<tr>
<td>$\beta^C$</td>
<td>0.00%</td>
<td>14.8%</td>
<td>12.6%</td>
<td>9.5%</td>
<td><strong>18.8%</strong></td>
<td>100%</td>
</tr>
</tbody>
</table>

$\text{EMD}^*$ decreases with larger $H$, thereby improving the model accuracy.
Evaluation of Dubhe

Registry sparsity

There are 3 out of 1000 clients who have class 2 and class 9 dominated.

\[ \kappa(u = (2, 9)) = 3 \]

• An example of the overall registry and its corresponding participated class proportion.

The sparsity of the overall registry is the primary cause of this deviation from expectation.
Conclusions

- The impact of data skewness on the performance degradation in FL is mathematically demonstrated.
- We propose **Dubhe**, a proactive client selection system to balance skewed data
  - pluggable, adaptive and robust to various FL settings
  - with negligible encryption and communication overhead
  - improves the training performance without bringing security threats.
Thanks for listening!

Any questions? Comments? Please contact:
Shulai Zhang: zslzsl1998@sjtu.edu.cn