About me

- Ph.D. candidate at NUS and funded by A*STAR,
  - advised by Prof. Bryan Kian Hsiang Low and Dr. Chuan-Sheng Foo
- Research interests: collaborative machine learning, data & model valuation
- Homepage: [https://xinyi-xu.com](https://xinyi-xu.com)
Agenda:

1. Motivations and overview of Collaborative Machine Learning (CML)
2. Our NeurIPS’21 paper: “Gradient-Driven Rewards to Guarantee Fairness in Collaborative Machine Learning”
3. Brief follow-up on our most recent works addressing some open issues.
4. Q & A
Importance of data in ML

❖ “Supervised learning, while successful in a wide variety of tasks, typically requires a large amount of human-labeled data …” - Yoshua Bengio, Geoffrey Hinton, and Yann LeCun [1].

❖ “In many industries where giant data sets simply don’t exist, I think the focus has to shift from big data to good data …” - Andrew Ng [2].

For ML to be effective, a large amount of good/high-quality data are needed.

Motivations for CML

- **Quantity** - Distributing the burden of data collection (in cross-silo FL) or effectively utilizing naturally distributed data (e.g., in cross-device FL).

- **Quality** - Data valuation to identify good/high-quality data in the specific ML use-cases.
  - The same data are not equally valuable for different ML algorithms; the same data are not equally valuable if others have access.
  - Another application is for pricing in AI marketplace.
Collaborative Machine Learning (CML)
Gradient-Driven Rewards to Guarantee Fairness in Collaborative Machine Learning

Xinyi Xu¹⁵, Lingjuan Lyu², Xingjun Ma³, Chenglin Miao⁴, Chuan Sheng Foo⁵, Bryan Kian Hsiang Low¹

Department of Computer Science, National University of Singapore¹
Sony AI²
School of Computer Science, Fudan University³
Department of Computer Science, University of Georgia⁴
Institute of Infocomm Research, Agency for Science, Technology and Research (A*STAR)⁵
Federated learning (FL)

Suppose $N$ self-interested and honest agents, each with a local dataset $D_i$. The federated objective is:

$$ w^* = \arg \min_w \sum_i p_i F(w; D_i) $$

In iteration $t$:

For Agent $i$:

$$ \Delta w_{i,t} \leftarrow -\eta \nabla F(w_{i,t}; D_i) $$

$$ w_{i,t+1} \leftarrow w_{i,t} + u_{N,t} $$

For Server:

$$ u_{N,t} \leftarrow \sum_i p_i \Gamma \frac{\Delta w_{i,t}}{||\Delta w_{i,t}||} $$

$p_i$ is an importance coefficient, $\Gamma$ is a normalizing constant and $N := \{i; 1 \leq i \leq N\}$ denotes all the agents.

Should every agent be rewarded equally? otherwise how?
Fair training-time rewards

Instead of rewarding all the agents equally, reward them fairly: Agents that upload more valuable gradients are rewarded better.

- Incentivize the agents to collect more data of higher quality.

1. How to determine the values of (the gradients of) the agents fairly?
2. How to guarantee the rewards are fair?
Fair training-time rewards

1. How to determine the values of (the gradients of) the agents fairly?

The **Shapley value (SV)** with several intuitive fairness properties.

- **null player**: if an agent uploads non-valuable gradients, the corresponding SV is zero.
- **symmetry**: if two agents upload identical (equally valuable) gradients, their corresponding SVs are equal.
2. How to guarantee the rewards are fair?

A higher $SV$ leads to a better downloaded gradient.

For an agent $i$:

- contributing more (while others remain the same) leads to a better reward;
- contributing more than agent $j$ leads to a better reward than agent $j$. 
Fair **training-time** rewards

2. How to guarantee the rewards are fair?

In each iteration, the agents are **rewarded with carefully managed gradients**.

- inherent rewards: no need for additional external resources;
- the agents do not need to wait till the end [1,2];
- **local-to-global**: fairness in each iteration $\rightarrow$ **fairness** overall (Theorem 2).

Cosine gradient Shapley value (CGSV)

Definition 1 (Cosine gradient Shapley value (CGSV)). Let $\Pi_N$ be a set of all possible permutations of $N$ and $S_{\pi,i}$ be the coalition of agents preceding agent $i$ in permutation $\pi \in \Pi_N$. The CGSV of agent $i \in N$ is defined as

$$\phi_i := (1/N!) \sum_{\pi \in \Pi_N} [\nu(S_{\pi,i} \cup \{i\}) - \nu(S_{\pi,i})].$$  

(2)

The gradient valuation function: $\nu(S) = \cos(u_S, u_N)$ where $u_i \leftarrow \Gamma \frac{\Delta w_i}{\|\Delta w_i\|}$, $u_S \leftarrow \sum_{i \in S} p_i u_i$.

- $\nu(S) = \cos(u_S, u_N) = \cos(\theta_S)$
- $\nu(S') = \cos(u_{S'}, u_N) = \cos(\theta_{S'})$
- $\nu(S) > \nu(S')$

The CGSV $\phi_i$ of an uploaded gradient $u_i$ (i.e., contribution from agent $i$) is evaluated via the vector alignment between $u_i$ and $u_N$, via the cosine similarity [1].

Efficiently Approximating CGSV

- Computing the exact CGSV incurs $O(2^N D)$ which is practically infeasible for larger $N$.

- We provide an efficient approximation (with a bounded error) as:

  $$\phi_i \approx \psi_i = \cos(u_i, u_N)$$

**Theorem 1 (Approximation Error).** Let $I \in \mathbb{R}^+$. Suppose that $\|u_i\| = \Gamma$ and $|\langle u_i, u_N \rangle| \geq 1/I$ for all $i \in N$. Then, $\phi_i - L_i \psi_i \leq I \Gamma^2$ where the multiplicative factor $L_i$ can be normalized away.

- Intuition: exploit linearity of CGSV and linearity of cosine similarity to “branch and bound”.

- It reduces the complexity to $O(ND)$ and we empirically demonstrate its effectiveness against a Monte Carlo sampling-based $(\epsilon, \delta)$—approximation.
Efficiently Approximating CGSV

- We compare $\ell_1, \ell_2$ errors with the exact value and runtime against $N$ and $D$.
- Solid lines denote our approximation and lower is better.
- Our approximation performs better for all 3 metrics and the performance gap widens as $N$ increases.
Server-Side Training-Time Gradient Reward Mechanism

- **Gradient aggregation** *(by Server)*
  - Update the contribution:
    
    $$ r_{i,t} \leftarrow \alpha r_{i,t-1} + (1 - \alpha) \psi_{i,t} , \quad r_{i,t} \leftarrow \frac{r_{i,t}}{\sum_{i' \in \mathcal{N}} r_{i',t}} $$
    
    - The cumulative update over iterations helps reduce fluctuations and provide a smoother estimate of the contributions of the agents.
  - Compute the aggregate gradient:
    
    $$ u_{\mathcal{N},t} \leftarrow \sum_i r_{i,t} u_{i,t} $$
    
    - $r_{i,t}$ is then used as the importance coefficient to aggregate the gradient.
Server-Side Training-Time Gradient Reward Mechanism

- **Gradient download** (for Agent \( i \))
  - Calculate the fair gradient reward s.t., “A higher SV leads to a better downloaded gradient.”
  
  \[
  \mathbf{v}_{i,t} \leftarrow \text{mask}(\mathbf{u}_{N,t}, q_{i,t}) \quad q_{i,t} \leftarrow \frac{D \tanh(\beta r_{i,t})}{\max_i \tanh(\beta r_{i,t})} \in [0, 1]
  \]

  - **sparsification**: \( \text{mask}(\mathbf{u}, q) \) retains the largest \( \max(0, q) \) components in magnitude of \( \mathbf{u} \) and zeros out all the rest. Lower sparsification (higher \( q_{i,t} \)) \( \Leftrightarrow \) better downloaded gradient.
  
  - \( q_{i,t} \) is max-normalized cumulative SV: higher SV \( \Leftrightarrow \) higher \( r_{i,t} \) \( \Leftrightarrow \) higher \( q_{i,t} \).
  
  - **altruism degree** \( \beta \) quantifies how much an agent with lower contributions benefit larger \( \beta \) \( \Leftrightarrow \) more altruistic/equitable while smaller \( \beta \) \( \Leftrightarrow \) stricter fairness.

- Update local model: \( \mathbf{w}_{i,t} \leftarrow \mathbf{w}_{i,t-1} + \mathbf{v}_{i,t} \)
Putting it all together

\[ \Delta w_{1,t+1} = -\eta \nabla F(w_{1,t}; D_1) \]

\[ w_{1,t+1} = w_{1,t} + \nu_{1,t} \]

\[ r_{1,t} = \alpha r_{1,t-1} + (1 - \alpha)\psi_{1,t} \]

\[ q_{1,t} = \left[ \frac{D \tanh(\beta r_{1,t})}{\max_t \tanh(\beta r_{1,t})} \right] \]

\[ \psi_{1,t} = \cos(\theta_1) \]

\[ \mathbf{u}_N,t : \begin{bmatrix} w_1 \ w_2 \ w_3 \ w_4 \ \vdots \ w_D \end{bmatrix} \]

\[ \text{mask}(\mathbf{u}_N,t, q_{1,t}) : \begin{bmatrix} 1 & 0 & 0 & 1 \ \vdots \ & \end{bmatrix} \]

\[ \mathbf{v}_{1,t} : \begin{bmatrix} w_1 \ 0 \ 0 \ w_4 \ \vdots \ w_D \end{bmatrix} \]
Global Fairness Guarantee

Theorem 2 (Fairness in Model Performance). Define $\delta_{i,t} := \|w_{N,t} - w_{i,t}\|$ and $w_{N,t}$ is near a stationary point of $F()$ and some regularity conditions on the objective function $F(\cdot)$. For any $t \in \mathbb{Z}^+$ and $\forall i, i' \in N$, if $r_{i,t} \geq r_{i',t}$ and $\delta_{i',t-1} - \delta_{i,t-1} \geq 2\|v_{i,t}\|$, then $F(w_{i,t}) \leq F(w_{i',t}).$

- Local fairness to global fairness:
  - An agent that uploads better gradients can download better gradients (locally fair), and as a result, this agent receives a better-performing model (globally fair).

- Intuition:
  - all agents start with the same model: $w_0$
  - agents with higher $r_{i,t}$ have less deviation from the trajectory: $\{w_0 + \sum_{l=1}^{t} u_{N,l}\}_t$
Experimental setup & baselines

- Datasets
  - MNIST, CIFAR-10, Movie Reviews, Stanford Sentiment Treebank

- Comparison baselines
  - FedAvg [1], and its variants
  - q-FFL [2], CFFL [3]
  - Shapley value-based: Extended contribution index (ECI) [4]
  - Euclidean distance variant instead of cosine similarity

- Data partitions
  - uniform (UNI)
  - powerlaw (POW)
    - Individual datasets of different sizes
  - classimbalance (CLA)
    - Individual datasets with different available classes
      - e.g. MNIST, for N=5, the agents have {1,3,5,7,10} classes respectively

Fairness evaluation metric

Pearson correlation coefficient between standalone performance & final local model performance.

- **Standalone performance** provides an estimate of the quality of the local dataset and thus the quality of the contribution (via uploaded gradients) by the agents.
- **Final local model performance** represents the rewards the agents receive at the end.

A correlation close to 1 indicates the rewards are commensurate with the contributions (i.e., fair), and validates Theorem 2.
## Fairness results

| No. Agents | MNIST | | | CIFAR-10 | | | MR | | | SST | |
|------------|-------|---|---|-----------|---|---|---|---|---|---|
|            | 10    | 20 | | 10       | | 5  | | 5  | | |
| Data Partition | UNI | POW | CLA | UNI | POW | CLA | UNI | POW | CLA | UNI | POW |
| FedAvg     | -45.60 | 55.24 | 24.12 | 0.85 | -32.58 | 40.83 | 18.47 | 97.48 | 98.75 | 48.68 | 57.50 |
| q-FFL      | -44.73 | 39.00 | 22.38 | -22.01 | 38.71 | 48.07 | -17.64 | 51.33 | 94.06 | 56.43 | -75.92 |
| CFFL       | 83.57 | 91.80 | 81.24 | 82.52 | 94.70 | 85.71 | 78.25 | 72.55 | 81.31 | 96.85 | 93.34 |
| ECI        | 85.26 | **99.83** | **99.98** | 80.95 | **99.41** | 95.21 | 75.85 | 79.50 | 99.55 | 97.69 | 95.00 |
| DW         | 89.15 | 98.93 | 65.34 | 86.94 | 99.63 | 35.21 | -23.14 | 91.97 | 45.45 | **99.20** | 97.12 |
| RR         | 83.77 | 71.17 | -26.75 | -18.64 | 25.47 | 95.86 | 30.67 | 0.70 | 90.67 | 44.16 | -25.11 |
| Ours (EU)  | 84.25 | 98.25 | 99.82 | 80.55 | 97.77 | **99.97** | 78.25 | 94.24 | 94.95 | 97.58 | 93.21 |
| Ours ($\beta = 1$) | 94.03 | 95.74 | 94.54 | 84.47 | 96.39 | 97.23 | **98.80** | **98.78** | **99.89** | 96.01 | 98.20 |
| Ours ($\beta = 1.2$) | 94.75 | 97.28 | 96.23 | 90.52 | 97.72 | 95.21 | 91.07 | 91.59 | 99.82 | 96.12 | **98.47** |
| Ours ($\beta = 1.5$) | **96.34** | 86.99 | 95.37 | 82.68 | 90.94 | 98.75 | 93.55 | 93.78 | 95.89 | 95.32 | 97.88 |
| Ours ($\beta = 2$) | 94.66 | 91.20 | 95.38 | **96.90** | 91.33 | 94.32 | 89.80 | 88.78 | 93.39 | 92.22 | 95.74 |
Fairness results

- **MNIST (UNI) β = 2**
- **MNIST (POW) β = 2**
- **MNIST (CLA) β = 2**

- **CIFAR-10 (UNI) β = 2**
- **CIFAR-10 (POW) β = 2**
- **CIFAR-10 (CLA) β = 2**
Fairness results

Increasing *altruism degree* $\beta$ “pushes” the training losses of all agents to be more equitably low, and it improves the performance of agents with relatively lower contributions.
# Accuracy results (on test set)

<table>
<thead>
<tr>
<th>No. Agents</th>
<th>MNIST</th>
<th></th>
<th></th>
<th>CIFAR-10</th>
<th></th>
<th></th>
<th>MR</th>
<th></th>
<th></th>
<th>SST</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Partition</td>
<td>10</td>
<td>20</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Standalone</td>
<td>UNI</td>
<td>POW</td>
<td>CLA</td>
<td>UNI</td>
<td>POW</td>
<td>CLA</td>
<td>UNI</td>
<td>POW</td>
<td>CLA</td>
<td>UNI</td>
<td>POW</td>
</tr>
<tr>
<td>FedAvg</td>
<td>91 (91)</td>
<td>88 (92)</td>
<td>53 (92)</td>
<td>91 (91)</td>
<td>89 (92)</td>
<td>48 (90)</td>
<td>46 (47)</td>
<td>43 (49)</td>
<td>31 (44)</td>
<td>47 (56)</td>
<td>31 (34)</td>
</tr>
<tr>
<td>q-FFL</td>
<td>93 (94)</td>
<td>92 (94)</td>
<td>53 (93)</td>
<td>93 (93)</td>
<td>92 (94)</td>
<td>49 (92)</td>
<td>48 (48)</td>
<td>47 (50)</td>
<td>32 (47)</td>
<td>51 (63)</td>
<td>33 (35)</td>
</tr>
<tr>
<td>CFFL</td>
<td>85 (91)</td>
<td>27 (45)</td>
<td>44 (64)</td>
<td>88 (91)</td>
<td>48 (53)</td>
<td>40 (59)</td>
<td>41 (46)</td>
<td>36 (36)</td>
<td>22 (28)</td>
<td>12 (18)</td>
<td>23 (25)</td>
</tr>
<tr>
<td>ECI</td>
<td>90 (92)</td>
<td>85 (90)</td>
<td>34 (44)</td>
<td>91 (93)</td>
<td>88 (91)</td>
<td>39 (46)</td>
<td>39 (41)</td>
<td>35 (45)</td>
<td>22 (40)</td>
<td>44 (53)</td>
<td>31 (32)</td>
</tr>
<tr>
<td>DW</td>
<td>94 (94)</td>
<td>92 (94)</td>
<td>53 (94)</td>
<td>94 (94)</td>
<td>92 (94)</td>
<td>49 (92)</td>
<td>49 (49)</td>
<td>47 (51)</td>
<td>31 (46)</td>
<td>56 (61)</td>
<td>33 (34)</td>
</tr>
<tr>
<td>RR</td>
<td>93 (94)</td>
<td>92 (94)</td>
<td>53 (94)</td>
<td>93 (93)</td>
<td>92 (94)</td>
<td>49 (92)</td>
<td>48 (48)</td>
<td>47 (50)</td>
<td>32 (47)</td>
<td>51 (62)</td>
<td>33 (35)</td>
</tr>
<tr>
<td>Ours (EU)</td>
<td>94 (94)</td>
<td>94 (94)</td>
<td>54 (94)</td>
<td>94 (94)</td>
<td>94 (94)</td>
<td>49 (92)</td>
<td>49 (49)</td>
<td>49 (51)</td>
<td>32 (46)</td>
<td>54 (59)</td>
<td>34 (36)</td>
</tr>
<tr>
<td>Ours ($\beta = 1$)</td>
<td>96 (97)</td>
<td>94 (95)</td>
<td>74 (95)</td>
<td>95 (96)</td>
<td>96 (97)</td>
<td>65 (93)</td>
<td>61 (62)</td>
<td>60 (62)</td>
<td>35 (54)</td>
<td>62 (76)</td>
<td>35 (36)</td>
</tr>
<tr>
<td>Ours ($\beta = 1.2$)</td>
<td>94 (95)</td>
<td>95 (95)</td>
<td>75 (95)</td>
<td>96 (96)</td>
<td>96 (97)</td>
<td>65 (93)</td>
<td>61 (62)</td>
<td>60 (62)</td>
<td>35 (54)</td>
<td>62 (75)</td>
<td>34 (37)</td>
</tr>
<tr>
<td>Ours ($\beta = 1.5$)</td>
<td>97 (97)</td>
<td>95 (95)</td>
<td>75 (95)</td>
<td>96 (97)</td>
<td>94 (95)</td>
<td>65 (93)</td>
<td>61 (62)</td>
<td>59 (62)</td>
<td>35 (54)</td>
<td>62 (74)</td>
<td>35 (37)</td>
</tr>
<tr>
<td>Ours ($\beta = 2$)</td>
<td>96 (96)</td>
<td>95 (96)</td>
<td>73 (94)</td>
<td>97 (97)</td>
<td>95 (96)</td>
<td>66 (95)</td>
<td>62 (62)</td>
<td>61 (62)</td>
<td>36 (54)</td>
<td>62 (75)</td>
<td>35 (37)</td>
</tr>
</tbody>
</table>

Average (maximum) test accuracies over all agents.
## Runtime results

| No. Agents | MNIST | | | CIFAR-10 | | | MR | | | SST |
|------------|-------|| | | | | | | | | |
|            | 5     | 10 | 20 | 5     | 10 | 5     | | | | |
| FedAvg     | 1.17 (7e-3) | 1.05 (1e-2) | 4.29 (1e-2) | 1.66 (7e-3) | 7.41 (1e-2) | 1.3 (1e-4) | 1.31 (6e-4) |
| q-FFL      | 6.14 (4e-2) | 4.97 (5e-2) | 91.20 (0.3) | 97.28 (0.4) | 58.94 (7e-2) | 90.01 (8e-3) | 82.85 (4e-2) |
| CFFL       | 32.15 (0.2) | 21.79 (0.3) | 500.03 (1.6) | 570.12 (2.0) | 302.44 (0.4) | 479.12 (0.2) | 487.71 (2e-1) |
| ECI        | 2377.33 (16) | 11937.80 (141) | 23749.06 (74) | 3571.75 (15) | 58835.83 (84) | 422.85 (4e-2) | 801.20 (0.4) |
| DW         | 0.89 (6e-3) | 0.79 (9e-3) | 1.60 (5e-3) | 1.21 (5e-3) | 5.29 (7e-3) | 0.99 (1e-5) | 0.98 (5e-4) |
| RR         | 0.89 (6e-3) | 0.82 (9e-3) | 1.60 (5e-3) | 3.31 (1e-2) | 5.41 (7e-3) | 1.01 (5e-4) | 0.99 (5e-4) |
| Ours (EU)  | 0.89 (6e-3) | 0.81 (9e-3) | 1.61 (5e-3) | 1.22 (5e-3) | 5.33 (7e-3) | 1.01 (5e-4) | 0.99 (5e-4) |
| Ours (Cosine) | 6.34 (4e-2) | 4.94 (5e-2) | 94.30 (0.3) | 98.39 (0.4) | 54.94 (7e-2) | 89.81 (8e-3) | 82.87 (4e-2) |

Number of seconds (ratio w.r.t. training time).
Discussion

Fairness in rewards in action

- Each agent’s interest is protected, i.e., they get rewarded commensurately with their contributions measured in Shapley values.
- Flexibly control the proportionality between rewards and contributions, via $\beta$.
- Computational overhead at server is small.
Collaborative Machine Learning (CML)
## Latest Publications

<table>
<thead>
<tr>
<th></th>
<th>ML Algorithm</th>
<th>Resource</th>
<th>Desiderata</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeurIPS’21 Fair CML</td>
<td>gradient descent</td>
<td>functionals of data</td>
<td>fairness and training-time rewards</td>
</tr>
<tr>
<td>NeurIPS’21 Volume</td>
<td>ML model agnostic</td>
<td>statistics of data</td>
<td>fairness, replication robustness</td>
</tr>
<tr>
<td>AAAI’22</td>
<td>generative modeling</td>
<td>statistics of data (distributional divergence)</td>
<td>fairness, synthetic data generation</td>
</tr>
<tr>
<td>ICML’22</td>
<td>Bayesian parameter learning</td>
<td>statistics of data (Fisher information)</td>
<td>asymptotic fairness</td>
</tr>
</tbody>
</table>

- **IJCAI-ECAI2022 - Survey on Data Valuation**
  - Use cases: interpretable ML, active learning, adversarial data detection
  - Data valuation principles and desiderata
Thank you!