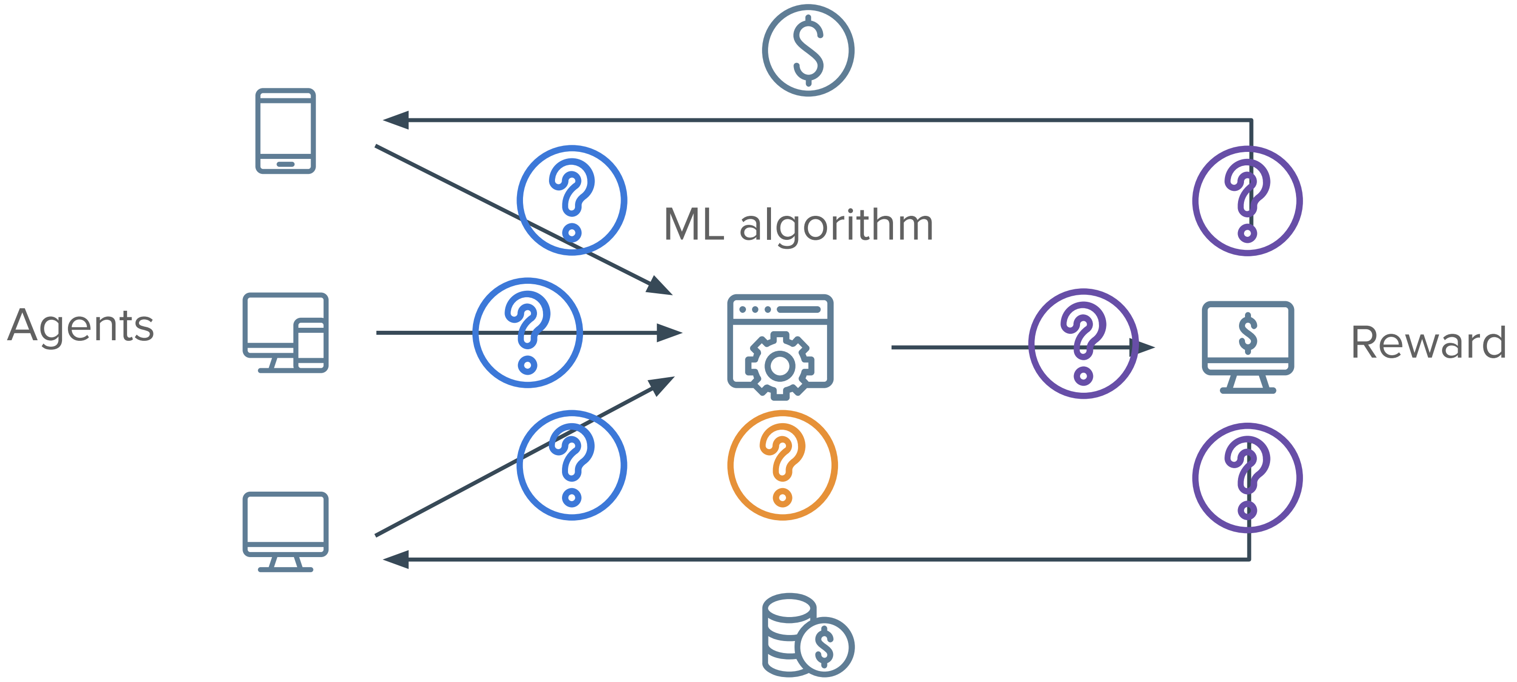
Beyond Standard FL: Gradient-Driven Rewards to Guarantee Fairness in Collaborative Machine Learning

Xinyi Xu[1,5], Lingjuan Lyu[2], Xingjun Ma[3], Chenglin Miao[4], Chuan Sheng Foo[5], Bryan Kian Hsiang Low[1]

Department of Computer Science, National University of Singapore[1]
Sony AI[2]
School of Computer Science, Fudan University[3]
Department of Computer Science, University of Georgia[4]
Institute of Infocomm Research, Agency for Science, Technology and Research (A*STAR)[5]

Collaborative Machine Learning (CML)



Federated learning (FL) with cross-siloed setting

Suppose *N* self-interested and honest agents, each with a local dataset $\mathcal{D}_i$. The federated objective is:

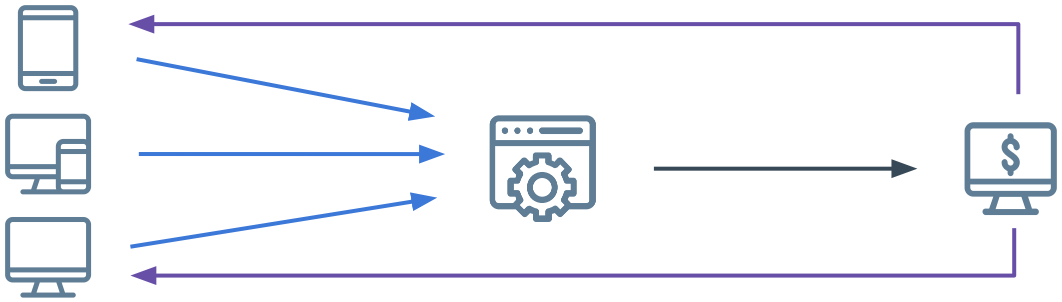
$$\boldsymbol{w}^* = \arg\min_w \sum_i p_i \mathbf{F}(\boldsymbol{w}; \mathcal{D}_i)$$

In <u>iteration *t*:</u>

For *Agent i:* $\boxed{\Delta \boldsymbol{w}_{i,t}} \leftarrow -\eta \nabla \mathbf{F}(\boldsymbol{w}_{i,t}; \mathcal{D}_i)$

For *Server:* $\boxed{\boldsymbol{u}_{\mathcal{N},t}} \leftarrow \sum_i p_i \Gamma \frac{\Delta \boldsymbol{w}_{i,t}}{\|\Delta \boldsymbol{w}_{i,t}\|}$

$$\boldsymbol{w}_{i,t+1} \leftarrow \boldsymbol{w}_{i,t} + \boxed{\boldsymbol{u}_{\mathcal{N},t}}$$



Should every agent be rewarded equally?

otherwise how?

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

Different notions of fairness in FL

- Algorithmic fairness [1]: whether the trained model makes predictions in a biased way towards certain sensitive features
- Equitable fairness [2]: whether the distribution of the performance of the agents/devices is highly spread out (the best are much better than the worst)
- Collaborative fairness [3,4]: whether the rewards the agents receive are commensurate with the contributions that they make

[1] A Survey on Bias and Fairness in Machine Learning. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan, ACM Computing Survey. 2022.
[2] Fair Resource Allocation in Federated Learning. Tian Li, Maziar Sanjabi, Ahmad Beirami, Virginia Smith. 2020, ICLR.
[3] Collaborative fairness in federated learning. Lingjuan Lyu, Xinyi Xu, Qian Wang. 2020, LNCS.
[4] Profit Allocation for Federated Learning. Tianshu Song, Yongxin Tong, Shuyue Wei, IEEE Big Data, 2019.

Fair training-time rewards

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

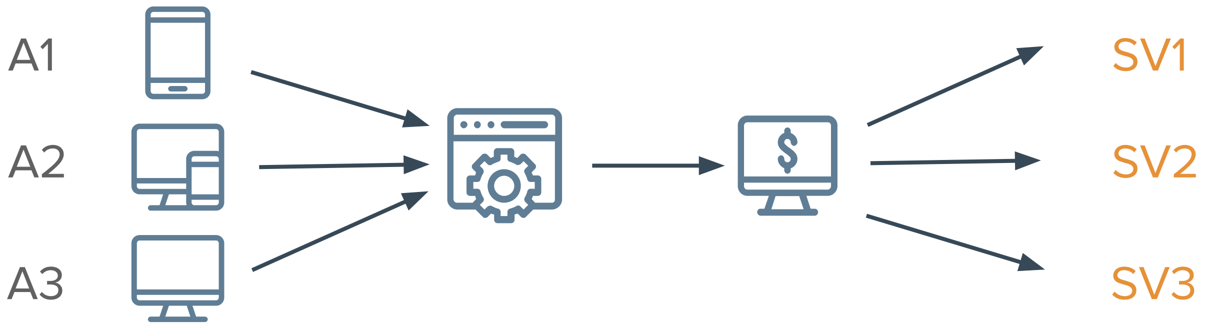
- Incentivize the agents to collect more data of higher quality.

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.



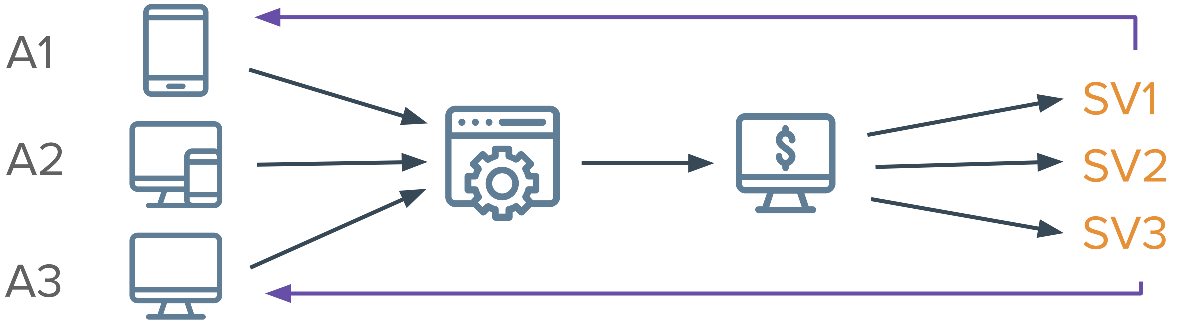
null player: if an agent uploads non-valuable gradients, the corresponding SV is zero.

symmetry: if two agents upload identical (equally valuable) gradients, their corresponding SVs are equal.

Fair training-time rewards

2. How to guarantee the rewards are fair?

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



For an agent *i*:

- <u>contributing more</u> (while others remain the same) leads to a <u>better</u> reward;
- <u>contributing more</u> than agent *j* leads to a <u>better</u> 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 ➔ **fairness** overall (Theorem 2).

[1] Profit Allocation for Federated Learning. Tianshu Song, Yongxin Tong, Shuyue Wei, IEEE Big Data, 2019.
[2] A Principled Approach to Data Valuation for Federated Learning. Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, Dawn Song, 2020, LNCS.

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

[1] Communication-Efficient Learning of Deep Networks from Decentralized Data. H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agüera y Arcas, 2017, AISTATS.
[2] Fair Resource Allocation in Federated Learning. Tian Li, Maziar Sanjabi, Ahmad Beirami, Virginia Smith. 2020, ICLR.
[3] Collaborative fairness in federated learning. Lingjuan Lyu, Xinyi Xu, Qian Wang. 2020, LNCS.
[4] Profit Allocation for Federated Learning. Tianshu Song, Yongxin Tong, Shuyue Wei, IEEE Big Data, 2019.

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 higher correlation (i.e., closer to 1) indicates better fairness: the **rewards** are commensurate with the **contributions**.

Fairness results (correlation * 100)

	MNIST						CIFAR-10			MR	SST
No. Agents	10			20			10			5	5
Data Partition	UNI	POW	CLA	UNI	POW	CLA	UNI	POW	CLA	POW	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

Takeaway: Our method achieves best/competitive fairness.

Accuracy results (on test set)

	MNIST						CIFAR-10			MR	SST
No. Agents	10			20			10			5	5
Data Partition	UNI	POW	CLA	UNI	POW	CLA	UNI	POW	CLA	POW	POW
Standalone	91 (91)	88 (92)	53 (92)	91 (91)	89 (92)	48 (90)	46 (47)	43 (49)	31 (44)	47(56)	31(34)
FedAvg	93 (94)	92 (94)	53 (93)	93 (93)	92 (94)	49 (92)	48 (48)	47 (50)	32 (47)	51(63)	33(35)
q-FFL	85 (91)	27 (45)	44 (64)	88 (91)	48 (53)	40 (59)	41 (46)	36 (36)	22 (28)	12(18)	23(25)
CFFL	90 (92)	85 (90)	34 (44)	91 (93)	88 (91)	39 (46)	39 (41)	35 (45)	22 (40)	44(53)	31(32)
ECI	94 (94)	92 (94)	53 (94)	94 (94)	92 (94)	49 (92)	49 (49)	47 (51)	31 (46)	56(61)	33(34)
DW	93 (94)	92 (94)	53 (93)	93 (93)	92 (94)	49 (92)	48 (48)	47 (50)	32 (47)	51(62)	33(35)
RR	94 (95)	**95** (95)	64 (72)	94 (95)	94 (95)	50 (56)	47 (59)	49 (51)	26 (29)	**63**(65)	**36**(36)
Ours (EU)	94 (94)	94 (94)	54 (94)	94 (94)	94 (94)	49 (92)	49 (49)	49 (51)	32 (46)	54(59)	34(36)
Ours ($\beta = 1$)	96 (97)	94 (95)	74 (**95**)	95 (96)	96 (97)	65 (93)	61 (**62**)	**60** (**62**)	35 (**54**)	62(**76**)	35(36)
Ours ($\beta = 1.2$)	94 (95)	**95** (95)	**75** (**95**)	96 (96)	**96** (**97**)	65 (93)	61 (**62**)	**60** (**62**)	35 (**54**)	62(75)	34(**37**)
Ours ($\beta = 1.5$)	**97** (**97**)	**95** (95)	**75** (**95**)	**96** (**97**)	94 (95)	65 (93)	61 (**62**)	59 (**62**)	35 (**54**)	62(74)	35(**37**)
Ours ($\beta = 2$)	96 (96)	**95** (**96**)	73 (94)	**97** (**97**)	95 (96)	**66** (**95**)	**62** (**62**)	**61** (**62**)	**36** (**54**)	62(75)	35(**37**)

Average (maximum) test accuracies over all agents.

Takeaway: Our method does not sacrifice predictive performance.

Runtime results

	MNIST			CIFAR-10		MR	SST
No. Agents	5	10	20	5	10	5	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)

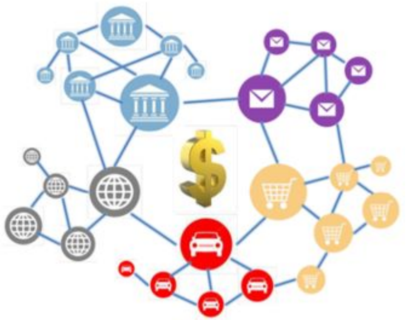
Number of seconds (ratio w.r.t. training time).

Takeaway: Our method is computationally efficient.

Discussion

One-liner summary: The server analyzes the **uploaded gradients** of the agents and carefully manages the **gradients the agents download**, to ensure what the agents **receive/download** is commensurate to what they **contribute/upload**.

- The commensurate relationship, i.e., fairness is via the **Shapley values**.
- Computational overhead at server is small.
- Predictive performance remains competitive.



Future directions

- How does fairness affect other properties: privacy, other notions of fairness, convergence?
- How to include the server (e.g., platform/algorithm provider) into the consideration instead of restricting to only the agents (e.g., clients)? E.g., how to fairly incentivize the server?

Thank you!

Find me at: https://xinyi-xu.com

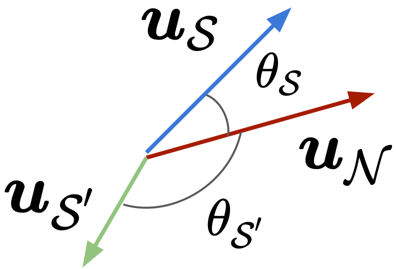
and poster **19**.

Cosine gradient Shapley value (CGSV)

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

$$\phi_i := (1/N!) \sum_{\pi \in \Pi_{\mathcal{N}}} \left[\nu(\mathcal{S}_{\pi,i} \cup \{i\}) - \nu(\mathcal{S}_{\pi,i})\right] . \quad (2)$$

The *gradient valuation function:* $\nu(\mathcal{S}) = \cos(\boldsymbol{u}_{\mathcal{S}}, \boldsymbol{u}_{\mathcal{N}})$ where $\boldsymbol{u}_i \leftarrow \Gamma \frac{\Delta \boldsymbol{w}_i}{\|\Delta \boldsymbol{w}_i\|}$, $\boldsymbol{u}_{\mathcal{S}} \leftarrow \sum_{i \in \mathcal{S}} p_i \boldsymbol{u}_i$



$$\nu(\mathcal{S}) = \cos(\boldsymbol{u}_{\mathcal{S}}, \boldsymbol{u}_{\mathcal{N}}) = \cos(\theta_{\mathcal{S}})$$

$$\nu(\mathcal{S}) > \nu(\mathcal{S}')$$

$$\nu(\mathcal{S}') = \cos(\boldsymbol{u}_{\mathcal{S}'}, \boldsymbol{u}_{\mathcal{N}}) = \cos(\theta_{\mathcal{S}'})$$

- The CGSV ϕ_i of an uploaded gradient $\boldsymbol{u}_i$ (i.e., contribution from agent *i*) is evaluated via the vector alignment between $\boldsymbol{u}_i$ and $\boldsymbol{u}_{\mathcal{N}}$, via the cosine similarity [1].

[1] A Reputation Mechanism Is All You Need: Collaborative Fairness and Adversarial Robustness in Federated Learning. Xinyi Xu, Lingjuan Lyu. 2021 FL-ICML workshop (Oral).

Efficiently Approximating CGSV

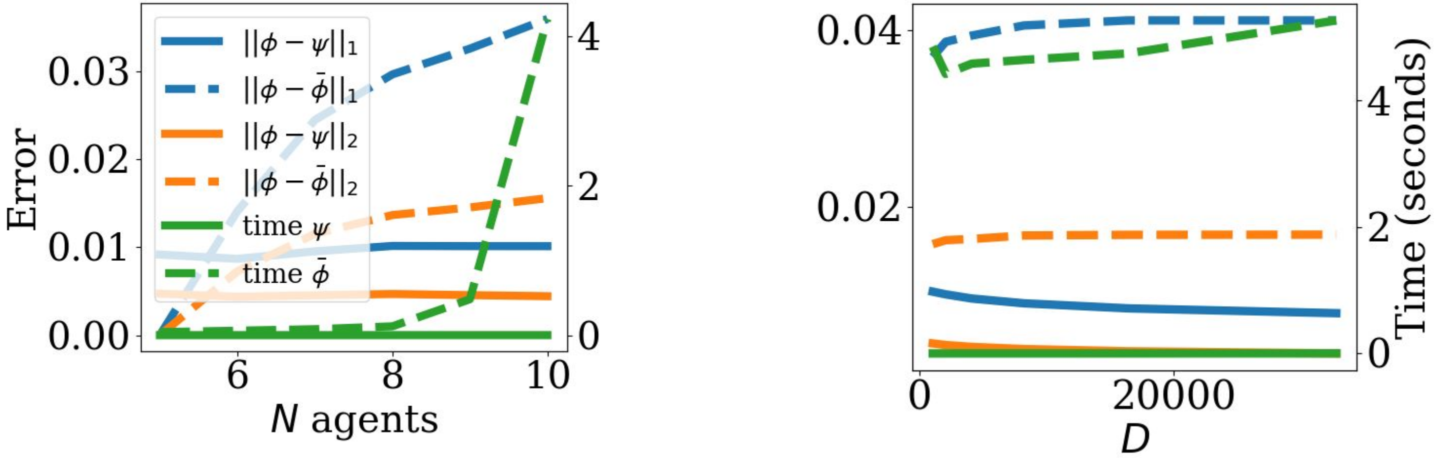
- Computing the exact CGSV incurs $\mathcal{O}(2^N D)$ which is practically infeasible for larger *N*.
- We provide an efficient approximation (with a bounded error) as:

$$\phi_i \approx \psi_i = \cos(\boldsymbol{u}_i, \boldsymbol{u}_{\mathcal{N}})$$

Theorem 1 (Approximation Error). *Let* $I \in \mathbb{R}^+$. *Suppose that* $\|\boldsymbol{u}_i\| = \Gamma$ *and* $|\langle \boldsymbol{u}_i, \boldsymbol{u}_{\mathcal{N}} \rangle| \geq 1/I$ *for all* $i \in \mathcal{N}$. *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 $\mathcal{O}(ND)$ and we empirically demonstrate its effectiveness against a Monte Carlo sampling-based $(\epsilon, \delta)-$approximation.

Efficiently Approximating CGSV



- We compare ℓ_1, ℓ_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:

$$r_{i,t} \leftarrow \alpha\, r_{i,t-1} + (1-\alpha)\, \psi_{i,t}\ ,\ r_{i,t} \leftarrow \frac{r_{i,t}}{\sum_{i' \in \mathcal{N}} r_{i',t}}$$

    - The cumulative update over iterations helps reduce fluctuations and provide a smoother estimate of the contributions of the agents.

  - Compute the aggregate gradient:

$$\boldsymbol{u}_{\mathcal{N},t} \leftarrow \sum_i r_{i,t} \boldsymbol{u}_{i,t}$$

    - $r_{i,t}$ is then used as the importance coefficient to aggregate the gradient.

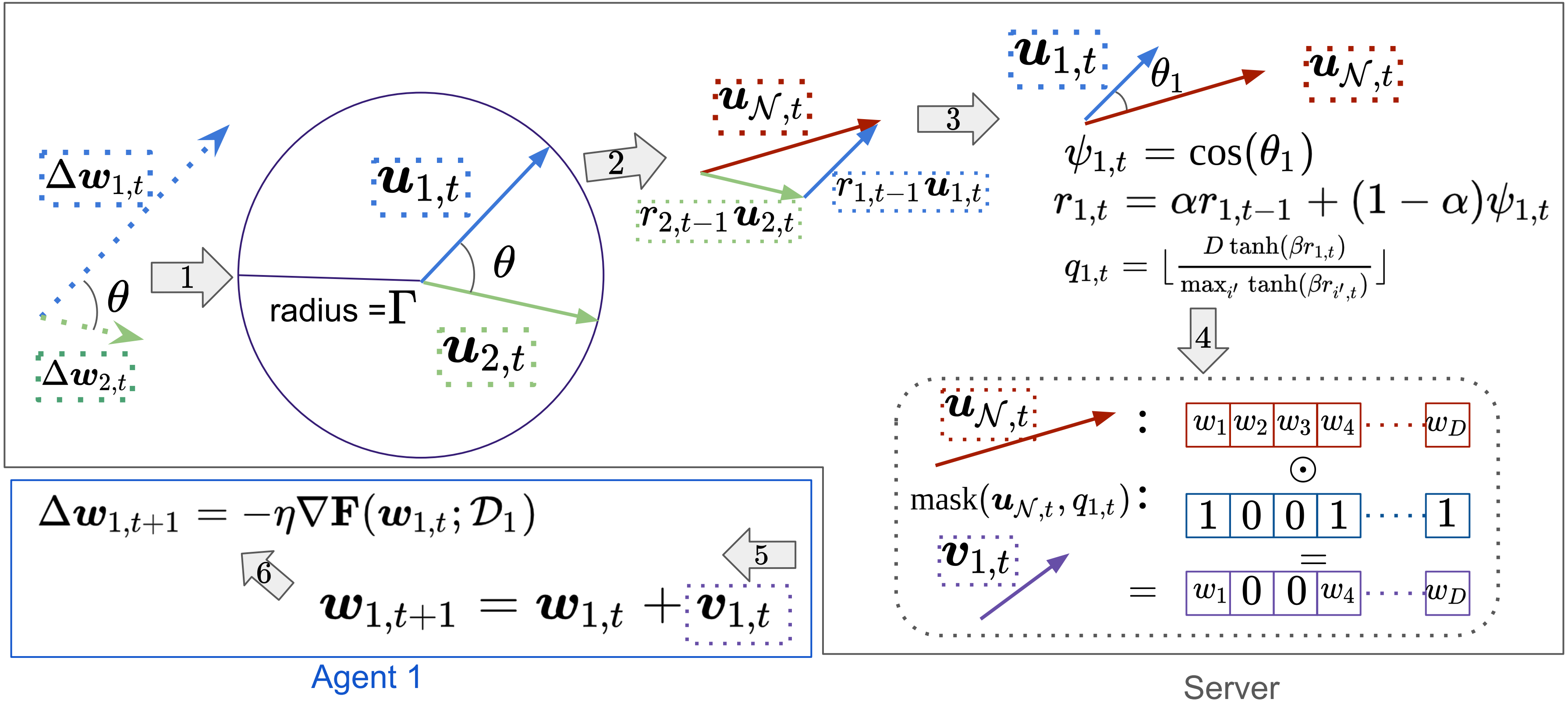
Server-Side Training-Time Gradient Reward Mechanism

- **Gradient download** (for *Agent i*)
  - Calculate the fair gradient reward s.t., "A higher **SV** leads to a better **downloaded gradient**."

$$\boldsymbol{v}_{i,t} \leftarrow \text{mask}(\boldsymbol{u}_{\mathcal{N},t}, q_{i,t}) \qquad q_{i,t} \leftarrow \left\lfloor \frac{D \boxed{\tanh(\beta r_{i,t})}}{\max_{i'} \tanh(\beta r_{i',t})} \right\rfloor \quad \in [0,1]$$

    - *sparsification:* $\text{mask}(\boldsymbol{u}, q)$ retains the largest $\max(0, q)$ components in magnitude of $\boldsymbol{u}$ and zeros out all the rest. Lower sparsification (higher $q_{i,t}$) ⇔ better **downloaded gradient**.
    - $q_{i,t}$ is max-normalized cumulative **SV**: higher **SV** ⇔ higher $r_{i,t}$ ⇔ higher $q_{i,t}$.
    - *altruism degree β* **quantifies** how much an agent with *lower* contributions benefit larger β ⇔ more altruistic/equitable while smaller β ⇔ stricter fairness.
  - Update local model: $\boldsymbol{w}_{i,t} \leftarrow \boldsymbol{w}_{i,t-1} + \boldsymbol{v}_{i,t}$

Putting it all together



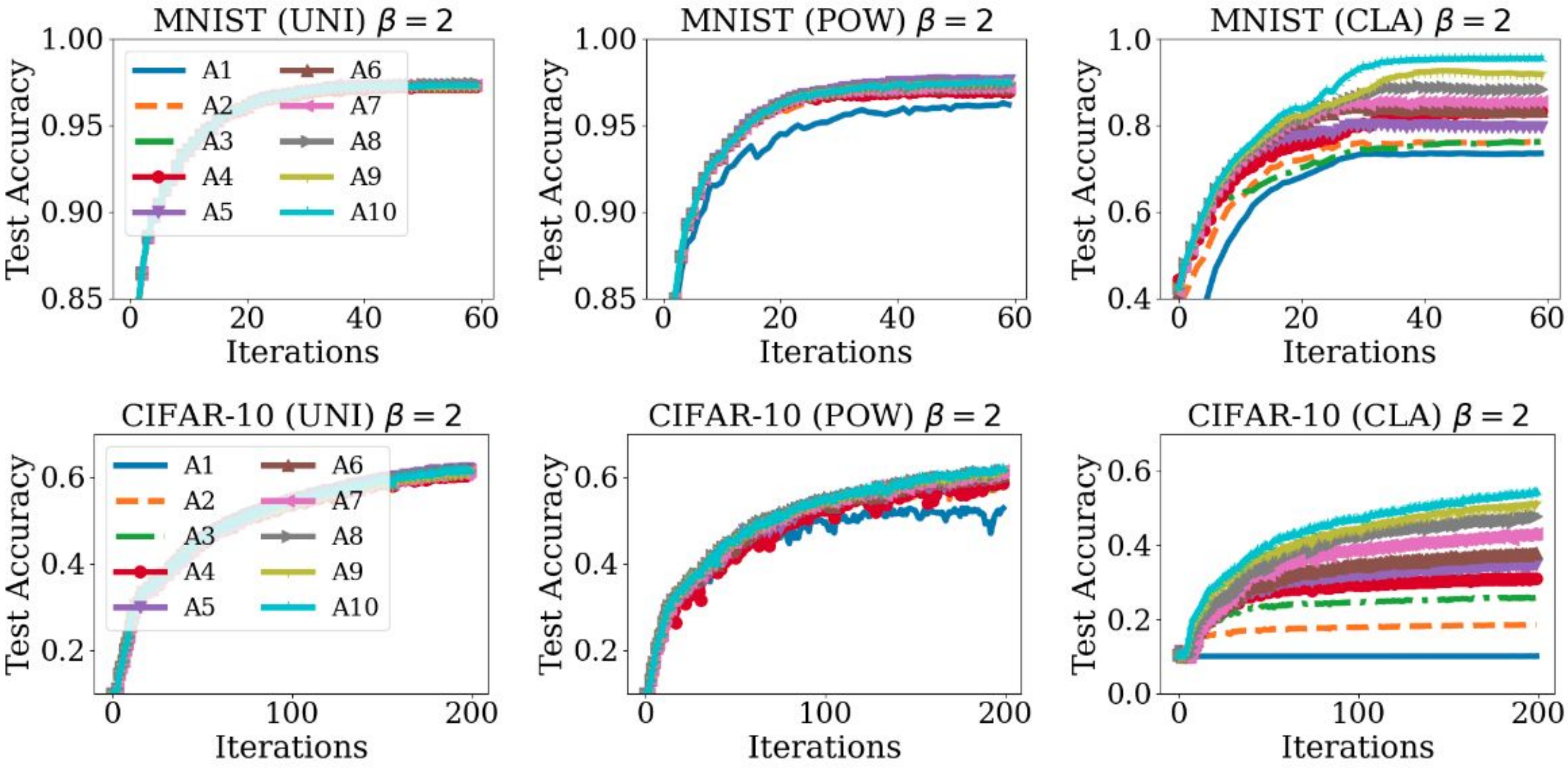
Global Fairness Guarantee

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

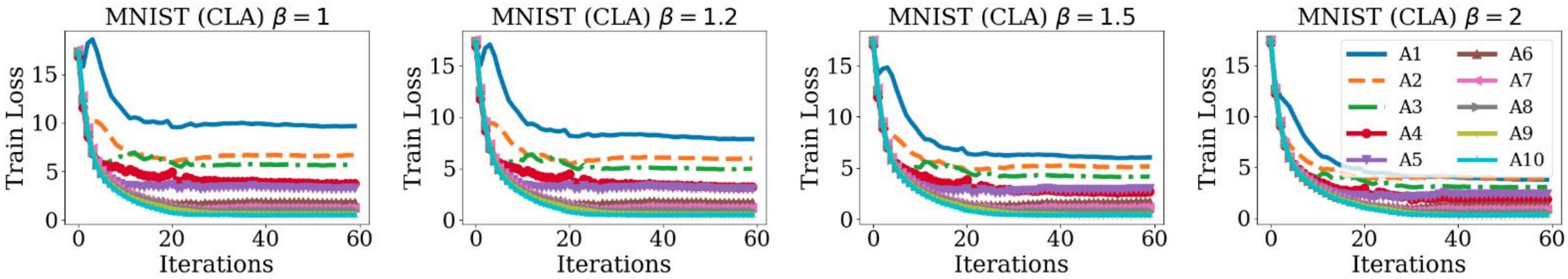
- Local fairness to global fairness:
  - An agent that uploads better gradients can download better gradients (locally fair), and as a result, this agent receives a better-performing model (globally fair).

- Intuition:
  - all agents start with the same model: $\boldsymbol{w}_0$
  - agents with higher $r_{i,t}$ have less deviation from the trajectory: $\{\boldsymbol{w}_0 + \sum_{l=1}^{t} \boldsymbol{u}_{\mathcal{N},l}\}_t$

Fairness results (convergence trajectories)



Fairness results (effect of β)



Increasing *altruism degree* β leads to more equitable performance, and in particular improves the performance of agents with relatively lower contributions.

Offers a fairness vs. equality trade-off.