ICML
International Conference
On Machine Learning

Pursuing Overall Welfare in Federated Learning through Sequential Decision Making

Seok-Ju Hahn, Gi-Soo Kim, Junghye Lee

UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
2 0 0 7

VERI LUX
TAS MEA

Introduction

- **Client-Level Fairness in FL** [1,2]
  - Uniform performance distributions of a global model across participating clients
    - i.e., a global model ($\boldsymbol{\theta}$) can be biased toward different clients.

$$F_1(\boldsymbol{\theta}) = 0.1, F_2(\boldsymbol{\theta}) = 2.3, F_3(\boldsymbol{\theta}) = 9.5, F_4(\boldsymbol{\theta}) = 0.6, F_5(\boldsymbol{\theta}) = 1.1$$

[1] Chen, H., Zhu, T., Zhang, T., Zhou, W., & Yu, P. S. (2023). Privacy and Fairness in Federated Learning: on the Perspective of Trade-off. ACM Computing Surveys.
[2] Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2019, September). Fair Resource Allocation in Federated Learning. In International Conference on Learning Representations.

Introduction

- **Stop Using Static Mixing Coefficient**

$$\min_{\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d} F(\boldsymbol{\theta}) := \sum_{i=1}^{K} w_i F_i(\boldsymbol{\theta}), \qquad w_i = \frac{n_i}{\sum_{j=1}^{K} n_j}$$

  - Simple solution: imposing larger coefficients to the clients with larger losses
  - Use adaptive mixing coefficient $\boldsymbol{p} = [p_1, \dots, p_K]^\top$ instead!
    - This adaptive decision is sequentially made by the server.

Previous Works

- **Research Gap: Truly Adaptive?**
  - Server only receives a single response vector (e.g., local losses $[F_1(\boldsymbol{\theta}), \dots, F_K(\boldsymbol{\theta})]^\top$)

    … for deciding another single mixing coefficient, $\boldsymbol{p} = [p_1, \dots, p_K]^\top$.
    - i.e., a sample-deficient situation!

Method	**Adaptive Mixing Coefficients**
`FedAvg[3]`	$p_i \propto w_i$
`q-FedAvg[2]`	$p_i \propto w_i F_i^q(\boldsymbol{\theta}), q \in \mathbb{R}_{>0}$
`AFL[4] & FedMGDA[5]`	(a special case of `q-FedAvg` when $q \to \infty$)
`TERM[6]`	$p_i \propto w_i \exp\big(\lambda F_i(\boldsymbol{\theta})\big), \lambda \in \mathbb{R}$
`PropFair[7]`	$p_i \propto \frac{w_i}{M - F_i(\boldsymbol{\theta})}, M \in \mathbb{R}_{>0}$

[2] Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2019, September). Fair Resource Allocation in Federated Learning. In International Conference on Learning Representations.
[3] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
[4] Mohri, M., Sivek, G., & Suresh, A. T. (2019, May). Agnostic federated learning. In International Conference on Machine Learning (pp. 4615-4625). PMLR.
[5] Hu, Z., Shaloudegi, K., Zhang, G., & Yu, Y. (2022). Federated learning meets multi-objective optimization. IEEE Transactions on Network Science and Engineering, 9(4), 2039-2051.
[6] Li, T., Beirami, A., Sanjabi, M., & Smith, V. (2020). Tilted empirical risk minimization. arXiv preprint arXiv:2007.01162.
[7] Zhang, G., Malekmohammadi, S., Chen, X., & Yu, Y. (2022). Proportional Fairness in Federated Learning. arXiv preprint arXiv:2202.01666.

Research Question

How can we improve the scheme of deciding **mixing coefficients** so that it is *truly adaptive* even under the sample-deficient condition?

Discovery

- **Online Convex Optimization (OCO) as a Unified Language**
  - Exponentiated Gradient (EG [8])
    - For all $t = 1, \ldots, T$, suppose we want to minimize a *decision loss* $\ell^{(t)}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \boldsymbol{r}^{(t)} \rangle$ sequentially, which is defined by a *response vector* $\boldsymbol{r}^{(t)} \in \mathbb{R}^K$ and a *decision variable* $\boldsymbol{p} \in \Delta_{K-1}$. $R(\boldsymbol{p})$ is a *regularizer* multiplied by a constant *step size* $\eta \in \mathbb{R}_{\geq 0}$.

$$\boldsymbol{p}^{(t+1)} = \underset{\boldsymbol{p} \in \Delta_{K-1}}{\operatorname{argmin}} \, \ell^{(t)}(\boldsymbol{p}) + \eta R(\boldsymbol{p})$$

    - As long as the regularizer $R(\boldsymbol{p})$ is fixed as the negative entropy, i.e., $R(\boldsymbol{p}) = \sum_{i=1}^{K} p_i \log p_i$, it has a closed-form update:

$$p_i^{(t+1)} = \frac{p_i^{(t)} \exp\left(r_i^{(t)}/\eta\right)}{\sum_{j=1}^{K} p_j^{(t)} \exp\left(r_j^{(t)}/\eta\right)}.$$

[8] Helmbold, D. P., Schapire, R. E., Singer, Y., & Warmuth, M. K. (1998). On-line portfolio selection using multiplicative updates. Mathematical Finance, 8(4), 325-347.

Discovery

- **EG Subsumes Existing Fair FL Algorithms**

$$p_i^{(t+1)} = \frac{p_i^{(t)} \exp\left(r_i^{(t)}/\eta\right)}{\sum_{j=1}^{K} p_j^{(t)} \exp\left(r_j^{(t)}/\eta\right)}.$$

Method	Response, $r_i^{(t)}$	Last Decision, $p_i^{(t)}$	Step Size, η	New Decision, $p_i^{(t+1)}$
`FedAvg[3]`	0	w_i	1	$\propto w_i$
`q-FedAvg[2]` (`AFL[4]` if $q \to \infty$)	$q \log F_i\left(\boldsymbol{\theta}^{(t)}\right)$	w_i	1	$\propto w_i F_i^q\left(\boldsymbol{\theta}^{(t)}\right)$
`TERM[6]`	$F_i\left(\boldsymbol{\theta}^{(t)}\right)$	w_i	$1/\lambda$	$\propto w_i \exp\left(\lambda F_i\left(\boldsymbol{\theta}^{(t)}\right)\right)$
`PropFair[7]`	$-\log\left(M - F_i\left(\boldsymbol{\theta}^{(t)}\right)\right)$	w_i	1	$\propto \frac{w_i}{M - F_i(\boldsymbol{\theta}^{(t)})}$

[2] Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2019, September). Fair Resource Allocation in Federated Learning. In International Conference on Learning Representations.
[3] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
[4] Mohri, M., Sivek, G., & Suresh, A. T. (2019, May). Agnostic federated learning. In International Conference on Machine Learning (pp. 4615-4625). PMLR.
[6] Li, T., Beirami, A., Sanjabi, M., & Smith, V. (2020). Tilted empirical risk minimization. arXiv preprint arXiv:2007.01162.
[7] Zhang, G., Malekmohammadi, S., Chen, X., & Yu, Y. (2022). Proportional Fairness in Federated Learning. arXiv preprint arXiv:2202.01666.

Proposed Methods

- **Fixing Suboptimal Designs in Existing Methods as EG**
  - i) Stateless decision making: $p_i^{(t)} = w_i$
  - ii) Fixed step size: η
  - iii) Decision loss without Lipchitz continuity and strict convexity guarantee: $\ell^{(t)}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \boldsymbol{r}^{(t)} \rangle$
    - The local loss $F_i(\boldsymbol{\theta}^{(t)})$ corresponded to $r_i^{(t)}$ is usually unbounded above, e.g., cross-entropy loss.

Proposed Methods

- **Follow-The-Regularized-Leader (FTRL [9-12])**

$$\boldsymbol{p}^{(t+1)} = \underset{\boldsymbol{p}\in \Delta_{K-1}}{\operatorname{argmin}} \sum_{\tau=1}^{t} \ell^{(\tau)}(\boldsymbol{p}) + \eta^{(t+1)} R(\boldsymbol{p})$$

  - i) Stateful as mirroring all previous decision losses: $\sum_{\tau=1}^{t} \ell^{(\tau)}(\boldsymbol{p})$
  - ii) Time-varying step size: $\eta^{(t+1)}$
    - …or time-varying regularizer: $R^{(t+1)}(\boldsymbol{p})$

[9] Abernethy, J. D., Hazan, E., & Rakhlin, A. (2009). Competing in the dark: An efficient algorithm for bandit linear optimization.
[10] Hazan, E., & Kale, S. (2010). Extracting certainty from uncertainty: Regret bounded by variation in costs. Machine learning, 80, 165-188.
[11] Agarwal, A., & Hazan, E. (2005). New algorithms for repeated play and universal portfolio management. Princeton University Technical Report TR-740-05.
[12] Shalev-Shwartz, S., & Singer, Y. (2006, June). Online learning meets optimization in the dual. In International Conference on Computational Learning Theory (pp. 423-437). Berlin, Heidelberg: Springer Berlin Heidelberg.

Proposed Methods

- **Logarithmic Growth from Online Portfolio Selection [13]**
  - Metaphor: OPS sequentially assigns higher portfolio weights to bullish assets, to maximize:
    - Logarithmic growth: $\sum_{t=1}^{T} \log\left(1 + \langle \boldsymbol{p}^{(t)}, \boldsymbol{r}^{(t)} \rangle\right)$
  - The negative logarithmic growth as our decision loss to minimize:

$$\ell^{(t)}(\boldsymbol{p}) = -\log\left(1 + \langle \boldsymbol{p}, \boldsymbol{r}^{(t)} \rangle\right)$$

    - Lipschitz continuous and strictly convex *(please see Lemma 4.1 and Lemma A.1)*
    - Loosely related to (rectified) min-max fairness notion

[13] Cover, T. M. (1991). Universal portfolios. Mathematical finance, 1(1), 1-29.

Proposed Methods

- **Doubly Robust Estimator for Partially Observed Responses**
  - Client sampling (especially in the cross-device FL setting)
    - The server can only observe partial entries of a response $\boldsymbol{r}^{(t)}$…
  - Doubly Robust Estimator [14-16]
    - Denote $C = P(i \in S^{(t)})$ as a client sampling probability, $S^{(t)}$ is an index set of selected clients:

$$\check{r}_i^{(t)} = \left(1 - \frac{\mathbb{I}\left(i \in S^{(t)}\right)}{C}\right)\bar{\mathrm{r}}^{(t)} + \frac{\mathbb{I}\left(i \in S^{(t)}\right)}{C} r_i^{(t)},$$

where $\bar{\mathrm{r}}^{(t)} = \frac{1}{|S^{(t)}|}\sum_{i \in S^{(t)}} r_i^{(t)}$. *(Please see Lemma 4.3)*

[14] Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427), 846-866.
[15] Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4), 962-973.
[16] Dimakopoulou, M., Zhou, Z., Athey, S., & Imbens, G. (2019, July). Balanced linear contextual bandits. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 3445-3453).

Proposed Methods

- **Practical FL Settings Require Different Conditions**
  - Cross-silo FL (number of clients < number of rounds, i.e., $K \ll T$)
    - e.g., $K = 20$ hospitals with $T = 200$ rounds [17]
    - All clients can usually be participated in each round.
  - Cross-device FL (number of clients > number of rounds, i.e., $K \gg T$)
    - e.g., $K = 1.5 \times 10^6$ users with $T = 3{,}000$ rounds [18]
    - Client sampling is inevitably required.

[17] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., ... & Ramage, D. (2018). Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.
[18] Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., ... & Li, Q. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature medicine, 27(10), 1735-1743.

Proposed Methods

- AAggFF: **<u>A</u>daptive <u>Agg</u>regation for <u>F</u>air <u>F</u>ederated Learning**
  - AAggFF-S: for cross-silo FL setting – Online Newton Step [19,20]

$$\boldsymbol{p}^{(t+1)} = \underset{\boldsymbol{p} \in \Delta_{K-1}}{\operatorname{argmin}} \sum_{\tau=1}^{t} \tilde{\ell}^{(\tau)}(\boldsymbol{p}) + \frac{\alpha}{2} \|\boldsymbol{p}\|_2^2 + \frac{\beta}{2} \sum_{\tau=1}^{t} \left(\langle \boldsymbol{g}^{(\tau)}, \boldsymbol{p} - \boldsymbol{p}^{(\tau)} \rangle \right)^2,$$

where $\tilde{\ell}^{(t)}(\boldsymbol{p})$ is a linearized loss defined as $\tilde{\ell}^{(t)}(\boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{g}^{(t)} \rangle$ and $\boldsymbol{g}^{(t)} = \nabla \ell^{(t)}(\boldsymbol{p}^{(t)})$.

(Please see pseudocodes in Appendix D)

  - Runtime: $\mathcal{O}(K^2 + K^3)$
    - $\mathcal{O}(K^3)$ for weighted projection to a simplex [21]
    - Empirically moderate for the cross-silo setting

[19] Agarwal, A., Hazan, E., Kale, S., & Schapire, R. E. (2006, June). Algorithms for portfolio management based on the newton method. In Proceedings of the 23rd international conference on Machine learning (pp. 9-16).
[20] Hazan, E., Agarwal, A., & Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. Machine Learning, 69, 169-192.
[21] Yurii Nesterov. Introductory lectures on convex optimization: A basic course, volume 87. Springer Science & Business Media, 2003.

Proposed Methods

- AAggFF: **Adaptive Aggregation for Fair Federated Learning**
  - AAggFF-D: for cross-device FL setting – FTRL [9-12]

$$\boldsymbol{p}^{(t+1)} = \underset{\boldsymbol{p}\in \Delta_{K-1}}{\operatorname{argmin}} \sum_{\tau=1}^{t} \tilde{\ell}^{(\tau)}(\boldsymbol{p}) + \frac{L_\infty \sqrt{t+1}}{\sqrt{\log K}} \sum_{i=1}^{K} p_i \log p_i \,,$$

where $\tilde{\ell}^{(t)}(\boldsymbol{p})$ is a linearized loss defined as $\tilde{\ell}^{(t)}(\boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{g}^{(t)} \rangle$ and $\boldsymbol{g}^{(t)} = \nabla \ell^{(t)}(\boldsymbol{p}^{(t)})$.

(Please see closed-form update in Remark 4.5 and pseudocodes in Appendix D)

  - Runtime: $\mathcal{O}(K)$
    - Linear; favorable to the dross-device setting

[9] Abernethy, J. D., Hazan, E., & Rakhlin, A. (2009). Competing in the dark: An efficient algorithm for bandit linear optimization.
[10] Hazan, E., & Kale, S. (2010). Extracting certainty from uncertainty: Regret bounded by variation in costs. Machine learning, 80, 165-188.
[11] Agarwal, A., & Hazan, E. (2005). New algorithms for repeated play and universal portfolio management. Princeton University Technical Report TR-740-05.
[12] Shalev-Shwartz, S., & Singer, Y. (2006, June). Online learning meets optimization in the dual. In International Conference on Computational Learning Theory (pp. 423-437). Berlin, Heidelberg: Springer Berlin Heidelberg.

Theoretical Guarantee

- **Regret Upper Bound for** AAggFF-S
  - **Theorem** (Regret Upper Bound for AAggFF-S)

Suppose $\forall \boldsymbol{p} \in \Delta_{K-1}$, let the decision $\{\boldsymbol{p}^{(t)}: t \in [T]\}$ be derived by AAggFF-S for K clients during T rounds.

Then, the regret can be bounded above as:

$$\text{Regret}^{(T)}(\boldsymbol{p}^{\star}) \leq 2L_{\infty}K\left(1 + \log\left(1 + \frac{T}{16K}\right)\right),$$

where $\alpha = 4KL_{\infty}$ and $\beta = \frac{1}{4L_{\infty}}$ in the objective, and L_{∞} can be adjusted by the range of a response.

Theoretical Guarantee

- **Regret Upper Bound for** AAggFF-D
  - **Theorem** (Regret Upper Bound for AAggFF-D with Full Client Participation)

Suppose $\forall \boldsymbol{p} \in \Delta_{K-1}$, let the decision $\{\boldsymbol{p}^{(t)}: t \in [T]\}$ be derived by AAggFF-D for K clients during T rounds with client sampling probability $C = 1$.

Then, the regret can be bounded above as:

$$\mathrm{Regret}^{(T)}(\boldsymbol{p}^{\star}) \leq 2L_{\infty}\sqrt{T \log K},$$

where L_{∞} can be adjusted by the range of a response.

Theoretical Guarantee

- **Regret Upper Bound for** AAggFF-D
  - **Corollary (**Regret Upper Bound for AAggFF-D with Partial Client Participation)

Suppose $\forall \boldsymbol{p} \in \Delta_{K-1}$, let the decision $\{\boldsymbol{p}^{(t)}: t \in [T]\}$ be derived by AAggFF-D for K clients during T rounds with client sampling probability $C \in (0,1)$.

Being equipped with the doubly robust estimator $\breve{\boldsymbol{r}}^{(t)}$, the regret can be bounded above in expectation as:

$$\mathbb{E}\left[\text{Regret}^{(T)}(\boldsymbol{p}^{\star})\right] \leq \mathcal{O}\left(L_{\infty}\sqrt{T \log K}\right),$$

where L_{∞} can be adjusted by the range of a response.

Experimental Results

- **Setup**
  - Cross-silo (number of clients (K) < number of rounds (T))
    - Berka (tabular): loan default prediction (2 classes)
    - MQP (text): medical sentence similarity classification (2 classes)
    - ISIC (image): skin cancer classification (8 classes)
  - Cross-device (number of clients (K) > number of rounds (T))
    - CelebA (image): smiling face recognition (2 classes)
    - Reddit (text): language modeling (10,000 sentence tokens)
    - SpeechCommands (audio): speech recognition (35 classes)

Cross-silo		
Dataset	K	T
Berka	7	100
MQP	11	100
ISIC	1	50

Cross-device		
Dataset	K	T
CelebA	9,343	3,000
Reddit	817	300
Speech Commands	2,005	500

Experimental Results

- **Boosted Performance in Both Cross-Silo and Cross-Device Settings**
  - Improved worst-case performance as well as little compromise on the average performance
    - Low Gini coefficient: uniform performance distribution

Dataset	Berka (AUROC)				MQP (AUROC)				ISIC (Acc. 5)			
Method	Avg. (↑)	Worst (↑)	Best (↑)	Gini (↓)	Avg. (↑)	Worst (↑)	Best (↑)	Gini (↓)	Avg. (↑)	Worst (↑)	Best (↑)	Gini (↓)
FedAvg [3]	80.09	48.06	**99.03**	10.87	56.06	41.03	76.31	8.63	87.42	69.92	92.57	4.84
	(2.45)	(25.15)	(1.37)	(4.11)	(0.06)	(4.33)	(8.42)	(0.91)	(2.11)	(6.78)	(2.56)	(1.17)
AFL [4]	79.70	49.02	98.55	10.58	56.01	41.28	75.54	8.56	87.39	68.17	93.33	4.80
	(4.14)	(25.89)	(2.05)	(5.03)	(0.30)	(3.92)	(6.77)	(1.24)	(2.31)	(10.09)	(2.18)	(1.74)
q-FedAvg [2]	79.98	49.44	98.07	10.62	**56.89**	40.22	**79.38**	8.68	41.59	20.38	58.08	22.25
	(3.89)	(26.15)	(2.73)	(5.22)	(0.42)	(3.06)	(9.09)	(0.57)	(16.22)	(23.24)	(28.52)	(10.02)
TERM [6]	80.11	48.96	**99.03**	10.86	56.47	40.73	76.80	8.67	87.89	77.32	96.00	3.77
	(3.08)	(25.79)	(1.37)	(4.73)	(0.19)	(4.36)	(8.30)	(1.43)	(1.69)	(5.84)	(3.27)	(0.94)
FedMGDA [5]	79.24	46.38	**99.03**	11.64	53.02	34.91	69.65	10.33	42.36	21.44	59.21	22.25
	(2.96)	(24.11)	(1.37)	(4.84)	(1.67)	(2.22)	(3.89)	(0.44)	(14.94)	(21.30)	(28.52)	(10.02)
PropFair [7]	79.61	49.44	98.07	10.47	56.60	41.71	79.09	8.74	83.88	58.36	91.35	7.91
	(4.49)	(26.15)	(2.73)	(5.04)	(0.39)	(3.80)	(7.40)	(0.87)	(2.50)	(11.63)	(2.48)	(2.10)
AAggFF-S	**80.93**	**52.08**	**99.03**	**10.16**	56.63	**41.79**	75.56	**8.38**	**89.76**	**85.17**	**98.22**	**2.52**
(Proposed)	(2.96)	(23.59)	(1.37)	(3.80)	(0.54)	(4.43)	(6.53)	(0.77)	(1.03)	(3.87)	(1.66)	(0.38)

Dataset	CelebA (Acc. 1)				Reddit (Acc. 1)				SpeechCommands (Acc. 5)			
Method	Avg. (↑)	Worst 10% (↑)	Best 10%(↑)	Gini (↓)	Avg. (↑)	Worst 10%(↑)	Best 10%(↑)	Gini (↓)	Avg. (↑)	Worst 10%(↑)	Best 10%(↑)	Gini (↓)
FedAvg [3]	90.79	55.76	100.00	7.86	10.76	2.50	20.86	25.66	75.51	7.93	100.00	24.58
	(0.53)	(0.84)	(0.00)	(0.30)	(1.45)	(0.21)	(3.64)	(0.49)	(1.08)	(2.87)	(0.00)	(1.34)
q-FedAvg [2]	90.88	55.73	100.00	7.82	12.76	3.38	21.81	23.34	73.34	11.19	100.00	23.16
	(0.19)	(0.85)	(0.00)	(0.21)	(0.32)	(0.20)	(0.19)	(0.34)	(0.47)	(0.47)	(0.00)	(0.13)
TERM [6]	90.71	55.66	100.00	7.90	12.02	2.85	20.74	24.15	70.90	5.98	100.00	26.37
	(0.65)	(0.93)	(0.00)	(0.38)	(0.16)	(0.41)	(0.65)	(1.05)	(2.96)	(1.10)	(0.00)	(1.32)
FedMGDA [5]	88.33	48.60	100.00	9.75	10.58	2.35	19.09	25.20	72.45	9.65	100.00	23.68
	(0.63)	(25.85)	(0.00)	(0.59)	(0.18)	(0.20)	(0.62)	(0.22)	(1.88)	(2.90)	(0.00)	(1.27)
PropFair [7]	87.25	48.11	100.00	10.39	11.26	1.95	21.33	25.97	73.64	7.30	100.00	24.97
	(5.01)	(10.03)	(0.00)	(3.43)	(0.71)	(0.32)	(0.92)	(1.02)	(3.31)	(1.02)	(0.00)	(1.09)
AAggFF-D	**91.27**	**56.71**	100.00	**7.54**	**12.95**	**4.75**	**22.81**	**22.59**	**76.68**	**14.54**	100.00	**21.42**
(Proposed)	(0.07)	(0.08)	(0.00)	(0.04)	(0.39)	(0.76)	(1.36)	(0.28)	(0.80)	(2.58)	(0.00)	(0.81)

[2] Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2019, September). Fair Resource Allocation in Federated Learning. In International Conference on Learning Representations.
[3] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
[4] Mohri, M., Sivek, G., & Suresh, A. T. (2019, May). Agnostic federated learning. In International Conference on Machine Learning (pp. 4615-4625). PMLR.
[5] Hu, Z., Shaloudegi, K., Zhang, G., & Yu, Y. (2022). Federated learning meets multi-objective optimization. IEEE Transactions on Network Science and Engineering, 9(4), 2039-2051.
[6] Li, T., Beirami, A., Sanjabi, M., & Smith, V. (2020). Tilted empirical risk minimization. arXiv preprint arXiv:2007.01162.
[7] Zhang, G., Malekmohammadi, S., Chen, X., & Yu, Y. (2022). Proportional Fairness in Federated Learning. arXiv preprint arXiv:2202.01666.

Conclusion

- AAggFF finds better **mixing coefficients**
  - through improved online convex optimization objectives.
- AAggFF is specialized into practical FL settings
  - `AAggFF-S` for the cross-silo setting, and `AAggFF-D` for the cross-device setting; both guarantee vanishing regrets.
- AAggFF pursues overall welfare in the federated system
  - not only inducing uniform performances, but also maintaining decent average performances.

The End.