Federated Fairness without Access to Demographics

Afroditi Papadaki
University College London
a.papadaki.17@ucl.ac.uk

Natalia Martinez
IBM Research
natalia.martinez.gil@ibm.com

Martin Bertran
Amazon Web Services
maberlop@amazon.com

Guillermo Sapiro
Duke University; Apple
guillermo.sapiro@duke.edu

Miguel Rodrigues
University College London
m.rodrigues@ucl.ac.uk

Abstract

Existing federated learning approaches address demographic group fairness assuming that clients are aware of the sensitive groups. Such approaches are not applicable in settings where sensitive groups are unidentified or unavailable. In this paper, we address this limitation by focusing on federated learning settings of fairness without demographics. We present a novel objective that allows trade-offs between (worst-case) group fairness and average utility performance through a hyper-parameter and a group size constraint. We show that the proposed objective recovers existing approaches as special cases and then provide an algorithm to efficiently solve the proposed optimization problem. We experimentally showcase the different solutions that can be achieved by our proposed approach and compare it against baselines on various standard datasets.

1 Introduction

Federated learning (FL) enables different entities to collaboratively learn a statistical model through an iterative procedure that is coordinated by a central server in a decentralized manner. The clients keep their raw data locally to preserve privacy, and share only focused updates with the server, which limits the information shared with the server while still achieving some global learning objective [11, 12]. These learning setups are characterized by data heterogeneity and unbalancedness across participants [10, 25].

A key challenge in federated learning is fairness. Existing literature is predominately focused on group fairness scenarios, where the goal is to learn a model that performs well across pre-defined sensitive groups.[1] The definition of a group varies across different federated learning approaches. The majority of these works [3, 9, 14, 18, 21, 27] characterize the clients as the collection of groups in the federation and propose methods to control performance disparities across them. Others [2, 28] suggest that groups are associated to the available within-client populations and provide methods to tackle within-client group disparities. More recently, the use of global demographics was proposed in [8, 19] for guaranteeing fairness across any sensitive groups included in the federation, thereby enabling clients that have access only to a subset of the required groups to benefit from the training procedure.

[1] We use the terms (protected/sensitive/demographic) group/population, interchangeably.

Workshop on Federated Learning: Recent Advances and New Challenges, in Conjunction with NeurIPS 2022 (FL-NeurIPS'22). This workshop does not have official proceedings and this paper is non-archival.

The aforementioned approaches assume that the participants have knowledge about the sensitive demographic groups during training time and that the group memberships have been correctly assigned to each data point. Unfortunately, the assumption that every client has access to predefined and accurate sensitive groups might be unrealistic for various federated learning scenarios. For example, consider a scenario where multiple medical institutions collaborate to learn a model to accurately perform a particular medical diagnosis that is fair across individuals of different demographic (sub)groups. Each institution owns a database consisting of patient data that cannot directly share with other participants due to privacy regulations (e.g. GDPR [1]). Furthermore, the documents facilitated at each local database might lack particular information (e.g. race, religion, sexual orientation, etc.) that the patients did not wish to disclose. The absence of such information during the training process makes existing FL fairness methods, such as [2, 8, 19], inappropriate for learning a model that is fair to patients of different demographics. We illustrate this example in Figure 1.

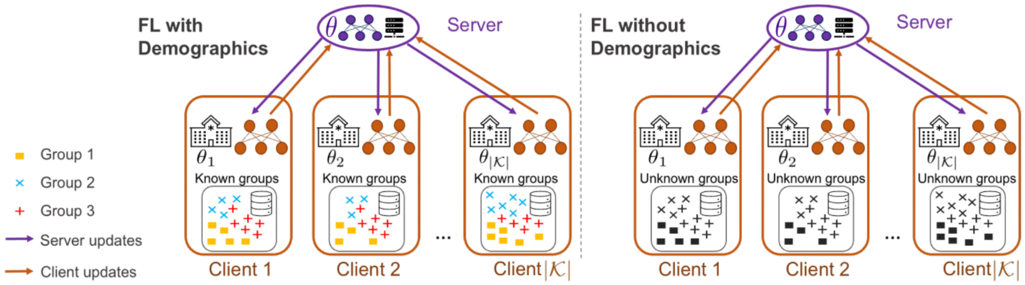

Figure 1: (*Left*) Each hospital $k \in \mathcal{K}$ has prior knowledge about the sensitive populations in its dataset and the global model h, parametrized by a vector $\boldsymbol{\theta}$, is trained by incorporating the information about demographics in the training procedure using existing FL fairness methods [2, 8, 19]. (*Right*) Clients are unaware of the local demographics and existing approaches cannot be deployed.

In this paper, we address the problem of group fairness in federated learning scenarios where the clients are unaware about the demographic group composition in their data. We propose a flexible optimization objective that admits a family of solutions that depend on a single hyperparameter ϵ. For $\epsilon = 1$ our objective reduces to the standard ERM problem which does not cater for group fairness. For $\epsilon \approx 0$ it recovers an optimization problem that seeks to minimize the worst-tail risk, subject to a probability-level constraint which is predefined based on some common set of policies and/or preferences across clients, that guarantees subgroup robustness and good performance on unidentified groups.

Our approach helps the participants (a) identify their global vulnerable (testing) population, even if it does not exist on its local distribution during training time and (b) learn a global hypothesis that allows a trade-off between average performance and subgroup robustness. We also present links to popular robustness formulations in centralized machine learning and provide an algorithm that solves our proposed objective.We empirically study the different trade-offs that can be achieved by the proposed approach and experimentally demonstrate the efficiency of the proposed method against other baselines in centralized and federated learning settings.

2 Related Work

Fairness without Demographics in ML. There is limited research in centralized machine learning trying to address group fairness without explicit demographics. One way to deal with data when group labels are not known is through proxy fairness [6]. Such approaches [6, 24, 29] develop a new proxy variable to substitute the true sensitive group variable so that conventional group fairness methods can be deployed. However, these approaches require the true group variable to be known while the samples group label are considered unavailable, which is not feasible for many applications.

An alternative line of work that addresses group fairness when demographics are completely unknown is (sub)group robustness [7, 13, 15]. These works aim to improve the performance on any possible sensitive population that can be generated that exceeds a predefined size. For example, the authors in [7] aim to minimize the worst-case risk that exceeds a high-risk threshold. Similarly, blind Pareto fairness (BPF) [15] aims to minimize the worst case risk that induced by the worst possible group distribution, subject to a sufficient group size, while ensuring that the produced solution is also Pareto optimal [17]. Our work departs from these methods since they require the data to be aggregated in a single entity that can be accessed anytime, while we focus on distributed settings where data is hosted in different entities and cannot be shared or accessed directly. Building upon these prior works, we propose a relaxed superquantile formulation that is flexible, in the sense that, it allows achieving different levels of (minimax) group fairness through a hyperparameter ϵ.

Fair Federated Learning. There are various fairness definitions in the federated learning literature. A class of popular approaches [3, 9, 18, 27] focuses on ensuring client fairness – or equivalently client-robustness – by learning a model that optimizes for the worst performing client (or cluster of clients). Nevertheless, as formally shown in [19] fairness across clients does not necessarily guarantee fairness across different demographic groups except if the client owns data from a single demographic population.

Several recent works [2, 28] consider group fairness to be a unique objective for each client that is associated only to the groups available in each client, while others [8, 19, 23] propose frameworks for learning group fair models independently of the demographics available in each client. Similar to [2, 8, 19, 23, 28], our goal is to learn a model that ensures (demographic) group fairness across any groups that exist in the clients data. The key difference though is that we explicitly focus on scenarios where the local sensitive populations are completely unknown to the clients during the training.

3 Problem Formulation

3.1 Preliminaries

Let the pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ represent the input features, and targets, generated from a distribution $p(X, Y)$. Let also a convex loss function $\ell : \Delta^{|\mathcal{Y}|-1} \times \Delta^{|\mathcal{Y}|-1} \to \mathbb{R}_+$ and a convex hypothesis h drawn from a hypothesis class $\mathcal{H} = \{h : \mathcal{X} \to \Delta^{|\mathcal{Y}|-1}\}$, where $\Delta^{|\mathcal{Y}|-1}$ is the probability simplex of dimension $|\mathcal{Y}| - 1$. In this work we consider scenarios where there is not any prior knowledge about the underlying group associated with a particular pair of features-target at training time. We can address this challenge by optimizing the expected risk of the input features-target pairs that exceed a certain threshold, akin to [7, 15]. In particular, let $L_{X,Y} := \ell(h(X), Y)$ denote a random variable representing the loss associated with a hypothesis $h \in \mathcal{H}$. For some predefined probability $\rho \in (0, 1)$, the $(1 - \rho)$-quantile function is defined as

$$q_{L_{X,Y}}(1-\rho) := \inf \left\{ \beta \in \mathbb{R} : p(L_{X,Y} \leq \beta) \geq 1 - \rho \right\}, \tag{1}$$

and the $(1 - \rho)$-superquantile[2] function at confidence level $(1 - \rho)$ is defined as

$$CVaR_{(1-\rho)}(L_{X,Y}) \quad = \underset{(X,Y)\sim p(X,Y)}{\mathbb{E}} [L_{X,Y} | L_{X,Y} \geq q_{L_{X,Y}}(1-\rho)]. \tag{2}$$

We note that Eq. 2 is a measure of the upper tail behavior of the distribution $p(L_{X,Y})$ and, as shown in [22], it can be expressed as a (variational) optimization problem

$$CVaR_{(1-\rho)}(L_{X,Y}) = \min_{c\in\mathbb{R}} \left\{ c + \tfrac{1}{\rho} \underset{(X,Y)\sim p(X,Y)}{\mathbb{E}} \left[(L_{X,Y} - c)_+ \right] \right\}, \tag{3}$$

where $(\cdot)_+ := \max\{0, \cdot\}$ and the second term in the objective represents the regret of any realizations of $L_{X,Y}$ that are positive. Note that the argument that minimizes the objective in Eq. 3 corresponds to the quantile $q_{L_{X,Y}}(1 - \rho)$. We also note that if the selected loss function is bounded, i.e. $0 \leq \ell(h(x), y) \leq B, \forall (x, y) \in \mathcal{X} \times \mathcal{Y}$, we can equivalently optimize over $c \in [0, B]$. If we consider that the unknown demographic groups can have a minimum size ρ, following [15, 26] we can formulate the problem of learning a minimax group fair hypothesis h^*, as

$$\min_{h\in\mathcal{H}, c\in\mathbb{R}} \left\{ c + \frac{1}{\rho} \underset{(X,Y)\sim p(X,Y)}{\mathbb{E}} [(L_{X,Y} - c)_+] \right\} \tag{4}$$

The problem in Eq. 4 can be applied in scenarios where the data are gathered in a single entity and can be accessed directly throughout the training process. We next show how this problem can be modified for federated learning scenarios.

[2] We use the terms conditional value at risk (CVaR), expected tail loss, and superquantile, interchangeably.

3.2 Federated Fairness with Unknown Groups Setting

In the context of federated learning, we consider an additional random variable $K \in \{1, \dots, |\mathcal{K}|\}$ which represents the clients participating in the federation. Each client $k \in \mathcal{K}$ holds data modelled by its own local distribution $p(X,Y|K=k) = p(X|K=k)p(Y|X,K=k)$. Therefore, the data of the entire federation can be described with the mixture distribution $p(X,Y) = \sum_{k\in\mathcal{K}} p(K=k)p(X,Y|K=k)$. Let $L_{X,Y|K=k} := \ell(h(X),Y)$, with $(X,Y) \sim p(X,Y|K=k)$, denote a random variable representing the local loss induced by a hypothesis h. Then the problem in Eq. 3 can be equivalently expressed for federated learning settings as

$$\min_{h\in\mathcal{H},c\in\mathbb{R}} CVaR_{(1-\rho)}(L_{X,Y}) = \min_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{K\sim p(K)}\left[\mathop{\mathbb{E}}_{X,Y\sim p(X,Y|K=k)}[(L_{X,Y|K=k}-c)_+]\right] \tag{5}$$

The $CVaR_{(1-\rho)}$ in RHS of Eq. 5 ignores any data that is not considered high-risk which in turn might unnecessarily reduce the utility performance of the obtained solution on it. Especially, if there are regions in the input space where there is no uncertainty about the target class, i.e., the data is perfectly separable. Motivated by this issue, we propose a relaxed version of Eq. 5, namely R-CVaR, where a hyperparameter $\epsilon \in (0,1]$ accommodates a compromise between the average performance and worst-samples performance as follows

$$\begin{gathered}\min_{h\in\mathcal{H}}\left\{(1-\epsilon)CVaR_{(1-\rho)}(L_{X,Y}) + \epsilon \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)}[L_{X,Y}]\right\} = \\ \min_{h\in\mathcal{H}c\in\mathbb{R}} \mathop{\mathbb{E}}_{K\sim p(K)}\left[(1-\epsilon)\mathop{\mathbb{E}}_{X,Y|K=k}[(L_{X,Y|K=k}-c)_+] + \epsilon \mathop{\mathbb{E}}_{X,Y|K=k}[L_{X,Y|K=k}]\right]\end{gathered} \tag{6}$$

The main benefit of objective in Eq. 6 is that it supports a continuum of solutions that depend on the different values of ϵ. Setting $\epsilon \approx 0$ allows to focus on the worst-performing sample at the cost of the low risk samples. For such small ϵ values we assert that the hypothesis solving Eq. 6 is minimax properly Pareto optimal [5, 17]. On the other hand, picking $\epsilon = 1$, the proposed objective generalizes prior work in federated learning settings and becomes the vanilla-ERM problem in FedAvg [16]. For any other intermediate value of $\epsilon \in (0,1)$, we achieve a trade-off between utility[3] and subgroup robustness. To better understand the set of trade-offs achieved by the proposed objective, we offer an illustrative example in Figure 2. We note that the value of ϵ is predefined and fixed, and therefore, we leave it to up to the policy maker(s) to determine it.

Remark 1. *The hypothesis that determines the $(1-\rho)$-quantile c, for which $c \geq L_{X,Y}$, is the uniform classifier $\overline{h} : \overline{h}_y(X) = \frac{1}{|\mathcal{Y}|} \forall y \in \mathcal{Y}$.*

4 Optimization

In real applications, each client holds a finite dataset $D_k = \{(x_i, y_i)\}_{i=1,\dots,n_k}$ sampled from the true distribution $p(X,Y|K=k)$, with $D = \bigcup_{k\in\mathcal{K}} D_k$ being the dataset containing all the data samples available across clients of size $n = \sum_{k\in\mathcal{K}} n_k$. We explicitly express the empirical form of the proposed R-CVaR objective, as follows

$$\min_{\theta\in\Theta,c\in[0,B]} \sum_{k\in\mathcal{K}} \frac{n_k}{n}\left[(1-\epsilon)[c + \frac{1}{n_k\rho}\sum_{i=1}^{n_k}(\ell(x_{i,k},y_{i,k};\theta)-c)_+] + \frac{\epsilon}{n_k}\sum_{i=1}^{n_k}\ell(x_{i,k},y_{i,k};\theta)\right], \tag{7}$$

where $\theta \in \Theta$ is the vector parametrizing the hypothesis $h \in \mathcal{H}$.

We now introduce an algorithm to solve the empirical objective of Eq. 7, in a federated way. Our algorithm consists of two main steps: (a) model parameters update and (b) periodic calculation of the threshold c.

Model update. Learning the model is a simple procedure where the participating clients receive the current round model parameters θ^t from the server, perform an optimization step and return the updated model parameters θ^t_k to the server. Then, the server produces the new model parameters θ^{t+1} by averaging the received client model parameters.

[3] We define as utility the mean model performance.

Threshold calculation. The global model parameters θ and quantile c can only be computed in the server, since it requires a collection of relevant information from the clients. One key challenge imposed by the proposed objective is the efficient calculation of the quantile c in a federated fashion without vastly increasing the communication overhead per client. Here, we propose a practical technique for estimating this parameter.

Let c be the estimated quantile. We denote $\rho(c) = \sum_{k \in \mathcal{K}} p(K = k)\rho_k(c)$ the estimated probability, with $\rho_k(c) = \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y|K=k)} [\mathbb{1}(L_{X,Y|K=k} \geq c)]$. Also, by definition, we have that the objective's $\rho = \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)} [\mathbb{1}(L_{X,Y} \geq q_{L_{X,Y}}(1-\rho))]$ is equal to $\rho(c)$. Thus, we can compute the quantile c that satisfies for a fixed ρ by ensuring that the objective's and estimated probabilities at each communication round are the same, i.e. $\rho = \rho(c)$. This can realized by the optimization procedure $\min_c \big(\rho - \rho(c)\big)^2$. As a result, we can update the estimated threshold c according to

$$c^{t+1} \leftarrow \prod_{c\in[0,B]} \Big(c^t - \eta_c \operatorname{sign}(\rho - \rho(c^t))\Big). \tag{8}$$

Note that η_c captures the product of the learning rate and the absolute value of the derivative $\big(\rho - \rho(c)\big)^2$ w.r.t. c. We summarize the proposed solver in Algorithm 1.

Algorithm 1 FEDRCVAR ALGORITHM

Inputs: $\mathcal{K}$: set of clients, T: communication rounds, $\eta_{\boldsymbol{\theta}}$: model learning rate, η_c: learning rate for quantile c, $\epsilon \in (0,1)$: trade-off parameter, $\rho \in (0,1)$: parameter for probability-level, c^0: initial threshold set to B.

```
Server initializes $\boldsymbol{\theta}^0$ randomly.
for $t = 1$ to $T$ do
    Server broadcasts $\boldsymbol{\theta}^{t-1}$ and $c^t$
    for each client $k \in \mathcal{K}$ in parallel do
        $\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_\theta \Big\{ \frac{1-\epsilon}{n_k} \sum_{i=1}^{n_k} (\ell(h(x_k^i), y_k^i) - c^t)_+ + \frac{\epsilon}{n_k} \sum_{i=1}^{n_k} \ell(h(x_k^i), y_k^i) \Big\}$
        Return local model $\boldsymbol{\theta}_k^t$ and $\rho_k(c^t) = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{1}(\ell(h(x_k^i), y_k^i) \geq c^t)$ to server
    end for
    $\boldsymbol{\theta}^t \leftarrow \sum_{k\in\mathcal{K}} \frac{n_k}{n} \boldsymbol{\theta}_k^t$
    $c^{t+1} \leftarrow \prod_{c\in[0,B]} \Big(c^t - \eta_c \operatorname{sign}(\rho - \sum_{k\in\mathcal{K}} \frac{n_k}{n} \rho_k(c^t))\Big)$
end for
```

Output: $\boldsymbol{\theta}^T$

5 Experimental Results

We show the benefits of the proposed approach in a *synthetic dataset* (see A.1 for dataset generation details) and two real world datasets: (a) *eICU dataset* [20] critical care dataset to predict patient mortality using real patient documents from different medical centers. For the FL approaches, we split the data to 11 clients, each representing one of the hospitals in the dataset. (b) *ACS Employment dataset* [4] for employment classification based on 14 input features. In FL settings, we distribute the data to 3 clients so that each client represents a race from {Black, White, Others}. For the centralized settings, there are no clients involved and we use all the available training data during the training process. We provide more information about experimental details in Appendix A.

First, we empirically explore the various levels of group fairness that can be achieved through ϵ, for a given size ρ, when we optimize the R-CVaR in Eq.6. This is shown in Figure 2. The R-CVaR objective magnifies the impact of worst performing group and decreases the focus on the average performance for $\epsilon \approx 0$. Setting $\epsilon = 1$ recovers the ERM objective. We also observe that the smaller the value of ρ the larger the trade-offs between fairness and utility we can achieve through ϵ. Note that when ϵ is closer to 0, and ρ is sufficiently small the worst-group risk is closer to the uniform classifier risk. This is consistent to the observations made in [15] about the existence of a critical ρ.

Next, we compare the proposed approach against centralized ML approaches and relevant FL approaches in Figure 3. We benchmark against (a) BPF which guarantees subgroup robustness, (b) empirical risk minimization (ERM), (c) AFL which ensures client-fairness (or equivalently client robustness), and (d) FedAvg which optimizes for utility in FL settings. We omit comparisons to approaches that require predefined groups. As expected, FedRCVaR admits the Pareto optimal subgroup robust solution for $\epsilon \approx 0$, also given by BPF, and provide similar solution to the centralized ERM for $\epsilon = 1$. Interestingly, we observe that it also increases the model's fairness and utility performance simultaneously, outperforming the FL baselines in most settings. For particular values of ϵ and ρ it is also able to recover a solution that achieves client robustness, even though we did not design our objective to explicitly provide it.

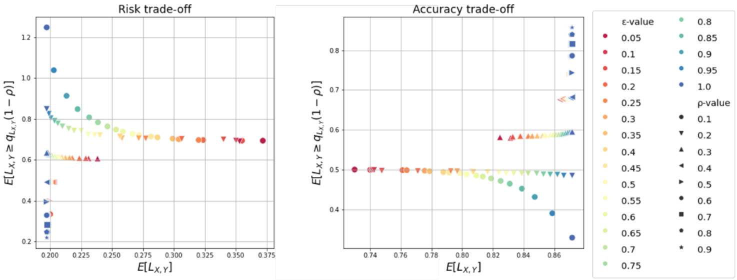

Figure 2: Toy example illustrating the flexibility of R-CVaR objective for hyperparameter $\epsilon \in (0, 1]$ and $\rho \in (0, 1)$ on synthetic data. FedRCVaR is trained for $\rho = \{0.1, \ldots, 0.9\}$ and $\epsilon = \{0.05, 0.1, \ldots, 0.95, 1.0\}$. Different colors describe various ϵ values, while the markers define a particular ρ value. We report the CVaR and average risks and accuracies.

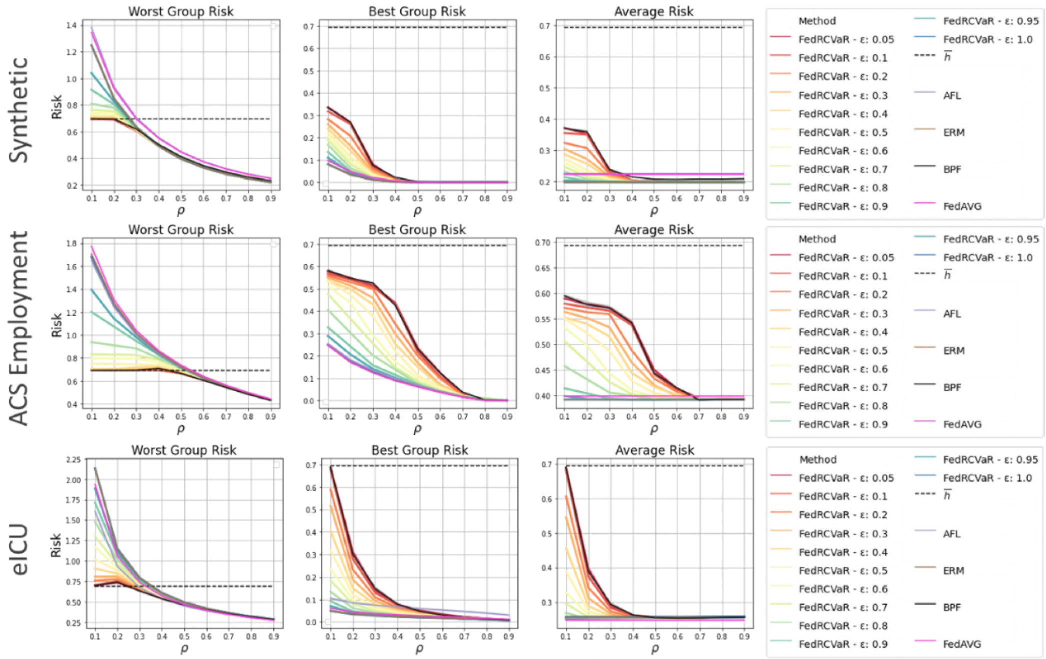

Figure 3: Cross Entropy risks comparison on synthetic, ACS Employment and eICU datasets. FedRCVaR recovers solutions equivalent to centalized machine learning for $\epsilon = \{0.05, 1.0\}$, while improves both utility and accuracy compared to FL baselines in many settings.

6 Conclusion

In this work, we address group fairness without access to demographics in FL settings. We design a optimization objective that admits different trade-offs between (sub)group robustness and average utility. We develop an algorithm that efficiently solves the proposed objective. Finally, we perform experiments to illustrate the solutions that can be achieved through the proposed objective and empirically evaluate the efficiency compared with existing works.

Acknowledgments

UCL authors were supported by Cisco under grant #217462. Duke University authors were partially supported by Cisco, DoD, NSF, and Simons Foundation.

References

[1] European Commission. Reform of eu data protection rules 2018. `https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf`.

[2] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[3] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[5] Arthur M Geoffrion. Proper efficiency and the theory of vector maximization. *Journal of mathematical analysis and applications*, 22(3):618–630, 1968.

[6] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness, 2018.

[7] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018.

[8] Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Fair federated learning via bounded group loss, 2022.

[9] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Fedmgda+: Federated learning meets multi-objective optimization. *CoRR*, abs/2006.11489, 2020.

[10] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019.

[11] Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.

[12] Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NeurIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[13] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc., 2020.

[14] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

[15] Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. Blind pareto fairness and subgroup robustness. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7492–7501. PMLR, 18–24 Jul 2021.

[16] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

[17] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12. Springer Science & Business Media, 2012.

[18] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *36th International Conference on Machine Learning, ICML 2019*, 36th International Conference on Machine Learning, ICML 2019, pages 8114–8124. International Machine Learning Society (IMLS), January 2019. 36th International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019.

[19] Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 142–159, New York, NY, USA, 2022. Association for Computing Machinery.

[20] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5, 09 2018.

[21] Jae Ro, Mingqing Chen, Rajiv Mathews, Mehryar Mohri, and Ananda Theertha Suresh. Communication-Efficient Agnostic Federated Averaging. In *Proc. Interspeech 2021*, pages 871–875, 2021.

[22] R.T. Rockafellar, J.O. Royset, and S.I. Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014.

[23] Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers, 2021.

[24] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19339–19352. Curran Associates, Inc., 2020.

[25] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Agüera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas N. Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konečný, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtárik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake E. Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. A field guide to federated optimization. *CoRR*, abs/2107.06917, 2021.

[26] Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR, 09–15 Jun 2019.

[27] Xubo Yue, Maher Nouiehed, and Raed Al Kontar. GIFAIR-FL: an approach for group and individual fairness in federated learning. *CoRR*, abs/2108.02741, 2021.

[28] Fengda Zhang, Kun Kuang, Yuxuan Liu, Chao Wu, Fei Wu, Jiaxun Lu, Yunfeng Shao, and Jun Xiao. Unified group fairness on federated learning, 2021.

[29] Yan Zhang. Assessing fair lending risks using race/ethnicity proxies. *Manage. Sci.*, 64(1):178–197, jan 2018.

A Experimental Details

A.1 Synthetic Dataset

We develop a dataset to learn a binary classification task, $Y \in \{0,1\}$, in a federation with two clients. Each client owns features sampled from a truncated normal distributions with means $\{\mu_0, \mu_1\} = \{-1, 1\}$, common variance $\sigma^2 = 1$ and that lie in the intervals $\{(1-, 0.5), (-0.5, 1)\}$, respectively. We consider $p(Y|X) = l\mathbb{1}[x \leq -0.5] + u\mathbb{1}[x \geq 0.5] + m\mathbb{1}[-0.5 < x < 0.5]$, with $\{l, m, u\} = \{0, 0.5\sin\frac{\pi}{2}x + 0.5, 1\}$. We assume that the testing distribution is a uniform $p_{test}(X) = \mathcal{U}(-1, 1)$.

A.2 Experimental Settings

We preprocess ACS Employment as described in [4] and eICU akin to [20]. eICU dataset requires credentialed access and the procedure for requesting access is described on the dataset's website https://eicu-crd.mit.edu/gettingstarted/access/. For the synthetic and ACS Employment datasets we use a MLP of an hidden layer with size 512 and for eICU we use logistic regression. We train using Cross Entropy in all cases. FedRCVaR, is trained using batch size equal to the size of client dataset for a single epoch. We use the same options for AFL. FedAvg is trained using batches of sample size 128 and local epochs $E = 8$. In all scenarios we use model learning rates $\eta_\theta = 0.001$, adversary/threshold learning rate $\eta_c = \eta_{adv} = 0.001$ (where relevant).

For the proposed approach, FedRCVaR, we report the results for group size $\rho = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and trade-off hyperparamenter $\epsilon = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ unless stated otherwise. In the figures we present the mean performance over three runs and in separate splits. The splits are generated using 3-fold cross validation.

A.3 Implementation & Training Devices

The experiments were conducted in Python using PyTorch. We produce results for BPF [15] using the original code available at github.com/natalialmg/BlindParetoFairness. The experiments were realised using $4\times$ NVIDIA Tesla V100 GPUs.