Federated Fairness without Access to Sensitive Groups

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 and Apple
guillermo.sapiro@duke.edu

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

Abstract

Current approaches to group fairness in federated learning assume the existence of predefined and labeled sensitive groups during training. However, due to factors ranging from emerging regulations to dynamics and location-dependency of protected groups, this assumption may be unsuitable in many real-world scenarios. In this work, we propose a new approach to guarantee group fairness that does not rely on any predefined definition of sensitive groups or additional labels. Our objective allows the federation to learn a Pareto efficient global model ensuring worst-case group fairness and it enables, via a single hyper-parameter, trade-offs between fairness and utility, subject only to a group size constraint. This implies that any sufficiently large subset of the population is guaranteed to receive at least a minimum level of utility performance from the model. The proposed objective encompasses existing approaches as special cases, such as empirical risk minimization and subgroup robustness objectives from centralized machine learning. We provide an algorithm to solve this problem in federation that enjoys convergence and excess risk guarantees. Our empirical results indicate that the proposed approach can effectively improve the worst-performing group that may be present without unnecessarily hurting the average performance, exhibits superior or comparable performance to relevant baselines, and achieves a large set of solutions with different fairness-utility trade-offs.

1 Introduction

Federated learning (FL) is a paradigm that allows multiple entities/clients, usually coordinated by a central server, to collaboratively train a model to achieve some common learning objective on their combined data. The clients do not share their raw data, but rather limited information (e.g., model updates, risk values, etc.) during the training procedure [Konecný et al., 2016a, Konecný et al., 2016b]. Such a learning paradigm has been widely used for high-stakes decision-making applications such as open banking [Long et al., 2020] and genomics research [Weinstein et al., 2013] to guarantee that data is kept decentralized and private.

A key concern in federated learning is ensuring the fairness of the resulting model across various sensitive

groups [Konecný et al., 2016a], where these groups may be present in different proportions across clients. This challenge has been the focus of many works – such as [Hu et al., 2022a, Papadaki et al., 2022] – where it is assumed that clients are aware of the sensitive groups during training and can accurately assign group memberships to each data point. Nevertheless, knowledge of the sensitive groups and access to group memberships might not always be feasible for various reasons. For instance, in a scenario where various hospitals collaborate to learn a group-fair diagnostic model through federated learning, it is often unrealistic to expect that sensitive groups are identified, or that medical records include accurate information about patients' race, religion, or sexual orientation. This is because obtaining these labels can be costly [Gebru et al., 2017], require specialized knowledge, or breach privacy regulations (e.g., GDPR [European-Commission, 2018] or CCPA [Mancini, 2021]) that restrict the collection and utilization of certain types of personal information.

In this work, we address the challenge of achieving federated group fairness for *any* potential definition of sensitive groups, even those defined *after* the model is deployed, assuming the group definitions cover a significant, but configurable, subset of the overall population. We focus on the Rawlsian max-min notion of fairness [Rawls, 2001] and on the no-harm principle where degrading the performance of a particular group can only be justified if it improves a disadvantaged group. We introduce a new learning objective, the Relaxed Conditional Value-at-Risk (RCVaR), designed to enhance the performance of the worst-off subset of data samples without unnecessarily reducing the performance of the remaining ones. Our objective depends on two parameters: (a) the trade-off parameter ϵ that allows to flexibly define the importance added to the average utility versus (min-imax) fairness; and (b) the constraint ρ which bounds the size of the worst-case group, depending on some common policy/preference. Our approach enables clients to (a) identify any global and potentially critical sensitive group, independently of whether it exists on its local distribution during training; and (b) learn a global hypothesis that allows a trade-off between mean performance and fairness. An example of our method is provided in Figure 1.

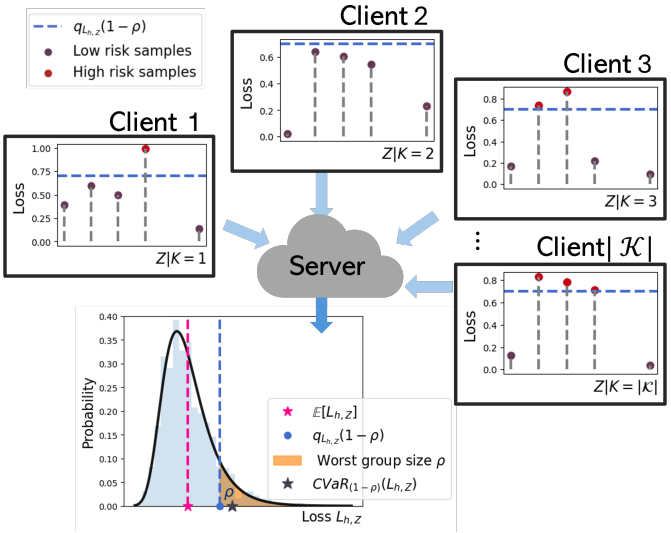

Figure 1: We consider a federated learning setting with $K \in \mathcal{K}$ clients. Our method maximizes the performance of the worst possible group of size ρ that can be formulated from the union of the local individuals/samples z, that is, within a model class, no other model performs better on its worst ρ fraction of the samples. Equivalently, no other model has a lower $(1-\rho)$-th superquantile of its loss distribution. We achieve this objective at the lowest possible cost to the non-critical samples. We make no assumptions on the distribution of the worst-performing samples amongst clients, and note that (a) worst-performing sensitive groups might not align with a single conventional demographic, and (b) 'groups' and 'clients' are not synonyms in our setting.

Main Contributions. To the best of our knowledge, we are the first to address the challenge of (minimax) Pareto federated group fairness with inhomogeneous and unknown sensitive groups, where clients and groups are not aligned. We introduce a new fairness-aware objective – RCVaR – that allows improving the performance of the high-risk samples, subject to only a group size constraint, while ensuring the best possible performance on the remaining samples. We draw formal connections between the proposed objective and existing ones, such as DRO [Hashimoto et al., 2018] and BPF [Martinez et al., 2021], demonstrating that RCVaR can also be used for learning Pareto subgroup robust models in centralized settings. We then introduce an algorithm – FedSRCVaR – that solves a smoothed approximation of RCVaR in federated learning settings. We establish the algorithm's convergence and excess risk properties, and show that the proposed objective can be easily federalized compared to centralized learning objectives. Finally, we empirically study the wide range of solutions that can be achieved by our approach through the trade-off parameter for various group sizes, and compare our method against other relevant baselines in centralized and federated learning settings using real datasets.

2 Related Work

Minimax Fairness in ML. Minimax fairness criterion – or Rawlsian max-min fairness from a utility point of view [Rawls, 2001, Martínez et al., 2020, Diana et al., 2020]–, is the no-harm (Pareto optimal) approach to equality of errors [Diana et al., 2020] since it aims to improve the model's utility for the worst-performing group without unjustifiably diminishing the performance of other groups. In the case of unknown sensitive groups, fairness is measured as the utility perceived by the worst-served subset of individuals / samples [Hashimoto et al., 2018], not as a difference in performance or outcomes between groups. In this work, we leverage this notion to learn a model that improves the worst-performing group in the most general formulation possible, that is, the worst group (subset) of samples distributed across all the clients in the federation. We note that other fairness definitions such as statistical parity, equality of odds, or equality of risks can conflict with each other, leading to sub-optimal outcomes where certain groups are harmed without any improvement to other groups, as discussed in [Kleinberg et al., 2016, Chen et al., 2018a, Barsotti and Koçer, 2022].

Fairness without Sensitive Groups in Centralized ML. A recent body of literature on centralized machine learning deals with group fairness without explicit protected groups. Early works [Gupta et al., 2018, Zhang, 2018] address the problem of unknown group annotations by designing a proxy variable that replaces the true sensitive group variable so that conventional group fairness methods can be deployed. These approaches require knowledge of the true sensitive groups, though the sample's group labels are considered unavailable, which is hard to obtain for many applications.

Subsequent research addresses fairness without group labels through (sub)group robustness. DRO [Hashimoto et al., 2018] optimizes the performance of samples that exceed a specified risk threshold. Similarly, blind Pareto fairness (BPF) [Martinez et al., 2021] minimizes the worst-case risk across any possible group distribution formulated by the training data, subject to a group size constraint, while ensuring that the model is Pareto efficient [Miettinen, 2012]. [Lahoti et al., 2020] and [Sohoni et al., 2020] use auxiliary models to discover the worst-performing group.

These works are designed for centralized settings where data is collected and processed by a single entity. Our research focuses on federated learning settings where the data is heterogeneously distributed across multiple clients and cannot be shared. We build upon DRO and BPF, due to the interpretability of their adversary, to propose a relaxed superquantile criterion that allows achieving different levels of (minimax) group fairness through a hyperparameter ϵ for a fixed group size ρ. Both DRO [Hashimoto et al., 2018] and BPF [Martinez et al., 2021] can be seen as a particular case of our proposed RCVaR formulation, as discussed in Section 3. Similar to BPF and differing from DRO, RCVaR considers properly Pareto optimal solutions and incorporates trade-offs between average performance and fairness.

Fair Federated Learning. The federated learning literature explores various notions of fairness, with a significant portion of these works [Mohri et al., 2019, Deng et al., 2020, Hu et al., 2022b, Yue et al., 2021] focusing on achieving fairness across clients. This is typically done by optimizing the model to enhance the performance of the client (or cluster of clients) that exhibits the lowest performance. However, as [Papadaki et al., 2022] formally demonstrates, fairness across clients does not guarantee fairness across different sensitive populations within those clients, except under specific circumstances, such as when each participant's dataset exclusively represents a single sensitive group.

The works in [Cui et al., 2021, Zhang et al., 2021] propose approaches to achieve group fairness within each individual client (hence, targeting only groups available within the client at training and testing time), while the works in [Rodríguez-Gálvez et al., 2021, Papadaki et al., 2022, Hu et al., 2022a] propose methods for learning models that are fair across all the known sensitive groups available in the clients, even if some participants have access to a subset of them during training. To our knowledge, the only work that considers scenarios, where group data cannot be leveraged, is [Juarez and Korolova, 2023], but it assumes that the collection of sensitive groups is known apriori (though not used due to privacy concerns). Hence, they recommend local differential private mechanisms to alleviate privacy issues, while allowing information about the group memberships to be used for learning fair models.

Similar to the aforementioned approaches, our goal is to learn a model that ensures group fairness across any sensitive groups that exist in the clients' data. However, we distinguish ourselves by focusing on a more complex scenario, where the sensitive populations are defined in terms of the performance of a given model, and cannot be labeled before the model learning process takes place. Hence, no a-priori group information can be incorporated into the training phase.

3 Problem Formulation

3.1 Minimax Fairness for Worst Case Scenarios

Let the pair of random variables $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ represent the input features and categorical targets, generated from a distribution $p(Z) = p(X, Y)$. Let also $\ell : \Delta^{|\mathcal{Y}|-1} \times \Delta^{|\mathcal{Y}|-1} \to \mathbb{R}_+$ be a loss function and h be a hypothesis drawn from the hypothesis class $\mathcal{H} = \{h : \mathcal{X} \to \Delta^{|\mathcal{Y}|-1}\}$, where $\Delta^{|\mathcal{Y}|-1}$ is the probability simplex over $\mathcal{Y}$. We assume there is no prior knowledge about the groups or sensitive labels associated with any z.

In particular, let $L_{h,Z} := \ell(h; Z)$ denote a random variable representing the loss associated with a

hypothesis $h \in \mathcal{H}$. For a predefined probability $\rho \in (0, 1)$, the $(1-\rho)$-quantile function is defined as

$$q_{L_{h,Z}}(1-\rho) := \inf \big\{\beta \in \mathbb{R} : p(L_{h,Z} \leq \beta) \geq 1-\rho\big\}, \tag{1}$$

and the $(1-\rho)$-superquantile, also known as the conditional value-at-risk (CVaR), function at confidence level $(1-\rho)$ is defined as

$$CVaR_{(1-\rho)}(L_{h,Z}) = \mathop{\mathbb{E}}_{Z}[L_{h,Z} | L_{h,Z} \geq q_{L_{h,Z}}(1-\rho)]. \tag{2}$$

The quantity in Eq. 2 is a measure of the upper tail behaviour of the distribution $p(L_{h,Z})$ and, as shown in [Rockafellar et al., 2014], it can be expressed for a bounded loss function, i.e., $0 \leq \ell(h; z) \leq B$, $B > 0$, $\forall z \in \mathcal{Z}$, as

$$CVaR_{(1-\rho)}(L_{h,Z}) = \min_{c \in [0,B]} c + \frac{1}{\rho} \mathop{\mathbb{E}}_{Z \sim p(Z)}[(L_{h,Z} - c)_+], \tag{3}$$

where $(\cdot)_+ := \max\{0, \cdot\}$ and the second term represents the regret of any positive realizations of $L_{h,Z}$. Note that the argument that minimizes the objective in Eq. 3 is the quantile $q_{L_{h,Z}}(1-\rho)$.

Therefore, we can formulate the problem of learning a minimax group fair hypothesis with no knowledge of sensitive group populations as follows,

$$h^*, c^* = \arg \min_{h \in \mathcal{H}, c \in [0,B]} c + \frac{1}{\rho} \mathop{\mathbb{E}}_{Z \sim p(Z)}[(L_{h,Z} - c)_+]. \tag{4}$$

The optimization problem in Eq. 4 allows for minimax fair solutions, since it optimizes for the worst tail risk with sample size ρ, or equivalently the worst performing samples that exceed threshold c.

Nevertheless, Eq. 4 ignores any data that is not considered high-risk (i.e., samples that are below the threshold c) and hence allows for solutions that are weakly Pareto optimal [Miettinen, 2012]. The formal definition of weakly Pareto optimality is provided in Definition 3.1.

Definition 3.1 (Weak Pareto optimality) *A hypothesis $h^* \in \mathcal{H}$ is weakly Pareto optimal if for any possible sensitive group $g \in \mathcal{G}$, $\nexists h \in \mathcal{H} : \mathop{\mathbb{E}}_{Z|g}[\ell(h; Z)] < \mathop{\mathbb{E}}_{Z|g}[\ell(h^*; Z)]$.*

Weakly Pareto optimal hypotheses can potentially compromise the performance of low-risk sensitive groups, especially when the input space exhibits regions of no uncertainty regarding the target class (i.e., the data is perfectly separable).

We next propose an objective that does not unnecessarily harm the low-risk sensitive group. We also focus on the more challenging scenario of federated learning where the data might be heterogeneously distributed across clients and the goal is to achieve a solution equivalent to centralized machine learning.

3.2 Federated Minimax Blind Fairness

In the context of federated learning, we consider an additional random variable K representing the clients in the federation. Each client $k \in \mathcal{K}$ holds data modelled by its own local distribution $p(Z|K=k) = p(X|K=k)p(Y|X, K=k)$. Therefore, the data of the entire federation can be described via the mixture distribution $p(Z) = \sum_{k \in \mathcal{K}} p(K=k)p(Z|K=K)$.

Let $L_{h,Z|K=k} := \ell(h; Z)$, with $Z \sim p(Z|K = K)$, denote a random variable representing the local loss on Z induced by a hypothesis h in client k. We formulate relaxed conditional value-at-risk (RCVaR), a generalization of the objective in Eq. 3 that (a) produces properly Pareto optimal solutions, and (b) is suitable for federated learning scenarios with inhomogeneously distributed data across clients, as follows,

$$\begin{aligned} &\min_{h\in\mathcal{H}} \Big\{(1-\epsilon)CVaR_{(1-\rho)}(L_{h,Z}) + \epsilon \mathop{\mathbb{E}}_{Z\sim p(Z)}[L_{h,Z}]\Big\} \\ &= \min_{h\in\mathcal{H}, c\in[0,B]} \mathop{\mathbb{E}}_{K}\Big[\mathop{\mathbb{E}}_{Z|k}\Big[(1-\epsilon)\Big(c + \tfrac{1}{\rho}(L_{h,Z|K=k} - c)_+\Big) + \epsilon L_{h,Z|K=k}\Big]\Big], \end{aligned} \tag{5}$$

where a hyperparameter $\epsilon \in [0, 1]$ induces a trade-off between the average and worst-case group performances. The threshold c is uniformly applied across all clients in the federation to identify the samples that belong to the *global* high-risk and the low-risk groups.[1] Also, it allows the selection of any partition of overall size ρ across clients, and to consider larger local group sizes in clients with worse performances from clients with high performance. Otherwise, assigning the same ρ across clients would yield a model that minimizes the server average of the worst per-client partition, which yields an overall partition of size ρ, but is a weaker adversary than our framework and does not guarantee the same performance as centralized settings.

Moreover, RCVaR ensures minimax properly Pareto fairness, where the worst possible group is formed by the high-risk samples subject to a predefined group size constraint ρ. This combination of minimax fairness and proper Pareto optimality is crucial for high-stakes decision-making to ensure that the produced model does not unnecessarily harm well-performing groups. The formal definition of proper Pareto optimality is offered in Definition 3.2.

Definition 3.2 (Proper Pareto optimality) *A hypothesis $h^* \in \mathcal{H}$ is properly Pareto optimal if for any possible sensitive group $g \in \mathcal{G}$, $\nexists h \in \mathcal{H} : \mathop{\mathbb{E}}_{Z|g}[\ell(h; Z)] \leq \mathop{\mathbb{E}}_{Z|g}[\ell(h^*; Z)]$, and $\exists g' \in \mathcal{G} : \mathop{\mathbb{E}}_{Z|g'}[\ell(h; Z)] < \mathop{\mathbb{E}}_{Z|g'}[\ell(h^*; Z)]$.*

By nature, and similar to DRO and BPF, RCVaR guarantees that no group partition that encompasses more than ρ of the total population, pre-defined or not, will experience a worse performance than the one obtained by optimizing Eq. 5.

Connections to CVaR and DRO. For $\epsilon = 0$, RCVaR is equivalent to CVaR in centralized learning settings, thereby allowing for minimax weakly Pareto optimal solutions. Also, CVaR is the dual formulation of DRO [Hashimoto et al., 2018] for specific uncertainty sets, as formally shown in Proposition 3 in [Hashimoto et al., 2018] and Lemma 2.1 in [Duchi et al., 2023]. Thus, RCVaR with $\epsilon = 0$ is the federated formulation of DRO [Hashimoto et al., 2018]. In contrast to standard CVaR and DRO, the additional utility term and trade-off parameter ϵ in Eq. 5 enables a larger set of achievable solutions that are proper Pareto optimal, as we discuss in Section 6. There are also algorithmic differences between optimizing DRO and RCVaR that we describe in Section 4.

[1] We focus on the high- and low-risk groups within the data distribution, accommodating any potential protected group, either binary or multigroup. The equivalence of minimax worst-case group performance on a two-group or n-group formulation, is shown in Lemma 3.1 in [Martinez et al., 2021].

Connections to BPF. Setting $\epsilon \approx 0$, sufficiently small but non-zero, produces a hypothesis that is minimax (properly) Pareto fair, since it focuses on the worst-performing samples, while still utilizing the remaining samples with a small priority ϵ. We argue that for such ϵ value, the LHS of Eq. 5 is a new expression for Pareto subgroup robustness suitable for centralized learning settings as well. We detail how RCVaR relates to BPF in Appendix A. Due to the connection of our objective with BPF, we argue that our objective inherits BPF's properties presented in [Martinez et al., 2021], including the fact that there exists a critical partition size ρ that leads to the uniform classifier for sufficiently small ϵ values.

Connections to ERM and FedAvg. If $\epsilon = 1$, our objective reduces to the vanilla-ERM objective in [McMahan et al., 2016]. For any other intermediate value of $\epsilon \in (0, 1)$, we obtain a trade-off between utility and subgroup robustness. To better understand the set of trade-offs achieved by RCVaR, we offer an illustrative example in Figure 3. We emphasize that the value of ϵ is predefined and fixed, and therefore, we leave it to the policy maker(s) to determine the fairness-utility compromise. An additional advantage of the objective in Eq. 5 is that it can be easily federalized, as shown in the sequel.

4 Optimization Method

In real applications, each client holds only a finite dataset $D_k = \{z_i^k\}_{i=1}^{n_k}$, with $z_i^k = (x_i^k, y_i^k)$, sampled from the true distribution $p(Z|K = k)$, with $D = \bigcup_{k\in\mathcal{K}} D_k$ being the dataset containing all the data samples available across clients of size $n = \sum_{k\in\mathcal{K}} n_k$. Hence, in the sequel we use the empirical form of RCVaR given by

$$\min_{\boldsymbol{\theta}\in\Theta, c\in[0,B]} \sum_{k\in\mathcal{K}} \frac{1}{n} \sum_{i=1}^{n_k} f(\boldsymbol{\theta}, c; z_i^k), \tag{6}$$

where

$$f(\boldsymbol{\theta}, c; z) = (1-\epsilon)[c + \frac{1}{\rho}(\ell(\boldsymbol{\theta}; z) - c)_+] + \epsilon\ell(\boldsymbol{\theta}; z), \tag{7}$$

and $\boldsymbol{\theta} \in \Theta$ is a vector that parametrizes the hypothesis $h \in \mathcal{H}$, and correspondingly change to the notation $\ell(\theta; z)$ instead of $\ell(h; z)$. We next offer a federated learning algorithm to solve Eq. 6 that relies on a smoothed version $\tilde{f}(\cdot)$ of the non-smooth function $f(\cdot)$.

In our federated learning setting (a) every client uses a batch size $b_k \leq n_k$ of data samples at each training iteration, (b) each client might use each local data sample more than once during the training, and (c) there are T communications between the clients and server. This realistic setting makes our algorithmic design and analysis challenging since – in order to develop a simple algorithm with strong theoretical guarantees – we need an objective that is continuously differentiable for all z. Unfortunately, even for smooth loss functions ℓ, the f in our current objective is non-smooth due to the plus function $(\cdot)_+$. To overcome this issue we consider a proxy problem of the RCVaR in Eq. 6, which relies on a smooth approximation.

4.1 Smooth Approximation of RCVaR

We consider the family of smoothed plus functions that satisfy Definition 4.1.

Definition 4.1 (Smooth Approximation [Xu and Zhang, 2009]) *For a smoothing parameter $\gamma \in \mathbb{R}_+$ and for any $m \in \mathbb{R}$, a $(\frac{2}{\gamma})-$smooth convex function $s : \mathbb{R} \to \mathbb{R}_+$ approximates a plus function $(\cdot)_+$, if it satisfies $0 \leq s(\cdot) - (\cdot)_+ \leq \gamma$.*

A smooth plus function $s(\cdot)$ becomes a more accurate approximation of the plus function, for small values of γ, as discussed in Section 5. The designed algorithm and its analysis support any function that is consistent with Definition 4.1 (e.g., soft ReLU [Peng, 1999], Zang smooth plus function [Zang, 1980], piecewise quadratic smoothed plus function [Alexander et al., 2003]), rather than a specific smoothed plus function. The empirical smooth approximation of RCVaR is formulated as

$$\min_{\boldsymbol{\theta}\in\Theta, c\in[0,B]} \frac{1}{n} \sum_{k\in\mathcal{K}} \sum_{i=1}^{n_k} \tilde{f}(\boldsymbol{\theta}, c; z_i^k) \tag{8}$$

where

$$\tilde{f}(\boldsymbol{\theta}, c; z) = (1-\epsilon)[c + \frac{1}{\rho} s(\ell(\boldsymbol{\theta}; z_i^k) - c)] + \epsilon \ell(\boldsymbol{\theta}; z_i^k).$$

4.2 FedSRCVaR: Federated Smoothed RCVaR Algorithm

Next, we introduce a federated learning algorithm designed to address Eq. 8, namely FedSRCVaR. The algorithm is outlined in Algorithm 1.

Our algorithm performs the following successive steps for T communication rounds: **(a)** the clients receive the global model-threshold pair $(\boldsymbol{\theta}^t, c^t)$ of the current round from the server; **(b)** The clients perform τ local updates on the model parameters and the threshold using b_k-samples; **(c)** The clients return the updated pair $(\boldsymbol{\theta}_k^{t+1}, c_k^{t+1})$ to the server; **(d)** Finally, the server produces the new model-threshold pair $(\boldsymbol{\theta}^{t+1}, c^{t+1})$ by averaging the received client updates.

We denote $proj_{[0,B]}$ the metric projection operator onto the set $[0, B]$ to ensure that the threshold c has a valid value within the specified range. The server averages the clients updates using the relative weights $\frac{b_k}{\sum_{k\in\mathcal{K}} b_k}$, with b_k being the batch size of client k which is proportional to n_k to allow clients to use a fraction of their local dataset, accommodating constraints such as computation limitations on the client side. The algorithm outputs the average model-threshold pair over the total communications $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ that is produced after $|\mathcal{K}|\tau T$ total updates.

Comparison to BPF and DRO methods: All methods use parameter ρ in their design. In addition to the flexibility of FedSRCVaR offered by the ϵ parameter discussed in Section 3, a key distinction between FedSRCVaR and DRO [Hashimoto et al., 2018] lies in their threshold learning approaches. Our algorithm employs (distributed) projected gradient descent with periodic averaging, while DRO relies on a (centralized) binary search method. Furthermore, FedSRCVaR can easily be deployed in dynamic learning settings, such as online learning settings, where the global model is trained using a continuous stream of new data arriving sequentially in real-time. BPF requires estimating and optimising per-sample adversarial weights at each optimization round, managing and accessing the last risk evaluation for every sample and adjusting the set from which adversarial weights are selected become computationally expensive. Finally, FedSRCVaR is lightweight even for $\tau = 1$, since it requires only the exchange of the updated model-threshold pair between clients and the server, which makes its communication overhead

Algorithm 1 FedSRCVaR Algorithm

Inputs: $\mathcal{K}$: set of clients, T: communication rounds, τ: local rounds, η: learning rate for model θ and quantile c, $\epsilon \in (0,1]$: trade-off parameter, $\rho \in (0,1)$: parameter for probability-level, b_k: local batch size.

Server initializes $\boldsymbol{\theta}^1$ randomly and sets $c^1 = B = 1$.
for $t = 1$ to T **do**
  Server **broadcasts** model-threshold $(\boldsymbol{\theta}^t, c^t)$
  for each client $k \in \mathcal{K}$ **in parallel do**
    Randomly sample a data batch of size b_k
    for $j = 1$ to τ **do**
      Set $(\boldsymbol{\theta}^{t,j=1}, c^{t,j=1}) = (\boldsymbol{\theta}^t, c^t)$
      $\boldsymbol{\theta}_k^{t,j+1} \leftarrow \boldsymbol{\theta}^{t,j} - \eta \nabla_\theta \left\{ \sum_{i=1}^{b_k} \frac{\tilde{f}(\boldsymbol{\theta}^{t,j}, c^{t,j}; z_i^k)}{b_k} \right\}$, $c_k^{t,j+1} \leftarrow c^{t,j} - \eta \nabla_c \left\{ \sum_{i=1}^{b_k} \frac{\tilde{f}(\boldsymbol{\theta}^{t,j}, c^{t,j}; z_i^k)}{b_k} \right\}$
      Return local pair $(\boldsymbol{\theta}_k^{t,\tau}, c_k^{t,\tau})$
    end for
  end for
  Server computes $\boldsymbol{\theta}^{t+1} \leftarrow \sum_{k \in \mathcal{K}} \frac{b_k \boldsymbol{\theta}_k^{t,\tau}}{\sum_{k \in \mathcal{K}} b_k}$, $c^{t+1} \leftarrow \prod_{c \in [0,B]} \left(\sum_{k \in \mathcal{K}} \frac{b_k c_k^{t,\tau}}{\sum_{k \in \mathcal{K}} b_k} \right)$
end for

Outputs: $\overline{\boldsymbol{\theta}}_T = \frac{1}{T} \sum_{t \in [T]} \boldsymbol{\theta}^t$ and $\overline{c}_T = \frac{1}{T} \sum_{t \in [T]} c^t$

insignificant compared to the communication costs and additional privacy concerns that are required for the federalization of BPF. We share more information about this comparison in Appendix A.1.

5 Algorithmic Analysis

We now examine the performance of Algorithm 1 by assessing the associated convergence rate and expected excess risk. Our analysis relies on the following assumptions.

Assumption 5.1 *The loss function $\ell(\boldsymbol{\theta}, z)$ is convex wrt z, $G-$Lipschitz, and $\beta-$smooth function of range $[0, B]$, with $B = 1$, for all z and $\boldsymbol{\theta} \in \Theta$.*

Assumption 5.2 *The set $\Theta \subseteq \mathbb{R}^d$ is convex with $||\boldsymbol{\theta} - \boldsymbol{\theta}'|| \leq M$, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$.*

Assumption 5.3 *For any model-threshold pair $(\boldsymbol{\theta}, c) \in \Theta \times [0, B]$, each client $k \in \mathcal{K}$ can query an unbiased stochastic gradient, i.e., $\mathbb{E}\big[\nabla\{\sum_{i=1}^{b} \frac{1}{b} \tilde{f}(\boldsymbol{\theta}, c; z_i^k)\}\big] = \nabla \mathop{\mathbb{E}}_{Z|k} \big[\tilde{f}(\boldsymbol{\theta}, c; Z)\big]$, with σ^2 – uniformly bounded variance, i.e.,*

$$\mathbb{E}\left[\left|\left|\tfrac{1}{b}\nabla\{\sum_{i=1}^{b} \tilde{f}(\boldsymbol{\theta}, c; z_i^k)\} - \nabla \mathop{\mathbb{E}}_{Z|k} \big[\tilde{f}(\boldsymbol{\theta}, c; Z)\big]\right|\right|^2\right] \leq \sigma^2.$$

Assumption 5.4 *The difference between local and global gradients is $\mu-$uniformly bounded, meaning that*

$$\max_{k \in \mathcal{K}} \sup_{(\boldsymbol{\theta}, c) \in \Theta \times [0,B]} \left|\left| \nabla \mathop{\mathbb{E}}_{Z|k} \big[\tilde{f}(\boldsymbol{\theta}, c; Z)\big] - \nabla \mathop{\mathbb{E}}_{Z} \big[\tilde{f}(\boldsymbol{\theta}, c; Z)\big] \right|\right| \leq \mu. \tag{9}$$

The definitions of these properties are provided in Appendix B. Under these assumptions, we can establish the core properties of the smooth and non-smooth functions, f and $\tilde{f}$, required for our analysis, in Lemma 5.5. The proof is provided in Appendix C.1.

Lemma 5.5 *Let Assumption 5.1 hold. Let also $s : \mathbb{R} \to \mathbb{R}_+$ be a $\frac{2}{\gamma}-$smooth convex function. Then,*

1. *The functions f and $\tilde{f}$ are convex for every z.*
2. *The function f and the smoothed function $\tilde{f}$ are $G_{\rho,\epsilon}-$ Lipschitz for all z with*

$$G_{\rho,\epsilon} = \max\left\{\frac{1}{\rho}\sqrt{G^2(1-\epsilon+\epsilon\rho)^2 + (1-\epsilon)^2(\rho-1)^2}, \sqrt{G^2\epsilon^2 + (1-\epsilon)^2}\right\}.$$

3. *The function $\tilde{f}$ is $\left(\frac{(1-\epsilon)}{\rho}(\beta + \frac{2}{\gamma}G^2) + \epsilon\beta\right)-$smooth.*
4. *For any model $\boldsymbol{\theta} \in \Theta$ we have that*

$$f(\boldsymbol{\theta}, c; z) \leq \tilde{f}(\boldsymbol{\theta}, c; z) \leq f(\boldsymbol{\theta}, c; z) + \tfrac{(1-\epsilon)}{\rho}\gamma. \tag{10}$$

We remark that Eq. 10 bounds the smooth function $\tilde{f}$ using f, which allows us to express our guarantees in terms of f, but importantly prove them in terms of the more analytically tractable $\tilde{f}$.

5.1 Convergence of Algorithm 1

We begin by characterizing the optimization error given by

$$\mathcal{E}_{opt} = \mathop{\mathbb{E}}_{\mathcal{A},D}\left[\sum_{z\in D}\tfrac{f(\overline{\boldsymbol{\theta}}_T, \overline{c}_T; z)}{n}\right] - \mathop{\mathbb{E}}_{D}\left[\sum_{z\in D}\tfrac{f(\boldsymbol{\theta}^*_D, c^*_D; z)}{n}\right],$$

where $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ is the average model-threshold pair after T rounds of Algorithm 1, $(\boldsymbol{\theta}^*_D, c^*_D)$ is the model-threshold pair that minimizes the smoothed objective in Eq. 8 and the outer expectation in the first term is taken over the randomness induced by our randomized algorithm $\mathcal{A}$ and the samples D, and in the second term with respect to the dataset D. This error captures how well the produced pair $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ approximates the optimal empirical pair $(\boldsymbol{\theta}^*_D, c^*_D)$ in terms of the true (non-smooth) objective function.

The next lemma offers a bound to the optimization error $\mathcal{E}_{opt}$. The proof – detailed in Appendix C.2 – leverages results for local-update gradient-based algorithms presented in Theorem 1 in [Wang et al., 2021].

Lemma 5.6 (Convergence of FedSRCVaR) *Let the assumptions 5.1 - 5.4 hold, $(\boldsymbol{\theta}^\star_D, c^\star_D)$ be the minimizer of Eq. 8, and a learning rate*

$$\eta = \min\left\{\frac{\sqrt{|\mathcal{K}|}\sqrt{M^2+1^2}}{\sigma\sqrt{\tau T}}, \left(\frac{M^2+1^2}{\sigma^2\tau^2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)T}\right)^{\frac{1}{3}}, \right.$$
$$\left. \frac{1}{4\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)}, \frac{(M^2+1^2)^{\frac{1}{3}}}{\tau(\mu^2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)T)^{\frac{1}{3}}}\right\} \tag{11}$$

Then, for the model-threshold pair $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ provided by Algorithm 1 after T rounds, we have

$$\mathcal{E}_{opt} \leq \frac{2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)(M^2+B^2)}{\tau T} + \frac{2\sigma\sqrt{M^2+B^2}}{\sqrt{|\mathcal{K}|\tau T}} + \frac{(1-\epsilon)\gamma}{\rho}$$
$$+ \left(\frac{\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)(M^2+B^2)^2}{T^2}\right)^{\frac{1}{3}}\left(5(\frac{\sigma^2}{\tau})^{\frac{1}{3}} + 19\mu^{\frac{2}{3}}\right) \tag{12}$$

Interpretation of Lemma 5.6 For $\tau = 1$, our algorithm finds a model-threshold pair $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ after T communication rounds that guarantees an optimization error of order $O\Big(\frac{1}{\tau T} + \frac{\sigma}{\sqrt{|\mathcal{K}|\tau T}} + \frac{(1-\epsilon)\gamma}{\rho}\Big)$. The first term corresponds to the deterministic convergence and the second term refers to the standard statistical noise term encountered by any algorithm that uses $|\mathcal{K}|\tau T$ total stochastic gradients. The third term depends on how accurately the smooth plus function approximates the plus function. When γ is sufficiently small, we recover the upper bound for synchronous SGD [Wang et al., 2021]. For $\tau > 1$, the last two terms in Eq. 12 appear, leading to an optimization error diminishing at a rate of $O\big(T^{-\frac{2}{3}}\big)$.

The guarantees also establish that in the presence of high data heterogeneity (i.e., $\mu \gtrsim \sigma$), the maximum number of local steps we can perform is $\tau = O(|\mathcal{K}|^{-1}(|\mathcal{K}|\tau T)^{\frac{1}{4}})$. When there is no heterogeneity, the local rounds increase to $\tau = O(|\mathcal{K}|^{-2}(|\mathcal{K}|\tau T)^{\frac{1}{2}})$. Therefore, for an appropriate selection of local rounds we can handle the error induced by the data heterogeneity across clients when we adopt multiple rounds in *lieu* of a single round per client.

5.2 Expected Excess Risk

Next, we characterize the excess risk, given by

$$\mathcal{E}_r = \mathop{\mathbb{E}}_{\mathcal{A},D}\left[\mathop{\mathbb{E}}_{K}\Big[\mathop{\mathbb{E}}_{Z|K=k}[f(\overline{\boldsymbol{\theta}}_T, \overline{c}_T; Z)]\Big]\right] - \mathop{\mathbb{E}}_{D}\left[\sum_{z\in D}\frac{f(\boldsymbol{\theta}^*_D, c^*_D; z)}{n}\right],$$

where $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ is the average model-threshold pair given by Algorithm 1 using dataset D and $(\boldsymbol{\theta}^*_D, c^*_D)$ is the optimal solutions pair that minimizes the smoothed empirical objective in Eq. 8 for the given dataset D. The outer expectation of the first term is taken over the randomness induced by our algorithm $\mathcal{A}$ and of samples D, and in the second term with respect to the samples D. The excess risk measures the difference between the expected population risk computed using the produced $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ and the expected minimum empirical risk given by the empirical optimal pair $(\boldsymbol{\theta}^*_D, c^*_D)$.

The following lemma – which relies on the excess risk analysis for stochastic gradient methods in [Hardt et al., 2016] – offers a characterization of $\mathcal{E}_r$. The proof is available in Appendix C.3.

Lemma 5.7 (Excess Risk Analysis) *Let assumptions 5.1 and 5.2 hold. Let also the learning rate* $\eta = \sqrt{n\big(\sum_{k\in\mathcal{K}} b_k\big)}\frac{\sqrt{M^2+B^2}}{G_{\rho,\epsilon}\sqrt{T(n+2T)}}$, $\tau = 1$, *and* $\gamma = \frac{2G^2_{\rho,\epsilon}}{(1-\epsilon+\epsilon\rho)^2}\eta$. *Then, for* T *communication rounds of Algorithm 1 that satisfy*

$$n\Big(\sum_{k\in\mathcal{K}} b_k\Big)\big(M^2+B^2\big)\left(\frac{\beta(1+\epsilon\rho)}{\rho G_{\rho,\epsilon}}\right)^2 \leq T(n+2T),$$

we have that

$$\mathcal{E}_r \leq \frac{G_{\rho,\epsilon}\sqrt{(M^2+B^2)\big(\frac{2}{n}+\frac{1}{T}\big)}}{\sqrt{\sum_{k\in\mathcal{K}} b_k}} + \frac{(1-\epsilon)\gamma}{\rho}.$$

Interpretation of Lemma 5.7: The bound in Lemma 5.7 indicates that, for a fixed step-size η and for choices of γ and T that satisfy the conditions stated above, our algorithm produces a pair $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ that yields an excess risk behaving as $O\left(\frac{1}{\sqrt{\sum_{k\in\mathcal{K}} b_k}}\sqrt{\frac{2}{n}+\frac{1}{T}}\right)$. This bound shows how to effectively improve

the overall performance by balancing the trade-off between optimization and generalization, since excess risk can be decomposed into a stability term[2] and an empirical optimization error term (see for example [Chen et al., 2018b]). Thus, we can directly get from Lemma 5.7 that FedSRCVaR has uniform stability of $\zeta \leq \frac{TG^2_{\rho,\epsilon}\eta}{n\sum\limits_{k\in\mathcal{K}} b_k}$. In contrast to the optimization error, the stability term scales with the communication rounds. For $T = n$, the result of Lemma 5.7 is of order $O\left(\frac{1}{\sqrt{T\sum_{k\in\mathcal{K}} b_k}}\right)$, and decreases with the square root of communication rounds times the total of batch size. Additionally, if we also have $T = \sum\limits_{k\in\mathcal{K}} b_k$ this quantity further improves and becomes $O\left(\frac{1}{T}\right)$. On the other hand, for $T \to \infty$, our bound is of order $O\left(\frac{1}{\sqrt{n\sum_{k\in\mathcal{K}} b_k}}\right)$, indicating that the excess risk scales down with square root of the number of data samples times the total batch size, meaning that we need a large number of client samples to reduce the excess risk. Moreover, if we also pick $n = \sum\limits_{k\in\mathcal{K}} b_k$ we can yield a bound that behaves as $O\left(\frac{1}{n}\right)$. We note, however, that $T \to \infty$ creates a communication bottleneck in federated learning systems, since there is a considerably large amount of messages that are exchanged between clients and server.

6 Experimental Results

We empirically demonstrate the advantages of the proposed approach on four datasets: (a) *eICU* [Pollard et al., 2018], a dataset with records from various medical centres that we use to predict patient mortality. The data is distributed to 11 clients and each client is mapped to a single hospital in the dataset. (b) *ACS Employment* [Ding et al., 2021] for employment classification based on 14 input features. The data is assigned to 51 clients based on geolocation. (c) *MNIST* [Deng, 2012], a grayscale image dataset that we use to classify 10 handwritten digits where each digit is allocated to a client. (d) *Celeb-A* [Liu et al., 2015], a dataset with facial images from celebrities. The target task is gender prediction and the data is randomly assigned to two clients.

For all methods, the *worst group* refers to the subset of test samples with losses higher than the $(1-\rho)$-quantile of the empirical test loss distribution. The *best group* comprises the remaining test samples. We measure *utility/mean* as the average risk across all test samples and define the *group risk disparity* as the risk difference between the best and worst groups. Further experimental details and additional experiments are provided in Appendix D.

6.1 Comparison to ML and FL Baselines

To our knowledge, this is the first work that addresses fairness without access to sensitive groups in FL settings. Hence, we compare our approach with (a) the centralized ML fairness baselines DRO and BPF that aim to achieve fairness without relying on group information, but also with ERM that optimizes for utility; and (b) the FL approaches AFL [Mohri et al., 2019] which ensures client fairness, and FedAvg [McMahan et al., 2016] which optimizes for utility, disregarding fairness. We note that the DRO and BPF are the hardest baselines to compare with, since they use a centralized dataset and can obtain the optimal solution. We report our results in Figure 2.

For $\epsilon \approx 0$, FedRCVaR provides a model with the best performance on the worst group risk, along with BPF

[2]We use algorithmic stability (see Definition B.8) to control the generalization error.

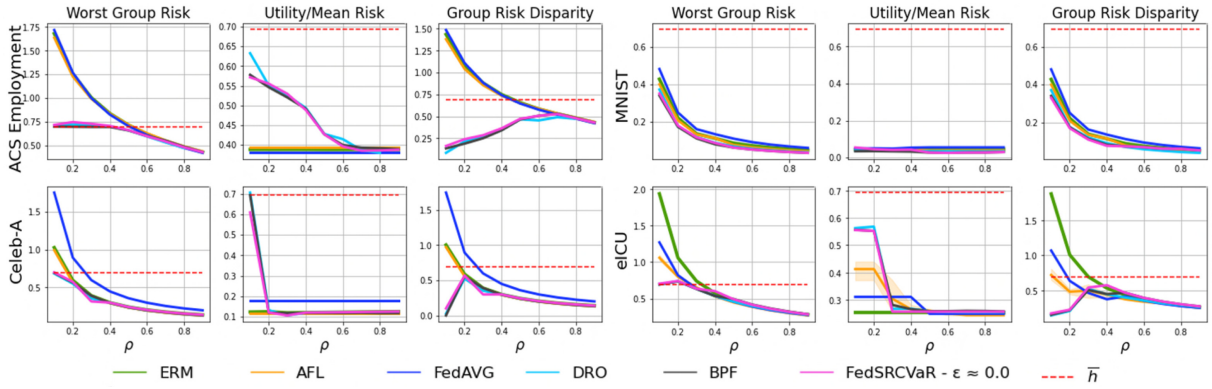

Figure 2: Comparison of worst group risk, utility risk and group risk disparity between the best and worst groups on different datasets. $\overline{h}$ denotes the uniform classifier. FedRCVaR recovers solutions equivalent to centralized settings, while improving both utility and fairness compared to FL baselines in many settings. ρ is a hyperparameter of FedSRCVaR, DRO and BPF. Differences in average performance as a function of ρ for ERM, FedAVG and AFL are due to the variation of the training hyperparameters, since for each ρ we report the hyperparameter combination producing the model with the best performance for each method.

and DRO, confirming that our approach effectively produces minimax fair and robust solutions. In some cases, DRO exhibits a higher average risk compared to FedRCVaR and BPF, despite similar performances in the worst group risk, which indicates that DRO underperforms on the remaining population for these cases.

AFL and FedAvg underperform on the worst group compared to our approach and perform better on the utility task for low ρ values, as expected, since they put different and possibly higher weights on the low-risk samples than FedSRCVaR. We note that AFL maximizes performance over the worst client, while FedSRCVaR optimizes for the worst ρ-sized partition across all samples and clients in the federation. We also notice that FedRCVaR improves both worst group fairness and utility performance simultaneously in some datasets, outperforming AFL and FedAvg. This suggests that for small ϵ, minimizing the right-tail risk of the samples is more effective and overall better in handling heterogeneity within the federation but also between training and testing sets. The results of FedSRCVaR and FedAVG vastly vary in most settings except for (a) high worst group size ρ since the worst group will consist of most samples, and/or (b) high ϵ values for which the utility term has more importance than the fairness term. For $\rho = 1$ our method is theoretically equivalent to FedAvg.

When the produced models are examined on larger values worst group size ρ, the risk variance across all different approaches is low, as expected. As we discuss in Remark A.1, Appendix A, there is a critical partition size ρ that leads to the uniform classifier $\overline{h}$ for sufficiently small ϵ values, which, in conjunction with the generalisation error, justifies the superior performance of $\overline{h}$ for small ρs, as illustrated in our results.

6.2 Achieving Various Trade-Offs through FedSRCVaR

We empirically assess the trade-offs FedSRCVaR can accomplish for various combinations of $\epsilon \in \{0.01, 0.1, \ldots, 0.9, 1.0\}$ and $\rho \in \{0.1, \ldots, 0.9\}$ in Figure 3. The different colours indicate a particu-

lar ϵ value and we report results for models that were trained individually for each pair of (ϵ, ρ) values. For ACS Employment we distribute the data to 3 clients based on the race classes: {Black, White, Others}.

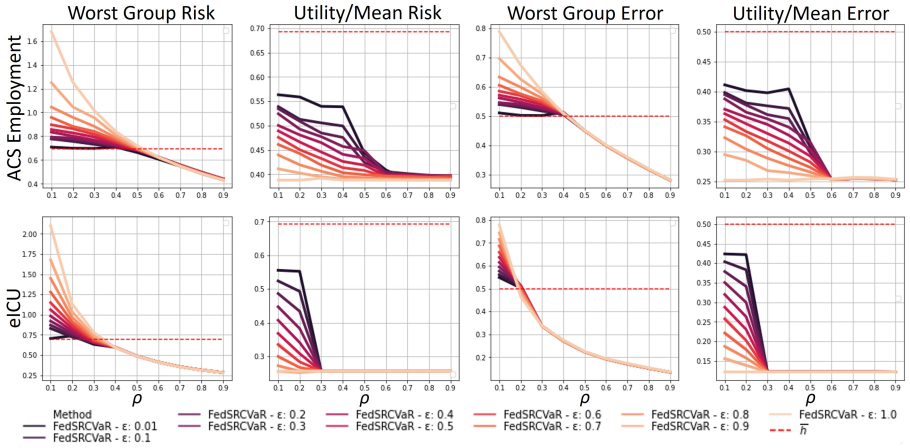

Figure 3: Performance trade-offs among worst group and utility for different pairs of $(\epsilon, \rho) \in \{0.01, 0.1, \ldots, 0.9, 1.0\} \times \{0.1, \ldots, 0.9\}$ values on real datasets. The different colours indicate different ϵ values. $\overline{h}$ denotes the uniform classifier. A lower score indicates better performance. We report the worst group and average/utility risks, as a function of ρ.

Figure 3 shows that ϵ effectively acts as a tuning parameter between worst group fairness and average performance. For small ρ values, ϵ has a significant impact on the worst group and the utility performance. We observe that the larger the ϵ the lower the average utility errors and risks, while as we decrease ϵ we boost the performance on worst group. Note that for $\epsilon \approx 0$ and $\rho \approx 0$, the worst-group risk is close to the uniform classifier risk which is consistent with Remark A.1 in Appendix A.1, and conclusions drawn about the existence of a critical worst-group size under which we yield the uniform classifier in [Martinez et al., 2021]. On the other hand, for large ρs we notice that all solutions are equivalent and the parameter ϵ has almost no influence on the solution. Interestingly, for particular values of ϵ and ρ, FedSRCVaR can recover client robustness solutions (akin to AFL), even though our objective does not explicitly aim for that.

6.3 FedSRCVaR with Multiple Local Rounds

In Figure 4 we compare the performance of FedSRCVaR for $\tau \in \{1, 5, 10\}$, and FedAVG, on the ACS Employment and eICU datasets. For small ρ values, FedSRCVaR for $\tau \in \{5, 10\}$ exhibits a higher worst-group risk compared to the FedSRCVaR with $\tau = 1$. This suggests that conducting multiple local rounds may result in inferior performance for the worst group when ρ is small, as indicated in our convergence guarantees. Moreover, as ρ increases, FedSRCVaR for $\tau \in \{5, 10\}$ demonstrates improvement in worst-case fairness similar to FedSRCVaR with $\tau = 1$. This implies that the impact of performing additional local epochs becomes less significant as the worst-group size becomes larger. For sufficiently large values of ρ, both methods converge to the same solution as FedAVG.

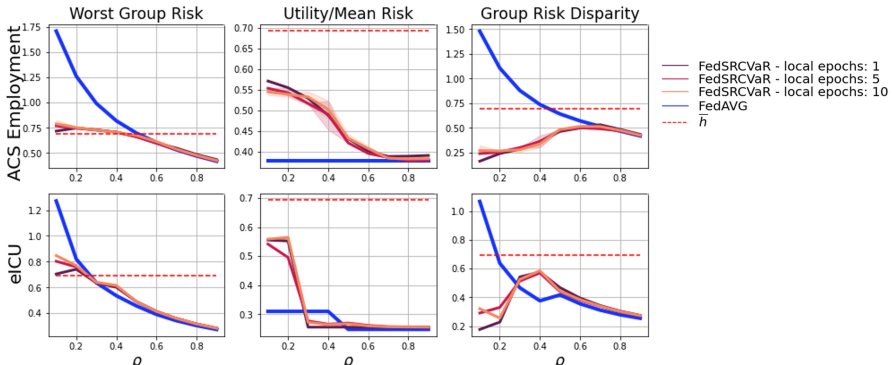

Figure 4: Performance comparison between FedSRCVaR for local epochs $\tau \in \{1, 5, 10\}$ and FedAVG. $\overline{h}$ denotes the uniform classifier. We report the worst group and average/utility risks, and the group risk disparity between the worst-performing samples and the remaining population, as a function of ρ.

7 Conclusions and Limitations

Federated learning is crucial to obtaining large, representative datasets across every sensitive group. Ensuring fairness across protected groups is essential for responsible machine learning, but prior knowledge of such groups is not always available, due to privacy constraints and evolving fairness requirements.

This is the first work to present a flexible federated learning objective to ensure minimax Pareto fairness with respect to any group of sufficient size. We propose an algorithm that solves a proxy of the proposed objective, providing performance guarantees in the convex setting. Experimentally, our approach surpasses relevant FL baselines, exhibits comparable performance to centralized ML approaches, and demonstrates the ability to achieve a diverse range of solutions. For a single local epoch, FedSRCVaR is robust to data imbalances and heterogeneity across clients, yielding the same solution as centralized ML settings. For multiple local epochs, FedSRCVaR improves communication costs, but might yield a suboptimal solution when data are highly non-iid across clients.

Acknowledgments

UCL authors were supported by Cisco under grant #217462. GS is partially supported by NSF, ONR, NGA, and the Simons Foundation.

References

[Alexander et al., 2003] Alexander, S., Coleman, T. F., and Li, Y. (2003). Derivative portfolio hedging based on cvar. *New Risk Measures in Investment and Regulation: Wiley*.

[Barsotti and Koçer, 2022] Barsotti, F. and Koçer, R. G. (2022). Minmax fairness: from rawlsian theory of justice to solution for algorithmic bias. *AI & SOCIETY*, pages 1–14.

[Bousquet and Elisseeff, 2002] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Chen et al., 2018a] Chen, I., Johansson, F. D., and Sontag, D. (2018a). Why is my classifier discriminatory?

[Chen et al., 2018b] Chen, Y., Jin, C., and Yu, B. (2018b). Stability and convergence trade-off of iterative optimization algorithms.

[Cui et al., 2021] Cui, S., Pan, W., Liang, J., Zhang, C., and Wang, F. (2021). Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34:26091–26102.

[Deng, 2012] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

[Deng et al., 2020] Deng, Y., Kamani, M. M., and Mahdavi, M. (2020). Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122.

[Diana et al., 2020] Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. (2020). Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 32:35–95.

[Ding et al., 2021] Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490.

[Duchi et al., 2023] Duchi, J., Hashimoto, T., and Namkoong, H. (2023). Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664.

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

[Gebru et al., 2017] Gebru, T., Krause, J., Deng, J., and Fei-Fei, L. (2017). Scalable annotation of fine-grained categories without experts. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1877–1881, New York, NY, USA. Association for Computing Machinery.

[Gupta et al., 2018] Gupta, M., Cotter, A., Fard, M. M., and Wang, S. (2018). Proxy fairness.

[Hardt et al., 2016] Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The*

33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA. PMLR.

[Hashimoto et al., 2018] Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

[Hu et al., 2022a] Hu, S., Wu, Z. S., and Smith, V. (2022a). Fair federated learning via bounded group loss.

[Hu et al., 2022b] Hu, Z., Shaloudegi, K., Zhang, G., and Yu, Y. (2022b). Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4):2039–2051.

[Juarez and Korolova, 2023] Juarez, M. and Korolova, A. (2023). "you can't fix what you can't measure": Privately measuring demographic performance disparities in federated learning. In *Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, pages 67–85. PMLR.

[Kleinberg et al., 2016] Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807.

[Konecný et al., 2016a] Konecný, J., McMahan, H. B., Ramage, D., and Richtárik, P. (2016a). Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527.

[Konecný et al., 2016b] Konecný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016b). Federated learning: Strategies for improving communication efficiency. In *NeurIPS Workshop on Private Multi-Party Machine Learning*.

[Lahoti et al., 2020] Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc.

[Liu et al., 2015] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[Long et al., 2020] Long, G., Tan, Y., Jiang, J., and Zhang, C. (2020). *Federated Learning for Open Banking*, pages 240–254. Springer International Publishing, Cham.

[Mancini, 2021] Mancini, J. (2021). Data portability, interoperability and digital platform competition: Oecd background paper.

[Martinez et al., 2021] Martinez, N. L., Bertran, M. A., Papadaki, A., Rodrigues, M., and Sapiro, G. (2021). Blind pareto fairness and subgroup robustness. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7492–7501. PMLR.

[Martínez et al., 2020] Martínez, N., Bertran, M., and Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. *Proceedings of machine learning research*, 119:6755–6764.

[McMahan et al., 2016] McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. (2016). Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629.

[Miettinen, 2012] Miettinen, K. (2012). *Nonlinear Multiobjective Optimization*, volume 12. Springer Science & Business Media.

[Mohri et al., 2019] Mohri, M., Sivek, G., and Suresh, A. (2019). 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). 36th International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019.

[Papadaki et al., 2022] Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. (2022). 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. Association for Computing Machinery.

[Peng, 1999] Peng, J.-M. (1999). *A Smoothing Function and Its Applications*, pages 293–316. Springer US, Boston, MA.

[Pollard et al., 2018] Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5.

[Rawls, 2001] Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

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

[Rodríguez-Gálvez et al., 2021] Rodríguez-Gálvez, B., Granqvist, F., van Dalen, R., and Seigel, M. (2021). Enforcing fairness in private federated learning via the modified method of differential multipliers.

[Sohoni et al., 2020] Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. (2020). No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352.

[Wang et al., 2021] Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., y Arcas, B. A., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., Diggavi, S. N., Eichner, H., Gadhikar, A., Garrett, Z., Girgis, A. M., Hanzely, F., Hard, A., He, C., Horvath, S., Huo, Z., Ingerman, A., Jaggi, M., Javidi, T., Kairouz, P., Kale, S., Karimireddy, S. P., Konečný, J., Koyejo, S., Li, T., Liu, L., Mohri, M., Qi, H., Reddi, S. J., Richtárik, P., Singhal, K., Smith, V., Soltanolkotabi, M., Song, W., Suresh, A. T., Stich, S. U., Talwalkar, A., Wang, H., Woodworth, B. E., Wu, S., Yu, F. X., Yuan, H., Zaheer, M., Zhang, M., Zhang, T., Zheng, C., Zhu, C., and Zhu, W. (2021). A field guide to federated optimization. *CoRR*, abs/2107.06917.

[Weinstein et al., 2013] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

[Xu and Zhang, 2009] Xu, H. and Zhang, D. (2009). Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Math. Program.*, 119:371–401.

[Yue et al., 2021] Yue, X., Nouiehed, M., and Kontar, R. A. (2021). GIFAIR-FL: an approach for group and individual fairness in federated learning. *CoRR*, abs/2108.02741.

[Zang, 1980] Zang, I. (1980). A smoothing-out technique for min–max optimization. *Math. Program.*, 19(1):61–77.

[Zhang et al., 2021] Zhang, F., Kuang, K., Liu, Y., Wu, C., Wu, F., Lu, J., Shao, Y., and Xiao, J. (2021). Unified group fairness on federated learning.

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

A Formal Connection to Pareto Subgroup Robustness from Centralized ML

An existing approach to fair centralized machine learning through subgroup robustness is BPF [Martínez et al., 2020]. In this Appendix, we explore the relationship between the proposed RCVaR objective and BPF objective, by showing via the Lagrange dual that one can recover RCVaR from the BPF.

To establish this connection, we consider the setting described in Section 3.1. We also introduce an additional random variable G which indicates whether a certain input-target pair belongs to the worst-performing group, i.e., the group with the high-risk samples. We let $G = 1$ denote the samples belonging to the worst performing group and $G = 0$ for the remaining data that do not belong to the worst group. We refer to the group $G = 0$ as the best-performing group.

By setting the worst group size to be equal to the probability ρ, i.e., $p(G = 1) = \rho$, and constraining $p(G = 1|Z) > \epsilon$, the BPF objective is defined as

$$\min_{h\in\mathcal{H}} \max_{\substack{p(G=1|Z) \\ \text{s.t. } p(G=1)=\rho \\ p(G=1|Z)>\epsilon}} \mathop{\mathbb{E}}_{Z\sim p(Z)} \left[\frac{p(G=1|Z)}{p(G=1)} L_{h,Z}\right] = \min_{h\in\mathcal{H}} \max_{\substack{\lambda(Z)\in Q_{\epsilon,\rho} \\ \text{s.t. } \int_{\mathcal{Z}} \lambda(z)dz=1}} \int_{\mathcal{Z}} \lambda(z)\ell(h;z)dz, \tag{BPF}$$

where $Q_{\epsilon,\rho} = \left\{\lambda(\cdot) : \lambda(z) \in \left[\frac{p(z)}{\rho}\epsilon, \frac{p(z)}{\rho}\right]\right\}$ and $\lambda(Z) \in Q_{\epsilon,\rho}$ is the density of variable Z. We define the Lagrangian of BPF as

$$\mathrm{L}_{BPF}(\lambda(Z), \mu) = \int_{\mathcal{Z}} \lambda(z)\ell(h;z)dz + \mu^*\left(1 - \int_{\mathcal{Z}} \lambda(z)dz\right) = \int_{\mathcal{Z}} \lambda(z)(\ell(h;z) - \mu^*)dz + \mu^*,$$

where μ^* is the Lagrange multiplier of the constraint in BPF. For a fixed hypothesis $h \in \mathcal{H}$, the optimal density $\lambda^*(Z)$ satisfies

$$\lambda^*(Z) = \arg\max_{\lambda(Z)\in Q_{\epsilon,\rho}} \mathrm{L}_{BPF}(\lambda(Z), \mu) = \arg\max_{\lambda(Z)\in Q_{\epsilon,\rho}} \int_{\mathcal{Z}} \lambda(z)(\ell(h;z) - \mu^*)dz = \begin{cases} \frac{p(z)}{\rho}, & \text{if } \ell(h;z) > \mu^* \\ \frac{p(z)}{\rho}\epsilon, & \text{if } \ell(h;z) \le \mu^* \end{cases}$$

Furthermore, using the fact that $\int_{\mathcal{Z}} \lambda(z)dz = 1$, we can compute the Lagrange multiplier μ^* as follows

$$\int_{\ell(h;z)\le\mu^*} \frac{\epsilon}{\rho}p(z)dz + \int_{\ell(h;z)>\mu^*} \frac{p(z)}{\rho}dz = 1 \Longrightarrow \frac{\epsilon}{\rho}\int_{\mathcal{Z}} p(z)dz + \left(\frac{1-\epsilon}{\rho}\right)\int_{\ell(h;z)>\mu^*} p(z)dz = 1$$

$$\Longrightarrow \int_{\ell(h;z)>\mu^*} p(z)dz = \frac{\rho-\epsilon}{(1-\epsilon)} \Longrightarrow \int_{\ell(h;z)\le\mu^*} p(z)dz = 1 - \frac{\rho-\epsilon}{(1-\epsilon)} \Longleftrightarrow \mu^* = F^{-1}(1-\rho'),$$

$$(\text{w/ } \rho' = \tfrac{\rho-\epsilon}{(1-\epsilon)})$$

where $F^{-1}(\cdot)$ corresponds to the inverse of $F(\cdot)$, and $F(L_{h,Z})$ is the cumulative distribution function of $L_{h,Z}$.

Then, by substituting the optimal density $\lambda^*(Z)$ and μ^* in the BPF objective we get

$$
\begin{aligned}
& \min_{h\in\mathcal{H}} \int_{\ell(h;z)\leq\mu^*} \epsilon\tfrac{p(z)}{\rho}\ell(h;z)dz + \int_{\ell(h;z)>\mu^*} \tfrac{p(z)}{\rho}\ell(h;z)dz \\
= & \min_{h\in\mathcal{H}} \tfrac{\epsilon}{\rho}\int_{\mathcal{Z}} p(z)\ell(h;z)dz + \tfrac{1-\epsilon}{\rho}\int_{\ell(h;z)>F^{-1}(1-\rho')} p(z)\ell(h;z)dz \quad (\text{Recall } \rho = \textstyle\int_{\ell(h;z)>F^{-1}(1-\rho')} p(z)dz) \\
= & \min_{h\in\mathcal{H}} \tfrac{\epsilon}{\rho} \mathop{\mathbb{E}}_{Z\sim p(Z)}[L_{h,Z}] + (1-\epsilon) \mathop{\mathbb{E}}_{Z\sim p(Z)}\left[L_{h,Z}|L_{h,Z} > F^{-1}(1-\rho')\right] \\
= & \min_{h\in\mathcal{H}} \tfrac{\epsilon}{\rho} \mathop{\mathbb{E}}_{Z\sim p(Z)}[L_{h,Z}] + (1-\epsilon)CVaR_{1-\rho'}(L_{h,Z}) \\
= & \min_{h\in\mathcal{H}} \underbrace{\epsilon \mathop{\mathbb{E}}_{Z}[L_{h,Z}] + (1-\epsilon)CVaR_{1-\rho'}(L_{h,Z})}_{\text{RCVaR for } \rho'=\frac{\rho-\epsilon}{(1-\epsilon)} \text{ and trade-off } \epsilon\approx 0} + \underbrace{\frac{\epsilon(1-\rho)}{\rho}\mathop{\mathbb{E}}_{Z}[L_{h,Z}]}_{\text{Add'l term}} \\
\approx & \min_{h\in\mathcal{H}} \epsilon \mathop{\mathbb{E}}_{Z\sim p(Z)}[L_{h,Z}] + (1-\epsilon)CVaR_{1-\rho'}(L_{h,Z}) \quad (\text{since } \epsilon \approx 0)
\end{aligned}
\tag{13}
$$

Eq. 13 shows that BPF is equivalent to RCVaR plus an additional error term for the same ϵ and $\rho' = \frac{\rho-\epsilon}{(1-\epsilon)}$. We remark that for sufficiently small ϵ values (i.e., $\epsilon \approx 0$) the additional term is negligible and probability $\rho' \approx \rho$, which makes the two objectives equivalent.

Due to this equivalence, we argue that the results presented in Lemma 3.2 in [Martínez et al., 2020], which proves the existence of a critical partition size ρ that leads to the uniform classifier for sufficiently small ϵ values; and Lemma 3.3 in [Martínez et al., 2020], that studies the penalty in performance when we use subgroup robustness instead of known groups, apply also to our case for $\epsilon \approx 0$. In Remark A.1, we state the impact of the threshold c in the resulting hypothesis in our objective, RCVaR.

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

A.1 Federalization of BPF

One of the merits of RCVaR compared to BPF is that it can easily be deployed in dynamic machine learning settings, such as online learning settings, where we (continue to) optimize the global model using a stream of new data arriving sequentially in real-time both in centralized and federated learning settings. In contrast, deploying BPF may present complexities and limitations due to the need for estimating and optimizing per-sample adversarial weights at each optimization round. Since data arrives sequentially, managing and accessing the last risk evaluation for every sample and adjusting the set from which adversarial weights are selected become computationally expensive. These factors can impede the efficiency of the learning process.

Another advantage of RCVaR is its suitability for federated learning. In Algorithm 1 we showed that the federalization of RCVaR requires only the exchange of the updated model-threshold pair between clients and the server. This efficient exchange allows for achieving a solution equivalent to our centralized BPF objective in [Martinez et al., 2021]. In contrast, federalizing BPF poses challenges, such as (a) failure to provide guarantees that the produced global model is equivalent to centralized BPF; or (b) significantly amplifying the computation and communication costs of the federated learning procedure, and raising additional privacy concerns.

In particular, one approach to federalize BPF is to assign a common value of ρ across all clients (i.e., $\rho = \rho_k \forall k \in \mathcal{K}$), resulting in an overall partition of size ρ. However, this approach has a weaker adversary compared to our proposed framework. In our framework, the adversary has the flexibility to choose any partition of size ρ and allocate more weight to clients with worse performances by shifting their budget (ρ_k) from clients with higher utility. Also, by assigning a fixed $\rho = \rho_k$, we have no guarantee that the solution is equivalent to (centralized) BPF, while our proposal is guaranteed to produce an equivalent solution.

Another way to federalize BPF would be for clients to share information about their local loss distributions with the server. This requires the clients to know their local group sizes (i.e., ρ_k) to ensure global Pareto subgroup robustness. However, due to the data heterogeneity across clients, there is no guarantee or prior knowledge about how the global worst group is distributed across clients at each communication round t. Thus, acquiring this information would involve additional communication rounds, increased computations on the client side, and additional computation on the server side to correctly hash and share the local group sizes. This not only amplifies the computation and communication costs but also raises privacy concerns as clients need to share information about their local loss distributions.

B Basic Definitions

We first provide some standard definitions and remarks. In what follows, the norm $||\cdot||$ denotes the Euclidean norm.

Definition B.1 (Convex Set) *A set Θ is convex if for any two points $\theta_1, \theta_2 \in \Theta$ we have that their convex combination*

$$\theta = \mu\theta_1 + (1-\mu)\theta_2, \text{ with } \mu \in [0,1],$$

also belongs to Θ, i.e. $\theta \in \Theta$.

Definition B.2 (Convex Function) *A function f with domain $dom(f)$ is convex if and only if $dom(f)$ is a convex set and for all $v, w \in dom(f)$ we have that*

$$f(\mu w + (1-\mu)v) \leq \mu f(w) + (1-\mu)f(v), \text{ with } \mu \in [0,1]. \tag{14}$$

Definition B.3 (Lipschitzness) *A function $f : dom(f) \to \mathbb{R}$ is G-Lipschitz if for all $v, w \in dom(f)$ we have that*

$$||\nabla f(w)|| \leq G \quad \text{and} \quad |f(w) - f(v)| \leq G||w - v||, \text{ for some } G > 0. \tag{15}$$

Lemma B.4 (First Order Condition) *A differentiable function f with domain $dom(f)$ is convex if $dom(f)$ is convex and for any $v, w \in dom(f)$ we have that*

$$f(w) \leq f(v) + \langle \nabla f(v), w - v \rangle.$$

Proof. For proof see section 3.1.3 in [Boyd and Vandenberghe, 2004]. □

Definition B.5 (Smoothness) *A function $f : dom(f) \to \mathbb{R}$ is β-smooth if it is continuously differentiable and its gradient is β-Lipschitz, i.e. $\exists \beta : \forall v, w \in dom(f)$ we have that*

$$||\nabla f(w) - \nabla f(v)|| \leq \beta ||w - v||.$$

Definition B.6 (Sub-gradient) *Let $f : dom(f) \to \mathbb{R}$, with $dom(f) \subseteq \mathbb{R}^d$. Then $g \in \mathbb{R}^d$ is a subgradient of f at point v if for any $w \in dom(f)$ we have that*

$$f(w) - f(v) \geq \langle g, w - v \rangle.$$

Note that the subgradient g might not be unique. We denote $\partial f(v)$ the set of subgradients computed at a point v, also called subdifferential of f at v, where $g \in \partial f(v)$. We also note that when a function $f : dom(f) \to \mathbb{R}$ is G-Lipschitz and convex, Eq. 15 becomes $||g|| \leq G$. We provide this elementary proof in Lemma B.7.

Lemma B.7 *Let a function $f : dom(f) \to \mathbb{R}$ be a $G-$Lipschitz continuous and $\partial f(v) \neq \emptyset$. Then, for any $v \in dom(f)$ we have that*

$$||g|| \leq G, \quad \text{with} \quad g \in \partial f(v) \tag{16}$$

Proof. Since f is $G-$Lipschitz we have that

$$|f(w) - f(v)| \leq G||w - v||, \text{ for some } G > 0$$

Also from the subgradient definition we know that

$$f(w) - f(v) \geq \langle g, w - v \rangle.$$

Combining the two inequalities we get that $||g|| \leq G$.

□

Definition B.8 (Uniform Stability, [Bousquet and Elisseeff, 2002]) *Let $f : dom(f) \to \mathbb{R}$. A randomized algorithm $\mathcal{A}$ is $\zeta-$uniformly stable if for any datasets D, D' that differ in at most a single sample, we have that*

$$\sup_z \mathbb{E}[f(\mathcal{A}(D); z) - f(\mathcal{A}(D'); z)] \leq \zeta \tag{17}$$

where $\zeta > 0$ and the expectation is w.r.t. the randomness of the algorithm and the samples.

C Analysis of Algorithm 1

In this section, we analyze the properties of Algorithm 1, as stated in Lemmas 5.6 and 5.7.

C.1 Smooth Approximation of Eq. 6

In order to provide an algorithmic analysis for our setting, we require the auxiliary function f, defined as

$$f(\boldsymbol{\theta}, c; z) = (1-\epsilon)[c + \frac{1}{\rho}(\ell(\boldsymbol{\theta}; z) - c)_+] + \epsilon\ell(\boldsymbol{\theta}; z)$$

to be smooth. For this reason, we define the smoothed version of the auxiliary function f as

$$\tilde{f}(\boldsymbol{\theta}, c; z) = (1-\epsilon)[c + \frac{1}{\rho}s(\ell(\boldsymbol{\theta}; z) - c)] + \epsilon\ell(\boldsymbol{\theta}; z),$$

where s is a convex and $(\frac{2}{\gamma})-$smooth function.

Given a function s that satisfies the conditions given in Definition 4.1, we provide the properties for the auxiliary function f and the smoothed function $\tilde{f}$ in 5.5.

Proof of Lemma 5.5. We use the same numbering to prove each case.

1. For the convexity of f we just need to show that $(\cdot)_+$ is convex. For a fixed $j \in \{1, \dots, m\}$, with $m > 0$, and $\lambda \in [0, 1]$, we have that

$$y_j \leq \max_i y_i, \quad x_j \leq \max_i x_i$$
$$\text{and thus} \quad \lambda x_j + (1-\lambda)y_j \leq \lambda \max_i x_i + (1-\lambda)\max_i y_i.$$

Consequently, we also have $\max_j[\lambda x_j + (1-\lambda)y_j] \leq \lambda \max_i x_i + (1-\lambda)\max_i y_i$. Note that in our scenario $m = 2$ and $y_j \in \{0, \ell(\boldsymbol{\theta}, z_i) - c\}$ and $x_j \in \{0, \ell(\boldsymbol{\theta}, z_l) - c\}$ with $j \in [m]$ and $i, l \in [n]$.

Note that the smoothed plus function s is convex w.r.t. z by definition. Thus, the convexity of the function $\tilde{f}$ is immediate since $\tilde{f}$ is a linear combination of convex terms.

2. For the second property, we let g denote the subgradient of f for a fixed pair of $(\boldsymbol{\theta}, c)$ (i.e. $g \in \partial_{(\boldsymbol{\theta},c)} f(\boldsymbol{\theta}, c; z)$). As we see in the previous appendix, the Euclidean norm of the subgradient of a convex and $G_{\rho,\epsilon}-$Lipschitz function, is upper bounded by $G_{\rho,\epsilon}$, i.e. $||g|| \leq G_{\rho,\epsilon}$. Thus, we work out a Lipschitzness parameter by finding an upper bound for the subgradient of f, $\quad \forall g \in \partial_{(\boldsymbol{\theta},c)} f(\boldsymbol{\theta}, c; z)$.

We remark that the plus function $(\cdot)_+$ in function f, induces three scenarios for any z: (i) $\ell(\boldsymbol{\theta}; z) > c$, (ii) $\ell(\boldsymbol{\theta}; z) = c$, or (iii) $\ell(\boldsymbol{\theta}; z) < c$. Thus, we define the set of subgradients as

$$\partial_{(\boldsymbol{\theta},c)} f(\boldsymbol{\theta}, c; z) = \begin{cases} \begin{bmatrix} \left(\frac{(1-\epsilon)}{\rho} + \epsilon\right) \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; z) \\ (1-\epsilon)(1-\frac{1}{\rho}) \end{bmatrix}, & \text{if } \ell(\boldsymbol{\theta}; z) > c \\ \begin{bmatrix} \left(\frac{(1-\epsilon)t}{\rho} + \epsilon\right) \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; z) \\ (1-\epsilon)(1-\frac{t}{\rho}) \end{bmatrix}, & \text{if } \ell(\boldsymbol{\theta}; z) = c,\ t \in [0,1] \\ \begin{bmatrix} \epsilon \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; z) \\ 1-\epsilon \end{bmatrix}, & \text{if } \ell(\boldsymbol{\theta}; z) < c \end{cases} \tag{18}$$

Consequently, $\forall g \in \partial_{(\boldsymbol{\theta},c)} f(\boldsymbol{\theta}, c; z)$ we get

$$||g||^2 \leq \begin{cases} G^2 \left(\frac{(1-\epsilon)}{\rho} + \epsilon\right)^2 + (1-\epsilon)^2 (1-\frac{1}{\rho})^2, & \text{if } \ell(\boldsymbol{\theta}; z) > c \\ \max\limits_{t \in [0,1]} \left[G^2 \left(\frac{(1-\epsilon)t}{\rho} + \epsilon\right)^2 + (1-\epsilon)^2 (1-\frac{t}{\rho})^2 \right], & \text{if } \ell(\boldsymbol{\theta}; z) = c \\ G^2 \epsilon^2 + (1-\epsilon)^2, & \text{if } \ell(\boldsymbol{\theta}; z) < c \end{cases}$$

$$\Rightarrow ||g|| \leq \max \left\{ \sqrt{\frac{G^2(1-\epsilon+\epsilon\rho)^2 + (1-\epsilon)^2(\rho-1)^2}{\rho^2}}, \sqrt{G^2\epsilon^2 + (1-\epsilon)^2} \right\}$$

Using similar reasoning, we can show that the smoothed auxiliary function $\tilde{f}(\boldsymbol{\theta}, c; z)$ is $G_{\rho,\epsilon}-$Lipschitz for all z. Let s' be the derivative of the smoothed plus function s. We have that

$$\nabla \tilde{f}(\boldsymbol{\theta}, c; z) = \begin{bmatrix} \left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta};z)-c)}{\rho} + \epsilon\right) \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; z) \\ (1-\epsilon)\left(1 - \frac{s'(\ell(\boldsymbol{\theta};z)-c)}{\rho}\right) \end{bmatrix} \tag{19}$$

Since $\ell \in [0,1]$, we also have that $s' \in [0,1]$ and thus $||\nabla \tilde{f}||^2 \leq \max\limits_{t \in [0,1]} \left[G^2 \left(\frac{(1-\epsilon)t}{\rho} + \epsilon\right)^2 + (1-\epsilon)^2 (1-\frac{t}{\rho})^2 \right] \leq G^2_{\rho,\epsilon}$.

3. Finally, we recall that by assumption the loss function ℓ is $\beta-$smooth and the smoothing plus function s is $\frac{2}{\gamma}-$smooth. Then, for any pairs $m_1 = (\boldsymbol{\theta}_1, c_1)$ and $m_2 = (\boldsymbol{\theta}_2, c_2)$, for the first coordinate of $\nabla \tilde{f}$ we

obtain

$$\left|\left|\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_1;z)-c_1)}{\rho}+\epsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_1;z)-\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_2;z)-c_2)}{\rho}+\epsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)\right|\right|$$

$$\begin{aligned}=&\left|\left|\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_1;z)-c_1)}{\rho}+\epsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_1;z)-\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_1;z)-c_1)}{\rho}+\epsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)\right.\right.\\&\left.\left.+\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_1;z)-c_1)}{\rho}+\epsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)-\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_2;z)-c_2)}{\rho}+\epsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)\right|\right|\end{aligned}$$

$$\begin{aligned}\leq&\left(\frac{(1-\epsilon)\left|s'(\ell(\boldsymbol{\theta}_1;z)-c_1)\right|}{\rho}+\epsilon\right)\cdot||\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_1;z)-\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)||\\&+||\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)||\cdot\left|\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_1;z)-c_1)}{\rho}+\epsilon\right)-\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_2;z)-c_2)}{\rho}+\epsilon\right)\right|\end{aligned}\tag{20}$$

We know that:

1. Since $\ell\in[0,1]$, we have that $s'\in[0,1]$ and

$$\left(\frac{(1-\epsilon)\left|s'(\ell(\boldsymbol{\theta}_1;z)-c_1)\right|}{\rho}+\epsilon\right)\leq\left(\frac{(1-\epsilon)}{\rho}+\epsilon\right).\tag{21}$$

2. The loss function ℓ is $\beta-$smooth, i.e.

$$||\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_1;z)-\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)||\leq\beta||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||.\tag{22}$$

3. The loss function ℓ is convex and $G-$Lipschitz, thus

$$||\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)||\leq G.\tag{23}$$

4. The smoothed plus function s is $\frac{2}{\gamma}$-smooth, i.e.

$$||s'(\ell(\boldsymbol{\theta}_1;z)-c_1)-s'(\ell(\boldsymbol{\theta}_2;z)-c_2)||\leq\frac{2}{\gamma}||(\ell(\boldsymbol{\theta}_1;z)-c_1)-(\ell(\boldsymbol{\theta}_2;z)-c_2)||\tag{24}$$

By substituting 21-24 into Eq. 20 we get

$$\begin{aligned}&\left(\frac{(1-\epsilon)\left|s'(\ell(\boldsymbol{\theta}_1;z)-c_1)\right|}{\rho}+\epsilon\right)\cdot||\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_1;z)-\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)||\\&+||\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_2;z)||\cdot\left|\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_1;z)-c_1)}{\rho}+\epsilon\right)-\left(\frac{(1-\epsilon)s'(\ell(\boldsymbol{\theta}_2;z)-c_2)}{\rho}+\epsilon\right)\right|\end{aligned}$$

$$\leq\left(\frac{(1-\epsilon)}{\rho}+\epsilon\right)\beta||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||+G\frac{(1-\epsilon)}{\rho}\left|s'(\ell(\boldsymbol{\theta}_1;z)-c_1)-s'(\ell(\boldsymbol{\theta}_2;z)-c_2)\right|$$

$$\leq \left(\frac{(1-\epsilon)}{\rho}+\epsilon\right)\beta||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2|| + G\frac{(1-\epsilon)}{\rho}\frac{2}{\gamma}\big|(\ell(\boldsymbol{\theta}_1;z)-c_1)-(\ell(\boldsymbol{\theta}_2;z)-c_2)\big|$$

$$\leq \left(\frac{(1-\epsilon)}{\rho}+\epsilon\right)\beta||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2|| + G^2\frac{(1-\epsilon)}{\rho}\frac{2}{\gamma}||(\boldsymbol{\theta}_1,c_1)-(\boldsymbol{\theta}_2,c_2)|| \qquad \text{(by } G\text{−Lipschitzness of } \ell\text{)}$$

$$= \left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)||m_1-m_2||,$$

with $m_1=(\boldsymbol{\theta}_1,c_1)$ and $m_2=(\boldsymbol{\theta}_2,c_2)$.

4. From Definition 4.1 we have that for any fixed pair of $(\boldsymbol{\theta},c)$:

$$\begin{aligned}
&(\ell(\boldsymbol{\theta};z)-c)_+ \leq s(\ell(\boldsymbol{\theta};z)-c) \leq (\ell(\boldsymbol{\theta};z)-c)_+ + \gamma \\
\Rightarrow &(1-\epsilon)[c+\frac{1}{\rho}(\ell(\boldsymbol{\theta};z)-c)_+] + \epsilon\ell(\boldsymbol{\theta};z) \leq (1-\epsilon)[c+\frac{1}{\rho}s(\ell(\boldsymbol{\theta};z)-c)] + \epsilon\ell(\boldsymbol{\theta};z) \\
&\leq (1-\epsilon)[c+\frac{1}{\rho}(\ell(\boldsymbol{\theta};z)-c)_+] + \epsilon\ell(\boldsymbol{\theta};z) + \frac{(1-\epsilon)}{\rho}\gamma \\
\Rightarrow &f(\boldsymbol{\theta},c;z) \leq \tilde{f}(\boldsymbol{\theta},c;z) \leq f(\boldsymbol{\theta},c;z) + \frac{(1-\epsilon)}{\rho}\gamma
\end{aligned}$$

□

C.2 Convergence of Algorithm 1

For the convergence analysis of Algorithm 1 we leverage the standard results for FedAvg presented in Theorem 1 in [Wang et al., 2021]. We apply them to our setting and provide the proof of Lemma 5.6 below.

Proof of Lemma 5.6. We apply the results from Theorem 1 in [Wang et al., 2021] to our setting by substituting the properties of the loss function ℓ with those of the non-smooth and smoothed auxiliary functions, f and $\tilde{f}$ respectively, given by Lemma 5.5. In particular,

1. β is changed to the smoothness parameter of $\tilde{f}$, which is $\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta$;
2. we use σ^2 to represent the bounded variance in the case that each client uses a batch with size b instead of a single sample;
3. M is changed to $\sqrt{M^2+B^2}$ since we optimize over $v\in\Theta\times[0,B]$.

Considering also Lemma 5.5, property 4, we finally obtain

$$\mathbb{E}\left[\sum_{k\in\mathcal{K}}\sum_{i=1}^{n_k}\frac{f(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;z_i^k)}{n}\right] \leq \mathbb{E}\left[\sum_{k\in\mathcal{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;z_i^k)}{n}\right]$$

$$\leq \mathbb{E}\left[\sum_{k\in\mathcal{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\boldsymbol{\theta}_D^*, c_D^*; z_i^k)}{n}\right] + \frac{2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)(M^2+B^2)}{\tau T} + \frac{2\sigma\sqrt{M^2+B^2}}{\sqrt{|\mathcal{K}|b\tau T}}$$
$$+ 5\left(\frac{\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)\sigma^2(M^2+B^2)^2}{\tau b T^2}\right)^{\frac{1}{3}} + 19\left(\frac{\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)\mu^2(M^2+B^2)^2}{T^2}\right)^{\frac{1}{3}}$$

$$= \mathbb{E}\left[\sum_{k\in\mathcal{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\boldsymbol{\theta}_D^*, c_D^*; z_i^k)}{n}\right] + \frac{2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)(M^2+B^2)}{\tau T}$$
$$+ \frac{2\sigma\sqrt{M^2+B^2}}{\sqrt{|\mathcal{K}|b\tau T}} + \left(\frac{\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)(M^2+B^2)^2}{T^2}\right)^{\frac{1}{3}}\left(5(\frac{\sigma^2}{b\tau})^{\frac{1}{3}} + 19\mu^{\frac{2}{3}}\right)$$

$$\leq \mathbb{E}\left[\sum_{k\in\mathcal{K}}\sum_{i=1}^{n_k}\frac{f(\boldsymbol{\theta}_D^*, c_D^*; z_i^k)}{n}\right] + \frac{2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)(M^2+B^2)}{\tau T} + \frac{(1-\epsilon)\gamma}{\rho}$$
$$+ \frac{2\sigma\sqrt{M^2+B^2}}{\sqrt{|\mathcal{K}|b\tau T}} + \left(\frac{\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)(M^2+B^2)^2}{T^2}\right)^{\frac{1}{3}}\left(5(\frac{\sigma^2}{b\tau})^{\frac{1}{3}} + 19\mu^{\frac{2}{3}}\right).$$

Finally, the learning rate becomes

$$\eta = \min\left\{\frac{(M^2+B^2)^{\frac{1}{3}}}{\tau(\mu^2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)T)^{\frac{1}{3}}}, \frac{\sqrt{b|\mathcal{K}|}\sqrt{M^2+B^2}}{\sigma\sqrt{\tau T}}, \left(\frac{b(M^2+B^2)}{\sigma^2\tau^2\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)T}\right)^{\frac{1}{3}},\right.$$
$$\left.\frac{1}{4\left(\frac{(1-\epsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\right)}\right\}$$

□

C.3 Excess Risk Analysis of Algorithm 1

In order to derive an upper bound in excess risk $\mathcal{E}_r$, we use the results for stochastic gradient methods from Proposition 5.4. in [Hardt et al., 2016], which we also repeat using our notation in Lemma C.1 for convenience.

Lemma C.1 (Proposition 5.4. in [Hardt et al., 2016]) *Let assumptions 5.1 and 5.2 hold. Let also* $\overline{\boldsymbol{\theta}}_T = \sum_{t=1}^{T}\boldsymbol{\theta}_t$ *and* $\boldsymbol{\theta}_D^\star = \arg\min_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}; z_i)$. *Suppose we run* T *steps of SGD with a learning rate* $\eta = \frac{M\sqrt{n}}{G\sqrt{T(n+2T)}}$. *Then,*

$$\mathbb{E}\left[\mathop{\mathbb{E}}_{Z\sim p(Z)}[\ell(\overline{\boldsymbol{\theta}}_T; Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}_D^\star; z_i)\right] \leq \frac{1}{2}\left(\frac{M^2}{\eta T} + \frac{G^2\eta}{n}(2T+n)\right) \tag{25}$$

where the outer expectation is taken w.r.t. the internal randomness of the algorithm and the randomness of samples D.

Note that the selected step-size in Lemma C.1 satisfies $\eta \leq \frac{2}{\beta}$ (see Theorem 3.7 in [Hardt et al., 2016] for more details). This condition is required to be satisfied in our analysis as well. Next, we apply these results in our setting and provide the formal proof of Lemma 5.7.

Proof of Lemma 5.7 . For the excess risk analysis, we use the same substitutions as in Lemma 5.6. Thus, Eq. 25 for our setting becomes

$$\begin{aligned} & \mathbb{E}\Bigg[\mathop{\mathbb{E}}_{K\in\mathcal{K}} \Big[\mathop{\mathbb{E}}_{Z|k} [\tilde{f}(\overline{\boldsymbol{\theta}}_T, \bar{c}_T; Z|k)]\Big] - \tfrac{1}{n} \sum_{k\in\mathcal{K}} \sum_{i=1}^{n_k} \tilde{f}(\boldsymbol{\theta}_D^*, c_D^*; z_i^k)\Bigg] \\ = \; & \mathbb{E}\Bigg[\mathop{\mathbb{E}}_{Z}[\tilde{f}(\overline{\boldsymbol{\theta}}_T, \bar{c}_T; Z)]\Bigg] - \mathbb{E}\Bigg[\tfrac{1}{n} \sum_{i=1}^{n} \tilde{f}(\boldsymbol{\theta}_D^*, c_D^*; z_i)\Bigg] \\ \leq \; & \tfrac{1}{2}\Bigg(\tfrac{M^2+B^2}{\eta T} + \tfrac{G_{\rho,\epsilon}^2 \eta}{\big(\sum_{k\in\mathcal{K}} b_k\big) n} (2T+n)\Bigg), \end{aligned} \tag{26}$$

with the learning rate being

$$\eta = \frac{\sqrt{M^2+B^2}\sqrt{n\big(\sum_{k\in\mathcal{K}} b_k\big)}}{G_{\rho,\epsilon}\sqrt{T(n+2T)}}. \tag{27}$$

Based on the remark we made above about the learning rate in Lemma C.1 being at most $\frac{2}{\beta}$, we must ensure that the respective step size in Eq. 27 is at most $\frac{2\rho}{(1-\epsilon)(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\rho}$ as well.

We know that for any $v, u > 0$ we have that $2\max\{v,u\} \geq v+u \Rightarrow \frac{2}{v+u} \geq \frac{1}{\max\{v,u\}}$. Thus, given also that $\epsilon \in (0,1]$, we obtain

$$\frac{2\rho}{(1-\epsilon)(\beta+\frac{2}{\gamma}G^2)+\epsilon\beta\rho} \geq \frac{2\rho}{\beta+\frac{2}{\gamma}G^2+\epsilon\beta\rho} \geq \frac{\rho}{\max\{\beta(1+\epsilon\rho), \frac{2}{\gamma}G^2\}} \geq \min\left\{\frac{\rho}{\beta(1+\epsilon\rho)}, \frac{\rho\gamma}{2G^2}\right\}.$$

Thus, it is sufficient to ensure that (i) $\eta \leq \frac{\rho}{\beta(1+\epsilon\rho)}$, and (ii) $\eta \leq \frac{\rho\gamma}{2G^2}$.

We can satisfy the first case, $\eta \leq \frac{\rho}{\beta(1+\epsilon\rho)}$, by the choice of the rounds number T, that is

$$\eta = \sqrt{M^2+B^2}\frac{\sqrt{n\big(\sum_{k\in\mathcal{K}} b_k\big)}}{G_{\rho,\epsilon}\sqrt{T(n+2T)}} \leq \tfrac{\rho}{\beta(1+\epsilon\rho)}$$

$$\Longrightarrow n\big(\sum_{k\in\mathcal{K}} b_k\big)\big(M^2+B^2\big)\left(\tfrac{\beta(1+\epsilon\rho)}{\rho G_{\rho,\epsilon}}\right)^2 \leq T(n+2T).$$

For the case $\eta \leq \frac{\rho\gamma}{2G^2}$, since $\rho \in (0,1)$, $\epsilon \in (0,1]$ and

$$G_{\rho,\epsilon} \;\; = \max\left\{\sqrt{G^2\epsilon^2+(1-\epsilon)^2}, \sqrt{\tfrac{G^2(1-\epsilon+\epsilon\rho)^2+(1-\epsilon)^2(\rho-1)^2}{\rho^2}}\right\} \geq \max\left\{\epsilon G, \tfrac{(1-\epsilon+\epsilon\rho)G}{\rho}\right\} \geq \tfrac{(1-\epsilon+\epsilon\rho)G}{\rho}$$

we have that

$$\frac{\rho\gamma}{2G^2} \geq \frac{\rho^2\gamma}{2G^2} \geq \frac{(1-\epsilon+\epsilon\rho)^2\gamma}{2G_{\rho,\epsilon}^2}. \tag{28}$$

Thus, by setting $\gamma = \frac{2G^2_{\rho,\epsilon}}{(1-\epsilon+\epsilon\rho)^2}\eta$ we can satisfy condition (ii).

Finally, we yield the proposed bound by using the learning rate in Eq. 27 and the results in Lemma 5.5 for which Eq. 26 becomes

$$\begin{aligned}
&\mathbb{E}\big[\mathop{\mathbb{E}}_{Z}[f(\overline{\boldsymbol{\theta}}_T, \overline{c}_T; Z)]\big] \leq \mathbb{E}\big[\mathop{\mathbb{E}}_{Z}[\tilde{f}(\overline{\boldsymbol{\theta}}_T, \overline{c}_T; Z)]\big] \\
&\leq \mathbb{E}\big[\tfrac{1}{n}\sum_{i=1}^{n} \tilde{f}(\boldsymbol{\theta}^*_D, c^*_D; z_i)\big] + G_{\rho,\epsilon}\sqrt{\frac{(M^2+B^2)(\frac{2}{n}+\frac{1}{T})}{\big(\sum_{k\in\mathcal{K}} b_k\big)}} \\
&\leq \mathbb{E}\big[\tfrac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}^*_D, c^*_D; z_i)\big] + G_{\rho,\epsilon}\sqrt{\frac{(M^2+B^2)(\frac{2}{n}+\frac{1}{T})}{\big(\sum\limits_{k\in\mathcal{K}} b_k\big)}} + \tfrac{(1-\epsilon)\gamma}{\rho}
\end{aligned} \tag{29}$$

□

D Experimental Details

D.1 Experimental Setup

We preprocess ACS Employment as described in [Ding et al., 2021] and eICU akin to [Pollard et al., 2018]. 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 MNIST we used a setting akin to FashionMNIST splits in [Deng et al., 2020, Mohri et al., 2019]. For ACS Employment we consider two settings: (a) a setting with 51 clients, where each client represents a different geo-location and (b) the hardest data partitioning using a sensitive group, proposed in [Papadaki et al., 2022], where the data is split to 3 clients based on the race classes: {Black, White, Others} to make the comparison in Appendix D.3 fairer. For Celeb-A, we randomly assign the partitions across two clients using three different seeds. For the ACS Employment and MNIST datasets, we use a MLP with a single hidden layer with 512 neurons. For eICU we use logistic regression and for Celeb-A we use a ResNet-18. We select cross entropy to be the loss function in every training scenario.

We report results such as the worst performing group and average utility that directly relate to the notion of minimax fairness. We also present the risk disparity between best and worst groups, conditioned on group size ρ. We train FedSRCVaR using local batch size $b_k = \{32, 64, 128\}$ and $\epsilon \approx 0.0$ is set as $\epsilon = 0.01$, except for ACS Employment that we pick the best solution from $\epsilon = \{0.001, 0.005, 0.01, 0.05\}$. We use the same batch size options for AFL. BPF is trained using $\epsilon = \{0.001, 0.005, 0.01, 0.05\}$. FedAvg is trained using batches of sample size 128 and local epochs $E = \{3, 8, 15\}$. We train all approaches using learning rates $\eta = \{0.01, 0.001, 0.0001, 0.00001\}$, adversary/threshold learning rate η_{adv} = 0.001 (where relevant). We pick the combination with the best solution for each case. For the proposed approach, FedSRCVaR, 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 parameter $\epsilon = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ 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. If a fixed test set is provided by the authors of the dataset we use

that for testing. For the experiments, we use soft-ReLU as a smoothed plus function with $\gamma = 0.05$.

D.2 Implementation & Training Devices

The experiments were implemented in Python using PyTorch. We produce results for BPF using the original code available at `github.com/natalialmg/BlindParetoFairness`. The experiments were realised using 4× NVIDIA Tesla V100 GPUs.

D.3 Additional Empirical Results

We examine the cost in (minimax) group fairness when considering sensitive groups that are (potentially incorrectly) anticipated in the testing phase. This scenario reflects realistic situations where we might lack access to information regarding the future evolution of demographics, even if we are aware of the sensitive demographic groups during training time. Hence, we demonstrate how a model trained on known demographic groups will perform when faced with the most challenging, worst-case scenario within these groups.

Table D.1: Cross Entropy risks comparison of minimax Pareto federated group fairness with real demographics (FedMinMax), unknown demographics (FedSRCVaR, ours) and baseline (FedAvg) on ACS Employment dataset. Results are averaged over 3 runs.

Group		**FedMinMax**	**FedSRCVaR** ($\epsilon = 0.05$)		**FedAvg**
			$\rho = 0.1$	$\rho = 0.3$	(Baseline)
Worst	$\rho = 0.1$	1.593±0.23	0.713±0.07	0.724±0.06	1.768±0.04
group	$\rho = 0.3$	1.176±0.08	0.698±0.02	0.695±0.05	1.037±0.09

We compare against the optimal minimax group fair solution for known sensitive groups, generated using FedMinMax [Papadaki et al., 2022], and present the results in Table D.1. We leverage ACS Employment dataset and allocate data on 3 clients based on the races {Black, White, Others}, as in [Papadaki et al., 2022]. We observe that FedMinMax has significantly poor performance on the worst groups of the selected ρ sizes. On the other hand, FedSRCVaR outperforms FedMinMax in the worst group generated by all samples. We emphasize that our approach does not utilize existing groups. FedSRCVaR optimizes for the worst-case group that can be formulated from the training data, ensuring that *no (real) group will perform worse than that*. However, the worst-case group may not be easily described as one of the commonly defined demographic groups; this subset of bad-performing samples may be distributed across the predefined group categories. On the other hand, FedAVG focuses solely on (average) utility, which means it will perform poorly on the worst possible subgroup.

These results indicate that the price of optimizing for unknown demographics is lower than the cost of optimizing for wrong demographics, given by the groups-agnostic approach. Hence, we argue that FedSRCVaR is not only beneficial for settings where the sensitive groups are completely unknown, but also it is preferable when the known sensitive groups could potentially change in the future; or in the general case that we are not completely certain that the sensitive demographics remain the same during training and testing time.

For reference, we also provide the proportion of the various demographics in the predicted worst group that is generated by our approach for $\epsilon \approx 0$ and low ρ values; and of the actual group populations in Table D.2. We notice that the composition of the worst group is very close to the actual size of the sensitive groups, especially as ρ grows.

Table D.2: Worst group formation and actual group size on the test split on ACS Employment dataset. FedSRCVaR is evaluated for $\epsilon = 0.05$ and $\rho = \{0.1, 0.3\}$. We denote as {U, E} the labels {Unemployed, Employed}.

		White (%)		**Black** (%)		**Other** (%)	
$\boldsymbol{\epsilon}$	$\boldsymbol{\rho}$	U	E	U	E	U	E
0.05	0.1	48.6	14.9	4.3	1.2	24.1	6.9
	0.3	30.2	30.7	2.6	2.	17.8	16.7
Actual size		33.4	27.9	2.9	1.9	18.3	15.6