

Minimax Demographic Group Fairness in Federated Learning

Afroditi Papadaki Natalia Martinez Martin Bertran
 University College London Duke University Duke University

Guillermo Sapiro Miguel Rodrigues
 Duke University and Apple Inc. University College London

{a.papadaki.17, m.rodrigues}@ucl.ac.uk
 {natalia.martinez, martin.bertran, guillermo.sapiro}@duke.edu

Abstract

Federated learning is an increasingly popular paradigm that enables a large number of entities to collaboratively learn better models. In this work, we study minimax group fairness in federated learning scenarios where different participating entities may only have access to a subset of the population groups during the training phase. We formally analyze how our proposed group fairness objective differs from existing federated learning fairness criteria that impose similar performance across participants instead of demographic groups. We provide an optimization algorithm – FedMinMax – for solving the proposed problem that provably enjoys the performance guarantees of centralized learning algorithms. We experimentally compare the proposed approach against other state-of-the-art methods in terms of group fairness in various federated learning setups, showing that our approach exhibits competitive or superior performance.

1 Introduction

Machine learning models are being increasingly adopted to make decisions in a range of domains, such as finance, insurance, medical diagnosis, recruitment, and many more [2]. Therefore, we are often confronted with the need – sometimes imposed by regulatory bodies – to ensure that such machine learning models do not lead to decisions that discriminate individuals from a certain demographic group.

The development of machine learning models that are fair across different (demographic) groups has been well studied in traditional learning setups where there is a single entity responsible for learning a model based on a local dataset holding data from individuals of the various groups. However, there are settings where the data representing different demographic groups is spread across multiple entities rather than concentrated on a single entity/server. For example, consider a scenario where various hospitals wish to learn a diagnostic machine learning model that is fair (or performs reasonably well) across different demographic groups but each hospital may only contain training data from certain groups because – in view of its geo-location – it serves predominantly individuals of a given demographic [5]. This new setup along with the conventional centralized one are depicted in Figure 1.

These emerging scenarios however bring about various challenges. The first challenge relates to the fact that each individual entity may not be able to learn locally by itself a fair machine learning model because it may not hold (or hold little) data from certain demographic groups. The second challenge relates to that fact that each individual entity may also not be able to directly share their own data with other entities due to legal or regulatory challenges such as GDPR [4]. Therefore, the conventional machine learning fairness *ansatz* – relying on the fact that the learner has access to the overall data – does not generalize from the centralized data setup to the new distributed one.

It is possible to address these challenges by adopting federated learning (FL) approaches. These learning approaches enable multiple entities (or clients¹) coordinated by a central server to iteratively learn in a decentralized manner a single global model to carry out some task [23, 24]. The clients do not share data with one another or with the server; instead the clients only share focused updates with the server, the server then updates a global model, and distributes the updated model to the clients, with the process carried out over multiple rounds or iterations. This learning approach enables different clients with limited local training data to learn better machine learning models.

However, with the exception of some recent works such as [5, 48], which we will discuss later, federated learning is not typically used to learn models that exhibit performance guarantees for different demographic groups served by a client (i.e., *group fairness* guarantees); instead, it is primarily used to learn models that exhibit specific performance guarantees for each client involved in the federation (i.e., *client fairness* guarantees). Importantly, in view of the fact that a machine learning model that is *client fair* is not necessarily *group fair* (as we formally demonstrate in this work), it becomes crucial to understand

how to develop new federated learning techniques leading up to models that are also fair across different demographic groups.

This work develops a new federated learning algorithm that can be adopted by multiple entities coordinated by a single server to learn a *global minimax group fair* model. We show that our algorithm leads to the same (minimax) group fairness performance guarantees of centralized approaches such as [8, 32], which are exclusively applicable to settings where the data is concentrated in a single client. Interestingly, this

¹Clients are different user devices, organisations or even geo-distributed datacenters of a single company [21]. In this manuscript we use the terms participants, clients, and entities, interchangeably.

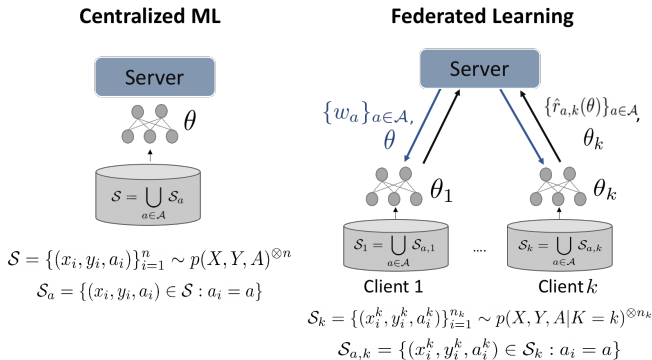


Figure 1: Centralized Learning vs. Federated Learning group fairness. *Left:* A single entity holds the dataset \mathcal{S} in a single server that is responsible for learning a model h parameterized by θ . *Right:* Multiple entities hold different datasets \mathcal{S}_k , sharing restricted information with a server that is responsible for learning a model h parameterized by θ , and the group importance weights $\mathbf{w} = \{w_a\}_{a \in \mathcal{A}}$. See also Section 3.

also applies to scenarios where certain clients do not hold any data from some of the demographic groups.

The rest of the paper is organized as follows: Section 2 overviews related work. Section 3 formulates our proposed distributed group fairness problem. Section 4 formally demonstrates that traditional federated learning approaches such as [6, 7, 27, 36] may not always solve group fairness. In Section 5 we propose a new federated learning algorithm to collaboratively learn models that are minimax group fair. Section 6 illustrates the performance of our approach in relation to other baselines. Finally, Section 7 draws various conclusions.

2 Related Work

Fairness in Machine Learning. The development of fair machine learning models in the standard *centralized learning setting* – where the learner has access to all the data – is underpinned by fairness criteria. One popular criterion is *individual fairness* [13] that dictates that the model is fair provided that people with similar characteristics/attributes are subject to similar model predictions/decisions. Another family of criteria – known as *group fairness* – requires the model to perform similarly on different demographic groups. Popular group fairness criteria include equality of odds, equality of opportunity [18], and demographic parity [30], that are usually imposed as a constraint within the learning problem. More recently, [32] introduced *minimax group fairness*; this criterion requires the model to optimize the prediction performance of the worst demographic group without unnecessarily impairing the performance of other demographic groups (also known as no-harm fairness) [8, 32]. In this work we leverage minimax group fairness criterion to learn a model that is (demographic) group fair across any groups included in the clients distribution in federated learning settings. However, the overall concepts here introduced can also be extended to other fairness criteria.

Fairness in Federated Learning. The development of fair machine learning models in *federated learning settings* has been building upon the group fairness literature. The majority of these works has concentrated predominantly on *client-fairness* which targets the development of algorithms leading to models that exhibit similar performance across different clients [27].

One such approach is agnostic federated learning (AFL) [36], whose aim is to learn a model that optimizes the performance of the worst performing client. Extensions of AFL [7, 41] improve its communication-efficiency by enabling clients to perform multiple local optimization steps. Another FL approach proposed in [27], uses an extra fairness constraint to flexibly control performance disparities across clients. Similarly, tilted empirical risk minimization [26] uses a hyperparameter called tilt to enable fairness or robustness by magnifying or suppressing the impact of individual client losses. FedMGDA+ [20] is an algorithm that combines minimax optimization coupled with Pareto efficiency [33] and gradient normalization to ensure fairness across users and robustness against malicious clients.

The works in [19, 37] enable fairness across clients with different hardware computational capabilities by allowing any participant to train a submodel of the original deep neural network (DNN) in order to contribute to the global model. The authors in [43] observe that unfairness across clients is caused by conflicting gradients that may significantly reduce the performance of some clients and therefore propose an algorithm for detecting and mitigating such conflicts. Finally, GIFAIR-FL [45] uses a regularization

term to penalize the spread in the aggregated loss to enforce uniform performance across the participating entities. Our work naturally departs from these fairness federated learning approaches since, as we prove in Section 4, client-fairness ensures fairness across all demographic groups included across clients datasets only under some special conditions.

Another fairness concept in federated learning is collaborative fairness [15, 47, 31, 38], which proposes each client’s performance compensation to correspond to its contribution on the utility task of the global model. Larger rewards to high-contributing clients motivate their participation in the federation while lower rewards prevent free-riders [31]. However, such approaches might further penalize clients that have access to the worst performing demographic groups resulting to a even more unfair global model.

There are some recent complementary works that consider group fairness within client distributions. Group distributional robust optimization (G-DRFA) [48], aims to optimize for the worst performing group by learning a weighting coefficient for each local group, even if there are shared groups across clients. In our work, we combine the statistics received from the clients sharing the same groups to learn a global model, since, as we experimentally show in Section 6, considering duplicates of the same group might lead to worst generalization in some FL scenarios. FCFL [5] focuses on improving the worst performing client while ensuring a level of local group fairness defined by each client, by employing gradient-based constrained multi-objective optimization. Our primary goal is to learn a model solving (demographic) group fairness across any groups included in the clients distribution, independently of the groups representation in a particular client.

Finally, some recent approaches study the effects of (demographic) group fairness in FL using metrics such as demographic parity and/or equality in opportunity [11, 5, 42, 46, 14, 10, 3]. Compared to these methods, our approach can support scenarios with multiple group attributes and targets without any modifications on the optimization procedure. Also, even though comparing different fairness metrics is out of the scope of this work,² the aforementioned methods enforce some type of zero risk disparity across groups³ and thus degrade the performance of the good performing groups. In this work, we consider minimax group fairness criterion [32, 8], and due to its no-unnecessary harm property, we do not disadvantage any demographic groups except if absolutely necessary, making it suitable for applications such as healthcare and finance. Our formulation is complemented by theoretical results connecting minimax client and minimax group fairness and by proposing a provably convergent optimization algorithm.

Robustness in Federated Learning. Works dealing with robustness to distributional shifts in user data, such as [22, 40], also relate to group fairness. One work that closely relates to group fairness is FedRobust [40], that aims to learn a model for the worst case affine shift, by assuming that a client’s data distribution is an affine transformation of a global one. However, it requires each client to have enough data to estimate the local worst case shift else the global model performance on the worst group hinders [28].

Our Contributions. To recap, our core contributions compared to the literature are:

- We formulate minimax group fairness in federated learning settings where some clients might only have access to a subset of the demographic groups during the training phase.

²There are various studies discussing the effects of different fairness metrics. See for example [16].

³The risk considered in the fairness constraints is different across fairness definitions.

- We formally show under what conditions minimax group fairness is equivalent to minimax client fairness so that optimizing for any of the two notions results into a model that is both group and client fair.
- We propose a provably convergent optimization algorithm to collaboratively learn a minimax fair model across any demographic groups included in the federation, that allows clients to have high, low or no representation of a particular group. We show that our federated learning algorithm leads to a global model that is equivalent to a model yielded by a centralized learning algorithm.

3 Problem Formulation

3.1 Group Fairness in Centralized Machine Learning

We first describe the standard minimax group fairness problem in a centralized machine learning setting [8, 32], where there is a single entity/server holding all relevant data and responsible for learning a group fair model (see Figure 1). We concentrate on classification tasks, though our approach also applies to other learning tasks such as regression. Let the triplet of random variables $(X, Y, A) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{A}$ represent input features, target, and demographic groups. Let also $p(X, Y, A) = p(A) \cdot p(X, Y|A)$ represent the joint distribution of these random variables where $p(A)$ represents the prior distribution of the different demographic groups and $p(X, Y|A)$ their data conditional distribution.

Let $\ell : \Delta^{|\mathcal{Y}|-1} \times \Delta^{|\mathcal{Y}|-1} \rightarrow \mathbb{R}_+$ be a loss function where Δ represents the probability simplex. We now consider that the entity will learn an hypothesis h drawn from an hypothesis class $\mathcal{H} = \{h : \mathcal{X} \rightarrow \Delta^{|\mathcal{Y}|-1}\}$, that solves the optimization problem given by

$$\min_{h \in \mathcal{H}} \max_{a \in \mathcal{A}} r_a(h), r_a(h) = \mathbb{E}_{(X, Y) \sim p(X, Y|A=a)} [\ell(h(X), Y) | A = a]. \quad (1)$$

Note that this problem involves the minimization of the expected risk of the worst performing demographic group.

Importantly, under the assumption that the loss is a convex function w.r.t the hypothesis⁴ and the hypothesis class is a convex set, solving the minimax objective in Eq. 1 is equivalent to solving

$$\min_{h \in \mathcal{H}} \max_{a \in \mathcal{A}} r_a(h) \geq \min_{h \in \mathcal{H}} \max_{\mu \in \Delta_{\geq \epsilon}^{|\mathcal{A}|-1}} \sum_{a \in \mathcal{A}} \mu_a r_a(h) \quad (2)$$

where $\Delta_{\geq \epsilon}^{|\mathcal{A}|-1}$ represent the vectors in the simplex with all of their components larger than ϵ . Note that if $\epsilon = 0$ the inequality in Eq. 2 becomes an equality, however, allowing zero value coefficients may lead to models that are weakly, but not strictly, Pareto optimal [17, 35].

The minimax objective over the linear combination of the sensitive groups can be achieved by alternating between projected gradient ascent or multiplicative weight updates to optimize the weights given the model, and stochastic gradient descent to optimize the model given the weighting coefficients [1, 8, 32].

⁴This is true for the most common functions in machine learning settings such as Brier score and cross entropy.

3.2 Group Fairness in Federated Learning

We now describe our proposed group fairness federated learning problem; this problem differs from the previous one because the data is now distributed across multiple clients but each client (or the server) do not have direct access to the data held by other clients. See also Figure 1.

In this setting, we incorporate a categorical variable $K \in \mathcal{K}$ to our data tuple (X, Y, A, K) to indicate the clients participating in the federation. The joint distribution of these variables is $p(X, Y, A, K) = p(K) \cdot p(A|K) \cdot p(X, Y|A, K)$, where $p(K)$ represents a prior distribution over clients – which in practice is the fraction of samples that are acquired by client K relative to the total number of data samples –, $p(A|K)$ represents the distribution of the groups conditioned on the client, and $p(X, Y|A, K)$ represents the distribution of the distribution of the input and target variables conditioned on the group and client. We assume that the group-conditional distribution is the same across clients, meaning $p(X, Y|A, K) = p(X, Y|A)$. Note, however, that our model explicitly allows for the distribution of the demographic groups to depend on the client (via $p(A|K)$), accommodating for the fact that certain clients may have a higher (or lower) representation of certain demographic groups over others.

We now aim to learn a model $h \in \mathcal{H}$ that solves the minimax fairness problem as presented in Eq. 1, but considering that the group loss estimates are split into $|\mathcal{K}|$ estimators associated with each client. We therefore re-express the linear weighted formulation of Eq. 2 using importance weights, allowing to incorporate the role of the different clients, as follows:

$$\begin{aligned} \min_{h \in \mathcal{H}} \max_{\mu \in \Delta_{\geq \epsilon}^{|\mathcal{A}|-1}} \sum_{a \in \mathcal{A}} \mu_a r_a(h) &= \min_{h \in \mathcal{H}} \max_{\mu \in \Delta_{\geq \epsilon}^{|\mathcal{A}|-1}} \sum_{a \in \mathcal{A}} p(A = a) w_a r_a(h) \\ &= \min_{h \in \mathcal{H}} \max_{\mu \in \Delta_{\geq \epsilon}^{|\mathcal{A}|-1}} \sum_{k \in \mathcal{K}} p(K = k) \sum_{a \in \mathcal{A}} p(A = a | K = k) w_a r_a(h) \quad (3) \\ &= \min_{h \in \mathcal{H}} \max_{\mu \in \Delta_{\geq \epsilon}^{|\mathcal{A}|-1}} \sum_{k \in \mathcal{K}} p(K = k) r_k(h, \mathbf{w}), \end{aligned}$$

where $r_k(h, \mathbf{w}) = \sum_{a \in \mathcal{A}} p(A = a | K = k) w_a r_a(h)$ is the expected client risk and $w_a = \mu_a / p(A = a)$ denotes the importance weight for a particular demographic group.

There is an immediate non-trivial challenge that arises within this proposed federated learning setting in relation to the centralized one described earlier: we need to devise an algorithm that solves the objective in Eq. 3 under the constraint that the different clients cannot share their local data with the server or with one another, but – in line with conventional federated learning settings [7, 27, 34, 36]– only local model updates of a global model (or other quantities such as local risks) are shared with the server. This will be addressed later in this paper by the proposed federated optimization.

4 Client Fairness vs. Group Fairness in Federated Learning

Before proposing a federated learning algorithm to solve our proposed group fairness problem, we first reflect whether a model that solves the more widely used client fairness objective in federated learning settings given by [36]

$$\min_{h \in \mathcal{H}} \max_{k \in \mathcal{K}} r_k(h) = \min_{h \in \mathcal{H}} \max_{\lambda \in \Delta^{|\mathcal{K}|-1}} \mathbb{E}_{\mathcal{D}_\lambda} [\ell(h(X), Y)], \quad (4)$$

where $\mathcal{D}_\lambda = \sum_{k=1}^{|\mathcal{K}|} \lambda_k p(X, Y|K = k)$ denotes a joint data distribution over the clients and $\lambda = \{\lambda_k\}_{k \in \mathcal{K}}$ is the vector consisting of client weighting coefficients, also solves our proposed minimax group fairness objective given by

$$\min_{h \in \mathcal{H}} \max_{a \in \mathcal{A}} r_a(h) = \min_{h \in \mathcal{H}} \max_{\mu \in \Delta^{|\mathcal{A}|-1}} \mathbb{E}_{\mathcal{D}_\mu} [\ell(h(X), Y)], \quad (5)$$

where $\mathcal{D}_\mu = \sum_{a=1}^{|\mathcal{A}|} \mu_a p(X, Y|A = a)$ denotes a joint data distribution over sensitive groups and $\mu = \{\mu_a\}_{a \in \mathcal{A}}$ is the vector of the group weights.

The following lemma illustrates that a model that is minimax fair with respect to the clients is equivalent to a relaxed minimax fair model with respect to the (demographic) groups.

Lemma 1 *Let $\mathbf{P}_\mathcal{A}$ denote a matrix whose entry in row a and column k is $p(A = a|K = k)$ (i.e., the prior of group a in client k). Then, given a solution to the minimax problem across clients*

$$h^*, \lambda^* \in \arg \min_{h \in \mathcal{H}} \max_{\lambda \in \Delta^{|\mathcal{K}|-1}} \mathbb{E}_{\mathcal{D}_\lambda} [\ell(h(X), Y)], \quad (6)$$

$\exists \mu^* = \mathbf{P}_\mathcal{A} \lambda^*$ that is solution to the following constrained minimax problem across sensitive groups:

$$h^*, \mu^* \in \arg \min_{h \in \mathcal{H}} \max_{\mu \in \mathbf{P}_\mathcal{A} \Delta^{|\mathcal{K}|-1}} \mathbb{E}_{\mathcal{D}_\mu} [\ell(h(X), Y)], \quad (7)$$

where the weighting vector μ is constrained to belong to the simplex subset defined by $\mathbf{P}_\mathcal{A} \Delta^{|\mathcal{K}|-1} \subseteq \Delta^{|\mathcal{A}|-1}$. In particular, if the set $\Gamma = \{\mu' \in \mathbf{P}_\mathcal{A} \Delta^{|\mathcal{K}|-1} : \mu' \in \arg \min_{h \in \mathcal{H}} \max_{\mu \in \Delta^{|\mathcal{A}|-1}} \mathbb{E}_{\mathcal{D}_\mu} [\ell(h(X), Y)]\} \neq \emptyset$, then $\mu^* \in \Gamma$, and the minimax fairness solution across clients is also a minimax fairness solution across demographic groups.

Lemma 1 proves that being minimax with respect to the clients is equivalent to finding the group minimax model constraining the weighting vectors μ to be inside the simplex subset $\mathbf{P}_\mathcal{A} \Delta^{|\mathcal{K}|-1}$. Therefore, if this set already contains a group minimax weighting vector, then the group minimax model is equivalent to client minimax model. Another way to interpret this result is that being minimax with respect to the clients is the same as being minimax for any group assignment \mathcal{A} such that linear combinations of the groups distributions are able to generate all clients distributions, and there is a group minimax weighting vector in $\mathbf{P}_\mathcal{A} \Delta^{|\mathcal{K}|-1}$.

Being minimax at the client and group level relies on $\mathbf{P}_\mathcal{A} \Delta^{|\mathcal{K}|-1}$ containing the minimax weighting vector. In particular, if for each sensitive group there is a client comprised entirely of this group ($\mathbf{P}_\mathcal{A}$ contains a identity block), then $\mathbf{P}_\mathcal{A} \Delta^{|\mathcal{K}|-1} = \Delta^{|\mathcal{A}|-1}$ and group and client level fairness are guaranteed to be fully compatible. Another trivial example is when at least one of the client's group priors is equal to a group minimax weighting vector. This result also suggests that client level fairness may also differ from group level fairness. This motivates us to develop a new federated learning algorithm to guarantee group fairness that – where the conditions of the lemma hold – also results in client fairness. We experimentally validate the insights deriving from Lemma 1 in Section 6. The proof for Lemma 1 is provided in the supplementary material, Appendix A.

5 MiniMax Group Fairness Federating Learning Algorithm

We now propose an optimization algorithm – Federated Minimax (FedMinMax) – to solve the group fairness problem in Eq. 3.

We let each client k have access to a dataset $\mathcal{S}_k = \{(x_i^k, y_i^k, a_i^k); i = 1, \dots, n_k\}$ containing various data points drawn i.i.d according to $p(X, Y, A|K = k)$. We also define three additional sets: (a) $\mathcal{S}_{a,k} = \{(x_i^k, y_i^k, a_i^k) \in \mathcal{S}_k : a_i = a\}$ is a set containing all data examples associated with group a in client k ; (b) $\mathcal{S}_a = \bigcup_{k \in \mathcal{K}} \mathcal{S}_{k,a}$ is the set containing all data examples associated with group a across the various clients; and (c) $\mathcal{S} = \bigcup_{k \in \mathcal{K}} \mathcal{S}_k = \bigcup_{a \in \mathcal{A}} \mathcal{S}_a = \bigcup_{k \in \mathcal{K}} \bigcup_{a \in \mathcal{A}} \mathcal{S}_{a,k}$ is containing all data examples across groups and across clients. Note again that – in view of our modelling assumptions – it is possible that $\mathcal{S}_{a,k}$ can be empty for some k and some a implying that such a client does not have data realizations for such group.

We will also let the model h be parameterized via a vector of parameters $\boldsymbol{\theta} \in \Theta$, i.e., $h(\cdot) = h(\cdot; \boldsymbol{\theta})$.⁵ Then, one can approximate the relevant statistical risks using empirical risks as

$$\hat{r}_k(\boldsymbol{\theta}, \mathbf{w}) = \sum_{a \in \mathcal{A}} \frac{n_{a,k}}{n_k} \hat{w}_a \hat{r}_{a,k}(\boldsymbol{\theta}), \quad \hat{r}_a(\boldsymbol{\theta}) = \sum_{k \in \mathcal{K}} \frac{n_{a,k}}{n_a} \hat{r}_{a,k}(\boldsymbol{\theta}), \quad (8)$$

where $\hat{r}_{a,k}(\boldsymbol{\theta}) = \frac{1}{n_{a,k}} \sum_{(x,y) \in \mathcal{S}_{a,k}} \ell(h(x; \boldsymbol{\theta}), y)$, $\hat{w}_a = \mu_a / (n_a/n)$, $n_k = |\mathcal{S}_k|$, $n_a = |\mathcal{S}_a|$, $n_{a,k} = |\mathcal{S}_{a,k}|$, and $n = |\mathcal{S}|$. Note that $\hat{r}_k(\boldsymbol{\theta}, \mathbf{w})$ is an estimate of $r_k(\boldsymbol{\theta}, \mathbf{w})$, $\hat{r}_a(\boldsymbol{\theta})$ is an estimate of $r_a(\boldsymbol{\theta})$, and $\hat{r}_{a,k}(\boldsymbol{\theta})$ is an estimate of $r_{a,k}(\boldsymbol{\theta}) = \mathbb{E}_{(X,Y) \sim p(X,Y|A=a,K=k)}[\ell(h(X), Y)|A = a, K = k]$.

We consider the importance weighted empirical risk \hat{r}_k since the clients do not have access to the data distribution but instead to a dataset with finite samples. Therefore, the clients in coordination with the central server attempt to solve the optimization problem given by:

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\mu} \in \Delta_{\geq \epsilon}^{|\mathcal{A}|-1}} \hat{r}_a(\boldsymbol{\theta}) := \sum_{a \in \mathcal{A}} \mu_a \hat{r}_a(\boldsymbol{\theta}) \equiv \min_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\mu} \in \Delta_{\geq \epsilon}^{|\mathcal{A}|-1}} \sum_{k \in \mathcal{K}} \frac{n_k}{n} \hat{r}_k(\boldsymbol{\theta}, \mathbf{w}). \quad (9)$$

The objective in Eq. 9 can be interpreted as a zero-sum game between two players: the learner aims to minimize the objective by optimizing the model parameters $\boldsymbol{\theta}$ and the adversary seeks to maximize the objective by optimizing the weighting coefficients $\boldsymbol{\mu}$.

We use a non-stochastic variant of the stochastic-AFL algorithm introduced in [36]. Our version, provided in Algorithm 1, assumes that all clients are available to participate in each communication round t . In each round t , the clients receive the latest model parameters $\boldsymbol{\theta}^{t-1}$, the clients then perform one gradient descent step using all their available data, and the clients then share the updated model parameters along with certain empirical risks with the server. The server (learner) then performs a weighted average of the client model parameters $\boldsymbol{\theta}^t = \sum_{k \in \mathcal{K}} \frac{n_k}{n} \boldsymbol{\theta}_k^t$. The server also updates the weighting coefficient using a projected gradient ascent step in order to guarantee that the weighting coefficient updates are consistent with the constraints. We use the Euclidean algorithm proposed in [12] in order to implement the projection operation ($\Pi_{\Delta^{|\mathcal{A}|-1}}(\cdot)$).

We can show that our proposed algorithm can exhibit convergence guarantees.

⁵This vector of parameters could for example correspond to the set of weights / biases in a neural network.

Algorithm 1 FEDERATED MINIMAX (FEDMINMAX)

Input: \mathcal{K} : Set of clients, T : total number of communication rounds, η_θ : model learning rate, η_μ : global adversary learning rate, $\mathcal{S}_{a,k}$: set of examples for group a in client k , $\forall a \in \mathcal{A}$ and $\forall k \in \mathcal{K}$.

1: Server **initializes** $\mu^0 \leftarrow \rho = \{|\mathcal{S}_a|/|\mathcal{S}|\}_{a \in \mathcal{A}}$ and θ^0 randomly.
2: **for** $t = 1$ **to** T **do**
3: Server **computes** $w^{t-1} \leftarrow \mu^{t-1}/\rho$
4: Server **broadcasts** θ^{t-1}, w^{t-1}
5: **for** each client $k \in \mathcal{K}$ **in parallel do**
6: $\theta_k^t \leftarrow \theta^{t-1} - \eta_\theta \nabla_{\theta} \hat{r}_k(\theta^{t-1}, w^{t-1})$
7: Client- k **obtains** and **sends** $\{\hat{r}_{a,k}(\theta^{t-1})\}_{a \in \mathcal{A}}$ and θ_k^t to server
8: **end for**
9: Server **computes:** $\theta^t \leftarrow \sum_{k \in \mathcal{K}} \frac{n_k}{n} \theta_k^t$
10: Server **updates:** $\mu^t \leftarrow \prod_{\Delta|\mathcal{A}|-1} (\mu^{t-1} + \eta_\mu \nabla_{\mu} \langle \mu^{t-1}, \hat{r}_a(\theta^{t-1}) \rangle)$
11: **end for**
Outputs: $\frac{1}{T} \sum_{t=1}^T \theta^t$

Lemma 2 Consider our federated learning setting (Figure 1, right) where each entity k has access to a local dataset $\mathcal{S}_k = \bigcup_{a \in \mathcal{A}} \mathcal{S}_{a,k}$, and a centralized machine learning setting (Figure 1, left) where there is a single entity that has access to a single dataset $\mathcal{S} = \bigcup_{k \in \mathcal{K}} \mathcal{S}_k = \bigcup_{k \in \mathcal{K}} \bigcup_{a \in \mathcal{A}} \mathcal{S}_{a,k}$ (i.e., this single entity in the centralized setting has access to the data of the various clients in the distributed setting). Then, Algorithm 1 (federated) and Algorithm 2 (non-federated, in supplementary material, Appendix B) lead to the same global model provided that learning rates and model initialization are identical.

The proof for Lemma 2 is provided in Appendix A. This lemma shows that our federated learning algorithm inherits any convergence guarantees of existing centralized machine learning algorithms. In particular, assuming that one can model the single gradient descent step using a δ -approximate Bayesian Oracle [1], we can show that a centralized algorithm converges and hence our FedMinMax one converges too (under mild conditions on the loss function, hypothesis class, and learning rates). See Theorem 7 in [1].

6 Experimental Results

In this section we empirically showcase the applicability and competitive performance of the proposed federated learning algorithm. We apply FedMinMax to diverse federated learning scenarios by utilizing common benchmark datasets with multiple targets and sensitive groups. In particular, we perform experiments on the following datasets:

- **Synthetic.** We generated a synthetic dataset for binary classification involving two sensitive groups (i.e., $|\mathcal{A}| = 2$). Let $\mathcal{N}(\mu, \sigma^2)$ be the normal distribution with μ being the mean and σ^2 being the variance, and $Ber(p)$ Bernoulli distribution with probability p . The data were generated assuming the group variable $A \sim Ber(\frac{1}{2})$, the input features variable $X \sim \mathcal{N}(0, 1)$ and the target variable $Y|X, A = a \sim Ber(h_a^*)$, where $h_a^* = u_a^l \mathbb{1}[x \leq 0] + u_a^h \mathbb{1}[x > 0]$ is the optimal hypothesis for group $A = a$. We select $\{u_0^h, u_1^h, u_0^l, u_1^l\} = \{0.6, 0.9, 0.3, 0.1\}$. As illustrated in Figure 2, left side, the optimal hypothesis h is equal to the optimal model for group $A = 0$.
- **Adult [29].** Adult is a binary classification dataset consisting of 32,561 entries for predicting yearly

income based on twelve input features such as age, race, education and marital status. We consider four sensitive groups (i.e., $|\mathcal{A}| = 4$) created by combining the gender labels and the yearly income as follows: {Male w/ income > 50K, Male w/ income <= 50K, Female w/ income > 50K, Female w/ income <= 50K}.

- **FashionMNIST [44]**. FashionMNIST is a grayscale image dataset which includes 60,000 training images and 10,000 testing images. The images consist of 28×28 pixels and are classified into 10 clothing categories. In our experiments we consider each of the target categories to be a sensitive group too, (i.e., $|\mathcal{A}| = 10$).
- **CIFAR-10 [25]**. CIFAR-10 is a collection of 60,000 colour images of 32×32 pixels. Each image contains one out of 10 object classes. There are 50,000 training images and 10,000 test images. We use all ten target categories, which we assign both as targets and sensitive groups (i.e., $|\mathcal{A}| = 10$).
- **ACS Employment [9]**. ACS Employment is a recent dataset constructed using ACS PUMS data for predicting whether an individual is employed or not. For our experiments we use the 2018 1-Year data for all the US states and Puerto Rico. We combine race and utility labels to generate the following sensitive groups: : {Employed White, Employed Black, Employed Other, Unemployed White, Unemployed Black, Unemployed Other} (i.e., $|\mathcal{A}| = 6$). We also conduct experiments where the sensitive class is race using the original 9 labels that we report in supplementary material, Appendix D.

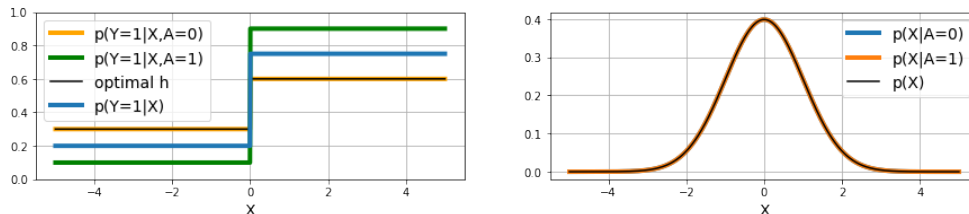


Figure 2: Illustration of the optimal hypothesis h and the conditional distributions $p(Y|X)$ and $p(X|A)$ for the generated synthetic dataset. (Left:) The worst group is $A = 0$ and the minimax optimal hypothesis h (black line) is equal to the optimal model for the worst group (orange line). (Right:) The distributions $p(X)$, and conditional distributions $p(X|A = 0)$ and $p(X|A = 1)$ are overlapping.

We also examine three federated learning settings, that we categorize based on the sensitive group allocation on clients as follows:

1. *Equal access to Sensitive Groups (ESG)*, where every client has access to all sensitive groups but does not have enough data to train a model individually. Each client in the federation has access to the same amount of the sensitive classes (i.e., $n_i = n_j \forall i, j \in \mathcal{K}, i \neq j$ and $n_{a,i} = n_{a,j} \forall i, j \in \mathcal{K}, a \in \mathcal{A}, i \neq j$). Here we examine a case where group and client fairness are not equivalent.
2. *Partial access to Sensitive Groups (PSG)*, where each participant has access to a subset of the available groups memberships. In particular, the data distribution is unbalanced across participants since the size of local datasets differs (i.e., $n_i \neq n_j \forall i, j \in \mathcal{K}, i \neq j$). Akin to ESG, this is a scenario where group and client fairness are incompatible. We use this scenario to compare the performances when there is low or no local representation of particular groups.

3. *Access to a Single Sensitive Group (SSG)*, where each client holds data from one sensitive group, for showcasing the group and client fairness objectives equivalence derived from Lemma 1. Similarly to PSG setting, the size of the local dataset varies across clients.

Note that ESG is an i.i.d. data scenario while PSG and SSG are non-i.i.d. data settings. Also note that each client’s data is unique, meaning that there are no duplicated examples across clients. In all experiments we consider a federation consisting of 40 clients and a single server that orchestrates the training procedure. We benchmark our approach against AFL [36], q -FedAvg [27], TERM [26] and FedAvg [34]. Further, as a baseline, we also run FedMinMax with one client (akin to centralized ML), that we denote *Centralized Minmax Baseline*, to confirm Lemma 2. We do not compare to baselines that explicitly employ a different fairness metric (e.g., demographic parity) since this not the focus of this work. For all the datasets, we compute the means and standard deviations of the accuracies and risks over three runs. We assume that every client is available to participate at each communication round for every method to make the comparison more fair. More details about model architectures and experiments are provided in Appendix C.

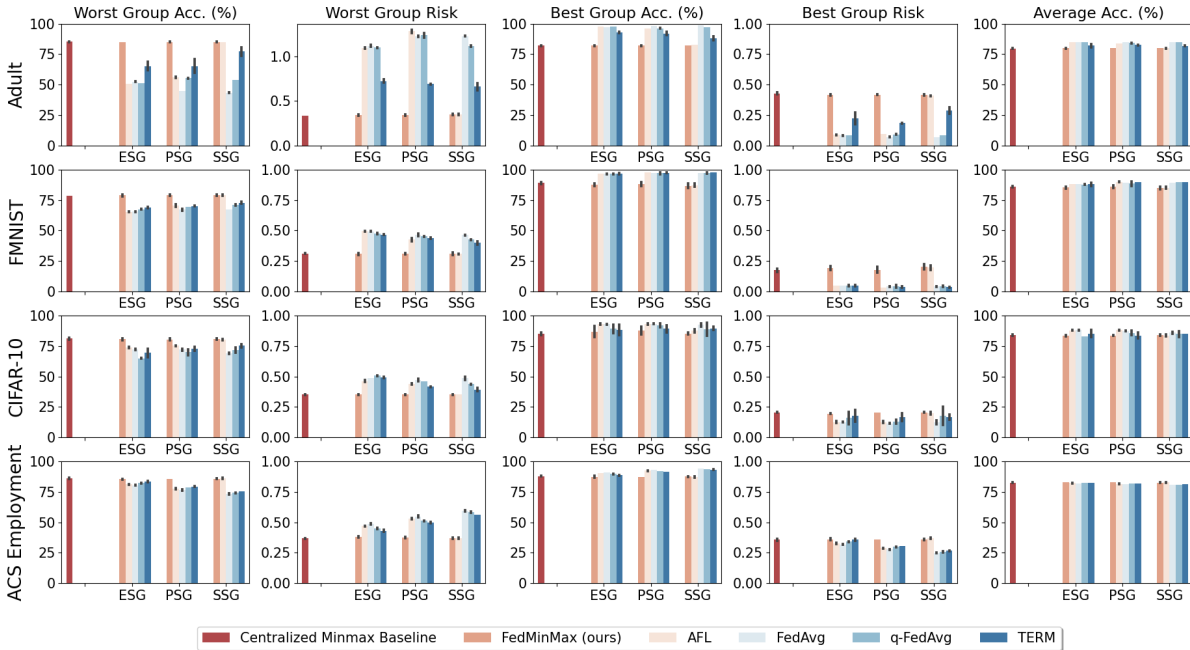


Figure 3: Comparison of the worst group, best group, and average risks and accuracies across three runs for AFL, FedAvg, q -FedAvg, TERM, FedMinMax and Centralized Minmax Baseline, across the different federated learning scenarios. Each bar reports the mean and standard deviation of the respective metric on the testing set. The numerical values, showing the advantages of the proposed framework, are provided in Tables D.2, D.4, D.5, D.7 and D.9, in the supplementary material.

We begin by investigating the worst group, the best group and the average utility performance for the Adult, FashionMNIST, CIFAR-10 and ACS Employment datasets in Figure 3. We present the mean and standard deviation of the accuracies and risks on the test dataset. FedMinMax enjoys a similar accuracy to the Centralized Minimax Baseline in all settings, as proved in Lemma 2. AFL is similar to FedMinMax and Centralized Minmax Baseline only in SSG, where group fairness is implied by client fairness, in line

with Lemma 1. FedAvg has similar best accuracy across federated settings, however the accuracy of the worst group decreases as the local data becomes more heterogeneous (i.e., in PSG and SSG). In many datasets, q -FedAvg and TERM have superior performance on the worst group compared to AFL and FedAvg in PSG and ESG, but do not to achieve minimax group fairness on any of the FL settings. Note that FedMinMax has the best worst group performance in all settings as expected.

For the numerical values, illustrating the efficiency of the proposed approach for every setting and dataset, see Tables D.2, D.4, D.5, D.7 and D.9, in the supplementary material.

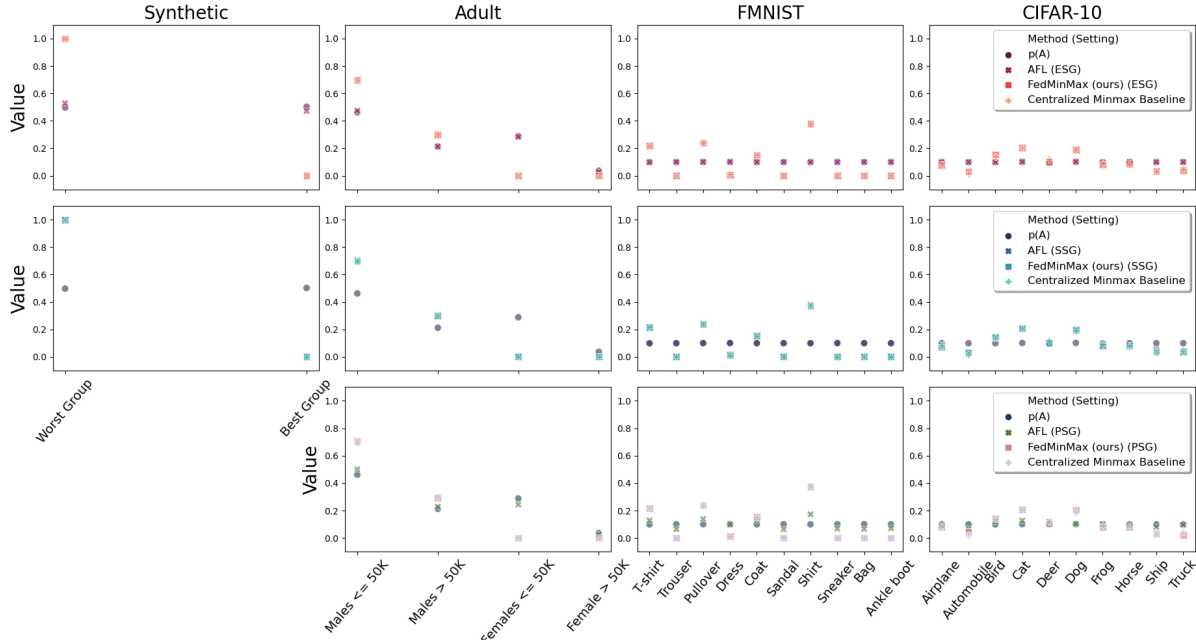


Figure 4: Sensitive group weighting coefficients for every minimax approach considered across different datasets calculated during the training time. We also provide the prior group distribution $p(A)$. Note that the weighting coefficients were produced based on the group risks on the training dataset and might not necessarily correspond to the group risks on the test set.

Next we show the final group weighting coefficients for the minimax approaches AFL, FedMinMax, and Centralized Minmax Baseline in Figure 4. Note that PSG scenario is valid only for datasets where $|\mathcal{A}| > 2$, else its equivalent to SSG setting.

The proposed approach yields similar group weights across all settings. FedMinMax also achieves the same weighting coefficients to Centralized Minmax Baseline, akin to Lemma 2. AFL produces weights similar to the group priors in ESG that move towards the minimax weighting coefficients the more we increase the heterogeneity w.r.t. the sensitive groups. AFL achieves the similar weights to FedMinMax and Centralized Minmax Baseline only in SSG scenario where each participant has access to exactly one group, following Lemma 1. Note that the group weighting coefficients are updated based on the risks calculated on the training set and might not generalize to the testing set for every dataset. We provide a complete description of the weighting coefficients for each approach in Tables D.1, D.3, D.6, D.8 and D.10, in the supplementary material.

Finally, we illustrate the efficiency of considering global demographics across entities instead of multiple

local ones, as in [48]. For these experiments, we re-purpose our algorithm – we call the adjusted version LocalFedMinMax – so that the adversary proposes a weighting coefficient for each group located in a client (i.e., $\boldsymbol{\mu} = \{\{\mu_{a,k}\}_{a \in \mathcal{A}}\}_{k \in \mathcal{K}}$). Recall that the adversary in our proposed algorithm uses a single weighting coefficient for every common demographic group (i.e., $\boldsymbol{\mu} = \{\mu_a\}_{a \in \mathcal{A}}$). We provide the detailed description of LocalFedMinMax in Algorithm 3, Appendix E.

In Table 1 we report results for both approaches on two federations consisting of 10 and 40 participants, respectively. LocalFedMinMax and FedMinMax offer similar improvement on the worst group on SSG regardless the number of clients. We also notice a similar behavior in the smaller federated network for the ESG scenario. In the remaining settings, LocalFedMinMax, leads to a worst performance as the amount of client increases and the number of data for each group per client reduces. On the other hand, FedMinMax is not effected by the local group representation since it aggregates the statistics received by each client and updates the weights for (global) demographics, leading up to a better generalization performance.

Table 1: Comparison of the worst group risk achieved for FedMinMax and LocalFedMinMax on Fashion-MNIST and CIFAR-10 datasets. We highlight the worst values. Extended versions for both datasets can be found in [Tables E.1 and E.2](#).

FashionMNIST						
Method	10 Clients			40 Clients		
	ESG	PSG	SSG	ESG	PSG	SSG
LocalFedMinMax	0.316±0.092	0.331±0.007	0.309±0.013	0.346±0.081	0.331±0.021	0.31±0.005
FedMinMax	0.31±0.005	0.308±0.012	0.308±0.003	0.307±0.01	0.31±0.008	0.309±0.011
CIFAR-10						
Method	10 Clients			40 Clients		
	ESG	PSG	SSG	ESG	PSG	SSG
LocalFedMinMax	0.358±0.008	0.353±0.042	0.352±0.0	0.381±0.004	0.378±0.005	0.352±0.007
FedMinMax	0.352±0.02	0.351±0.005	0.351±0.0	0.351±0.002	0.351±0.009	0.351±0.002

7 Conclusion

In this work, we formulate (demographic) group fairness in federated learning setups where different participating entities may only have access to a subset of the population groups during the training phase (but not necessarily the testing phase), exhibiting minmax fairness performance guarantees akin to those in centralized machine learning settings.

We formally show how our fairness definition differs from the existing fair federated learning works, offering conditions under which conventional client-level fairness is equivalent to group-level fairness. We also provide an optimization algorithm, FedMinMax, to solve the minmax group fairness problem in federated setups that exhibits minmax guarantees akin to those of minmax group fair centralized machine learning algorithms.

We empirically confirm that our method outperforms existing federated learning methods in terms of group fairness in various learning settings and validate the conditions under which the competing approaches yield the same solution as our objective.

References

- [1] Robert S. Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [2] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, April 2020.
- [3] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning, 2021.
- [4] European Commission. Reform of eu data protection rules 2018. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.
- [5] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- [6] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Fair and consistent federated learning. *CoRR*, abs/2108.08435, 2021.
- [7] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33, 2020.
- [8] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 2020.
- [9] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- [10] Wei Du and Xintao Wu. Robust fairness-aware learning under sample selection bias. *CoRR*, abs/2105.11570, 2021.
- [11] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. *CoRR*, abs/2010.05057, 2020.
- [12] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l_1 -ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 272–279, New York, NY, USA, 2008. Association for Computing Machinery.
- [13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.
- [14] Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: Enabling group fairness in federated learning, 2021.

- [15] Zhenan Fan, Huang Fang, Zirui Zhou, Jian Pei, Michael P. Friedlander, Changxin Liu, and Yong Zhang. Improving fairness for data valuation in federated learning, 2021.
- [16] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery.
- [17] Arthur M Geoffrion. Proper efficiency and the theory of vector maximization. *Journal of mathematical analysis and applications*, 22(3):618–630, 1968.
- [18] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.
- [19] Samuel Horváth, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Donald Lane. FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- [20] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Fedmgda+: Federated learning meets multi-objective optimization. *CoRR*, abs/2006.11489, 2020.
- [21] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019.
- [22] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020.
- [23] Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.
- [24] Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NeurIPS Workshop on Private Multi-Party Machine Learning*, 2016.

- [25] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).
- [26] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.
- [27] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.
- [28] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021.
- [29] M. Lichman. UCI machine learning repository, 2013.
- [30] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- [31] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. *Collaborative Fairness in Federated Learning*, pages 189–204. Springer International Publishing, Cham, 2020.
- [32] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6755–6764. PMLR, 13–18 Jul 2020.
- [33] Andreu Mas-Colell, Michael Whinston, and Jerry Green. *Microeconomic Theory*. Oxford University Press, 1995.
- [34] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.
- [35] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12. Springer Science & Business Media, 2012.
- [36] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *36th International Conference on Machine Learning, ICML 2019*, 36th International Conference on Machine Learning, ICML 2019, pages 8114–8124. International Machine Learning Society (IMLS), January 2019. 36th International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019.
- [37] Muhammad Tahir Munir, Muhammad Mustansar Saeed, Mahad Ali, Zafar Ayyub Qazi, and Ihsan Ayyub Qazi. Fedprune: Towards inclusive federated learning, 2021.
- [38] Lokesh Nagalapatti and Ramasuri Narayanam. Game of gradients: Mitigating irrelevant clients in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9046–9054, May 2021.

- [39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [40] Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21554–21565. Curran Associates, Inc., 2020.
- [41] Jae Ro, Mingqing Chen, Rajiv Mathews, Mehryar Mohri, and Ananda Theertha Suresh. Communication-Efficient Agnostic Federated Averaging. In *Proc. Interspeech 2021*, pages 871–875, 2021.
- [42] Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers, 2021.
- [43] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. In *IJCAI*, 2021.
- [44] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.
- [45] Xubo Yue, Maher Nouiehed, and Raed Al Kontar. GIFAIR-FL: an approach for group and individual fairness in federated learning. *CoRR*, abs/2108.02741, 2021.
- [46] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning, 2021.
- [47] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060, 2020.
- [48] Fengda Zhang, Kun Kuang, Yuxuan Liu, Chao Wu, Fei Wu, Jiaxun Lu, Yunfeng Shao, and Jun Xiao. Unified group fairness on federated learning, 2021.

A Appendix: Proofs

Lemma 1. Let $\mathbf{P}_{\mathcal{A}}$ denote a matrix whose entry in row a and column k is $p(A = a|K = k)$ (i.e., the prior of group a in client k). Then, given a solution to the minimax problem across clients

$$h^*, \boldsymbol{\lambda}^* \in \arg \min_{h \in \mathcal{H}} \max_{\boldsymbol{\lambda} \in \Delta^{|\mathcal{K}|-1}} \mathbb{E} [\ell(h(X), Y)], \quad (10)$$

$\exists \boldsymbol{\mu}^* = \mathbf{P}_{\mathcal{A}} \boldsymbol{\lambda}^*$ that is solution to the following constrained minimax problem across sensitive groups:

$$h^*, \boldsymbol{\mu}^* \in \arg \min_{h \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1}} \mathbb{E} [\ell(h(X), Y)], \quad (11)$$

where the weighting vector $\boldsymbol{\mu}$ is constrained to belong to the simplex subset defined by $\mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1} \subseteq \Delta^{|\mathcal{A}|-1}$. In particular, if the set $\Gamma = \{\boldsymbol{\mu}' \in \mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1} : \boldsymbol{\mu}' \in \arg \min_{h \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \Delta^{|\mathcal{A}|-1}} \mathbb{E} [\ell(h(X), Y)]\} \neq \emptyset$, then $\boldsymbol{\mu}^* \in \Gamma$, and the minimax fairness solution across clients is also a minimax fairness solution across demographic groups.

PROOF. The objective for optimizing the global model for the worst mixture of client distributions is:

$$\min_{h \in \mathcal{H}} \max_{\boldsymbol{\lambda} \in \Delta^{|\mathcal{K}|-1}} \mathbb{E} [l(h(X), Y)] = \min_{h \in \mathcal{H}} \max_{\boldsymbol{\lambda} \in \Delta^{|\mathcal{K}|-1}} \sum_{k=1}^{|\mathcal{K}|} \lambda_k \mathbb{E}_{p(X, Y|K=k)} [l(h(X), Y)], \quad (12)$$

given that $\mathcal{D}_{\boldsymbol{\lambda}} = \sum_{k=1}^{|\mathcal{K}|} \lambda_k p(X, Y|K = k)$. Since $p(X, Y|K = k) = \sum_{a \in \mathcal{A}} p(A = a|K = k) p(X, Y|A = a)$ with $p(A = a|K = k)$ being the prior of $a \in \mathcal{A}$ for client k , and $p(X, Y|A = a)$ is the distribution conditioned on the sensitive group $a \in \mathcal{A}$, Eq. 12 can be re-written as

$$\begin{aligned} & \min_{h \in \mathcal{H}} \max_{\boldsymbol{\lambda} \in \Delta^{|\mathcal{K}|-1}} \sum_{k=1}^{|\mathcal{K}|} \lambda_k \sum_{a \in \mathcal{A}} p(A = a|K = k) \mathbb{E}_{p(X, Y|A=a)} [l(h(X), Y)] = \\ & \min_{h \in \mathcal{H}} \max_{\boldsymbol{\lambda} \in \Delta^{|\mathcal{K}|-1}} \sum_{a \in \mathcal{A}} \mathbb{E}_{p(X, Y|A=a)} [l(h(X), Y)] \left(\sum_{k=1}^{|\mathcal{K}|} p(A = a|K = k) \lambda_k \right) = \\ & \min_{h \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1}} \sum_{a \in \mathcal{A}} \mu_a \mathbb{E}_{p(X, Y|A=a)} [l(h(X), Y)]. \end{aligned} \quad (13)$$

where $\mu_a = \sum_{k=1}^{|\mathcal{K}|} p(A = a|K = k) \lambda_k$, $\forall a \in \mathcal{A}$. Note that this creates the vector $\boldsymbol{\mu} = \mathbf{P}_{\mathcal{A}} \boldsymbol{\lambda} \subseteq \mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1}$. It holds that the set of possible $\boldsymbol{\mu}$ vectors satisfies $\mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1} \subseteq \Delta^{|\mathcal{A}|-1}$, since $\mathbf{P}_{\mathcal{A}} = \{p(A = a|K = k)\}_{a \in \mathcal{A}}\}_{k \in \mathcal{K}} \in \mathbb{R}_+^{|\mathcal{A}| \times |\mathcal{K}|}$, with $\sum_{a \in \mathcal{A}} p(A = a|K = k) = 1 \forall k$ and $\boldsymbol{\lambda} \in \Delta^{|\mathcal{K}|-1}$.

Then, from the equivalence in Equation 13 we have that

$$h^*, \boldsymbol{\lambda}^* \in \arg \min_{h \in \mathcal{H}} \max_{\boldsymbol{\lambda} \in \Delta^{|\mathcal{K}|-1}} \mathbb{E} [\ell(h(X), Y)], \quad (14)$$

and

$$h^*, \boldsymbol{\mu}^* \in \arg \min_{h \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1}} \mathbb{E}_{\mathcal{D}_{\boldsymbol{\mu}}} [\ell(h(X), Y)], \quad (15)$$

with $\boldsymbol{\mu}^* = \mathbf{P}_{\mathcal{A}} \boldsymbol{\lambda}^*$ have the same minimax risk, that is

$$\mathbb{E}_{\mathcal{D}_{\boldsymbol{\mu}^*}} [\ell(h^*(X), Y)] = \mathbb{E}_{\mathcal{D}_{\boldsymbol{\lambda}^*}} [\ell(h^*(X), Y)]. \quad (16)$$

In particular, if the space $\mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1}$ contains any group minimax fair weights, meaning that the set $\Gamma = \{\boldsymbol{\mu}' \in \mathbf{P}_{\mathcal{A}} \Delta^{|\mathcal{K}|-1}: \boldsymbol{\mu}' \in \arg \min_{h \in \mathcal{H}} \max_{\boldsymbol{\mu} \in \Delta^{|\mathcal{A}|-1}} \mathbb{E}_{\mathcal{D}_{\boldsymbol{\mu}}} [\ell(h(X), Y)]\}$ is not empty, then it follows that any $\boldsymbol{\mu}^*$ (solution to Equation 15) is already minimax fair with respect to the groups $\boldsymbol{\mu}^* \in \Gamma$, and the client-level minimax solution is also a minimax solution across sensitive groups.

Lemma 2. Consider our federated learning setting (Figure 1, right) where each entity k has access to a local dataset $\mathcal{S}_k = \bigcup_{a \in \mathcal{A}} \mathcal{S}_{a,k}$, and a centralized machine learning setting (Figure 1, left) where there is a single entity that has access to a single dataset $\mathcal{S} = \bigcup_{k \in \mathcal{K}} \mathcal{S}_k = \bigcup_{k \in \mathcal{K}} \bigcup_{a \in \mathcal{A}} \mathcal{S}_{a,k}$ (i.e., this single entity in the centralized setting has access to the data of the various clients in the distributed setting). Then, Algorithm 1 (federated) and Algorithm 2 (non-federated, in supplementary material, Appendix B) lead to the same global model provided that learning rates and model initialization are identical.

PROOF. We will show that FedMinMax, in Algorithm 1 is equivalent to the centralized algorithm, in Algorithm 2 under the following conditions:

1. the dataset on client k , in FedMinMax is $\mathcal{S}_k = \bigcup_{a \in \mathcal{A}} \mathcal{S}_{a,k}$ and the dataset in centralized MinMax is $\mathcal{S} = \bigcup_{k \in \mathcal{K}} \mathcal{S}_k = \bigcup_{k \in \mathcal{K}} \bigcup_{a \in \mathcal{A}} \mathcal{S}_{a,k}$; and
2. the model initialization θ^0 , the number of adversarial rounds T ,⁶ learning rate for the adversary η_{μ} , and learning rate for the learner η_{θ} , are identical for both algorithms.

This can then be immediately done by showing that steps lines 3-7 in Algorithm 1 are entirely equivalent to step 3 in Algorithm 2. In particular, note that we can write

$$\begin{aligned} \hat{r}(\boldsymbol{\theta}, \boldsymbol{\mu}) &= \sum_{a \in \mathcal{A}} \mu_a \hat{r}_a(\boldsymbol{\theta}) \\ &= \sum_{a \in \mathcal{A}} \mu_a \sum_{k \in \mathcal{K}} \frac{n_{a,k}}{n_a} \hat{r}_{a,k}(\boldsymbol{\theta}) \\ &= \sum_{a \in \mathcal{A}} \mu_a \frac{n}{n_a} \frac{1}{n} \sum_{k \in \mathcal{K}} n_{a,k} \hat{r}_{a,k}(\boldsymbol{\theta}) \\ &= \sum_{a \in \mathcal{A}} w_a \frac{1}{n} \sum_{k \in \mathcal{K}} \frac{n_{a,k}}{n_k} n_k \hat{r}_{a,k}(\boldsymbol{\theta}) \end{aligned}$$

⁶In the federated Algorithm 1, we also refer to the adversarial rounds as communication rounds.

$$\begin{aligned}
&= \sum_{k \in \mathcal{K}} \frac{n_k}{n} \sum_{a \in \mathcal{A}} w_a \frac{n_{a,k}}{n_k} \hat{r}_{a,k}(\boldsymbol{\theta}) \\
&= \sum_{k \in \mathcal{K}} \frac{n_k}{n} \hat{r}_k(\boldsymbol{\theta}, \mathbf{w}),
\end{aligned} \tag{17}$$

$$\text{where } \hat{r}_k(\boldsymbol{\theta}, \mathbf{w}) = \sum_{a \in \mathcal{A}} \frac{n_{a,k}}{n_k} w_a \hat{r}_{a,k}(\boldsymbol{\theta}), \text{ with } w_a = \frac{\mu_a}{n}, \text{ and } \hat{r}_a(\boldsymbol{\theta}) = \sum_{k \in \mathcal{K}} \frac{n_{a,k}}{n_a} \hat{r}_{a,k}(\boldsymbol{\theta}). \tag{18}$$

Therefore, the model update

$$\boldsymbol{\theta}^t = \sum_{k \in \mathcal{K}} \frac{n_k}{n} \boldsymbol{\theta}_k^t = \sum_{k \in \mathcal{K}} \frac{n_k}{n} (\boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_{\boldsymbol{\theta}} \hat{r}_k(\boldsymbol{\theta}^{t-1}, \mathbf{w}^{t-1})) \tag{19}$$

associated with step in 7 at round t of Algorithm 1, is entirely equivalent to the model update

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_{\boldsymbol{\theta}} \hat{r}(\boldsymbol{\theta}^{t-1}, \mathbf{w}^{t-1}) \tag{20}$$

associated with step in line 3 at round t of Algorithm 2, provided that $\boldsymbol{\theta}^{t-1}$ is the same for both algorithms.

It follows therefore by induction that, provided the initialization $\boldsymbol{\theta}^0$ and learning rate η_θ are identical in both cases the algorithms lead to the same model. Also, from Eq. 18, we have that the projected gradient ascent step in line 4 of Algorithm 2 is equivalent to the step in line 10 of Algorithm 1.

B Appendix: Centralized Minimax Algorithm

We provide the centralized version of FedMinMax in Algorithm 2.

Algorithm 2 CENTRALIZED MINMAX BASELINE

Input: T : total number of adversarial rounds, η_θ : model learning rate, η_μ : adversary learning rate, \mathcal{S}_a : set of examples for group a , $\forall a \in \mathcal{A}$.

- 1: Server **initializes** $\boldsymbol{\mu}^0 \leftarrow \{|\mathcal{S}_a|/|\mathcal{S}|\}_{a \in \mathcal{A}}$ and $\boldsymbol{\theta}^0$ randomly.
- 2: **for** $t = 1$ **to** T **do**
- 3: Server **computes** $\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_{\boldsymbol{\theta}} \hat{r}_k(\boldsymbol{\theta}^{t-1}, \boldsymbol{\mu}^{t-1})$
- 4: Server **updates**

$$\boldsymbol{\mu}^t \leftarrow \prod_{\Delta \in \mathcal{A}} \left(\boldsymbol{\mu}^{t-1} + \eta_\mu \nabla_{\boldsymbol{\mu}} \langle \boldsymbol{\mu}^{t-1}, \hat{r}_a(\boldsymbol{\theta}^{t-1}) \rangle \right)$$
- 5: **end for**

Outputs: $\frac{1}{T} \sum_{t=1}^T \boldsymbol{\theta}^t$

C Appendix: Experimental Details

Experimental Setting and Model Architectures. For AFL and FedMinMax the batch size is equal to the number of examples per client while for TERM, FedAvg and q -FedAvg is equal to 100. For the synthetic

dataset, we use an MLP architecture consisting of four hidden layers of size 512. In the experiments for Adult we use a single layer MLP with 512 neurons. For FashionMNIST we use a CNN architecture with two 2D convolutional layers with kernel size 3, stride 1, and padding 1. Each convolutional layer is followed with a maxpooling layer with kernel size 2, stride 2, dilation 1, and padding 0. For CIFAR-10 we use a ResNet-18 architecture without batch normalization. Finally for ACS Employment dataset we use a single layer MLP with 512 neurons for the experiments where the sensitive label is the combination of race and employment, and Logistic Regression for the experiments with the original 9 races. For training we use either cross entropy or Brier score loss function. We perform a grid search over the following hyperparameters: tilt- $t = \{0.01, 0.1, 0.5, 0.8, 1.0\}$, $q = \{0.2, 0.5, 1.0, 2.0, 5.0\}$, local epochs $E = \{3, 10, 15\}$ and $\eta_\theta = \eta_\mu = \eta_\lambda = \{0.001, 0.005, 0.01, 0.05, 0.1\}$ (where appropriate). We report a summary of the experimental setup in Table C.1. During the training process we tune the hyperparameters based on the validation set for each approach. The mean and standard deviation reported on the results are calculated over three runs. We use 3-fold cross validation to split the data into training and validation for each run.

Table C.1: Summary of parameters used in the training process for all experiments. Epochs refer to the local iterations performed at each client, n_k is the number of local data examples in client k , η_θ is the model’s learning rate and η_μ or η_λ is the adversary learning rates.

Dataset	Setting	Method	η_θ	Batch Size	Loss	Hypothesis Type	Epochs	η_μ or η_λ
Synthetic	ESG,SSG	AFL	0.1	n_k	Brier Score	MLP (4x512)	-	0.1
		FedAvg	0.1	100	Brier Score	MLP (4x512)	15	-
		q -FedAvg	0.1	100	Brier Score	MLP (4x512)	15	-
		FedMinMax (ours)	0.1	n_k	Brier Score	MLP (4x512)	-	0.1
		Centralized Minmax	0.1	n	Brier Score	MLP (4x512)	-	0.1
Adult	ESG,SSG,PSG	AFL	0.01	n_k	Cross Entropy	MLP (512)	-	0.01
		FedAvg	0.01	100	Cross Entropy	MLP (512)	15	-
		q -FedAvg	0.01	100	Cross Entropy	MLP (512)	15	-
		FedMinMax (ours)	0.01	n_k	Cross Entropy	MLP (512)	-	0.01
		Centralized Minmax	0.01	n	Cross Entropy	MLP (512)	-	0.01
FashionMNIST	ESG,SSG,PSG	AFL	0.1	n_k	Brier Score	CNN	-	0.1
		FedAvg	0.1	100	Brier Score	CNN	15	-
		q -FedAvg	0.1	100	Brier Score	CNN	15	-
		FedMinMax (ours)	0.1	n_k	Brier Score	CNN	-	0.1
		Centralized Minmax	0.1	n	Brier Score	CNN	-	0.1
CIFAR-10	ESG,SSG,PSG	AFL	0.1	n_k	Brier Score	ResNet-18 w/o BN	-	0.01
		FedAvg	0.1	100	Brier Score	ResNet-18 w/o BN	3	-
		q -FedAvg	0.1	100	Brier Score	ResNet-18 w/o BN	3	-
		FedMinMax (ours)	0.1	n_k	Brier Score	ResNet-18 w/o BN	-	0.01
		Centralized Minmax	0.1	n	Brier Score	ResNet-18 w/o BN	-	0.01
ACS Employment (6 sensitive groups)	ESG,SSG,PSG	AFL	0.01	n_k	Cross Entropy	MLP (512)	-	0.01
		FedAvg	0.01	100	Cross Entropy	MLP (512)	10	-
		q -FedAvg	0.01	100	Cross Entropy	MLP (512)	10	-
		FedMinMax (ours)	0.01	n_k	Cross Entropy	MLP (512)	-	0.01
		Centralized Minmax	0.01	n	Cross Entropy	MLP (512)	-	0.01
ACS Employment (9 sensitive groups)	ESG,SSG,PSG	AFL	0.01	n_k	Cross Entropy	Logistic Regression	-	0.01
		FedAvg	0.01	100	Cross Entropy	Logistic Regression	10	-
		q -FedAvg	0.01	100	Cross Entropy	Logistic Regression	10	-
		FedMinMax (ours)	0.01	n_k	Cross Entropy	Logistic Regression	-	0.01
		Centralized Minmax	0.01	n	Cross Entropy	Logistic Regression	-	0.01

Software & Hardware. The proposed algorithms and experiments are written in Python, leveraging PyTorch [39]. The experiments were realised using $1 \times$ NVIDIA Tesla V100 GPU.

D Appendix: Additional Results

Experiments on Synthetic dataset. Recall that we consider two sensitive groups (i.e., $|\mathcal{A}| = 2$) in the synthetic dataset. In the *Equal access to Sensitive Groups (ESG)* setting, we distribute the two groups on 40 clients, while for the *Single access to Sensitive Groups (SSG)* case, every client has access to a single group, each group is distributed to 20 clients, and the amount of samples on each local dataset varies across clients. There is no *Partial access to Sensitive Groups (PSG)* setting for binary sensitive group scenarios since it is equivalent to SSG. A comparison of the testing group risks is provided in Table D.2 and the weighting coefficients for the groups are given by Table D.1.

Table D.1: Final group weighting coefficients for AFL and FedMinmax across different federated learning scenarios on the synthetic dataset for binary classification involving two sensitive groups.

Setting	Method	Worst Group	Best Group
ESG	AFL	0.528	0.472
	FedMinMax (ours)	0.999	0.001
SSG	AFL	0.999	0.001
	FedMinMax (ours)	0.999	0.001
Centralized Minmax Baseline		0.999	0.001

Table D.2: Testing Brier score risks for FedAvg, AFL, q -FedAvg, TERM, and FedMinmax across different federated learning scenarios on the synthetic dataset for binary classification involving two sensitive groups. PSG scenario is not included because for $|\mathcal{A}| = 2$ it is equivalent to SSG.

Setting	Method	Worst Group Risk	Best Group Risk
ESG	AFL	0.485±0.0	0.216±0.001
	FedAvg	0.487±0.0	0.214±0.002
	q -FedAvg ($q=0.2$)	0.479±0.002	0.22±0.002
	q -FedAvg ($q=5.0$)	0.478±0.002	0.223±0.004
	TERM ($t=1.0$)	0.469±0.0	0.261±0.001
	FedMinMax (ours)	0.451±0.0	0.31±0.001
SSG	AFL	0.451±0.0	0.31±0.001
	FedAvg	0.483±0.002	0.219±0.001
	q -FedAvg ($q=0.2$)	0.476±0.001	0.221±0.002
	q -FedAvg ($q=5.0$)	0.468±0.005	0.274±0.004
	TERM ($t=1.0$)	0.461±0.004	0.272±0.001
	FedMinMax (ours)	0.451±0.0	0.309±0.003
Centralized Minmax Baseline		0.451±0.0	0.308±0.001

Experiments on Adult dataset. In the *Equal access to Sensitive Groups (ESG)* setting, we distribute the 4 groups equally on 40 clients. In the *Partial access to Sensitive Groups (PSG)* setting, 20 clients have access to *Males* subgroups, and the other 20 to subgroups relating to *Females*. In the *Single access to Sensitive Groups (SSG)* setting, every client has access to a single group and each group is distributed to 10 clients. We show the testing group risks in Table D.4 and the group weights in Table D.3.

Table D.3: Final group weighting coefficients for AFL and FedMinmax across different federated learning scenarios on the Adult dataset. We round the weights values to the last three decimal places.

Setting	Method	Males earning $\leq 50K$	Males earning $> 50K$	Females earning $\leq 50K$	Females earning $> 50K$
ESG	AFL	0.475	0.214	0.284	0.028
	FedMinMax (ours)	0.697	0.301	0.001	0.001
SSG	AFL	0.705	0.293	0.003	0.001
	FedMinMax (ours)	0.697	0.301	0.001	0.001
PSG	AFL	0.500	0.229	0.244	0.027
	FedMinMax (ours)	0.705	0.293	0.001	0.001
Centralized Minmax Baseline		0.697	0.301	0.001	0.001

Table D.4: Cross entropy risks for FedAvg, AFL, q-FedAvg, TERM, and FedMinmax across different federated learning settings on adult dataset.

Setting	Method	Males earning \leq 50K	Males earning $>$ 50K	Females earning \leq 50K	Females earning $>$ 50K
ESG	AFL	0.263 \pm 0.002	0.701 \pm 0.003	0.086 \pm 0.002	1.096 \pm 0.008
	FedAvg	0.255 \pm 0.002	0.697 \pm 0.004	0.081 \pm 0.001	1.121 \pm 0.009
	q-FedAvg	0.263 \pm 0.003	0.697 \pm 0.004	0.084 \pm 0.001	1.1 \pm 0.006
	TERM	0.381 \pm 0.101	0.607 \pm 0.04	0.224 \pm 0.06	0.725 \pm 0.021
	FedMinMax (ours)	0.414 \pm 0.003	0.453 \pm 0.003	0.415 \pm 0.008	0.347\pm0.007
SSG	AFL	0.418 \pm 0.006	0.452 \pm 0.009	0.416 \pm 0.002	0.349\pm0.007
	FedAvg	0.263 \pm 0.001	0.704 \pm 0.002	0.07 \pm 0.0	1.23 \pm 0.002
	q-FedAvg	0.261 \pm 0.001	0.683 \pm 0.002	0.082 \pm 0.001	1.117 \pm 0.01
	TERM	0.358 \pm 0.016	0.579 \pm 0.002	0.286 \pm 0.031	0.693 \pm 0.071
	FedMinMax (ours)	0.413 \pm 0.002	0.453 \pm 0.005	0.414 \pm 0.006	0.348\pm0.01
PSG	AFL	0.274 \pm 0.003	0.757 \pm 0.009	0.094 \pm 0.002	1.285 \pm 0.022
	FedAvg	0.263 \pm 0.001	0.7 \pm 0.001	0.069 \pm 0.001	1.226 \pm 0.007
	q-FedAvg	0.263 \pm 0.004	0.752 \pm 0.014	0.09 \pm 0.004	1.239 \pm 0.032
	TERM	0.485 \pm 0.195	0.581 \pm 0.108	0.367 \pm 0.316	0.69 \pm 0.003
	FedMinMax (ours)	0.411 \pm 0.002	0.452 \pm 0.006	0.417 \pm 0.001	0.346\pm0.008
Centralized Minmax Baseline		0.412 \pm 0.004	0.453 \pm 0.005	0.416 \pm 0.012	0.347\pm0.004

Experiments on FashionMNIST dataset. For the *Equal access to Sensitive Groups (ESG)* setting, each client in the federation has access to the same amount of the 10 classes. In the *Partial access to Sensitive Groups (PSG)* setting, 20 of the participants have access only to groups *T-shirt, Trouser, Pullover, Dress* and *Coat*. The remaining 20 clients own data from groups *Sandal, Shirt, Sneaker, Bag* and *Ankle Boot*. Finally, in the *Single access to Sensitive Groups (SSG)* setting, every group is owned by 4 clients only and all clients have access to just one group membership. The group risks are provided in Table D.5. We also show the weighting coefficients for each sensitive group in Table D.6.

Table D.5: Brier score risks for FedAvg, AFL, q-FedAvg, TERM, and FedMinmax across different federated learning settings on FashionMNIST dataset.

Setting	Method	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
ESG	AFL	0.239 \pm 0.003	0.046 \pm 0.0	0.262 \pm 0.001	0.159 \pm 0.001	0.252 \pm 0.004	0.06 \pm 0.0	0.494 \pm 0.004	0.067 \pm 0.001	0.049 \pm 0.0	0.07 \pm 0.001
	FedAvg	0.243 \pm 0.003	0.046 \pm 0.0	0.262 \pm 0.001	0.158 \pm 0.003	0.253 \pm 0.002	0.061 \pm 0.0	0.492 \pm 0.003	0.068 \pm 0.0	0.049 \pm 0.0	0.069 \pm 0.0
	q-FedAvg	0.268 \pm 0.051	0.047 \pm 0.005	0.312 \pm 0.016	0.164 \pm 0.029	0.306 \pm 0.052	0.039 \pm 0.003	0.477 \pm 0.006	0.074 \pm 0.001	0.036 \pm 0.005	0.056 \pm 0.008
	TERM	0.256 \pm 0.066	0.048 \pm 0.008	0.31 \pm 0.083	0.175 \pm 0.022	0.294 \pm 0.016	0.041 \pm 0.012	0.467 \pm 0.002	0.066 \pm 0.019	0.038 \pm 0.011	0.062 \pm 0.018
	FedMinMax (ours)	0.261 \pm 0.006	0.191 \pm 0.016	0.256 \pm 0.027	0.217 \pm 0.013	0.223 \pm 0.031	0.207 \pm 0.027	0.307\pm0.01	0.172 \pm 0.016	0.193 \pm 0.021	0.156 \pm 0.011
SSG	AFL	0.267 \pm 0.009	0.194 \pm 0.023	0.236 \pm 0.013	0.226 \pm 0.012	0.262 \pm 0.012	0.201 \pm 0.026	0.307\pm0.003	0.178 \pm 0.033	0.205 \pm 0.025	0.162 \pm 0.021
	FedAvg	0.227 \pm 0.003	0.039 \pm 0.001	0.236 \pm 0.004	0.143 \pm 0.003	0.232 \pm 0.003	0.051 \pm 0.001	0.463 \pm 0.003	0.067 \pm 0.0	0.041 \pm 0.0	0.063 \pm 0.001
	q-FedAvg	0.24 \pm 0.001	0.041 \pm 0.008	0.246 \pm 0.026	0.142 \pm 0.014	0.257 \pm 0.028	0.036 \pm 0.001	0.425 \pm 0.002	0.059 \pm 0.014	0.027 \pm 0.002	0.042 \pm 0.007
	TERM	0.251 \pm 0.011	0.034 \pm 0.003	0.26 \pm 0.017	0.144 \pm 0.005	0.242 \pm 0.034	0.04 \pm 0.004	0.399 \pm 0.017	0.05 \pm 0.003	0.026 \pm 0.001	0.044 \pm 0.001
	FedMinMax (ours)	0.269 \pm 0.012	0.2 \pm 0.026	0.238 \pm 0.017	0.231 \pm 0.013	0.252 \pm 0.034	0.2 \pm 0.024	0.309\pm0.011	0.177 \pm 0.03	0.205 \pm 0.032	0.169 \pm 0.013
PSG	AFL	0.244 \pm 0.007	0.032 \pm 0.001	0.257 \pm 0.066	0.122 \pm 0.006	0.209 \pm 0.098	0.045 \pm 0.002	0.425 \pm 0.019	0.059 \pm 0.001	0.041 \pm 0.001	0.062 \pm 0.001
	FedAvg	0.229 \pm 0.008	0.039 \pm 0.0	0.236 \pm 0.004	0.142 \pm 0.002	0.232 \pm 0.003	0.052 \pm 0.001	0.464 \pm 0.011	0.067 \pm 0.001	0.042 \pm 0.001	0.063 \pm 0.001
	q-FedAvg	0.278 \pm 0.062	0.04 \pm 0.013	0.256 \pm 0.083	0.16 \pm 0.026	0.311 \pm 0.044	0.045 \pm 0.013	0.453 \pm 0.002	0.063 \pm 0.02	0.029 \pm 0.007	0.047 \pm 0.004
	TERM	0.226 \pm 0.007	0.037 \pm 0.005	0.233 \pm 0.004	0.153 \pm 0.007	0.255 \pm 0.016	0.038 \pm 0.0	0.439 \pm 0.007	0.053 \pm 0.003	0.026 \pm 0.001	0.043 \pm 0.002
	FedMinMax (ours)	0.263 \pm 0.013	0.177 \pm 0.026	0.228 \pm 0.011	0.21 \pm 0.019	0.238 \pm 0.025	0.182 \pm 0.03	0.31\pm0.008	0.16 \pm 0.027	0.184 \pm 0.031	0.154 \pm 0.018
Centralized Minmax Baseline		0.259 \pm 0.01	0.173 \pm 0.015	0.239 \pm 0.051	0.213 \pm 0.008	0.24 \pm 0.063	0.182 \pm 0.024	0.311\pm0.006	0.168 \pm 0.018	0.18 \pm 0.013	0.151 \pm 0.012

Table D.6: Final group weighting coefficients for AFL, Centralized Minmax Baseline, and FedMinmax across different federated learning scenarios on the FashionMNIST dataset. Note that the weighting coefficients are rounded to the last three decimal places. We highlight the weighting coefficient for the worst group.

Setting	Method	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
ESG	AFL	0.099	0.100	0.101	0.101	0.100	0.100	0.099	0.100	0.100	0.100
	FedMinMax (ours)	0.217	0.001	0.241	0.007	0.151	0.001	0.380	0.001	0.001	0.001
SSG	AFL	0.217	0.001	0.241	0.007	0.151	0.001	0.379	0.001	0.001	0.001
	FedMinMax (ours)	0.216	0.001	0.237	0.017	0.155	0.001	0.370	0.001	0.001	0.001
PSG	AFL	0.128	0.064	0.138	0.099	0.129	0.063	0.173	0.069	0.066	0.071
	FedMinMax (ours)	0.216	0.001	0.238	0.014	0.154	0.001	0.372	0.001	0.001	0.001
Centralized Minmax Baseline		0.217	0.001	0.240	0.010	0.152	0.001	0.377	0.001	0.001	0.001

Experiments on CIFAR-10 dataset. In the *Equal access to Sensitive Groups (ESG)* setting, the 10 classes are equally distributed across the clients, creating a scenario where each client has access to the same amount of data examples and groups. In the *Partial access to Sensitive Groups (PSG)* setting, 20 clients own data from groups *Airplane*, *Automobile*, *Bird*, *Cat* and *Deer* and the rest hold data from *Dog*, *Frog*, *Horse*, *Ship* and *Truck* groups. Finally, in the *Single access to Sensitive Groups (SSG)* setting, every client owns only one sensitive group and each group is distributed to only 4 clients. We report the risks on the test set in Table D.7 and the final group weighting coefficients in Table D.8.

Table D.7: Brier score risks for FedAvg, AFL, q-FedAvg, TERM, and FedMinmax across different federated learning settings on CIFAR-10 dataset.

Setting	Method	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
ESG	AFL	0.14±0.001	0.104±0.009	0.289±0.011	0.461±0.01	0.243±0.01	0.28±0.016	0.151±0.009	0.14±0.009	0.125±0.012	0.132±0.009
	FedAvg	0.148±0.014	0.108±0.006	0.283±0.011	0.487±0.002	0.237±0.002	0.256±0.002	0.144±0.005	0.148±0.008	0.123±0.003	0.128±0.004
	q-FedAvg	0.178±0.065	0.118±0.047	0.308±0.099	0.507±0.003	0.311±0.054	0.41±0.01	0.179±0.012	0.119±0.013	0.158±0.07	0.182±0.05
	TERM	0.217±0.087	0.115±0.006	0.311±0.057	0.491±0.007	0.274±0.055	0.272±0.026	0.176±0.041	0.166±0.013	0.175±0.069	0.12±0.006
	FedMinMax (ours)	0.257±0.003	0.189±0.009	0.324±0.015	0.351±0.002	0.291±0.004	0.291±0.03	0.231±0.007	0.309±0.008	0.194±0.002	0.158±0.008
SSG	AFL	0.283±0.027	0.259±0.001	0.18±0.008	0.352±0.0	0.285±0.002	0.328±0.008	0.231±0.043	0.212±0.031	0.198±0.012	0.159±0.007
	FedAvg	0.189±0.011	0.102±0.009	0.253±0.005	0.485±0.017	0.239±0.079	0.339±0.074	0.148±0.021	0.166±0.029	0.121±0.019	0.138±0.022
	q-FedAvg	0.18±0.026	0.11±0.017	0.29±0.016	0.437±0.002	0.334±0.069	0.345±0.009	0.161±0.03	0.175±0.057	0.176±0.105	0.129±0.013
	TERM	0.149±0.015	0.146±0.014	0.378±0.042	0.392±0.021	0.262±0.039	0.307±0.02	0.192±0.052	0.176±0.003	0.167±0.032	0.119±0.029
	FedMinMax (ours)	0.258±0.01	0.187±0.005	0.332±0.005	0.351±0.002	0.293±0.007	0.334±0.017	0.216±0.009	0.305±0.009	0.205±0.002	0.154±0.005
PSG	AFL	0.158±0.019	0.121±0.01	0.289±0.015	0.439±0.006	0.247±0.01	0.28±0.014	0.151±0.016	0.168±0.011	0.125±0.013	0.118±0.009
	FedAvg	0.167±0.005	0.098±0.004	0.32±0.009	0.471±0.014	0.224±0.036	0.304±0.009	0.15±0.009	0.162±0.028	0.113±0.003	0.121±0.013
	q-FedAvg	0.173±0.008	0.132±0.027	0.303±0.001	0.46±0.001	0.259±0.038	0.297±0.009	0.178±0.037	0.147±0.013	0.129±0.025	0.114±0.017
	TERM	0.177±0.034	0.137±0.025	0.4±0.066	0.415±0.006	0.303±0.074	0.33±0.029	0.172±0.036	0.172±0.076	0.164±0.044	0.18±0.005
	FedMinMax (ours)	0.261±0.007	0.184±0.007	0.321±0.021	0.351±0.009	0.295±0.003	0.323±0.011	0.22±0.008	0.299±0.011	0.201±0.001	0.154±0.008
Centralized Minmax Baseline		0.263±0.013	0.187±0.005	0.325±0.016	0.352±0.003	0.293±0.007	0.334±0.017	0.216±0.009	0.305±0.009	0.205±0.002	0.154±0.005

Table D.8: Final group weighting coefficients for AFL, Centralized Minmax Baseline, and FedMinmax across different federated learning scenarios on the CIFAR-10 dataset. The weights are rounded to the last three decimal places and the weighting coefficients for the worst group are in bold.

Setting	Method	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
ESG	AFL	0.100	0.100	0.100	0.101	0.099	0.102	0.099	0.100	0.100	0.100
	FedMinMax (ours)	0.075	0.031	0.152	0.206	0.102	0.192	0.083	0.085	0.035	0.039
SSG	AFL	0.088	0.031	0.140	0.207	0.101	0.200	0.079	0.074	0.054	0.028
	FedMinMax (ours)	0.071	0.030	0.147	0.209	0.103	0.195	0.082	0.085	0.038	0.040
PSG	AFL	0.091	0.066	0.119	0.128	0.118	0.103	0.102	0.097	0.081	0.097
	FedMinMax (ours)	0.078	0.045	0.143	0.207	0.108	0.203	0.080	0.078	0.033	0.024
Centralized Minmax Baseline		0.082	0.017	0.139	0.205	0.118	0.190	0.091	0.080	0.032	0.046

Experiments on ACS Employment dataset (employment and race combination). In the *Equal access to Sensitive Groups (ESG)* setting, we split the 6 groups across the clients equally. In the *Partial access to Sensitive Groups (PSG)* setting, 20 clients own data from groups *Unemployed White*, *Employed Black*, *Employed White*, and the remaining own data from *Unemployed Other*, *Unemployed Black*, and *Employed Other*. Finally, in the *Single access to Sensitive Groups (SSG)* setting, every client has access to only one sensitive class. In particular, data for *Employed White* is owned by 10 clients and each of the remaining 5 groups is allocated to six clients. We report the risks on the test set in Table D.9 and the group weighting coefficients produced from the training process are in Table D.10.

Table D.9: Test risks for FedAvg, AFL, q-FFL, TERM, and FedMinmax across different federated learning settings on ACS Employment dataset.

Setting	Method	Unemployed White	Employed White	Employed Black	Unemployed Other	Unemployed Black	Employed Other
ESG	AFL	0.322±0.004	0.47±0.006	0.45±0.003	0.424±0.004	0.357±0.002	0.328±0.004
	FedAvg	0.312±0.003	0.486±0.005	0.459±0.002	0.435±0.004	0.351±0.002	0.317±0.003
	q-FedAvg	0.335±0.005	0.451±0.007	0.44±0.004	0.411±0.005	0.365±0.003	0.341±0.006
	TERM	0.349±0.006	0.431±0.008	0.429±0.004	0.396±0.006	0.373±0.003	0.357±0.007
	FedMinMax (ours)	0.383±0.003	0.374±0.005	0.381±0.001	0.366±0.008	0.374±0.001	0.36±0.01
SSG	AFL	0.386±0.01	0.374±0.004	0.384±0.007	0.365±0.009	0.377±0.009	0.362±0.007
	FedAvg	0.256±0.002	0.596±0.005	0.517±0.003	0.527±0.005	0.316±0.002	0.249±0.003
	q-FedAvg	0.261±0.003	0.582±0.007	0.51±0.005	0.513±0.007	0.32±0.002	0.258±0.004
	TERM	0.27±0.001	0.563±0.003	0.499±0.001	0.499±0.003	0.326±0.001	0.267±0.002
	FedMinMax (ours)	0.384±0.006	0.373±0.004	0.383±0.003	0.365±0.005	0.375±0.007	0.36±0.007
PSG	AFL	0.287±0.004	0.529±0.008	0.481±0.005	0.469±0.005	0.337±0.003	0.289±0.003
	FedAvg	0.278±0.005	0.548±0.011	0.491±0.006	0.485±0.004	0.331±0.004	0.277±0.003
	q-FedAvg	0.296±0.002	0.513±0.003	0.472±0.002	0.457±0.003	0.343±0.001	0.298±0.003
	TERM	0.303±0.004	0.5±0.008	0.466±0.005	0.447±0.001	0.347±0.003	0.306±0.001
	FedMinMax (ours)	0.385±0.004	0.375±0.005	0.384±0.006	0.364±0.001	0.376±0.003	0.36±0.002
Centralized Minmax Baseline		0.381±0.006	0.375±0.003	0.382±0.002	0.367±0.004	0.374±0.007	0.359±0.011

Table D.10: Final group weighting coefficients for AFL, Centralized Minmax Baseline, and FedMinmax for the ACS Employment dataset. The weights are rounded to the last three decimal places.

Setting	Method	Unemployed White	Employed White	Employed Black	Unemployed Other	Unemployed Black	Employed Other
ESG	AFL	0.419	0.351	0.038	0.078	0.062	0.052
	FedMinMax (ours)	0.461	0.356	0.041	0.044	0.070	0.029
SSG	AFL	0.461	0.355	0.040	0.045	0.070	0.029
	FedMinMax	0.461	0.356	0.040	0.045	0.070	0.028
PSG	AFL	0.431	0.343	0.040	0.072	0.067	0.048
	FedMinMax (ours)	0.461	0.356	0.041	0.044	0.070	0.029
Centralized Minmax Baseline		0.461	0.356	0.040	0.045	0.070	0.029

Experiments on ACS Employment dataset (race). We also use the original 9 races of the ACS Employment dataset to run experiments on the three federated learning settings. We refer to the available race groups using the following label tags: { *White*: White alone, *Black*: Black or African American alone, *American Indian*: American Indian alone, *Alaska Native*: Alaska Native alone, *A.I. &/or A.N. Tribes*: American Indian and Alaska Native tribes specified, or American Indian or Alaska Native, not specified and no other races, *Asian*: Asian alone, *N. Hawaiian & other P.I.*: Native Hawaiian and Other Pacific Islander alone, *Other*: Some Other Race alone, *Multiple*: Two or More Races}. In the *Equal access to Sensitive Groups (ESG)* setting, we split the 9 groups across the clients. In the *Partial access to Sensitive Groups (PSG)* setting, 20 clients own data from groups *White*, *Black /African American*, *American Indian*, *Alaska Native*, *A.I. &/or A.N. Tribes* and the remaining clients hold data from *Asian*, *N. Hawaiian & other P.I.*, *Other*, and *Multiple*. Finally, in the *Single access to Sensitive Groups (SSG)* setting, every client owns only one sensitive group and each group is distributed to only 4 clients, except *White* race that is distributed to 8 clients. We report the risks on the test set in Table D.11.

Table D.11: Risks for FedAvg, AFL, q-FFL, TERM, and FedMinmax across different federated learning settings on ACS Employment dataset.

Setting	Method	White	Black	American Indian	Alaska Native	A.I. &/or A.N. Tribes	Asian	N. Hawaiian & other P.I.	Other	Multiple	
ESG	AFL	0.47±0.003	0.477±0.002	0.499±0.002	0.438±0.009	0.555±0.001	0.487±0.001	0.526±0.003	0.468±0.006	0.363±0.001	
	FedAvg	0.471±0.004	0.477±0.002	0.501±0.002	0.437±0.012	0.556±0.001	0.488±0.001	0.526±0.004	0.471±0.007	0.363±0.002	
	q-FedAvg	0.47±0.001	0.476±0.001	0.499±0.0	0.436±0.005	0.554±0.001	0.487±0.0	0.525±0.001	0.468±0.002	0.363±0.0	
	TERM	0.47±0.004	0.483±0.005	0.504±0.007	0.398±0.043	0.553±0.001	0.488±0.001	0.527±0.004	0.469±0.008	0.365±0.003	
	FedMinMax (ours)	0.467±0.0	0.48±0.001	0.5±0.001	0.375±0.004	0.545±0.0	0.487±0.001	0.522±0.0	0.464±0.001	0.363±0.0	
	SSG	AFL	0.467±0.0	0.479±0.0	0.499±0.0	0.396±0.003	0.547±0.001	0.488±0.0	0.523±0.0	0.465±0.0	0.362±0.0
SSG	FedAvg	0.473±0.002	0.475±0.001	0.501±0.0	0.412±0.009	0.575±0.003	0.487±0.0	0.524±0.003	0.482±0.001	0.363±0.001	
	q-FedAvg	0.472±0.001	0.475±0.001	0.5±0.0	0.418±0.005	0.571±0.001	0.487±0.0	0.525±0.001	0.48±0.001	0.364±0.0	
	TERM	0.469±0.001	0.474±0.0	0.5±0.001	0.421±0.006	0.567±0.002	0.487±0.0	0.525±0.001	0.48±0.001	0.363±0.0	
	FedMinMax (ours)	0.467±0.0	0.479±0.001	0.499±0.0	0.383±0.004	0.546±0.001	0.487±0.001	0.522±0.0	0.465±0.001	0.363±0.0	
	PSG	AFL	0.468±0.0	0.475±0.0	0.503±0.002	0.424±0.0	0.563±0.0	0.49±0.001	0.529±0.002	0.481±0.002	0.365±0.001
	FedAvg	0.468±0.0	0.475±0.001	0.503±0.003	0.421±0.002	0.564±0.001	0.489±0.003	0.529±0.004	0.481±0.003	0.365±0.001	
PSG	q-FedAvg	0.468±0.0	0.475±0.0	0.503±0.001	0.43±0.011	0.561±0.001	0.49±0.001	0.53±0.002	0.48±0.002	0.365±0.001	
	TERM	0.471±0.006	0.476±0.003	0.502±0.003	0.434±0.009	0.559±0.001	0.489±0.002	0.528±0.005	0.474±0.009	0.364±0.002	
	FedMinMax (ours)	0.467±0.0	0.48±0.001	0.5±0.001	0.373±0.004	0.546±0.001	0.486±0.0	0.522±0.0	0.465±0.0	0.363±0.001	
	Centralized Minmax Baseline	0.467±0.0	0.48±0.0	0.5±0.001	0.372±0.002	0.545±0.0	0.486±0.001	0.522±0.001	0.465±0.001	0.364±0.0	

E Appendix: Complementary Algorithms

In the main text we refer to slightly different optimization objective and an algorithm that we use to compare the generalization efficiency of considering global demographics on some scenarios. Here we provide more information about this approach. In particular we consider the following empirical objective:

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{\mu \in \Delta_{\geq \epsilon}^{|\mathcal{A}| |\mathcal{K}| - 1}} \hat{r}_{a,k}(\boldsymbol{\theta}) := \sum_{a \in \mathcal{A}} \sum_{k \in \mathcal{K}} \mu_{a,k} \hat{r}_{a,k}(\boldsymbol{\theta}). \quad (21)$$

Note that this optimization objective assumes that each demographic group that a client has access to is treated as a unique sensitive group even if the same group exists in several clients. We extend our FedMinMax algorithm to solve the objective in Eq. 21. The adjusted algorithm is called LocalFedMinMax for which we share the pseudocode in Algorithm 3 and the full table of risks for FashionMNIST and CIFAR-10 in Tables E.1 and E.2, respectively. LocalFedMinMax and FedMinMax behave similarly on the worst group on SSG regardless for different number of clients, while LocalFedMinMax has higher worst group risks for the remaining settings compared to FedMinMax.

Algorithm 3 LOCAL FEDERATED MINIMAX (LOCALFEDMINMAX)

Input: \mathcal{K} : Set of clients, T : total number of communication rounds, $\eta_{\boldsymbol{\theta}}$: model learning rate, $\eta_{\boldsymbol{\mu}}$: global adversary learning rate, $\mathcal{S}_{a,k}$: set of examples for group a in client k , $\forall a \in \mathcal{A}$ and $\forall k \in \mathcal{K}$.

- 1: Server **initializes** $\boldsymbol{\mu}^0 \leftarrow \rho = \{ \{ |\mathcal{S}_{a,k}| / |\mathcal{S}| \}_{a \in \mathcal{A}} \}_{k \in \mathcal{K}}$ and $\boldsymbol{\theta}^0$ randomly.
 - 2: **for** $t = 1$ **to** T **do**
 - 3: Server **computes** $\boldsymbol{w}^{t-1} \leftarrow \boldsymbol{\mu}^{t-1} / \rho$
 - 4: Server **broadcasts** $\boldsymbol{\theta}^{t-1}, \boldsymbol{w}^{t-1}$
 - 5: **for** each client $k \in \mathcal{K}$ **in parallel do**
 - 6: $\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}^{t-1} - \eta_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \hat{r}_k(\boldsymbol{\theta}^{t-1}, \boldsymbol{w}^{t-1})$
 - 7: Client- k **obtains** and **sends** $\{ \hat{r}_{a,k}(\boldsymbol{\theta}^{t-1}) \}_{a \in \mathcal{A}}$ and $\boldsymbol{\theta}_k^t$ to server
 - 8: **end for**
 - 9: Server **computes:** $\boldsymbol{\theta}^t \leftarrow \sum_{k \in \mathcal{K}} \frac{n_k}{n} \boldsymbol{\theta}_k^t$
 - 10: Server **updates:** $\boldsymbol{\mu}^t \leftarrow \prod_{\Delta^{|\mathcal{K}| * |\mathcal{A}| - 1}} \left(\boldsymbol{\mu}^{t-1} + \eta_{\boldsymbol{\mu}} \nabla_{\boldsymbol{\mu}} \langle \boldsymbol{\mu}^{t-1}, \hat{r}_{a,k}(\boldsymbol{\theta}^{t-1}) \rangle \right)$
 - 11: **end for**
- Outputs:** $\frac{1}{T} \sum_{t=1}^T \boldsymbol{\theta}^t$
-

Table E.1: Brier Score risks for FedMinMax and LocalFedMinMax on FashionMNIST across the different federated learning scenarios.

Setting	Method	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
ESG (10 clients)	LocalFedMinMax	0.298±0.054	0.173±0.021	0.316±0.092	0.224±0.006	0.256±0.036	0.184±0.033	0.29±0.022	0.157±0.042	0.185±0.015	0.149±0.019
	FedMinMax (ours)	0.25±0.003	0.168±0.014	0.218±0.015	0.205±0.008	0.243±0.025	0.184±0.021	0.31±0.005	0.159±0.016	0.174±0.017	0.143±0.005
SSG (10 clients)	LocalFedMinMax	0.288±0.055	0.153±0.009	0.253±0.069	0.22±0.023	0.251±0.029	0.161±0.024	0.309±0.013	0.15±0.021	0.166±0.007	0.135±0.004
	FedMinMax (ours)	0.265±0.004	0.184±0.023	0.229±0.016	0.216±0.017	0.256±0.031	0.192±0.029	0.308±0.003	0.177±0.029	0.193±0.023	0.158±0.014
PSG (10 clients)	LocalFedMinMax	0.331±0.007	0.153±0.008	0.323±0.03	0.232±0.005	0.23±0.0	0.152±0.012	0.307±0.012	0.131±0.005	0.167±0.01	0.134±0.003
	FedMinMax (ours)	0.266±0.002	0.187±0.021	0.278±0.029	0.217±0.015	0.201±0.044	0.192±0.04	0.308±0.012	0.165±0.022	0.187±0.026	0.158±0.011
ESG (40 clients)	LocalFedMinMax	0.284±0.008	0.03±0.012	0.346±0.081	0.147±0.007	0.232±0.006	0.156±0.006	0.271±0.004	0.165±0.0	0.09±0.008	0.154±0.009
	FedMinMax (ours)	0.261±0.006	0.191±0.016	0.256±0.027	0.217±0.013	0.223±0.031	0.207±0.027	0.307±0.01	0.172±0.016	0.193±0.021	0.156±0.011
SSG (40 clients)	LocalFedMinMax	0.25±0.005	0.206±0.003	0.24±0.006	0.25±0.007	0.28±0.007	0.23±0.01	0.31±0.05	0.105±0.006	0.18±0.008	0.182±0.001
	FedMinMax (ours)	0.269±0.012	0.2±0.026	0.238±0.017	0.231±0.013	0.252±0.034	0.2±0.024	0.309±0.011	0.177±0.03	0.205±0.032	0.169±0.013
PSG (40 clients)	LocalFedMinMax	0.331±0.021	0.039±0.006	0.281±0.001	0.178±0.006	0.191±0.051	0.065±0.05	0.275±0.006	0.068±0.1	0.041±0.09	0.12±0.2
	FedMinMax (ours)	0.263±0.013	0.177±0.026	0.228±0.011	0.21±0.019	0.238±0.025	0.182±0.03	0.31±0.008	0.16±0.027	0.184±0.031	0.154±0.018

Table E.2: Brier score risks for LocalFedMinMax and FedMinmax on CIFAR-10 dataset across different federated learning scenarios.

Setting	Method	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
ESG (10 clients)	LocalFedMinMax	0.24±0.039	0.119±0.015	0.319±0.018	0.358±0.008	0.278±0.024	0.276±0.022	0.264±0.001	0.197±0.029	0.213±0.111	0.14±0.053
	FedMinMax (ours)	0.279±0.028	0.243±0.089	0.32±0.019	0.352±0.02	0.266±0.03	0.33±0.013	0.229±0.02	0.323±0.029	0.222±0.037	0.219±0.022
SSG (10 clients)	LocalFedMinMax	0.263±0.012	0.236±0.04	0.227±0.09	0.352±0.0	0.29±0.009	0.334±0.017	0.234±0.039	0.25±0.055	0.199±0.013	0.156±0.003
	FedMinMax (ours)	0.278±0.032	0.211±0.043	0.284±0.083	0.351±0.0	0.287±0.004	0.328±0.008	0.213±0.014	0.267±0.065	0.204±0.002	0.157±0.009
PSG (10 clients)	LocalFedMinMax	0.235±0.018	0.161±0.044	0.294±0.004	0.353±0.042	0.249±0.073	0.331±0.03	0.226±0.025	0.236±0.0	0.189±0.096	0.223±0.093
	FedMinMax (ours)	0.236±0.024	0.185±0.006	0.334±0.004	0.351±0.005	0.296±0.007	0.341±0.016	0.217±0.012	0.248±0.082	0.23±0.032	0.179±0.029
ESG (40 clients)	LocalFedMinMax	0.203±0.05	0.152±0.024	0.326±0.016	0.381±0.004	0.304±0.069	0.335±0.045	0.195±0.065	0.171±0.027	0.167±0.086	0.182±0.063
	FedMinMax (ours)	0.257±0.003	0.189±0.009	0.324±0.015	0.351±0.002	0.291±0.004	0.291±0.03	0.231±0.007	0.309±0.008	0.194±0.002	0.158±0.008
SSG (40 clients)	LocalFedMinMax	0.245±0.032	0.119±0.015	0.312±0.029	0.352±0.007	0.298±0.004	0.307±0.066	0.235±0.039	0.226±0.013	0.275±0.025	0.233±0.079
	FedMinMax (ours)	0.258±0.01	0.187±0.005	0.332±0.005	0.351±0.002	0.293±0.007	0.334±0.017	0.216±0.009	0.305±0.009	0.205±0.002	0.154±0.005
PSG (40 clients)	LocalFedMinMax	0.236±0.027	0.14±0.038	0.32±0.025	0.378±0.005	0.296±0.005	0.314±0.048	0.232±0.028	0.214±0.023	0.267±0.022	0.222±0.059
	FedMinMax (ours)	0.261±0.007	0.184±0.007	0.321±0.021	0.351±0.009	0.295±0.003	0.323±0.011	0.22±0.008	0.299±0.011	0.201±0.001	0.154±0.008