Fair Federated Learning

Afroditi Papadaki

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Electronic and Electrical Engineering
University College London

December 11, 2023

To my loving grandparents, *Zaharias* and *Afroditi*.

Declaration

I, Afroditi Papadaki, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

AFRODITI PAPADAKI

Acknowledgements

First, I would like to thank my primary advisor, Miguel Rodrigues. I deeply appreciate the trust and autonomy he granted me to explore and delve into areas of research that genuinely sparked my interest. His insightful feedback, constructive criticism, and willingness to engage in intellectual discourse have been instrumental in refining my research methodology, thinking bigger and aiming higher. I am sincerely grateful for his support, including his responsiveness as an advisor, and the valuable opportunities for growth and development he presented to me.

Throughout my PhD journey, I have had the privilege of collaborating with Guillermo Sapiro, Natalia Martinez, and Martin Bertran. Their invaluable insights, guidance, and contributions have greatly enriched and influenced the quality and trajectory of my work. Natalia and Martin have also played a pivotal role in shaping my mindset towards my PhD journey, fostering boundless curiosity and instilling a tireless work ethic. I deeply appreciate the opportunity to learn from their humility and dedication, and I aspire to embody these qualities throughout my career.

I would like to extend my heartfelt appreciation to my second advisor, Laura Toni, for her contagious optimism, wise counsel, and encouragement over the past few years. She consistently cared for me as part of her own research group, providing invaluable guidance and advice that have greatly contributed to my PhD journey.

I am also grateful to many others who have played a crucial role in shaping my academic experience at UCL. To my lab family, especially Chao Zhou, Maria Novitasari, Martin Ferianc, Zhaoyan Lyu and Zhuo Zhi, who have been with me throughout the majority of my PhD, I express my deep appreciation. Our insightful discussions and their feedback on my work, particularly during my final year,

have been invaluable. Furthermore, engaging in enjoyable and thought-provoking conversations with Alan Guedes, Edoardo Gruppi, Eduardo Pignatelli, Gianluca Anselmi, Nagham Osman (who holds a special place in my heart), Nithin Babu, Pedro Gomes, Sephora Madjiheurem, (our newest addition) Shyam Ramesh and Xinyue Liu has significantly broadened my perspectives and deepened my understanding of various subjects. I also want to acknowledge the contributions of the IT support team and other people from the EEE department, including Izzat Darwazeh and (the unsung hero) Lee Heagney. Their tireless efforts and prompt assistance have been instrumental in creating a seamless research environment for all EEE students.

My gratitude extends to Michael Filippakis and Dimosthenis Kyriazis from the University of Piraeus for fueling my enthusiasm for science and teaching, especially as an undergraduate student.

Lastly, I am grateful to the people closest to me. To my parents, Angeliki and George, I am indebted for their unwavering presence and unconditional love, which continuously motivate me to overcome challenges and pursue my dreams. I want to express my gratitude to my brother, Michalis, for his constant encouragement and unwavering belief in my abilities, and to my uncle Dimitris for celebrating my successes as if they were his own. I would also like to thank my friends Antonio Nascimento, Dimitrios Zarkadas, Elina Kasapantoniou, Giorgos Leventis, Grace Harrison, Nana Karkalou, Olga Maggiorou, Telma Guerreiro and Todd Sproule for their support and countless enjoyable moments we shared.

Lastly, but certainly not least, I am profoundly grateful to my life partner, Nikos, for being my unwavering pillar of support. His kindness, honesty, and goodness have been a constant source of strength and inspiration. Spending time with him has kept me grounded and reminded me of the truly important aspects of life.

Thank you all!!!

Abstract

Federated learning is a collaborative and distributed machine learning approach in which a statistical model is trained to solve an optimization problem using a federation of clients – such as different user devices or organizations – coordinated by a central server. During training, clients share only targeted updates designed to include the minimum information needed for the specific predictive task with the server, not the data itself. These updates are used by the server to improve the global model without directly accessing the clients' data. The server is responsible for aggregating these updates and uses them to improve the global model.

One of the key challenges in such learning settings is ensuring that the trained model is both accurate and unbiased with respect to various population groups that relate to demographics (e.g. gender, disability, sexual orientation or ethnicity). For instance, in the banking sector, federated learning is harnessed to develop more resilient models for credit score prediction, by aggregating information from multiple banks that hold data from different demographic backgrounds in a heterogeneous manner. Therefore, this work addresses federated demographic group fairness in two pragmatic federated learning scenarios.

In the first learning scenario, we study federated (minimax) global group fairness where the target sensitive groups are known but the participating clients may only have access to a subset of the population groups during training. We discuss how the proposed group fairness notion differs from existing federated fairness criteria that impose similar performance across participants instead of demographic groups. We provide an algorithm to solve the proposed problem that enjoys the performance guarantees of centralized learning algorithms. We empirically compare the proposed

approach against other methods in terms of group fairness in various setups, showing that our approach exhibits competitive or superior performance.

In the second setting, we assume that the parties engaging in the federation are unaware of the target demographic groups and their corresponding group labels. To address this issue, we first introduce an objective that allows to learn a Pareto efficient global hypothesis ensuring (worst-case) group fairness. Our objective enables, via a single hyper-parameter, trade-offs between fairness and utility, subject to a group size constraint. The proposed objective recovers existing approaches as special cases, such as empirical risk minimization and subgroup robustness objectives from centralized machine learning. Next, we provide an algorithm to solve in federation a smoothed version of the proposed problem and prove that it exhibits convergence and excess risk guarantees. Our experiments indicate that our approach effectively improves the worst-performing group without unnecessarily hurting the average performance and achieves a large set of solutions with different fairness-utility trade-offs. Finally, we demonstrate that its deployment can be beneficial even in some cases with known demographics.

The methods proposed in this thesis have a generic nature, allowing for their application in various federated learning domains such as medicine, insurance, finance, and college admissions, among others.

Impact Statement

Federated learning gains momentum in various high-stakes decision-making applications and consumer products. For instance, in medical applications such as brain tumour segmentation and hospitalization prediction, federated learning facilitates the development of a model by leveraging information from medical institutions in diverse geographical locations with data from varying demographic backgrounds.

Similarly, consider a scenario where a group of financial institutions aims to create an accurate fraud detection system without sharing sensitive customer information. Through federated learning, the institutions can harness collective intelligence by collaborating to train a model that detects and prevents fraud collectively, while improving privacy by keeping the data decentralized.

Another example is the deployment of federated learning in the development of a resume screening or candidate evaluation model. In this scenario, participating entities possess data from candidates representing diverse sensitive groups. It becomes crucial to eliminate any bias associated with demographic factors to ensure the establishment of a fair and inclusive hiring process that not only promotes diversity but also aligns with regulatory standards.

These real-world instances highlight the direct impact of federated learning on individuals and communities, underscoring the significance of averting biased and fallible decision-making of statistical models. The research included in this thesis endeavours to tackle the demographic group fairness challenges arising in such federated learning paradigms, and provides effective solutions that mitigate bias during the training process in such contexts, thereby advancing fairness and equity in the application of federated learning.

Contents

List of Publications

The work reported in this thesis has been published/submitted as the following papers.

Fair federated learning with known demographics (Chapter 3)

1. Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. (2022a). Minimax demographic group fairness in federated learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 142–159, New York, NY, USA. Association for Computing Machinery

2. Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. (2021). Federating for learning group fair models. In *New Frontiers in Federated Learning: Privacy, Fairness, Robustness, Personalization and Data Ownership at the 35th Conference on Neural Information Processing Systems*, NeurIPS '21, Virtual

Fair federated learning without demographics (Chapters 4 - 5)

1. Papadaki, A., Martinez, N., Bertran, M. A., Sapiro, G., and Rodrigues, M. R. D. (2023). Federated fairness without demographics. *Under Review*

2. Papadaki, A., Martinez, N., Bertran, M. A., Sapiro, G., and Rodrigues, M. R. D. (2022b). Federated fairness without access to demographics. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*

During my PhD, I also contributed to the following publications, the contents of which are either not extensively discussed or not considered in this thesis.

Additional contributions to fairness in centralized machine learning

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

2. Martinez, N. L., Bertran, M. A., Papadaki, A., Rodrigues, M., and Sapiro, G. (2020b). Pareto robustness for fairness beyond demographics. In *Fair AI in Finance at the 34th Conference on Neural Information Processing Systems*, NeurIPS '20, Virtual

Other topics in responsible machine learning

1. Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M., Reeves, G., and Sapiro, G. (2019). Adversarially learned representations for information obfuscation and inference. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 614–623. PMLR

2. Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M. R. D., and Sapiro, G. (2018). Learning representations for utility and privacy: An information-theoretic based approach. In *Workshop on Privacy Preserving Machine Learning (in Conjunction with NeurIPS 2018)*

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AFL	Agnostic Federated Learning
BPF	Blind Pareto Fairness
CNN	Convolutional Neural Network
(C)ML	(Centralized) Machine Learning
DNN	Deep Neural Networks
DP	Demographic Parity
DRO	Distributionally Robust Optimization
EOd	Equal Odds
EOp	Equal Opportunity
ESG	Equal access to Sensitive Groups
FPR	False Positive Rate
FL	Federated Learning
I.I.D.	Independent and Identically Distributed
LHS	Left-Hand Side
MLP	Multi-Layer Perceptron
PSG	Partial access to Sensitive Groups
(R)CVaR	(Relaxed) Conditional Value-at-Risk

RHS	Right-Hand Side
(S)GD	(Stochastic) Gradient Descent
SSG	Access to a Single Sensitive Group
TERM	Tilted Empirical Risk Minimization
TPR	True Positive Rate

List of Mathematical Notations and Symbols

Random variables are denoted using capital Latin letters (e.g., Z) and their values as lowercase letters (e.g., z). We use the lowercase p to annotate distributions. For example, $p(Z)$ is the distribution of random variable Z, $p(Z|K=k)$ is the distribution of Z given that the variable K takes value k, $p(Z,K)$ is the joint distribution of the variables Z and K. We annotate a set of possible values as a capital calligraphic letter (e.g., $\mathscr{Z}$). Finally, we use boldface lowercase letters to denote vectors (e.g., $\boldsymbol{\mu}$) and lowercase letters with subscripts to indicate the component of a vector (e.g., μ_i is the i-th element of vector $\boldsymbol{\mu}$). The following glossary provides symbols and notation that are universally valid across the thesis.

X	Input variable
Y	Target variable
K	Federation's clients variable
A	Group variable of known demographics
G	High-risk group membership variable
$\mathscr{H}$	Hypothesis set
h	A hypothesis from the hypothesis set $\mathscr{H}$
θ	A vector parametrizing hypothesis

Θ	Parametric vector space
ℓ	Loss function
r	Risk function
$\hat{r}$	Empirical risk function
$\boldsymbol{\mu}$	Weighting coefficients of known groups
$\boldsymbol{w}$	Importance weighting coefficients of known groups
ε	Trade-off parameter/ Lower bound of weighting coefficients
ρ	Group size for high-risk group
$\mathbb{R}$	Set of real numbers
$\mathbb{E}$	Expectation operator
Δ^{m-1}	Simplex over $\mathbb{R}^m$

Chapter 1

Introduction

"All social values – liberty and opportunity, income and wealth, and the bases of self-respect – are to be distributed equally unless an unequal distribution of any, or all, of these values is to everyone's advantage."

A THEORY OF JUSTICE, JOHN RAWLS

1.1 Overview and Motivation

Machine learning models are increasingly being deployed to make high-stakes decisions in a range of domains, such as finance, insurance, medical diagnosis, recruitment, and more [Chouldechova and Roth, 2020]. The significance of such applications, coupled with emerging regulations, demand that these decision-making models do not exhibit discriminatory behaviour towards individuals based on their demographic characteristics, so that everyone has similar access to opportunities and resources [Barocas et al., 2019b].

In order to ensure a fair predictive outcome across people with different demographic attributes, it is crucial to develop methods that address and mitigate potential biases and discriminatory patterns of such models. This has been the topic of extensive research in traditional learning settings, where a single central entity trains a model on a dataset containing information from individuals belonging to different sensitive demographic groups, e.g., see [Dwork et al., 2011, Hardt et al., 2016a, Martinez et al., 2020a]. Nevertheless, data representing different demographic groups may be distributed across multiple entities instead of being

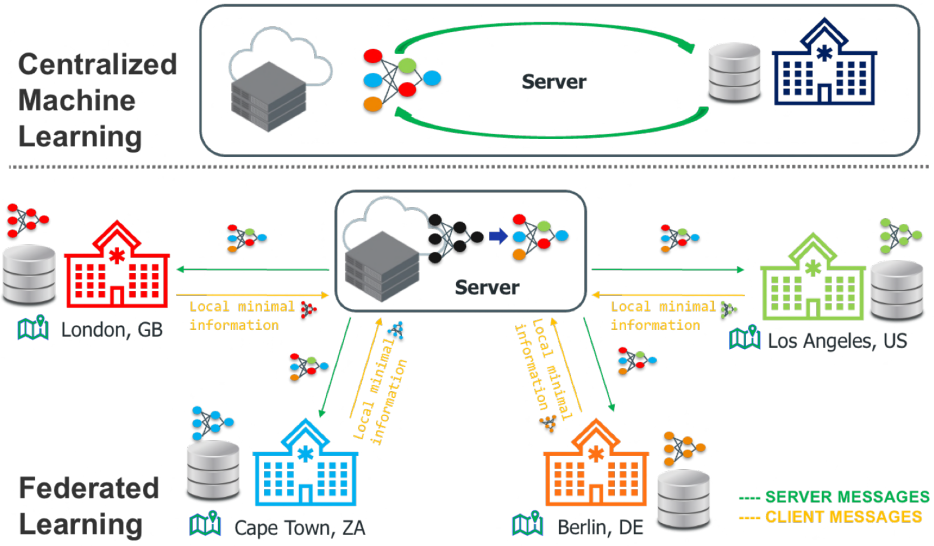

Figure 1.1: Centralized machine learning vs. federated learning. *Top:* In centralized machine learning, data is collected and stored in a single (centralized) location. The central entity (server) hosting the training process, has direct access to the entire dataset at any time during the training process. As a result, the server has the ability to perform the required computations to achieve the desired fairness objectives. *Bottom:* In federated learning, data is distributed across multiple clients or devices, and is kept decentralized. During training, only the required information is exchanged between the clients and the central server, such as the model parameters client associated losses.

stored on a single entity or server, and cannot be directly shared directly or aggregated centrally. This could be due to legal or regulatory constraints (e.g., the EU's General Data Protection Regulation (GDPR) [European-Commission,] and California Consumer Privacy Act (CCPA) [Mancini, 2021] in the US), or for improving an application's performance such as its bandwidth and response time [Steen and Tanenbaum, 2016].

In such scenarios, a model is usually trained using federated learning (FL) [McMahan et al., 2016a] approaches. Federated learning is a distributed learning paradigm, usually coordinated by a central server, that enables multiple entities – referred to as clients – to learn a single global model in a decentralized manner [Konečný et al., 2016a, Konečný et al., 2016b]. The clients participating in the federation do not share their data with one another or with the server; instead, they send focused updates to the server, which then updates the global model and re-distributes it to the clients over multiple rounds or iterations. This approach allows clients with

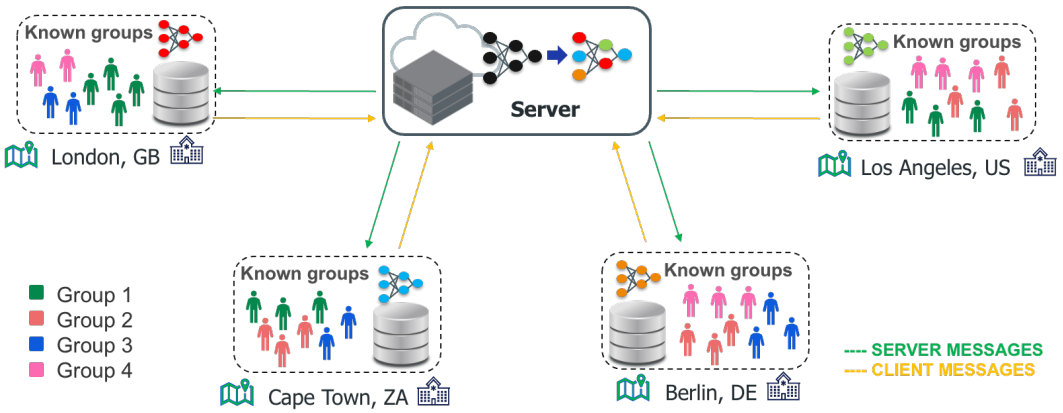

Figure 1.2: Example of a federated learning scenario where the participating entities aim to collectively train a global model that attains group fairness for specific demographic populations. The clients' local group representation varies during training, influenced by factors like geographic location. However, during deployment time, medical centres handle individuals with diverse demographics compared to the training phase.

limited local training data to learn better machine learning models while preserving the privacy of their data. We visualize the two learning settings in Figure 1.1.

Training a model in a federation presents a unique challenge for developing group fair models that can be trained on limited data from different groups and clients. In this thesis, we identify and study two realistic problems arising from such learning paradigms: (a) (global)[1] group fairness with known demographics, and (b) (global) group fairness without access to sensitive groups.

1.1.1 Fairness with Known Demographic Groups in FL

The first problem considers federated learning scenarios where clients possess data from diverse demographics and the goal is to train a global model that is group fair with respect to predefined demographic populations. During the training phase, the local groups available on each client may vary based on factors such as the geographical location and the specific demographics they serve. However, during the testing phase, we assume that there is a common set of groups among the clients.

For instance, consider the context of developing a machine learning model for medical diagnosis in federated learning settings involving multiple hospitals. Each

[1]The term "global" refers to treating any demographic group that is shared across clients as a unified single group.

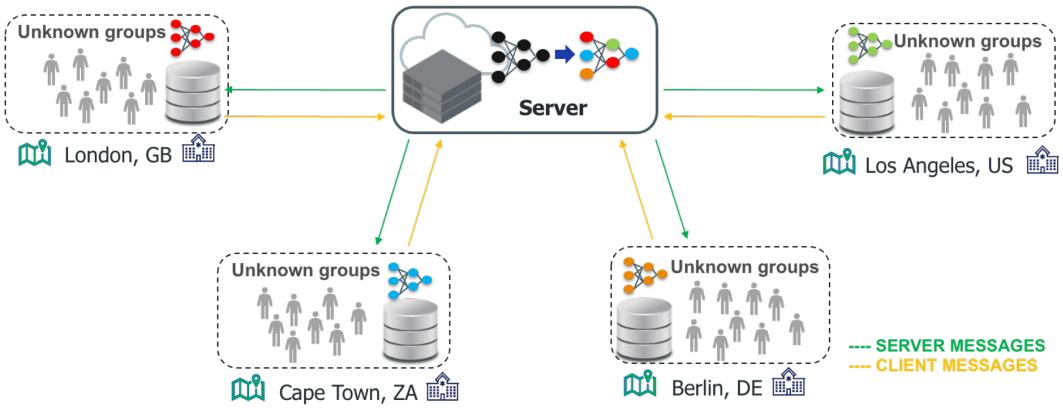

Figure 1.3: Illustration of a federated learning scenario where the clients are unaware of the sensitive demographic groups and the corresponding labels of the patients they handle. This lack of knowledge is primarily attributed to factors such as privacy regulations or the significant cost involved in acquiring accurate information about these demographics. Instead of sharing specific demographic group information (e.g., group risks), clients need to contribute to the collaborative training process by providing alternative types of descriptive information about their local loss distribution.

hospital may have data from specific demographic groups based on their geographical locations and the populations they serve. However, despite the variations in local group representation, the participating clients aim to achieve high performance across patients of all demographics, considering factors such as seasonality or increased patient influx during certain periods (e.g., tourism seasons). This example is illustrated in Figure 1.2.

1.1.2 Federated Group Fairness without Access Demographics

The previous challenge focuses on achieving group fairness in federated learning scenarios where participants are assumed to have knowledge about the sensitive demographic groups during the training, and accurate group memberships are assigned to each data point. However, it is important to recognize that in various federated learning scenarios, it is not always feasible to assume that every client has access to predefined and accurate sensitive groups.

For instance, consider again the previous example where medical institutions collaborate to learn a group fair diagnostic model using federated learning techniques. It is possible that the medical institutions are unaware of the true sensitive groups

or possess medical records that lack sensitive group labels. This could be due to factors such as the high cost of acquiring labels [Gebru et al., 2017], the need for specialized expertise, or privacy regulations like GDPR [European-Commission,] and CCPA CCPA [Mancini, 2021], which limit the collection and use of certain personal information. This scenario is depicted in Figure 1.3.

1.2 Contributions

The preceding section highlights the significance of ensuring demographic fairness in federated learning settings and describes the two major challenges in achieving fair machine learning models in federated learning settings.

The first challenge pertains to achieving (global) group fairness when the sensitive demographic groups are known, while the second challenge involves achieving (global) group fairness without access to the sensitive groups. This thesis focuses on addressing both of these challenges. We summarize the contributions of this thesis as follows:

Fairness with Known Demographic Groups in FL (Chapter 3)

1. We introduce a novel fairness definition for federated learning setting, namely *global group fairness*. We provide a formulation for achieving global group fairness by leveraging the notion of minimax Pareto fairness. The proposed formulation takes into account the possibility that certain clients in the federation may only have access to a subset of the demographic groups during the training phase.

2. We present an optimization algorithm that can be adopted by multiple entities coordinated by a single server to learn a global minimax group fair model and exhibit convergence guarantees. Our algorithm accommodates clients with varying levels of representation for specific groups, whether high, low, or none at all. Furthermore, we demonstrate that the global model obtained through our federated learning approach is equivalent to a model produced by a centralized learning algorithm, ensuring consistency and effectiveness across both frameworks.

3. We perform extensive experiments, comparing the proposed approach against existing state-of-the-art federated learning methods in terms of (global) group fairness in various federated learning setups, confirming that our approach exhibits competitive or superior performance. We also empirically validate the conditions under which the competing approaches yield the same solution as our objective.

Federated Group Fairness without Access Demographics (Chapters 4 - 5)

1. We address the problem of (minimax) Pareto federated group fairness with inhomogeneous and unknown demographics. We introduce a new fairness-aware objective – RCVaR – that allows improving the performance of the high-risk (low-utility) samples, subject to a group size constraint, while ensuring the best possible performance on the remaining samples. To the best of our knowledge, we are the first to address this problem in federated learning settings.

2. We draw connections between the proposed objective and existing ones, such as DRO [Hashimoto et al., 2018] and BPF [Martinez et al., 2021], demonstrating that RCVaR can also be used for learning Pareto subgroup robust models in centralized settings.

3. We then introduce an algorithm – FedSRCVaR – that solves a smoothed approximation of RCVaR in federated learning settings. We establish a number of guarantees associated with this algorithm, including its convergence and excess risk properties, and show that the proposed objective can be easily federalized.

4. Finally, we empirically study the wide range of solutions that can be achieved by our approach through the trade-off parameter for various group sizes. We also empirically compare our method against other relevant baselines in centralized and federated learning settings using real datasets, considering also scenarios with known demographics.

1.3 Thesis Outline

In this chapter, we provide an overview of the thesis, discussing its motivation and the specific problems that it addresses. We outline the main objectives and goals of our research, highlighting the significance and relevance of the chosen topics. The remainder of this document is structured as follows.

In Chapter 2, we begin by presenting an overview of the standard supervised learning problem and its associated objective. We introduce the key fairness definitions that are commonly used in centralized machine learning, and discuss their relevance to the challenges addressed in our research. We explain how the supervised learning problem is extended for federated learning settings and present the different fairness notions that have been proposed specifically for distributed learning settings.

In Chapter 3, we introduce a novel approach to address the challenge of achieving minimax Pareto group fairness in federated learning settings where participating entities may have limited access to population groups during training. We highlight how our objective differs from existing federated fairness criteria. Furthermore, we develop an optimization algorithm – FedMinMax – for solving the proposed fairness problem and establish that it enjoys the convergence and fairness guarantees that are typically associated with centralized learning algorithms. We present empirical results that support our theoretical analysis and demonstrate that our method outperforms existing baselines.

In Chapter 4, we present a new objective, called Relaxed Conditional Value-at-Risk (RCVaR), within the context of centralized learning settings, that allows to learn a Pareto efficient hypothesis ensuring (worst-case) group fairness when demographic groups are unavailable. Our objective allows for trade-offs between fairness and utility through a single hyper-parameter and subject to a protected demographic group size constraint. Additionally, we show that this objective encompasses and extends existing approaches from centralized machine learning, including the widely used empirical risk minimization and subgroup robustness objectives.

In Chapter 5, we address the challenge of achieving federated group fairness without awareness of the existing demographic groups by the parties engaging in

the federation. We extend the objective introduced in Chapter 4 to accommodate this scenario and propose an algorithm that operates within the federation to solve a smoothed approximation of the problem. We provide theoretical guarantees regarding the convergence of the algorithm and its performance in terms of excess risk. We empirically show that it successfully improves the worst-performing group without unnecessarily hurting the average performance and that it achieves a diverse set of solutions with varying fairness-utility trade-offs, allowing practitioners to choose the solution that best aligns with their specific requirements. Furthermore, we highlight that even in cases where demographic groups are known, our approach still delivers improvements.

Finally, in Chapter 6, we draw conclusions based on the findings and contributions of this thesis. We summarize the main results and insights gained from each chapter and highlight their significance in the context of fairness in federated learning. Additionally, we offer final remarks that reflect upon the broader implications and potential impact of our research. We acknowledge open issues and identify areas that could be further explored or improved upon in future work.

Chapter 2

Background

In this chapter, we provide a comprehensive introduction to the standard supervised learning problem and its underlying objective. We recognize the importance of fairness considerations in machine learning and delve into the key fairness definitions that are widely employed in centralized machine learning. We discuss the significance of these definitions in relation to the challenges we aim to tackle in our research. Moreover, we explore the extension of the supervised learning problem to federated learning settings, where data is distributed across multiple clients. Within this context, we present the various fairness notions that have been proposed specifically for distributed learning settings. By examining these fairness notions, we lay the groundwork for our subsequent research, focusing on addressing demographic fairness challenges in the federated learning paradigm.

2.1 Supervised Machine Learning

Machine learning can be categorized into different types: supervised, semi-supervised, unsupervised, and reinforcement learning. In this thesis, our focus is on supervised classification learning, which is a type of machine learning where a statistical model is trained using a dataset that contains labeled data examples. Each data example consists of input features and their corresponding target labels. The goal of supervised machine learning is to develop a predictive model that can accurately map the input features to their respective target labels. This is achieved by optimizing a loss function – or equivalently a cost function – that measures the

discrepancy between the predicted labels generated by the model and the actual target labels in the training dataset. Once the model is trained, it can be applied to make predictions on new, unseen data by utilizing the learned mapping from input features to target labels.

We can formally express a supervised classification setting as follows. Let the random variables $(X,Y) \in \mathscr{X} \times \mathscr{Y}$, where X represents the input features and Y represents the target variable. These random variables are drawn from a joint distribution $p(X,Y)$. The specific form of X can vary and can include vectors, abstract objects like images, or any other suitable representation for the problem at hand. We denote the realizations of X and Y as $x_i \in \mathscr{X}$ and $y_i \in \mathscr{Y}$, respectively.

To quantify the discrepancy between predicted and actual labels, we introduce a loss function $\ell : \Delta^{|\mathscr{Y}|-1} \times \Delta^{|\mathscr{Y}|-1} \to \mathbb{R}_+$, where Δ represents the probability simplex over $\mathbb{R}^{|\mathscr{Y}|}$. The loss function measures the dissimilarity between the predicted label distribution and the true label distribution.

The objective of supervised classification learning is to learn a classifier $h \in \mathscr{H}$ that can effectively predict the target variable $Y \in \mathscr{Y}$ based on the input features $X \in \mathscr{X}$. The classifier h is chosen from a hypothesis space $\mathscr{H}$, which contains a set of possible classifiers. The goal is to find the classifier h that minimizes the expected loss over the joint distribution $p(X,Y)$. The problem of expected risk minimization is formulated as

$$h^* \in \arg\min_{h \in \mathscr{H}} \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)} [\ell(h(X),Y)]. \tag{2.1}$$

It is important to acknowledge that in practical applications, we have access to a finite dataset $D = \{x_i, y_i\}_{i=1,\ldots,n}$ rather than the true data distribution. Consequently, instead of dealing with the expected risk as shown in Eq. 2.1, we rely on the empirical risk computed on the dataset D and aim to solve the problem of empirical risk minimization

$$\min_{h \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i). \tag{2.2}$$

To address the optimization problem stated in Eq. 2.2, various techniques are commonly employed, such as gradient descent (GD), stochastic gradient descent

(SGD), minibatch-stochastic gradient descent (minibatch-SGD), or their respective variants [Boyd and Vandenberghe, 2004].

2.2 Fairness in Centralized Machine Learning

2.2.1 Group Fairness with Demographics

In the field of centralized machine learning, there are various fairness criteria that aim to capture different notions of fairness and non-discrimination. In this section, we discuss the three primary categories of statistical fairness: (a) independence, (b) separability, and (c) minimax group fairness. For other notions studied in centralized machine learning, such as sufficiency, individual fairness, and counterfactual fairness, we refer the reader to the following works [Dwork et al., 2011, Kusner et al., 2017, Barocas et al., 2019a, Caton and Haas, 2020, Gupta and Kamble, 2021, Makhlouf et al., 2021]. While these works may not directly relate to the scope of this thesis, they contribute to the overall understanding of fairness considerations and solutions explored in machine learning.

To address fairness considerations, we introduce an additional variable $A \in \mathscr{A}$ to represent sensitive demographic groups. The random variables (X, Y, A) are jointly distributed according to $p(X,Y,A) = p(A) \cdot p(X,Y|A)$. Here, $p(A)$ represents the prior distribution of the known demographic groups, and $p(X,Y|A)$ represents the conditional distribution of the data given the demographic groups. Furthermore, we denote the outcome variable predicted by the model h based on the input variable X as $\hat{Y}$. The predicted outcome $\hat{Y}$ belongs to the set $\mathscr{Y}$, which represents the possible target label values.

2.2.1.1 Independence

The independence criterion is one of the most popular fairness criteria, which states that the sensitive group variable A should be statistically independent of the outcome $\hat{Y} \in \mathscr{Y}$. The formal definition is given in Definition 2.1.

Definition 2.1 (Independence)**.** *The decision variable $\hat{Y}$ and group variable A satisfy independence if $\hat{Y} \perp A$, or equivalently if $p(\hat{Y}|A) = p(\hat{Y})$.*

In binary classification scenarios, the concept of independence is commonly referred to as demographic parity, statistical parity or disparate impact [Dwork et al., 2011, Barocas et al., 2019b]. It can be defined as the condition where the probability of the outcome $\hat{Y}$ being equal to 1 is the same across all groups $a, a' \in \mathscr{A}$. We offer the formal definition in Definition 2.2.

Definition 2.2 (Demographic Parity [Dwork et al., 2011])**.** *The decision variable $\hat{Y}$ and the demographic group variable A satisfy demographic parity, if $p(\hat{Y} = 1|A = a) = p(\hat{Y} = 1|A = a')$, with $a, a' \in \mathscr{A}$.*

To allow for some flexibility in the fairness constraint, a relaxation is introduced with the introduction of a positive slack variable ξ. This relaxation allows for a certain degree of deviation from strict equality in the fairness criterion. Specifically, it requires that $p(\hat{Y} = 1|A = a) \geq p(\hat{Y} = 1|A = a') - \xi$ for all groups $a, a' \in \mathscr{A}$. This inequality ensures that the probability of the predicted outcome being 1, given a particular sensitive group a, is greater than or equal to the probability of the predicted outcome being 1, given another sensitive group a', with an allowed deviation of ξ. Alternatively, this constraint can be expressed as $\frac{p(\hat{Y}=1|A=a)}{p(\hat{Y}=1|A=a')} = 1 - \xi$.

Another way to characterize the independence condition stated in Definition 2.1 is through the concept of mutual information. If the mutual information between the sensitive group variable A and the decision variable $\hat{Y}$ is zero, denoted as $I(A;\hat{Y}) = 0$, it implies that there is no information flow between these variables. In the relaxation of this constraint, we allow for a positive value ξ that indicates the maximum allowed mutual information between A and $\hat{Y}$. Therefore, the relaxation can be defined as $I(A;\hat{Y}) \leq \xi$, indicating that the information flow between the sensitive group and the decision variable is upper bounded by ξ.

Independence criterion assumes that all groups should be treated equally in terms of access to opportunities and resources [Barocas et al., 2019b], advocating for proportional allocation, and is easy to ensure [Calders et al., 2009, Zliobaite, 2015, Zafar et al., 2017]. Nevertheless, decisions produced by a classifier that adheres to the independence criterion can still exhibit undesirable properties, especially in the common scenario where the group conditional priors are different, i.e.,

$p(Y = 1|A = a) \neq p(Y = 1|A = b)$, $\forall a,b \in \mathscr{A}$. For example, in the context of university admissions, enforcing independence may lead to unfair outcomes. One drawback is that it could result in the acceptance of multiple unqualified applicants solely to ensure that the proportion of positive outcomes aligns with each group's representation. This compromises the selection of the most qualified candidates.

Additionally, if the predicted outcome $\hat{Y}$ is correlated with the sensitive group A, enforcing independence can negatively impact utility performance, since the optimal predictor $\hat{Y} = Y$ is not allowed [Hardt et al., 2016a]. For instance, in the case of hiring a senior software engineer, the optimal candidate would ideally possess substantial experience in software engineering. However, if independence is strictly enforced, the consideration of an applicant's experience may be limited or ignored, potentially leading to suboptimal hiring decisions.

2.2.1.2 Separation

The criterion of separation is designed to address the limitations of the independence criterion. Unlike demographic parity, separation aims to minimize any discriminatory impact that the correlation between $\hat{Y}$ and A may have on decision-making processes. This is accomplished by ensuring that the predicted outcome $\hat{Y}$ depends on A but only through the target variable Y, rather than directly affecting the prediction.

Definition 2.3 (Separation)**.** *The decision variable $\hat{Y}$ and group variable A satisfy separation if they satisfy conditional independence, that is $\hat{Y} \perp A|Y$.*

The fairness concept of separation, also known as conditional independence, permits the perfect predictor where $\hat{Y} = Y$ [Hardt et al., 2016a]. In the context of binary classification, separation can be defined as the constraint described in Definition 2.4 for the relevant groups.

Definition 2.4 (Equal Odds (EOd) [Hardt et al., 2016a])**.** *The decision variable $\hat{Y}$ and the demographic group variable A satisfy equal odds, if $p(\hat{Y} = 1|A = 0, Y = y) = p(\hat{Y} = 1|A = 1, Y = y)$, $\forall y \in \{0,1\}$.*

Equal Odds aims to achieve parity in error rates among different groups. This means that each group has an equal chance of being correctly recognized as positive

instances (true positive rate – TPR) and an equal chance of being incorrectly assigned a positive outcome (false positive rate – FPR). A commonly used relaxation of Equal Odds, known as Equal Opportunity is introduced in Definition 2.5. Unlike Equal Odds, which aims for non-discrimination across all outcome groups, Equal Opportunity focuses specifically on non-discrimination within the advantaged outcome group, thus allowing for more utility.

Definition 2.5 (Equal opportunity (EOp) [Hardt et al., 2016a])**.** *A binary classifier satisfies equal opportunity w.r.t. A, if its decision is independent of the demographic group, i.e.,* $p(\hat{Y} = 1|A = a, Y = 1) = p(\hat{Y} = 1|Y = 1), \forall a \in \mathscr{A}$.

Both EOd and EOp are relatively easier to achieve compared to notions of independence since they are more aligned with improving utility. For example, a completely accurate classification model automatically satisfies the requirements for EOd [Barocas et al., 2019b]. In many cases, achieving separation as a fairness objective can be addressed as a post-processing step by adjusting the decision threshold of a trained model to meet the desired fairness constraints [Hardt et al., 2016a, Iosifidis et al., 2020, Awasthi et al., 2020]. However, it's important to note that such post-processing methods typically require access to sensitive group memberships during the testing phase [Agarwal et al., 2018], which can raise privacy concerns and lead to unintended consequences [Chen et al., 2018a]. The authors in [Li and Liu, 2022] proposed an in-process mechanism for sample re-weighting during training, while other approaches focus on data preprocessing techniques to balance the data per group before training the model [Iosifidis et al., 2020, Yu, 2021].

2.2.1.3 Minimax Group Fairness

Minimax group fairness [Rawls, 2001, Martinez et al., 2020a, Diana et al., 2020], also known as maximin group fairness from a utility standpoint, is a fairness criterion aiming to address the disparity in performance among sensitive groups by focusing on improving the outcomes for the worst-performing group without unnecessarily degrading the performance of other groups.

This criterion is particularly relevant in high-stakes decision-making scenarios where it is crucial to avoid disadvantaging well-performing groups. For instance, consider the application of predictive models in the criminal justice system, where these models are used to assess the likelihood of a convicted individual re-offending, which in turn will influence the sentencing severity. In such cases, it is essential to ensure that the predictive model improves outcomes for individuals in the worst-performing group without harming the utility for well-performing groups, unless necessary. Furthermore, fairness definitions such as demographic parity, equality of odds, or equality of risks can sometimes conflict with each other, leading to sub-optimal outcomes where certain groups are harmed without any improvement for other groups, as discussed in previous works [Kleinberg et al., 2016, Chen et al., 2018a, Barsotti and Kocer, 2022].

We introduce the formal definition for minimax fairness in Definition 2.6.

Definition 2.6 (Minimax Fairness [Martinez et al., 2020a])**.** *A predictive model h^* is minimax group fair, if it minimizes the worst performing demographic group risk, i.e.,*

$$h^* \in \arg\min_{h\in\mathscr{H}} \max_{a\in\mathscr{A}} r_a(h) \coloneqq \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y|A=a)} [\ell(h(X),Y)]. \tag{2.3}$$

We note that in practical scenarios, the true data distribution is inaccessible, and instead, we leverage the empirical risk estimated from a finite dataset $D = \{x_i, y_i\}_{i=1,\dots,n}$. The empirical minimax fairness objective is described as

$$\min_{h\in\mathscr{H}} \max_{a\in\mathscr{A}} \hat{r}_a(h) \coloneqq \frac{1}{n_a} \sum_{i=1}^{n_a} \ell(h(x_i), y_i). \tag{2.4}$$

Minimax fairness can be attained by employing alternating optimization techniques. This involves alternating between projected gradient ascent or multiplicative weight updates to optimize the weighting coefficients while considering the model, and utilizing stochastic gradient descent to optimize the model while considering the weighting coefficients [Chen et al., 2017b, Diana et al., 2020, Martinez et al., 2020a].

2.2.2 Fairness without Sensitive Groups

Next, we discuss two approaches in centralized machine learning dealing with group fairness without explicit demographics: (a) proxy fairness, and (b) subgroup robustness.

2.2.2.1 Proxy-Fairness

Several works [Gupta et al., 2018, Zhang, 2018] have focused on scenarios where the group annotations are either unknown or inaccurate. These works propose methods to construct a proxy variable that can serve as a substitute for the true sensitive group variable. Once a suitable proxy variable is obtained, the group fairness approaches discussed in Section 2.2.1 can be applied to ensure fairness.

One notable example is the Bayesian Improved Surname Geocoding (BISG) method [Elliott et al., 2008, Elliott et al., 2009], which utilizes data from the US decennial census to estimate the race membership based on surnames and information about geographical location (e.g., ZIP code). The method proposed in [Kilbertus et al., 2017] involves identifying suitable proxy groups by assuming and analyzing causal relationships within the available data. Once identified, they leverage these proxy groups to mitigate or eliminate problematic causal relationships.

Nevertheless, many studies, including [Zhang, 2016, Kallus et al., 2019, Chen et al., 2019], have emphasized the potential negative consequences of utilizing inappropriate proxy groups. These works highlight that such approaches can perpetuate or even amplify existing inequalities in performance across groups. Moreover, proxy methods require knowledge of the true demographics though the samples group labels are considered unavailable, which is hard to obtain for many applications.

2.2.2.2 Subgroup Robustness

Another line of research tackles more complex learning scenarios characterized by the lack of knowledge about both the demographic groups and group labels through (sub)group robustness.

The most popular approach leverages distributionally robust optimization

(DRO) [Ben-Tal et al., 2013, Lam and Zhou, 2015, Namkoong and Duchi, 2016, Hashimoto et al., 2018, Duchi et al., 2020]. DRO aims to address this problem by minimizing the worst-case loss over all distributions in a ball around the data distribution. This method has been studied for different types of divergences, including chi-squared (χ^2), empirical likelihood, and Kullback-Leibler divergences. We define the DRO objective using the χ^2-divergence as follows.

Definition 2.7 (Distributonally Robust Optimization [Hashimoto et al., 2018, Duchi et al., 2020])**.** *Let $D_{\chi^2}(Q \parallel P) = \int \left(\frac{dQ}{dP} - 1\right)^2 dP$ denote the χ^2-divergence between two probability distributions Q and P. Let also $\mathscr{B}(p(X,Y),r) = \left\{q(X,Y) \ll p(X,Y) : D_{\chi^2}(q(X,Y) \parallel p(X,Y)) \leq r\right\}$ be the chi-squared ball around the data probability distribution $p(X,Y)$ of radius r. A predictive model h^*achieves distributional robustness, if it optimizes the worst case loss over all r-perturbations around data distribution $p(X,Y)$, that is*

$$h^* \in \min_{h \in \mathscr{H}} \max_{q(X,Y) \in \mathscr{B}(p(X,Y),r)} \mathbb{E}_{(X,Y) \sim q(X,Y)} \left[\ell(h(X),Y)\right]. \tag{2.5}$$

The authors in [Hashimoto et al., 2018] proposed a gradient-descent-based optimization procedure that minimizes the empirical dual objective of Eq. 2.5, expressed as

$$\min_{h \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} (\ell(h(x_i),y_i) - c)_+^2, \tag{2.6}$$

that takes as an input a predefined and fixed dual variable c.

Similarly, our approach called blind Pareto fairness (BPF) presented in [Martinez et al., 2021], aims to learn a model that minimizes the worst-case risk on any possible group distribution formulated by the training data, subject to a certain group size, while ensuring that the produced model is properly Pareto efficient [Miettinen, 2012].

Definition 2.8 (Blind Pareto Fairness [Martinez et al., 2021])**.** *Let the random variable G denote the group with the high-risk samples. We let $G = 1$ denote data belonging to the worst-performing group and $G = 0$ the remaining data. A predictive*

model h^* is blind Pareto fair, if it optimizes the objective

$$h^* \in \min_{h\in\mathscr{H}} \max_{\substack{p(G=1|X,Y) \\ \text{s.t. } p(G=1)=\rho \\ p(G=1|X,Y)>0}} \mathbb{E}_{(X,Y)\sim p(X,Y)} \left[\frac{p(G=1|X,Y)}{p(G=1)} \ell(h(X),Y) \right], \tag{2.7}$$

where ρ is the size of the high-risk group, i.e., $p(G=1)=\rho$.

To address the problem of blind Pareto fairness in Definition 2.8, we focused on the empirical importance weighted problem stated as

$$\min_{h\in\mathscr{H}} \max_{\boldsymbol{w}\in Q_\rho} \sum_{i=1}^{n} \frac{w_i}{n\rho} \ell(h(x_i),y_i), \tag{2.8}$$

where $Q_\rho = \{\boldsymbol{w} : w_i \in [\varepsilon,1] \sum_{i=1}^{n} \frac{w_i}{n\rho} = 1\}$ and propose a variant of projected gradient ascent-descent algorithm to solve it. Other approaches such as in [Lahoti et al., 2020, Sohoni et al., 2020] rely on a neural network or a clustering model to discover the worst performing group and aim to improve fairness by leveraging this group information.

The aforementioned methods optimize the worst-case risk, which naturally emerges as an objective for ensuring minimax fairness (see the formal definition in Section 2.2.1.3). In scenarios where the demographics are unknown, fairness is evaluated based on the utility experienced by the subset of individuals who are worst-served, rather than focusing on performance differences between groups. It is worth noting that subgroup robustness, in addition to group fairness, has found applications in various other domains such as model regularization [Namkoong and Duchi, 2016] and defence against adversarial attacks [Sinha et al., 2018].

2.3 Supervised Federated Learning

When considering federated learning settings, a categorical variable $K \in \mathscr{K}$ is introduced to represent the clients participating in the federation. Thus, we consider problems involving at least the random variables (X,Y,K) that are jointly distributed according to $p(X,Y,K) = p(K) \cdot p(X,Y|K)$, where, $p(K)$ reflects the proportion of data samples contributed by each client relative to the total number of data samples

and $p(X,Y|K)$ denotes the distribution of the input and target variables conditioned on the client. The average utility problem in federated learning is formulated as

$$\min_{h\in\mathscr{H}} \sum_{k\in\mathscr{K}} p(K=k) r_k(h), \tag{2.9}$$

where $r_k(h) = \mathop{\mathbb{E}}_{X,Y\sim p(X,Y|K=k)}[\ell(h;X,Y)]$ is the local objective or risk on a client k. Usually, the formulation of the local objective is consistent across clients, ensuring a unified objective for model training. However, the local data distribution $p(X,Y|K=k)$ can vary among clients, reflecting the heterogeneity of the data they possess.

In practical scenarios, the true local data distribution is inaccessible, and instead, clients are limited to a finite local dataset $D_k = \{x_i^k, y_i^k\}_{i=1,\ldots,n_k}$. Hence, we rely on the empirical risk estimated on the dataset D and the utility optimization objective becomes

$$\min_{h\in\mathscr{H}} \sum_{k\in\mathscr{K}} \frac{n_k}{n} \hat{r}_k(h) := \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(h(x_i^k), y_i^k). \tag{2.10}$$

There are two widely used algorithms for tackling the optimization problem described in Eq. 2.10: (a) Federated Stochastic Gradient Descent (FedSGD) and (b) Federated Averaging (FedAVG). Originally proposed in [McMahan et al., 2016b], both FedSGD and FedAVG draw inspiration from parallel SGD techniques [Chen et al., 2017a]. Both approaches involve applying local SGD to a randomly selected subset of clients during each communication round. In FedSGD, each client performs a single optimization step based on its local data. On the other hand, FedAVG extends this approach by allowing clients to perform multiple local optimization steps (usually denoted as $\tau > 1$) per communication round.

While FedAVG addresses challenges related to the limited communication capabilities of clients in federated learning, it is important to note that performing multiple local updates and reducing server-client communication can pose challenges for achieving good global model performance in scenarios where the data is non-I.I.D., (e.g., when clients are geographically distributed or possess data from different time periods) [Kairouz et al., 2019].

2.4 Fairness in Federated Learning

In the context of federated learning, there are various fairness notions that have been explored in the literature, building upon the existing concepts of group fairness.

For these scenarios, we let the random variables (X,Y,A,K) be drawn from the joint distribution of the variables in is expressed as $p(X,Y,A,K) = p(K) \cdot p(A|K) \cdot p(X,Y|A,K)$. Here, $p(K)$ denotes the prior distribution over the clients, $p(A|K)$ captures the distribution of the groups conditioned on the client, while $p(X,Y|A,K)$ represents the distribution of the input and target variables conditioned on the group and client.

2.4.1 Client Fairness

The field of fair federated learning has primarily focused on the concept of *client-fairness*. Research works in this area propose methods and techniques that ensure models trained through federated learning exhibit comparable performance across different clients (i.e., $A \equiv K$). This notion is depicted in Figure 2.1.

A popular approach is agnostic federated learning (AFL) [Mohri et al., 2019], whose aim is to learn a model that optimizes the performance of the most disadvantaged client (or cluster of clients). We provide the formal definition of AFL in Definition 2.9, which is essentially a restatement of Definition 2.6, where the group variable A is replaced with the clients variable K.

Definition 2.9 (Minimax Client Fairness [Mohri et al., 2019])**.** *A hypothesis h^* is minimax client fair, if it minimizes the worst performing client (or cluster of clients), as follows*

$$h^* = \arg\min_{h \in \mathscr{H}} \max_{k \in \mathscr{K}} r_k(h), \tag{2.11}$$

where $r_k(h) = \mathop{\mathbb{E}}_{X,Y \sim p(X,Y|K=k)} [\ell(h;X,Y)]$ is the expected client risk.

The authors of [Mohri et al., 2019] propose a stochastic mirror-prox-based algorithm, which combines the mirror descent method with proximal updates [Nemirovski et al., 2009, Juditsky et al., 2011], to address the optimization problem described in Eq. 2.11 in the finite data regime. In this setting, each client possesses a dataset $D_k = \{x_i^k, y_i^k\}_{i=1,\ldots,n_k}$, and empirical minimax client fairness objective is

stated as

$$\min_{h\in\mathscr{H}} \max_{k\in\mathscr{K}} \hat{r}_k(h) := \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(h(x_i^k), y_i^k). \tag{2.12}$$

Extensions of AFL [Deng et al., 2020, Ro et al., 2021] improve its communication efficiency by enabling clients to perform multiple local optimization steps. Another FL approach proposed in [Li et al., 2020b], uses an extra fairness constraint to flexibly control performance disparities across clients. Similarly, tilted empirical risk minimization [Li et al., 2021] uses a hyperparameter called tilt to enable fairness or robustness by magnifying or suppressing the impact of individual client losses. FedMGDA+ [Hu et al., 2020] is an algorithm that combines minimax optimization coupled with Pareto efficiency [Mas-Colell et al., 1995] and gradient normalization to ensure fairness across users and robustness against malicious clients.

GIFAIR-FL [Yue et al., 2021] incorporates a regularization term that penalizes the spread in the aggregated loss, aiming to achieve uniform performance across all participating entities. Other works, such as FJORD [Horváth et al., 2021] and FedPrune [Munir et al., 2021], address the challenge of fairness across clients with varying hardware computational capabilities. They allow participants to train sub-models of the original deep neural network (DNN) and contribute to the global model, ensuring fairness while accommodating different computational resources. The authors in [Wang et al., 2021b] remark that unfairness across clients can arise from conflicting gradients, which significantly impact the performance of certain clients, and propose a method for detecting and mitigating these conflicts, thereby promoting client fairness.

A considerable body of client fairness research, such as [Mohri et al., 2019, Deng et al., 2020, Hu et al., 2020, Reisizadeh et al., 2020, Pillutla et al., 2021, Sharma et al., 2023], focuses on developing models that promote fairness across clients, by prioritizing the client (or cluster of clients) with the lowest performance during optimization. This is also referred to as client robustness. While these efforts are valuable for enhancing fairness across clients, our formal findings presented in [Papadaki et al., 2022a] highlight that achieving fairness across clients does not automatically guarantee fairness across different demographic populations within

client distributions, unless certain conditions are satisfied. Specifically, one such condition is that each participant's data should exclusively represent a single demographic group. Our finding suggests that fairness considerations must extend beyond client-level fairness and take into account the demographic composition of the data.

2.4.2 Within-Client Fairness

Recent research studies, including [Cui et al., 2021, Zhang et al., 2021], have proposed methods to achieve group fairness within individual clients, referred to as *within-client fairness*, as illustrated in Figure 2.1. These approaches specifically target the groups that are available within each client during both the training and testing phases.

One such approach is group-distributional robust optimization (G-DRFA) [Zhang et al., 2021], which focuses on optimizing the performance of the worst-performing demographic groups by learning weighting coefficients for each local group. Another approach, called fairness-constrained federated learning (FCFL) [Cui et al., 2021], aims to improve the performance of the worst-performing client while ensuring a certain level of local group fairness defined by each client. This is achieved through gradient-based constrained multi-objective optimization. Complementary works, such as [Du et al., 2020, Zeng et al., 2021, Du and Wu, 2021, Chu et al., 2021] have proposed methods that target both client and within-client fairness.

Most of these approaches operate under the assumption that each participating client holds data from all the groups that are involved in the testing phase [Cui et al., 2021, Chu et al., 2021]. However, this assumption may not reflect the realistic scenario where different clients have access to different subsets of groups during the training phase. Moreover, it allows clients to have different local fairness objectives that are independent of each other [Wang et al., 2021b], so there are no formal guarantees. Each client aims to optimize fairness within its own data, but the desired local fairness guarantees may not be fully met due to limited access to the overall population. This can lead to disparities in fairness across different clients. Furthermore, existing approaches often treat each demographic group available to a

client as a separate sensitive group (i.e., $A \equiv A|K$), regardless of whether the same group exists in other clients [Du and Wu, 2021, Zhang et al., 2021]. This can result in poor generalization performance, as we will demonstrate in Chapter 3.

2.4.3 Global Group Fairness

Recently, the notion of *global group fairness* has been studied in [Papadaki et al., 2021, Papadaki et al., 2022a, Gálvez et al., 2022, Hu et al., 2022]. Unlike within-client fairness approaches, which focus on fairness within each individual client, global group fairness aims to achieve fairness across the groups available within the distributions of all participating clients and produce a solution that is comparable to centralized settings, where all data is available centrally. This concept is visually presented in Figure 2.1.

Our work, FedMinMax [Papadaki et al., 2021, Papadaki et al., 2022a], proposes and addresses global group fairness in federated learning. Specifically, we propose a federated learning method to achieve minimax properly Pareto group fairness across all groups included in the clients' data distributions, irrespective of how these groups are represented within individual clients, while maintaining the utility performance on other groups. In addition, other studies, such as [Gálvez et al., 2022], propose methods for ensuring privacy and global group fairness by constraining the difference in group losses to satisfy fairness notions such as demographic parity, equal opportunity, or equal odds. Finally, the work presented in [Hu et al., 2022] extends the concept of bounded group loss (BGL) [Agarwal et al., 2019] to the federated learning setting. This approach aims to enforce global group fairness by incorporating a static bias term into the group-conditional loss terms. This approach can be reduced to FedMinMax (ours in [Papadaki et al., 2022a]) for specific hyperparameter values in their objective function, as discussed in [Hu et al., 2022].

In Section 3, we provide more comprehensive details about our work, FedMinMax, and elaborate on how our approach achieves global group fairness in federated learning. We summarize the three main types of fairness in Figure 2.1.

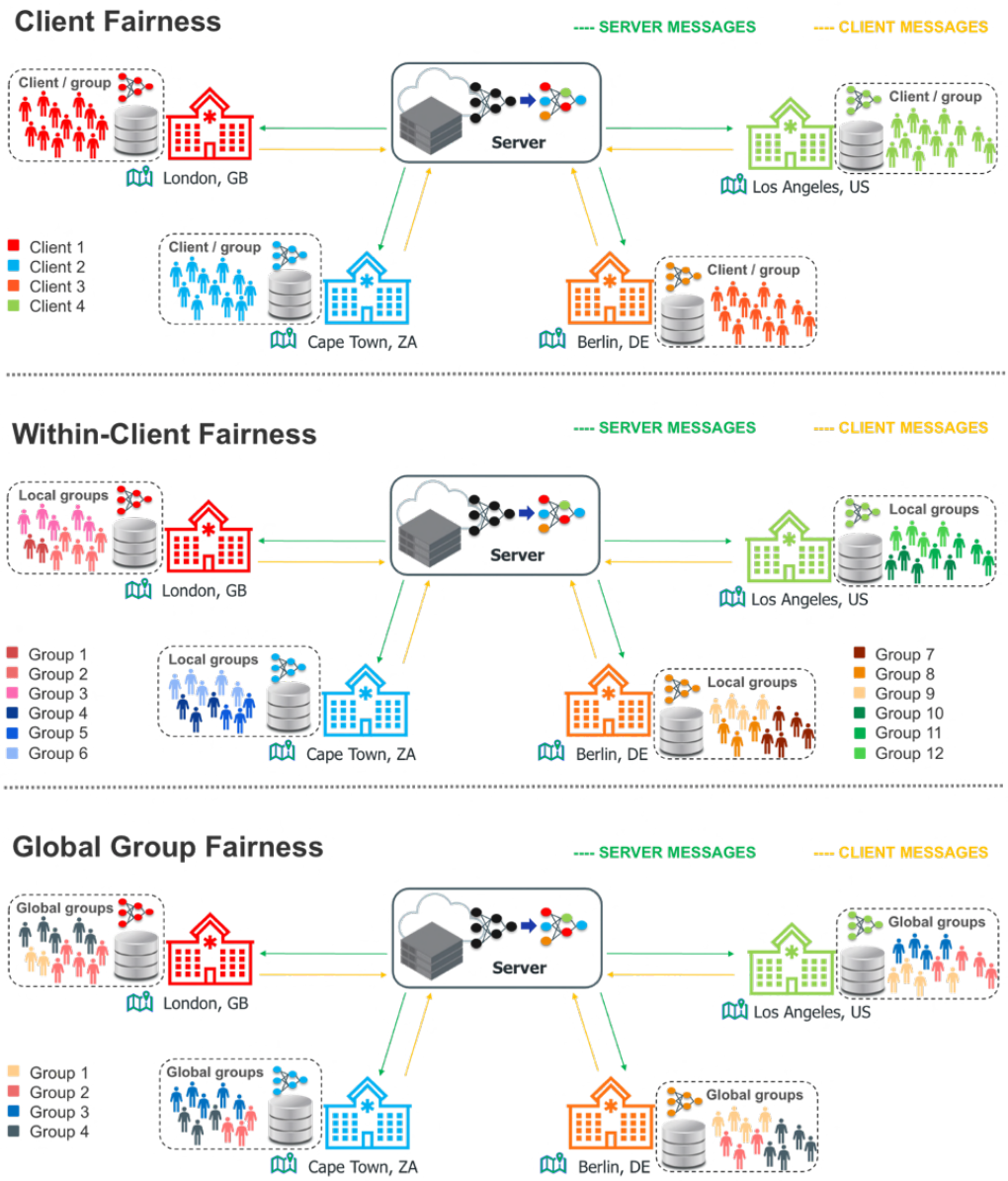

Figure 2.1: Illustration of the three main group fairness notions in federated learning. *Top:* Client fairness aims to achieve comparable performance across the different participating clients (i.e., $A \equiv K$). *Middle:* Within-client fairness targets local fairness across the demographic groups that are available within each client. Often, the local sensitive groups available to a client are considered distinct groups (i.e., $A \equiv A|K$). *Bottom:* Global group fairness aims to achieve fairness across the groups available within the union of clients distributions, regardless of local group representation, and produce a solution that is comparable to centralized settings, where all data is available centrally.

2.4.4 Other Types of Fairness in Federated Learning

Finally, we provide other relevant literature and references in the broader field of fairness in federated learning settings. These works do not directly align with the specific concepts and contributions of this thesis but can provide additional insights into fairness considerations in federated learning. Some notable references in this area include:

Fairness through Personalization: In the pursuit of accommodating competing client fairness objectives, various approaches advocate for the adoption of personalized local models instead of a single global model [Divi et al., 2021].

One such approach is Ditto [Li et al., 2020a], which introduces a lightweight regularization method for learning personalized local models within the framework of federated learning. The method involves training the global model and the local personalized models in an alternating fashion, allowing for customized models that align with individual client fairness requirements. Another approach, lp-proj [Lin et al., 2022], leverages L_p-regularization and low-dimensional random projection to project local models into a shared-and-fixed low-dimensional random subspace. By employing infimal convolution to control the deviation between the reference model and the projected local models, lp-proj facilitates the realization of personalized models in the federated learning setting. To address the challenge of high local training costs associated with personalized federated learning, Personalization-aware Federated Learning (PaFL) [Iacob et al., 2023] allows clients to forego the maintenance of a local model instance by incorporating a personalization term in the local objective. This approach reduces the computational burden on clients while still enabling personalization in the federated learning process.

While fairness through personalization gained attention in the field of federated learning for addressing diverse client fairness objectives, it is important to note that our thesis focuses on scenarios where clients do not have access to sensitive testing populations but rather different subsets of groups during the training phase. Nonetheless, the federated learning frameworks proposed in Chapters 3 and 5 can integrate and leverage the aforementioned approaches, to allow any clients in the

federation to benefit of additional personalization.

Collaborative Fairness: Collaborative fairness [Zhang et al., 2020, Lyu et al., 2020, Nagalapatti and Narayanam, 2021, Fan et al., 2021], introduces the concept of compensating each client based on their contribution to the utility task of the global model. This approach aims to incentivize high-contributing clients to actively participate in the federated learning process by offering them larger rewards, while discouraging free-riding behaviour through lower rewards [Lyu et al., 2020]. While collaborative fairness addresses the issue of client fairness and promotes active client participation, it is important to consider its potential impact on fairness across demographic groups. It is worth noting that in some cases, such approaches may inadvertently penalize clients that have access to the worst-performing demographic groups. This can result in an even more unfair global model, where the performance of disadvantaged demographic groups is further compromised.

Fairness without Demographics: To our knowledge, the only work that considers federated learning scenarios where demographic data cannot be leveraged is [Juárez and Korolova, 2022]. This work assumes that the collection of sensitive groups is known in advance, but this information is not utilized due to privacy considerations. Thus, they propose the use of local differential privacy mechanisms to protect the sensitive information of individual participants while still allowing for the utilization of group membership information to train fair machine learning models. In contrast to this approach, in Chapter 5 we explicitly focus on a more challenging case where the local sensitive populations and the corresponding labels are completely unknown to the clients. Hence, no group information can be incorporated into the training phase.

2.5 Summary

In this chapter, we provide an introductory guide to the vanilla supervised classification problem. We explore the fundamental fairness definitions commonly utilized in centralized machine learning, focusing on those that are most relevant to the challenges addressed in our research. Additionally, we expand the scope of the

supervised learning problem to the domain of federated learning, where data is distributed among multiple clients, and present fairness notions specifically tailored for such distributed learning settings. Figure 2.2 serves as a summary of the different fairness categories and concepts discussed for both centralized and federated learning settings, providing key references for further exploration.

Building upon this foundation, the subsequent chapter delves into the challenge of achieving group fairness across global demographics within the federated learning paradigm. We propose a novel approach designed to address this issue, present an optimization algorithm that effectively solves the proposed fairness problem, and conduct extensive experimental studies, examining its performance and capabilities in various scenarios.

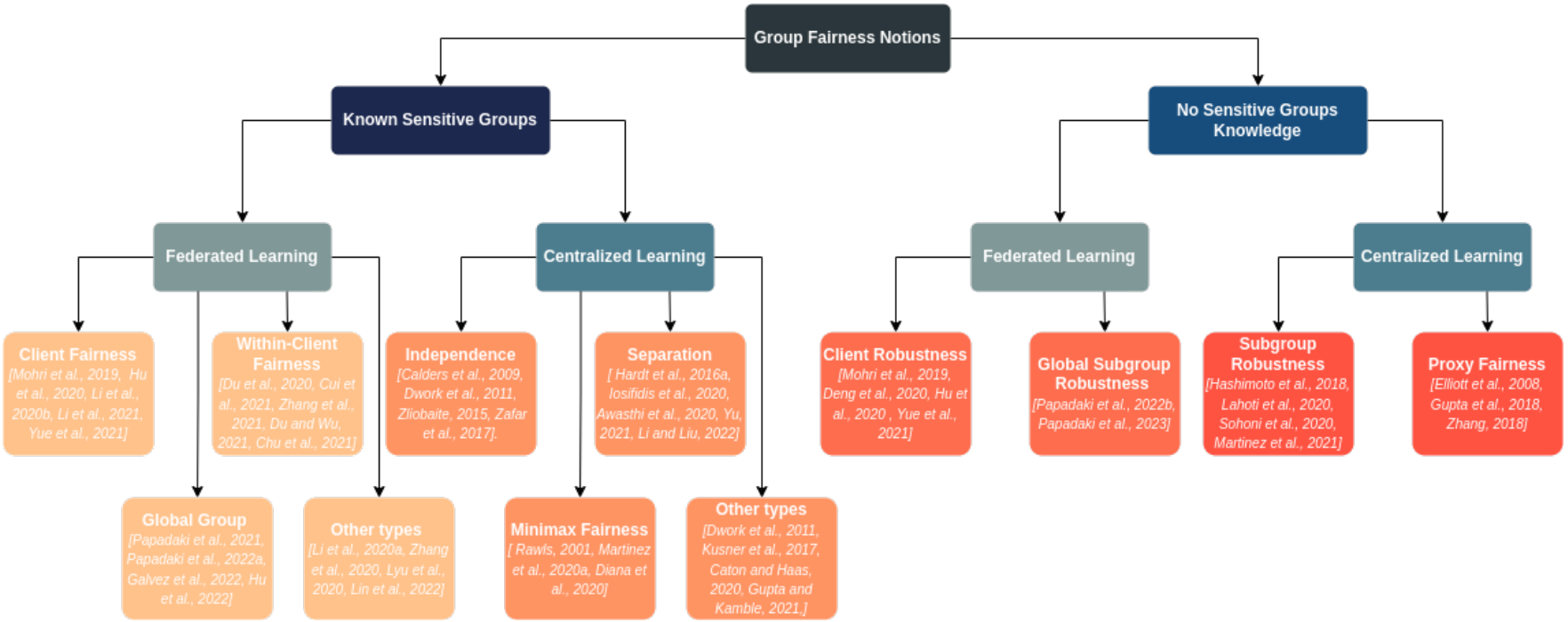

Figure 2.2: Comprehensive overview of the concepts and ideas discussed in this chapter for fairness in the context of both federated learning and centralized machine learning. We note that we include only the key relevant references that contribute to each notion.

Chapter 3

Global Group Fairness in Federated Learning

In this chapter, we formally formulate minimax global group fairness in federated learning settings where the demographic populations are predefined and known by the clients participating in the federation. In particular, the goal is to learn a global model that improves the worst-performing global group regardless of the local group representation of some clients during the training phase. 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. Finally, we empirically 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. The content of this chapter is presented in [Papadaki et al., 2021, Papadaki et al., 2022a].

3.1 Minimax Pareto Fairness in Centralized ML

In order to address the challenge of achieving fairness across all groups included in the clients' distribution in federated learning settings, we leverage the minimax group fairness criterion presented in Definition 2.6. This criterion is chosen due to its

property of avoiding unnecessary harm to any demographic group unless absolutely necessary, making it suitable for sensitive domains such as healthcare and finance. Prior to introducing our proposed method though, we provide an overview of how the minimax group fairness problem in a centralized machine learning setting is adapted to generate solutions that are properly Pareto optimal [Geoffrion, 1968].

A hypothesis is considered (Geoffrion) properly Pareto optimal if it satisfies the condition that improving one group's risk cannot be achieved without significantly worsening the risk of other groups [Geoffrion, 1968, Miettinen, 2012]. We can achieve properly Pareto optimal solutions by optimizing the linear combination of the group-conditional risks, under some mild assumptions, as also discussed in [Miettinen, 2012, Martinez et al., 2020a, Diana et al., 2020].

In particular, let the loss function $\ell : \Delta^{|\mathscr{Y}|-1} \times \Delta^{|\mathscr{Y}|-1} \to \mathbb{R}_+$ be a convex function with respect to the hypothesis h, or equivalently the classifier's output. We note that this is a realistic assumption since the most common loss functions in machine learning, such as Brier score and cross-entropy, are convex. Let also the hypothesis class $\mathscr{H}$ solving the minimax objective provided in Eq. 2.3, Definition 2.6, be a convex set. Then, according to Theorem 3.1.6 in [Miettinen, 2012], the minimax fairness problem presented in Eq. 2.3 is equivalent to solving

$$\begin{aligned} \min_{h\in\mathscr{H}} \max_{a\in\mathscr{A}} \mathop{\mathbb{E}}_{X,Y|A=a}[\ell(h(X),Y)] &= \min_{h\in\mathscr{H}} \max_{\mu\in\Delta^{|\mathscr{A}|-1}} \sum_{a\in\mathscr{A}} \mu_a \mathop{\mathbb{E}}_{X,Y|A=a}[\ell(h(X),Y)] \\ &\geq \min_{h\in\mathscr{H}} \max_{\mu\in\Delta^{|\mathscr{A}|-1}_{\geq\varepsilon}} \sum_{a\in\mathscr{A}} \mu_a \mathop{\mathbb{E}}_{X,Y|A=a}[\ell(h(X),Y)] \end{aligned} \tag{3.1}$$

where $\Delta^{|\mathscr{A}|-1}_{\geq\varepsilon}$ denotes a set of vectors with components at least ε, with ε being a small value close to zero, and $\Delta^{|\mathscr{A}|-1}$ ing the probability simplex that permits coefficients with zero values (i.e., $\varepsilon = 0$). We note that solving the objective in the LHS of the inequality in Eq. 3.1 can result in weakly Pareto optimal solutions. This means that the resulting model may have the same worst-case risk as the model achieving the RHS of the inequality in Eq. 3.1, but the other sensitive group risks could be further improved. In contrast, by allowing strictly positive weighting coefficients, we can obtain properly Pareto optimal solutions [Geoffrion, 1968, Miettinen, 2012].

3.2 Minimax Fairness across Global Demographics

We now consider a supervised federated learning scenario where the random variables $(X,Y,A,K) \in \mathscr{X} \times \mathscr{Y} \times \mathscr{A} \times \mathscr{K}$ represent the inputs, the target, the unique/global demographic groups and the clients participating in the federation. Our goal is to learn a hypothesis h that ensures (demographic) group fairness across all sensitive populations present in the distribution of clients in federated learning settings.

Informally, a global model, produced by a federated learning procedure is global group fair, if it is fair with respect to the global sensitive demographics available across clients, regardless of the local group representation. To formalize this concept, we utilize the minimax properly Pareto group fairness criterion in Eq. 3.1. Nevertheless, we highlight that the concept of global group fairness can be realized using other fairness criteria, such as equal opportunity and risk parity, as demonstrated in [Gálvez et al., 2022, Hu et al., 2022]. Additionally, we consider the distribution of demographic groups on each client, denoted as $p(A|K)$, to be dependent on the client. This accommodates the possibility that certain clients may have a higher, lower or no representation of specific demographic groups compared to others.

We also note that the conditional distribution of the target variable Y given the group variable A and the feature variable X remains the same across clients, meaning that $p(Y|X,A,K) = p(Y|X,A)$. This assumption holds in many federated learning scenarios and implies that the optimal classifier for a particular demographic group is shared among all clients. It provides a basis for collaborative learning and ensures that the insights gained from one client's data can be beneficial for improving the utility of the global model on the same group across all clients.

In our formulation, we assume that the group loss estimates are split into $|\mathscr{K}|$ estimators associated with each client. To incorporate the role of different clients, we express the linear weighted formulation of Equation 3.1 using importance weights as follows:

$$\min_{h \in \mathscr{H}} \max_{\mu \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}} \sum_{a \in \mathscr{A}} \mu_a \mathop{\mathbb{E}}_{X,Y|A=a} [\ell(h(X),Y)]$$

$$
\begin{aligned}
&= \min_{h\in\mathscr{H}} \max_{\mu\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} \sum_{a\in\mathscr{A}} p(A=a)\frac{\mu_a}{p(A=a)} \mathop{\mathbb{E}}_{X,Y|A=a}[\ell(h(X),Y)] \\
&= \min_{h\in\mathscr{H}} \max_{\mu\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} \sum_{k\in\mathscr{K}} p(K=k) \sum_{a\in\mathscr{A}} \frac{p(A=a|K=k)}{p(A=a)}\mu_a \mathop{\mathbb{E}}_{X,Y|A=a,K=k}[\ell(h(X),Y)].
\end{aligned} \tag{3.2}
$$

We describe the concept of minimax global group fairness in federated learning formally in Definition 3.1.

Definition 3.1 (Minimax Pareto Global Group Fairness)**.** *A convex hypothesis h^* is minimax (Pareto) global group fair, if it minimizes the group importance weighted client risks*

$$
h^*, \boldsymbol{\mu}^* = \arg\min_{h\in\mathscr{H}} \max_{\mu\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} \sum_{k\in\mathscr{K}} p(K=k) r_k(h,\boldsymbol{w}), \tag{3.3}
$$

with $\boldsymbol{\mu}^*$ being the corresponding minimax fair weighting vector, $w_a = \mu_a / p(A=a)$ denoting the importance weight for a particular demographic group and $r_k(h,\boldsymbol{w}) = \sum_{a\in\mathscr{A}} p(A=a|K=k) w_a \mathop{\mathbb{E}}_{X,Y|A=a}[\ell(h(X),Y)]$ being the expected client risk.

The objective function in Eq. 3.3 encompasses the concept of global group fairness and is designed to achieve fairness across all groups in the federated learning setting. It aims to create a solution that is comparable to the centralized machine learning setting, where all data is available in a single location.

Under certain conditions, our proposed objective exhibits a relationship with (minimax) client fairness – or equivalently client robustness – studied in [Mohri et al., 2019, Deng et al., 2020], as formally demonstrated in Lemma 1 in [Papadaki et al., 2022a]. One such condition is having a federation of clients consisting only of a single sensitive group. Another scenario is when a client's group priors align with a group minimax weighting vector.

This compatibility result implies that client-level fairness may differ from group-level fairness. For a more detailed discussion and the formal proof, we refer the reader to Lemma 1 in our work [Papadaki et al., 2022a]. Nevertheless, we provide an experimental validation of these insights in Section 3.4.

3.2.1 Finite sample (Minimax) Global Group Fairness

In real applications, clients do not have access to the data distribution but instead, each client k owns a dataset $D_k = \{(x_i^k, y_i^k, a_i^k)\}_{i=1\dots n_k}$ sampled independently and identically from the distribution $p(X,Y,A|K=k)$.

To describe the data subsets related to specific demographic groups, we define the set $D_a = \bigcup_{k\in\mathscr{K}} D_{a,k}$, which includes the collection of the sets $D_{a,k}$ with examples associated with group a within client k. Similarly, the set D represents the complete dataset, comprising the data examples across all demographic groups and clients, i.e. $D = \bigcup_{k\in\mathscr{K}} D_k = \bigcup_{a\in\mathscr{A}} D_a = \bigcup_{k\in\mathscr{K}} \bigcup_{a\in\mathscr{A}} D_{a,k}$. Recall that in some cases, certain clients may not have any data instances associated with specific demographic groups. As a result, there may be empty sets $D_{a,k}$ for certain combinations of demographic groups a and clients k within the federation.

Let the hypothesis h be parametrized via a vector of parameters $\boldsymbol{\theta} \in \Theta$, i.e., $h(\cdot) = h(\cdot;\boldsymbol{\theta})$. For instance, $\boldsymbol{\theta}$ could correspond to the set of weights/biases in a neural network. We remark that convexity is assumed for the hypothesis class $\mathscr{H}$, which refers to the set of models or functions, not the vector space Θ in which the model parameters are defined. Thus, the vector space Θ that represents the parameter space of the models can be non-convex. We can express the empirical formulation of the optimization problem in Eq. 3.3 as

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} \sum_{a\in\mathscr{A}} \mu_a \hat{r}_a(\boldsymbol{\theta}) \equiv \min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} \sum_{k\in\mathscr{K}} \frac{n_k}{n} \hat{r}_k(\boldsymbol{\theta},\boldsymbol{w}), \tag{3.4}$$

where n_k is the number of data points in client k, n is the total number of data points in the dataset, and the empirical risks

$$\hat{r}_k(\boldsymbol{\theta},\boldsymbol{w}) = \sum_{a\in\mathscr{A}} \frac{n_{a,k}}{n_k} \hat{w}_a \hat{r}_{a,k}(\boldsymbol{\theta}), \quad \hat{r}_a(\boldsymbol{\theta}) = \sum_{k\in\mathscr{K}} \frac{n_{a,k}}{n_a} \hat{r}_{a,k}(\boldsymbol{\theta}),$$
$$\text{and} \quad \hat{r}_{a,k}(\boldsymbol{\theta}) = \frac{1}{n_{a,k}} \sum_{i=1}^{n_{a,k}} \ell(\theta; x_i^k, y_i^k).$$

Here, $\hat{r}_k(\boldsymbol{\theta},\boldsymbol{w})$ is an estimate of $r_k(\boldsymbol{\theta},\boldsymbol{w})$, $\hat{r}_{a,k}(\boldsymbol{\theta})$ is an estimate of $r_{a,k}(\boldsymbol{\theta}) =$

$\underset{X,Y|A=a,K=k}{\mathbb{E}}[\ell(h(X),Y)]$, n_a is the number of data points associated with group a, and $n_{a,k}$ is the number of data points in client k that belong to group a.

3.3 Federated Minimax Global Group Fairness Algorithm

To address the problem stated in Eq. 3.4, we introduce an optimization algorithm called Federated Minimax (FedMinMax). We note that the objective function in Eq. 3.4 can be formulated as a zero-sum game, involving a learner and an adversary. The learner's goal is to minimize the objective by finding optimal model parameters $\boldsymbol{\theta}$, while the adversary aims to maximize the objective by determining the weighting coefficients $\boldsymbol{\mu}$.

Algorithm 1 FEDERATED MINIMAX (FEDMINMAX) ALGORITHM

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

Server **initializes** $\boldsymbol{\mu}^0 \leftarrow \{|D_a|/|D|\}_{a\in\mathscr{A}}$ and $\boldsymbol{\theta}^0$ randomly.
for $t = 1$ **to** T **do**
  Server **computes** $\boldsymbol{w}^{t-1} \leftarrow \boldsymbol{\mu}^{t-1}/\boldsymbol{\mu}^0$
  Server **broadcasts** $\boldsymbol{\theta}^{t-1}$, $\boldsymbol{w}^{t-1}$
  for each client $k \in \mathscr{K}$ **in parallel do**
    $\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_\theta \hat{r}_k(\boldsymbol{\theta}^{t-1}, \boldsymbol{w}^{t-1})$
    Client-k **obtains** and **sends** $\{\hat{r}_{a,k}(\boldsymbol{\theta}^{t-1})\}_{a\in\mathscr{A}}$ and $\boldsymbol{\theta}_k^t$ to server
  end for
  Server **computes**: $\boldsymbol{\theta}^t \leftarrow \sum_{k\in\mathscr{K}} \frac{n_k}{n}\boldsymbol{\theta}_k^t$
  Server **updates:**
  $\boldsymbol{\mu}^t \leftarrow proj_{\Delta^{|\mathscr{A}|-1}_{\geq\varepsilon}}\left(\boldsymbol{\mu}^{t-1} + \eta_{\boldsymbol{\mu}}\nabla_{\boldsymbol{\mu}}\langle\, \boldsymbol{\mu}^{t-1}, \hat{r}_a(\boldsymbol{\theta}^{t-1})\rangle\right)$
end for

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

The FedMinMax algorithm, outlined in Algorithm 1, assumes that all clients are available to participate in each communication round t. It operates through an iterative process, where both the learner and the adversary take turns updating their respective variables. In particular, at each communication round $t \in T$,

1. the clients receive the latest model parameters $\boldsymbol{\theta}^{t-1}$ and the importance weights

$\boldsymbol{w}^{t-1}$;

2. The clients perform one gradient descent step on the model using n_k- samples;

3. The clients share with the server the updated model parameters $\boldsymbol{\theta}_k^t$ and the local empirical group risks $\{\hat{r}_{a,k}(\boldsymbol{\theta}^{t-1})\}_{a\in\mathscr{A}}$;

4. Finally, the server then performs a weighted average of the client model parameters, and updates the weighting coefficient using a projected gradient ascent step.

In the proposed algorithm, we denote $proj_{\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}}(\cdot)$ the projection operator onto the simplex $\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}$ and employ the Euclidean algorithm introduced in [Duchi et al., 2008] to implement the projection operation. Furthermore, it is worth noting that in practice, clients have the flexibility to use a batch of samples during the local optimization step, rather than utilizing the entire local dataset. The chosen batch size b_k must be proportional to the dataset size n_k.

3.3.1 Algorithmic Analysis

In order to establish convergence guarantees for our proposed algorithm, we begin by demonstrating in Lemma 3.1, the equivalence between each optimization step of our algorithm and the centralized version of our algorithm that is presented in Algorithm 2.

Algorithm 2 CENTRALIZED MINIMAX BASELINE ALGORITHM

Input: T : total number of adversarial rounds, $\eta_{\boldsymbol{\theta}}$: model learning rate, $\eta_{\boldsymbol{\mu}}$: adversary learning rate, D_a: set of examples for group a, $\forall a \in \mathscr{A}$.

1: Server **initializes** $\boldsymbol{\mu}^0 \leftarrow \{|D_a|/|D|\}_{a\in\mathscr{A}}$ and $\boldsymbol{\theta}^0$ randomly.
2: **for** $t = 1$ **to** T **do**
3: Server **computes** $\boldsymbol{\theta}^t \leftarrow \boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_\theta \hat{r}(\boldsymbol{\theta}^{t-1}, \boldsymbol{\mu}^{t-1})$
4: Server **updates**
$\boldsymbol{\mu}^t \leftarrow proj_{\Delta^{|\mathscr{A}|-1}}\left(\boldsymbol{\mu}^{t-1} + \eta_{\boldsymbol{\mu}} \nabla_{\boldsymbol{\mu}} \langle\, \boldsymbol{\mu}^{t-1}, \hat{r}_a(\boldsymbol{\theta}^{t-1})\rangle\right)$
5: **end for**

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

Lemma 3.1. *Consider our federated learning setting where each entity k has access to a local dataset $D_k = \bigcup_{a\in\mathscr{A}} D_{a,k}$, and a centralized machine learning setting where there is a single entity that has access to a single dataset $D = \bigcup_{k\in\mathscr{K}} D_k = \bigcup_{k\in\mathscr{K}}\bigcup_{a\in\mathscr{A}} D_{a,k}$ (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) lead to the same global model provided that learning rates and model initialization are identical.*

Lemma 3.1 asserts that the optimization steps performed in our federated algorithm, as outlined in Algorithm 1, achieve the same objective as the steps in the centralized algorithm presented in Algorithm 2. Therefore, both algorithms converge towards the same solution, despite the decentralized and collaborative nature of federated learning.

By establishing this equivalence, we can leverage the 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 [Chen et al., 2017b], 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). We provide the convergence guarantees in Lemma 3.2.

Assumption 3.1 (δ-approximate Bayesian Oracle)**.** *We assume that the server approximates the hypothesis h, parametrized by $\boldsymbol{\theta}$, for any group weights $\boldsymbol{\mu}$ using δ-approximate Bayesian solver*

$$M(\boldsymbol{\mu}) \simeq \arg\min_{h\in\mathscr{H}} R(h,\boldsymbol{\mu}) \coloneqq \sum_{a\in\mathscr{A}} \mu_a \hat{r}_a(h). \tag{3.5}$$

Assumption 3.2 (Lipschitzness Condition)**.** *We also assume that $R(h,\boldsymbol{\mu}) = \sum_{a\in\mathscr{A}} \mu_a \hat{r}_a(h)$ is a 1-Lipschitz function w.r.t. $\boldsymbol{\mu}$.*

Lemma 3.2 (Adjusted from Theorem 7 in [Chen et al., 2017b])**.** *Let assumptions 3.1, 3.2 hold. Let also $\mathscr{D}$ be a uniform distribution over a set of hypotheses $\{h^1,\ldots,h^T\}$*

and $\eta = \max_{\boldsymbol{\mu} \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}} \frac{||\boldsymbol{\mu}||_2}{\sqrt{2T}}$. *Then, given the equivalence in Lemma 3.1, Algorithms 1 and 2 output a distribution* $\mathscr{D}$*, such that*

$$\max_{\boldsymbol{\mu} \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}} \mathbb{E}_{h \sim \mathscr{D}}[R(h, \boldsymbol{\mu})] \leq \delta r^* + \max_{\boldsymbol{\mu} \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}} ||\boldsymbol{\mu}||_2 \sqrt{\frac{2}{T}}$$

where the optimal risk $r^* = \arg\min_{h \in \mathscr{H}} \max_{\boldsymbol{\mu} \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}} R(h, \boldsymbol{\mu})$.

3.4 Empirical Results

Next, 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.

3.4.1 Experimental Setup

We conducted experiments using three different federated learning settings, each characterized by specific properties related to the allocation of sensitive groups among clients. We present the various settings, datasets, and model architectures used for the experiments below.

3.4.1.1 Learning Settings

We explored the following three federated learning settings based on the allocation of sensitive groups:

1. **Equal access to Sensitive Groups (ESG):** In this setting, all clients have equal access to all sensitive groups, but each client has an insufficient amount of data to train a model individually. Also, the distribution of sensitive classes is the same across clients (i.e., $n_i = n_j \forall i, j \in \mathscr{K}, i \neq j$ and $n_{a,i} = n_{a,j} \forall i, j \in \mathscr{K}, a \in \mathscr{A}, i \neq j$). This setting illustrates a case where group fairness and client fairness are not equivalent.

2. **Partial access to Sensitive Groups (PSG):** In this setting, each participant has access to a subset of the available sensitive groups. The data allocation

across participants is unbalanced, meaning that the size of the local datasets varies among clients (i.e., $n_i \neq n_j \forall i, j \in \mathscr{K}, i \neq j$). This setting represents another scenario where group fairness and client fairness are incompatible. We used this setting to compare performance when there is low or no local representation of specific groups.

3. **Access to a Single Sensitive Group (SSG):** In this setting, each client possesses data from only one sensitive group. The local dataset size varies among clients, similar to the PSG setting. This setting demonstrates the equivalence between group fairness and client fairness objectives, as derived from Lemma 1 in [Papadaki et al., 2022a].

It is worth noting that the ESG setting represents an i.i.d. (independent and identically distributed) data scenario, while PSG and SSG settings involve non-i.i.d. data. Each client has a unique dataset, and there are no duplicated examples across clients. Our experiments included a federation of 40 clients and a single server that orchestrates the training process.

3.4.1.2 Datasets

For our experiments, we utilized several datasets to evaluate the performance of the FedMinMax algorithm.

- **Synthetic Dataset:** We generated a synthetic dataset for binary classification with two sensitive groups, i.e., $|\mathscr{Y}| = 2$ and $|\mathscr{A}| = 2$. The data generation process involves the following distributions: the group variable A follows a Bernoulli distribution $Ber(\frac{1}{2})$, the input features X follow a normal distribution $\mathscr{N}(0,1)$, and the target variable Y given X and $A = a$ follows a Bernoulli distribution $Ber(h_a^*)$, where h_a^* is the optimal hypothesis for group $A = a$. The specific values used for h_a^* were $\{u_0^h, u_1^h, u_0^l, u_1^l\} = \{0.6, 0.9, 0.3, 0.1\}$. This dataset allowed us to demonstrate the performance of FedMinMax in scenarios with different sensitive groups and optimal hypotheses. The optimal hypothesis h for this dataset aligns with the optimal model specifically designed for group $A = 0$, as shown on the left side of Figure 3.1. This means that

the model tailored for group $A = 0$ represents the optimal solution for that particular group. In the scenario where equal access to sensitive groups (ESG) is considered, we evenly distribute the two groups among 40 clients. On the other hand, in the case of single access to sensitive groups (SSG), each client is assigned only one group, and each group is distributed to 20 clients. It is important to note that the number of samples in each local dataset may vary across clients in the SSG setting. However, for binary sensitive group scenarios, there is no partial access to the sensitive groups (PSG) setting since it is equivalent to the SSG setting.

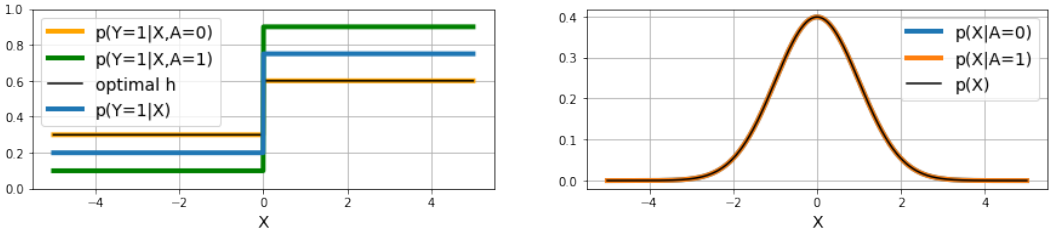

Figure 3.1: Illustration of the optimal hypothesis h and the conditional distributions $p(Y|X)$ and $p(X|A)$ for the generated synthetic dataset. *Left*: The worst group is $A = 0$ and the minimax optimal hypothesis h (black line) is equal to the optimal model for the worst group (orange line). *Right*: The distributions $p(X)$, and conditional distributions $p(X|A = 0)$ and $p(X|A = 1)$ are overlapping.

- **Adult Dataset [Lichman, 2013]:** The Adult dataset is a binary classification dataset, i.e., $|\mathscr{Y}| = 2$, comprising $32,561$ entries used for predicting annual income based on 12 input features such as age, race, education, and marital status. We considered four sensitive groups, $|\mathscr{A}| = 4$, created by combining gender labels and income levels. In the ESG setting, we evenly distributed the four groups among 40 clients. In the PSG setting, 20 clients have access to male subgroups, and the other 20 have access to female subgroups. In the SSG setting, each client has access to only one sensitive group.

- **ACS Employment Dataset [Ding et al., 2021]:** The ACS Employment Dataset is a recent dataset derived from ACS PUMS data, used for predicting an individual's employment status, i.e., $|\mathscr{Y}| = 2$. For our experiments, we utilize the 2018 1-year data encompassing all US states and Puerto Rico. For

this dataset, we consider the following two scenarios based on the sensitive group:

1. By combining race and utility labels, we generate the sensitive groups {Employed White, Employed Black, Employed Other, Unemployed White, Unemployed Black, Unemployed Other}, resulting in a total of $|\mathscr{A}| = 6$ groups. In the PSG setting, 20 clients have access to data from the *Unemployed White*, *Employed Black*, and *Employed White* groups, while the remaining clients hold data from the *Unemployed Other*, *Unemployed Black*, and *Employed Other* groups. Finally, in the SSG setting, each client is allocated only one sensitive group, with the *Employed White* group owned by 10 clients, and the remaining 5 groups distributed among 6 clients each.

2. The sensitive group is the race attribute with the original nine labels { *White*, *Black / African American*, *American Indian*, *Alaska Native*, *A.I. &/or A.N. Tribes*, *Asian*, *N. Hawaiian & other P.I.*, *Other*, *Multiple*}. In the PSG setting, 20 clients possess data from the *White*, *Black / African American*, *American Indian*, *Alaska Native*, and *A.I. &/or A.N. Tribes* groups, while the other clients hold data from the remaining groups. Finally, in the SSG setting, each client owns only one sensitive group, with each group allocated to only four clients, except for the *White* race group, which is distributed among eight clients.

- **FashionMNIST Dataset [Xiao et al., 2017]:** FashionMNIST is a grayscale image dataset containing $60,000$ training images and $10,000$ testing images. Each image is a 28×28 pixel grayscale image categorized into one of ten clothing categories. In our experiments, we treat each target category as a sensitive group ($|\mathscr{A}| = 10$). In the PSG setting, 20 participants have access to a subset of five clothing categories, while the remaining 20 participants have access to the other five categories.

- **CIFAR-10 Dataset [Krizhevsky et al.,]:** CIFAR-10 is a dataset containing

60,000 color images of size 32×32 pixels. Each image belongs to one of ten object classes. We considered the target categories as sensitive groups ($|\mathscr{A}| = 10$). In the PSG setting, half of the clients have access to a subset of five object classes, while the other half have access to the remaining five classes.

Similar to other standard baselines like AFL [Mohri et al., 2019] and DRFA [Deng et al., 2020], we adopt the practice of assigning the target categories as sensitive attributes in the image datasets, rather than using arbitrary labels.

3.4.1.3 Benchmarks and Hyperparameters

As part of our baseline comparison, we consider the Centralized Minmax Baseline, for which we run FedMinMax with only one client to simulate the centralized machine learning setting described in Algorithm 2. This allows us to verify the findings stated in Lemma 3.1. We also compare against AFL [Mohri et al., 2019], TERM [Li et al., 2021], FedAvg [McMahan et al., 2016b], and q-FedAvg [Li et al., 2020b]. We do not compare our approach to baselines that explicitly optimize for a different fairness metric, such as demographic parity, as our focus in this work is not on comparing fairness metrics. Instead, we aim to demonstrate the effectiveness of considering global demographics across entities rather than multiple local ones, as highlighted in [Zhang et al., 2021]. To explore this further, we propose a variation of our algorithm called LocalFedMinMax, where optimization is performed separately for each local group instead of globally. More detailed discussion on LocalFedMinMax can be found in Section 3.4.3.

To ensure a fair comparison, we assume that every client is available to participate in each communication round for all methods. Additionally, we conduct a grid search over hyperparameters to identify the best combination for each algorithm. For AFL and FedMinMax, we set the batch size equal to the number of examples per client, while for TERM, FedAvg, and q-FedAvg, the batch size is set to 100. The grid search is performed 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 applicable). During the

training process, we tune the hyperparameters based on the validation set for each approach. The mean and standard deviation reported in the results are calculated over three runs. We use 3-fold cross-validation to split the data into training and validation sets for each run.

3.4.1.4 Model Architectures and Loss Functions

For the synthetic dataset, we use a Multi-Layer Perceptron (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 Convolutional Neural Network (CNN) architecture with two 2D convolutional layers with kernel size 3, stride 1, and padding 1. Each convolutional layer is followed by a max-pooling 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 the 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 functions.

3.4.2 Global Group Fairness vs. CML and FL Baselines

We begin by comparing FedMinMax with the centralized Minmax baseline as well as other relevant federated learning methods. We assess the performance of the worst group, the best group, and the average utility for the Adult, FashionMNIST, CIFAR-10, and ACS Employment datasets. The results are presented in Figure 3.2.

FedMinMax demonstrates similar performance to the Centralized Minmax Baseline across all settings, consistent with the findings in Lemma 3.1. AFL achieves comparable results to FedMinMax and the Centralized Minmax Baseline in the Single access to Sensitive Groups (SSG) setting, where group fairness is implied by client fairness, aligning with Lemma 1 in [Papadaki et al., 2022a]. We also observe that FedAvg exhibits similar error rates for the best group across federated settings, but the risk for the worst group increases as the local data becomes more heterogeneous (i.e., in PSG and SSG).

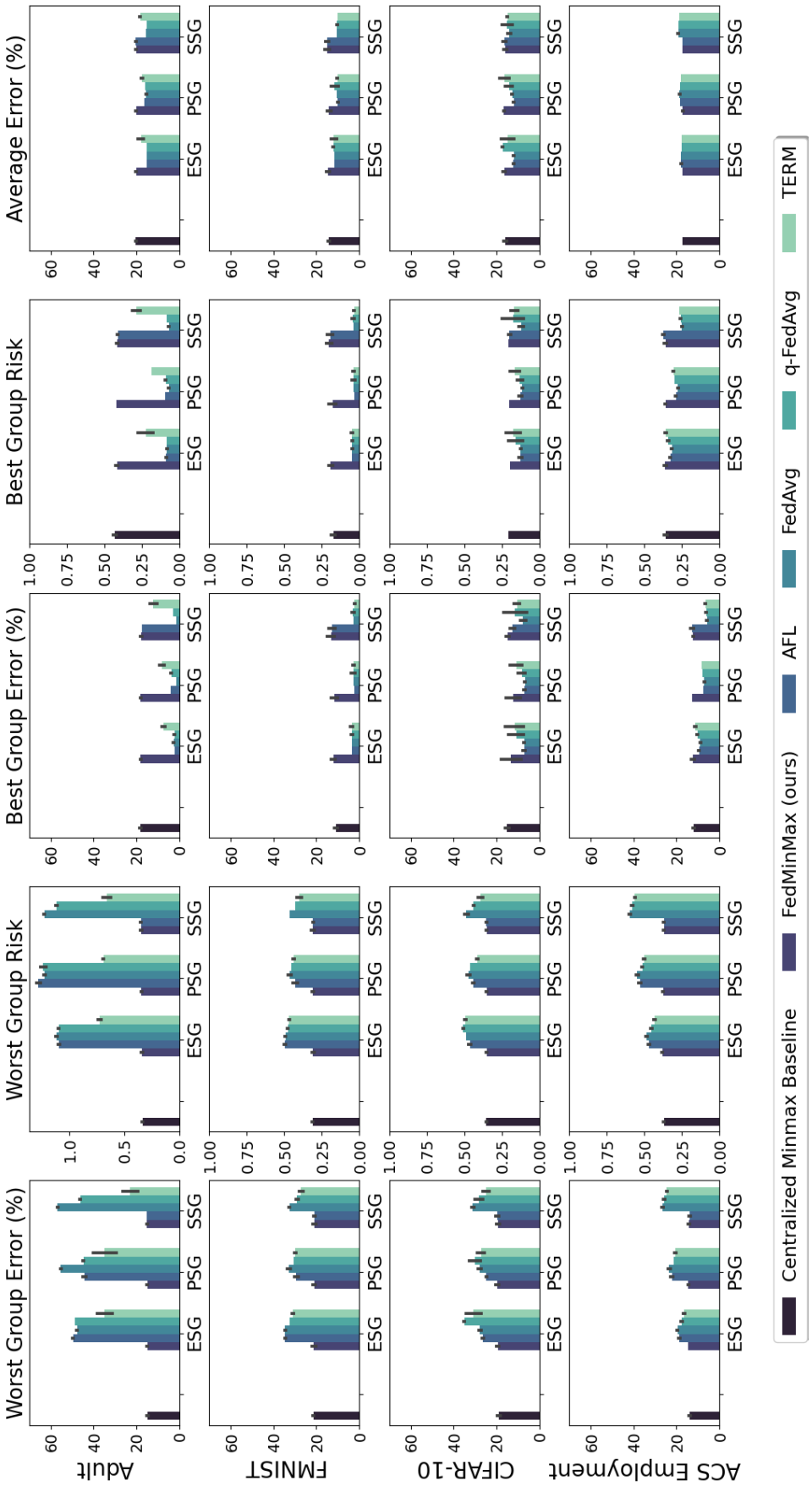

Figure 3.2: The graphs present a comparison of the worst group, best group, and average risks and errors for AFL, FedAvg, q-FedAvg, TERM, FedMinMax, and the Centralized Minmax Baseline. The comparison is conducted across three runs for each method, considering different federated learning scenarios. Each bar on the graph represents the mean and standard deviation of the corresponding metric evaluated on the testing set.

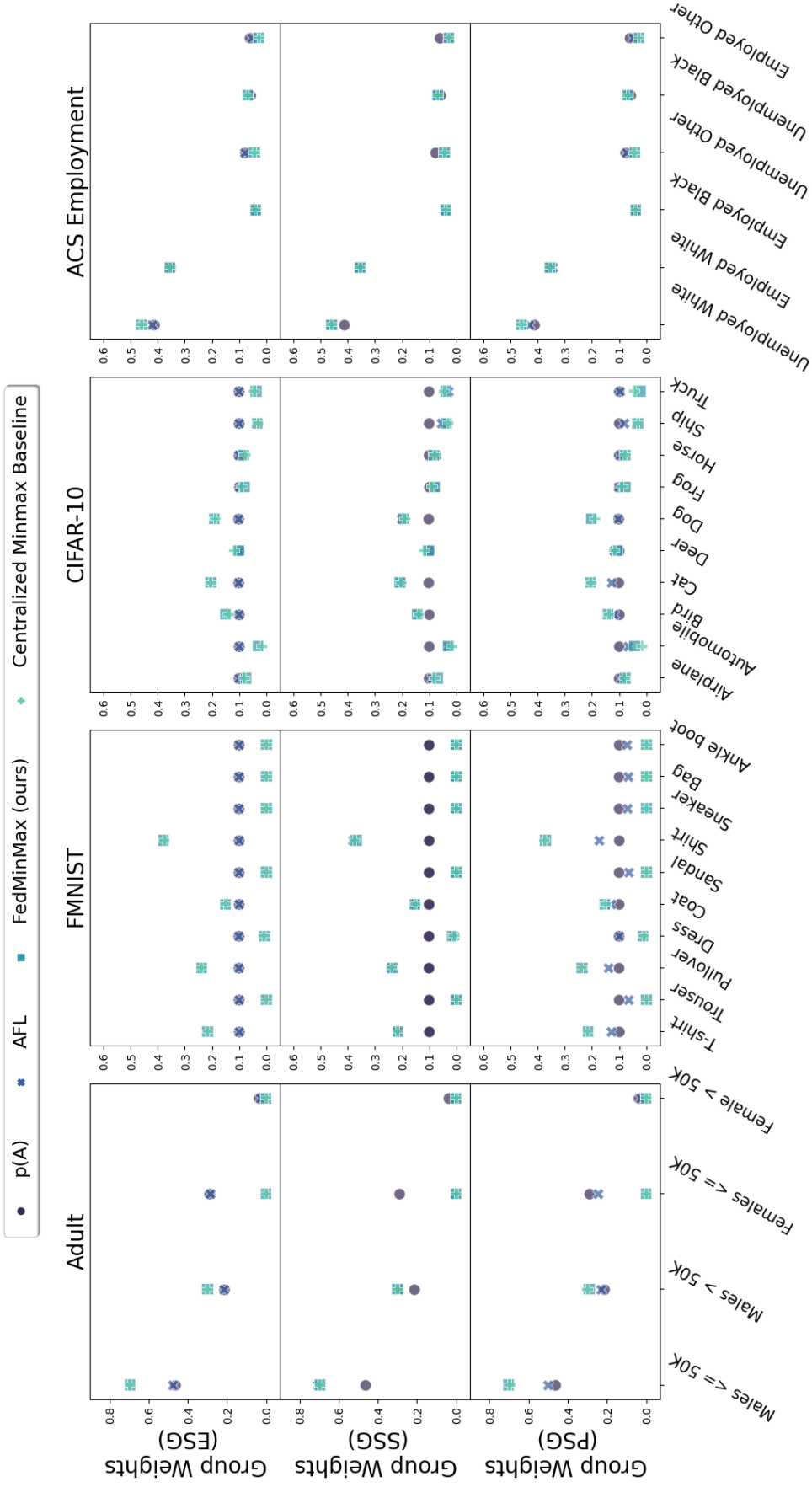

Figure 3.3: Sensitive group weighting coefficients for each minimax approach across different datasets. These results are calculated during the training time. The prior group distribution $p(A)$ is also illustrated. The weighting coefficients were determined based on the group risks observed during training and may not necessarily align with the group risks observed during testing.

On the other hand, q-FedAvg and TERM outperform AFL and FedAvg in terms of the worst group performance in PSG and Equal access to Sensitive Groups (ESG) settings in various datasets. However, they do not achieve minimax group fairness in any of the FL settings. We highlight that FedMinMax consistently delivers the best worst-group performance across all settings, as anticipated.

Additionally, we present the final group weighting coefficients for the minimax approaches AFL, FedMinMax, and Centralized Minmax Baseline to explore the relationship between global group fairness and client fairness in federated learning settings. The results are summarized in Figure 3.3. We note that some of the illustrated weighting coefficients may overlap.

The proposed approach demonstrates similar group weights across all settings for the same dataset. FedMinMax also achieves identical weighting coefficients as the Centralized Minmax Baseline, confirming the findings stated in Lemma 3.1. AFL produces weights that are comparable to the group priors in the ESG setting and progressively converge towards the minimax weighting coefficients as the heterogeneity with respect to the sensitive groups increases. AFL achieves similar weights to FedMinMax and Centralized Minmax Baseline only in the SSG scenario, where each participant has access to exactly one group, as indicated by Lemma 1 in [Papadaki et al., 2022a]. It is important to note that the group weighting coefficients are updated based on the risks calculated on the training set and may not generalize to the testing set for every dataset.

In Appendix B, we offer a comprehensive description of the weighting coefficients for each approach in Tables B.2, B.4, B.5, B.9, and B.11; and the detailed numerical values showcasing the efficacy of the proposed approach in each setting and dataset in Tables B.1, B.3, B.6, B.8, and B.10.

3.4.3 Global Group Fairness vs. Within-Client Fairness

To investigate the efficiency of considering global demographics across entities instead of multiple local ones, as proposed in [Zhang et al., 2021], we introduce an adjusted version of our algorithm called LocalFedMinMax. In this version, the adversary proposes a weighting coefficient for each group located in a client, denoted

as $\boldsymbol{\mu} = \{\{\mu_{a,k}\}_{a\in\mathscr{A}}\}_{k\in\mathscr{K}}$.

Specifically, we compare our approach to the following within-client fairness optimization problem:

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|\times|\mathscr{K}|-1}} \sum_{a\in\mathscr{A}} \sum_{k\in\mathscr{K}} \mu_{a,k}\hat{r}_{a,k}(\boldsymbol{\theta}), \tag{3.6}$$

where $\boldsymbol{\theta}$ represents the model parameters in the optimization space Θ, and $\hat{r}_{a,k}(\boldsymbol{\theta})$ denotes the empirical risk associated with group a and client k.

In order to solve Eq. 3.6, we modify the original FedMinMax algorithm. This version of the algorithm is described in detail in Algorithm 3. We remark that, while the adversary in our proposed algorithm uses a single weighting coefficient for every common demographic group (i.e., $\boldsymbol{\mu} = \{\mu_a\}_{a\in\mathscr{A}}$), the optimization objective in Eq. 3.6 treats each demographic group that a client has access to, as a unique sensitive group, even if the same group exists in several clients. This adjustment allows us to explore the efficiency of considering global demographics across entities rather than treating each local demographic group independently.

Algorithm 3 LOCAL FEDERATED MINIMAX (LOCALFEDMINMAX) ALGORITHM

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

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

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

Table 3.1 presents the results of both LocalFedMinMax and FedMinMax on

FashionMNIST and CIFAR-10 datasets, and two federated networks with 10 and 40 participating clients, respectively. We observe that in the SSG scenario, both LocalFedMinMax and FedMinMax achieve similar improvements in the worst group performance, regardless of the number of clients. This indicates that considering local group demographics in LocalFedMinMax leads to comparable results to FedMinMax, even with varying client numbers.

In the ESG scenario, we also observe similar behaviour in the smaller federated network, where LocalFedMinMax and FedMinMax show similar improvements in the worst group performance. However, as the number of clients increases and the amount of data per group per client reduces, LocalFedMinMax tends to have a worse performance compared to FedMinMax. This suggests that as the local group representation becomes sparser, LocalFedMinMax may struggle to generalize well. In contrast, FedMinMax is not affected by the local group representation since it aggregates the statistics received from each client and updates the weights for global demographics. This leads to better generalization performance.

These results indicate that local methods ensuring within-client fairness, such as LocalFedMinMax, can lead to global group fairness in certain scenarios with a limited number of clients and dense local group representation. In contrast, FedMinMax demonstrates consistent performance across different settings, making it more robust and suitable for generalization. In Appendix B, we provide the complete table of risks for FashionMNIST and CIFAR-10 in Tables B.12 and B.13, respectively.

3.5 Proofs

3.5.1 Analysis of Algorithm 1

In this section, we establish the convergence of the proposed federated learning algorithm (a) by first showing its equivalence to its centralized variant; and (b) by leveraging existing convergence guarantees for centralized machine learning algorithms to demonstrate FedMinMax convergence.

3.5.1.1 Equivalence to Centralized Algorithm 2

Proof [Lemma 3.1]

Table 3.1: Comparison of the worst group risk achieved for FedMinMax and LocalFedMinMax on FashionMNIST and CIFAR-10 datasets. We highlight the worst risk values. Lower values indicate better performance.

	FashionMNIST					
	10 **Clients**			40 **Clients**		
Method	**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					
	10 **Clients**			40 **Clients**		
Method	**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

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 $D_k = \bigcup_{a\in\mathscr{A}} D_{a,k}$ and the dataset in centralized MinMax is $D = \bigcup_{k\in\mathscr{K}} D_k = \bigcup_{k\in\mathscr{K}}\bigcup_{a\in\mathscr{A}} D_{a,k}$; and
2. the model initialization θ^0, the number of adversarial rounds T,[1] 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

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

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

where

$$\hat{r}_k(\boldsymbol{\theta}, \boldsymbol{w}) = \sum_{a \in \mathscr{A}} \frac{n_{a,k}}{n_k} w_a \hat{r}_{a,k}(\boldsymbol{\theta}),$$

with $w_a = \frac{\mu_a}{\frac{n_a}{n}}$, and $\hat{r}_a(\boldsymbol{\theta}) = \sum_{k \in \mathscr{K}} \frac{n_{a,k}}{n_a} \hat{r}_{a,k}(\boldsymbol{\theta})$. Therefore, the model update

$$\boldsymbol{\theta}^t = \sum_{k \in \mathscr{K}} \frac{n_k}{n} \boldsymbol{\theta}_k^t = \sum_{k \in \mathscr{K}} \frac{n_k}{n} \left(\boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_\theta \hat{r}_k(\boldsymbol{\theta}^{t-1}, \boldsymbol{w}^{t-1}) \right)$$

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

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

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. 3.5.1.1, 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. □

3.5.1.2 Convergence Analysis of Algorithm 1

For the convergence analysis of Algorithm 1 we leverage the results for improper robust optimization presented in [Chen et al., 2017b] that show convergence using the regret guarantees of the projected gradient ascent algorithm.

Proof [Lemma 3.2] We define the risk induced by the adversary picking $\boldsymbol{\mu} \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}$ and the modeller picking $h \in \mathscr{H}$ in round $t \in [T]$, as $R(h^t, \boldsymbol{\mu}^t)$, where $h^t = M(\boldsymbol{\mu}^t)$ and $M(\cdot)$ being a δ-approximate Bayesian solver.

The regret guarantees of the projected gradient ascent algorithm gives that:

$$\frac{1}{T} \sum_{t \in [T]} R(h^t, \boldsymbol{\mu}^t) \geq \max_{\boldsymbol{\mu} \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}} \frac{1}{T} \sum_{t \in [T]} R(h^t, \boldsymbol{\mu}) - \max_{\boldsymbol{\mu} \in \Delta_{\geq \varepsilon}^{|\mathscr{A}|-1}} ||\boldsymbol{\mu}||_2 \sqrt{\frac{2}{T}}.$$

Furthermore, from the distributional oracle guarantee we have that:

$$\begin{aligned} r^* &= \{\arg\} \min_{h\in\mathscr{H}} \max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} R(h,\boldsymbol{\mu}) \\ &\geq \tfrac{1}{T}\sum_{t\in[T]} \min_{h\in\mathscr{H}} R(h,\boldsymbol{\mu}^t) \geq \tfrac{1}{T}\sum_{t\in[T]} \tfrac{1}{\delta} R(h^t,\boldsymbol{\mu}^t) \\ &\geq \tfrac{1}{\delta}\left(\max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} \tfrac{1}{T}\sum_{t\in[T]} R(h^t,\boldsymbol{\mu}) - \max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} ||\boldsymbol{\mu}||_2 \sqrt{\tfrac{2}{T}} \right). \end{aligned}$$

Thus, we have shown that for a uniform distribution over a set of hypotheses $\mathscr{D}$, the following inequality holds

$$\max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} \mathbb{E}_{h\sim\mathscr{D}}[R(h,\boldsymbol{\mu})] \leq \delta r^* + \max_{\boldsymbol{\mu}\in\Delta_{\geq\varepsilon}^{|\mathscr{A}|-1}} ||\boldsymbol{\mu}||_2 \sqrt{\frac{2}{T}}.$$

□

3.6 Summary

In this chapter, we present an approach aimed at addressing the challenge of achieving global group fairness in federated learning settings, particularly when there are limitations on participating entities' access to population groups during the training process. We highlight the distinctive features of our proposed objective and provide an optimization algorithm that effectively solves the introduced fairness problem. We assess the efficacy of our approach by conducting extensive experimental studies. The empirical findings illustrate that the suggested approach exhibits comparable or superior performance to relevant centralized and federated learning methods.

Moving forward, our focus shifts to the issue of fairness without demographics. We begin by examining this problem within the context of centralized machine learning and introduce a new flexible objective that offers various possible solutions. In subsequent chapters, we extend this objective to federated learning settings, exploring its applicability and effectiveness in decentralized learning scenarios.

Chapter 4

No Demographics, No Cry: Relaxed Conditional Value-at-Risk (RCVaR)

This chapter focuses on optimizing the performance for the worst-case group when there is no prior information available about the demographics. Building on the concepts discussed in Chapter 3, we address this challenge through the lens of minimax fairness, or equivalently subgroup robustness, where fairness is measured based on the utility perceived by the worst-served subset of individuals. We study this problem in the context of centralized machine learning and propose an alternative optimization objective that aims to improve the performance of the worst-performing group while considering a minimum group size constraint. By incorporating a trade-off parameter, we show that existing approaches that balance fairness and overall performance can be recovered. Through our analysis, we establish the connection between our proposed optimization objective and prior works in the field, including distributionally robust optimization (DRO) [Hashimoto et al., 2018] and our work, blind Pareto fairness (BPF) [Martinez et al., 2021]. The findings presented in this chapter are based on work in [Papadaki et al., 2022b, Papadaki et al., 2023].

4.1 Properly Pareto Subgroup Robustness through CVaR

Similar to previous learning settings, we consider the random variables $(X, Y) \in \mathscr{X} \times \mathscr{Y}$, where X represents the input features and Y represents the categorical targets.

These variables are generated from a distribution $p(X,Y)$. We assume that there is no prior knowledge available about the underlying groups or the sensitive labels associated with specific feature-target pairs. To address this challenge, we adopt the concept of subgroup robustness, similar to approaches in [Hashimoto et al., 2018, Williamson and Menon, 2019, Martinez et al., 2021].

Specifically, we let the random variable $L_{h,X,Y} = \ell(h(X),Y)$ represent the loss associated with hypothesis $h \in \mathscr{H}$. For a predefined probability $\rho \in (0,1)$ that determines the tail quantile level, the $(1-\rho)$-quantile function is defined as

$$q_{L_{h,X,Y}}(1-\rho) := \inf\left\{\beta \in \mathbb{R} : p(L_{h,X,Y} \leq \beta) \geq 1-\rho\right\}. \tag{4.1}$$

and the $(1-\rho)$-superquantile function, also known as Conditional Value-at-Risk (CVaR), at confidence level $(1-\rho)$ is defined as

$$CVaR_{(1-\rho)}(L_{h,X,Y}) = \mathop{\mathbb{E}}_{X,Y}[L_{h,X,Y} | L_{h,X,Y} \geq q_{L_{h,X,Y}}(1-\rho)]. \tag{4.2}$$

The quantity in Equation 4.2 measures the upper tail behaviour of the distribution $p(L_{h,X,Y})$. Furthermore, according to Theorem 1 in [Rockafellar et al., 2014], CVaR can be expressed as the following variational optimization problem:

$$CVaR_{(1-\rho)}(L_{h,X,Y}) = \min_{c \in \mathbb{R}} c + \frac{1}{\rho} \mathop{\mathbb{E}}_{X,Y}\left[(L_{h,X,Y} - c)_+\right], \tag{4.3}$$

where $(\cdot)+ = \max\{0,\cdot\}$ and the second term in the objective represents the regret of positive realizations of $L_{h,X,Y}$. The optimal argument that minimizes the objective in Equation 4.3 corresponds to the quantile $q_{L_{h,X,Y}}(1-\rho)$. If the selected loss function is bounded, i.e., $0 \leq \ell(h(x),y) \leq B$ for all $(x,y) \in \mathscr{X} \times \mathscr{Y}$, where $B > 0$, we can equivalently optimize over $c \in [0,B]$. Therefore, we can formulate the problem of learning a minimax group fair hypothesis without prior knowledge of demographics as

$$h^*, c^* = \arg\min_{h \in \mathscr{H}} \min_{c \in [0,B]} \left\{ c + \frac{1}{\rho} \mathop{\mathbb{E}}_{(X,Y) \sim p(X,Y)} \left[(L_{h,X,Y} - c)_+\right] \right\}. \tag{4.4}$$

Figure 4.1 depicts an example loss distribution and showcases how the measures (a)

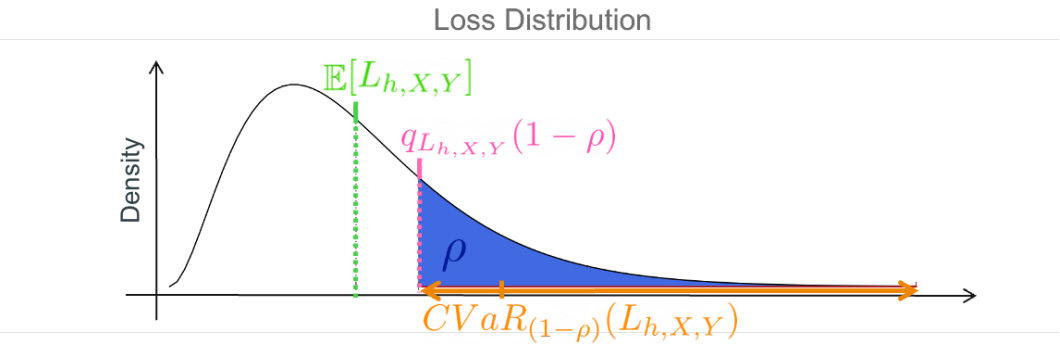

Figure 4.1: Example distribution of a random variable $L_{h,X,Y}$ associated with hypothesis $h \in \mathscr{H}$. We illustrate the measures (a) expected risk $\mathbb{E}[L_{h,X,Y}]$; (b) quantile $q_{L_{h,X,Y}}(1-\rho)$; and (c) Conditional Value-at-Risk $CVaR_{(1-\rho)}(L_{h,X,Y})$, for a tail quantile level ρ.

expected risk, (b) quantile, and (c) CVaR characterize it.

The optimization problem in Eq. 4.4 produces minimax fair solutions by optimizing for the worst tail risk with a sample size of ρ, or equivalently, the worst-performing samples that exceed the threshold c [Williamson and Menon, 2019]. However, this formulation allows for weakly Pareto optimal solutions [Miettinen, 2012], as it disregards any data that is not considered high-risk. Consequently, the learned model's utility on the low-risk population may be unnecessarily reduced, especially in scenarios where certain regions of the input space exhibit perfect separability.

Motivated by the concern of unnecessarily reducing the model's utility on the low-risk population, we propose a generalization of the CVaR objective, called relaxed CVaR (RCVaR). RCVaR aims to strike a balance between the worst-case group performance and the average performance by introducing a trade-off parameter $\varepsilon \in [0,1]$. The RCVaR objective is formulated as follows:

$$\min_{h\in\mathscr{H}} \left\{ (1-\varepsilon) CVaR_{(1-\rho)}(L_{h,X,Y}) + \varepsilon \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)} [L_{h,X,Y}] \right\} \tag{4.5}$$

$$= \min_{h\in\mathscr{H}} \min_{c\in[0,B]} \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)} \left[\left[(1-\varepsilon)\left(c + \tfrac{1}{\rho}(L_{h,X,Y}-c)_+\right) + \varepsilon L_{h,X,Y} \right], \right.$$

The proposed objective focuses on ensuring minimax fairness, where the worst

possible group is formed by the high risk samples subject to a predefined group size constraint ρ, while allowing for minimax properly Pareto optimal solutions [Geoffrion, 1968, Miettinen, 2012]. The minimax group performance guarantees imply that no group partition comprising more than a fraction ρ of the total population will experience a performance worse than the one obtained by a minimax model.

By varying the value of ε, we can control the trade-off between the worst-case group performance and the average performance. When $\varepsilon = 0$, the objective focuses solely on minimizing the worst-case group performance, while when $\varepsilon = 1$, it focuses solely on the average performance. Intermediate values of ε allow for different trade-offs between these two objectives.

Another way to control the trade-off between fairness and utility is by adjusting the worst group size parameter ρ, which determines the tail quantile level. In particular, as we also highlight later in Remark 4.1, there exists a critical value of ρ where the hypothesis solving Eq. 4.5 becomes a random classifier, resulting in no utility. Furthermore, as we increase ρ, the worst group size approaches the entire data population, leading to optimization for average performance. We emphasize that the values of ε and ρ are predefined and fixed, and therefore, we leave it to the policy maker(s) to determine the compromise between fairness and utility.

4.2 RCVaR Connection to Distributionally Robust Optimization

The (vanilla) CVaR formulation presented in Eq. 4.3 serves as the dual representation of DRO, as demonstrated in Proposition 3 in [Hashimoto et al., 2018] and Lemma 2.1 in [Duchi et al., 2020]. In particular, according to [Shapiro et al., 2009, Duchi and Namkoong, 2020], the minimization of CVaR can be expressed as

$$\min_{h\in\mathscr{H}} \sup_{q(X,Y)\in\mathcal{Q}_\rho} \mathbb{E}_{(X,Y)\sim q(X,Y)} [L_{h,X,Y}] \tag{4.6}$$

where the uncertainty set Q_ρ is defined as:

$$\begin{aligned} Q_\rho = & \left\{ q(X,Y) | D_\infty(p(X,Y) \parallel q(X,Y)) \leq \log \tfrac{1}{\rho} \right\} \\ = & \left\{ q(X,Y) | \exists q'(X,Y) : p(X,Y) = \beta q(X,Y) + (1-\beta) q'(X,Y), \beta \in [\rho, 1] \right\} \end{aligned}$$

This uncertainty set corresponds to distributions with the worst group size of at least ρ.

Given that the RCVaR objective is equivalent to CVaR when $\varepsilon = 0$, it also encompasses DRO as a special case. Nevertheless, the RCVaR objective offers an additional advantage compared to DRO. For small values of ε, specifically when $\varepsilon \approx 0$, RCVaR can achieve minimax fairness solutions that are also properly Pareto optimal, which is not guaranteed by the existing formulation of DRO.

4.3 RCVaR Connection to Blind Pareto Fairness

Next, we formally define the relation of the proposed objective, RCVaR, and our BPF objective in [Martinez et al., 2021]. To establish this connection, we consider the setting described in Section 4.1.

We introduce another random variable G, which indicates whether a specific input-target pair belongs to the worst-performing group. We define $G = 1$ for data points belonging to the worst group and $G = 0$ for the remaining data points that do not belong to the worst group.

Lemma 4.1 (RCVAR equivalence to BPF)**.** *Let $\varepsilon \approx 0$ be the lower bound constraint for the worst partition density, $p(G = 1|X,Y) > \varepsilon$. Let also the probability ρ represent the worst group size, $p(G = 1) = \rho$. Then, we have that*

$$\min_{h \in \mathscr{H}} \max_{\substack{p(G=1|X,Y) \\ \text{s.t. } p(G=1)=\rho \\ p(G=1|X,Y)>\varepsilon}} \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)} \left[\frac{p(G=1|X,Y)}{p(G=1)} L_{h,X,Y} \right] \tag{BPF}$$

$$= \min_{h \in \mathscr{H}} (1-\varepsilon) CVaR_{1-\rho}(L_{h,X,Y}) + \varepsilon \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)} [L_{h,X,Y}] \tag{RCVaR}$$

The proof is presented in Section 4.4. Due to this equivalence, we argue

that the results presented in Lemma 3.2 in [Martinez et al., 2021], which proves the existence of a critical partition size ρ that leads to the uniform classifier for sufficiently small ε values; and Lemma 3.3 in [Martinez et al., 2021], that studies the penalty in performance when we use subgroup robustness instead of known groups, apply also to our case for $\varepsilon \approx 0$.

In Remark 4.1, we re-purpose the results of Lemma 3.2 in [Martinez et al., 2021] for the existence of a critical ρ value that leads to the uniform classifier for sufficiently small ε values, by stating the impact of the threshold c in the resulting hypothesis in our objective, RCVaR.

Remark 4.1. *The hypothesis that determines the* $(1-\rho)$*-quantile* c*, for which any realizations of* $L_{h,X,Y}$ *are at most* c*, is the uniform classifier* $\overline{h} : \overline{h}_y(X) = \frac{1}{|\mathscr{Y}|} \forall y \in \mathscr{Y}$.

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

Another advantage of RCVaR is its suitability for federated learning. Compared to BPF, RCVaR can be easily federalized, allowing for efficient implementation in distributed settings. In the following Chapter, we will delve into the specifics of how the RCVaR objective can be adapted to the federated setting, highlighting its advantages over BPF in distributed learning scenarios.

4.4 Proofs

In this section, we analyze the relation of the proposed objective, RCVaR, and our BPF objective in [Martinez et al., 2021]. To establish this connection, we consider

the setting described in Section 4.1.

Proof [**Lemma 4.1**] We first slightly adjust the notation and instead of representing the input features and categorical targets as separate variables, we denote them as a single variable $Z = (X,Y)$. Let the worst group size be equal to the probability ρ, i.e., $p(G=1) = \rho$. Let also the constraint $p(G=1|Z) > \varepsilon$. We can re-formulate the BPF objective, given in Eq. 2.7, as

$$\min_{h\in\mathscr{H}} \max_{\substack{p(G=1|Z) \\ \text{s.t. } p(G=1)=\rho \\ p(G=1|Z)>\varepsilon}} \mathop{\mathbb{E}}_{Z\sim p(Z)} \left[\frac{p(G=1|Z)}{p(G=1)} L_{h,Z}\right] = \min_{h\in\mathscr{H}} \max_{\substack{\lambda(Z)\in Q_{\varepsilon,\rho} \\ \text{s.t. } \int_{\mathscr{Z}} \lambda(z)dz=1}} \int_{\mathscr{Z}} \lambda(z)\ell(h;z)dz, \tag{4.7}$$

where $Q_{\varepsilon,\rho} = \left\{\lambda(\cdot) : \lambda(z) \in \left[\frac{p(z)}{\rho}\varepsilon, \frac{p(z)}{\rho}\right]\right\}$ and $\lambda(Z) \in Q_{\varepsilon,\rho}$ represents the density of the input-target pair variable Z.

Next, we leverage the Lagrange duality approach to demonstrate that the BPF objective can be transformed into the RCVaR objective. We define the Lagrangian of the RHS of Eq. 4.7 as

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

where μ^* is the Lagrange multiplier of the constraint in Eq. 4.7. For a fixed hypothesis $h \in \mathscr{H}$, the optimal density $\lambda^*(Z)$ satisfies

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

Furthermore, using the fact that $\int_{\mathscr{Z}} \lambda(z)dz = 1$, we can compute the Lagrange

multiplier μ^* as follows

$$\int_{\ell(h;z)\leq\mu^*} \frac{\varepsilon}{\rho} p(z)dz + \int_{\ell(h;z)>\mu^*} \frac{p(z)}{\rho} dz = 1$$

$$\implies \frac{\varepsilon}{\rho} \int_{\mathscr{Z}} p(z)dz + \left(\frac{1-\varepsilon}{\rho}\right) \int_{\ell(h;z)>\mu^*} p(z)dz = 1 \implies \int_{\ell(h;z)>\mu^*} p(z)dz = \frac{\rho-\varepsilon}{(1-\varepsilon)}$$

$$\implies \int_{\ell(h;z)\leq\mu^*} p(z)dz = 1 - \frac{\rho-\varepsilon}{(1-\varepsilon)} \iff \mu^* = F^{-1}(1-\rho'), \qquad (\text{with } \rho' = \tfrac{\rho-\varepsilon}{(1-\varepsilon)})$$

where $F^{-1}(\cdot)$ corresponds to the inverse of $F(\cdot)$, and $F(L_{h,Z})$ is the cumulative distribution function of $L_{h,Z}$. Then, by substituting the optimal density $\lambda^*(Z)$ and μ^* in the BPF objective in Eq. 4.7 we get

$$\min_{h\in\mathscr{H}} \int_{\ell(h;z)\leq\mu^*} \varepsilon \frac{p(z)}{\rho} \ell(h;z)dz + \int_{\ell(h;z)>\mu^*} \frac{p(z)}{\rho} \ell(h;z)dz$$

$$= \min_{h\in\mathscr{H}} \frac{\varepsilon}{\rho} \int_{\mathscr{Z}} p(z)\ell(h;z)dz + \frac{(1-\varepsilon)}{\rho} \int_{\ell(h;z)>F^{-1}(1-\rho')} p(z)\ell(h;z)dz$$

Recall that $\rho = \int_{\ell(h;z)>F^{-1}(1-\rho')} p(z)dz$. Thus, we have that

$$\min_{h\in\mathscr{H}} \tfrac{\varepsilon}{\rho} \textstyle\int_{\mathscr{Z}} p(z)\ell(h;z)dz + \tfrac{(1-\varepsilon)}{\rho} \int_{\ell(h;z)>F^{-1}(1-\rho')} p(z)\ell(h;z)dz$$

$$= \min_{h\in\mathscr{H}} \frac{\varepsilon}{\rho} \mathop{\mathbb{E}}_{Z\sim p(Z)} [L_{h,Z}] + (1-\varepsilon) \mathop{\mathbb{E}}_{Z\sim p(Z)} \left[L_{h,Z} | L_{h,Z} > F^{-1}\left(1-\rho'\right)\right]$$

$$= \min_{h\in\mathscr{H}} \frac{\varepsilon}{\rho} \mathop{\mathbb{E}}_{Z\sim p(Z)} [L_{h,Z}] + (1-\varepsilon) CVaR_{1-\rho'}(L_{h,Z})$$

$$= \min_{h\in\mathscr{H}} \underbrace{\varepsilon \mathop{\mathbb{E}}_{Z\sim p(Z)}[L_{h,Z}] + (1-\varepsilon)CVaR_{1-\rho'}(L_{h,Z})}_{\text{RCVaR for probability } \rho'=\frac{\rho-\varepsilon}{(1-\varepsilon)} \text{ and trade-off } \varepsilon} + \underbrace{\frac{\varepsilon(1-\rho)}{\rho} \mathop{\mathbb{E}}_{Z\sim p(Z)}[L_{h,Z}]}_{\text{Add'l term}} \tag{4.8}$$

Eq. 4.8 shows that BPF is equivalent to RCVaR plus an additional error term for the same ε and $\rho' = \frac{\rho-\varepsilon}{(1-\varepsilon)}$. We remark that for sufficiently small ε values (i.e., $\varepsilon \approx 0$) the additional term is negligible and probability $\rho' \approx \rho$, making the two objectives equivalent. □

4.5 Summary

In this chapter, we tackle the challenge of optimizing the performance of the worst-case group in centralized machine learning settings when demographic information is unavailable. We introduce an alternative optimization objective that aims to improve the performance of the worst-performing group while considering a minimum group size constraint. We establish connections between our objective and prior works in the field.

In the next chapter, we extend the proposed objective to ensure global group fairness in federated learning settings in the absence of demographic information. We develop an algorithm that effectively solves a smooth approximation of the proposed fairness objective within a federation. We thoroughly examine the performance guarantees of our algorithm in terms of optimization error and generalization. Furthermore, we conduct a series of experiments to validate the efficiency and effectiveness of our approach.

Chapter 5

Federated Global Group Fairness without Demographics

In this chapter, we shift our focus from centralized learning settings, where data is concentrated in a single entity, to federated learning settings characterized by distributed data across multiple clients who do not share their data. Our primary objective is to tackle the challenge of achieving minimax global group fairness in federated learning when the local sensitive populations and their corresponding labels are completely unknown to the participating parties.

To address this problem, we extend the concept of RCVaR introduced in Chapter 4 and incorporate the unique roles of individual clients in the federated learning context. We present an algorithm that solves a smoothed version of the proposed problem within the federation, that enjoys convergence and excess risk guarantees. Through extensive empirical evaluations, we demonstrate that our approach effectively improves the performance of the worst-performing group without unnecessarily sacrificing the average performance. Furthermore, our method outperforms relevant baselines and offers a diverse set of solutions, allowing for flexible fairness-utility trade-offs. Finally, we showcase the practicality and benefits of our approach even in cases where demographic information is available. The content of this chapter is presented in [Papadaki et al., 2022b, Papadaki et al., 2023].

5.1 RCVaR for Federated Global Group Fairness

We adopt the learning setting discussed in Section 4.1, incorporating a random variable $K \in \{1,\dots,|\mathscr{K}|\}$ representing the clients participating in the federation. In particular, we let each client $k \in \mathscr{K}$ possess data modelled by its own local distribution $p(X,Y|K=k) = p(X|K=k)p(Y|X,K=k)$. Therefore, the data of the entire federation can be described via the mixture distribution $p(X,Y) = \sum_{k\in\mathscr{K}} p(K=k)p(X,Y|K=K)$.

Additionally, we introduce a random variable $L_{h,X,Y|K=k} := \ell(h(X),Y)$ representing the local loss induced by a hypothesis h in client k. Then, we can extend the RCVaR optimization problem from centralized learning, as given in Eq. 4.5, to the federated learning setting as follows,

$$\begin{aligned}\min_{h\in\mathscr{H}} \left\{ (1-\varepsilon) CVaR_{(1-\rho)}(L_{h,X,Y}) + \varepsilon \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y)} [L_{h,X,Y}] \right\} = \\ \min_{h\in\mathscr{H}, c\in[0,B]} \mathop{\mathbb{E}}_{K} \left[\mathop{\mathbb{E}}_{X,Y|K=k} \left[(1-\varepsilon)\left(c + \tfrac{1}{\rho}(L_{h,X,Y|K=k} - c)_+\right) + \varepsilon L_{h,X,Y|K=k} \right] \right],\end{aligned} \tag{5.1}$$

where the threshold c is assumed to be shared across the clients participating in the federation. Selecting a common threshold across clients allows us to identify the samples belonging to the global (across the various clients) worst group and the non-high-risk group. Therefore, the adversary has the flexibility to allocate more weight to clients with worse performances by shifting a portion of its budget, denoted as ρ_k, from clients with higher utility, while ensuring that the overall partition size is ρ.

5.1.1 Finite sample RCVaR formulation:

In real-world scenarios, each client holds only a finite dataset $D_k = \{(x_i^k, y_i^k)\}_{i=1,\dots,n_k}$, sampled from the true distribution $p(X,Y|K=k)$, with $D = \bigcup_{k\in\mathscr{K}} D_k$ being the dataset containing all the data samples available across clients with total size $n = \sum_{k\in\mathscr{K}} n_k$. Hence, in the sequel we will be using the empirical form of the RCVaR learning

problem given by

$$\min_{\boldsymbol{\theta}\in\Theta}\min_{c\in[0,B]}\frac{n_k}{n}\sum_{k\in\mathscr{K}}\frac{1}{n_k}\sum_{i=1}^{n_k}f(\boldsymbol{\theta},c;x_i^k,y_i^k), \tag{5.2}$$

where

$$f(\boldsymbol{\theta},c;x,y)=(1-\varepsilon)[c+\frac{1}{\rho}(\ell(\boldsymbol{\theta};x,y)-c)_+]+\varepsilon\ell(\boldsymbol{\theta};x,y), \tag{5.3}$$

and $\boldsymbol{\theta}\in\Theta$ is the vector parametrizing the hypothesis $h\in\mathscr{H}$. Note that we slightly adjust the notation and parametrize the loss function ℓ using the vector $\boldsymbol{\theta}$ instead of h.

Next, we propose a federated learning algorithm to solve the empirical RCVaR optimization problem that relies on a smoothed version $\tilde{f}(\cdot)$ of the original non-smooth function $f(\cdot)$. We also provide an alternative heuristic method for optimizing Eq. 5.2, along with the corresponding empirical results, in Appendix C.

5.2 Smooth Approximation of RCVaR

We consider a federated learning setting where each client might use the local data samples more than once during the training process. This realistic scenario poses challenges in terms of algorithmic design and analysis, since in order to develop a simple algorithm with strong theoretical guarantees, we require a continuously differentiable objective for all data samples $z=(x,y)$. Nevertheless, even with a smooth loss functions ℓ, the auxiliary function f in our current objective is non-smooth due to the presence of the ReLU function $(\cdot)_+$.

To overcome this issue, we introduce a proxy problem for the RCVaR in Eq. 5.2 that relies on a smooth approximation. We first offer the definition of a smoothed plus function $s:\mathbb{R}\to\mathbb{R}_+$.

Definition 5.1 (Smoothed plus function)**.** *For a $\gamma\in\mathbb{R}_+$ and for any $m\in\mathbb{R}$, a $(\frac{2}{\gamma})-$smooth convex function $s:\mathbb{R}\to\mathbb{R}_+$ is defined such that*

$$0\leq s(m)-(m)_+\leq\gamma. \tag{5.4}$$

Our optimization method and analysis are applicable to a wide range of functions, e.g., soft ReLU, the piecewise-quadratic smoothed plus function, etc., that adhere to the conditions outlined in Definition 5.1, rather than a very specific smoothed plus function. Note that smaller values of parameter γ will be a better approximation to the original plus function.

We now introduce the empirical smooth approximation of RCVaR as follows

$$\min_{\boldsymbol{\theta}\in\Theta}\min_{c\in[0,B]}\frac{1}{n}\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\tilde{f}(\boldsymbol{\theta},c;x_i^k,y_i^k), \tag{5.5}$$

where

$$\tilde{f}(\boldsymbol{\theta},c;x,y)=(1-\varepsilon)[c+\frac{1}{\rho}s(\ell(\boldsymbol{\theta};x,y)-c)]+\varepsilon\ell(\boldsymbol{\theta};x,y).$$

5.3 Federated Smoothed RCVaR Algorithm

Next, we introduce a federated learning algorithm called FedSRCVaR, that solves the optimization problem defined in Eq. 5.5 by leveraging multi-pass minibatch stochastic gradient descent (SGD).

We assume that (1) all clients are available to participate during the training process, (2) every client uses a batch size $b_k \leq n_k$ of data samples at each training iteration, and (3) the training process involves a total of T communication rounds between the clients and the central server. FedSRCVaR algorithm, summarized in Algorithm 4, performs the following successive steps for T communication rounds:

1. the clients receive the global model-threshold pair $(\boldsymbol{\theta}^t, c^t)$ of the current round from the server;

2. The clients perform a single local gradient descent update on the model parameters and the threshold using b_k-samples;

3. The clients return the updated pair $(\boldsymbol{\theta}_k^{t+1}, c_k^{t+1})$ to the server;

4. Finally, the server produces the new model-threshold pair $(\boldsymbol{\theta}^{t+1}, c^{t+1})$ by averaging the received client updates.

Algorithm 4 FEDERATED SMOOTHED-RCVAR (FEDSRCVAR) ALGORITHM

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

Server initializes $\boldsymbol{\theta}^1$ randomly.
Server sets $c^1 = B = 1$.
for $t = 1$ to T **do**
  Server **broadcasts** global model-threshold pair $(\boldsymbol{\theta}^t, c^t)$
  for each client $k \in \mathscr{K}$ **in parallel do**
    Randomly sample a data batch with size b_k
    $\boldsymbol{\theta}_k^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta_t \nabla_{\theta^t} \left\{ \frac{1}{b_k} \sum_{i=1}^{b_k} \tilde{f}(\boldsymbol{\theta}^t, c^t; x_i^k, y_i^k) \right\}$
    $c_k^{t+1} \leftarrow c^t - \eta_t \nabla_{c^t} \left\{ \frac{1}{b_k} \sum_{i=1}^{b_k} \tilde{f}(\boldsymbol{\theta}^t, c^t; x_i^k, y_i^k) \right\}$
    Return local model-threshold pair $(\boldsymbol{\theta}_k^{t+1}, c_k^{t+1})$
  end for
  Server computes
  $$\boldsymbol{\theta}^{t+1} \leftarrow \sum_{k \in \mathscr{K}} \frac{b_k}{\sum_{k \in \mathscr{K}} b_k} \boldsymbol{\theta}_k^{t+1}, \quad c^{t+1} \leftarrow proj_{[0,B]} \left(\sum_{k \in \mathscr{K}} \frac{b_k}{\sum_{k \in \mathscr{K}} b_k} c_k^{t+1} \right)$$
end for

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

We denote $proj_{[0,B]}$ the metric projection operator, where $[0,B]$ represents the desired range for c, to ensure that the threshold c remains within a valid range. Note that the server uses the relative weights $\frac{b_k}{\sum_{k \in \mathscr{K}} b_k}$, with b_k being the batch size of client k. These weights are proportional to the respective client's dataset size n_k. Furthermore, we consider an averaging scheme, which instead of returning the model-threshold pair $(\boldsymbol{\theta}^T, c^T)$ from the last communication round, our algorithm outputs the average of the total communications $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$. We note that SGD-based algorithms often exhibit oscillations around global minimizers, and the method of averaging iterates helps mitigate this oscillation effect, leading to a solution that is close to the optimal solution up to certain constants in the case of convex optimization, as discussed in [Shamir and Zhang, 2013].

Our algorithm does not vastly increase the communication overhead in the federated learning procedure, since the communication between clients and server is limited to exchanging the updated model – which is common practice in many

federated learning approaches –, and the estimated threshold. This is insignificant compared to the communication costs and additional privacy concerns of sharing the per-sample (naive) federalization of the BPF objective, as we extensively discuss in the sequel.

5.3.1 Benefits of RCVaR over BPF Federalization

RCVaR offers advantages in terms of easy federalization compared to BPF. In Algorithm 4, the federalization of RCVaR requires only the exchange of the updated model-threshold pair between clients and the server. This efficient exchange allows for achieving a solution equivalent to our centralized BPF objective in [Martinez et al., 2021]. In contrast, federalizing BPF poses challenges, such as (a) failure to provide guarantees that the produced global model is equivalent to centralized BPF; or (b) significantly amplifying the computation and communication costs of the federated learning procedure, and raising additional privacy concerns.

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

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

need to share information about their local loss distributions.

5.3.2 Algorithmic Analysis

In this section, we examine the performance of Algorithm 4 by assessing the associated convergence rate and expected excess risk. Our analysis relies on two key assumptions.

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

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

We note that several losses in machine learning (e.g. logistic loss) satisfy – or can be adjusted to satisfy – Assumption 5.1. For instance, hinge loss is non-smooth, but we can use instead the generalized smooth hinge loss. Common examples for Assumption 5.2 are logistic regression and neural networks with only linear layers.

Under these two assumptions, we can establish the core properties of the smooth and non-smooth functions, f and $\tilde{f}$, required for our analysis, in Lemma 5.1.

Lemma 5.1. *Let Assumption 5.1 hold. Let also s be a $\frac{2}{\gamma}-$smooth convex function. Then,*

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

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

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

$$f(\boldsymbol{\theta}, c; x, y) \leq \tilde{f}(\boldsymbol{\theta}, c; x, y) \leq f(\boldsymbol{\theta}, c; x, y) + \frac{(1-\varepsilon)}{\rho}\gamma. \tag{5.6}$$

We remark that Eq. 5.6 bounds the smoothed function $\tilde{f}$ using f which allows us to express our guarantees in terms of f, but importantly prove them in terms of the smooth function $\tilde{f}$. We offer the proof of Lemma 5.1 in Section 5.6.1.

5.3.2.1 Convergence Guarantees of FedSRCVaR

We begin by characterizing the optimization error given by

$$\mathscr{E}_{opt} = \mathop{\mathbb{E}}_{\mathscr{S},D}\left[\sum_{(x,y)\in D}\frac{f(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;x,y)}{n}\right] - \mathop{\mathbb{E}}_{D}\left[\sum_{(x,y)\in D}\frac{f(\boldsymbol{\theta}^\star_D,c^\star_D;x,y)}{n}\right],$$

where $(\overline{\boldsymbol{\theta}}_T,\bar{c}_T)$ is the average model-threshold pair after T rounds of Algorithm 4, $(\boldsymbol{\theta}^*_D,c^\star_D)$ is the model-threshold pair that minimizes the smoothed objective in Eq. 5.5 and the outer expectation in the first term is taken over the randomness induced by our randomized algorithm $\mathscr{S}$ and the samples D, and in the second term with respect to the dataset D. This error captures how well the produced pair $(\overline{\boldsymbol{\theta}}_T,\bar{c}_T)$ approximates the optimal empirical pair $(\boldsymbol{\theta}^\star_D,c^\star_D)$ in terms of the true (non-smooth) objective function.

The next lemma offers a bound to the optimization error $\mathscr{E}_{opt}$. The proof leverages results for stochastic (non-strongly) convex optimization using first-order oracles obtained in the seminal work of Nemirovski and Yudin [Nemirovski and Yudin, 1983], also presented in Lemma 5.7, for convenience.

Lemma 5.2 (Convergence of FedSRCVaR – convex setting)**.** *Let assumptions 5.1 - 5.2 hold, $(\boldsymbol{\theta}^\star_D,c^\star_D)$ be the minimizer of Eq. 5.5, and a constant learning rate η_t. Then, for the model-threshold pair $(\overline{\boldsymbol{\theta}}_T,\bar{c}_T)$ provided after T communication rounds of Algorithm 4, we have that*

$$\mathscr{E}_{opt} \leq \frac{1}{2}\left(\frac{M^2+B^2}{\eta T} + \frac{G^2_{\rho,\varepsilon}\eta}{\sum_{k\in\mathscr{K}} b_k}\right) + \frac{(1-\varepsilon)\gamma}{\rho}. \tag{5.7}$$

The proof of Lemma 5.2 is provided in Section 5.6.2.1. By picking the step size which minimizes the RHS of Eq. 5.7, that is $\eta_{opt} = \frac{1}{G_{\rho,\varepsilon}\sqrt{T}}\sqrt{\sum_{k\in\mathscr{K}} b_k(M^2+B^2)}$, the

upper bound becomes

$$\mathscr{E}_{opt} \leq \frac{\sqrt{M^2+B^2}}{\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}} b_k}} G_{\rho,\varepsilon} + \frac{(1-\varepsilon)\gamma}{\rho}. \tag{5.8}$$

For a constant learning-rate policy, Lemma 5.2 shows that our algorithm finds a model-threshold pair $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ after T communication rounds that guarantees an optimization error of order $O\left(\frac{1}{\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}} b_k}}\right)$, hence, it decays with the square root of the product of the number of communication rounds times the total batch size. For $T = \sum_{k\in\mathscr{K}} b_k$, this bound becomes $O\left(\frac{1}{T}\right)$, and reduces with the number of communication rounds.

We highlight that the process of saving model parameters at each communication round necessitates sufficient memory resources, particularly when conducting a large number of iterations or dealing with a significant number of model parameters. In scenarios where the server's resources are limited, we provide an upper bound on the optimization error in Lemma 5.3. This bound quantifies how effectively a model-threshold pair $(\overline{\boldsymbol{\theta}}_{V:T}, \overline{c}_{V:T})$, averaged from round V to round T using Algorithm 4 (with $V \leq T$), approximates the optimal empirical pair $(\boldsymbol{\theta}_D^\star, c_D^\star)$ in terms of the true (non-smooth) objective function. The proof builds upon the results for stochastic gradient descent established in [Nemirovski et al., 2009], which are summarized in Lemma 5.8.

Lemma 5.3 (Convergence of FedSRCVaR – server-side efficient and convex settings)**.** *Let assumptions 5.1 - 5.2 hold, $(\boldsymbol{\theta}_D^\star, c_D^\star)$ be the minimizer of Eq. 5.5, and a constant learning rate $\eta_t = \frac{1}{G_{\rho,\varepsilon}\sqrt{T}}\sqrt{\sum_{k\in\mathscr{K}} b_k(M^2+B^2)}$. Then, for the model-threshold pair $(\overline{\boldsymbol{\theta}}_{V:T}, \overline{c}_{V:T})$ provided from V to T communication rounds of Algorithm 4, with $V \leq T$,*

we have that

$$\mathop{\mathbb{E}}_{\mathscr{S},D}\left[\sum_{(x,y)\in D}\frac{f(\overline{\boldsymbol{\theta}}_{V:T},\bar{c}_{V:T};x,y)}{n}\right]-\mathop{\mathbb{E}}_{D}\left[\sum_{(x,y)\in D}\frac{f(\boldsymbol{\theta}^{\star}_{D},c^{\star}_{D};x,y)}{n}\right]$$

$$\leq\frac{\sqrt{M^2+B^2}}{\sqrt{T}\sqrt{\sum\limits_{k\in\mathscr{K}}b_k}}G_{\rho,\varepsilon}\left(\frac{1}{2}+\frac{2T}{T-V+1}\right)+\frac{(1-\varepsilon)\gamma}{\rho}. \tag{5.9}$$

The proof of Lemma 5.3 can be found in Section 5.6.2.1. We remark that the outer expectation in the first term of the LHS in Eq. 5.9 is taken over the randomness induced by our randomized algorithm $\mathscr{S}$ and the samples D, and in the second term with respect to the dataset D.

Lemma 5.3 demonstrates that our algorithm discovers a model-threshold pair $(\overline{\boldsymbol{\theta}}_{V:T},\bar{c}_{V:T})$, averaged from round V to round T with $V\leq T$, that ensures an optimization error of the order $O\left(\frac{1}{\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}}b_k}}\left(\frac{1}{2}+\frac{2T}{T-V+1}\right)\right)$. The error bound in Lemma 5.3 is less favourable compared to the error bound presented in Lemma 5.2. However, it offers a more resource-efficient solution compared to the previous one. Additionally, for a round $V\leq\frac{T}{2}$, Lemma 5.3 yields an optimization error that behaves as $O\left(\frac{1}{\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}}b_k}}\right)$, which is also the optimal bound derived for the scenario where the output is averaged over T rounds, discussed in Lemma 5.2. This is due to the equivalence, up to a constant factor, of the first term of the expression on the right-hand side (RHS) in Eq. 5.9 with the first term of the expression on the RHS in Eq. 5.8. It should be noted that the bound relies on utilizing the same optimal constant learning rate η_{opt} as in Eq. 5.8.

Non-Convex Settings: In our previous analysis, we formally established that Fed-SRCVaR algorithm convergences to the global minima for convex functions f and $\tilde{f}$. Now, we extend our investigation by relaxing the assumption of convexity and expanding the performance guarantees to encompass settings where these functions are non-convex. This extension becomes particularly relevant when training deep neural networks (DNNs), where dealing with a non-convex loss function is prevalent.

Nevertheless, finding the global minima for non-convex functions is known to

be NP-hard [Murty and Kabadi, 1987]. Instead, we can leverage the fact that for a β–smooth loss function ℓ and a $\frac{2}{\gamma}$-smooth plus function s, the auxiliary function $\tilde{f}$ is $\left(\frac{(1-\varepsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\right)$–smooth, to establish the convergence of the FedSRCVaR algorithm to a stationary point. In particular, the smoothness property of $\tilde{f}$ ensures that as we approach a stationary point, the gradient of the function tends to zero. Consequently, it suffices to show that the norm of the gradient decreases in order to show convergence to a stationary point.

In Lemma 5.4 we prove the convergence of the FedSRCVaR algorithm, as summarized in Algorithm 4, for a non-convex and smooth function $\tilde{f}$ that satisfies the lower bound condition, by demonstrating that the model θ^T obtained from the last iteration of the proposed algorithm converges to a stationary point with probability 1.

For the proof, we take into account the assumption of bounded variance of the stochastic gradient, and the Robins and Monro conditions [Robbins and Monro, 1951] for the positive learning rates that are used in the standard SGD analysis for non-convex smooth functions.

Assumption 5.3 (Bounded Variance of the Stochastic Gradient)**.** *Every client $k \in \mathcal{K}$ queries an unbiased minibatch stochastic gradient such that $\nabla\left\{\sum\limits_{k\in\mathcal{K}}\sum\limits_{i=1}^{b_k}\frac{\tilde{f}(\boldsymbol{\theta},c;x_i^k,y_i^k)}{\sum_{k\in\mathcal{K}} b_k}\right\}$ of $\hat{F}$, i.e., $\mathbb{E}_{x,y}\left[\frac{1}{\sum_{k\in\mathcal{K}} b_k}\nabla\left\{\sum\limits_{k\in\mathcal{K}}\sum\limits_{i=1}^{b_k}\tilde{f}(\boldsymbol{\theta},c;x_i^k,y_i^k)\right\}\right]=\nabla\hat{F}(\boldsymbol{\theta},c)$, has $\frac{\sigma^2}{\sum_{k\in\mathcal{K}} b_k}$−uniformly bounded variance, i.e., $\forall(\boldsymbol{\theta},c)\in\Theta\times[0,B]$*

$$\mathbb{E}\left[\left|\left|\frac{1}{\sum\limits_{k\in\mathcal{K}} b_k}\nabla\left\{\sum_{k\in\mathcal{K}}\sum_{i=1}^{b_k}\tilde{f}(\boldsymbol{\theta},c;x_i^k,y_i^k)\right\}-\nabla\hat{F}(\boldsymbol{\theta},c)\right|\right|^2\right]\leq\frac{\sigma^2}{\sum\limits_{k\in\mathcal{K}} b_k}.$$

Assumption 5.4 (Robins and Monro conditions [Robbins and Monro, 1951])**.** *The learning rate sequence $\{\eta_t\}_{t=1}^{\infty}$ is a deterministic sequence that satisfy $\eta_t>0\forall t$, $\sum\limits_{t=1}^{\infty}\eta_t=\infty$ and $\sum\limits_{t=1}^{\infty}\eta_t^2<\infty$.*

We are now ready to show that for non-convex smooth functions, FedSRCVaR finds an approximate stationary point.

Lemma 5.4 (Probabilistic convergence of FedSRCVaR – non-convex setting)**.** *Let the feasible region Θ be $\mathbb{R}^d$, a $\beta-$smooth loss function of range $[0,B]$, with $B=1$,*

and the assumptions 5.3 - 5.4 hold. Then, for the model-threshold pair $(\boldsymbol{\theta}^T, c^T)$ *provided in the last round of Algorithm 4, we have that* $\left|\left|\nabla \hat{F}(\boldsymbol{\theta}^T, c^T)\right|\right| \to 0$ *as* $T \to \infty$ *with probability* 1.

The proof, outlined in Section 5.6.2.1, builds upon the results presented in [Orabona, 2020], which in turn rely on the seminal work of Bertsekas and Tsitsiklis [Bertsekas and Tsitsiklis, 1999]. We highlight that Lemma 5.4 offers a notable advantage by demonstrating the convergence of the final iteration's pair $(\boldsymbol{\theta}^T, c^T)$ of FedSRCVaR to zero with high probability, which holds practical significance. This is favourable compared to other studies of stochastic gradient methods in non-convex settings (e.g., [Allen-Zhu, 2018, Mertikopoulos et al., 2020, Soma and Yoshida, 2020, Sebbouh et al., 2021]), which typically establish convergence rates for the best iterate (without specifying which one it is) or some randomly selected model from the T iterations, rather than guaranteeing the performance of a specific round model.

5.3.2.2 Expected Excess Risk

Next, we characterize the excess risk, given by [Hardt et al., 2016b]

$$\mathscr{E}_r = \mathop{\mathbb{E}}_{\mathscr{S},D}\left[\mathop{\mathbb{E}}_{K}\left[\mathop{\mathbb{E}}_{X,Y}[f(\overline{\boldsymbol{\theta}}_T, \overline{c}_T; X, Y|k)]\right]\right] - \mathop{\mathbb{E}}_{D}\left[\sum_{(x,y)\in D} \frac{f(\boldsymbol{\theta}^\star_D, c^\star_D; x, y)}{n}\right],$$

where $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ is the average model-threshold pair given by Algorithm 4 using dataset D and $(\boldsymbol{\theta}^\star_D, c^\star_D)$ is the optimal solutions pair that minimizes the smoothed empirical objective in Eq. 5.5. The outer expectation of the first term is taken over the randomness induced by our algorithm $\mathscr{S}$ and of samples D, and in the second term with respect to the samples D. The excess risk of measures the difference between the expected population risk computed using the produced $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ and the expected minimum empirical risk that is computed using the empirical optimal pair $(\boldsymbol{\theta}^\star_D, c^\star_D)$.

The following lemma – which relies on the excess risk analysis for stochastic gradient methods in [Hardt et al., 2016b] – offers a characterization of $\mathscr{E}_r$.

Lemma 5.5 (Excess Risk Analysis of FedSRCVaR). *Let assumptions 5.1 and 5.2 hold. Let also the learning rate* $\eta = \sqrt{n\Big(\sum_{k\in\mathscr{K}} b_k\Big)}\frac{\sqrt{M^2+B^2}}{G_{\rho,\varepsilon}\sqrt{T(n+2T)}}$ *and* $\gamma = \frac{2G_{\rho,\varepsilon}^2}{(1-\varepsilon+\varepsilon\rho)^2}\eta$. *Then, for* T *communication rounds of Algorithm 4 that satisfy*

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

we have that

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

The formal proof is provided in Section 5.6.2.2. The bound in Lemma 5.5 indicates that, for a fixed step-size η and for choices of γ and T that satisfy the conditions stated above, our algorithm produces a model-threshold pair $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ that satisfies an excess risk behaving as $O\left(\frac{1}{\sqrt{\sum_{k\in\mathscr{K}} b_k}}\sqrt{\frac{2}{n}+\frac{1}{T}}\right)$. Since the excess risk can be decomposed into a stability term and an empirical optimization error term (see for example [Chen et al., 2018b]), the upper bound above shows how to effectively improve the overall performance by balancing the trade-off between optimization and generalization.[1]

We note that the tuned learning rate that minimizes the excess risk is smaller than the optimal step size we selected for our convergence guarantees, i.e., $\eta = \eta_{opt}\sqrt{\frac{n}{n+2T}}$. From [Hardt et al., 2016b], we have that the bound above is composed by the empirical optimization error and algorithmic stability. Thus, we can directly get from Lemma 5.5 that FedSRCVaR has uniform stability of $\zeta \leq \frac{TG_{\rho,\varepsilon}^2\eta}{n\sum_{k\in\mathscr{K}} b_k}$. In contrast to the optimization error, the stability term scales with the communication rounds. For $T = n$, the result of Lemma 5.5 is of order $O\left(\frac{1}{\sqrt{T\sum_{k\in\mathscr{K}} b_k}}\right)$, and decreases with the square root of communication rounds times the total of batch size. Additionally, if we also have $T = \sum_{k\in\mathscr{K}} b_k$ this quantity further improves and becomes $O\left(\frac{1}{T}\right)$.

On the other hand, for $T \to \infty$, our bound is of order $O\left(\frac{1}{\sqrt{n\sum_{k\in\mathscr{K}} b_k}}\right)$, indicating that the excess risk scales down with square root of the number of data samples times

[1] We use algorithmic stability (see Definition A.6) to control the generalization error.

the total batch size, meaning that we need a large number of client samples to reduce the excess risk. Moreover, if we also pick $n = \sum_{k \in \mathcal{K}} b_k$ we can yield a bound that behaves as $O\left(\frac{1}{n}\right)$. We note, however, that $T \to \infty$ creates a communication bottleneck in federated learning systems, since there is a considerably large amount of messages that are exchanged between clients and server.

Next, we show in Section 5.4, how we can reduce the communication rounds by performing multiple local iterations.

5.4 Federated Smoothed RCVaR Algorithm with Multiple Local Epochs

Previously, we proposed an algorithm that leverages multi-pass minibatch SGD for solving Eq. 5.5. Next, we provide a communication-efficient version of FedSRCVaR algorithm that utilizes periodic averaging to reduce the communication rounds between server and clients. We also offer its convergence analysis, at the cost of some additional assumptions.

For convenience, we begin by discussing the changes in the initial algorithm that are required to support multiple local rounds, also presented in Algorithm 5. We call this new version Multi-Round FedSRCVaR. Akin the original version, all clients participate at each communication round. The training steps, also provided in Algorithm 5 are as follows:

1. Each communication round $t \geq 1$ starts with the server sending the current global model-threshold pair $(\boldsymbol{\theta}^t, c^t)$ to the clients.

2. Each client receives the global pair and performs τ local SGD updates on the model parameters and the threshold. At each local epoch $j \in [\tau]$ a new batch with size $b_k = b$ is used.

3. Clients return the updated pair $(\boldsymbol{\theta}_k^{(t,\tau)}, c_k^{(t,\tau)})$ to the server.

4. Finally, the server produces the new model-threshold pair $(\boldsymbol{\theta}^{t+1}, c^{t+1})$ by averaging the received client updates using the relative weight $\frac{1}{|\mathcal{K}|}$.

Algorithm 5 MULTI-ROUND FEDERATED SMOOTHED-RCVAR (MULTI-ROUND FEDSRCVAR) ALGORITHM

Inputs: $\mathscr{K}$: set of clients, T: communication rounds, τ: local epochs, η: learning rate for model $\boldsymbol{\theta}$ and quantile c, $\varepsilon \in (0,1]$: trade-off parameter, $\rho \in (0,1)$: parameter for probability-level, c^1: initial threshold set to $B=1$, b: local batch size.

Server initializes $\boldsymbol{\theta}^1$ randomly.
for $t = 1$ to T **do**
  Server **broadcasts** global model-threshold pair $(\boldsymbol{\theta}^t, c^t)$
  for each client $k \in \mathscr{K}$ **in parallel do**
    Set $(\boldsymbol{\theta}^{t,j=1}, c^{t,j=1}) = (\boldsymbol{\theta}^t, c^t)$
    for $j = 1$ to τ **do**
      Sample a batch with size b and compute

$$\boldsymbol{\theta}_k^{t,j+1} \leftarrow \boldsymbol{\theta}_k^{t,j} - \eta \nabla_{\boldsymbol{\theta}} \left\{ \frac{1}{b} \sum_{i=1}^{b} \tilde{f}(\boldsymbol{\theta}_k^{t,j}, c^t; x_i^k, y_i^k) \right\},$$

$$c_k^{t,j+1} \leftarrow c_k^{t,j} - \eta \nabla_c \left\{ \frac{1}{b} \sum_{i=1}^{b} \tilde{f}(\boldsymbol{\theta}_k^{t,j}, c_k^{t,j}; x_i^k, y_i^k) \right\}$$

    end for
    Return local model-threshold pair $(\boldsymbol{\theta}_k^{t,\tau}, c_k^{t,\tau})$ to server
  end for
  Server computes $\boldsymbol{\theta}^{t+1} \leftarrow \frac{1}{|\mathscr{K}|} \sum_{k \in \mathscr{K}} \boldsymbol{\theta}_k^{t,\tau}$ and $c^{t+1} \leftarrow proj_{c \in [0,B]} \left(\frac{1}{|\mathscr{K}|} \sum_{k \in \mathscr{K}} c_k^{t,\tau} \right)$
end for

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

5.4.1 Algorithmic Analysis

Next, we provide the convergence rate of Algorithm 5. In this multi-round setting, we perform multiple local iterates and we want to ensure that all of the generated pairs are approaching the minimizing pair $(\boldsymbol{\theta}_D^*, c_D^*)$. For a fixed $\boldsymbol{\theta} \in \Theta$, we denote

$$\tilde{F}(\boldsymbol{\theta}, c) = \sum_{k \in \mathscr{K}} \frac{1}{|\mathscr{K}|} \tilde{F}_k(\boldsymbol{\theta}, c) \quad \text{with} \quad \tilde{F}_k(\boldsymbol{\theta}, c) = \sum_{i=1}^{n_k} \frac{1}{n_k} \tilde{f}(\boldsymbol{\theta}, c; x_i^k, y_i^k).$$

We consider the assumptions presented in 5.1 and 5.2 and state some additional assumptions required for our guarantees below.

Assumption 5.5. *Every client $k \in \mathscr{K}$ queries an unbiased minibatch stochastic gradient $\nabla \left\{ \sum_{i=1}^{b} \frac{1}{b} \tilde{f}(\boldsymbol{\theta}, c; x_i^k, y_i^k) \right\}$ of $\tilde{F}_k$, i.e., $\mathbb{E}_{x,y} \left[\frac{1}{b} \nabla \left\{ \sum_{i=1}^{b} \tilde{f}(\boldsymbol{\theta}, c; x_i^k, y_i^k) \right\} \right] = \nabla \tilde{F}_k(\boldsymbol{\theta}, c)$,*

with $\frac{\sigma^2}{b}-$*uniformly bounded variance, i.e.,*

$$\mathbb{E}\left[\left|\left|\frac{1}{b}\nabla\{\sum_{i=1}^{b}\tilde{f}(\boldsymbol{\theta},c;x_i^k,y_i^k)\}-\nabla\tilde{F}_k(\boldsymbol{\theta},c)\right|\right|^2\right]\leq\frac{\sigma^2}{b}\quad\forall(\boldsymbol{\theta},c)\in\Theta\times[0,B],k\in\mathscr{K}.$$

Assumption 5.6. *Let* $\nabla\tilde{F}_k(\boldsymbol{\theta},c)$ *be the (local) gradient at client* $k\in\mathscr{K}$ *and* $\nabla\tilde{F}(\boldsymbol{\theta},c)$ *the global gradient such that* $\tilde{F}(\boldsymbol{\theta},c)=\frac{1}{|\mathscr{K}|}\sum_{k\in\mathscr{K}}\tilde{F}_k(\boldsymbol{\theta},c)$*. We assume that the difference between local and global gradients is* $\mu-$*uniformly bounded, meaning that*

$$\max_{k\in\mathscr{K}}\sup_{(\boldsymbol{\theta},c)\in\Theta\times[0,B]}\left|\left|\nabla\tilde{F}_k(\boldsymbol{\theta},c)-\nabla\tilde{F}(\boldsymbol{\theta},c)\right|\right|\leq\mu. \tag{5.10}$$

5.4.1.1 Convergence Guarantees

For the convergence analysis of Algorithm 5 we leverage the standard results for FedAVG presented in Theorem 1 in [Wang et al., 2021a]. We apply them to our setting and provide the proof of Lemma 5.6 below.

Lemma 5.6 (Convergence of Multi-Round FedSRCVaR)**.** *Let the assumptions 5.1, 5.2, 5.5 and 5.6 hold; and a constant learning rate*

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

Then, for the model-threshold pair $(\overline{\boldsymbol{\theta}}_T,\bar{c}_T)$ *provided by Algorithm 5 after* T *rounds we have*

$$\begin{aligned}&\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;x,y_i^k)-\tilde{f}(\boldsymbol{\theta}_D^*,c_D^*;x,y_i^k)}{n}\right]\\ \leq\;&\frac{2\left(\frac{(1-\varepsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\right)(M^2+B^2)}{\tau T}+\left(\frac{\left(\frac{(1-\varepsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\right)(M^2+B^2)^2}{T^2}\right)^{\frac{1}{3}}\left(5(\frac{\sigma^2}{b\tau})^{\frac{1}{3}}+19\mu^{\frac{2}{3}}\right)\\ &+\frac{2\sigma\sqrt{M^2+B^2}}{\sqrt{|\mathscr{K}|b\tau T}}+\frac{(1-\varepsilon)\gamma}{\rho}.\end{aligned}$$

The proof for Lemma 5.6 is provided in Section 5.6.3.

5.5 Empirical Results

5.5.1 Experimental Setup

5.5.1.1 Datasets and Model Architectures

In our experiments, we utilized the following standard datasets to evaluate the performance of the FedSRCVaR algorithm.

- **eICU Dataset [Pollard et al., 2018]:** The eICU dataset contains medical records from various medical centres and the target task is patient mortality. We distribute the data to 11 clients with each client representing a unique hospital in the dataset. To preprocess the data, we followed the instructions provided in [Pollard et al., 2018]. Note that accessing the eICU dataset requires proper credentials, and the procedure for requesting access is described on the dataset's website.[2]

- **ACS Employment Dataset [Ding et al., 2021]:** As discussed in Section 3.4.1.2, the ACS Employment dataset is used for employment classification and consists of 14 input features. We conducted experiments in two different settings: (a) a setting where data is allocated to 51 clients so that each client represents a different geo-location; (b) a setting where the data is split into 3 clients based on the race classes: {Black, White, Others}. We follow the same data preprocessing steps as in [Ding et al., 2021].

- **MNIST Dataset [Deng, 2012]:** The MNIST dataset consists of 28×28 pixel grayscale images of handwritten digits. The utility task is to classify the digits into 10 target classes. To create a federation, we assigned each digit to a different client, resulting in 10 clients, where each client represents a target class. This is a similar data splitting strategy as used in FashionMNIST splits in [Deng et al., 2020, Mohri et al., 2019], but also as described in Section 3.4.1.2.

- **Celeb-A Dataset [Liu et al., 2015]**: Celeb-A is a visual dataset containing

[2] eICU dataset website: https://eicu-crd.mit.edu/gettingstarted/access/.

facial images of celebrities, and the target task is to predict gender. We randomly assigned the data to two clients, using three different random seeds.

For the ACS Employment and MNIST datasets, we use a MLP with a single hidden layer with 512 neurons. For eICU we use Logistic Regression and for Celeb-A we use a ResNet-18. We select cross entropy to be the loss function in every method.

5.5.1.2 Benchmarks

We conducted experiments comparing the proposed FedSRCVaR approach against relevant baselines from (a) centralized machine learning DRO, BPF and ERM; and (b) federated learning FedAVG and AFL.

For each federated learning method, we assume that every client is available to participate in each communication round to ensure a fair comparison. We performed a grid search over hyperparameters to identify the best combination for each method. In particular, we train FedSRCVaR using local batch size $b_k = \{32, 64, 128\}$. We use the same options for AFL. BPF is trained using $\varepsilon = \{0.001, 0.005, 0.01, 0.05\}$. FedAVG is trained using batches of sample size 128 and local epochs $E = \{3, 8, 15\}$. We train all approaches using learning rates $\eta = \{0.01, 0.001, 0.0001, 0.00001\}$, adversary/threshold learning rate $\eta_{adv} = 0.001$ (where relevant). We pick the combination with the best solution for each case. For the proposed approach, FedSRCVaR, we report the results for group size $\rho = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and trade-off parameter $\varepsilon = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ unless stated otherwise.

In the reported results we present the mean performance over three runs and in separate splits. The splits are generated using 3-fold cross-validation. If a fixed test set is provided by the authors of the dataset we use that for testing. In the experiments, we used soft-ReLU as the smoothed plus function with a parameter value of $\gamma = 0.05$. This choice of function and parameter helps address the non-smoothness of the original plus function, as discussed earlier.

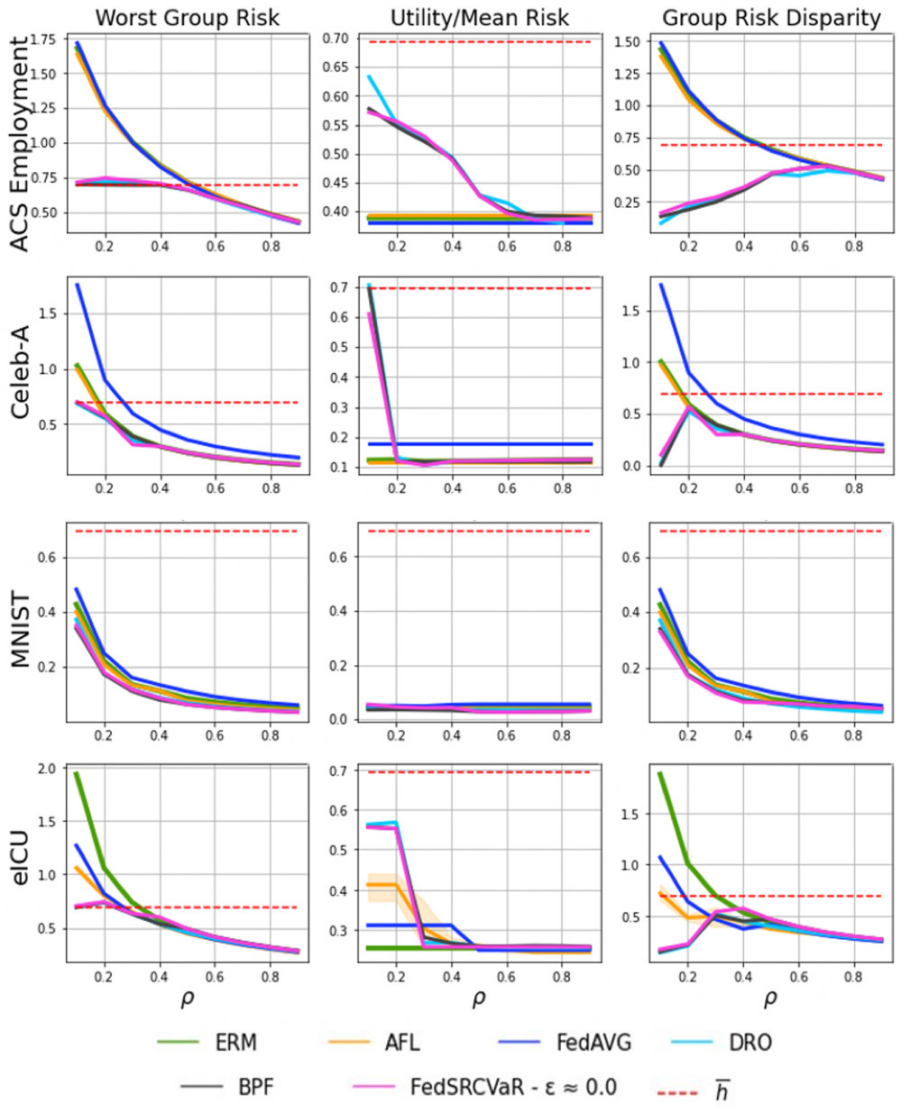

Figure 5.1: Comparison of worst group risk, utility risk and group risk disparity between the best and worst groups on different datasets. $\overline{h}$ denotes the uniform classifier. FedRCVaR recovers solutions equivalent to centralized machine learning, while improving both utility and fairness compared to FL baselines in many settings. For all datasets $\varepsilon \approx 0.0$ is set as $\varepsilon = 0.01$, except for ACS Employment that we pick the best solution from $\varepsilon = \{0.001, 0.005, 0.01, 0.05\}$.

5.5.2 Comparison to ML and FL Baselines

In Figure 5.1, we present a comprehensive comparison of our proposed approach, FedRCVaR, against several methods, assessing the effectiveness of the proposed approach in terms of group fairness and utility across different datasets and scenarios.

The results demonstrate that for small values of ε (i.e., $\varepsilon \approx 0$), FedRCVaR achieves the best performance in terms of worst group risk, along with BPF and DRO, confirming that our approach effectively produces subgroup robust solutions. Furthermore, we observe that DRO exhibits higher average risk compared to FedRCVaR and BPF at the same worst group performance in some scenarios. This suggests that DRO may underperform on the remaining population in these scenarios, emphasizing the importance of considering both worst group fairness and overall average performance.

Interestingly, in certain datasets, FedRCVaR improves both worst group fairness and utility performance simultaneously, outperforming AFL, which focuses on client fairness (or client robustness), and FedAVG, which primarily optimizes for utility and disregards fairness in federated learning settings. This indicates that for small values of ε, minimizing the right-tail risk of the samples is more effective in handling heterogeneity within the federation and achieving better performance on both fairness and utility measures. It also highlights the significance of addressing heterogeneity between the training and testing sets. We note that the performance difference observed between centralized ERM and FedAVG in certain scenarios can be attributed to the non-iidness of client data and the number of local iterations performed in the latter approach.

5.5.3 Global Group Fairness with Demographics vs without Demographics

Next, we examine the cost in (minimax) group fairness when being unaware of any demographics during the training phase, against the case of considering sensitive groups that are (potentially incorrectly) anticipated in the testing phase. We compare against the optimal minimax group fair solution for known sensitive groups, gener-

ated our approach FedMinMax in [Papadaki et al., 2022a], and present the results in Table 5.1. We leverage the ACS Employment dataset and allocate data on 3 clients based on the races {Black, White, Others}.

Table 5.1: Cross Entropy risks comparison of minimax Pareto federated group fairness with known demographics (FedMinMax), unknown demographics (FedSRCVaR, ours) and baseline (FedAVG) on ACS Employment dataset. {U, E} stand for {Unemployed, Employed} statuses, respectively. Results are averaged over 3 runs.

Group		**FedMinMax**	**FedSRCVaR** ($\varepsilon = 0.05$)		**FedAVG**
			$\rho = 0.1$	$\rho = 0.3$	(Baseline)
White	U	0.386±0.07	0.447±0.29	0.412±0.3	0.214±0.06
	E	0.382±0.03	0.697±0.17	0.635±0.12	0.601±0.05
Black	U	0.396±0.04	0.561±0.27	0.508±0.28	0.305±0.02
	E	0.374±0.04	0.696±0.0	0.635±0.0	0.648±0.04
Other	U	0.371±0.02	0.481±0.32	0.521±0.33	0.559±0.05
	E	0.37±0.05	0.696±0.0	0.634±0.1	0.209±0.03
Worst	$\rho = 0.1$	1.593±0.23	0.713±0.07	0.724±0.06	1.768±0.04
group	$\rho = 0.3$	1.176±0.08	0.698±0.02	0.695±0.05	1.037±0.09

We observe that FedMinMax has better performance on the known demographic groups compared to FedSRCVaR (as expected), but significantly poor performance on the worst groups of the selected ρ sizes. On the other hand, FedSRCVaR outperforms FedMinMax in the worst group generated by all samples, and performs relatively well on the predefined groups.

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

FedAVG performs better on many of the predefined groups compared to FedSRCVaR but underperforms FedSRCVaR in terms of worst group risk. Also, FedAVG's outcomes are significantly unbalanced across predefined demographic groups, and

its performance on the worst possible subgroup formulated for small group sizes $\rho = \{0.1, 0.3\}$ shows that it suffers substantially larger risk compared to our approach, as expected.

Table 5.2: Worst group formation of ACS Employment and actual group size on the test split. We evaluate FedSRCVaR for $\varepsilon = 0.05$ and $\rho = \{0.1, 0.3\}$. The labels {Unemployed, Employed} are denoted as {U, E}.

		White (%)		**Black** (%)		**Other** (%)	
ε	ρ	U	E	U	E	U	E
0.05	0.1	48.6	14.9	4.3	1.2	24.1	6.9
	0.3	30.2	30.7	2.6	2.	17.8	16.7
Actual size		33.4	27.9	2.9	1.9	18.3	15.6

For reference, we also provide the proportion of the various demographics in the predicted worst group that is generated by our approach for $\varepsilon \approx 0$ and low ρ values; and of the actual group populations in Table 5.2. We notice that the composition of the worst group is very close to the actual size of the sensitive groups, especially as ρ grows.

5.5.4 Achieving Various Trade-Offs using FedSRCVaR

Next, we empirically assess different trade-offs that can be accomplished using FedSRCVaR for various combinations of ε and ρ, Figure 5.2. In particular, we examine $\varepsilon = \{0.01, 0.1, \ldots, 0.9, 1.0\}$ and group sizes $\rho = \{0.1, \ldots, 0.9\}$.

The different colors indicate a particular ε value and we report results for models that were trained individually for each pair of (ε, ρ) values. For ACS Employment we use the same setting as in Section 5.5.3, where the data is split to 3 clients based on the race classes: {Black, White, Others}.

Figure 5.2 shows that ε effectively acts as a tuning parameter between worst group fairness and average performance. For small ρ values, ε has a significant impact on the worst group and the utility performance. We observe that the larger the ε the lower the average utility errors and risks, while as we decrease ε we boost the performance of the worst group. Note that for $\varepsilon \approx 0$ and $\rho \approx 0$, the worst-group risk is close to the uniform classifier risk which is consistent with Remark 4.1 in

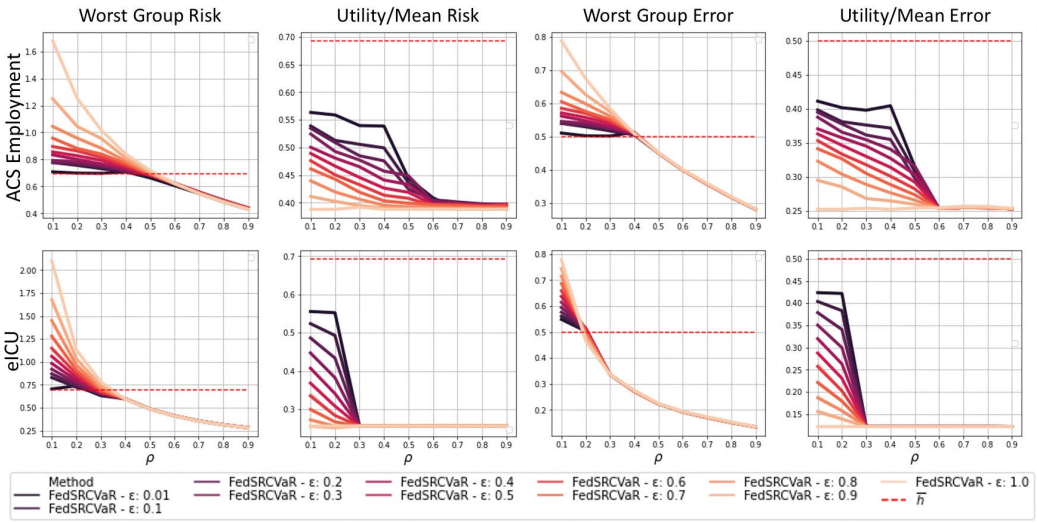

Figure 5.2: Performance trade-offs among the worst group and utility risks and errors. We examine different pairs of (ε, ρ) values on ACS Employment and eICU datasets. $\overline{h}$ denotes the uniform classifier. A lower score indicates better performance.

Section 4.3, and conclusions drawn about the existence of a critical worst-group size under which we yield the uniform classifier in [Martinez et al., 2021].

On the other hand, for large ρs we notice that all solutions are equivalent and the parameter ε has almost no influence on the solution. Interestingly, for particular values of ε and ρ, FedSRCVaR can recover client robustness solutions (akin to AFL), even though our objective does not explicitly aim for that.

5.5.5 FedSRCVaR vs. Multi-Round FedSRCVaR

Lastly, we compare the performance of three methods: FedSRCVaR, its communication-efficient variant called Multi-Round FedSRCVaR (described in Algorithm 5), and FedAVG, on the ACS Employment and eICU datasets. For the ACS Employment dataset, we adopt the setup of Section 5.5.2, where the data is distributed among 51 clients representing different US states. The models are trained for $\rho = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The results are presented in Figure 5.3.

For small ρ values, we notice that Multi-Round FedSRCVaR for $\tau \in \{5, 10\}$ exhibits a higher worst-group risk compared to the standard FedSRCVaR. This suggests that conducting multiple local rounds may result in inferior performance

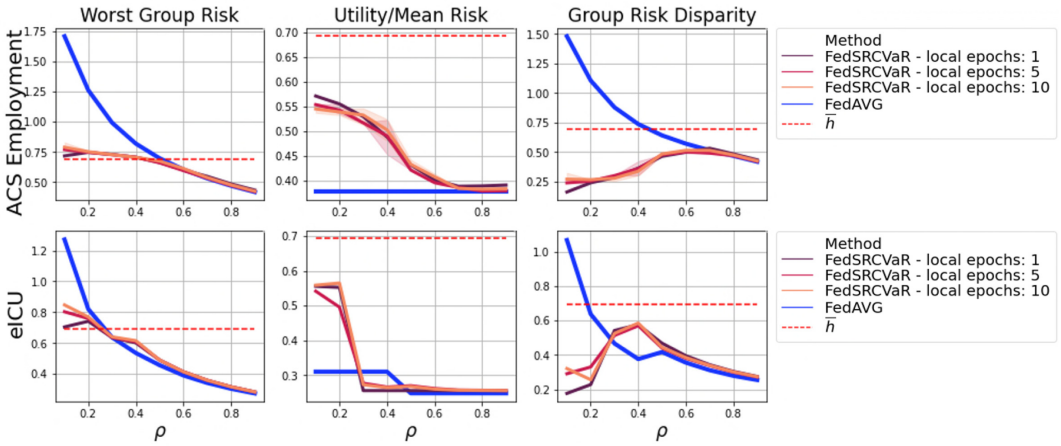

Figure 5.3: Performance comparison between FedAVG, (vanilla) FedSRCVaR and Multi-Round FedSRCVaR for local epochs $\tau \in \{5, 10\}$ on ACS Employment and eICU datasets. $\bar{h}$ denotes the uniform classifier. We report the worst group and average/utility cross entropy risks, and the group risk disparity between the worst performing group and the remaining population, as a function of ρ.

for the worst group when ρ is small, as indicated in Lemma 5.6. Moreover, as ρ increases, Multi-Round FedSRCVaR demonstrates improvement in worst-case fairness similar to the original FedSRCVaR. This implies that the impact of performing local epochs becomes less significant as the worst-group size becomes larger. Additionally, for sufficiently large values of ρ, both methods converge to the same solution as FedAVG.

5.6 Proofs

5.6.1 Smooth Approximation of Eq. 5.2

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

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

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

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

where $z = (x, y)$ and s is a convex and $(\frac{2}{\gamma})-$smooth function.

Given a function s that satisfies the conditions given in Definition 5.1, we provide the properties for the auxiliary function f and the smoothed function $\tilde{f}$ in Lemma 5.3.

Proof [Lemma 5.1]

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

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

Consequently, we also have $\max_j [\lambda x_j + (1-\lambda) y_j] \leq \lambda \max_i x_i + (1-\lambda) \max_i y_i$. Note that in our scenario $m = 2$ and $y_j \in \{0, \ell(\boldsymbol{\theta}, z_i) - c\}$ and $x_j \in \{0, \ell(\boldsymbol{\theta}, z_l) - c\}$ with $j \in [m]$ and $i, l \in [n]$. Note that the smoothed plus function s is convex with respect to z by definition. Thus, the convexity of the function $\tilde{f}$ is immediate since $\tilde{f}$ is a linear combination of convex terms.

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

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

define the set of subgradients as

$$\partial_{(\boldsymbol{\theta},c)} f(\boldsymbol{\theta},c;z) = \begin{cases} \begin{bmatrix} \left(\frac{(1-\varepsilon)}{\rho}+\varepsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};z) \\ (1-\varepsilon)(1-\frac{1}{\rho}) \end{bmatrix}, & \text{if } \ell(\boldsymbol{\theta};z) > c \\ \begin{bmatrix} \left(\frac{(1-\varepsilon)t}{\rho}+\varepsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};z) \\ (1-\varepsilon)(1-\frac{t}{\rho}) \end{bmatrix}, t \in [0,1], & \text{if } \ell(\boldsymbol{\theta};z) = c \\ \begin{bmatrix} \varepsilon\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};z) \\ 1-\varepsilon \end{bmatrix}, & \text{if } \ell(\boldsymbol{\theta};z) < c \end{cases} \tag{5.11}$$

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

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

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

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

$$\nabla\tilde{f}(\boldsymbol{\theta},c;z) = \begin{bmatrix} \left(\frac{(1-\varepsilon)s'(\ell(\boldsymbol{\theta};z)-c)}{\rho}+\varepsilon\right)\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};z) \\ (1-\varepsilon)\left(1-\frac{s'(\ell(\boldsymbol{\theta};z)-c)}{\rho}\right) \end{bmatrix} \tag{5.12}$$

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

$$||\nabla \tilde{f}||^2 \leq \max_{t\in[0,1]} \left[G^2 \left(\frac{(1-\varepsilon)t}{\rho} + \varepsilon \right)^2 + (1-\varepsilon)^2 (1-\frac{t}{\rho})^2 \right] \leq G^2_{\rho,\varepsilon}.$$

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

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

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

$$\begin{aligned} \leq & \left(\tfrac{(1-\varepsilon)\left|s'(\ell(\boldsymbol{\theta}_1;z)-c_1)\right|}{\rho} + \varepsilon \right) \cdot ||\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_1;z) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_2;z)|| \\ & + ||\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_2;z)|| \cdot \left| \left(\tfrac{(1-\varepsilon)s'(\ell(\boldsymbol{\theta}_1;z)-c_1)}{\rho} + \varepsilon \right) - \left(\tfrac{(1-\varepsilon)s'(\ell(\boldsymbol{\theta}_2;z)-c_2)}{\rho} + \varepsilon \right) \right| \end{aligned} \tag{5.13}$$

We know that:

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

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

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

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

3. The loss function ℓ is convex and $G-$Lipschitz, thus $||\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_2;z)|| \leq G$.

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

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

By substituting 1-4 into Eq. 5.13 we get

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

$$\begin{aligned}
\leq\ & \left(\tfrac{(1-\varepsilon)}{\rho}+\varepsilon\right)\beta||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||+G\tfrac{(1-\varepsilon)}{\rho}\left|s'(\ell(\boldsymbol{\theta}_1;z)-c_1)-s'(\ell(\boldsymbol{\theta}_2;z)-c_2)\right| \\
\leq\ & \left(\tfrac{(1-\varepsilon)}{\rho}+\varepsilon\right)\beta||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||+G\tfrac{(1-\varepsilon)}{\rho}\tfrac{2}{\gamma}\left|(\ell(\boldsymbol{\theta}_1;z)-c_1)-(\ell(\boldsymbol{\theta}_2;z)-c_2)\right| \\
\leq\ & \left(\tfrac{(1-\varepsilon)}{\rho}+\varepsilon\right)\beta||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||+G^2\tfrac{(1-\varepsilon)}{\rho}\tfrac{2}{\gamma}||(\boldsymbol{\theta}_1,c_1)-(\boldsymbol{\theta}_2,c_2)|| \\
=\ & \left(\tfrac{(1-\varepsilon)}{\rho}(\beta+\tfrac{2}{\gamma}G^2)+\varepsilon\beta\right)||m_1-m_2||, \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by } G\text{−Lipschitzness of } \ell)
\end{aligned}$$

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

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

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

□

5.6.2 Analysis of Algorithm 4

In this section, we analyze the properties of Algorithm 4, stated in Lemmas 5.2 and 5.5. To achieve this, we also introduce the following two definitions.

Definition 5.2. *We let $z=(x,y)$ and $(\boldsymbol{\theta}^t, c^t)$ be the model-threshold pair at round t of Algorithm 4. We also let $(\overline{\boldsymbol{\theta}}_T, \overline{c}_T)$ be the average model-threshold pair after T rounds of Algorithm 4, with $\overline{\boldsymbol{\theta}}_T = \frac{1}{T}\sum_{t\in[T]} \boldsymbol{\theta}^t$ and $\overline{c}_T = \frac{1}{T}\sum_{t\in[T]} c^t$.*

Definition 5.3. *We let $(\boldsymbol{\theta}^\star_D, c^\star_D)$ be the model-threshold pair that optimizes the empirical objective $\min_{\boldsymbol{\theta}\in\Theta}\min_{c\in[0,1]} \frac{1}{n}\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k} \tilde{f}(\boldsymbol{\theta}, c; z_i^k)$.*

5.6.2.1 Convergence of Algorithm 4

Convex Setting: For the convergence analysis of Algorithm 4 we use the standard results for SGD in [Nemirovski and Yudin, 1983] that we repeat in Lemma 5.7 for convenience.

Lemma 5.7 (Chapter 5 in [Nemirovski and Yudin, 1983])**.** *Let assumptions 5.1, and 5.2 hold. Let also $\overline{\boldsymbol{\theta}}_T = \sum_{t=1}^{T} \boldsymbol{\theta}_t$ and $\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_Z[\ell(\boldsymbol{\theta}; Z)]$. Suppose we run T steps of SGD with a fixed learning rate η. Then,*

$$\mathop{\mathbb{E}}_{Z\sim p(Z)}[\ell(\overline{\boldsymbol{\theta}}_T; Z)] - \mathop{\mathbb{E}}_{Z\sim p(Z)}[\ell(\boldsymbol{\theta}^\star; Z)] \leq \frac{1}{2}\left(\frac{M^2}{\eta T} + G^2\eta\right).$$

Next, we apply the results of SGD described in Lemma 5.7 to our setting and provide the proof of Lemma 5.2 below.

Proof [Lemma 5.2]

We apply the results from Lemma 5.7 in our setting by substituting the properties of the loss function ℓ with those of the non-smooth and smoothed auxiliary functions, f and $\tilde{f}$ respectively, given by Lemma 5.1. In particular,

1. M is changed to $\sqrt{M^2+B^2}$ since we optimize over $v\in\Theta\times[0,B]$.
2. G^2 is substituted with $\frac{G^2_{\rho,\varepsilon}}{\left(\sum_{k\in\mathscr{K}} b_k\right)}$ due to the Lipschitz conditions in Lemma 5.1 and the use of minibatches which reduce the Lipschitz parameter by a factor equal to the batch size $\left(\sum_{k\in\mathscr{K}} b_k\right)$ since $\left|\left|\nabla \frac{1}{\sum_{k\in\mathscr{K}} b_k}\sum_{i=1}^{b_k} \tilde{f}(\cdot, z_i^k)\right|\right|^2 \leq \frac{1}{\sum_{k\in\mathscr{K}} b_k} G^2_{\rho,\varepsilon}$.

Thus, since Algorithm 4 optimizes for the auxiliary function $\tilde{f}$, we get

$$\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;z_i^k)-\tilde{f}(\boldsymbol{\theta}_D^*,c_D^*;z_i^k)}{n}\right]\leq\frac{1}{2}\left(\frac{M^2+B^2}{\eta T}+\frac{G_{\rho,\varepsilon}^2}{\sum_{k\in\mathscr{K}}b_k}\eta\right)$$

Considering also Lemma 5.1 property (4), we finally obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{f(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;z_i^k)}{n}\right]&\leq\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;z_i^k)}{n}\right]\\&\leq\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\boldsymbol{\theta}_D^*,c_D^*;z_i^k)}{n}\right]+\frac{1}{2}\left(\frac{M^2+B^2}{\eta T}+\frac{G_{\rho,\varepsilon}^2}{\sum_{k\in\mathscr{K}}b_k}\eta\right)\\&\leq\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{f(\boldsymbol{\theta}_D^*,c_D^*;z_i^k)}{n}\right]+\frac{1}{2}\left(\frac{M^2+B^2}{\eta T}+\frac{G_{\rho,\varepsilon}^2}{\sum_{k\in\mathscr{K}}b_k}\eta\right)\\&\quad+\frac{(1-\varepsilon)\gamma}{\rho}.\end{aligned}$$

□

For the convergence analysis of Algorithm 4, for a model-threshold pair $(\overline{\boldsymbol{\theta}}_{V:T},\bar{c}_{V:T})$ that is averaged from round V to round T, with $1<V\leq T$, we use the standard results for SGD in [Nemirovski et al., 2009] that we repeat in Lemma 5.8 for completeness.

Lemma 5.8 (Section 2.2 in [Nemirovski et al., 2009])**.** *Let assumptions 5.1, and 5.2 hold. Let also* $\overline{\boldsymbol{\theta}}_{V:T}=\sum_{t=V}^{T}\boldsymbol{\theta}_t$*, with* $V\leq T$ *and* $\boldsymbol{\theta}^\star=\arg\min_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_Z[\ell(\boldsymbol{\theta};Z)]$*. Suppose we run* T *steps of SGD with a fixed learning rate* η*. Then,*

$$\mathbb{E}_{Z\sim p(Z)}[\ell(\overline{\boldsymbol{\theta}}_{V:T};Z)]-\mathbb{E}_{Z\sim p(Z)}[\ell(\boldsymbol{\theta}^\star;Z)]\leq\frac{4M^2+G^2\eta^2(T-V+1)}{2\eta(T-V+1)}.$$

Building upon the findings of SGD outlined in Lemma 5.8, we apply these results to our specific scenario and present the proof for Lemma 5.3.

Proof [Lemma 5.3]

We apply the same substitutions akin to Lemma 5.2. Therefore, we get

$$\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\overline{\boldsymbol{\theta}}_{V:T},\bar{c}_{V:T};z_i^k)-\tilde{f}(\boldsymbol{\theta}_D^\star,c_D^\star;z_i^k)}{n}\right] \leq \frac{4(M^2+B^2)\sum_{k\in\mathscr{K}}b_k+G_{\rho,\varepsilon}^2\eta^2(T-V+1)}{2\eta(T-V+1)\sum_{k\in\mathscr{K}}b_k}.$$

By selecting as a step size $\eta=\frac{1}{G_{\rho,\varepsilon}\sqrt{T}}\sqrt{\sum_{k\in\mathscr{K}}b_k(M^2+B^2)}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\frac{\tilde{f}(\overline{\boldsymbol{\theta}}_{V:T},\bar{c}_{V:T};z_i^k)-\tilde{f}(\boldsymbol{\theta}_D^\star,c_D^\star;z_i^k)}{n}\right] &\leq \frac{4(M^2+B^2)\sum_{k\in\mathscr{K}}b_k+G_{\rho,\varepsilon}^2\eta^2(T-V+1)}{2\eta(T-V+1)\sum_{k\in\mathscr{K}}b_k}\\ &= \frac{2TG_{\rho,\varepsilon}\sqrt{M^2+B^2}}{(T-V+1)\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}}b_k}}+\frac{G_{\rho,\varepsilon}\sqrt{M^2+B^2}}{2\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}}b_k}}\\ &= \frac{G_{\rho,\varepsilon}\sqrt{M^2+B^2}}{\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}}b_k}}\left(\frac{1}{2}+\frac{2T}{T-V+1}\right)\end{aligned}$$

Finally, we yield the proposed bound leveraging the results in Lemma 5.1 as follows

$$\begin{aligned}\mathbb{E}\big[\mathop{\mathbb{E}}_{Z}[f(\overline{\boldsymbol{\theta}}_{V:T},\bar{c}_{V:T};Z)]\big] &\leq \mathbb{E}\big[\mathop{\mathbb{E}}_{Z}[\tilde{f}(\overline{\boldsymbol{\theta}}_{V:T},\bar{c}_{V:T};Z)]\big]\\ &\leq \mathbb{E}\big[\tfrac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{\theta}_D^\star,c_D^\star;z_i)\big]+\frac{G_{\rho,\varepsilon}\sqrt{M^2+B^2}}{\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}}b_k}}\left(\frac{1}{2}+\frac{2T}{T-V+1}\right)\\ &\leq \mathbb{E}\big[\tfrac{1}{n}\sum_{i=1}^{n}f(\boldsymbol{\theta}_D^\star,c_D^\star;z_i)\big]+\frac{G_{\rho,\varepsilon}\sqrt{M^2+B^2}}{\sqrt{T}\sqrt{\sum_{k\in\mathscr{K}}b_k}}\left(\frac{1}{2}+\frac{2T}{T-V+1}\right)\\ &\quad+\frac{(1-\varepsilon)\gamma}{\rho}.\end{aligned}$$

□

Non-Convex Setting: To establish the probabilistic convergence of FedSRCVaR, we depend on a lemma derived from the findings in [Orabona, 2020], which encapsulates the outcomes of the seminal work by Bertsekas and Tsitsiklis [Bertsekas and Tsitsiklis, 1999].

Lemma 5.9 (Lemma 1 [Orabona, 2020]). *Let* $\{\lambda_t\}_{t\geq 1}$ *and* $\{\eta_t\}_{t\geq 1}$ *be two non-negative sequences and* $\{\alpha_t\}_{t\geq 1}$ *be a sequence of vectors. Assume* $\sum_{t=1}^{\infty} \eta_t \lambda_t^p < \infty$ *and* $\sum_{t=1}^{\infty} \eta_t = \infty$, *with* $p \geq 1$. *Assume also there exists* $L \geq 0$ *such that*

$$|\lambda_{t+\tau} - \lambda_t| \leq L\left(\sum_{i=t}^{t+\tau-1} \eta_i \lambda_i + \Big|\Big| \sum_{i=t}^{t+\tau-1} \eta_i \alpha_i \Big|\Big|\right),$$

where α_t *is such that* $|| \sum_{t=1}^{\infty} \eta_t \alpha_t|| < \infty$. *Then* λ_t *converges to* 0.

Proof [Lemma 5.4] Following the proof in Theorem 2 [Orabona, 2020], we let $\lambda_t = \left|\left|\nabla \hat{F}(\boldsymbol{\theta^t}, c^t)\right|\right|$. From the smoothness assumption of the loss function, but also Lemma 5.1.3, we get

$$\left|\left|\left|\nabla \hat{F}(\boldsymbol{\theta}^{t+\tau}, c^{t+\tau})\right|\right| - \left|\left|\nabla \hat{F}(\boldsymbol{\theta^t}, c^t)\right|\right|\right|$$

$$\leq \left|\left|\nabla \hat{F}(\boldsymbol{\theta}^{t+\tau}, c^{t+\tau}) - \nabla \hat{F}(\boldsymbol{\theta^t}, c^t)\right|\right| \qquad \text{(by reverse triangle inequality)}$$

$$\leq \left(\frac{(1-\varepsilon)}{\rho}(\beta + \frac{2}{\gamma}G^2) + \varepsilon\beta\right) ||(\boldsymbol{\theta}^{t+\tau}, c^{t+\tau}) - (\boldsymbol{\theta}^t, c^t)||$$

(by smoothness property in Lemma 5.1.3)

$$= \left(\frac{(1-\varepsilon)}{\rho}(\beta + \frac{2}{\gamma}G^2) + \varepsilon\beta\right) \left|\left| \sum_{i=t}^{t+\tau-1} \eta_i \nabla \left\{ \sum_{k\in\mathscr{K}} \sum_{i=1}^{b_k} \frac{\tilde{f}(\boldsymbol{\theta}^i, c^i; x_i^k, y_i^k)}{\sum_{k\in\mathscr{K}} b_k} \right\} \right|\right|$$

$$\text{(by using } (\boldsymbol{\theta}^{t+1}, c^{t+1}) = (\boldsymbol{\theta}^t, c^t) - \eta_t \nabla \left\{ \sum_{k\in\mathscr{K}} \sum_{i=1}^{b_k} \frac{\tilde{f}(\boldsymbol{\theta}^t, c^t; x_i^k, y_i^k)}{\sum_{k\in\mathscr{K}} b_k} \right\})$$

$$= \left(\frac{(1-\varepsilon)}{\rho}(\beta + \frac{2}{\gamma}G^2) + \varepsilon\beta\right) \left|\left| \sum_{i=t}^{t+\tau-1} \eta_i \nabla \left\{ \sum_{k\in\mathscr{K}} \sum_{i=1}^{b_k} \frac{\tilde{f}(\boldsymbol{\theta}^i, c^i; x_i^k, y_i^k)}{\sum_{k\in\mathscr{K}} b_k} \right\} \right.\right.$$

$$\left.\left. + \eta_i \nabla \hat{F}(\boldsymbol{\theta}^i, c^i) - \eta_i \nabla \hat{F}(\boldsymbol{\theta}^i, c^i) \right|\right| \qquad \text{(by adding and subtracting } \eta_i \nabla \hat{F}(\boldsymbol{\theta}^i, c^i))$$

$$\leq \left(\frac{(1-\varepsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\right)\sum_{i=t}^{t+\tau-1}\eta_i\left|\left|\nabla\hat{F}(\boldsymbol{\theta}^i,c^i)\right|\right| + \left(\frac{(1-\varepsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\right)\left|\left|\sum_{i=t}^{t+\tau-1}\eta_i\nabla\left\{\sum_{k\in\mathscr{K}}\sum_{i=1}^{b_k}\frac{\tilde{f}(\boldsymbol{\theta}^i,c^i;x_i^k,y_i^k)}{\sum_{k\in\mathscr{K}}b_k}\right\}-\eta_i\nabla\hat{F}(\boldsymbol{\theta}^i,c^i)\right|\right|.$$

(by triangle inequality)

By letting $\alpha_i = \nabla\left\{\sum_{k\in\mathscr{K}}\sum_{i=1}^{b_k}\frac{\tilde{f}(\boldsymbol{\theta}^i,c^i;x_i^k,y_i^k)}{\sum_{k\in\mathscr{K}}b_k}\right\}-\nabla\hat{F}(\boldsymbol{\theta}^i,c^i)$, we finally get

$$\left|\left|\left|\nabla\hat{F}(\boldsymbol{\theta}^{t+\tau},c^{t+\tau})\right|\right|-\left|\left|\nabla\hat{F}(\boldsymbol{\theta}^{t},c^{t})\right|\right|\right| \leq \left(\frac{(1-\varepsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\right)\left(\sum_{i=t}^{t+\tau-1}\eta_i\left|\left|\nabla\hat{F}(\boldsymbol{\theta}^i,c^i)\right|\right|+\left|\left|\sum_{i=t}^{t+\tau-1}\eta_i\alpha_i\right|\right|\right).$$

Based on Lemma 5.9 with the choice of $p=2$, we can conclude that $\sum_{t=1}^{\infty}\eta_t||\nabla\hat{F}(\boldsymbol{\theta}^t,c^t)||^2<\infty$ with a probability of 1. Additionally, considering that $\sum_{t=1}^{T}\eta_t\alpha_t$ for $T=1,2,\ldots$ forms a martingale with variance bounded by $\left(\frac{\sigma^2}{\sum_{k\in\mathscr{K}}b_k}\right)\sum_{t=1}^{\infty}\eta_t<\infty$, we can conclude that $||\sum_{t=1}^{\infty}\eta_t\alpha_t||<\infty$ with a probability of 1. As a result, $\sum_{t=1}^{T}\eta_t\alpha_t$ for $T=1,2,\ldots$ is a martingale in $\left(\frac{(1-\varepsilon)}{\rho}(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\right)^2$, indicating that it converges in the same space with a probability 1. This convergence implies that the gradients of FedSRCVaR converge to 0 with a probability of 1.

□

5.6.2.2 Excess Risk Analysis of Algorithm 4

In order to derive an upper bound in excess risk $\mathscr{E}_r$, we use the results for stochastic gradient methods from Proposition 5.4. in [Hardt et al., 2016b], which we also repeat using our notation in Lemma 5.10 for convenience.

Lemma 5.10 (Proposition 5.4. in [Hardt et al., 2016b]). *Let assumptions 5.1 and 5.2 hold. Let also* $\overline{\boldsymbol{\theta}}_T=\sum_{t=1}^{T}\boldsymbol{\theta}_t$ *and* $\boldsymbol{\theta}_D^\star=\arg\min_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta};z_i)$. *Suppose we run T*

steps of SGD with a learning rate $\eta = \frac{M\sqrt{n}}{G\sqrt{T(n+2T)}}$. *Then,*

$$\mathbb{E}\left[\mathop{\mathbb{E}}_{Z\sim p(Z)}[\ell(\overline{\boldsymbol{\theta}}_T;Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}_D^\star;z_i)\right] \leq \frac{1}{2}\left(\frac{M^2}{\eta T} + \frac{G^2\eta}{n}(2T+n)\right) \tag{5.14}$$

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

Note that the selected step-size in Lemma 5.10 satisfies $\eta \leq \frac{2}{\beta}$ (see Theorem 3.7 in [Hardt et al., 2016b] for more details). This condition is required to be satisfied in our analysis as well. Next, we apply these results in our setting and provide the formal proof of Lemma 5.5.

Proof [Lemma 5.5]

For the excess risk analysis, we use the same substitutions as in 5.2. Additionally, the loss function smoothness parameter β is changed to the smoothness parameter of $\tilde{f}$, which is $\frac{(1-\varepsilon)}{\rho}(\beta + \frac{2}{\gamma}G^2) + \varepsilon\beta$. Thus, Eq. 5.14 for our setting becomes

$$\begin{aligned}
&\mathbb{E}\left[\mathop{\mathbb{E}}_{K\in\mathscr{K}}\left[\mathop{\mathbb{E}}_{Z|k}[\tilde{f}(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;Z|k)]\right] - \tfrac{1}{n}\sum_{k\in\mathscr{K}}\sum_{i=1}^{n_k}\tilde{f}(\boldsymbol{\theta}_D^*,c_D^*;z_i^k)\right] \\
&= \mathbb{E}\left[\mathop{\mathbb{E}}_{Z}[\tilde{f}(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;Z)]\right] - \mathbb{E}\left[\tfrac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{\theta}_D^*,c_D^*;z_i)\right] \qquad (5.15)\\
&\leq \tfrac{1}{2}\left(\tfrac{M^2+B^2}{\eta T} + \tfrac{G_{\rho,\varepsilon}^2\eta}{\left(\sum_{k\in\mathscr{K}} b_k\right)n}(2T+n)\right),
\end{aligned}$$

with the learning rate being

$$\eta = \frac{\sqrt{M^2+B^2}}{G_{\rho,\varepsilon}\sqrt{T(n+2T)}}\sqrt{n\sum_{k\in\mathscr{K}} b_k}. \tag{5.16}$$

Based on the remark we made above about the learning rate in Lemma 5.10 being at most $\frac{2}{\beta}$, we must ensure that the respective step size in Eq. 5.16 is at most $\frac{2\rho}{(1-\varepsilon)(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\rho}$ as well.

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

$$\frac{2\rho}{(1-\varepsilon)(\beta+\frac{2}{\gamma}G^2)+\varepsilon\beta\rho} \geq \frac{2\rho}{\beta+\frac{2}{\gamma}G^2+\varepsilon\beta\rho} \geq \frac{\rho}{\max\{\beta(1+\varepsilon\rho),\frac{2}{\gamma}G^2\}}$$

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

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

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

$$\eta = \sqrt{M^2+B^2}\frac{\sqrt{n\left(\sum_{k\in\mathscr{K}} b_k\right)}}{G_{\rho,\varepsilon}\sqrt{T(n+2T)}} \leq \frac{\rho}{\beta(1+\varepsilon\rho)}$$

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

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

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

$$\geq \max\left\{\varepsilon G,\frac{(1-\varepsilon+\varepsilon\rho)G}{\rho}\right\} \geq \frac{(1-\varepsilon+\varepsilon\rho)G}{\rho}$$

we have that

$$\frac{\rho\gamma}{2G^2} \geq \frac{\rho^2\gamma}{2G^2} \geq \frac{(1-\varepsilon+\varepsilon\rho)^2\gamma}{2G_{\rho,\varepsilon}^2}. \tag{5.17}$$

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

Finally, we yield the proposed bound by using the learning rate in Eq. 5.16 and

the results in Lemma 5.1 for which Eq. 5.15 becomes

$$\begin{aligned}\mathbb{E}\big[\mathop{\mathbb{E}}_{Z}[f(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;Z)]\big] &\leq \mathbb{E}\big[\mathop{\mathbb{E}}_{Z}[\tilde{f}(\overline{\boldsymbol{\theta}}_T,\bar{c}_T;Z)]\big] \\ &\leq \mathbb{E}\big[\tfrac{1}{n}\sum_{i=1}^{n}\tilde{f}(\boldsymbol{\theta}^*_D,c^*_D;z_i)\big] + G_{\rho,\varepsilon}\sqrt{\frac{(M^2+B^2)(\frac{2}{n}+\frac{1}{T})}{\sum_{k\in\mathscr{K}} b_k}} \\ &\leq \mathbb{E}\big[\tfrac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}^*_D,c^*_D;z_i)\big] + G_{\rho,\varepsilon}\sqrt{\frac{(M^2+B^2)(\frac{2}{n}+\frac{1}{T})}{\sum_{k\in\mathscr{K}} b_k}} + \frac{(1-\varepsilon)\gamma}{\rho}\end{aligned} \tag{5.18}$$

□

5.6.3 Analysis of Algorithm 5

Proof [Lemma 5.6]

We apply the results from Theorem 1 in [Wang et al., 2021a] to our setting by substituting the properties of the loss function ℓ with those of the non-smooth and smoothed auxiliary functions, f and $\tilde{f}$ respectively, given by Lemma 5.1.

In particular,

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

Considering also Lemma 5.1, property 4, we finally obtain

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

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

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

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

Finally, the learning rate becomes

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

□

5.7 Summary

In this chapter, we tackle the challenge of achieving federated group fairness in the absence of demographic awareness among the participating parties. To address this issue, we extend RCVaR – introduced in Chapter 4 – and propose an algorithm specifically designed for federated learning settings. This algorithm solves a smoothed approximation of the fairness problem and is accompanied by theoretical guarantees related to convergence and excess risk. Through extensive empirical evaluations, we demonstrate the effectiveness of our approach in improving the performance of the worst-performing group while maintaining a satisfactory average performance. Our method offers a diverse set of solutions with varying fairness-utility trade-offs, empowering practitioners to select the solution that best suits their specific requirements. Furthermore, we emphasize that our approach continues to provide benefits even when demographic information is available, showcasing its practicality and versatility.

Chapter 6

Conclusions and Future Work

Federated learning has emerged as an indispensable approach for enhancing model utility and acquiring large, inclusive datasets that encompass diverse demographics. Nevertheless, ensuring fairness across demographics has become paramount in responsible machine learning practices. We acknowledge the significance of fairness considerations in machine learning, particularly when dealing with large-scale and inclusive datasets representing a wide range of population groups. This research specifically focuses on addressing the unique challenges posed by federated learning in achieving fairness.

6.1 Summary of Contributions

In this work, we introduce and address the problem of achieving global group fairness in federated learning settings, considering scenarios where the sensitive groups can be either known or unknown. We recognize that in federated learning, different participating entities (clients) may only have access to a subset of the population groups during the training phase, while the testing phase might involve the entire population. Our goal is to develop techniques and algorithms that can ensure fairness across all population groups, even in such data-constrained and distributed settings.

Specifically, Chapter 3 delves into the formal definition of global group fairness in federated learning and highlights the distinctiveness of our fairness definition compared to existing works on fair federated learning. We establish conditions under which conventional client-level fairness aligns with group-level fairness, shedding

light on the intricate relationship between client fairness and group fairness in federated settings. Furthermore, we propose an optimization algorithm called FedMinMax that specifically addresses the minimax group fairness problem in federated setups. Our algorithm demonstrates competitive minimax guarantees comparable to those achieved by centralized machine learning algorithms in achieving group fairness. Through extensive empirical evaluations, we substantiate the superiority of our method in terms of group fairness across various learning settings. Additionally, we validate the conditions under which competing approaches yield the same solutions as our proposed objective.

In Chapter 4, we introduce a novel optimization objective, called Relaxed Conditional Value-at-Risk (RCVaR), in the context of centralized machine learning. Our objective enables the enforcement of Pareto (sub)group robustness by optimizing the right tail risk for a sufficiently large sample size. RCVaR sets bounds on the worst-case group disparity that can be expected in the testing phase and offers flexibility in achieving effective trade-offs between fairness and average utility. We establish the connections between RCVaR and existing literature on fairness without demographics in machine learning, providing a comprehensive discussion on its merits and limitations.

Lastly, in Chapter 5, we extend the formulation of RCVaR to accommodate federated learning scenarios where access to demographic information is not available. We propose a simple yet effective algorithm that solves a smoothed approximation of the proposed fairness objective. We provide theoretical guarantees regarding the convergence and generalization properties of our algorithm. We experimentally demonstrate that our approach achieves comparable solutions to centralized machine learning counterparts, while surpassing relevant federated learning approaches in terms of group fairness.

Overall, these chapters contribute to advancing the field of fairness in machine learning, particularly in the context of federated learning, by introducing novel objectives, optimization algorithms, and theoretical foundations. The empirical evaluations validate the effectiveness and practicality of our proposed approaches,

emphasizing the importance of achieving group fairness in diverse and distributed data settings. We note that the methods proposed in this thesis have broad applicability and can be deployed in various federated learning applications. These applications span different domains, including the medical field, insurance, finance, college admissions, and more.

6.2 Limitations, Open Problems and Future Work

The field of fair federated learning presents numerous open problems and promising directions for future research, which have significant relevance to our work and can contribute to the advancement of federated fair learning as a whole. These areas of exploration hold great potential for further understanding and improving the fairness outcomes in federated learning settings.

6.2.1 Trade-off between Global Group Fairness and Privacy Preservation

One of the key advantages of federated learning is that data remains on the client and only local updates are shared with the server which significantly enhances privacy. Nevertheless, it is still possible to deduce information about the local raw data. For instance, having knowledge of the previous model and the current gradient update from a client allows the inference of the client's training sample(s) [Kairouz et al., 2019]. Therefore, privacy concerns in federated learning are a crucial aspect that requires investigation and we leave it for future work. As the field progresses, it becomes essential to develop methods and techniques that not only aim for global group fairness in federated learning but also ensure the privacy of the generated models and sensitive data.

In particular, one potential direction for future work involves studying how to achieve minimax fairness while simultaneously safeguarding the privacy of the participating entities' local models and data. This entails examining the trade-offs and synergies between privacy preservation and fairness objectives and identifying strategies to reconcile these often competing goals.

To achieve this, we might need to adjust the frameworks and optimization

algorithms in Chapters 3 and 5 to incorporate privacy-aware mechanisms, ensuring that fairness objectives are pursued while preserving the privacy of sensitive information. It should be investigated how different privacy-preserving mechanisms and constraints affect the achievement of fairness objectives in federated learning. This requires a careful analysis of the interplay between privacy measures and fairness criteria, as well as the potential trade-offs or compromises that may arise.

6.2.2 Federated Fairness with Partial Demographic Group Knowledge

Another potential avenue for future research involves conducting a more in-depth exploration of the insights and outcomes outlined in Section 5.5.3. Our proposed approach, RCVaR, as discussed in Section 4.1, establishes upper bounds on the maximum potential group disparity that may arise during the testing phase. However, it is important to acknowledge that these bounds might be overly pessimistic in scenarios where the demographics are known in advance. Therefore, a valuable direction for future work entails investigating methods to incorporate partial knowledge of existing demographics into the federated fair learning procedure.

This research direction would involve investigating strategies for handling situations where only partial knowledge of demographics is available. This could entail developing algorithms that can handle missing or incomplete demographic information and devise approaches to make accurate predictions or estimates based on the available data. Such techniques would enable federated learning systems to adapt and perform well even in scenarios where comprehensive demographic information is not accessible.

6.2.3 Dynamic Adaptation to Fairness Definitions

Given the dynamic and evolving nature of federated learning, there exists a promising area for exploration focused on developing methods that can dynamically adapt group definitions within the federated learning framework. This direction of research aims to address the challenge of aligning the group definitions with the ever-changing social contexts, client preferences, or other contextual factors that influence fairness

considerations in federated learning [Kairouz et al., 2019].

For example, this could be addressed by incorporating feedback from marginalized or underrepresented groups into the process of refining and updating group definitions. By actively involving these groups, their perspectives and experiences can contribute to the development of more inclusive and fair group definitions. This iterative feedback loop would enable continuous improvement and fine-tuning of fairness objectives, ensuring that the evolving needs and aspirations of marginalized communities are effectively addressed.

To implement such an approach, various mechanisms for soliciting feedback from marginalized groups should be explored, such as conducting surveys, interviews, or participatory design processes. By actively engaging with these groups, we can gain insights into the specific challenges and concerns they face in the context of federated learning. This feedback can then be used to refine the existing group definitions, ensuring that they capture the nuances and complexities of the social dynamics at play.

Furthermore, the development of adaptive algorithms and techniques for updating group definitions in response to evolving social contexts and client preferences represents an important research direction. These algorithms would need to incorporate mechanisms for detecting changes in the social landscape, tracking emerging fairness considerations, and adapting the group definitions accordingly. This adaptability would allow the federated learning system to stay responsive and aligned with the evolving needs and values of the communities it serves.

The aforementioned open challenges and potential future directions offer exciting opportunities to advance the broader field of federated learning. In summary, forthcoming research efforts will concentrate on developing methodologies that achieve both group fairness and privacy preservation, exploring the dynamic adaptation of group definitions to promote inclusivity and alignment with societal values, and investigating approaches to incorporate partial knowledge of existing demographics into the federated fair learning procedure. These endeavours will contribute to the ongoing progress and enhancement of federated learning as a whole.

Bibliography

[Agarwal et al., 2018] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

[Agarwal et al., 2019] Agarwal, A., Dudík, M., and Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. *CoRR*, abs/1905.12843.

[Allen-Zhu, 2018] Allen-Zhu, Z. (2018). Natasha 2: Faster non-convex optimization than sgd. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[Awasthi et al., 2020] Awasthi, P., Kleindessner, M., and Morgenstern, J. (2020). Equalized odds postprocessing under imperfect group information. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1770–1780. PMLR.

[Barocas et al., 2019a] Barocas, S., Hardt, M., and Narayanan, A. (2019a). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[Barocas et al., 2019b] Barocas, S., Hardt, M., and Narayanan, A. (2019b). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

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

[Ben-Tal et al., 2013] Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Manage. Sci.*, 59(2):341–357.

[Bertran et al., 2019] Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M., Reeves, G., and Sapiro, G. (2019). Adversarially learned representations for information obfuscation and inference. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 614–623. PMLR.

[Bertran et al., 2018] Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M. R. D., and Sapiro, G. (2018). Learning representations for utility and privacy: An information-theoretic based approach. In *Workshop on Privacy Preserving Machine Learning (in Conjunction with NeurIPS 2018)*.

[Bertsekas and Tsitsiklis, 1999] Bertsekas, D. and Tsitsiklis, J. (1999). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10.

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

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

[Calders et al., 2009] Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. pages 13–18.

[Caton and Haas, 2020] Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *CoRR*, abs/2010.04053.

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

[Chen et al., 2019] Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019). Fairness under unawareness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

[Chen et al., 2017a] Chen, J., Pan, X., Monga, R., Bengio, S., and Jozefowicz, R. (2017a). Revisiting distributed synchronous sgd.

[Chen et al., 2017b] Chen, R. S., Lucier, B., Singer, Y., and Syrgkanis, V. (2017b). Robust optimization for non-convex objectives. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

[Chouldechova and Roth, 2020] Chouldechova, A. and Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Commununications of the ACM*, 63(5):82–89.

[Chu et al., 2021] Chu, L., Wang, L., Dong, Y., Pei, J., Zhou, Z., and Zhang, Y. (2021). Fedfair: Training fair models in cross-silo federated learning.

[Cui et al., 2021] Cui, S., Pan, W., Liang, J., Zhang, C., and Wang, F. (2021). Addressing algorithmic disparity and performance inconsistency in federated learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

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

[Deng et al., 2020] Deng, Y., Kamani, M. M., and Mahdavi, M. (2020). Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33.

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

[Ding et al., 2021] Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

[Divi et al., 2021] Divi, S., Lin, Y., Farrukh, H., and Celik, Z. B. (2021). New metrics to evaluate the performance and fairness of personalized federated learning. *CoRR*, abs/2107.13173.

[Du and Wu, 2021] Du, W. and Wu, X. (2021). Robust fairness-aware learning under sample selection bias. *CoRR*, abs/2105.11570.

[Du et al., 2020] Du, W., Xu, D., Wu, X., and Tong, H. (2020). Fairness-aware agnostic federated learning. *CoRR*, abs/2010.05057.

[Duchi and Namkoong, 2020] Duchi, J. and Namkoong, H. (2020). Learning models with uniform performance via distributionally robust optimization.

[Duchi et al., 2008] Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l1-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. Association for Computing Machinery.

[Duchi et al., 2020] Duchi, J. C., Hashimoto, T., and Namkoong, H. (2020). Distributionally robust losses for latent covariate mixtures. *CoRR*, abs/2007.13982.

[Dwork et al., 2011] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2011). Fairness through awareness. *CoRR*, abs/1104.3913.

[Elliott et al., 2009] Elliott, M., Morrison, P., Fremont, A., Mccaffrey, D., Pantoja, P., and Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:252–253.

[Elliott et al., 2008] Elliott, M. N., Fremont, A. M., Morrison, P. A., Pantoja, P. M., and Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research*, 43 5 Pt 1:1722–36.

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

[Fan et al., 2021] Fan, Z., Fang, H., Zhou, Z., Pei, J., Friedlander, M. P., Liu, C., and Zhang, Y. (2021). Improving fairness for data valuation in federated learning.

[Gálvez et al., 2022] Gálvez, B. R., Granqvist, F., van Dalen, R., and Seigel, M. (2022). Enforcing fairness in private federated learning via the modified method of differential multipliers.

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

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

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

[Gupta and Kamble, 2021] Gupta, S. and Kamble, V. (2021). Individual fairness in hindsight. *Journal of Machine Learning Research*, 22(144):1–35.

[Hardt et al., 2016a] Hardt, M., Price, E., and Srebro, N. (2016a). 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. Curran Associates Inc.

[Hardt et al., 2016b] Hardt, M., Recht, B., and Singer, Y. (2016b). Train faster, generalize better: Stability of stochastic gradient descent. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA. PMLR.

[Hashimoto et al., 2018] Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR.

[Horváth et al., 2021] Horváth, S., Laskaridis, S., Almeida, M., Leontiadis, I., Venieris, S., and Lane, N. D. (2021). FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

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

[Hu et al., 2020] Hu, Z., Shaloudegi, K., Zhang, G., and Yu, Y. (2020). Fedmgda+: Federated learning meets multi-objective optimization. *CoRR*, abs/2006.11489.

[Iacob et al., 2023] Iacob, A., de Gusmao, P. P. B., and Lane, N. D. (2023). Can fair federated learning reduce the need for personalization?

[Iosifidis et al., 2020] Iosifidis, V., Fetahu, B., and Ntoutsi, E. (2020). FAE: A fairness-aware ensemble framework. *CoRR*, abs/2002.00695.

[Juárez and Korolova, 2022] Juárez, M. and Korolova, A. (2022). You can't fix what you can't measure: Privately measuring demographic performance disparities in federated learning. *CoRR*, abs/2206.12183.

[Juditsky et al., 2011] Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17 – 58.

[Kairouz et al., 2019] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2019). Advances and open problems in federated learning. *CoRR*, abs/1912.04977.

[Kallus et al., 2019] Kallus, N., Mao, X., and Zhou, A. (2019). Assessing algorithmic fairness with unobserved protected class using data combination. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

[Kilbertus et al., 2017] Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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

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

[Krizhevsky et al.,] Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research).

[Kusner et al., 2017] Kusner, M., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4069–4079, Red Hook, NY, USA. Curran Associates Inc.

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

[Lam and Zhou, 2015] Lam, H. and Zhou, E. (2015). Quantifying uncertainty in sample average approximation. In *2015 Winter Simulation Conference (WSC)*, pages 3846–3857.

[Li and Liu, 2022] Li, P. and Liu, H. (2022). Achieving fairness at no utility cost via data reweighing with influence. In Chaudhuri, K., Jegelka, S., Song, L.,

Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12917–12930. PMLR.

[Li et al., 2021] Li, T., Beirami, A., Sanjabi, M., and Smith, V. (2021). Tilted empirical risk minimization. In *International Conference on Learning Representations*.

[Li et al., 2020a] Li, T., Hu, S., Beirami, A., and Smith, V. (2020a). Federated multi-task learning for competing constraints. *CoRR*, abs/2012.04221.

[Li et al., 2020b] Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2020b). Fair resource allocation in federated learning. In *International Conference on Learning Representations*.

[Lichman, 2013] Lichman, M. (2013). UCI machine learning repository.

[Lin et al., 2022] Lin, S., Han, Y., Li, X., and Zhang, Z. (2022). Personalized federated learning towards communication efficiency, robustness and fairness. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30471–30485. Curran Associates, Inc.

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

[Lyu et al., 2020] Lyu, L., Xu, X., Wang, Q., and Yu, H. (2020). *Collaborative Fairness in Federated Learning*, pages 189–204. Springer International Publishing, Cham.

[Makhlouf et al., 2021] Makhlouf, K., Zhioua, S., and Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642.

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

[Martinez et al., 2020a] Martinez, N., Bertran, M., and Sapiro, G. (2020a). Minimax pareto fairness: A multi objective perspective. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6755–6764. PMLR.

[Martinez et al., 2020b] Martinez, N. L., Bertran, M. A., Papadaki, A., Rodrigues, M., and Sapiro, G. (2020b). Pareto robustness for fairness beyond demographics. In *Fair AI in Finance at the 34th Conference on Neural Information Processing Systems*, NeurIPS '20, Virtual.

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

[Mas-Colell et al., 1995] Mas-Colell, A., Whinston, M., and Green, J. (1995). *Microeconomic Theory*. Oxford University Press.

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

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

[Mertikopoulos et al., 2020] Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. (2020). On the almost sure convergence of stochastic gradient descent in non-convex problems. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1117–1128. Curran Associates, Inc.

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

[Mohri et al., 2019] Mohri, M., Sivek, G., and Suresh, A. (2019). Agnostic federated learning. In *36th International Conference on Machine Learning, ICML 2019*, 36th International Conference on Machine Learning, ICML 2019, pages 8114–8124. International Machine Learning Society (IMLS). 36th International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019.

[Munir et al., 2021] Munir, M. T., Saeed, M. M., Ali, M., Qazi, Z. A., and Qazi, I. A. (2021). Fedprune: Towards inclusive federated learning.

[Murty and Kabadi, 1987] Murty, K. G. and Kabadi, S. N. (1987). Some np-complete problems in quadratic and nonlinear programming. *Math. Program.*, 39(2):117–129.

[Nagalapatti and Narayanam, 2021] Nagalapatti, L. and Narayanam, R. (2021). Game of gradients: Mitigating irrelevant clients in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9046–9054.

[Namkoong and Duchi, 2016] Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[Nemirovski et al., 2009] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609.

[Nemirovski and Yudin, 1983] Nemirovski, A. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley Interscience.

[Orabona, 2020] Orabona, F. (2020). Almost sure convergence of sgd on smooth non-convex functions. https://parameterfree.com/2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions/. Accessed: 2023-06-21.

[Papadaki et al., 2021] Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. (2021). Federating for learning group fair models. In *New Frontiers in Federated Learning: Privacy, Fairness, Robustness, Personalization and Data Ownership at the 35th Conference on Neural Information Processing Systems*, NeurIPS '21, Virtual.

[Papadaki et al., 2022a] Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. (2022a). Minimax demographic group fairness in federated learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 142–159, New York, NY, USA. Association for Computing Machinery.

[Papadaki et al., 2022b] Papadaki, A., Martinez, N., Bertran, M. A., Sapiro, G., and Rodrigues, M. R. D. (2022b). Federated fairness without access to demographics. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*.

[Papadaki et al., 2023] Papadaki, A., Martinez, N., Bertran, M. A., Sapiro, G., and Rodrigues, M. R. D. (2023). Federated fairness without demographics. *Under Review*.

[Pillutla et al., 2021] Pillutla, K., Laguel, Y., Malick, J., and Harchaoui, Z. (2021). Federated learning with heterogeneous data: A superquantile optimization approach. *CoRR*, abs/2112.09429.

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

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

[Reisizadeh et al., 2020] Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. (2020). Robust federated learning: The case of affine distribution shifts.

[Ro et al., 2021] Ro, J., Chen, M., Mathews, R., Mohri, M., and Suresh, A. T. (2021). Communication-Efficient Agnostic Federated Averaging. In *Proc. Interspeech 2021*, pages 871–875.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.

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

[Sebbouh et al., 2021] Sebbouh, O., Gower, R. M., and Defazio, A. (2021). Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball.

[Shamir and Zhang, 2013] Shamir, O. and Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79, Atlanta, Georgia, USA. PMLR.

[Shapiro et al., 2009] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on stochastic programming. Modeling and theory*.

[Sharma et al., 2023] Sharma, P., Panda, R., and Joshi, G. (2023). Federated minimax optimization with client heterogeneity.

[Sinha et al., 2018] Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*.

[Sohoni et al., 2020] Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. (2020). No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H.,

editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19339–19352. Curran Associates, Inc.

[Soma and Yoshida, 2020] Soma, T. and Yoshida, Y. (2020). Statistical learning with conditional value at risk.

[Steen and Tanenbaum, 2016] Steen, M. and Tanenbaum, A. S. (2016). A brief introduction to distributed systems. *Computing*, 98(10):967–1009.

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

[Wang et al., 2021b] Wang, Z., Fan, X., Qi, J., Wen, C., Wang, C., and Yu, R. (2021b). Federated learning with fair averaging. In *IJCAI*.

[Williamson and Menon, 2019] Williamson, R. and Menon, A. (2019). Fairness risk measures. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR.

[Xiao et al., 2017] Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[Yu, 2021] Yu, Z. (2021). Fair balance: Mitigating machine learning bias against multiple protected attributes with data balancing. *CoRR*, abs/2107.08310.

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

[Zafar et al., 2017] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification.

[Zeng et al., 2021] Zeng, Y., Chen, H., and Lee, K. (2021). Improving fairness via federated learning.

[Zhang et al., 2020] Zhang, D. Y., Kou, Z., and Wang, D. (2020). 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.

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

[Zhang, 2016] Zhang, Y. (2016). Assessing fair lending risks using race/ethnicity proxies. *Comparative Political Economy: Regulation eJournal*.

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

[Zliobaite, 2015] Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *CoRR*, abs/1505.05723.

Appendices

Appendix A

Basic Definitions

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

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

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

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

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

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

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

$$||\nabla f(w)|| \leq G \quad \text{and} \quad |f(w) - f(v)| \leq G||w - v||, \text{ for some } G > 0. \tag{A.2}$$

Lemma A.1 (First Order Condition)**.** *A differentiable function f with domain $dom(f)$ is convex if $dom(f)$ is convex and for any $v, w \in dom(f)$ we have that*

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

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

Definition A.4 (Smoothness)**.** *A function $f : dom(f) \to \mathbb{R}$ is β-smooth if it is continuously differentiable and its gradient is β-Lipschitz, i.e. $\exists \beta : \forall v, w \in dom(f)$ we have that*

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

Definition A.5 (Sub-gradient)**.** *Let $f : dom(f) \to \mathbb{R}$, with $dom(f) \subseteq \mathbb{R}^d$. Then $g \in \mathbb{R}^d$ is a subgradient of f at point v if for any $w \in dom(f)$ we have that*

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

Note that the subgradient g might not be unique. We denote $\partial f(v)$ the set of subgradients computed at a point v, also called subdifferential of f at v, where $g \in \partial f(v)$. We also note that when a function $f : dom(f) \to \mathbb{R}$ is G-Lipschitz and convex, Eq. A.2 becomes $||g|| \leq G$. We provide this elementary proof in Lemma A.2.

Lemma A.2. *Let a function $f : dom(f) \to \mathbb{R}$ be a $G-$Lipschitz continuous and $\partial f(v) \neq \emptyset$. Then, for any $v \in dom(f)$ we have that*

$$||g|| \leq G, \quad \text{with} \quad g \in \partial f(v) \tag{A.3}$$

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

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

Also from the subgradient definition, we know that

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

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

□

Definition A.6 (Uniform Stability, [Bousquet and Elisseeff, 2002]). *Let $f : dom(f) \to \mathbb{R}$. A randomized algorithm $\mathscr{A}$ is $\zeta-$uniformly stable if, for any datasets D, D' that differ in at most a single sample, we have that*

$$\sup_{z} \mathbb{E}[f(\mathscr{A}(D); z) - f(\mathscr{A}(D'); z)] \leq \zeta \tag{A.4}$$

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

Appendix B

Additional Material for Chapter 3

B.1 Analytical Results

In this subsection, we present analytical tables and provide detailed numerical values that correspond to the results and approaches discussed in Chapter 3.

B.1.1 Experiments on a Synthetic dataset

The risks associated with different groups are presented in Table B.1, along with the corresponding weighting coefficients for each sensitive group, which are displayed in Table B.2.

Table B.1: 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 $|\mathscr{A}| = 2$ it is equivalent to SSG.

Setting	Method	Worst Group	Best Group
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**
Centalized Minmax Baseline		**0.451±0.0**	**0.308±0.001**

Table B.2: 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
Centalized Minmax Baseline		0.999	0.001

B.1.2 Experiments on the Adult dataset

We show the testing group risks in Table B.3 and the group weights in Table B.4.

Table B.3: Cross entropy risks for FedAvg, AFL, q-FedAvg, TERM, and FedMinmax across different federated learning settings on the Adult dataset.

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

Table B.4: 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, $\leq$ 50K	Males, $>$ 50K	Females, $\leq$ 50K	Females, $>$ 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
Centalized Minmax Baseline		0.697	0.301	0.001	0.001

B.1.3 Experiments on the ACS Employment dataset (employment and race combination)

We report the risks on the test set in Table B.6 and the group weighting coefficients produced from the training process are in Table B.5.

Table B.5: Final group weighting coefficients for AFL, Centalized 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
Centalized Minmax Baseline		0.461	0.356	0.040	0.045	0.070	0.029

Table B.6: 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
Centalized 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

B.1.4 Experiments on the ACS Employment dataset (race)

We report the risks on the test set in Table B.7.

Table B.7: Risks for FedAvg, AFL, q-FFL, TERM, and FedMinmax across different federated learning settings on ACS Employment dataset.

Setting	Method	White	Black/ African American	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
	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
	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
Centalized 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

B.1.5 Experiments on the FashionMNIST dataset

The group risks are provided in Table B.8. We also show the weighting coefficients for each sensitive group in Table B.9.

Table B.8: Brier score risks for FedAvg, AFL, q-FedAvg, TERM, and FedMinmax across different federated learning settings on the FashionMNIST dataset.

Setting	Method	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
ESG	AFL	0.239±0.003	0.046±0.0	0.262±0.001	0.159±0.001	0.252±0.004	0.06±0.0	0.494±0.004	0.067±0.001	0.049±0.0	0.07±0.001
	FedAvg	0.243±0.003	0.046±0.0	0.262±0.001	0.158±0.003	0.253±0.002	0.061±0.0	0.492±0.003	0.068±0.0	0.049±0.0	0.069±0.0
	q-FedAvg	0.268±0.051	0.047±0.005	0.312±0.016	0.164±0.029	0.306±0.052	0.039±0.003	0.477±0.006	0.074±0.001	0.036±0.005	0.056±0.008
	TERM	0.256±0.066	0.048±0.008	0.31±0.083	0.175±0.022	0.294±0.016	0.041±0.012	0.467±0.002	0.066±0.019	0.038±0.011	0.062±0.018
	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	AFL	0.267±0.009	0.194±0.023	0.236±0.013	0.226±0.012	0.262±0.012	0.201±0.026	**0.307±0.003**	0.178±0.033	0.205±0.025	0.162±0.021
	FedAvg	0.227±0.003	0.039±0.001	0.236±0.004	0.143±0.003	0.232±0.003	0.051±0.001	0.463±0.003	0.067±0.0	0.041±0.0	0.063±0.001
	q-FedAvg	0.24±0.001	0.041±0.008	0.246±0.026	0.142±0.014	0.257±0.028	0.036±0.001	0.425±0.002	0.059±0.014	0.027±0.002	0.042±0.007
	TERM	0.251±0.011	0.034±0.003	0.26±0.017	0.144±0.005	0.242±0.034	0.04±0.004	0.399±0.017	0.05±0.003	0.026±0.001	0.044±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	AFL	0.244±0.007	0.032±0.001	0.257±0.066	0.122±0.006	0.209±0.098	0.045±0.002	0.425±0.019	0.059±0.001	0.041±0.001	0.062±0.001
	FedAvg	0.229±0.008	0.039±0.0	0.236±0.004	0.142±0.002	0.232±0.003	0.052±0.001	0.464±0.011	0.067±0.001	0.042±0.001	0.063±0.001
	q-FedAvg	0.278±0.062	0.04±0.013	0.256±0.083	0.16±0.026	0.311±0.044	0.045±0.013	0.453±0.002	0.063±0.02	0.029±0.007	0.047±0.004
	TERM	0.226±0.007	0.037±0.005	0.233±0.004	0.153±0.007	0.255±0.016	0.038±0.0	0.439±0.007	0.053±0.003	0.026±0.001	0.043±0.002
	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
Centalized Minmax Baseline		0.259±0.01	0.173±0.015	0.239±0.051	0.213±0.008	0.24±0.063	0.182±0.024	**0.311±0.006**	0.168±0.018	0.18±0.013	0.151±0.012

Table B.9: Final group weighting coefficients for AFL, Centalized 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
Centalized Minmax Baseline		0.217	0.001	0.240	0.010	0.152	0.001	**0.377**	0.001	0.001	0.001

B.1.6 Experiments on the CIFAR-10 dataset

We report the risks on the test set in Table B.10 and the final group weighting coefficients in Table B.11.

Table B.10: Brier score risks for FedAvg, AFL, q-FedAvg, TERM, and FedMinmax across different federated learning settings on the 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
Centalized 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 B.11: Final group weighting coefficients for AFL, Centalized 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
Centalized Minmax Baseline		0.082	0.017	0.139	**0.205**	0.118	0.190	0.091	0.080	0.032	0.046

B.1.7 Empirical Results Comparing LocalFedMinMax and FedMinMax

Table B.12: Brier Score risks for FedMinMax and LocalFedMinMax on FashionMNIST across the different federated learning scenarios.

Setting	Method	T-shirt	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
ESG	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
(10 clients)	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	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
(10 clients)	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	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
(10 clients)	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	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
(40 clients)	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	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
(40 clients)	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	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
(40 clients)	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 B.13: 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	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
(10 clients)	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	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
(10 clients)	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	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
(10 clients)	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	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
(40 clients)	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	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
(40 clients)	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	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
(40 clients)	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

Appendix C

Additional Material for Chapter 5

C.1 Alternative Algorithm for optimizing RCVaR

C.1.1 Federated RCVaR Algorithm

We present an algorithm to solve the empirical objective of Eq. 5.2, in a federated way. The algorithm consists of two main steps: (a) model parameters update and (b) periodic calculation of the threshold c.

Model update: The model learning process involves a straightforward procedure. At each round, the server sends the current model parameters $\boldsymbol{\theta}^t$ to the participating clients. The clients then perform an optimization step using the received parameters and return their updated model parameters $\boldsymbol{\theta}_k^t$ to the server. The server aggregates the client parameters to obtain the new model parameters $\boldsymbol{\theta}^{t+1}$ by averaging them.

Threshold calculation: The computation of the global model parameters $\boldsymbol{\theta}$ and quantile c is performed exclusively on the server side, since it involves aggregating relevant information from the clients. Nevertheless, as discussed in Section 5.3.1, efficiently calculating the quantile c in a federated manner poses a significant challenge without significantly increasing the communication overhead per client. To address this challenge, we propose a practical technique for estimating the quantile parameter, which allows for an effective and scalable approach within the federated learning framework.

In particular, let c be the estimated quantile. We denote $\rho(c) = \sum_{k \in \mathcal{K}} p(K = k)\rho_k(c)$ the estimated probability, with $\rho_k(c) = \mathop{\mathbb{E}}_{(X,Y)\sim p(X,Y|K=k)}[\mathbb{1}(L_{h,X,Y|K=k} \geq c)]$.

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

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

Note that η_c captures the product of the learning rate and the absolute value of the derivative $\big(\rho - \rho(c)\big)^2$ w.r.t. c. We summarize the proposed solver in Algorithm 6.

Algorithm 6 FEDERATED RCVAR (FEDRCVAR) ALGORITHM

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

1: Server initializes $\boldsymbol{\theta}^0$ randomly.
2: **for** $t = 1$ to T **do**
3: Server **broadcasts** $\boldsymbol{\theta}^{t-1}$ and c^t
4: **for** each client $k \in \mathscr{K}$ **in parallel do**
5: $\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}^{t-1} - \eta_\theta \nabla_\theta \left\{ \frac{1-\varepsilon}{n_k} \sum_{i=1}^{n_k} (\ell(h(x_i^k), y_i^k) - c^t)_+ + \frac{\varepsilon}{n_k} \sum_{i=1}^{n_k} \ell(h(x_i^k), y_i^k) \right\}$
6: Return local model $\boldsymbol{\theta}_k^t$ and $\rho_k(c^t) = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{1}(\ell(h(x_i^k), y_i^k) \geq c^t)$ to server
7: **end for**
8: $\boldsymbol{\theta}^t \leftarrow \sum_{k\in\mathscr{K}} \frac{n_k}{n} \boldsymbol{\theta}_k^t$
9: $c^{t+1} \leftarrow \prod_{c\in[0,B]} \left(c^t - \eta_c \operatorname{sign}(\rho - \sum_{k\in\mathscr{K}} \frac{n_k}{n} \rho_k(c^t)) \right)$
10: **end for**

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

C.1.2 Experimental Results

We show the benefits of the proposed approach in a *synthetic dataset* (see C.1.3 for dataset generation details) and two real-world datasets: *eICU dataset* [Pollard et al., 2018] and *ACS Employment dataset* [Ding et al., 2021]. We offer the description and setup of these datasets in Section 5.5. For the centralized settings,

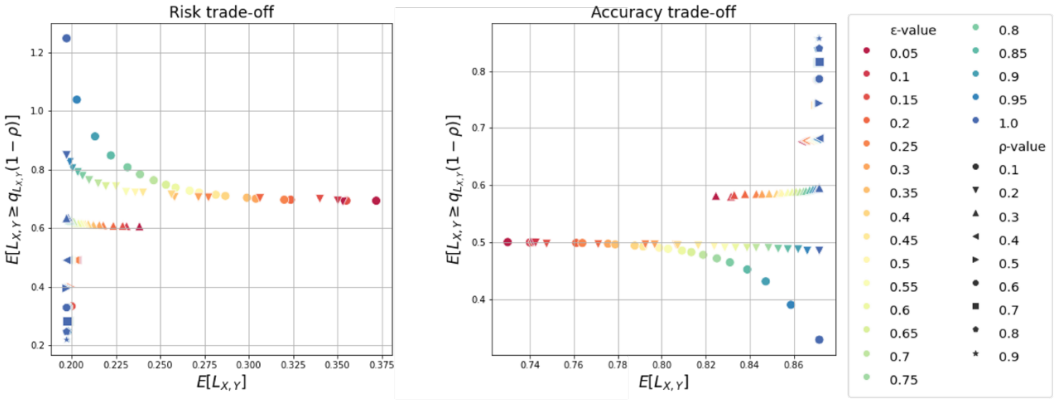

Figure C.1: Toy example illustrating the flexibility of RCVaR objective for hyperparameter $\varepsilon \in (0,1]$ and $\rho \in (0,1)$ on synthetic data. FedRCVaR is trained for $\rho = \{0.1,\ldots,0.9\}$ and $\varepsilon = \{0.05,0.1,\ldots,0.95,1.0\}$. Different colors describe various ε values,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

there are no clients involved and we use all the available training data during the training process.

In our empirical analysis, we first investigate the impact of different levels of group fairness achievable through the parameter ε, considering a fixed group size ρ, when optimizing the RCVaR objective. Figure C.1 illustrates this exploration. Notably, the RCVaR objective emphasizes the influence of the worst-performing group while reducing the emphasis on average performance when $\varepsilon \approx 0$. When $\varepsilon = 1$, the objective reverts to the standard Empirical Risk Minimization (ERM). We also observe that as the value of ρ decreases, there is a greater trade-off between fairness and utility achievable through ε. It is worth noting that when ε approaches 0 and ρ is sufficiently small, the risk for the worst-performing group aligns more closely with the risk of the uniform classifier. This observation is consistent with previous findings in [Martinez et al., 2021] but also Remark 4.1, which highlight the existence of a critical ρ that corresponds to a particular threshold.

In Figure C.2, we present a comprehensive comparison between our proposed approach and several baseline methods, including centralized ML approaches: ERM, BPF; and relevant FL approaches: AFL, FedAvg. The proposed approach achieves the Pareto optimal subgroup robust solution when $\varepsilon \approx 0$, which is also achieved by BPF. Furthermore, it produces a solution similar to the centralized ERM approach

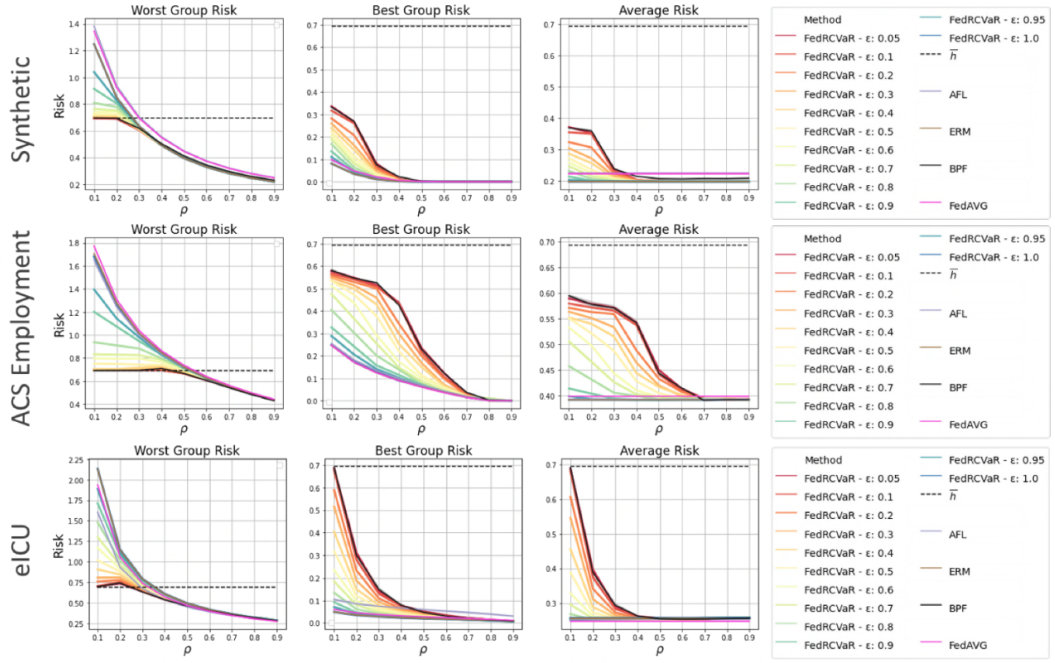

Figure C.2: Cross entropy risks comparison on synthetic, ACS Employment and eICU datasets. FedRCVaR recovers solutions equivalent to centralized machine learning for $\varepsilon = \{0.05, 1.0\}$, while improving both utility and accuracy compared to FL baselines in many settings.

when $\varepsilon = 1$. Interestingly, for specific values of ε and ρ, FedRCVaR also achieves client robustness, even though our objective was not explicitly designed for this purpose.

C.1.3 Synthetic Dataset

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