Robust Personalized Federated Learning under Demographic Fairness Heterogeneity

Alycia N. Carey, Wei Du, Xintao Wu
Department of Computer Science and Computer Engineering
University of Arkansas
Fayetteville, Arkansas, USA
{ancarey, wd005, xintaowu}@uark.edu

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020554

***Abstract*—Personalized federated learning (PFL) gives each client in a federation the power to obtain a model tailored to their specific data distribution or task without the client forfeiting the benefits of training in a federated manner. However, the concept of demographic group fairness has not been widely studied in PFL. Further, fairness heterogeneity – when not all clients enforce the same local fairness metric – has not been studied at all. To fill this gap, we propose *Fair Hypernetworks* (FHN), a personalized federated learning architecture based on hypernetworks that is robust to statistical (e.g., non-IID and unbalanced data) and fairness heterogeneity. We theoretically show that granting clients the ability to independently choose multiple (possibly conflicting) fairness constraints, such as demographic parity or equalized odds, does not break previously proven generalization bounds on hypernetworks used in the federated setting. Additionally, we empirically test FHN against several baselines in multiple fair federated learning settings, and we find that FHN outperforms all other federated baselines when handling clients with heterogeneous fairness metrics. We further demonstrate the scalability of FHN to show that minimal degradation to the accuracy and the fairness of the clients occurs when the federation grows in size. Additionally, we empirically validate our theoretical analysis to show FHN generalizes well to new clients. To our knowledge, our FHN architecture is the first to consider tolerance to fairness heterogeneity which gives clients the freedom to personalize the fairness metric enforced during local training.**

***Index Terms*—fairness, hypernetworks, personalized federated learning**

I. Introduction

In federated learning (FL), distributed clients, who each own a private local dataset, jointly train a global machine learning model without having to share their private data with either the other clients or with the global server [1]. While FL provides many benefits including edge computation, increased privacy protection, and communication efficiency, it also has several shortcomings. First, most FL architectures assume that every client in the federation possesses data that are uniformly drawn from the same data distribution. When this assumption of independent and identically distributed (IID) data does not hold, the accuracy of the learned model can degrade up to ∼55% [2]. Unfortunately, real world datasets are often comprised of non-IID data which presents a challenge for deploying FL architectures. Additionally, most federated learning architectures rely on the assumption that each client implements the same task (e.g., architecture and optimization function) and task heterogeneity – when clients have different local tasks – is often not considered. These restrictions constrain the clients in terms of how they can personalize their local model to their specific problem and dataset.

For an illustrative example, consider a federation of hospitals across different countries. Most countries have distinct laws and regulations in regards to privacy (e.g., the General Data Protection Regulation in the European Union vs. Health Insurance Portability and Accountability Act in the USA), and it is likely that regulations will eventually be implemented to require demographic fairness when client data is utilized in a learning process. However, the definition of fairness is highly subjective and changes country to country due to differences in culture. This dissonance would manifest in the regulations enforced in one country disagreeing with or even contradicting the regulations of another. These differences in policy would result in Hospital A (e.g., in the USA) enforcing demographic parity while Hospital B (e.g., in Spain) enforces equalized odds. In this setting, a global model trained through standard federated learning would not align with the wanted fairness constraint of any individual client. In addition, the global model would have poor accuracy due to statistical and fairness heterogeneity.

To overcome the challenges present in standard FL, personalized federated learning (PFL) has been proposed. The main goal of PFL is to allow each client to train a personalized local model while still receiving the benefits of standard FL (e.g., overcoming data limitations and continual learning). Many different methods for PFL have been proposed [3]–[6], and one such method is the use of a hypernetwork as the global model [4]. Instead of performing a standard machine learning task such as classification or regression, the main purpose of a hypernetwork is to generate the network parameters for other models. Hypernetworks are naturally suitable for PFL as they learn a diverse set of personalized models by conditioning on an input tailored to each individual client.

In addition to the increased research on PFL, FL has also become concerned with the issue of fairness. In centralized learning, fairness refers to treating members of different demographic groups equally. On the other hand, in FL fairness is often equated to client parity [7]–[10] and only a small portion of current works focus on demographic group fairness [11]–[16]. Additionally, these few approaches are not personalizable

and they require each client to enforce the same fairness constraint.

To alleviate these restrictions, we propose *Fair Hypernetworks* (FHN). FHN is based on the work [4] of Shamsian et al. who first proposed the use of hypernetworks for PFL. We extend their work and incorporate fairness by formulating the client's chosen fairness metric as a linear constraint on the local optimization function [17]. In FHN, each client is able to individually choose which fairness metric to enforce during local training. In other words, FHN is robust against fairness heterogeneity – even in the case of conflicting fairness constraints, for instance when some clients use demographic parity while other clients use equalized odds.

Our major contributions include:

1) The presentation of FHN: an architecture for fair personalized federated learning based on hypernetworks that is robust against statistical and fairness heterogeneity;
2) A theoretical analysis into why FHN can produce accurate client network parameters even when clients enforce different fairness metrics on their local optimization function; and
3) The empirical evaluation of FHN against several baselines in the non-IID setting to show how FHN can handle heterogeneous fairness with only minor accuracy degradation while other methods fall short.

The remainder of the paper is organized as follows. We begin by introducing closely related works in Section II. In Section III we formulate and present our FHN framework and provide theoretical analysis on the generalizability of FHN to new clients. In Section IV we empirically show that FHN is able to produce accurate personalized models under statistical and fairness heterogeneity. Finally, in Section V we provide our concluding remarks.

II. Related Work

Federated learning was first proposed in 2016 by McMahan et al. [1], who desired to train a high-quality centralized model without requiring aggregation of the distributed clients' private data. Specifically, the authors proposed the Federated Averaging (FedAvg) architecture. In each round of FedAvg, a subset of clients obtain the global model's parameters, train their local model for a set number of rounds, and then return the newly updated parameters to the global model. The global model then performs a weighted aggregation of the received parameters based on the amount of data each client has, and then sets the averaged parameters as the new global parameters. This process continues until the desired level of fairness is reached by the global model. Many of the proposed FL architectures use FedAvg as a foundation and improve upon it by decreasing communication costs or increasing client parity (i.e., achieving the same accuracy for each client). Despite these advancements in the FL field, there are still several open problems and challenges to be solved, such as incorporating fairness [18], dealing with non-IID data [19]–[23], and giving clients the ability to obtain personalized models [24].

A. Fair Federated Learning

A major focus of recent machine learning research is demographic fairness – ensuring that a machine learning model treats individuals from different demographic groups (e.g., based on race, gender, etc.) similarly. Despite there not being one set definition of fairness [25]–[27], multiple approaches to achiving both individual [25], [28] and group [29]–[31] demographic fairness have been proposed, including pre-processing [26], [32], in-processing [17], [27], [33], and post-processing [26], [34], [35] methods. However, most of the proposed techniques require access to the sensitive variable of each data point, making them unsuitable for the federated setting [16]. While the majority of research on fairness in FL is centered around client parity [7]–[10], works focusing on demographic group fairness are consistently gaining popularity. Specifically, [11]–[16] all aim to achieve demographic group fairness in a federated setting.

One of the first fairness-aware federated learning approaches is [11]. To achieve their three concurrent goals of fairness, accuracy, and privacy, the authors proposed FairFL, a fair federated learning framework based on deep multi-agent reinforcement learning and secure information aggregation. Instead of using reinforcement learning, [12], [14], [36] and [37] all approached demographic group fairness in FL through solving a fairness-constrained optimization problem using a modified version of FedSGD or FedAvg [1]. [13] and [38] additionally approached the problem through enforcing fairness on the optimization function, but instead of using a modified version of FedSGD or FedAvg, they each constructed a new optimization procedure which are named Alternating Gradient Projection and Federated Mirror Descent Ascent with Momentum Acceleration, respectively. [15] and [16] achieved fairness through a slight modification of the FedAvg aggregation weights to help guarantee demographic group fairness. But while [16] only applies in cases with a single binary sensitive attribute, [15] can be applied in more general cases. While all the mentioned approaches require each client to enforce the same fairness metric, our architecture of FHN allows each client to independently choose which metric they enforce locally. Additionally, while all the mentioned approaches use standard federated learning, we approach the problem using a hypernetwork – a personalized federating learning technique.

B. Personalized Federated Learning

Personalized federated learning (PFL) allows each client in a federation to modify their local model to better fit their data or task. Many different approaches to PFL have been proposed, such as: user clustering and collaborating [6], [39], transfer learning [40], multi-task learning [41], and meta-learning [3]. One emerging approach to PFL is the use of hypernetworks instead of a standard federated learning architecture. The idea of the hypernetwork was originally proposed by Schmidhuber in 1992 when he suggested that one network can be used to produce context-dependent parameters for another [42].

In [43], the hypernetwork is trained end-to-end with gradient descent in conjunction with the main network. The hypernetwork takes in an input that contains information about the structure of the network layers and generates the parameters for each layer. In [4], the authors proposed pFedHN which builds off the idea of [43] and uses a global hypernetwork model to generate the parameters for each of the client's local models. In pFedHN, each client has a unique embedding vector stored on the global server, which is passed as input to the hypernetwork to produce the client's personalized model parameters. Additionally, similar to other FL works, in pFedHN the clients perform multiple rounds of local training to reduce communication costs. But, instead of sending the updated local gradients obtained at the end of training, the clients send the difference between their updated parameters and the parameters sent from the hypernetwork at the beginning of the round. The authors also theoretically analyzed why hypernetworks can generalize to new clients and how information is shared between the clients.

Despite PFL allowing clients to have more control over their local models, the concept of demographic group fairness, especially the idea of fairness heterogeneity, in PFL has not been explored. [16] proposed the idea of each client enforcing their own local debiasing strategy. But this is different from our approach as we focus on in-processing methods and not pre-processing. Additionally, the authors of [16] did not experimentally test their theory and left it as future work. To our knowledge, our work is the first to formally explore personalized federated learning under fairness heterogeneity.

III. Problem Formulation

In this section, we formulate and present our fair hypernetworks (FHN) framework. We begin by giving background information on in-processing techniques for fairness as well as for PFL and then conclude with the presentation of FHN. An overview of major symbols used can be seen in Table I and we give a brief description of them below.

We consider the following setting for FHN. Let n denote the number of clients and let each client i have their own private data distribution $\mathcal{P}^i_{\mathcal{X}^i\times\mathcal{Y}^i}$ which, for brevity, we shorten to $\mathcal{P}^i_{\mathcal{X}\times\mathcal{Y}}$. Let $\mathcal{D}^i$ denote each client's individual dataset, which contains m^i IID data samples drawn from $\mathcal{P}^i_{\mathcal{X}\times\mathcal{Y}}$. We note that while each client has IID samples locally, the data distribution between each client is non-IID. The k-th data point from $\mathcal{D}^i$ is of the form $(\boldsymbol{x}^i_k, y^i_k)$ with $\boldsymbol{x}^i_k \in \mathcal{X}^i$ denoting the feature vector and $y^i_k \in \mathcal{Y}^i$ denoting the label. Our main task is to formulate a personalized classification model $f_{\theta^i}: \mathcal{X}^i \to \mathcal{Y}^i$ for each individual client in a collaborative manner without revealing each client's dataset $\mathcal{D}^i$. Let h denote the global hypernetwork with parameters φ, $\ell(\cdot): \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}^+$ denote the loss on a single data point, and $\mathcal{L}(\cdot)$ denote the expected empirical loss. Further, let $\boldsymbol{V}$ denote the global embedding matrix whose rows contain the global embedding vectors $\boldsymbol{v}^i$, η denote the learning rate for the clients, and α denote the global hypernetwork learning rate.

TABLE I
Definition of symbols used.

Symbol	Meaning
n	Number of clients
K	Batch size
R	Number of rounds
T	Number of local steps
$[\cdot]^i$	Belonging to client i
$\mathcal{P}^i_{\mathcal{X}\times\mathcal{Y}}$	Data distribution
$\mathcal{D}^i$	Data set
m^i	Number of data points in $\mathcal{D}^i$
$(\boldsymbol{x}^i_k, y^i_k)$	(feature vector, label) of k-th data point
$f(\cdot;\theta^i)$	Classification model f with parameters θ^i
$h(\cdot;\varphi)$	Hypernetwork h with parameters φ
$\Theta = \{\theta^i\}_{i=1}^n$	Set of client classification network parameters
$\ell(\cdot)$	Loss on one data point
$\mathcal{L}(\cdot)$	Expected empirical loss
$\boldsymbol{v}^i$	Global embedding vector
$\boldsymbol{V}$	Global embedding matrix
η	Client learning rate
α	Hypernetwork learning rate
$\boldsymbol{\lambda}$	Lagrangian multipliers
$\boldsymbol{M}$	Linear constraint for fairness
$\boldsymbol{\mu}$	Conditional moments
$\boldsymbol{c}$	Linear constraint for fairness

A. Personalized Federated Learning

In FL, each client i has their own data distribution $\mathcal{P}^i_{\mathcal{X},\mathcal{Y}}$ and has access to m^i IID samples drawn from $\mathcal{P}^i_{\mathcal{X},\mathcal{Y}}$. In PFL, the goal is to train a personalized model for each client in a collaborative way while accounting for data disparities across clients. The general objective function for PFL can be written as follows:

$$\arg\min_{\Theta} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m^i}\sum_{j=1}^{m^i}\ell(f(\boldsymbol{x}^i_j;\theta^i), y^i_j) \tag{1}$$

where Θ denotes the collection of all personal parameters, $\{\theta^1, \ldots, \theta^n\}$.

In [4], the authors proposed pFedHN, which uses a hypernetwork to produce network parameters for each client. A hypernetwork architecture consists of a pair of collaborating neural networks $h: \boldsymbol{\mathcal{V}} \to \Theta$ and $f: \mathcal{X} \to \mathbb{R}$, such that for an input $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}$, h produces the parameters $\theta = h(\boldsymbol{v};\varphi)$ of predictor f, where φ are the parameters of h. The prediction network f takes an input $\boldsymbol{x}$ and returns an output $f(\boldsymbol{x};\theta)$ that depends on both $\boldsymbol{x}$ and the task specific input $\boldsymbol{v}$. In practice, h is typically a large neural network while f is a small [44]. Shamsian et al. modify the general PFL objective function (1) to be:

$$\arg\min_{\varphi,\boldsymbol{v}} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m^i}\sum_{j=1}^{m^i}\ell^i(f(\boldsymbol{x}^i_j; h(\boldsymbol{v^i};\varphi)), y^i_j)) \tag{2}$$

where ℓ^i is the loss function for client i. The descriptor can be an arbitrarily-sized trainable embedding vector, or

can be a fixed vector provided that a good client representation is known a-priori [4]. In FHN, we choose to treat $\boldsymbol{v}^i \in \boldsymbol{V}$ as a trainable embedding vector. By the end of training, the hypernetwork will learn a family of models $\{\theta^i = h(\boldsymbol{v}^i; \varphi) \mid i \in [n]\}$. I.e., h learns the model parameters for each participating client.

B. Demographic Fairness Heterogeneity

There are three main approaches for generating a fair machine learning algorithm: pre-processing, in-processing, and post-processing. In our work, we use in-processing to enforce demographic parity (DP; all groups have equal probability of being assigned to the positive class) and equalized odds (EO; equal false positive and equal false negative rates across groups). In-processing aims to improve the fairness of an algorithm by adding a constraint, or a regularization term, to the existing objective function. Specifically, we use a constrained optimization problem based on the reduction approach presented in [17] to enforce fairness on the local model. We follow their formulation of writing fairness metrics as linear constraints. This formulation allows the hypernetwork to find the hypothesis space that contains the best hypotheses for the clients – specifically in the case where the clients use different fairness metrics such as demographic parity or equalized odds. We discuss this line of reasoning further in Section III-D.

More formally, [17] proposed that demographic parity and equalized odds (among other fairness metrics) can be described as a set of linear constraints of the form:

$$\boldsymbol{M}\boldsymbol{\mu}(h) \leq \boldsymbol{c} \tag{3}$$

where h is the classifier defined as in [17] (i.e., here h is not a hypernetwork), $\boldsymbol{M} \in \mathbb{R}^{|\mathcal{K}|\times|\mathcal{J}|}$ and $\boldsymbol{c} \in \mathbb{R}^{|\mathcal{K}|}$ define the linear constraints, $\boldsymbol{\mu}(h) \in \mathbb{R}^{|\mathcal{J}|}$ is a collection of conditional moments, $|\mathcal{K}|$ is the number of constraints, and $|\mathcal{J}|$ denotes the number of sensitive features being considered. The goal is to learn the most accurate classifier while still satisfying the desired fairness constraints, which equates to solving the constrained optimization problem:

$$\min_{h\in\mathcal{H}} \mathcal{L}(h) \text{ s.t. } \boldsymbol{M}\boldsymbol{\mu}(h) \leq \boldsymbol{c} \tag{4}$$

The authors of [17] additionally proposed to solve the constrained optimization problem by formulating it in the saddle point form:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{|\mathcal{K}|},\, ||\boldsymbol{\lambda}||_1\leq B} \min_{h\in\mathcal{H}} \mathcal{L}(h) + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{M}\boldsymbol{\mu}(h) - \boldsymbol{c}) \tag{5}$$

where $\lambda_j \geq 0$ is the Lagrangian multiplier for the j-th constraint. For statistical and computational reasons, the authors additionally place a bound on the L_1 norm of $\boldsymbol{\lambda}$ in that it must be less than or equal to $B = \frac{1}{eps}$, where eps represents the allowable fairness violation.

In their approach, Agarwal et al. solve (5) through searching for an equilibrium in a zero-sum game. In our work, we take a simpler approach that is more conducive to the federated setting. Since their approach requires computationally expensive ensemble training of the model, we instead preform gradient ascent on $\boldsymbol{\lambda}$ alongside gradient descent on θ^i during the local training phase. In other words, we perform:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{|\mathcal{K}|},\, ||\boldsymbol{\lambda}||_1\leq B} \min_{\Theta} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m^i}\sum_{j=1}^{m^i}(\ell^i(f(\boldsymbol{x}_j^i;\theta^i), y_j^i) + (\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}(\theta^i) - \boldsymbol{c})) \tag{6}$$

For brevity, we refer readers to [17] for an explanation on how $\boldsymbol{M}$, $\boldsymbol{\mu}(\cdot)$, and $\boldsymbol{c}$ are set.

C. Fair Federated Learning via Hypernetworks

We now present our approach to fair personalized federated learning through using hypernetworks (FHN). As mentioned previously, pFedHN uses a unique embedding vector specific to each client. These embedding vectors are stored globally as rows in an embedding matrix which is updated alongside the hypernetwork's parameters, meaning that the global model should learn the correct embedding vectors to produce the best parameters for each client [4]. Using (2) and (6) we formulate the objective function for FHN as follows:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{|\mathcal{K}|},\, ||\boldsymbol{\lambda}||_1\leq B} \min_{\varphi,\boldsymbol{v}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m^i}\sum_{j=1}^{m^i}\left(\ell^i(f(\boldsymbol{x}_j^i; h(\boldsymbol{v}^i;\varphi)), y_j^i)\right) + (\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}(h(\boldsymbol{v}^i;\varphi)) - \boldsymbol{c})\right] \tag{7}$$

We present the pseudocode of our technique in Algorithm 1 and an overview of the FHN architecture in Fig. 1.

Algorithm 1 Fair Hypernetwork (FHN)

Input: Number of rounds R, number of local steps T, batch size K, hypernetwork learning rate α, local learning rate η

1: Randomly initialize all $\boldsymbol{v}^i \in \boldsymbol{V}$
2: **for** $r = 1, \ldots, R$ **do**
3: Uniformly sample client $i \in [n]$
4: Set $\theta^i = h(\boldsymbol{v}^i; \varphi)$ and $\widetilde{\theta}_i = \theta_i$
5: **for** $t = 1, \ldots, T$ **do**
6: Sample mini-batch $\mathcal{B}_K^i \subset \mathcal{D}^i$
7: $\widetilde{\theta}^i \leftarrow \widetilde{\theta}^i - \eta\nabla_{\widetilde{\theta}^i}(\mathcal{L}^i(\mathcal{B}_K^i) + (\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\mu}(\widetilde{\theta}^i) - \boldsymbol{c}))$
8: **end for**
9: $\Delta\theta^i = \widetilde{\theta}^i - \theta^i$
10: $\varphi = \varphi - \alpha(\nabla_\varphi \theta^i)^T \Delta\theta^i$
11: $\boldsymbol{V} = \boldsymbol{V} - \alpha(\nabla_{\boldsymbol{V}}\varphi)^T(\nabla_\varphi\theta^i)^T\Delta\theta^i$
12: **end for**

In Algorithm 1 we begin by randomly initializing the global embedding vector $\boldsymbol{v}^i$ for each client in the federation. Then, in lines 2 through 10 we perform the entire training procedure. In each round R, we first uniformly sample a client i and obtain their parameters by feeding their embedding vector $\boldsymbol{v}^i$ to the hypernetwork (lines 3-4). The parameters are then sent to the client who proceeds to perform T rounds of fairness-enforced training on their local model (lines 5-7). Specifically, in each of the T local steps, the client first selects a mini-batch

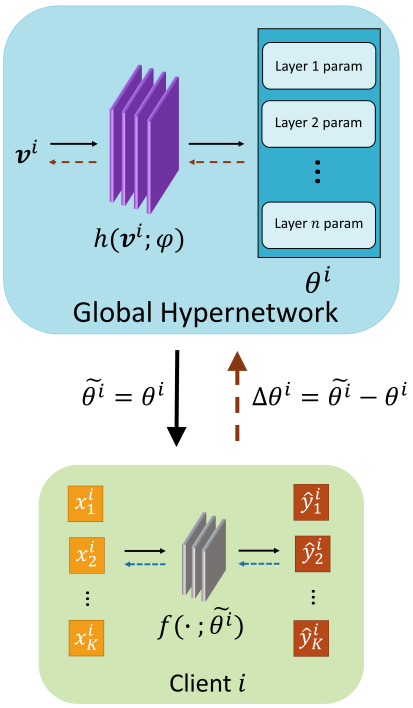


Fig. 1. Our proposed FHN architecture. First, the hypernetwork is used to generate client i's classification model parameters by using the client's embedding vector $\boldsymbol{v}^i$ as input. The generated parameters are then sent to the client who performs T rounds of local training by propagating the loss along the blue dashed lines. The client then passes $\Delta\theta^i = \widetilde{\theta}^i - \theta^i$ to the global server which is then used to update the weights of the hypernetwork and the client embedding matrix (red dashed lines).

of data of size K and proceeds to update the local model parameters using the fairness enforced loss function. After the T rounds of training are complete, the client calculates the difference between the parameters they were sent by the hypernetwork at the beginning of the round (θ^i) and the values of their parameters at the end of local training round T ($\widetilde{\theta}^i$) and sends it to the hypernetwork (line 8). Finally, the hypernetwork updates both its parameters φ and the client embedding matrix $\boldsymbol{V}$ using the obtained update $\Delta\theta^i$ from client i (lines 9-10).

We note that our update procedure for the hypernetwork is written slightly different than originally proposed in [4]. There, instead of updating the entire embedding matrix $\boldsymbol{V}$, the authors depict updating the client's embedding vector only: $\boldsymbol{v}^i = \boldsymbol{v}^i - \alpha(\nabla_{\boldsymbol{v}^i}\varphi)^T(\nabla_\varphi\theta^i)^T\Delta\theta^i$. Despite their pseudocode insinuating that only the client's embedding vector was updated, in their source code (which we used as the base of our experiments) the entire embedding matrix was updated during the backward pass of $\Delta\theta^i$.

D. Theoretical Analysis

In this section, we prove why FHN can accommodate multiple fairness constraints while still aligning with the generalization bounds presented in [4]. To begin, we recall the intuition behind hypernetworks as presented in [45] – namely that of inductive bias learning. In normal probably approximately correct (PAC) learning, a learner is given a hypothesis space $\mathcal{H}$ and a set of training points $\boldsymbol{z} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$ drawn independently according to some underlying distribution $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. Based on the information contained in $\boldsymbol{z}$, the learner's goal is to select a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{H}$ minimizing some measure of expected loss with respect to $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. In our setting, rather than each client learning the correct $h \in \mathcal{H}$, they let the hypernetwork find it based on some conditioning input. In other words, the hypernetwork learns a hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$, that minimizes the expected loss for each client. Specifically, the hypernetwork is given a family of hypothesis spaces $\mathbb{H} = \{\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3, \ldots\}$, and the goal is to find a bias (i.e., hypothesis space $\mathcal{H} \in \mathbb{H}$) that is appropriate for the entire group of clients.

All of the hypotheses in $\mathcal{H}$ have a degree of complexity associated with them. The degree of complexity can be measured in the number of features or the polynomial degree of the learning function. Given a hypothesis space $\mathcal{H}$, consider all functions in $\mathcal{H}$ with complexity at most r:

$$\mathcal{H}_r = \{h \in \mathcal{H} \mid \Omega(h) \leq r\} \tag{8}$$

In order to enforce fairness, we formulate the fairness metric as a linear constraint on the loss function. If all of the clients have the same base network and loss function, then each client adding a linear constraint (no matter if it is the linear constraint for equalized odds, demographic parity, or any other fairness metric) does not change the overall degree of complexity. Therefore clients enforcing different fairness constraints does not change the hypernetwork's ability to find a hypothesis space $\mathcal{H}$ that suits all of the clients, and the generalizability proof (Theorem 1) given in [4] holds as written.

IV. Evaluation

We now assess the empirical performance of FHN against several baselines on the COMPAS and Adult datasets. Specifically, we analyze the ability of FHN to learn accurate and fair personalized models for all clients using either demographic parity or equalized odds. We implement all experiments in Pytorch and execute them on a Tesla V100-SXM2 32GB GPU. The code used for the experimentation is publicly available at: https://bit.ly/3yVZ5rZ.

A. Datasets

We utilize two popular fair machine learning datasets in our experimentation: the Adult [46] and the COMPAS [47] datasets. The Adult dataset is comprised of 48,842 instances and the task is to predict whether someone's income exceeds $\$50,000$ a year based on collected U.S. census data. The COMPAS dataset is comprised of 7,214 instances and the task is to predict whether a criminal defendant will reoffend in the two years after the original offense. To simulate covariate shift, we split the Adult dataset among the clients based on type of employment (government or private sector) which varies the marginal distribution $\mathcal{P}_{\mathcal{X}}^i$ across clients. For the COMPAS dataset, we split the data among clients based on the listed age

($\leq 31, > 31$). For both cases, we treat gender as the binary sensitive attribute.[1] Additionally, after dividing the data among the clients, we use a split of $85/15$ training/testing for the COMPAS dataset and use the default split given for the Adult dataset. We also note that in this setting each client within the same split partition (e.g., all clients who have data with ages ≤ 31) has approximately the same amount of training data and testing data, but clients in different split partitions do not necessarily have the same amount. After cleaning (following the procedures taken by the original COMPAS analysis [47] and general good data practices such as dropping rows with 'N/A' values), a total of 6,172 and 40,807 data points remain for the COMPAS and the Adult datasets, respectively.

B. Baselines

We compare FHN to two fair FL algorithms as well as against a decentralized fair learning approach (DFL) where each client only trains their model locally.

Fair Federated Learning via FedAvg (FFLvFA) was first proposed in [15] as a baseline for testing fair federated learning architectures. For our FFLvFA baseline, we borrow this idea and alter the implementation of FedAvg [1] to allow clients to enforce fairness on their local models. The idea behind FFLvFA is that if each client runs a fair learning algorithm on their own data, and the locally trained models are aggregated via FedAvg, then one might hope to obtain a model that is accurate and fair on the overall data distribution [15]. Specifically, we implement the same linear constraint on the local loss function as we do for FHN. Each client sends their parameter updates to the global model as usual, and the global model does not do any kind of fairness enforcement while aggregating and updating the global parameters.

FairFed [16] is a slight adaption of the FedAvg architecture that adaptively modifies the aggregation weights for the clients at the server each round. Specifically, the aggregation weights are formulated based on the mismatch between the global fairness value (which is an aggregation of client calculated values) and the local fairness value at the client, favoring clients whose local value match closely to the global fairness value. The aggregation weights are calculated as:

$$\bar{w}_k^t = \exp(-\beta \mid F_k^t - F_{\text{global}}^t \mid) \cdot \frac{n_k}{\sum_{i=1}^K n_k}, \forall k \in \{1, \ldots, K\} \quad (9)$$

where $\bar{w}_k^t$ is the k^{th} client's aggregation weight at round t, F_k^t is client k's fairness value at round t, F_{global}^t is the global fairness value at round t, n_k is number of data points from client k, and β controls the trade-off between accuracy and fairness.

Decentralized Fair Learning (DFL): each client only trains on their data to produce a local model. No collaboration between clients takes place, meaning that no global model is produced. This gives perfect privacy at the expense of not learning from other clients' data.

[1]We acknowledge that gender is a highly diverse social category and reducing it to a binary classification is an oversimplification. We do so only to facilitate the analysis of FHN.

C. Metrics

We compare the baselines with our FHN architecture along accuracy, statistical parity difference (SPD), and equalized odds difference (EOD). SPD is measured as the absolute value of the difference between the positive prediction rate (PPR) of the minority class and the PPR of the majority class. Since we consider gender as our sensitive attribute, we compute SPD as:

$$\text{SPD} = |\text{PPR}_f - \text{PPR}_m| \quad (10)$$

where f/m represent female/male, respectively. EOD is measured as the absolute value of the difference between the false negative rate (FNR) of the minority group and the FNR of the majority group, plus the difference between the true positive rate (TPR) of the minority group and the TPR of the majority group, all divided by 2. Specifically, we compute:

$$\text{EOD} = \frac{|(\text{FNR}_f - \text{FNR}_m) + (\text{TPR}_f - \text{TPR}_m)|}{2} \quad (11)$$

D. Architecture and Hyperparameters

For FHN, the hypernetwork is made of 5 hidden layers with 100 neurons each using ReLU activations and binary cross entropy loss. We detail our hyperparameter search space and selected values in Table III. Specifically, we select the hyperparameter combination that produces the best average accuracy across the clients. For FairFed we follow the implementation details in the original paper [16], including the hyperparameter settings. For DFL and FFLvFA we use the same settings as FHN. Additionally, we set the constraints (c and eps in (6)) for demographic parity to be .1 for the COMPAS experiments and .01 for the Adult while the constraints for equalized odds are set to .01 for both datasets.

E. FHN Under No Fairness Constraint

For a control, we test how well FHN performs when no fairness is enforced. This is to give a baseline of the accuracy and fairness values before the fairness constraint is added. The results of this experiment can be seen in Table II (rows marked '–'). In the Adult experiment, FHN achieved accuracy on par with FFLvFA and DFL while in the COMPAS experiment it obtained the highest accuracy. In this experiment, the values for EOD and SPD are not relevant and were only reported for comparison with the later experiments where fairness is enforced.

F. FHN Under Fairness Homogeneity

In this experiment, we test how well FHN performs when all clients enforce the same fairness metric. We do this in order to directly compare to other fair federated learning implementations such as FFLvFA and FairFed. We split the dataset among four different clients in the manner described previously, and we note that all clients had training examples for each possible label. We test both our FHN and the baselines using demographic parity and equalized odds and the results can be seen in Table II (rows marked 'DP' and 'EO').

TABLE II

EXPERIMENTAL RESULTS ON THE COMPAS AND ADULT DATASETS WITH FOUR CLIENTS SPLIT BY AGE (COMPAS) AND TYPE OF EMPLOYMENT (ADULT). SENSITIVE ATTRIBUTE IS GENDER FOR BOTH DATASETS. EOD: EQUALIZED ODDS DIFFERENCE, SPD: STATISTICAL PARITY DIFFERENCE, –: NO FAIRNESS, DP: DEMOGRAPHIC PARITY, EO: EQUALIZED ODDS. GRAY HIGHLIGHT INDICATES A RESULT IS NOT RELEVANT TO THE TEST PERFORMED. BLUE HIGHLIGHT INDICATES RESULTS THAT DIRECTLY SUPPORT OUR HYPOTHESIS OF FHN UNDER FAIRNESS HETEROGENEITY. HIGHER VALUES ARE BETTER FOR ACCURACY AND THE LOWER THE BETTER FOR THE EOD AND SPD MEASURES. TOTAL TIME REPORTED IS FOR THE ENTIRE TRAINING PROCEDURE, NOT PER CLIENT.

Dataset	Fairness	Arch	Accuracy					EOD					SPD					Time
			Avg	Client 1	Client 2	Client 3	Client 4	Abs Avg	Client 1	Client 2	Client 3	Client 4	Abs Avg	Client 1	Client 2	Client 3	Client 4	
Adult	–	DFL	0.7997	0.7848	0.7755	0.8185	0.8198	0.1564	0.2535	0.0573	0.1612	0.1536	0.2040	0.3194	0.1704	0.165	0.161	47 sec
		FFLvFA	0.8089	0.7577	0.7446	0.8202	0.8183	0.1882	0.3208	0.2606	0.0753	0.0959	0.2695	0.4325	0.4026	0.1225	0.1203	18 min 50 sec
		FairFed	0.7227	0.681	0.6763	0.722	0.7392	0.0149	0.0034	0.0192	0.0214	0.0157	0.0292	0.0085	0.044	0.0297	0.0346	2 sec
		FHN	0.7822	0.7446	0.7624	0.7876	0.7871	0.1271	0.2023	0.08	0.1	0.1262	0.1596	0.2225	0.1683	0.1145	0.1332	10 min 11 sec
	DP	DFL	0.7850	0.7746	0.7334	0.8088	0.823	0.1228	0.0221	0.222	0.1335	0.1134	0.0448	0.093	0.0626	0.0001	0.0234	1 min 26 sec
		FFLvFA	0.8157	0.7624	0.7558	0.8274	0.8242	0.1258	0.0812	0.1645	0.132	0.1255	0.0307	0.0813	0.0078	0.0167	0.017	35 min 11 sec
		FairFed	0.7421	0.7109	0.6969	0.7475	0.7503	0.0233	0.0079	0.0142	0.0258	0.0452	0.0651	0.0664	0.0403	0.0725	0.0813	7 sec
		FHN	0.7806	0.7456	0.7661	0.7794	0.7906	0.0596	0.0012	0.046	0.0879	0.1034	0.0416	0.0739	0.0762	0.0107	0.0055	18 min
	EO	DFL	0.7928	0.7839	0.753	0.8217	0.8124	0.0420	0.0758	0.0186	0.0575	0.0159	0.0950	0.1682	0.0826	0.068	0.061	1 min 59 sec
		FFLvFA	0.8117	0.7792	0.7774	0.82	0.8153	0.0366	0.04	0.0624	0.012	0.0319	0.1069	0.1295	0.0636	0.1181	0.1164	48 min 16 sec
		FairFed	0.7415	0.7081	0.6978	0.748	0.7488	0.0271	0.0114	0.0192	0.0279	0.0499	0.0640	0.0655	0.0383	0.0724	0.0799	6 sec
		FHN	0.7779	0.7437	0.7586	0.7836	0.7822	0.0407	0.0918	0.0257	0.0239	0.0215	0.0970	0.1689	0.1188	0.0521	0.0481	24 min 26 sec
	DP & EO	DFL	0.7872	0.7746	0.753	0.8088	0.8124	0.0173	0.0221	0.0186	0.1335	0.0159	0.0466	0.093	0.0826	0.0001	0.061	1 min 39 sec
		FFLvFA	0.8267	0.7858	0.7928	0.8362	0.8307	0.0751	0.0257	0.1303	0.012	0.0198	0.1039	0.1154	0.0477	0.0923	0.099	41 min 7 sec
		FairFed	0.7406	0.7119	0.695	0.7453	0.7493	0.0314	0.0017	0.019	0.0229	0.0437	0.0671	0.0627	0.0421	0.0715	0.0813	6 sec
		FHN	0.7917	0.7362	0.8064	0.8031	0.7877	0.0222	0.0144	0.0238	0.1344	0.0205	0.0403	0.0593	0.141	0.0212	0.0547	21 min 13 sec
Compas	–	DFL	0.6701	0.657	0.657	0.6982	0.6682	0.2742	0.6478	0.1074	0.3118	0.0296	0.2882	0.6782	0.1293	0.3141	0.0312	27 sec
		FFLvFA	0.6512	0.5785	0.6405	0.6757	0.7182	0.2120	0.215	0.1369	0.3131	0.1828	0.2074	0.2151	0.1543	0.3199	0.1404	9 min 12 sec
		FairFed	0.6544	0.6488	0.6033	0.6937	0.6773	0.1700	0.1949	0.0409	0.2305	0.2136	0.1760	0.2459	0.055	0.2267	0.1762	2 sec
		FHN	0.6749	0.6446	0.6653	0.6892	0.7045	0.2506	0.4466	0.131	0.2389	0.1858	0.2570	0.4813	0.1531	0.2448	0.1487	6 min 4 sec
	DP	DFL	0.5668	0.4917	0.5083	0.5946	0.6727	0.0515	0.0517	0.085	0.0474	0.0217	0.0553	0.0615	0.0905	0.0462	0.023	1 min 17 sec
		FFLvFA	0.4806	0.642	0.5864	0.5309	0.3636	0.1158	0.3905	0.0727	0	0	0.1055	0.3615	0.0606	0	0	25 min 25 sec
		FairFed	0.6501	0.6529	0.595	0.6892	0.6682	0.0808	0.2205	0.0126	0.226	0.2212	0.1664	0.2716	0.0004	0.2264	0.1672	3 sec
		FHN	0.6771	0.6777	0.657	0.6802	0.6955	0.0808	0.0323	0.03	0.1304	0.1303	0.0674	0.0248	0.0084	0.1407	0.0957	13 min 53 sec
	EO	DFL	0.6256	0.5041	0.6364	0.6802	0.6818	0.0433	0.0254	0.1089	0.0388	0	0.0557	0.0403	0.0908	0.0917	0	1 min 26 sec
		FFLvFA	0.6609	0.624	0.6488	0.6712	0.7045	0.1448	0.1757	0.1301	0.1515	0.1219	0.1285	0.1477	0.1096	0.1723	0.0842	36 min 46 sec
		FairFed	0.6523	0.6198	0.6157	0.6937	0.6864	0.1939	0.2624	0.0887	0.2194	0.2051	0.1943	0.301	0.1047	0.2067	0.1647	3 sec
		FHN	0.6749	0.657	0.657	0.6667	0.7227	0.1087	0.0916	0.0643	0.1233	0.1557	0.0889	0.0495	0.043	0.1489	0.1142	19 min 19 sec
	DP & EO	DFL	0.6011	0.4917	0.6364	0.5946	0.6818	0.0545	0.0517	0.1089	0.0474	0	0.0539	0.0615	0.0908	0.0462	0	1 min 16 sec
		FFLvFA	0.6458	0.6157	0.5909	0.6892	0.6955	0.2559	0.4823	0.4932	0.201	0.0185	0.3009	0.4513	0.4791	0.1504	0.0237	30 min 49 sec
		FairFed	0.6544	0.6281	0.6157	0.6937	0.6864	0.1469	0.2604	0.0887	0.2194	0.2051	0.2539	0.301	0.1047	0.2067	0.1647	3 sec
		FHN	0.6663	0.6488	0.6364	0.6712	0.7139	0.0707	0.1108	0.0039	0.1522	0.1375	0.1170	0.07	0.023	0.1639	0.1027	16 min 18 sec

TABLE III

GRID SEARCH SPACE FOR THE HYPERPARAMETERS. BOTH THE CLIENT AND HYPERNETWORK OPTIMIZERS ARE SET TO THE SAME WEIGHT DECAY VALUE. K: BATCH SIZE, η: CLIENT LEARNING RATE, α: HYPERNETWORK LEARNING RATE, wd: WEIGHT DECAY.

Symbol	Values	COMPAS	Adult
K	{24, 64, 128, 256}	64	256
η	{1e-5, 3e-5, 5e-5,. . . , 1e-2, 3e-2, 5e-2}	5e-2	1e-3
α	{1e-5, 3e-5, 5e-5, . . . , 1e-2, 3e-2, 5e-2}	5e-5	1e-5
wd	{1e-6, 1e-8, 1e-10}	1e-10	1e-10

For the Adult dataset DFL, FFLvFA, and FHN achieved comparable accuracy and fairness values with FFLvFA performing the best out of the three. In the EO test, FairFed achieved the lowest EOD with a value of .0271. This was at a cost to the accuracy which was the lowest of all the metrics. For the COMPAS dataset on both DP and EO tests, FHN achieved the best accuracy and the second best fairness. DFL achieved the best values for EOD and SPD, but with much lower accuracy than FHN. Additionally, on both the DP and EO tests, there were clients that achieved an EOD and SPD value of zero. We note that this does not actually denote perfect fairness. It is known that demographic parity trades false negatives for false positives [48]. In these experiments, the clients had zero true positive values which caused the calculations of SPD (10) as well as EOD (11) to go to zero. Finally, it may seem odd that in the tests on the Adults dataset the accuracy of FairFed actually increased as a fairness constraint was added to the loss function when normally doing so causes the accuracy to degrade. We believe this is due to FairFed's optimization procedure being constructed specifically to enforce fairness. In the case where no fairness is enforced, FairFed reduces to FedAvg and their updated client weighing procedure is not used.

G. FHN Under Fairness Heterogeneity

In this experiment we formally test our hypothesis that FHN is robust against fairness heterogeneity. We will once again test against the baselines of DFL and FFLvFA, although

TABLE IV
ACCURACY, EOD, AND SPD WITH STANDARD DEVIATION OVER FIVE ROUNDS FOR AN INCREASING NUMBER OF CLIENTS.

# Clients	Accuracy	EOD	SPD
10	.780±.005	.010±.013	.020±.017
20	.774±.005	.013±.010	.019±.022
30	.772±.003	.023±.011	.032±.019
40	.770±.005	.059±.017	.034±.021
50	.773±.004	.037±.022	.021±.028
60	.770±.004	.040±.017	.037±.018
70	.766±.004	.063±.051	.027±.024
80	.772±.002	.073±.026	.020±.017
90	.769±.004	.037±.027	.047±.016
100	.774±.002	.046±.021	.045±.020

FFLvFA was designed with all clients enforcing the same fairness metric during training. Additionally, we will further test against FairFed which is closest in spirit to our design. The authors of FairFed proposed that each client can use different local pre-processing debiasing strategies, but they had no formal experimentation to provide proof of their hypothesis. The results of this experiment can also be seen in Table II (rows marked 'DP & EO').

In both the Adult and the COMPAS experiments, FHN performed best overall and achieved the highest accuracy while the values of EOD and SPD were close to zero compared to the other methods. In the Adult experiment, DFL achieved better EOD values and comparable SPD values, which is expected since the clients have no interaction during training. Also as expected, FFLvFA performed the worst in this experimental setting as no attention is payed to fairness during parameter aggregation which results in the global model not adhering to the clients wanted fairness constraints. This experiment shows that FHN is able to produce accurate network parameters that align with each clients' wanted fairness constraint with only a slight degradation to, or in some cases (e.g., the Adult experiment) an increase to, the accuracy of the model.

H. FHN Under Increasing Federation Size

In the previous three experiments we used a federation with four clients. In this experiment, we show how adding more clients does not degrade the ability of FHN to find parameters that align with the clients' chosen fairness constraints. To do so, we randomly sample with replacement the Adult dataset to assign data points to the clients. Each client's dataset varied in size with a minimum of 4,000 data points (10% of the training data) and a maximum of 13,378 (50% of the training data) for the training set and 1,400 and 7,025 points for testing. We note that in this experiment we no longer adhere to splitting the data among the clients based on the type of employment (i.e., no intentional covariate shift). We test a range of 10 to 100 clients to see how both the accuracy and the fairness changes with the addition of more clients. Half of the clients in our test use equalized odds while the remaining use demographic parity. We ran each setting five times and reported the average of the obtained values. The results for this experiment can be seen in Table IV and Fig. 2.

Table IV shows that the values for the accuracy, EOD, SPD, and standard deviation held relatively steady no matter the size of the federation. The accuracy always stayed over 76%, EOD always stayed under .099 away from zero, and SPD stayed under .063 away from zero. The most noticeable trend is the slight increase in EOD and SPD values as the number of clients increased. Specifically, the average EOD value for experiments with 30 or fewer clients was .015 while the average EOD value for experiments with 40 or more clients was .051. A similar trend occurs with the SPD values when the client size is 90 or more. However, these increases are small and are still acceptable in terms of the fairness provided. Fig. 2 reinforces the trend of accuracy, EOD, and SPD maintaining approximately the same values despite the growth in federation size.

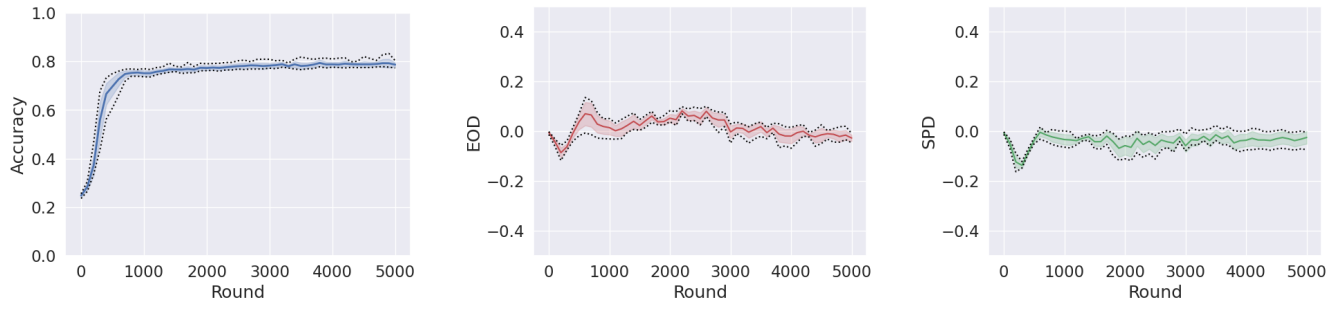
I. Generalization to Novel Clients

In this experiment, we empirically support our theoretical analysis in Section III-D. We follow the experimental design of [4] and use a federation size of 100 clients. 90 of the clients participate in training while the remaining 10 are held out to be introduced as novel clients. Since we perform binary classification, we draw data samples according to the Beta distribution rather than from the Dirichlet distribution as originally performed in [4]. Specifically, we set the percentage of data points with label zero as $x \sim \text{Beta}(.5, .5)$ and the percentage of points with label one as $1-x$ for each training client. Each training client is randomly assigned an amount of data points from 100 to 4,000 which are randomly sampled from the training set according to the probabilities x and $1-x$. The same sampling procedure is performed for the novel clients, but we vary the Beta distribution as Beta(ρ, ρ) where $\rho \in \{.1, .5, 1, 2, 3\}$. A similar process is performed for the test data. Like [4], we report the distance between a novel client's distribution and its nearest neighbor's distribution from the training set as Total Variation (TV):

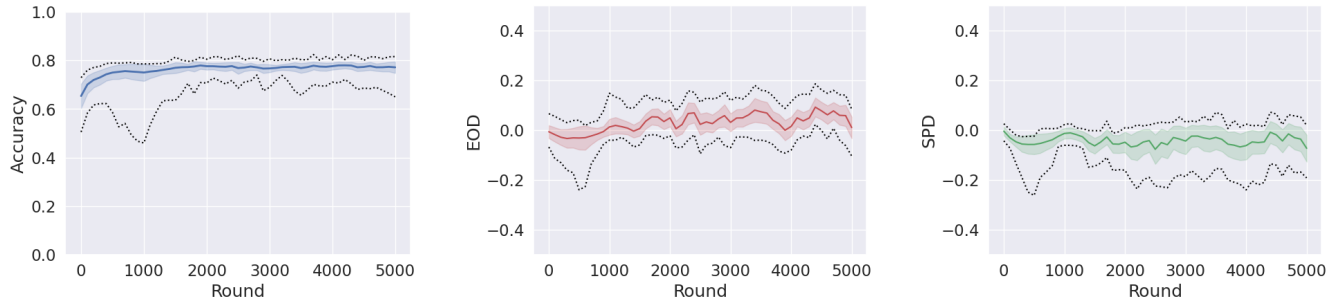
$$TV = \frac{1}{2}||P - Q||_1 \tag{12}$$

where P is the novel client's distribution and Q is the training client's distribution. We note that in this experiment, the local training was reduced to 10 rounds for both training and novel clients and only a total of 500 rounds of training was completed for the novel clients.

Fig. 3 presents the generalization gap for the accuracy, EOD, and SPD values ($\text{metric}_{novel} - \text{metric}_{train}$) between the training and novel clients. If the bar is above the x-axis then the novel clients have a higher average value (which is favorable for accuracy but not for EOD and SPD). Overall there is no consistent trend as TV increases. However, the gap between the novel and training clients for accuracy was always below $\pm.0504$ and for EOD/SPD below $\pm.04$. Therefore FHN under fairness heterogeneity can generalize well to new clients, and additionally, novel clients achieve their desired fairness metric at almost the same degree as clients used in training.



(a) 10 Clients: Accuracy (b) 10 Clients: EOD (c) 10 Clients: SPD



(d) 100 Clients: Accuracy (e) 100 Clients: EOD (f) 100 clients: SPD

Fig. 2. Accuracy, EOD, and SPD across all epochs for the 10 (row 1) and 100 (row 2) client settings. The solid line depicts the mean value of all clients, the color band shows the standard deviation, while the dotted line shows the maximum and minimum value obtained by any client in the federation for each round.

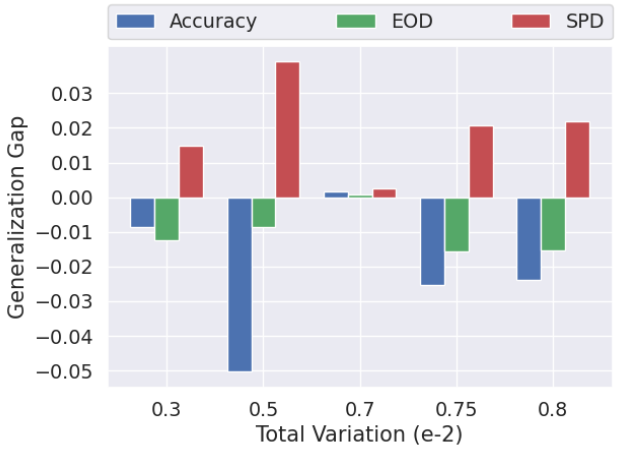


Fig. 3. Generalization gap for accuracy, EOD, and SPD between training clients and novel clients. The generalization gap is reported as metric$_{novel}$ – metric$_{train}$. Bars above the x-axis indicate the novel clients achieved a higher value.

J. Execution Time

The total time for FairFed on all single fairness tests (Table II rows marked 'DP' and 'EO') was much lower than all of the other baselines. This is due to using the same hyperparameters FairFed reported in the original paper. The authors report only using 20 total rounds where each client only trained locally for one epoch. This setting explains the lower accuracy values for all FairFed tests since the other methods were trained for 5,000 rounds with 50 epochs per local training iteration. For the scalability test (Section IV-H and Table 2), the total time stayed at a constant 22 minutes per total training time no matter how many clients were included in the federation. This is due to holding the amount of total global rounds at a constant 5,000, the amount of the inner rounds at 50, and continuing to select one client per global round.

V. Conclusion

In this work, we proposed *Fair Hypernetworks* (FHN), a personalized federated learning architecture based on hypernetworks that is robust to statistical and fairness heterogeneity. We theoretically showed that when the fairness metric is formalized as a linear constraint on the local optimization function, the hypernetwork is still able to generalize and find accurate network parameters – even when clients have heterogeneous data and/or fairness constraints. Additionally, we empirically showed that FHN is able outperform other baselines in under fairness heterogeneity, even when clients enforce different fairness metrics such as demographic parity and equalized odds. Further, we empirically showed that FHN is robust against federation size as well as empirically validated our theoretical analysis. While we specifically consider linear constraints, and use this assumption as part of our

theoretical analysis, we realize that not all fairness metrics can be formalized as such (e.g., calibration and predictive parity). But, we believe that FHN will still be robust against fairness heterogeneity when other fairness methods such as pre-processing or post-processing are used – which we leave as future work.

ACKNOWLEDGEMENTS

This work was supported in part by NSF 1910284, 1946391, 2137335, and 2147375.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[3] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[4] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, "Personalized federated learning using hypernetworks," *arXiv preprint arXiv:2103.04628*, 2021.

[5] V. Kulkarni, M. Kulkarni, and A. Pant, "Survey of personalization techniques for federated learning," in *WorldS4'20*. IEEE, 2020, pp. 794–797.

[6] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[7] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *ICML*. PMLR, 2021, pp. 6357–6368.

[8] X. Yue, M. Nouiehed, and R. A. Kontar, "Gifair-fl: An approach for group and individual fairness in federated learning," *arXiv preprint arXiv:2108.02741*, 2021.

[9] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[10] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *ICML*. PMLR, 2019, pp. 4615–4625.

[11] D. Y. Zhang, Z. Kou, and D. Wang, "Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models," in *IEEE Big Data'20*, 2020, pp. 1051–1060.

[12] B. Rodríguez-Gálvez, F. Granqvist, R. van Dalen, and M. Seigel, "Enforcing fairness in private federated learning via the modified method of differential multipliers," in *NeurIPS'21 Workshop Privacy in Machine Learning*, 2021.

[13] L. Chu, L. Wang, Y. Dong, J. Pei, Z. Zhou, and Y. Zhang, "Fedfair: Training fair models in cross-silo federated learning," *arXiv preprint arXiv:2109.05662*, 2021.

[14] W. Du, D. Xu, X. Wu, and H. Tong, "Fairness-aware agnostic federated learning," in *SDM'21*. SIAM, 2021, pp. 181–189.

[15] Y. Zeng, H. Chen, and K. Lee, "Improving fairness via federated learning," *arXiv preprint arXiv:2110.15545*, 2021.

[16] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr, "Fairfed: Enabling group fairness in federated learning," *arXiv preprint arXiv:2110.00857*, 2021.

[17] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *ICML*. PMLR, 2018, pp. 60–69.

[18] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[19] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *Neurocomputing*, vol. 465, pp. 371–390, 2021.

[20] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-iid data with reinforcement learning," in *IEEE INFOCOM 2020*. IEEE, 2020, pp. 1698–1707.

[21] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-iid data quagmire of decentralized machine learning," in *ICML*. PMLR, 2020, pp. 4387–4398.

[22] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," *arXiv preprint arXiv:2102.02079*, 2021.

[23] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[24] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM TIST*, vol. 10, no. 2, pp. 1–19, 2019.

[25] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ITCS'12*, 2012, pp. 214–226.

[26] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[27] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *WWW'26*, 2017, pp. 1171–1180.

[28] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[29] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[30] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, feb 2022. [Online]. Available: https://doi.org/10.1145/3494672

[31] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[32] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 962–970.

[34] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, "Decoupled classifiers for group-fair and efficient machine learning," in *FAT'18*. PMLR, 2018, pp. 119–133.

[35] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *FAT'18*. PMLR, 2018, pp. 107–118.

[36] A. Papadaki, N. Martinez, M. Bertran, G. Sapiro, and M. Rodrigues, "Minimax demographic group fairness in federated learning," *arXiv preprint arXiv:2201.08304*, 2022.

[37] S. Hu, Z. S. Wu, and V. Smith, "Provably fair federated learning via bounded group loss," *arXiv preprint arXiv:2203.10190*, 2022.

[38] F. Zhang, K. Kuang, Y. Liu, C. Wu, F. Wu, J. Lu, Y. Shao, and J. Xiao, "Unified group fairness on federated learning," *arXiv preprint arXiv:2111.04986*, 2021.

[39] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized Cross-Silo Federated Learning on Non-IID Data," *AAAI'21*, vol. 35, no. 9, pp. 7865–7873, May 2021, number: 9.

[40] K. Wang, R. Mathews, C. Kiddon, H. Eichner, F. Beaufays, and D. Ramage, "Federated evaluation of on-device personalization," *arXiv preprint arXiv:1910.10252*, 2019.

[41] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[42] J. Schmidhuber, "Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks," *Neural Computation*, vol. 4, no. 1, pp. 131–139, 01 1992.

[43] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016.

[44] T. Galanti and L. Wolf, "On the modularity of hypernetworks," in *NIPS'20*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[45] J. Baxter, "A model of inductive bias learning," *JAIR*, vol. 12, pp. 149–198, mar 2000. [Online]. Available: https://doi.org/10.1613%2Fjair.731

[46] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[47] ProPublica, "Compas recidivism risk score data and analysis," https://propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis, 2021.

[48] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. Online: fairmlbook.org, 2019.