RESEARCH INTERNSHIP REPORT

Fully decentralized learning under fairness and heterogeneity constraints

20 Mars 2023 - 18 Aout 2023

Ismail Labiad

ÉCOLE POLYTECHNIQUE

INSTITUT POLYTECHNIQUE DE PARIS

Inria

TABLE DES MATIÈRES

1 INTRODUCTION

Machine learning algorithms have gained massive attraction in different industries in the past few years, ranging from simple classifiers that differentiate between cat and dog images to algorithms used by big companies for automating CVs screenings and banks for risk management.

And in recent years it came into question the reliability and safety of these machine-learning algorithms, that affects our daily life. These machine-learning models are only as good as their training data allows them to be, and it becomes important when the human factor is involved to not only take into account the accuracy of the models but also their explainability, fairness, privacy, and robustness, these research areas are referred to as **Trustworthy ML**. In this report, we will focus on the joint study of fairness and decentralized learning.

Fairness in machine learning refers to the techniques and frameworks that try to mitigate bias in machine learning models, as the trained models can be discriminatory with respect to a sensitive attribute such as : race, gender, age, etc. And this line of research has gained a lot of attention from researchers, especially since the publication of the ProPublica report on the COMPAS software [18] [20] where the report shows that the software used nationwide in the US to assess the likelihood of an arrest to committing a crime is racially biased against African Americans even when the software didn't include race as one of its parameters. Machine learning models can exhibit bias through their training data, for example, OpenAI addresses these questions in GPT-4 to not perpetuate stereotypes [17] as the earlier versions of the Chatbot provided biased responses.

Approaches to mitigate bias and increase fairness can be divided into three main categories : pre-processing, in-processing, and post-processing. Pre-processing approaches act on the data level before the training of the algorithm to prevent bias from reaching the models. In-processing, incorporate fairness while training the models. And post-processing methods that act on already trained models to ensure the fairness of the new output.

Another aspect of machine learning safety is the sovereignty of the user's data. Machine learning in its classical form is called centralized, meaning all the data is accessible by the agent training the model. But in a variety of scenarios sharing and centralizing the data is simply not possible, either because it is illegal, for example, patients' data in hospitals. Or because the clients are not willing to share their data for competitive reasons, which may be the case for banks, and advertising agencies that may not want to reveal such information. There are two approaches to handling this problem : Federated Learning and **Decentralized Learning**. Federated learning is when there is a central server trusted by the clients that handle the training, and the communication topology in this case is a star graph (all clients are only connected to a trusted central server). Decentralized learning on the other hand is peer-to-peer, clients communicate only with their neighbors through a communication topology.

The joint study of group fairness and federated learning has seen some interest lately [9], [22], [7]. In [9], the authors propose FairFed, a fair version of FedAVG that relies on computing new averaging weight through SecAgg protocol. In [22], the authors provide FairFB a federated version of FairBatch, allowing for the acceptance of more than only two sensitive groups. And in [7] the authors develop FedFair which relies on solving a min-max problem through the AGP algorithm.

But, to the best of our knowledge, there is no work that tackles group fairness which will be the main focus of our work.

Motivation Most recent approaches that try to mitigate bias [ref some papers] rely on solving a constrained version of the original optimization problem by using a surrogate loss. The problem is that there is no way of ensuring fairness through minimizing the surrogate, as the authors demonstrate in [Too relaxed to be fair]. Moreover, fairness has received far less attention in the decentralized setting compared to the work done in the centralized case [ref some papers] or in the federated learning case [ref some papers]. Our goal is then to propose a fully decentralized approach to training fair models.

Contributions In this paper, we propose D-SearchFair, a fully decentralized approach to learning ε-fair models. Our approach combines the work done in [15], [12], [3]. We show that in a perfect scenario, the decentralized

algorithm works exactly as its centralized version, therefore outputting an exact fair model.

- We also derive a theroticial bound on the fairness level for the practical implementation where each client has only access to an empirical dataset and where we work 5.5 work.

We also propose a more cost-efficient version of the algorithm D-SearchFair with averaging, but that lacks a theoretical guarantee for its utility.

We show the utility of both approaches through experiments on synthetic datasets and real-world datasets with linear and deep models.

2 RELATED WORK

There exists a wide range of approaches that looks to incorporate fairness at some point in the training process of the model, we will refer the reader to the following surveys [10] [5] for an extensive list. We will present here some of the recent in-processing approaches. [19] proposes FairBatch a batch gradient approach that relies on oversampling disadvantaged groups to ensure fairness during training. [16] propose FairGrad that relies on computing new loss weights for each sensitive group at each iteration, and shows that sometimes negative weight can be necessary in order to obtain fair models. Both approaches offer a simple way to incorporate fairness during training, either by changing the sampling weights or the loss weights, but ... In [15], the authors show that minimizing the surrogate does not necessarily minimize the true fairness, and propose a new approach SearchFair that guarantees exact fairness in case we have access to the true distribution. SearchFair will be detailed more in section 3.2.

Comprehensive work has already been done on the convergence of D-SGD, [12] provides the upper bound on the convergence speed of the algorithm in the non-convex, convex, and strictly convex cases. Their framework allows the use of different communication topologies at each iteration, as long as we observe a decrease toward a consensus after a fixed number of iterations. In our work, we simply consider a fixed communication Topology, but we believe that it can be extended to cover the same range of topology distributions as in [12]. Moreover, the work of [13] introduces a new *neighborhood heterogeneity* concept that allows obtaining faster convergence speed when the correct communication topology is picked to minimize the distribution heterogeneity, and they also provide an algorithm to compute such a topology. We provide further technical details on D-SGD in section 3.3.

There have been some papers on fair decentralized models [6], but they don't consider group fairness and focus on other fairness notions.

3 BACKGROUND

We consider the fully decentralized learning setting where K clients collaborate (without directly sharing data) to solve a consensus problem. Precisely, the clients want to obtain a single "fair" model that is useful for the clients, in other words :

$$\arg\min_{\theta} \frac{1}{K} \sum_{i=1}^{K} l_i(\theta) \quad \text{s.t.} \quad |\hat{f}(\theta)| \leq \epsilon$$

Where $\hat{f}$ is an empirical fairness metric, that we will define in the following section, and l_i is the objective function to minimize (the loss) of each client.

Each client i has an unknown distribution D_i of data. We denote by $X \in \mathbb{R}^d$ the feature vector, $S \in \{-1, 1\}$ the sensitive attribute, $Y \in \{-1, 1\}$ the target (label of classification) and we write Z = (X, S, Y) $\sim D_i$.

In the rest of the report, we denote by $\hat{D}_i$ the known empirical dataset of client i and by $\hat{D}$ the concatenation of all the clients datasets. We also define n_i : the number of data points in $\hat{D}_i$ and $n = \sum_{i=1}^{K} n_i$. Note that the distributions D_i can be very different in both the features X and the sensitive attributes S.

We denote by H_θ a classifier parametrized with $\theta \in \mathbb{R}^d$.

3.1 Fairness Metrics

In this report, we adopt the following two fairness metrics :

Demographic Parity : Demographic Parity requires that the proportions of positively classified data be equal across the sensitive groups.
Formally, for a binary classifier H_θ is fair for demographic parity when its predictions are independent of the sensitive attribute (Calders et al., 2009 ; Calders Verwer, 2010). This can be written as :

$$\mathbb{P}(H_\theta(X) = 1|S = 1) = \mathbb{P}(H_\theta(X) = 1|S = -1)$$

Equality of Opportunity : Equality of Opportunity requires that the proportions of positively classified data for a preferred label/score are equal across, the sensitive groups. Formally, a binary classifier H_θ is fair with respect to equality of opportunity if the classification probability is indepent of the sensitive attribute for a subset ("opportunity"). Here the subset is simply positive labels :

$$\mathbb{P}(H_\theta(X) = 1|S = 1, Y = 1) = \mathbb{P}(H_\theta(X) = 1|S = -1, Y = 1)$$

Example : Let's say we have two schools denoted by "-1" and "1", and we have a classifier that classifies the student as admissible or not to a highly selective university. The school "-1" has 100 students, and 10 have the qualification (enough scores for example) to enter the said university. On the other hand, school "1" has 200 students and 50 of them have the required qualifications to get into the university (Y=1). If our model predicts exactly the label Y meaning those who are qualified are admitted then :

$$\mathbb{P}(H_\theta(X) = 1|S = -1) = \frac{1}{10} \neq \frac{1}{4} = \mathbb{P}(H_\theta(X) = 1|S = 1)$$

but : $\mathbb{P}(H_\theta(X) = 1|S = -1, Y = 1) = \mathbb{P}(H_\theta(X) = 1|S = 1, Y = 1) = 1$

Therefore, the model H_θ is fair with respect to equality of opportunity but not with respect to demographic parity.

Enforcing equality in Demographic Parity definition might come at a great cost to the loss function. Instead, we consider the following fairness score :

$$f_{DP}(\theta) = \mathbb{P}(H_\theta(X) = 1|S = 1) - \mathbb{P}(H_\theta(X) = 1|S = -1)$$

Moreover, in practice, we don't have access to the true distribution, so we consider the empirical estimate of the above quantity :

$$\hat{f}_{DP}(\theta) = \hat{\mathbb{E}}(\mathbb{I}_{H_\theta(x)=1}|S = 1) - \hat{\mathbb{E}}(\mathbb{I}_{H_\theta(x)=1}|S = -1) = \frac{1}{N_1} \sum_{\substack{(x,s,y)\in\hat{D} \\ s=1}} \mathbb{I}_{H_\theta(x)=1} - \frac{1}{N_{-1}} \sum_{\substack{(x,s,y)\in\hat{D} \\ s=-1}} \mathbb{I}_{H_\theta(x)=1}$$

where $N_j = |\{(x, s, y) \in \hat{D}|s = j\}|$ for $j \in \{-1, 1\}$

As explained above, we define, similarly to Demographic parity, the two following quantities for Equality of Opportunity :

$$f_{EO}(\theta) = \mathbb{P}(H_\theta(X) = 1|S = 1, Y = 1) - \mathbb{P}(H_\theta(X) = 1|S = -1, Y = 1)$$

$$\begin{aligned}\hat{f}_{EO}(\theta) &= \hat{\mathbb{E}}(\mathbb{I}_{H_\theta(x)=1}|S = 1, Y = 1) - \hat{\mathbb{E}}(\mathbb{I}_{H_\theta(x)=1}|S = -1, Y = 1) \\ &= \frac{1}{N_1} \sum_{\substack{(x,s,y)\in\hat{D} \\ s=1,y=1}} \mathbb{I}_{H_\theta(x)=1} - \frac{1}{N_{-1}} \sum_{\substack{(x,s,y)\in\hat{D} \\ s=-1,y=1}} \mathbb{I}_{H_\theta(x)=1}\end{aligned}$$

We will later, formally, define the surrogate version $\tilde{f}_{DP}$ and $\tilde{f}_{EO}$ of the above empirical fairness metric, which is widely used as it has all the "nice" properties and can be used as a constraint or regularization. But there is no immediate way of ensuring true fairness by minimizing the surrogate.

There are other fairness metrics that allow us to evaluate a different aspect of the learned model (section 6 in [10]). We choose to work with these two fairness metrics because they are widely used in most research papers [10].

3.2 SearchFair

In this subsection, we present SearchFair algorithm from [15] which we will construct a decentralized version of in this paper. The goal of the algorithm is to find a perfectly fair model. In the paper, the authors explain that the use of surrogate fairness constraints $\tilde{f}$ doesn't ensure the true fairness of the outputted model, and they consider a different optimization problem :

$$\theta^*(\lambda, \beta) = \arg\min_\theta l(\theta) + \lambda\tilde{f}(\theta) + \beta\Omega(\theta) \tag{1}$$

Where l is the loss function of the considered problem, $\tilde{f}$ is the surrogate fairness regularization (in the next theorem it is only required to be a convex function in θ) and Ω is a regularization.

The algorithm ensures exact fairness by finding a hyperparameter λ such that :

$$f(\theta^*) = \mathbb{P}(H_{\theta(X)^*} = 1|S = 1) - \mathbb{P}(H_{\theta^*}(X) = 1|S = -1) = 0$$

The following theorem and corollary provide the theoretical guarantee for the existence of such λ, under some assumptions that will be detailed in Section 4.2.

Theorem 3.1 (Continuity of f (too relaxed to be fair)) *Under some Assumptions the function $\lambda \to f(\theta^*(\lambda, \beta))$ is continuous*

Corollary 3.1.1 (Existence of a fair classifier) *Assume that Theorem 3.1 hold and that the convex approximation $\tilde{f}$ is chosen such that the optimization problem 1 there exists :*

(i) λ_+ such that $f(\theta^*(\lambda_+, \beta)) > 0$

(ii) λ_- such that $f(\theta^*(\lambda_-, \beta)) < 0$

Then there exists at least one value λ_0 in the interval $[\min(\lambda_-, \lambda_+), \max(\lambda_-, \lambda_+)]$ such that $f(\theta^*(\lambda_0, \beta)) = 0$.

SearchFair algorithm relies on the existence of λ_+ and λ_-, and runs a dichotomy search to narrow down the interval containing a λ_0. In [Scalable and Stable Surrogates for Flexible Classifiers with Fairness Constraints] the authors show that such λ_- and λ_+ don't exist for all fairness surrogates, but in our experiments, we always manage to find λ_- and λ_+ with our chosen surrogate.

The pseudo-algorithm is presented below, where equation 1 can be solved by an optimization algorithm (SGD for example).

Algorithm 1 SearchFair

Input : $\beta > 0$, number of dichotomy steps C, λ_- and λ_+ where $\hat{f}(\theta^\star(\lambda_-, \beta)) \leq 0$ and $\hat{f}(\theta^\star(\lambda_+, \beta)) \geq 0$
Output : a fair classifier
for c = 1, . . . , C **do**
  $\lambda = \frac{1}{2}(\lambda_+ + \lambda_-)$
  $\theta^\star(\lambda, \beta) = \arg\min_\theta l(\theta) + \lambda \tilde{f}(\theta) + \beta\Omega(\theta)$
  $\hat{f}_\lambda = \hat{f}(\theta^\star(\lambda, \beta))$
  if $\hat{f}_\lambda > 0$ **then**
    $\lambda_+ = \lambda$ and $\hat{f}_+ = \hat{f}_\lambda$
  else
    $\lambda_- = \lambda$ and $\hat{f}_- = \hat{f}_\lambda$
  end if
end for
if $|\hat{f}_-| < |\hat{f}_+|$ **then**
  return $\theta^\star(\lambda_-, \beta)$
else
  return $\theta^\star(\lambda_+, \beta)$
end if

Our main contribution will be to make a decentralized version of algorithm 1.
In the following subsection, we present the decentralized optimization algorithm that we will rely on.

3.3 D-SGD

In this subsection, we present the convergence result for decentralized SGD from [12] and the accelerated convergence when choosing a good communication topology from [13], which we will use to solve a decentralized optimization problem.

The algorithm aims to solve a consensus problem, where K clients with their own data distribution D_i want to minimize the global loss. More formally, the clients want to solve the following optimization problem :

$$\theta^* = \arg\min_\theta \left[g(\theta) := \frac{1}{K} \sum_{i=1}^{K} g_i(\theta) \right] \text{ where } g_i(\theta) = \mathbb{E}_{Z \sim D_i} \left(G_i(\theta, Z) \right)$$

D-SGD works by letting each client do their gradient step locally and then average their model parameters updates among their neighbors. Algorithm 2 shows a pseudo-implementation.

Algorithm 2 D-SGD

Input : Initialise $\forall i,\ \theta_i^{(0)}$, number of iterations T, step sizes $\{\eta_t\}_{t=0}^{T-1}$
for t = 0, . . . , T-1 **do**
  In parallel for each node i
  Sample $Z_i^{(t)} \sim D_i$
  $\theta_i^{(t+\frac{1}{2})} \leftarrow \theta_i^{(t)} - \eta_t \nabla G_i(\theta_i^{(t)}, Z_i^{(t)})$
  $\theta_i^{(t+1)} \leftarrow \sum_{j=1}^{n} W_{ij} \theta_j^{(t+\frac{1}{2})}$
end for

The following result from [kolskova paper] gives a theoretical convergence guarantee in the strongly-convex case (which will be our working case) under some assumptions that we will not detail here.

Theorem 3.2 (kolskova et al. Upper bound) *Under some assumptions on the losses g_i and communication topology W.For any target accuracy, $\varepsilon > 0$ There exists a constant stepsize (potentially depending on ε) such that the accuracy can be reached after at most the following number of iterations T :*

$$\sum_{t=0}^{T} \frac{w_t}{W_T}(\mathbb{E}g(\bar{\theta}^{(t)}) - g(\theta^*)) + \mu\mathbb{E}\|\bar{\theta}^{(T+1)} - \theta^*\|^2 \leq \varepsilon$$

After a certain number of iterations T depends on the problem parameters. Where w_t are positive weights and μ is such that g is μ-strongly convex.

In the paper, the authors derive an upper bound in $\tilde{\mathcal{O}}$ for the non-convex, convex, and strongly convex cases.

In [13], the authors prove that taking into account the choice of the communication topology in order to compensate for the heterogeneity in the data leads to faster convergence of the algorithm.

In the following subsection, we present theoretical results on estimating a mean in a decentralized setting, which will be useful for computing the fairness metric in a decentralized manner.

3.4 Gossip mean estimation

In this subsection, we present some results from [3] that we will need in order to compute a decentralized estimate of the fairness difference f.

The goal is to compute a mean $\bar{x}$ through a communication topology W which is a $K \times K$ matrix of K clients where each client i initially has a value $x_i^{(0)} \in \mathbb{R}$ ($x_i^{(0)}$ can also be a vector, but in our case, we only need the results for real numbers). We denote by $x^{(t)} = [x_1^{(t)}, \ldots, x_K^{(t)}]$.

As its name suggests, the algorithm is a gossip protocol, where each client computes its new estimate of the mean by weight averaging the values of its neighbors, we can write this as the following :

$$x^{(t+1)} = Wx^{(t)}$$

The following result from [Boyd paper] states that under some assumptions on W, each clients estimate converges to the mean and provides a bound to the speed of convergence.

Lemma 3.3 (Boyd result) *under the following assumptions :*

- *W doubly stochastic*
- *$\lambda_2(W) < 1$ where $\lambda_2(W)$ denotes the second largest eigenvalue of W*

The algorithm converges and after t iterations, it holds that :

$$\|x^{(t)} - \bar{x}\| \leq \lambda_2(W)^t \|x^{(0)} - \bar{x}\|$$

The second condition is to ensure the connectivity of the graph because otherwise, each connected component will converge to its own local mean, which is not necessarily equal to the global mean.

The obtained speed by Boyd considers the worst-case scenario where $x^{(0)} - \bar{x} \in$ Span$(u_2(W))$ where $u_2(W)$ is the eigenvector associated with $\lambda_2(W)$, but in practice other eigenvalues will have an impact.

4
DECENTRALIZED SEARCHFAIR

In this section, we present our main contribution D-SearchFair which relies on SearchFaire framework, D-SGD, and gossip algorithms to output a fair model.

4.1 Setup

We consider the same optimisation problem as in SearchFair, but assume that the surrogate fairness metric is separable on clients, that is : $\tilde{f}(\theta) = \frac{1}{K}\sum_{i=1}^{K} \tilde{f}_i(\theta)$. Therefore, we can write :

$$\theta^*(\lambda, \beta) = \arg\min_{\theta} \frac{1}{K} \sum_{i=1}^{K} (l_i(\theta) + \lambda \tilde{f}_i(\theta) + \beta\Omega(\theta)) \tag{2}$$

Where $\lambda, \beta \in \mathbb{R}^+$ are two hyperparameters, Ω a regularization (l2 regularization for example) and $\tilde{f}_i$ are the "local surrogate fairness metrics"

In order to be in the same optimization framework as D-SGD. we will assume the existence of pointwise loss and surrogate fairness : $l_i(\theta) = \mathbb{E}_{Z\sim D_i}(L_i(\theta, Z))$ and $\tilde{f}_i(\theta) = \mathbb{E}_{Z\sim D_i}(\tilde{F}_i(\theta, Z))$.

Choice of $\tilde{f}_i$ (and of $\tilde{f}$)
In [15] the authors provide an example for $\tilde{f}$ (denoted by $R_{D\hat{D}P}(f)$ in their paper). We follow the same scheme to define our $\tilde{f}$ depending on the sign of $\hat{f}(\theta^*(0, \beta))$. The idea is after solving problem 2 for $\lambda = 0$ we compute $\hat{f}(\theta^*(0, \beta))$ and if it is positive, meaning the model is biased against the group S=-1, then we choose $\tilde{f}$ such that the solution $\theta^*(1, \beta)$ will be biased against the group S=1, and vice versa.

$$\text{if} \hat{f}(\theta^*(0,\beta)) > 0 : \tilde{f}(\theta) = \hat{\mathbb{E}}(\tilde{F}(\theta, Z)) = \hat{\mathbb{E}}\left(L_F(h_\theta(X), -1)\frac{\mathbb{I}_{S=1}}{p_1} + L_F(h_\theta(X), 1)\frac{\mathbb{I}_{S=-1}}{1-p_1} - 1\right)$$
$$\text{if} \hat{f}(\theta^*(0,\beta)) < 0 : \tilde{f}(\theta) = \hat{\mathbb{E}}(\tilde{F}(\theta, Z)) = \hat{\mathbb{E}}\left((-1 + L_F(h_\theta(X), 1))\frac{\mathbb{I}_{S=1}}{p_1} + (-1 + L_F(h_\theta(X), 1))\frac{\mathbb{I}_{S=-1}}{1-p_1} + 1\right)$$

Where $\hat{\mathbb{E}}$ is the empirical mean taking over $(x, s, y) \in \hat{D}$ and L_F is a loss function.
In practice, we pick the smooth Hinge Loss as L_F.

Now, if $\tilde{f} = \frac{1}{n}\sum_{Z\in\hat{D}} \tilde{F}(\theta, Z)$ where n the number of all clients data points, which is the case here, then :

$$\tilde{F}_i(\theta, Z) = \frac{n_i K}{n}\tilde{F}(\theta, Z) \quad \text{so as to have :} \quad \tilde{f}(\theta) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{Z\in\hat{D}_i}\tilde{F}_i(\theta, Z)$$

4.2 Assumptions

We will state, in this subsection, all the assumptions that allow us to have access to the theoretical results of D-SGD, SearchFair, and Gossip mean estimation. And we will provide two example losses that verify the assumptions in the next subsection.

Assumption 1 (SearchFair assumptions) *With $\beta > 0$ fixed, we define the set of learnable model parameters as $\Theta_\Lambda = \{\theta \in \mathbb{R}^d, \exists\lambda \geq 0, \theta = \theta^*(\lambda, \beta)\}$*

- *Problem 2 is strongly convex in θ*
- *$\forall\theta \in \mathbb{R}^d$, $\tilde{f}(\theta)$ is bounded in [-B, B]*
- *$\exists\rho$, a metric such that (Θ_Λ, ρ) is a metric space*
- *$\forall x \in supp(X), g(\theta) : \theta \to h_\theta(x)$ is continuous*
- *$\forall\theta \in \Theta_\Lambda$, $h\theta$ is Lebesgue measurable and the set $\{x : (x, s, y) \in supp(Z), h_\theta(x) = 0\}$ is a Lebesgue null set*

— *The probability density function* $f_{Z|s=1}$ *and* $f_{Z|s=-1}$ *are Lebesgue measurable*

Assumption 1 is the required assumption in [15] in order to have the continuity of $\lambda \to f(\theta^*(\lambda, \beta))$ presented in Theorem 3.1.

Assumption 2 (L-smoothness) *We assume that each function* F_i, L_F *and* Ω *are differentiable for each* $z \in supp(Z)$, *and that they verify the following conditions :*

— $\|\nabla L_i(\theta, z) - \nabla L_i(\omega, z)\| \leq L_1 \|\theta - \omega\|$

— $\|\nabla L_F(\theta, z) - \nabla L_F(\omega, z)\| \leq L_2 \|\theta - \omega\|$

— $\|\nabla \Omega(\theta) - \nabla \Omega(\omega)\| \leq L_3 \|\theta - \omega\|$

Assumption 2 is required for the convergence result of D-SGD. A smaller Lipschitz constant implies faster convergence.

Assumption 3 (convexity) *We assume that* L_i *and* L_F *are convex in* θ. *And we assume that* Ω *is strongly convex in* θ

Assumption 3 along with $\lambda \geq 0,\ \beta > 0$ makes the problem 2 μ-strongly convex, which is needed as the first statement in assumption 1 in order to ensure that there is only one optimum for optimization problem 2. It is also needed to obtain the convergence result of D-SGD in the strongly convex case, and the parameter μ does come up in the convergence speed of the algorithm. A higher value for μ implies faster convergence.

Assumption 4 (Bounded noise and variance at the optimum) *We assume the following bound on the noise at the optimum for all* $\lambda, \beta \in \mathbb{R}^+$, *where we write* $\theta^* = \theta^*(\lambda, \beta)$

$$\|\nabla l_i(\theta^\star)\|_2^2,\ \|\nabla \tilde{f}_i(\theta^\star)\|_2^2,\ \textit{and}\ \|\nabla \Omega(\theta^\star)\|_2^2\ \textit{bounded.}$$

And we assume the following for the variance :

$$\mathbb{E}_{Z\sim D_i}(\|\nabla L_i(\theta^\star, Z) - \nabla l_i(\theta^\star)\|_2^2)\ \textit{and}\ \mathbb{E}_{Z\sim D_i}(\|\nabla \tilde{F}_i(\theta^\star, Z) - \nabla \tilde{f}_i(\theta^\star)\|_2^2)\ \textit{bounded.}$$

Assumption 4 is required for the convergence of D-SGD and the exact bounds on these quantities do come up in the convergence speed bound of the algorithm. Lower bounds imply faster convergence.

The bound result on l_i is assumed anyway to be able to use D-SGD and compute an unconstrained solution. We can obtain a similar bound on $\tilde{f}$ by picking $L_F = L_i$ for some i or a combination of L_i. The bounded assumption on Ω can be obtained by picking a differentiable regularization, such as l2-regularization.

The same comment above applies to the bounded assumption on the variance for L_i and L_F. There is no bounded variance assumption on the regularization as $\mathbb{E}_{Z\sim D_i}(\|\nabla \Omega(\theta^\star) - \nabla \mathbb{E}_{Z\sim D_i}(\Omega(\theta^\star))\|_2^2) = 0$.

Assumption 5 (Mixing parameter) W *is symmetric* ($W = W^T$) *doubly stochastic matrix* ($W\mathbf{1} = \mathbf{1}$ *and* $\mathbf{1}^T W = \mathbf{1}^T$). *We also assume that :*

$$\rho\left(W - \frac{\mathbf{1}\mathbf{1}^T}{N}\right) < 1 \qquad \textit{Where } \rho(M) \textit{ is the spectral radius of } M$$

In other words : $\lambda_2(W) < 1$, *the second-largest eigenvalue of* W *is strictly less than 1 meaning the graph is connected.*

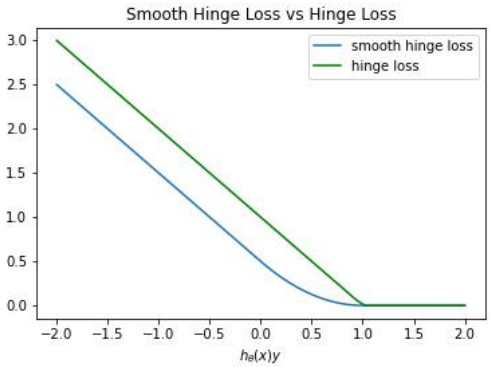
4.3 Example loss functions

In all the following examples, we assume that our model is of the form : $h_\theta(x) = \theta^T \Phi(x)$ where Φ is a **fixed bounded** transformation applied to x (doesn't depend on θ)

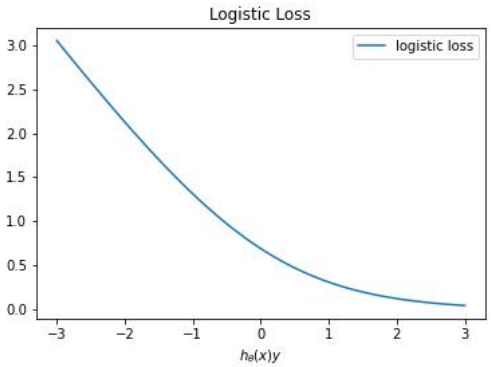
Both losses presented below are L-smooth and convex (for this linear classifier under the bounded assumption of Φ). The proof can be found in the appendix.

Example1 : Smooth Hinge loss (Rennie and Srebro's)

$$L_{sh}(h_\theta(x), y) = \begin{cases} \frac{1}{2} - h_\theta(x)y & \text{if } h_\theta(x)y \leq 0 \\ \frac{1}{2}(1 - h_\theta(x)y)^2 & \text{if } 0 < h_\theta(x)y < 1 \\ 0 & \text{if } h_\theta(x)y \geq 1 \end{cases}$$



(a) *Smooth Hinge loss vs classical hinge loss*



(b) *Logistic Loss figure*

FIGURE 1 – *Algorithm 2 demonstration (apex on tropical segment $[u, v]$)*

We set $L_F(\theta, Z) = L_{sh}(h_\theta(x), y)$ and we choose $\Omega(\theta) = \|\theta\|_2^2$ the l_2 regularization. We have that :

$$\nabla L_{sh}(h_\theta(x), y) = \begin{cases} -y\Phi(x) & \text{if } h_\theta(x)y \leq 0 \\ -y\Phi(x) + \Phi(x)\Phi(x)^T\theta & \text{if } 0 < h_\theta(x)y < 1 \\ 0 & \text{if } h_\theta(x)y \geq 1 \end{cases}$$

Example2 : Logistic loss

$$L_{logistic}(h_\theta(x), y) = \log(1 + \exp(-yh_\theta(x)))$$

$$\nabla L_F(\theta, Z) = \frac{-y}{(1 + \exp(y\Phi(x)^T\theta))}\Phi(x)$$

4.4 ALGORITHM

In this subsection, we present one of our main contributions, D-SearchFair. The pseudo algorithm 3 uses both D-SGD and Gossip mean estimation to make the algorithm fully decentralized over a communication Topology.

Approximation of $\hat{f}$ using Gossip algorithm We will use Demographic parity as an example here (Equality of opportunity is done similarly). Let θ be some model parameter :

$$\begin{aligned}\hat{f}(\theta) &= \sum_{(x,s,y)\in\hat{D}} \mathbb{I}_{h_\theta(x)\geq 0}\left(\frac{\mathbb{I}_{s=-1}}{N_1} - \frac{\mathbb{I}_{s=-1}}{N_{-1}}\right) \\ &= \frac{1}{K}\sum_{i=1}^{K} K \sum_{(x,s,y)\in\hat{D}_i} \mathbb{I}_{h_\theta(x)\geq 0}\left(\frac{\mathbb{I}_{s=-1}}{N_1} - \frac{\mathbb{I}_{s=-1}}{N_{-1}}\right) \\ \hat{f}(\theta) &= \frac{1}{K}\sum_{i=1}^{K}\varphi_i(\theta)\end{aligned}$$

And all that's left is to estimate $\hat{f}$ that we just rewrote into a decentralized mean using the Gossip algorithm. The only catch is that each client has their own θ_i, and so in practice, we compute the following quantity :

$$\hat{\varphi}(\theta_1,\ldots,\theta_K) = \frac{1}{K}\sum_{i=1}^{K}\varphi_i(\theta_i)$$

And we have that : $\hat{\varphi}(\theta,\ldots,\theta) = \hat{f}(\theta)$

We denote by $\hat{\varphi}_i^{(t)}(\theta_1,\ldots,\theta_K)$ the estimate of client i after t iterations of the Gossip algorithm. To alleviate notation we will simply write $\hat{\varphi}_i^{(t)}$ instead of $\hat{\varphi}_i^{(t)}(\theta_1,\ldots,\theta_K)$ and $\hat{\varphi}$ instead of $\hat{\varphi}(\theta_1,\ldots,\theta_K)$.

Algorithm 3 Decentralized SearchFair

Input : for each node $i \in [n]$ initialize $\theta_i^{(0)} \in \mathbb{R}^d$, number of D-SGD iterations T, step sizes $\{\eta_t\}_{t=0}^{T-1}$, communication topology W, number of dichotomy steps C, λ_- and λ_+ where $\hat{f}(\theta^\star(\lambda_-,\beta)) \leq 0$ and $\hat{f}(\theta^\star(\lambda_+,\beta)) \geq 0$
Output : K fair classifier
for c = 1, …, C **do**
  in parallel for each client i
  $\lambda = \frac{1}{2}(\lambda_+ + \lambda_-)$ ▷ Each client updates its λ locally
  $\theta_i^{(T)}$ = D-SGD(λ, β, T, $\{\eta_t\}_{t=0}^{T-1}$, W)
  Compute $\varphi_i^{(0)}(\theta_i^{(T)})$
  $\hat{\varphi}_i^{(t_{gossip})}$ = GossipMean($\varphi_1^{(0)}(\theta_1^{(T)}),\ldots,\varphi_K^{(0)}(\theta_K^{(T)})$)
  if $\hat{\varphi}_i^{(t_{gossip})} > 0$ **then** ▷ local updates
    $\lambda_+ = \lambda$ and $\hat{f}_+ = \hat{f}_\lambda$
  else
    $\lambda_- = \lambda$ and $\hat{f}_- = \hat{f}_\lambda$
  end if
end for

The algorithm assumes that we have found λ_- and λ_+ which can be done in an earlier step (in a decentralized manner). At the start of the algorithm, each client has the same λ hyperparameter.

Each client computes its local λ at each iteration, and D-SGD is then run to compute the new model parameters $\theta_1,\theta_2,\ldots,\theta_K$. $\hat{f}$ is then estimated by each client using the Gossip mean estimation algorithm [] before updating the values of λ_- and λ_+ locally.

In our implementation, we assume that the parameters K (number of clients), n (global number of data points), and $p_1 = P(S=1)$ in case of demographic parity (or $P(S=1,Y=1)$ and $P(S=-1,Y=1)$ in case of equality of opportunity) are known to all clients.

The total number of data points n and the number of clients K aren't sensitive and shouldn't be a problem to share before starting the algorithm. The sensitive attribute proportions on the other hand can be sensitive and

in that case, they could be estimated via a Gossip Algorithm (mean estimation) at the start of the algorithm as follows :

$$p_1 = \frac{1}{n} \sum_{(x,s,y)\in\hat{D}} \mathbb{I}_{s=1} = \frac{1}{K} \sum_{i=1}^{K} \frac{Kn_i}{n} \frac{1}{n_i} \sum_{(x,s,y)\in\hat{D}_i} \mathbb{I}_{s=1}$$

$$p_1 = \frac{1}{K} \sum_{i=1}^{K} \frac{Kn_i}{n} p_{1,i}$$

5 THEORETICAL ANALYSIS

In this section, we will first show that our algorithm converges to a fair model in case we can run D-SGD and Gossip algorithm infinitely. And in the following section, we will provide some theoretical guarantees for the practical implementation.

5.1 Theoretical algorithm analysis

In this subsection, we provide the theoretical guarantee to Algorithm 3 when we can run D-SGD and Gossip algorithm infinitely.

Theorem 5.1 (Fairness root guarantee) *Under Assumptions 1, 2, 3, 4, 5. Assume further that each client i has access to its true distribution D_i and that λ_- and λ_+ exists such that :*

(i) λ_+ such that $f(\theta^*(\lambda_+, \beta)) > 0$

(ii) λ_- such that $f(\theta^*(\lambda_-, \beta)) < 0$

And if we assume that we can run infinitely many D-SGD and Gossip iterations, that is if $T = +\infty$ and $t_{gossip} = +\infty$: $\exists C \in \mathbb{N} \cup \{+\infty\}$ such that the outputted model θ^*_{output} is perfectly fair : $f(\theta^*_{output}) = 0$

Proof : With assumptions 1 and 3 our optimization problem is strongly convex and satisfies the conditions of Theorem 1 in [Too relaxed to be fair paper], where the authors prove the continuity of $\lambda \to f(\theta^*(\lambda, \beta))$. And with the existence of λ_- and λ_+, we know that a zero exists and that a dichotomy method will find such a zero.

Assumptions 2, 3, 4, and 5 satisfy the necessary conditions for the convergence results in the strongly convex case of [unified theory of d-sgd paper] (with $\tau = 1$) and for the Gossip mean estimation. And if $T = +\infty$ and $t_{gossip} = +\infty$, all clients converges to the same θ^* and $\hat{f}(\theta^*)$ and our main algorithm works exactly as its centralized version and outputs a fair model.

This result is not impressive nor novel, but it is important to know that in a perfect scenario, we have exactly what we want. In the next subsection, we will explore the empirical errors and provide a theoretical guarantee up to ε in fairness.

5.2 Practical implementation analysis

In this subsection, we will analyze the practical implementation of the algorithm ($T < +\infty$ and $t_{gossip} < +\infty$). And we will only consider the linear classifier in our proofs : $h_\theta(x) = \theta^T \Phi(x)$ where Φ is a **fixed bounded** transformation applied to x (doesn't depend on θ).

In practice, there are 3 sources of error :
- The first source of errors is the error between the empirical estimate $\tilde{f}$ and the true fairness difference f.
- The second source of error comes from running D-SGD for a finite number of iterations and each client ends up with $\theta_i(\lambda, \beta) \approx \theta^\star(\lambda, \beta)$
- The third source of error comes from running Gossip algorithm for a finite number of iterations, causing each client to have a different estimate value $\hat{\varphi}_i^{(t)}$ and they might have different signs, which may occur if the empirical fairness value is close to 0, causing the clients to choose different λ while running the algorithms. Having a different λ for each client doesn't mean necessarily that the algorithm will fail, but it is outside our theoretical context.

Assumption 6 (Bounded transition) *We assume the following :*

$$\forall (x,s,y) \in \hat{D},\ \forall \lambda \in (\lambda_-, \lambda_+)\ :\ \sum_{(x,s,y)\in\hat{D}} \mathbb{I}_{h^*_\theta(\lambda,\beta)(x)=0} \leq m$$

Assumption 6 will allow us to bound the transition between values of $\hat{f}$ as it takes a limited number of values in $\mathbb{Q}$. We also rely on this assumption for a bound between the empirical fairness $\hat{f}$ and our estimate $\hat{\varphi}$ in the following lemma.

Lemma 5.2 (Fairness control) *Let θ and $\theta_1, \ldots, \theta_K \in \mathbb{R}^d$ be model parameters. Assume that :*

$$\sum_{(x,s,y)\in\hat{D}} \mathbb{I}_{h_\theta(x)=0} \leq m$$

We can derive the following bound between $\hat{f}(\theta)$ and $\hat{\varphi}(\theta_1, \ldots, \theta_K)$

$$|\hat{f}(\theta) - \hat{\varphi}(\theta_1, \ldots, \theta_K)| \leq \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \left(\sum_{i=1}^{K} \|\theta_i - \theta\| \sum_{\substack{(x,s,y)\in\hat{D}_i \\ h_\theta(x)\neq 0}} \frac{\|\Phi(x)\|}{|h_\theta(x)|} + m \right) \tag{3}$$

Lemma 5.2 gives us a bound to control the difference between what we try to estimate empirically $\hat{\varphi}(\theta_1, \ldots, \theta_K)$ and what we actually want to compute in theory $\hat{f}(\theta^*)$ if $\frac{\|\Phi(x)\|}{|h_\theta(x)|}$ is finite. When θ_i, which here represents each client model, converges to θ we will get a bound of $\varepsilon + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$.

Lemma 5.3 (Gossip rate of convergence Boyd et al. []]

$$\|(\hat{\varphi}_1^{(t)}, \ldots, \hat{\varphi}_K^{(t)})^T - \hat{\varphi}\mathbf{1}\|^2 \leq \lambda_2^{2t}(W) \|(\hat{\varphi}_1^{(0)}, \ldots, \hat{\varphi}_K^{(0)})^T - \hat{\varphi}\mathbf{1}\|^2 \tag{4}$$

Lemma 5.3 gives us a bound between the exact mean $\hat{\varphi}(\theta_1, \ldots, \theta_K)$ and the estimate $\hat{\varphi}_i^{(t)}$ each client has after t gossip iterations. We require such result as each client i will have an estimate $\hat{\varphi}_i^{(t)}$ of $\hat{\varphi}$ which approximates the quantity of interest $\hat{f}$ by lemma 5.2

Lemma 5.4 (No root dichotomy) *Let $(\lambda_-^{(t)}, \lambda_+^{(t)})$ denote the interval of dichotomy after t iterations. We assume Assumption 6 and that :*

(i) $\hat{f}(\theta^*(\lambda_+^{(0)}, \beta)) \geq 0$

(ii) $\hat{f}(\theta^*(\lambda_-^{(0)}, \beta)) \leq 0$

Then there exists a number of dichotomy iterations C such that :

$$\min(|\hat{f}(\theta^*(\lambda+^{(C)}, \beta))|, |\hat{f}(\theta^*(\lambda-^{(C)}, \beta))|) \leq m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

In practice, we run dichotomy on $\hat{f}$ which may not have a root. Lemma 5.4 tell us that one of the two values that we will have at the end of C iterations $|\hat{f}(\theta^*(\lambda+^{(C)},\beta))|$ or $\hat{f}(\theta^*(\lambda-^{(C)},\beta))|$ will be "small enough".

Proof of Lemma 5.2, 5.3, and 5.4 can be found in the supplementary material.

Our goal will be to have $sign(\hat{\varphi}_i^{(t)}) = sign(\hat{\varphi}(\theta_1,\dots,\theta_K)) = sign(\hat{f}(\theta^*))$ so that the dichotomy finds the zero (or a value close to 0) of $\hat{f}$. The next theorem is the main fairness guarantee, and it formalizes this idea.

Theorem 5.5 (Practical ε-fairness guarantee) *Under Assumptions 1, 2, 3, 4, 5, 6. We further assume that λ_- and λ_+ exists where :*

(i) $\hat{f}(\theta^*(\lambda_+,\beta)) \geq 0$

(ii) $\hat{f}(\theta^*(\lambda_-,\beta)) \leq 0$

Let $\varepsilon > 0$

There exists a number of iterations $t_{gossip}(\varepsilon)$ and $t_{D-SGD}(\varepsilon)$ such that there exists t_s where Algorithm 3 will have reached a value parameter θ_i for each client such that :

$$\forall i, \mathbb{E}(|\hat{f}(\theta_i)|) \leq \varepsilon + 2m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

Where the mean is taking over the random variable θ_i (due to the sampling randomness in D-SGD)

Proof : for some iteration c of the algorithm dichotomy (outer loop). λ is fixed during this iteration for each client and we simply write θ^* instead of $\theta^*(\lambda,\beta)$.

We first start by fixing t_{gossip}, we have that :

$$\begin{aligned}
\hat{\varphi}_i^{(0)}(\theta_i) &= K \sum_{(x,s,y)\in\hat{D}_i} \mathbb{I}_{h_\theta(x)\geq 0}\left(\frac{\mathbb{I}_{s=-1}}{N_1} - \frac{\mathbb{I}_{s=-1}}{N_{-1}}\right) \\
&= \frac{KN_{1,i}}{N_1} \sum_{(x,s,y)\in\hat{D}_i} \frac{\mathbb{I}_{h_\theta(x)\geq 0}\mathbb{I}_{s=1}}{N_{1,i}} - \frac{KN_{-1,i}}{N_{-1}} \sum_{(x,s,y)\in\hat{D}_i} \frac{\mathbb{I}_{h_\theta(x)\geq 0}\mathbb{I}_{s=-1}}{N_{-1,i}}
\end{aligned}$$

$$\forall i \in [K], \quad -\frac{KN_{-1,i}}{N_{-1}} \leq \hat{\varphi}_i^{(0)}(\theta_i) \leq \frac{KN_{1,i}}{N_1} \Rightarrow |\hat{\varphi}| \leq 1 \quad \left(\text{as } N_{\pm 1} = \sum_{i=1}^K N_{\pm 1,i}\right) \tag{5}$$

$$\begin{aligned}
\text{and : } \|(\hat{\varphi}_1^{(0)},\dots,\hat{\varphi}_K^{(0)})^T - \hat{\varphi}\mathbf{1}\|^2 &= \sum_{i=1}^K (\hat{\varphi}_i^{(0)})^2 - K\hat{\varphi}^2 \\
&\leq \sum_{i=1}^K (\hat{\varphi}_i^{(0)})^2 \\
&\overset{(5)}{\leq} \sum_{i=1}^K \max\left(\frac{K^2N_{-1,i}^2}{N_{-1}^2}, \frac{K^2N_{1,i}^2}{N_1^2}\right) \\
&\leq K^2\left(\frac{\sum_{i=1}^K N_{-1,i}^2}{N_{-1}^2} + \frac{\sum_{i=1}^K N_{1,i}^2}{N_1^2}\right) \\
&\leq K^2(1+1) = 2K^2
\end{aligned}$$

We therefore pick t_{gossip} such that :

$$\begin{aligned}
\lambda_2^{2t_{gossip}}(W)(2K^2) &\leq \left(\frac{\varepsilon}{4}\right)^2 \\
t_{gossip} &\geq \log(\frac{\varepsilon\sqrt{2}}{8K})/\log(\lambda_2(W)) \\
\Rightarrow \forall i \in [K], \quad |\hat{\varphi}_i^{(t_{gossip})} - \hat{\varphi}| &\leq \frac{\varepsilon}{4}
\end{aligned}$$

(ex : for K=50 clients, $\varepsilon = 10^{-2}$ and a ring topology we get : $t_{gossip} \geq 461$)
the result is established $\forall \theta_1, \dots, \theta_K$ therefore :

$$\forall i \in [K], \quad \mathbb{E}(|\hat{\varphi}_i^{(t_{gossip})} - \hat{\varphi}|) \leq \frac{\varepsilon}{4} \tag{6}$$

We have from lemma 5.2 :

$$|\hat{f}(\theta) - \hat{\varphi}(\theta_1, \dots, \theta_K)| \leq \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \left(\sum_{i=1}^{K} \|\theta_i - \theta\| \sum_{\substack{(x,s,y) \in \hat{D}_i \\ h_\theta(x) \neq 0}} \frac{\|\Phi(x)\|}{|h_\theta(x)|} + m \right) \tag{7}$$

$$\Rightarrow \mathbb{E}(|\hat{f}(\theta^\star) - \hat{\varphi}(\theta_1, \dots, \theta_K)|) \leq \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \left(\sum_{i=1}^{K} \|\theta_i - \theta\| \sum_{\substack{(x,s,y) \in \hat{D}_i \\ h_\theta(x) \neq 0}} \frac{\|\Phi(x)\|}{|h_\theta(x)|} + m \right) \tag{8}$$

and in [][D-SGD paper] we know that :

$$\sum_{i=1}^{K} \mathbb{E}(\|\theta_i - \theta^\star\|) \leq \sum_{i=1}^{K} \mathbb{E}(\|\theta_i - \theta^\star\|) + \mathbb{E}(\|\bar{\theta} - \theta^\star\|) \xrightarrow[t_{D-SGD} \to +\infty]{} 0$$

Then there exists $t_{D-SGD}^{(c)}$ such that :

$$\mathbb{E}(|\hat{f}(\theta^\star) - \hat{\varphi}(\theta_1, \dots, \theta_K)|) \leq \frac{\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$
$$\text{and : } \mathbb{E}(|\hat{f}(\theta^\star) - \hat{\varphi}(\theta_i, \theta_i, \dots, \theta_i)|) = \mathbb{E}(|\hat{f}(\theta^\star) - \hat{f}(\theta_i)|) \leq \frac{\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

Where θ_i here are obtained after $t_{D-SGD}^{(c)}$ iterations of D-SGD.

— if $\forall i \in [K], \quad \mathbb{E}(|\hat{f}(\theta_i)|) \leq \varepsilon + 2m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$ then we already have what we want ($C = 0$)

— if $\exists i \in [K], \quad \mathbb{E}(|\hat{f}(\theta_i)|) > \varepsilon + 2m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$ Then :

$$\mathbb{E}(|\hat{f}(\theta^\star) - \hat{f}(\theta_i)|) \leq \frac{\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \Rightarrow |\hat{f}(\theta^\star)| > \frac{3\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

$$|\hat{f}(\theta^\star) - \mathbb{E}(\hat{\varphi})| \leq \mathbb{E}(|\hat{f}(\theta^\star) - \hat{\varphi}|) \overset{(7)}{\leq} \frac{\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \Rightarrow |\mathbb{E}(\hat{\varphi})| \geq \frac{\varepsilon}{2}$$

$$\forall i \in [K], \quad |\hat{\varphi}_i^{(t_{gossip})} - \hat{\varphi}| \leq \mathbb{E}(|\hat{\varphi}_i^{(t_{gossip})} - \hat{\varphi}|) \leq \frac{\varepsilon}{4} \Rightarrow \forall i \in [K], \quad sign(\mathbb{E}(\hat{\varphi}_i^{(t_{gossip})})) = sign(\mathbb{E}(\hat{\varphi}))$$

$$|\hat{f}(\theta^\star)| > \frac{3\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \text{ and } |\hat{f}(\theta^\star) - \mathbb{E}(\hat{\varphi})| \leq \frac{\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$
$$\Rightarrow sign(\mathbb{E}(\hat{\varphi})) = sign(\hat{f}(\theta^\star))$$

And we, therefore, can run the dichotomy on the sign of $\mathbb{E}(\hat{\varphi})$) and use the result of lemma 5.4 : $\exists t_s$ such that $\exists i \in [K], \quad |\hat{f}(\theta^*)| \leq m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$. Then :

$$\mathbb{E}(|\hat{f}(\theta^\star) - \hat{f}(\theta_i)|) \leq \frac{\varepsilon}{4} + m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \Rightarrow \forall i \in [K], \ |\mathbb{E}(\hat{f}(\theta_i))| \leq \varepsilon + 2m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

This theorem tells us that under these assumptions, a number of iterations can be set before running the algorithm for both D-SGD and Gossip mean estimation such that the final models for each client achieve a certain level of fairness that depends on ε and m.

Assumption 6 states that the decision boundary of the optimized model θ^* will at most pass on m data points. In the case of a linear boundary, this translates to having at most m collinear data points.

The result of Theorem 5.5 tells us that we will be able to achieve fairness if hyperparameters are well chosen, but don't provide any upper bound on the speed at which we will achieve such fairness. In fact, in order to have access to such results we need to know some properties of the empirical fairness $\hat{f}$.

5.3 Cost of the algorithm

Algorithm 3 uses 2 loops, the outer loop is the dichotomy search for the hyperparameter λ while the inner loop runs the D-SGD optimization algorithm to find a new $\theta^*(\lambda, \beta)$ at each iteration, which makes the final algorithm very costly and slow.

A possible **practical** improvement is to take advantage of the continuity of the fairness metric with respect to θ and search through the segment $(\theta^*(\lambda_-, \beta), \theta^*(\lambda_+, \beta))$. This will allow us to get the new model parameters through simple averaging (in our case we pick the middle of the interval) and therefore largely accelerate the algorithm. The downside is that there is no reason why the average of the parameters will coincide (or be close) to $\theta^*(\lambda, \beta)$ for a certain λ. That being said, we do have the following result thanks to the convexity of the loss.

Lemma 5.6 *Under Assumption 3.*

$$\forall t \in [0,1],\ \frac{1}{K}\sum_{i=1}^{K} l_i(t\theta^*(\lambda_-,\beta) + (1-t)\theta^*(\lambda_+,\beta)) \leq t\frac{1}{K}\sum_{i=1}^{K} l_i(\theta^*(\lambda_-,\beta)) + (1-t)\frac{1}{K}\sum_{i=1}^{K} l_i(\theta^*(\lambda_+,\beta))$$

Meaning that the loss will be in the worst-case scenario as bad as for $\theta^*(\lambda_-, \beta)$ or $\theta^*(\lambda_+, \beta)$.

Proof : We simply use the convexity property of the loss.

Below is the averaging version of D-SearchFair pseudo-algorithm. We write $\theta_- := \theta^*(\lambda_-, \beta)$ and $\theta_+ := \theta^*(\lambda_+, \beta)$.

Algorithm 4 Decentralized SearchFair with Averaging

Input : number of D-SGD iterations T, step sizes $\{\eta_t\}_{t=0}^{T-1}$, Communication topology W, number of dichotomy steps C, λ_- and λ_+. For each node $i \in [n]$ $\theta_{i,-}^{(1)}$ and $\theta_{i,+}^{(1)}$ local estimates of θ_- and θ_+ where $\hat{f}(\theta_-) \leq 0$ and $\hat{f}(\theta_+) \geq 0$
Output : K fair classifiers
for c = 1, . . . , C **do**
  In parallel for each client i
  $\theta_i^{(c)} = \frac{1}{2}\theta_{i,-}^{(c)} + \frac{1}{2}\theta_{i,+}^{(c)}$ ▷ Each client updates its θ locally
  Compute $\varphi_i^{(0)}(\theta_i^{(c)})$
  $\hat{\varphi}_i^{(t_{gossip})} = \text{GossipMean}(\varphi_1^{(0)}, \ldots, \varphi_K^{(0)})$
  if $\hat{\varphi}_i^{(t_{gossip})}$ **then** ▷ local updates
    $\theta_{i,+}^{(c+1)} = \theta_i^{(c)}$
  else
    $\theta_{i,-}^{(c+1)} = \theta_i^{(c)}$
  end if
end for

5.4 Data heterogeneity

One major characteristic of our problem is the heterogeneity in the data. In practice, clients will have different data distributions, which can impact the convergence. The work done by [13] shows how to take advantage of such characteristics and choose a communication topology to provide a faster convergence rate.

In our case, we wanted to investigate the heterogeneity effect on achieving better fairness, but as the previous theorem shows, we can achieve exact fairness if we have access to the true distributions with infinite computing power. As we have no result on the speed at which the dichotomy converges, we cannot derive a faster convergence depending on the topology.

The only part of the algorithm that relies on the communication topology is the mean estimation through the gossip algorithm. Following the same intuition in [13], we thought that choosing the topology depending on the local percentage of the sensitive attribute would give the same level of fairness with lower gossip iterations, which was the case in some experiments. However, the following preliminary result tells us that it might not be possible to derive a faster convergence based only on the percentage of the sensitive attributes.

Theorem 5.7 *Let $p_1, \dots, p_K$ be the clients proportions of the sensitive attribute : for client i, $p_i = \hat{}$ $P_{\hat{D}_i}(S = 1)$. we assume that $\forall i,\ p_i \in]0, 1]$.*
Let W be a topology verifying assumption 4.
We fix a model θ.
Then $\forall \varepsilon > 0$ there exists a dataset $(\hat{D}_i, \dots, \hat{D}_K)$ that have the same proportions $p_1, \dots, p_K$, and such that the convergence speed of the gossip algorithm is lower bounded as follows :

$$\forall t \geq 0, \quad \|(\hat{\varphi}_1^{(t)}, \dots, \hat{\varphi}_K^{(t)})^T - \hat{\varphi}\mathbf{1}\| \geq \lambda_2^t(W)\|(\hat{\varphi}_1^{(0)}, \dots, \hat{\varphi}_K^{(0)})^T - \hat{\varphi}\mathbf{1}\| - \varepsilon$$

Proof of Theorem 5.7 can be found in the supplementary material.

Theorem 5.7 tells us that we can construct a parametric counter-example dataset, that depends on $p_1, \dots, p_K$ and W, such that the convergence speed of the gossip algorithm is as close as we want from Boyd result in lemma 3.3. It must be noted that we assumed that the model θ is fixed since we rely on a predefined decision bound to construct the counter-example, which is not the case when running algorithm 3.

6
EXPERIMENTS AND RESULTS

In this section, we empirically evaluate D-SearchFair with 3 baselines on 2 real-world datasets and one synthetic dataset. We also run the evaluations across two different models.

Datasets We consider CelbA [14], ACSincome [8] as real-world datasets, the exact data processing can be found in the supplementary materials. We also consider a simple synthetic dataset where we generate $X_{|S=i} \sim \mathcal{N}(\mu_i, \sigma_i)$ with $i \in \{-1, 1\}$ and control Y distribution such that the most accurate model isn't fair. The heterogeneity of the proportion of S for each client is controlled by $p \sim \text{Clipped}(\mathcal{N}(0.5, \alpha^2), 0, 1) : \alpha \to 0$ all clients have a group proportion of 0.5 and $\alpha \to +\infty$ clients have a group proportion uniformly sampled in the interval [0, 1].

Baselines In order to ensure that the final model obtained by each client is useful, we compare D-SearchFair and D-SearhcFair with averaging to the unconstrained decentralized model by simply running D-SGD with $\lambda = 0$ and $\beta = 0$.

Models We use two models. The first is a simple linear classifier $h_\theta(x) = \theta^T x$ (where x contains a column of 1 that represents the bias term). The second model is a simple Neural Network with one hidden layer and ReLU activation functions.

Metrics We are interested in learning a fair model that is as well useful. So our first metric is the empirical fairness value for Demographic Parity and Equality of opportunity, and we will consider the maximum absolute value of the clients fairness (lower is better). The second metric is the models Accuracy (averaged across all client models on the entire dataset) (higher is better).

Parameters batch size number of clients heterogeneity number of data points, number of each algorithm iterations.

6.1 Single run experiment

In this subsection, we present the result of a single run of Algorithm 3 to investigate the behavior of the algorithm through its iterations.

We choose to work with the linear model as it is the one covered by our theoretical results, and we will run the experiment on the synthetic dataset shown in Figure 2 as it offers us greater control over the distributions of each client.

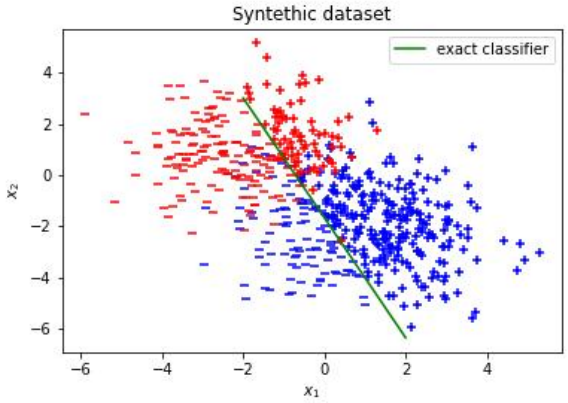
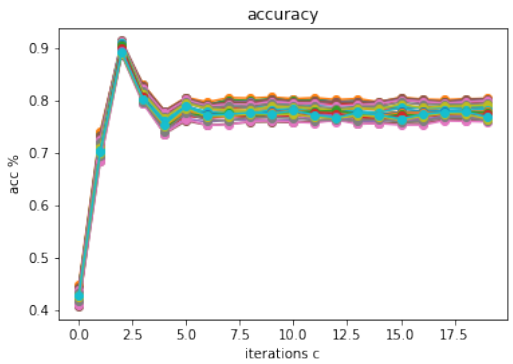
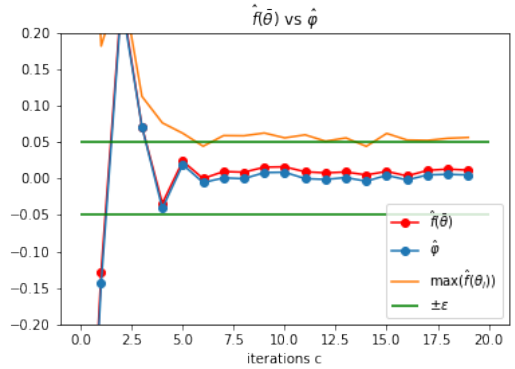


Figure 2 – *Synthetic dataset : blue means S=1, read means S=-1. plus sign corresponds to Y=1 and minus sign to Y=-1.* $\mu_1 = \begin{pmatrix}11\\8\end{pmatrix}$, $\sigma_1 = \begin{pmatrix}5&0\\0&5\end{pmatrix}$ *and* $\mu_2 = \begin{pmatrix}8\\11\end{pmatrix}$, $\sigma_2 = \begin{pmatrix}2&0\\0&2\end{pmatrix}$

We pick K=30 number of clients, C=20 number of dichotomy iterations, T = 1000 number of D-SGD iterations, $\forall t\ \eta_t = 0.1$, a batch size of 50, and we choose $\varepsilon = 0.05$ as the required level of fairness to be achieved by each client model. Heterogeneity level α is set to 0.3. We chose here Demographic Parity as a fairness metric.



(a) *each color represents the accuracy of a client over the global dataset*



(b) *Fairness score comparison between our estimate $\hat{\varphi}$ and the empirical fairness of the averaged model $\bar{\theta}$*

FIGURE 3 – single run experiment

In Figure 3a we achieve an average accuracy across clients of 78% with a standard deviation of 1.37%. In Figure 3b the maximum empirical fairness across clients by the end of the algorithm is 0.056 while the lowest achieved across dichotomy iterations is 0.044. We can also observe that adding a final averaging step for the clients models achieves a better fairness result. Figure 3b also shows that our use of $\hat{\varphi}$ is a good approximation for $\hat{f}$.

For comparison we run the unconstrained D-SGD ($\lambda = 0, \beta = 0$), and we get the following results : average accuracy across clients is 98.8% and the maximum fairness level across clients is 0.47. The averaged unconstrained model also has a level of fairness of 0.46.

6.2 EXPERIMENTS

We provide the results of our experiments in the following tables :

Dataset		Synthetic		CelebA		ACSincome	
alpha		0.01	0.3	0.01	0.3	0.01	0.3
Accuracy	Linear unconstrained D-SGD	0.97 ± 0	0.98 ± 0	0.85 ± 0	0.85 ± 0	67 ± 0	67 ± 0
	NN unconstrained D-SGD	0.98 ± 0	0.97 ± 0	0.85 ± 0	0.85 ± 0	0.67 ± 0	0.66 ± 0
	Linear D-SearchFair	0.79 ± 0	0.79 ± 0	0.85 ± 0	0.85 ± 0	0.67 ± 0	0.66 ± 0
	NN D-SearchFair	0.76 ± 0	0.77 ± 0	0.85 ± 0	0.85 ± 0	0.66 ± 0	0.66 ± 0
	Linear D-SearchFairAVG	0.79 ± 0	0.77 ± 0	0.85 ± 0	0.85 ± 0	0.67 ± 0	0.66 ± 0
	NN D-SearchFairAVG	0.75 ± 0	0.76 ± 0	0.85 ± 0	0.85 ± 0	0.66 ± 0	0.66 ± 0
max |DP|	Linear unconstrained D-SGD	0.43 ± 0	0.41 ± 0	0.21 ± 0	0.22 ± 0	2e-2 ± 0	2e-2 ± 0
	NN unconstrained D-SGD	0.49 ± 0	0.48 ± 0	0.2 ± 0	0.22 ± 0	2e-2 ± 0	2e-2 ± 0
	Linear D-SearchFair	2.1e-2 ± 5e-3	2e-2 ± 5e-3	0.17 ± 0	0.22 ± 0	1e-2 ± 0	1e-2 ± 0
	NN D-SearchFair	3e-2 ± 1e-2	2e-2 ± 7e-3	0.18 ± 0	0.24 ± 0	1e-2 ± 0	1e-2 ± 0
	Linear D-SearchFairAVG	1.6e-2 ± 3e-3	1.5e-2 ± 2e-3	0.17 ± 0	0.22 ± 0	1e-2 ± 0	1e-2 ± 0
	NN D-SearchFairAVG	1.2e-2 ± 2e-3	9e-3 ± 1e-3	0.21 ± 0	0.19 ± 0	1e-2 ± 0	1e-2 ± 0

Dataset		Synthetic		CelebA		ACSincome	
alpha		0.01	0.3	0.01	0.3	0.01	0.3
Accuracy	Linear unconstrained D-SGD	0.97 ± 0	0.98 ± 0	0.85 ± 0	0.85 ± 0	0.66 ± 0	0.66 ± 0
	NN unconstrained D-SGD	0.98 ± 0	0.97 ± 0	0.85 ± 0	0.85 ± 0	0.66 ± 0	0.66 ± 0
	Linear D-SearchFair	0.94 ± 0	0.91 ± 0	0.85 ± 0	0.85 ± 0	0.67 ± 0	0.67 ± 0
	NN D-SearchFair	0.97 ± 0	0.94 ± 0	0.85 ± 0	0.85 ± 0	0.67 ± 0	0.66 ± 0
	Linear D-SearchFairAVG	0.95 ± 0	0.95 ± 0	0.85 ± 0	0.85 ± 0	0.67 ± 0	0.67 ± 0
	NN D-SearchFairAVG	0.95 ± 0	0.95 ± 0	0.85 ± 0	0.85 ± 0	0.66 ± 0	0.66 ± 0
max |EO|	Linear unconstrained D-SGD	0.13 ± 0	0.09 ± 0	0.19 ± 0	0.19 ± 0	2e-2 ± 0	2e-2 ± 0
	NN unconstrained D-SGD	0.1 ± 0.03	0.07 ± 0.02	0.17 ± 0	0.19 ± 0	2e-2 ± 0	2e-2 ± 0
	Linear D-SearchFair	2e-2 ± 7e-3	2.4e-2 ± 9e-3	0.2 ± 0	0.2 ± 0	1e-2 ± 0	1e-2 ± 0
	NN D-SearchFair	2e-2 ± 1e-2	7e-2 ± 1e-1	0.22 ± 0	0.16 ± 0	2e-2 ± 0	2e-2 ± 0
	Linear D-SearchFairAVG	1e-2 ± 3e-3	1.1e-2 ± 4e-3	0.2 ± 0	0.21 ± 0	1.7e-2 ± 0	1e-2 ± 0
	NN D-SearchFairAVG	1e-2 ± 4e-3	9e-3 ± 4e-3	0.22 ± 0	0.15 ± 0	2e-2 ± 0	2e-2 ± 0

Results We can see that on the synthetic dataset, both versions of the algorithm manage to reduce the empirical fairness level and maintain a useful accuracy (the obtained model isn't constant). On the other hand, the methods don't impact the value of fairness in the case of ACSincome dataset as the fairness difference is already near zero for the unconstrained classifier.

7 LIMITATIONS AND EXTENSIONS

7.1 No Accuracy guarantee

If we go back to the original problem, our goal was to optimize the following :

$$\text{minimize } \frac{1}{K}\sum_{i=1}^{K} l_i(\theta) \quad \text{s.t.} \quad |\hat{f}(\theta)| \leq \epsilon$$

But our algorithm only provides a way to ensure that each client obtains a fair model, and we didn't address the accuracy (the utility of the obtained models). Precisely, we obtain at the end a solution to the following problem :

$$\theta^*(\lambda^*, \beta) = \arg\min_{\theta} \frac{1}{K}\sum_{i=1}^{K}(l_i(\theta) + \lambda^* \tilde{f}_i(\theta) + \beta\Omega(\theta))$$

Where λ^* is the hyperparameter such that $\theta^*(\lambda^*, \beta)$ is a perfectly fair model.

It is apparent that the utility of the obtained models will mainly depend on the choice of the surrogate loss $\tilde{f}$, the hyperparameter β, and the final λ^* (which also depends indirectly on λ_- and λ_+).

7.2 Stability

Our main result (theorem 5.5) states that there is a certain number of iterations t_{gossip} for Gossip mean estimation and T for D-SGD such that Algorithm 3 will reach a fair model for each client after a certain number of iterations C. But in practice, we don't know the exact value of T to pick as it depends on some unknown problem constants (Lifshitz constant, theoretical noise with respect to true distributions, etc).

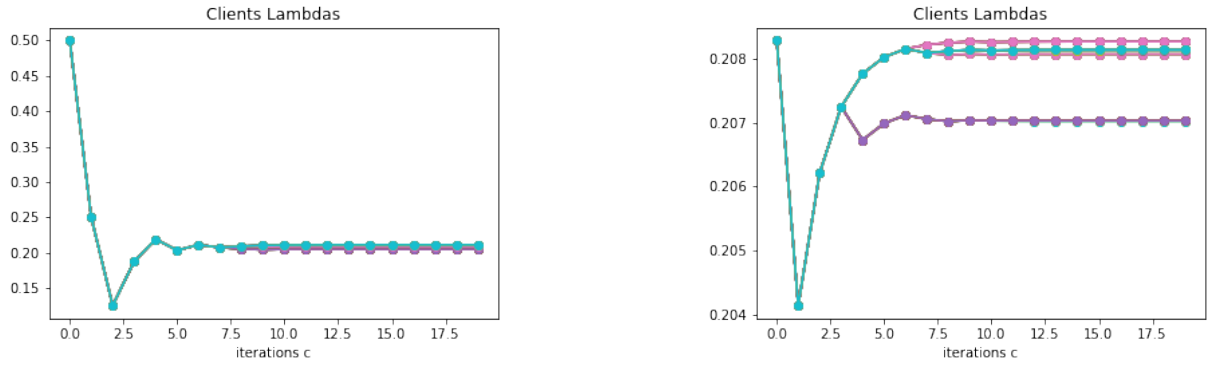
So in the practical implementation of the algorithm, and as with any machine learning hyperparameter choice, we often choose a "high enough" value for T and a certain number of iterations C.

The first instability arises if D-SGD didn't converge as expected and the clients parameters are far from the optimum (T not "high enough"), then we cannot guarantee that the sign of $\hat{\varphi}$ is the same as $\hat{f}(\theta^*)$, and D-SearchFair will not behave as expected

The second instability arises from the choice of C. Figure 4 illustrate the phenomenon : Figure 4a correspond to the single run experiment of section 6.1 where we can see that a slight difference in λ values occur after the 7th iteration of the algorithm. Figure 4b is for the same setup but starting with $\lambda_- = 0.2$ and $\lambda_+ \simeq 0.22$. Usually, the phenomenon shouldn't have a huge impact on the end result when it happens in the later iterations as the difference in values of λ will only be very small, but it might have an impact if it happens that we encounter λ^* right from the first iteration. Having different λ doesn't necessarily mean that the outputted model will not be fair, but it is outside our theoretical framework and can be a direction to investigate.

As in our proof of theorem 5.5, we can guarantee that all clients will have the same λ hyperparameter until $\exists i,\ s.t.\ |\varphi_i^{(t_{gossip})}| \leq \frac{\varepsilon}{4}$. But this would also mean that $\forall i \in [K],\ |\hat{f}(\theta_i)| \leq \varepsilon + 2m$ and that we can stop the algorithm before reaching C iterations. After this point, the clients might have different λ values, and we cannot guarantee that $|\hat{f}(\theta_i)|$ will remain less than $\varepsilon + m$ for further iterations.

The question is how do we stop a decentralized Algorithm ?



(a) λ *values*

(b) λ *values*

FIGURE 4 – *Algorithm 2 demonstration (apex on tropical segment* $[u, v]$*)*

One naive way to propagate a stop condition through the communication topology is by having the clients compute an additional mean using Gossip mean estimation from Section 3.4, where each client will compute the following sum at each iteration c of the dichotomy :

$$x_i^{(t)} = GossipMean(x_1^{(0)}, \ldots, x_K^{(0)}) \text{ Where : } x_i^{(0)} \in \{0, 1\}$$

This works by having all clients start with a 0, so the mean will always be 0 (regardless of the number of gossip iterations), and when a client i encounters a stop condition he picks $x_i^{(0)} = 1$ so that the mean estimate is now positive for all clients, after a sufficient number of gossip iterations of the same order as the longest path in the communication topology. Once the clients will all have found, a strictly positive estimate of the mean, they stop.

This is merely a naive way of doing it, and it opens the possibility of adversarial attacks when a malicious client can decide to make everyone stop whenever he wants even if the stop condition hasn't been reached.

7.3 Fairness limitation

The first fairness limitation that faces our approaches is the fact that we can only consider binary-sensitive attributes, limiting the use cases of our method. One possible research direction is to expand the framework of SearchFair to account for multiple λ and establish the existence of a zero in the higher dimension. Another possibility is to extend the work done in [19] as it allows multiple sensitive attributes to be considered, especially since a Federated version of the algorithm already exists [22].

The second limitation is the number of fairness metrics definitions the framework supports. In fact, we can only support one equality constraint, which is the case for Demographic Parity and Equality of Opportunity but not the case for other used metrics such as Equalized Odds.

8 CONCLUSION

In this paper, we provide a decentralized version of the SearchFair algorithm. We derive theoretical guarantees on the fairness level obtained by the practical implementation of the algorithm. To the best of our knowledge, our work is the first to tackle the problem of group fairness in a decentralized setting. We also explored the impact of the topology on fairness and hope that our preliminary result may give a better understanding on a possible improvement based on the choice of the Topology.

RÉFÉRENCES

[1] Alekh Agarwal et al. “A reductions approach to fair classification”. In : *International conference on machine learning*. PMLR. 2018, p. 60-69.

[2] Henry C Bendekgey et Erik Sudderth. “Scalable and stable surrogates for flexible classifiers with fairness constraints”. In : *Advances in Neural Information Processing Systems* 34 (2021), p. 30023-30036.

[3] Stephen Boyd et al. “Randomized gossip algorithms”. In : *IEEE transactions on information theory* 52.6 (2006), p. 2508-2530.

[4] Toon Calders et Sicco Verwer. “Three naive bayes approaches for discrimination-free classification”. In : *Data mining and knowledge discovery* 21 (2010), p. 277-292.

[5] Simon Caton et Christian Haas. “Fairness in machine learning : A survey”. In : *arXiv preprint arXiv :2010.04053* (2020).

[6] Zheyi Chen et al. “A fairness-aware peer-to-peer decentralized learning framework with heterogeneous devices”. In : *Future Internet* 14.5 (2022), p. 138.

[7] Lingyang Chu et al. “Fedfair : Training fair models in cross-silo federated learning”. In : *arXiv preprint arXiv :2109.05662* (2021).

[8] Frances Ding et al. “Retiring Adult : New Datasets for Fair Machine Learning”. In : *Advances in Neural Information Processing Systems* 34 (2021).

[9] Yahya H Ezzeldin et al. “Fairfed : Enabling group fairness in federated learning”. In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 37. 6. 2023, p. 7494-7502.

[10] Max Hort et al. “Bia mitigation for machine learning classifiers : A comprehensive survey”. In : *arXiv preprint arXiv :2207.07068* (2022).

[11] Peter Kairouz et al. “Advances and open problems in federated learning”. In : *Foundations and Trends® in Machine Learning* 14.1–2 (2021), p. 1-210.

[12] Anastasia Koloskova et al. “A unified theory of decentralized sgd with changing topology and local updates”. In : *International Conference on Machine Learning*. PMLR. 2020, p. 5381-5393.

[13] Batiste Le Bars et al. “Refined convergence and topology learning for decentralized sgd with heterogeneous data”. In : *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, p. 1672-1702.

[14] Ziwei Liu et al. “Deep learning face attributes in the wild”. In : *Proceedings of the IEEE international conference on computer vision*. 2015, p. 3730-3738.

[15] Michael Lohaus, Michael Perrot et Ulrike Von Luxburg. “Too relaxed to be fair”. In : *International Conference on Machine Learning*. PMLR. 2020, p. 6360-6369.

[16] Gaurav Maheshwari et Michaël Perrot. “Fairgrad : Fairness aware gradient descent”. In : *arXiv preprint arXiv :2206.10923* (2022).

[17] OpenAI. “GPT-4 Technical Report”. In : (mars 2023). arXiv : 2303.08774 [cs.CL].

[18] ProPublica. *Machine Bias*. 2016. url : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visité le 26/03/2023).

[19] Yuji Roh et al. “Fairbatch : Batch selection for model fairness”. In : *arXiv preprint arXiv :2012.01696* (2020).

[20] Wikipedia. *COMPAS (software)*. url : https://en.wikipedia.org/wiki/COMPAS_(software) (visité le 26/03/2023).

[21] Blake Woodworth et al. “Learning non-discriminatory predictors”. In : *Conference on Learning Theory*. PMLR. 2017, p. 1920-1953.

[22] Yuchen Zeng, Hongxu Chen et Kangwook Lee. “Improving fairness via federated learning”. In : *arXiv preprint arXiv :2110.15545* (2021).

APPENDIX

Datasets processing

We follow the data processing in [15] for the CelebA dataset, and we consider "Male" (-1 = female, 1 = male) as the sensitive attribute. We drop the image_id column and only consider the 40 numerical columns where the values are either -1 or 1. We pick a random number of data points for each client using a clipped normal distribution centered around 1000 with 100 standard deviation.

We use the attribute "SEX" as the sensitive attribute (-1 for male and 1 for female), and the predicted label Y is whether the person has an income higher than 50k a year (Y=1) or not (Y=-1). We pick a random number of data points for each client using a clipped normal distribution centered around 1000 with 100 standard deviation.

Proof of lemmas

Proof of lemma 5.2 We use here the fact that, $h_\theta(x) = \Phi(x)^T\theta$

$$\begin{aligned}
|\hat{f}(\theta) - \hat{\varphi}(\theta_1, \dots, \theta_K)| &= \left| \sum_{i=1}^{K} \sum_{(x,s,y)\in\hat{D}_i} (\mathbb{I}_{h_\theta(x)\geq 0} - \mathbb{I}_{h_{\theta_i}(x)\geq 0}) \left(\frac{\mathbb{I}_{s=1}}{N_1} - \frac{\mathbb{I}_{s=-1}}{N_{-1}} \right) \right| \\
&\leq \sum_{i=1}^{K} \sum_{(x,s,y)\in\hat{D}_i} \left| \mathbb{I}_{h_\theta(x)\geq 0} - \mathbb{I}_{h_{\theta_i}(x)\geq 0} \right| \left(\frac{\mathbb{I}_{s=1}}{N_1} + \frac{\mathbb{I}_{s=-1}}{N_{-1}} \right) \\
&\leq \max\left(\frac{1}{N_1}, \frac{1}{N_{-1}}\right) \sum_{i=1}^{K} \sum_{(x,s,y)\in\hat{D}_i} \left| \mathbb{I}_{h_\theta(x)\geq 0} - \mathbb{I}_{h_{\theta_i}(x)\geq 0} \right|
\end{aligned}$$

and we can write that : $\mathbb{I}_{h_{\theta_i}(x)\geq 0} = \mathbb{I}_{h_\theta(x)+h_{\theta_i}(x)-h_\theta(x)\geq 0}$

$$\begin{aligned}
\mathbb{I}_{h_{\theta_i}(x)\geq 0} &= \mathbb{I}_{h_\theta(x)\geq h_{\theta-\theta_i}(x)} \\
&\leq \mathbb{I}_{h_\theta(x)\geq -\|\Phi(x)\|\|\theta-\theta_i\|} \\
&\leq \mathbb{I}_{h_\theta(x)\geq 0} + \mathbb{I}_{-\|\Phi(x)\|\|\theta-\theta_i\|\leq h_\theta(x)<0}
\end{aligned}$$

and therefore : $\mathbb{I}_{h_{\theta_i}(x)\geq 0} - \mathbb{I}_{h_\theta(x)\geq 0} \leq \mathbb{I}_{-\|\Phi(x)\|\|\theta-\theta_i\|\leq h_\theta(x)<0}$

$$\mathbb{I}_{h_{\theta_i}(x)\geq 0} - \mathbb{I}_{h_\theta(x)\geq 0} \leq \mathbb{I}_{\left\{\frac{\|\Phi(x)\|\|\theta-\theta_i\|}{|h_\theta(x)|}>=1\right\}} \leq \frac{\|\Phi(x)\|\|\theta-\theta_i\|}{|h_\theta(x)|}$$

Similarly, we prove that :

$$\begin{aligned}
\mathbb{I}_{h_\theta(x)\geq 0} - \mathbb{I}_{h_{\theta_i}(x)\geq 0} = \mathbb{I}_{h_{\theta_i}(x)\leq 0} - \mathbb{I}_{h_\theta(x)\leq 0} &\leq \mathbb{I}_{0<h_\theta(x)\leq\|\Phi(x)\|\|\theta-\theta_i\|} \\
\mathbb{I}_{h_\theta(x)\geq 0} - \mathbb{I}_{h_{\theta_i}(x)\geq 0} &\leq \frac{\|\Phi(x)\|\|\theta-\theta_i\|}{|h_\theta(x)|}
\end{aligned}$$

Finally :

$$\left| \mathbb{I}_{h_\theta(x)\geq 0} - \mathbb{I}_{h_{\theta_i}(x)\geq 0} \right| \leq \frac{\|\Phi(x)\|\|\theta-\theta_i\|}{|h_\theta(x)|}$$

Proof of lemma 5.3 This is an immediate consequence of Boyd's result, lemma 3.3 in section 3.4, as we have that :

$$\hat{\varphi}(\theta_1^{(0)}, \dots, \theta_K^{(0)}) = GossipMean(\varphi_1(\theta_1^{(0)}), \dots, \varphi_K(\theta_K^{(0)}))$$

Proof of lemma 5.4 We denote by $(\lambda_-^{(0)}, \lambda_+^{(0)}) \supset (\lambda_-^{(1)}, \lambda_+^{(1)}) \supset \dots$ the sequence of intervals obtained by the dichotomy algorithm on $\hat{f}$ such that :

$$\hat{f}(\theta^*(\lambda_-^{(t)}, \beta)) \leq 0 \text{ and } \hat{f}(\theta^*(\lambda_+^{(t)}, \beta) \geq 0$$

As $\hat{f}$ takes a finite number of values in $\mathbb{Q}$ and the sequence of intervals is infinite, there exists a number of iterations t_0 such that all the subsequent values of $\hat{f}(\theta^*(\lambda_-^{(t)}, \beta))$ and $\hat{f}(\theta^*(\lambda_+^{(t)}, \beta)$ are stationary for $t \geq t_0$.

We denote by λ_0 the common limit of $\lambda_-^{(t)}$ and $\lambda_+^{(t)}$. We have that $\hat{f}(\theta^*(\lambda_0, \beta)) \in \{\hat{f}(\theta^*(\lambda_-^{(t_0)}, \beta), \hat{f}(\theta^*(\lambda_+^{(t_0)}, \beta)\}$ and without loss of generality we assume that $\hat{f}(\theta^*(\lambda_0, \beta)) = \hat{f}(\theta^*(\lambda_+^{(t_0)}, \beta)$.

With assumption 6 and by using Lemma 5.2 with $\theta = \theta^*(\lambda_0, \beta)$ and $\theta_1 = \dots = \theta_K = \theta^*(\lambda_-^{(C)}, \beta)$ with $C \geq t_0$ and :

$$|\hat{f}(\theta^*(\lambda_0, \beta)) - \hat{\varphi}(\theta^*(\lambda_-^{(C)}, \beta), \dots, \theta^*(\lambda_-^{(C)}, \beta))| \leq \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \left(\sum_{i=1}^{K} \|\theta^*(\lambda_-^{(C)}, \beta) - \theta^*(\lambda_0, \beta))\| \sum_{\substack{(x,s,y)\in \hat{D}_i \\ h_\theta(x)\neq 0}} \frac{\|\Phi(x)\|}{|h_\theta(x)|} + m \right)$$

$$\Rightarrow |\hat{f}(\theta^*(\lambda_+^{(t_0)}, \beta) - \hat{f}(\theta^*(\lambda_-^{(C)}, \beta))| \leq \max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \left(\sum_{i=1}^{K} \|\theta^*(\lambda_-^{(C)}, \beta) - \theta^*(\lambda_0, \beta))\| \sum_{\substack{(x,s,y)\in \hat{D}_i \\ h_\theta(x)\neq 0}} \frac{\|\Phi(x)\|}{|h_\theta(x)|} + m \right)$$

And we pick C sufficiently large as to have :

$$\max(\frac{1}{N_1}, \frac{1}{N_{-1}}) \sum_{i=1}^{K} \|\theta^*(\lambda_-^{(C)}, \beta) - \theta^*(\lambda_0, \beta))\| \sum_{\substack{(x,s,y)\in \hat{D}_i \\ h_\theta(x)\neq 0}} \frac{\|\Phi(x)\|}{|h_\theta(x)|} \leq m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

$$\Rightarrow |\hat{f}(\theta^*(\lambda_+^{(t_0)}, \beta) - \hat{f}(\theta^*(\lambda_-^{(C)}, \beta))| \leq 2m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

And as $\hat{f}(\theta^*(\lambda_+^{(C)}, \beta) = \hat{f}(\theta^*(\lambda_+^{(t_0)}, \beta) \geq 0$ and $\hat{f}(\theta^*(\lambda_-^{(C)}, \beta) \leq 0$ we have that :

$$\min(|\hat{f}(\theta^*(\lambda_-^{(C)}, \beta)|, |\hat{f}(\theta^*(\lambda_+^{(C)}, \beta)|) \leq m \times \max(\frac{1}{N_1}, \frac{1}{N_{-1}})$$

Proof of examples in 4.3

L-smoothness of smooth Hinge loss .
1^{st} **method** we can compute :

$$\nabla^2 L_{sh}(h_\theta(x), y) = \begin{cases} 0 & \text{if } h_\theta(x)y \leq 0 \\ \Phi(x)\Phi(x)^T & \text{if } 0 < h_\theta(x)y < 1 \\ 0 & \text{if } h_\theta(x)y \geq 1 \end{cases}$$

and therefore $\|\nabla^2 L_{sh}(h_\theta(x), y)\|_{op} \leq \|\Phi(x)\|^2$ which means that ∇L_{sh} is $\|\Phi(x)\|^2$-Lipschitz (where we used the operator-norm for the second derivative)
2^{nd} **method** Let θ and $\omega \in \mathbb{R}^d$, we verify all the 6 cases :

— if $h_\theta(x)y$ and $h_\omega(x)y$ are in the same case he have that $\|\nabla L_F(\theta, Z) - \nabla L_F(\omega, Z)\| \leq \|\Phi(x)\|^2 \|\theta - \omega\|$

— if $h_\theta(x)y \leq 0$ and $0 < h_\omega(x)y < 1$ then :

$$\begin{aligned} h_\omega(x)y &\leq h_\omega(x)y - h_\theta(x)y \\ y\Phi(x)^T\omega &\leq y\Phi(x)^T\omega - y\Phi(x)^T\theta \\ |\Phi(x)^T\omega| &\leq |\Phi(x)^T(\omega-\theta)| \\ |\Phi(x)^T\omega| &\leq \|\Phi(x)\|\|\omega-\theta\| \\ \|\nabla L_F(\theta,Z) - \nabla L_F(\omega,Z)\| &= \\ |\Phi(x)^T\omega|\|\Phi(x)\| &\leq \|\Phi(x)\|^2\|\omega-\theta\| \end{aligned}$$

— Other cases are done in the same manner

L-smoothness of logistic loss .
Let a : x $\to \frac{1}{1+\exp(x)}$
we have that $a'(x) = \frac{\exp(x)}{(1+\exp(x))^2} \Rightarrow |a'(x)| \leq \frac{1}{4}$, and therefore a is $\frac{1}{4}$-Lipschitz
Then :

$$\begin{aligned} \|\nabla L_F(\theta,Z) - \nabla L_F(\omega,Z)\| &\leq \|\Phi(x)\||a(yh_\theta(x)) - a(yh_\omega(x))| \\ \|\nabla L_F(\theta,Z) - \nabla L_F(\omega,Z)\| &\leq \frac{\|\Phi(x)\|^2}{4}\|\theta-\omega\| \end{aligned}$$

And therefore L_F is $\frac{\|\Phi(x)\|^2}{4}$-Lipschitz.

Convexity of logistic loss .
Let $f : x \to \log(1+\exp(x))$ and $g_y : \theta \to -y\Phi(x)^T\theta$, we have that :

$$f'(x) = \frac{\exp(x)}{1+\exp(x)} \geq 0 \quad \text{and} \quad f''(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} \geq 0$$

meaning that f is convex and increasing, and as g is linear in θ he have that $L_F(\theta, Z) = f(g_y(\theta))$ is convex in θ

Proof of Theorem 5.7

We pick c a constant in $\mathbb{N}$ such that :

$$\frac{\sqrt{K}}{c} \leq \varepsilon$$

Since p_i are the empirical proportions, then $\forall i \in [K],\ p_i \in \mathbb{Q}$
We choose our dataset such that :

$$\forall i,j \in [K], \quad p_i n_i = p_j n_j = c$$

Therefore : $N_1 = \sum_{i=1}^K N_{1,i} = \sum_{i=1}^K p_i n_i = Kc$

We know that : $\hat{\varphi}_i^{(0)}(\theta) = \frac{K}{N_1}\sum_{(x,y,z)\in\hat{D}_i} \mathbb{I}_{h_\theta(x)\geq 0}\mathbb{I}_{s=1} - \frac{K}{N_{-1}}\sum_{(x,y,z)\in\hat{D}_i} \mathbb{I}_{h_\theta(x)\geq 0}\mathbb{I}_{s=-1}$

We denote by : $l_i^+(\theta) = \sum_{(x,y,z)\in \hat{D}_i} \mathbb{I}_{h_\theta(x)\geq 0}\mathbb{I}_{s=1} \in \{0, 1, \ldots, c\}$

So that : $\frac{K}{N_1} l_i^+(\theta) = \frac{l_i^+(\theta)}{c} \in \{0, \frac{1}{c}, \ldots, 1\}$

In order to get the slow convergence speed, we need that :

$$(\hat{\varphi}_1^{(t)}, \ldots, \hat{\varphi}_K^{(t)})^T - \hat{\varphi}\mathbf{1} = v^{(t)} - \hat{\varphi}\mathbf{1} \in Span(u_2)$$

Where u_2 is the spectral vector associated with the second-largest spectral value of W.

* So we will look to have : $v^{(t)} = \frac{1}{2}u_2 + \frac{1}{2}\mathbf{1}$

Since we fixed the model θ available to each client, we know the model decision boundary. We therefore can pick a dataset for each client i, to have $\hat{\varphi}_i^{(0)}(\theta)$ be of any value we want in $\{0, \frac{1}{c}, \ldots, 1\}$, in the following manner : if we want, $\hat{\varphi}_i^{(0)}(\theta) = \frac{2}{c}$ then we choose 2 data points on the positive side of the decision boundary ($h_\theta(x) \geq 0$). We then pick c-2 data points on the negative side of the decision boundary ($h_\theta(x) \leq 0$). We put all the above data points in the sensitive group S=1. We then add $\frac{(1-p_i)}{p_i}c$ data points on the negative side of the decision boundary of the sensitive group S=-1 to obtain the right percentage p_i.

Now we choose $\hat{\varphi}_i^{(0)}(\theta)$ to be the closest value possible to $\frac{1}{2}u_{2,i} + \frac{1}{2} \in [0, 1]$ $(\|u_2\| = 1)$, where $u_{2,i}$ is the i_th coordinate of u_2. Therefor we have :

$$|\hat{\varphi}_i^{(0)}(\theta) - \frac{1}{2}u_{2,i} - \frac{1}{2}| \leq \frac{1}{2c}$$

* if $u_{2,i} \in \mathbb{Q}$ we can have an exact estimate by adjusting the value of c, we will in the following consider the general case of $u_{2,i} \in \mathbb{R}$.

We can write :

$$\begin{aligned}
\hat{\varphi}_i^{(0)}(\theta) &= \frac{1}{2}u_{2,i} + \frac{1}{2} + \varepsilon_i \quad (|\varepsilon_i| \leq \frac{1}{2c}) \\
\Rightarrow v^{(0)} &= \frac{1}{2}u_2 + \frac{1}{2}\mathbf{1} + \varepsilon \quad (\varepsilon = (\varepsilon_1, \ldots, \varepsilon_K)^T) \\
\Rightarrow \hat{\varphi} &= \frac{1}{2} + \bar{\varepsilon} \quad (u_2 \perp \mathbf{1}) \\
\Rightarrow v^{(0)} - \hat{\varphi} &= \frac{1}{2}u_2 + \varepsilon - \bar{\varepsilon}\mathbf{1} \simeq \frac{1}{2}u_2
\end{aligned}$$

We have that :

$$\begin{aligned}
v^{(t)} - \hat{\varphi}\mathbf{1} &= W^t(v^{(0)} - \hat{\varphi}\mathbf{1}) \\
&= \lambda_2^t \times \frac{1}{2}u_2 + W^t(\varepsilon - \bar{\varepsilon}\mathbf{1}) \\
&= \lambda_2^t(v^{(0)} - \hat{\varphi}\mathbf{1}) + (W^t - \lambda_2^t I)(\varepsilon - \bar{\varepsilon}\mathbf{1}) \quad (I \text{ is the identity matrix})
\end{aligned}$$

We therefore obtain the final inequality :

$$\begin{aligned}
\|v^{(t)} - \hat{\varphi}\mathbf{1}\| &\geq \lambda_2^t\|v^{(0)} - \hat{\varphi}\mathbf{1}\| - \|(W^t - \lambda_2^t I)(\varepsilon - \bar{\varepsilon}\mathbf{1})\| \\
&\geq \lambda_2^t\|v^{(0)} - \hat{\varphi}\mathbf{1}\| - \max(1 - \lambda_2^t, \lambda_2^t - \lambda_K^t)\|\varepsilon - \bar{\varepsilon}\mathbf{1}\| \\
&\geq \lambda_2^t\|v^{(0)} - \hat{\varphi}\mathbf{1}\| - \frac{\sqrt{K}}{c} \\
\|v^{(t)} - \hat{\varphi}\mathbf{1}\| &\geq \lambda_2^t\|v^{(0)} - \hat{\varphi}\mathbf{1}\| - \varepsilon
\end{aligned}$$