Crowdsourcing with Fairness, Diversity and Budget Constraints

Naman Goel, Boi Faltings
naman.goel@epfl.ch, boi.faltings@epfl.ch
Artificial Intelligence Lab, EPFL
Lausanne, Switzerland

Abstract

Recent studies have shown that the labels collected from crowdworkers can be discriminatory with respect to sensitive attributes such as gender and race. This raises questions about the suitability of using crowdsourced data for further use, such as for training machine learning algorithms. In this work, we address the problem of fair and diverse data collection from a crowd under budget constraints. We propose a novel algorithm which maximizes the expected accuracy of the collected data, while ensuring that the errors satisfy desired notions of fairness. We provide guarantees on the performance of our algorithm and show that the algorithm performs well in practice through experiments on a real dataset.

1 Introduction

Algorithmic decision-making is gaining popularity in many diverse application areas of social importance. Examples include criminal recidivism prediction, stop-and-frisk programs, university admissions, bank loan decisions, screening job candidates, fake news control, information filtering(personalization) and search engine rankings etc. Recently, questions were raised about the fairness of these algorithms. An investigation, led by (ProPublica 2017), found COMPAS (a popular software used by courts to predict criminal recidivism risk) racially discriminatory. Other software systems have also been found to be biased against people of different races, genders and political views (Kay, Matuszek, and Munson 2015; Bolukbasi et al. 2016; Otterbacher, Bates, and Clough 2017; Kulshrestha et al. 2017). This has led to a widespread and legitimate concern about the potential negative influence of such systems on the society (Barocas and Selbst 2016; White House 2016). One of the main reasons of algorithmic bias is the bias in the training datasets. In order to achieve algorithmic fairness, the issue of **data fairness** needs to be addressed first. In many interesting cases, data is directly or indirectly influenced by some kind of human feedback. The influence is obvious and direct if human assigned labels are used as a proxy for ground truth labels. However, human feedback can also indirectly influence the so-called "ground-truth" datasets (when the labels are not human assigned but observed in reality). This is because the ground truth labels can only be collected for a finite number of data points and the selection of data points is often influenced by humans. For example, there are no ground truth labels for recidivism of people who were never released by the judges. In this paper, we focus only on the direct influence of human feedback on data fairness i.e. the case in which humans assign labels for data.

Crowdsourcing is increasingly used to collect training data labels. Inevitably, crowdworkers have different biases, which are then reflected in the labels collected from the workers. A very recent study (Dressel and Farid 2018) conducted on Amazon Mechanical Turk showed that the crowdworkers were equally racially biased as COMPAS in predicting recidivism. The difference in false positive rates of crowd predictions for white and black defendants was significant and nearly equal to that of the predictions made by COMPAS. The same was true for false negative rates also. The bias didn't change much even when the crowdworkers were not explicitly displayed the race of the defendants.

We consider settings similar to (Dressel and Farid 2018). Workers are asked to provide their answers (or labels) about some tasks with unknown ground truth labels. Every task has some non-sensitive details that are shown to the workers and a sensitive attribute (for example, race) that is not explicitly shown. But the sensitive attribute may potentially be correlated with the non-sensitive task details. A worker inspects the tasks assigned to her and submits labels for the tasks. Each task is assumed to have a ground truth label but the workers don't have any way of accessing the ground truth. They can only use the task details, their prior knowledge and incomplete information from other sources to make an "educated guess" about the ground truth. The examples of such tasks are "Will a defendant with given personal history recidivate within the next two years or not?" or "Will a candidate with given CV be successful in the job applied for?" or "Is given political news item fake?". The sensitive attributes in these example tasks are race, gender and political group respectively. Every worker charges a fee for answering the assigned tasks. The requester has a budget constraint on the amount of fees that she can pay to the workers. In this paper, we make the following contributions:

1. We propose a novel algorithm for assigning tasks to the workers, which optimizes the expected accuracy of labels obtained from crowd while ensuring that the collected la-

bels satisfy desired notions of error fairness. The algorithm also ensures diversity of responses by limiting the probability of assigning many tasks to a single worker. Our algorithm works even when the values of the sensitive attribute of the tasks are unavailable or can't used because of ethical/legal reasons.

2. With a novel formulation of the task assignment strategy as a probability distribution over the workers, we can cast the optimization problem as a linear program and avoid the use of integer programming or other graph matching algorithms which are popular in the task assignment literature but are harder to solve exactly and analyze. This also makes our algorithm suitable for online settings in which the requester is not aware of the tasks in advance.
3. We use a limited number of gold tasks (tasks with known ground-truth answers) for estimating workers' parameters and then optimally assign non-gold tasks to the workers. We provide performance bounds for our algorithm and show empirical performance on a real dataset.

2 Related Work

Empirical Studies : (Dressel and Farid 2018) find racial discrimination in recidivism prediction tasks on Amazon Mechanical Turk (AMT). (Otterbacher 2015a) analyzes linguistic bias in labels collected through GWAP (Games with a Purpose) on AMT. (Otterbacher 2015b) analyzes the linguistic bias in collaboratively produced biographies. (Hannák et al. 2017) find discrimination in reputation crowdsourcing systems in online marketplaces.

Proposed Solutions : In an independent and pioneering work, (Valera, Singla, and Rodriguez 2018) consider the problem of fairness in human decision-making tasks like recidivism prediction, without budget and diversity constraints. Criminal cases with *known* race information arrive in batches of *known* sizes and an MDP based maximum weighted matching algorithm assigns each case to *exactly one* human judge such that the overall utility from decisions of releasing or keeping any defendant is maximized, while ensuring *demographic parity* of release decisions across two races. To the best of our knowledge this is the first and the most recent work to consider settings somewhat similar to ours but our work differs from theirs in several ways. We consider general crowdsourcing settings, in which several assumptions from their model don't hold. In particular, they assume that "true" risk scores of individual defendants are known to the human judges and the case assignment algorithm. In general crowdsourcing settings, one can only hope to have an overall label distribution for the population. In fact, finding the label probability for individual tasks is the very objective of crowdsourcing. Further, it is not immediately clear how their work can be extended for other important fairness definitions. In their model, given true risk scores of the defendants, judges only apply different thresholds for black and white defendants to predict recidivism. The threshold parameters alone can't capture unfairness measures such as unequal error rates. Even if one does improvise the model with more parameters, it remains an open question whether the theoretical conjectures made in the paper are still likely. This is because the conjectures assume that every time a judge gives a decision, the model parameters of the judge are updated. This becomes an issue with error rate parameters since the ground truth labels are not revealed for all tasks in crowdsourcing. (Neel and Roth 2018) consider a different but related problem of bias resulting from adaptive data gathering (when the choice of whether to collect more data of a given type depends on the data already collected) and propose a differentially private data collection process as a solution.

There is also a lot of work on task assignment in crowdsourcing, which doesn't consider fairness but is related to our work. (Tran-Thanh et al. 2014) propose a greedy knapsack approach to satisfy limits on budget and the number of tasks any worker can solve. (Karger, Oh, and Shah 2014; Ho and Vaughan 2012; Ho, Jabbari, and Vaughan 2013) consider task assignment problem when workers arrive online. (Bragg, Weld, and others 2016) propose optimal gold task assignment when workers' diligence change over time.

Beyond data collection, there is also a lot of recent work on making algorithms fair and robust to bias in the training data (Dwork et al. 2012; Hardt et al. 2016; Zafar et al. 2017; Kusner et al. 2017; Kleinberg, Mullainathan, and Raghavan 2017) and on correcting bias in training datasets (Feldman et al. 2015; Calmon et al. 2017). Correcting bias in a given dataset requires modifying the feature values and/or the labels in the dataset. In this paper, we aim to collect unbiased dataset to begin with, relaxing the responsibility and the overhead of such post-processing from data users (for e.g., data scientists and machine learning engineers).

We note that several classic AI papers consider fairness in different applications such as resource allocation (Bertsimas, Farias, and Trichakis 2011) and kidney exchange (Dickerson, Procaccia, and Sandholm 2014). Our work vaguely resembles them in the sense that we also propose a constrained optimization framework to balance fairness-utility tradeoff; but the nature of utility, fairness and application constraints are entirely different.

3 Model

Let there be a finite set of n workers and a large pool of tasks with unknown ground truth labels. The data requester randomly chooses tasks from the pool one by one and assigns each to one (or more) worker(s). The requester may not have knowledge of all the tasks in the pool (not even the number of tasks in the pool) in advance. A worker i charges a constant amount of fee c_i for every label she provides. The requester has a budget constraint for the maximum *expected* money to be spent on acquiring one label from a worker.

Let Z be a random variable denoting the sensitive attribute and Y denoting the (unknown) ground truth labels of the tasks such that $Z, Y \in \{0, 1\}$. For the tasks attempted by a worker i, let $\hat{Y}_i \in \{0, 1\}$ denote the labels submitted by the worker. We denote the realizations of random variables Z, Y and $\hat{Y}_i$ by lower case letters z, y and $\hat{y}_i$ respectively and will drop the subscripts for brevity when the context is clear. We will use $[n]$ to denote $\{1, 2, \ldots, n\}$. The workers are modeled using their accuracy matrices as follows:

Definition 1 (Accuracy Matrices of a Worker)**.** *The accuracy matrices $\mathcal{A}_{iz},\ z \in \{0,1\}$ of a worker i are two 2×2 row stochastic matrices such that, $\forall y, \hat{y_i} \in \{0,1\}$, the entry $\mathcal{A}_{iz}[y, \hat{y_i}]$ is the probability of the worker's label on a task being $\hat{y_i}$ given that the sensitive attribute of the task is z and the ground truth label is y.*

The two matrices $\mathcal{A}_{i0}$ and $\mathcal{A}_{i1}$ define the accuracy of the worker i for tasks belonging to the two different values of the sensitive attribute. The accuracy matrix model, also known as the Dawid-Skene model (Dawid and Skene 1979) in crowdsourcing literature, is strong enough to capture different errors (for e.g. false positive and false negative rates) that a worker may make for tasks belonging to a given sensitive attribute value. If a worker is unbiased in the sense that her errors don't depend on the value of sensitive attribute of the task, her two accuracy matrices are identical.

The requester uses a probabilistic policy to assign the tasks to workers and collects the labels from the workers.

Definition 2 (Crowdsourcing Policy)**.** *A crowdsourcing policy is an n-dimensional stochastic vector S, such that an element $S[i], i \in [n]$ is the probability of assigning any task to worker i, regardless of the sensitive attribute value of the task.*

Note that the requester's policy doesn't depend on the value of the sensitive attribute of the task. This is a thoughtful modeling choice to deal with the situations in which the sensitive attribute values of the tasks may not be available. It may be due to missing data, privacy reasons or legal/ethical requirements of not using the sensitive attribute.

For any task, the requester randomly selects one (or more than one) worker(s) with probabilities specified by the crowdsourcing policy vector S and assigns the task to the selected worker(s). The labels collected from the workers are obviously not guaranteed to be error free. We can define the accuracy matrices of the crowdsourcing policy in the same way as we defined the accuracy matrices of workers.

Definition 3 (Accuracy Matrices of a Crowdsourcing Policy)**.** *The accuracy matrices $\mathcal{A}_z,\ z \in \{0,1\}$ of a crowdsourcing policy are two 2×2 row stochastic matrices such that, $\forall y, \hat{y} \in \{0,1\}$, the entry $\mathcal{A}_z[y, \hat{y}]$ is the probability that a crowdsourced label for a task[1] is $\hat{y}$ given that the sensitive attribute of the task is z and the ground truth label is y.*

We use the letter $\mathcal{A}$ to denote accuracy matrices of crowdsourcing policy and of workers but readers can differentiate between the two by noting that $\mathcal{A}$ has an additional subscript i when referring to the matrix of a worker i. It is easy to see that we can express the accuracy matrices of a policy in terms of the accuracy matrices of the workers as follows:

$$\mathcal{A}_z = \sum_{i=1}^{n} S[i] \cdot \mathcal{A}_{iz} \qquad ,\forall z \in \{0,1\} \tag{1}$$

[1]We note that the accuracy of a crowdsourcing policy can also be defined in terms of aggregated label when multiple labels per task are collected. But such definitions depend on specific label aggregation algorithms used. It is sufficient in our case to assume that the accuracy of a policy with aggregated labels is an increasing function of this accuracy, which is a reasonable assumption.

The requester is interested in finding a crowdsourcing policy that maximizes the expected accuracy of the collected labels while ensuring that the data is *fair*, *diverse* and is acquired within budget constraints.

Crowd diversity is a subjective property and is generally defined in terms of the demographics of crowdworkers. In this paper, we work with a given set of crowdworkers and can't control such a measure of diversity. For settings like these, we define diversity as follows:

Definition 4 (β-Diverse Crowdsourcing Policy)**.** *A crowdsourcing policy is called β-diverse if and only if $\forall\ i \in [n], S[i]$ is upper bounded by β, where β is a diversity parameter such that $0 \leq \beta < 1$.*

This definition limits the influence of individual workers on the overall crowdsourced dataset and aims to distribute the influence across more workers.

Similar to diversity, fairness is also a subjective property. We use some standard definitions of fairness from the machine learning literature (Hardt et al. 2016; Zafar et al. 2017; Barocas, Hardt, and Narayanan 2018).

Definition 5 (False Positive Rate Parity)**.** *A crowdsourcing policy, with accuracy matrices $\mathcal{A}_0$ and $\mathcal{A}_1$, is said to satisfy false positive rate parity if and only*

$$\mathcal{A}_0[0,1] = \mathcal{A}_1[0,1]$$

One can similarly define false negative rate parity, which requires $\mathcal{A}_0[1,0] = \mathcal{A}_1[1,0]$.

Definition 6 (Error Rate Parity)**.** *A crowdsourcing policy, with accuracy matrices $\mathcal{A}_0$ and $\mathcal{A}_1$, is said to satisfy error rate parity if and only if it satisfies false positive rate parity and false negative rate parity, i.e.*

$$\mathcal{A}_0 = \mathcal{A}_1$$

It is easy to see that if all workers are unbiased, any crowdsourcing policy satisfies the above fairness definitions and one only need to select a policy that maximizes accuracy while satisfying budget and diversity constraints. In this paper, we address the general problem scenario (when workers are not necessarily unbiased).

4 Finding Optimal Crowdsourcing Policy

Let's first assume that the accuracy matrices of all the workers are known and the requester is interested in finding the optimal crowdsourcing policy maximizing the expected accuracy under budget, fairness and diversity constraints. We model this as a constrained optimization problem. The objective function in the minimization problem is the negative of the expected accuracy of the policy variable S:

$$-\mathbb{E}[\mathcal{A}(S)] = \\ -\sum_{z\in\{0,1\}} P(Z=z) \sum_{y\in\{0,1\}} P_z(Y=y) \sum_{i=1}^{n} S[i]\mathcal{A}_{iz}[y,y] \tag{2}$$

where $P(Z = z)$ is the known prior probability that any random task in the pool will have sensitive attribute value equal to z and $P_z(Y = y)$ is the known prior probability that any random task with sensitive attribute value z in the pool will have a ground truth label equal to y.

Together with the fairness and diversity constraints, we get the following optimization problem:

$$\begin{aligned}
\arg\min_{S} \quad & -\sum_{z\in\{0,1\}} P(Z=z)\sum_{y\in\{0,1\}} P_z(Y=y)\sum_{i=1}^{n} S[i]\mathcal{A}_{iz}[y,y] \\
\text{subject to} \quad & \sum_{i=1}^{n} S[i] = 1 \\
& S[i] \geq 0 \quad ,\forall i \in [n] \\
& S[i] \leq \beta \quad ,\forall i \in [n] \\
& \mathcal{A}_0[0,1] - \mathcal{A}_1[0,1] \leq \alpha \\
& -(\mathcal{A}_0[0,1] - \mathcal{A}_1[0,1]) \leq \alpha \\
& \sum_{i=1}^{n} S[i]\cdot c_i \leq C
\end{aligned} \tag{3}$$

The first two constraints are due to the fact that the crowdsourcing policy vectors are probabilistic and so, all elements must be positive and sum to 1. The third is the diversity constraint as formalized in Definition 4. The forth and fifth constraints together are equivalent to $\left|\mathcal{A}_0[0,1] - \mathcal{A}_1[0,1]\right| \leq \alpha$. For $\alpha = 0$, we get the exact fairness constraint (false positive rate parity) as formalized in Definition 5. Other fairness constraints can also be similarly included. The last constraint is due to the maximum expected budget (C) that can be spent on acquiring one answer from a worker.

Estimates of Worker Accuracy Matrices

Until now, we assumed that the accuracy matrices of the workers are known. However, in practice, we need to estimate them. As is common in the literature (Oleson et al. 2011), we assume that the requester has some limited number of gold standard tasks. Gold tasks are the tasks for which the requester not only knows the sensitive attribute value z but also the ground truth label y. We use gold tasks to estimate unknown worker accuracy matrices. Estimating all the entries of the worker accuracy matrices requires that every worker answers some gold tasks of each "type" (the type of a task is specified by its ground truth answer and its sensitive attribute value). We assign N_g tasks of every type to each worker to estimate their accuracy matrices. The estimation process is explained in the appendix. Let $\hat{\mathcal{A}}_{iz}$ be the estimate of the worker accuracy matrices $\mathcal{A}_{iz},\ \forall z \in \{0,1\}$. The optimization problem 3 can now be written as follows, by replacing the accuracy matrices with their estimates:

$$\begin{aligned}
\arg\min_{S} \quad & -\sum_{z\in\{0,1\}} P(Z=z)\sum_{y\in\{0,1\}} P_z(Y=y)\sum_{i=1}^{n} S[i]\hat{\mathcal{A}}_{iz}[y,y] \\
\text{subject to} \quad & \sum_{i=1}^{n} S[i] = 1 \\
& S[i] \geq 0 \quad ,\forall i \in [n] \\
& S[i] \leq \beta \quad ,\forall i \in [n] \\
& \hat{\mathcal{A}}_0[0,1] - \hat{\mathcal{A}}_1[0,1] \leq \alpha \\
& -(\hat{\mathcal{A}}_0[0,1] - \hat{\mathcal{A}}_1[0,1]) \leq \alpha \\
& \sum_{i=1}^{n} S[i]\cdot c_i \leq C
\end{aligned} \tag{4}$$

where,

$$\hat{\mathcal{A}}_z = \sum_{i=1}^{n} S[i]\cdot\hat{\mathcal{A}}_{iz} \qquad ,\forall z \in \{0,1\} \tag{5}$$

This is a linear program, which can be exactly solved in polynomial time. In practice, the simplex method (Chvatal 1983) can be used to find the optimal solution efficiently with common optimization libraries like IBM CPLEX and SciPy. Depending on the constraints, the cost and the accuracy matrices of workers, it is possible that no feasible solution exists for the optimization problem. In this case, the requester will have no choice but to relax the constraints.

We will now analyze our algorithm theoretically and empirically. Readers can find a summary of steps of our complete crowdsourcing algorithm in the appendix.

5 Theoretical Analysis

When worker accuracy matrices are known, our method is guaranteed to provide the optimal solution, satisfying constraints. However, when estimates of the accuracy matrices are used, two interesting questions arise:

1. Does the solution of problem 4 (which is optimal and satisfies fairness constraints only according to the estimated accuracy parameters) also satisfy fairness in reality?
2. How much does the requester lose in terms of actual expected accuracy of the policy because of using the estimated accuracy parameters in optimization?

Theorem 1. *With probability at least γ, the solution $\hat{S}$ to the optimization problem 4 satisfies*

$$\left|\mathcal{A}_0[0,1] - \mathcal{A}_1[0,1]\right| \leq \alpha + \delta$$

where

$$\delta = 2\sqrt{\frac{-\ln\left(1 - \sqrt[2n]{\gamma}\right) + \ln 2}{2N_g}}; \mathcal{A}_z = \sum_{i=1}^{n} \hat{S}[i]\mathcal{A}_{iz}, \forall z \in \{0,1\}$$

and N_g is number of gold tasks.

The theorem states that when we use estimates of the worker accuracy matrices instead of the real matrices, the obtained solution $\hat{S}$ doesn't violate the fairness constraints in reality by more than δ, with probability at least γ.

Theorem 2. *Assuming that the optimal solution $\hat{S}$ of problem 4 satisfies fairness constraints of problem 3 and the optimal solution S of problem 3 satisfies fairness constraints of problem 4, then with probability at least γ'*

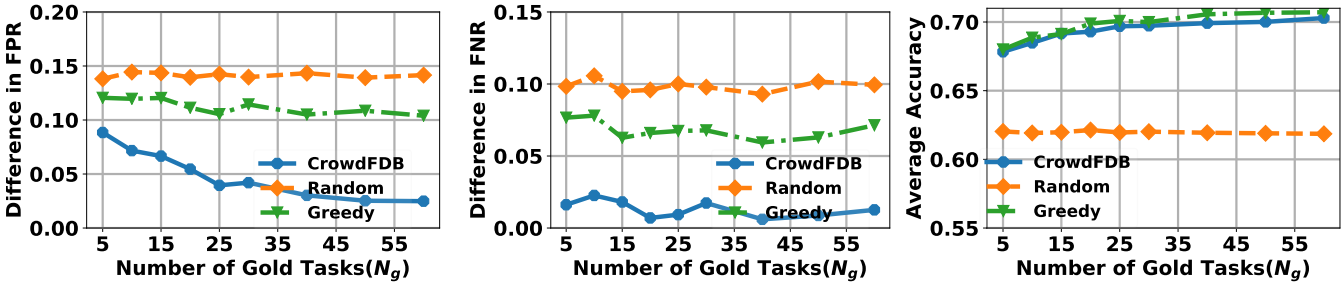
$$\mathbb{E}[\mathcal{A}(S)] - \mathbb{E}[\mathcal{A}(\hat{S})] \leq 2n\beta\sqrt{\frac{-\ln\left(1 - \sqrt[4n]{\gamma'}\right) + \ln 2}{2N_g}}$$

where

$$\mathbb{E}[\mathcal{A}(S)] = \sum_{z\in\{0,1\}} P(Z=z)\sum_{y\in\{0,1\}} P_z(Y=y)\sum_{i=1}^{n} S[i]\mathcal{A}_{iz}[y,y],$$

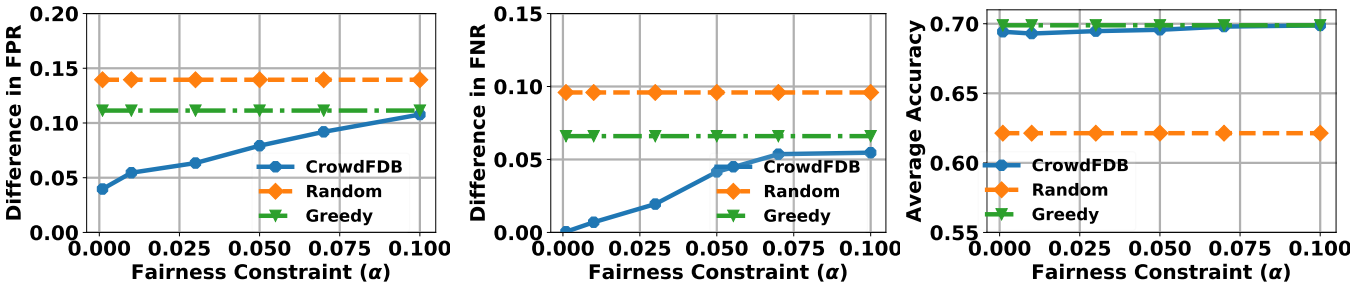
$$\mathbb{E}[\mathcal{A}(\hat{S})] = \sum_{z\in\{0,1\}} P(Z=z)\sum_{y\in\{0,1\}} P_z(Y=y)\sum_{i=1}^{n} \hat{S}[i]\mathcal{A}_{iz}[y,y]$$

The theorem provides an upper bound on the loss in real expected accuracy of the crowdsourcing policy, when we use the estimated worker matrices instead of the real accuracy matrices for optimization. Note that in both the theorems, the bounds get better with increasing number of gold tasks.



(a) Absolute Difference in FPR (b) Absolute Difference in FNR (c) Average Accuracy

Figure 1: Varying N_g (Number of gold tasks), Settings : Uniform Costs, $\beta = 0.01, \alpha = 0.01$



(a) Absolute Difference in FPR (b) Absolute Difference in FNR (c) Average Accuracy

Figure 2: Varying α (Fairness Constraint), Settings : Uniform Costs, $\beta = 0.01, N_g = 20$

6 Experimental Evaluation

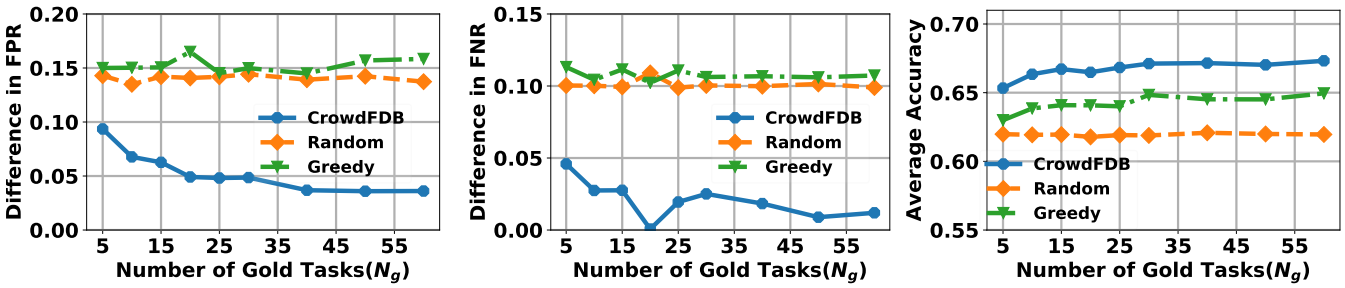
Datasets We use the following datasets in our experiments.

1. **Broward County Dataset** (ProPublica 2017) : This dataset contains information about 7214 defendants arrested in Broward County, Florida between 2013 and 2014. The information includes race of defendants among other non-sensitive attributes such as age, prior charges etc. The dataset also contains ground-truth whether the defendants recidivated within 2 years or not. There are 3696 black defendants and 2454 white defendants in the dataset and the base rate of recidivism is 51.43% among black defendants and 39.36% among white defendants .
2. **Crowd Judgment Dataset** : (Dressel and Farid 2018) randomly selected a subset of 1000 defendants from the Broward County dataset and asked 20 random workers on Amazon Mechanical Turk to predict recidivism for each individual. In total, 400 workers participated in their study and each worker submitted answers for 50 different defendants. The dataset contains these crowd answers.

Experiment Outline The idea is to split the set of defendants into two sets. The first set acts as the gold standard set, which we use to estimate worker accuracy matrices. Once we have the estimates of the worker accuracy matrices, we can solve the optimization problem 4 and learn optimal crowdsourcing policy. We then pick non-gold defendants one by one and assign it to one of the 400 workers, randomly selected according to the policy. The workers' responses are then compared with the ground-truth label to evaluate fairness and accuracy of our crowdsourcing policy.

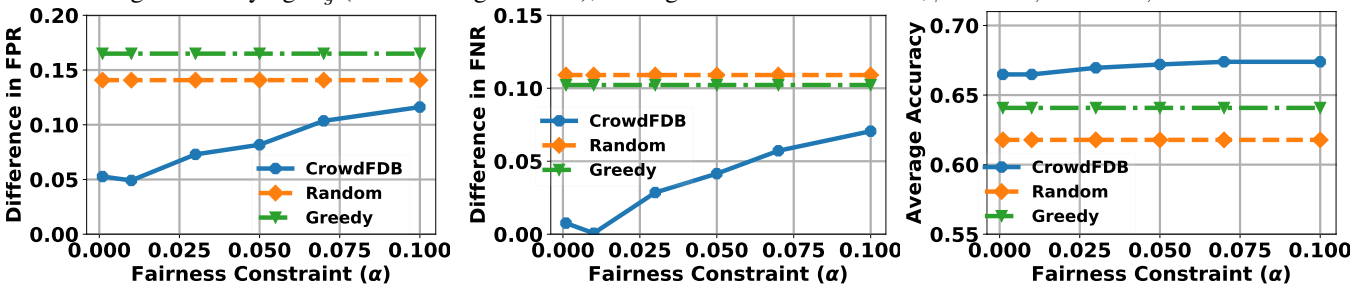
Handling Limitations of Datasets Unfortunately, none of the two datasets alone can be used for such experiment. The Broward County dataset contains ground truth labels but doesn't contain workers' answers. On the other hand, the Crowd Judgment dataset does contain worker answers but is very limited for the following reasons. In this dataset, tasks have already been assigned (randomly) to workers and for every defendant, we have responses of only a subset of 20 workers out of all 400 workers. If the crowdsourcing policy learned by our algorithm decides that a worker outside that subset of 20 workers should be assigned a task, then we will need to know the answer of that worker but the answer of this worker is not part of the dataset. The second reason is that every worker has submitted answers for 50 defendants, which is sufficient for getting good estimates of the accuracy parameters of the workers but not big enough to be further split into gold and non-gold sets.

To overcome these limitations, we first create a bigger synthetic dataset using the two real datasets as follows. We generate synthetic answers of all the 400 workers for all the 3696 black and 2454 white defendants in the Broward County dataset. The answers are generated using the worker accuracy parameters estimated from the entire Crowd Judgment dataset. Note that even though this is a synthetic dataset but none of the parameters of the dataset are synthetic. The worker accuracy parameters are derived from the entire real dataset of (Dressel and Farid 2018) and the base dataset (Broward County dataset) is used as it is. In other words, there is no parameter in this dataset generation process, which can be tuned to favor any algorithm.



(a) Absolute Difference in FPR (b) Absolute Difference in FNR (c) Average Accuracy

Figure 3: Varying N_g (Number of gold tasks), Settings : Non-Uniform Costs, $\beta = 0.01, \alpha = 0.01, C = 1.5$



(a) Absolute Difference in FPR (b) Absolute Difference in FNR (c) Average Accuracy

Figure 4: Varying α (Fairness Constraint), Settings : Non-Uniform Costs, $\beta = 0.01, N_g = 20, C = 1.5$

Worker Costs : The datasets also don't contain workers' costs. We create this information in two different ways. In the first setting, we associate a uniform cost of \$1 to each worker. In the second and more interesting setting, we probabilistically associate a cost of \$1 or \$3. The probability of a worker's cost being \$3 is equal to her average accuracy and of it being \$1 is equal to $1-$her average accuracy. Thus, the higher the average accuracy of a worker, the higher is the probability that she will charge a cost of \$3.

Now this complete dataset is ready to be used in the experiment outlined earlier in this section. The dataset will also be made public for reproducibility. We now compare our approach (called 'CrowdFDB' in the figures) with two baselines (called 'Random' and 'Greedy' (Tran-Thanh et al. 2014)). The baselines are described in the appendix.

Observations

Parameter β was set to 0.01 in all experiments. We use equal error rate parity (Definition 6) as the desired fairness. All results reported in the paper are averages over 100 repeated runs. In the uniform costs settings, C was set to \$1 and in non-uniform settings, $C = \$1.5$.

Uniform Costs In Figure 1, we keep the fairness constraint α to be fixed (0.01) and observe the effect of increasing number of gold tasks (N_g). Figures 1a and 1b show that as we increase N_g, the fairness i.e. the absolute difference in FPR (and FNR) for black and white populations, gets closer and closer to α. In other words, the δ of Theorem 1 gets closer to 0 as expected. Moreover, the margin between our algorithm and the baselines also increases. However, meeting the fairness constraints alone is not enough. This could also be done by a bad algorithm that collects equally wrong labels for both white and black populations. Hence, accuracy of the collected labels is also an important measure. Figure 1c shows that our algorithm has an accuracy competitive to the Greedy baseline method, which is a highly efficient baseline in the literature for accuracy optimization. Our algorithm can achieve same level of accuracy while also providing fairness. In Figure 2, we keep N_g fixed (20) and observe the effect of increasing value of α. As value of α increases, the fairness constraints are more relaxed and the algorithm can obtain better accuracy.

Non-Uniform Costs In the non-uniform costs settings, we observe similar patterns in Figure 3 and 4. There are a few notable differences. The accuracy of our algorithm as well as the Greedy baseline are lower. Our algorithm doesn't select more accurate workers because of budget constraints and the Greedy baseline also finds the density of the more accurate workers comparatively lower due to their higher costs and prefers to choose other high density workers. In this case, our algorithm beats Greedy in not just fairness but also in accuracy by better utilizing the available budget.

Some more experimental results, further attesting the above trends, are discussed in the appendix.

7 Conclusions and Future Work

In this paper, we addressed the problem of data fairness in crowdsourcing. We proposed a novel crowdsourcing algorithm that learns an optimal sampling probability distribution over the available set of workers to maximize the expected accuracy of collected data, while ensuring that the

errors in the data are not unfairly discriminatory towards any particular social group. When a limited number of gold tasks are used to estimate worker tasks, we provide bounds on the performance of our algorithm. Experimental analysis further confirms the performance under different parameter settings.

While this is an important step towards achieving data fairness in crowdsourcing, there also remain many challenges to be addressed. One particular challenge is to define data fairness for the case of subjective tasks, which have no ground truth labels and thus, no clear notion of errors. Ensuring fairness in subjective data collection is also likely to create a challenging problem of lying incentives for workers.

References

Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review 671*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2018. Fairness and machine learning : Limitations and opportunities.

Bertsimas, D.; Farias, V. F.; and Trichakis, N. 2011. The price of fairness. *Operations research*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

Bragg, J.; Weld, D. S.; et al. 2016. Optimal testing for crowd workers. In *International Conference on Autonomous Agents & Multiagent Systems*.

Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*.

Chvatal, V. 1983. *Linear programming*. Macmillan.

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* 20–28.

Dickerson, J. P.; Procaccia, A. D.; and Sandholm, T. 2014. Price of fairness in kidney exchange. In *international conference on Autonomous agents and multi-agent systems*.

Dressel, J., and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *3rd Innovations in Theoretical Computer Science Conference*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Hannák, A.; Wagner, C.; Garcia, D.; Mislove, A.; Strohmaier, M.; and Wilson, C. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *CSCW*, 1914–1933.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*.

Ho, C.-J., and Vaughan, J. W. 2012. Online task assignment in crowdsourcing markets. In *AAAI Conference*.

Ho, C.-J.; Jabbari, S.; and Vaughan, J. W. 2013. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning*.

Karger, D. R.; Oh, S.; and Shah, D. 2014. Budget-optimal task allocation for reliable crowdsourcing systems. *Opeations Research* 1–24.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *33rd Annual ACM Conference on Human Factors in Computing Systems*.

Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2017. Inherent trade-offs in the fair determination of risk scores. In *8th Conf. on Innovations in Theoretical Computer Science*.

Kulshrestha, J.; Eslami, M.; Messias, J.; Zafar, M. B.; Ghosh, S.; Gummadi, K. P.; and Karahalios, K. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*.

Neel, S., and Roth, A. 2018. Mitigating bias in adaptive data gathering via differential privacy. In *35th International Conference on Machine Learning*.

Oleson, D.; Sorokin, A.; Laughlin, G.; Hester, V.; Le, J.; and Biewald, L. 2011. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *AAAI Workshop on Human Computation*.

Otterbacher, J.; Bates, J.; and Clough, P. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *CHI Conference on Human Factors in Computing Systems*.

Otterbacher, J. 2015a. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *33rd Annual ACM Conference on Human Factors in Computing Systems*.

Otterbacher, J. 2015b. Linguistic bias in collaboratively produced biographies: Crowdsourcing social stereotypes? In *International AAAI Conference on Web and Social Media*.

ProPublica. 2017. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

Tran-Thanh, L.; Stein, S.; Rogers, A.; and Jennings, N. R. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*.

Valera, I.; Singla, A.; and Rodriguez, M. G. 2018. Enhancing the accuracy and fairness of human decision making. *arXiv preprint arXiv:1805.10318*.

White House. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.