On the Compatibility of Privacy and Fairness

Rachel Cummings*
Varun Gupta*
Dhamma Kimpara*
Jamie Morgenstern*

ABSTRACT

In this work, we investigate whether privacy and fairness can be simultaneously achieved by a single classifier in several different models. Some of the earliest work on fairness in algorithm design defined fairness as a guarantee of similar outputs for "similar" input data, a notion with tight technical connections to differential privacy. We consider whether tensions exist between differential privacy and statistical notions of fairness, namely Equality of False Positives and Equality of False Negatives (EFP/EFN). We show that even under full distributional access, there are cases where the constraint of differential privacy precludes exact EFP/EFN. We then turn to asking whether one can learn a differentially private classifier which approximately satisfies EFP/EFN, and show the existence of a PAC learner which is private and approximately fair with high probability. We then conclude with an discussion of several techniques for achieving privacy and fairness, describing which of the former and latter may be (in)compatible.

ACM Reference Format:
Rachel Cummings[1], Varun Gupta[1], Dhamma Kimpara[1], and Jamie Morgenstern[1]. 2018. On the Compatibility of Privacy and Fairness. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent applications of machine learning in human-relevant domains raise concerns that too much of some individual's information might be leaked through a model learned on some training data. For example, if one learns a model based on historical health data, many algorithms have some real possibility of outputting a model containing information about individuals' HIV status or other sensitive information. Resultingly, academia and industry have spent much effort designing and implementing *differentially private* machine learning methods.

Differential privacy gives a strong guarantee to individuals whose data we use to train a model; these human-centric uses of ML systems have also raised concerns of equity of the predictive power for a model on different populations. The *fairness* of a model can be thought of as some equitable performance guarantee for individuals who will be evaluated by the model, rather than a guarantee for someone participating in the model's training process. When phrasing privacy and fairness of a model this way, a natural question arises: in what settings can we learn a model which is private in the training data while guaranteeing equitable performance for multiple populations on which the model will ultimately run? More formally, will it be possible to guarantee privacy of training data, fairness for the predictions made on two populations, and some accuracy overall?

[Jamie: More stuff here]

1.1 Related Work

The focus on fairness in machine learning and its relationship to differential privacy was explored in early work in the community (cite fairness through awareness). This work introduced the concept of treating similar people similarly, where "similarity" is defined as some task-specific metric over individuals. The authors then point out that this desiderata can be formulated as a Lipschitz constraint and show how to satisfy it using tools from DP.

More recently, (citet the paper from fat) raised questions of whether statistical notions of equitable predictive power, such as equalized odds (cite moritz), are compatible with privacy. In a limiting sense, when one talks about feature and model selection, there appears to be some tension here: an additional feature might increase the possible privacy loss an individual faces, while the additional feature should only make EO easier to satisfy (as alluded to in Hanna's paper on exploration and exploitation). However, to the best of the authors' knowledge, no work has heretofore shown whether an ML algorithm can guarantee individuals differential privacy and their populations some kind of group fairness simultaneously.

The technical tools we use for this work come from differential price vacy, using the exponential mechanism (cite), private SGD for our empirical results, and FILL IN ANY OTHER TOOLS WE USE.

Related work we definitely want to cover: fairness through awareness, the paper proposing we study privacy + fairness, dwork's "it isn't private and it isn't fair" or whatever...

1.2 Our Contributions

We present two contributions to the study of the intersection of differential privacy and fairness in the binary classification setting. First, we show that it is impossible to achieve both differential privacy and exact fairness even in the setting where we have access to the full distribution of the data. We then consider a notion of approximate fairness in the finite sample access setting and show that there exists a private PAC learner that is differentially private and satisfies approximate fairness with high probability. Finally, we present some incompatibility results concerning current practical differentially private and fairness-ensuring learning algorithms (in-process methods).

*Georgia Institute of Technology. Email: {rachelc, vgupta74, dkimpara1, jamiemmt.cs}@gatech.edu.

Conference'17, July 2017, Washington, DC, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 PRELIMINARIES

Let $\mathcal{X}$ be our data universe consisting of elements $z = (x, a, y)$ where x is the features, a the protected (binary) attribute, and y the binary response. Note that x may have arbitrary correlation with a, including containing a copy of it. $\mathcal{A}(x)$ is the probability distribution over outputs of a randomized algorithm $\mathcal{A}$ on input x.

2.1 Differential Privacy

For our results, we use two notions of a database:

(1) a distribution D over $\mathcal{X}$;
(2) a finite sample, vector $Z = (z_1, ..., z_n)$ drawn i.i.d. from a distribution D over $\mathcal{X}$.

For each notion, there is an accompanying notion of a neighboring database. We present them here in respective order.

For notion (1), we adapt the definition of neighboring databases in the continuous case from [4]:

Definition 2.1. (ζ-closeness). Distributions D and D' taking values in $\mathcal{X}$ are ζ-close if the statistical distance between their distributions is at most ζ, i.e.,

$$||D - D'||_{SD} := \frac{1}{2} \sum_{d \in \mathcal{X}} |\Pr[D = d] - \Pr[D' = d]| \leq \zeta.$$

Neighboring databases are ζ-close distributions.

For notion (2), we define neighboring samples as samples that differ in one entry.

Definition 2.2. [2] Samples Z and Z' are neighboring if $z_i \neq z'_i$ for exactly one $i \in [n]$ (i.e. the Hamming distance between Z and Z' is 1).

Alternatively we could define a finite sample as a multi-set, and use the symmetric distance instead of the Hamming metric to measure distance. Note that Definition 2.1 is a generalization of Definition 2.2.

For each of these notions, the definition of differential privacy remains the same. That is, a randomized algorithm is private if neighboring databases induce close distributions on its outcomes:

Definition 2.3. (ϵ-differential privacy [1]) A randomized algorithm $\mathcal{A}$ is ϵ-differentially private if for all pairs of neighboring databases D, D' and for all sets $\mathcal{S}$ of outputs,

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{A}(D') \in \mathcal{S}].$$

[DK: need to talk about composition/post-processing theorems and/or exponenetial mechanism?]

2.2 Exact Fairness

In this setting our database notion is (1): a distribution D over X. This corresponds with the full distributional access setting of [cite eq of opp hardt] where we have access to the joint probability distribution $D = (X, A, Y)$ over $\mathcal{X}$, where X are the features, A sensitive attributes, and Y true labels, respectively. Hence we state our notion of exact fairness in these terms:

Definition 2.4. (Equal Opportunity [3]). We say that a binary predictor h satisfies equal opportunity with respect to A and Y if

$$\Pr\{h = 1 | A = 1, Y = 1\} = \Pr\{h = 1 | A = 0, Y = 1\}.$$

Note that this is the weaker of the two definitions of fairness in Hardt et. al. however, it is sufficient to prove impossibility for this notion since violating this notion implies violating the stronger notion, equal odds. These measures of fairness belong to a broader class of fairness constraints that is, constraints on functions of the confusion matrix of a classifier (convex constraints paper).

2.3 Approximate Fairness

We now turn to the approximate fairness setting with and define analogous definitions from the exact fairness setting. Here our notion of a database is (2) i.e. we have access to a finite sample Z consisting of entries drawn i.i.d. from an arbitrary distribution D over $\mathcal{X}$.

For ease of notation we define $Z_{ya} := \{z_i \in Z | y_i = y, a_i = a\}$.

We also have an analogous definition for the group-conditional true positive rates: [DK: clarify summation as such?: $\sum_{z_i \in Z_{ya}} h(x_i)$]

$$\gamma^Z_{ya}(h) = \frac{1}{|Z_{ya}|} \sum_{Z_{ya}} h(x).$$

Definition 2.5. (α-discrimination [5]). We say that a binary predictor h is α-discriminatory with respect to a binary protected attribute A on the population or on a sample Z if, respectively:

$$\Gamma(h) := \max_{y \in \{0,1\}} |\gamma_{y0}(h) - \gamma_{y1}(h)| \leq \alpha$$

or

$$\Gamma^Z(h) := \max_{y \in \{0,1\}} |\gamma^Z_{y0}(h) - \gamma^Z_{y1}(h)| \leq \alpha.$$

When $\alpha = 0$ we are in the exact fairness setting. Note that by [5] the sample fairness measure Γ^Z converges to the population measure Γ when n is large. For the analog of Definition 2.4 (Equality of Opportunity) where we only consider true positive rates, the definition reduces to:

$$\Gamma^Z(h) := |\gamma^Z_{10}(h) - \gamma^Z_{11}(h)| \leq \alpha.$$

In the remainder of this paper, Γ refers to this fairness measure.

[DK: talk about motivation for approx fairness because of finite sample (cannot achieve exact) and now also because of impossibility with privacy.]

2.4 Preliminaries from learning theory

[DK: do we need to factor a weight λ into the new error function? ie. $\ell + \lambda\Gamma$ talk in end of proof] [DK: very not sure how to write this section, help?] A concept is a function that labels examples taken from the domain X by the elements of the range Y. A *concept class* is a set of concepts. The domain and range in $\mathcal{C}$ are ensembles $X = \{X_d\}_{d \in \mathbb{N}}$, $Y = \{Y_d\}_{d \in \mathbb{N}}$ where the representation size of the elements in X_d, Y_d is at most d. Since we focus on binary classification problems, d measures the size of examples in X_d. When the size parameter is clear from the context or not important, we omit the subscript d.

In our particular setting, our distribution of labeled examples is arbitrary on $X_d \times Y_d$ and hence we consider the agnostic setting that removes the realizability assumption. The goal of a learner then is to output a hypothesis $h \in \mathcal{H}$ whose error with respect to

the distribution is close to the optimal possible by a function from C. The misclassification error of h on D is defined as:

$$err(h) = \Pr_{(x,y)\tilde{D}}[h(x) \neq y].$$

In our setting we consider the agnostic learnability setting. [DK: not sure if following def is correct (wrt. concept class C_d do we consider C in the agnostic setting?)]

Definition 2.6. (Agnostic PAC Learning) A concept class C over $\mathcal{X}$ is PAC learnable using hypothesis class $\mathcal{H}$ if there exists an algorithm $\mathcal{A}$ and a polynomial $poly(\cdot,\cdot)$ such that for all $d \in \mathbb{N}$, all concepts $c \in C_d$, all distributions D on $\mathcal{X}$ and all $\alpha, \beta \in (0, 1/2)$, given inputs α, β, and $Z = (z_1, \ldots z_n)$, where n = $poly(d, 1/\alpha, \log(1/\beta))$, $z_i = (x_i, a_i, y_i)$ drawn i.i.d. from D for $i \in [n]$, algorithm $\mathcal{A}$ outpus a hypothesis $h \in \mathcal{H}$ satisfying

$$\Pr[err(h) \leq OPT + \alpha] \geq 1 - \beta.$$

[DK: insert note about distribution-free learning]

2.5 Private and Approximately Fair Agnostic Learning

We define private and approximately fair PAC learners as algorithms that satisfy the definitions of differential privacy, approximate fairness (with high probability), and PAC learning.

Definition 2.7. (Private and Approximately Fair Agnostic PAC Learning) Let d, α, β be as in Definition 2.6 (agnostic PAC learnigng) and $\epsilon > 0$. Concept class C is (inefficiently) privately and approximately fair agnostically learnable using hypothesis class $\mathcal{H}$ if there exists an algorithm $\mathcal{A}$ that takes inputs $\epsilon, \alpha, \beta, Z$, where $|Z| = n$ is polynomial in $1/\epsilon, d, 1\alpha, \log(1/\beta)$ and satisfies

(1) [Fairness] Algorithm $\mathcal{A}$ satisfies $\Pr[\Gamma(h) \leq OPT + \alpha] \geq 1 - \beta$;
(2) [Privacy] For all $\epsilon > 0$, algorithm $\mathcal{A}(\epsilon, \cdot, \cdot, \cdot)$ is ϵ-differentially private;
(3) [Utility] Algorithm $\mathcal{A}$ agnostically PAC learns C using H.

3 ACHIEVING EXACT FAIRNESS AND DIFFERENTIAL PRIVACY IS IMPOSSIBLE

We start with the full distributional access setting (which can be translated to the finite sample setting later). Let $\mathcal{X}$ be the data universe with elements $d_i = (x_i, a_i, y_i) \in \mathcal{X}$ where each entry consists respectively of the response, features, and protected attribute. Note that x_i may have arbitrary correlation with a_i, including containing a copy of it.

[DK: actually we could do an impossibility result for any hard threshold for Γ in any setting. Hence motivating approx fairness with high probability]

Definition 3.1. For ease of notation we define

$$\mathcal{X}_{ya} := \{d_i \in \mathcal{X} | y_i = y, a_i = a\}.$$

Definition 3.2. (non-trivial hypothesis class). We say that a hypothesis class $\mathcal{H}$ is *non-trivial* if there exists $d_1, d_0 \in \mathcal{X}$ such that for some $h \in \mathcal{H}$, $h(x_1) > h(x_0)$.

Lemma 3.1. *Let $\mathcal{H}$ be a non-trivial hypothesis class. Then releasing an exactly fair hypothesis $h \in \mathcal{H}$ in a differentially private manner is impossible.*

[DK: (currently in weaker eq. of opp. fairness notion, can also be done for stronger notions.) TODO: general proof of impossibility given a 'reasonable' fairness constraint on the confusion matrix? Seems like it would be difficult since need to construct neighboring databases. Hence TODO: proof for eq. odds? or say easy extension into other 'reasonable' fairness notions.]

Proof. We will construct distributions D, D' that are ζ-close and show impossibility. Suppose that the fair (as in Definition 2) hypothesis h has a non-zero probability of being released by algorithm $\mathcal{A}$ on input of D. Let $d(\cdot)$ be the measure as induced by distribution D.

Pick $d_1, d_0 \in \mathcal{X}_{11}$ such that $h(x_1) > h(x_0)$ and assume that $d_1, d_0 \in supp(D)$ and $supp(D')$. Let $\zeta < \min_{i,D,D'}(d_D(d_i))$. D' be a database with ζ more probability mass of d_1 and ζ less mass of d_0. Then recalling that $d_i = (x_i, a_i, y_i)$:

$$\Pr_{D'}\{h(x) = 1 | a = 1, y = 1\} = \int_{\mathcal{X}_{11}} h(x)dx + \frac{\zeta}{d(\mathcal{X}_{11})}h(x_1) - \frac{\zeta}{d(\mathcal{X}_{11})}h(x_0)$$

$$= \Pr_{D}\{h(x) = 1 | a = 1, y = 1\} + \frac{\zeta}{d(\mathcal{X}_{11})}(h(x_1) - h(x_0))$$

$$> \Pr_{D}\{h(x) = 1 | a = 1, y = 1\}$$

where the last inequality is due to the assumption that $h(x_1) > h(x_0)$.

Since we have not modified the distributions over X_{10}, $\Pr_{D'}\{h(x) = 1 | a = 1, y = 1\} = \Pr_D\{h(x) = 1 | a = 1, y = 1\}$. Hence hypothesis h is unfair (in the sense of definition 2.4) when evaluated on distribution D' and cannot be released by the exactly fair algorithm $\mathcal{A}$ on input of database D'. So

$$\Pr\{\mathcal{A}(D) = h\} \not\leq \exp(\epsilon)\Pr\{\mathcal{A}(D') = h\} = 0$$

.

which violates the constraint of privacy in Definition 1. □

This result also applies to more general notions of statistical fairness. Definition 2.4, Equal Opportunity, is a strictly weaker notion of fairness than Equal Odds also defined in [DK: cite]. In other words, Equal Opportunity is a necessary condition for Equal Odds.

We can also use the same proof method to show impossibility of fairness and privacy when we consider notions of approximate fairness such as disparate impact and mean difference scores. Approximate in these notions of fairness meaning that the resulting classifier (or anything released by the randomized algorithm) satisfies some hard threshold on the 'level of discrimination' on the distribution. One needs to construct neighboring distribution where when considering a hypothesis $h \in \mathcal{H}$, h is fair on one distribution but not the other. [DK: use 'database' or 'distribution' terminology?]

Lemma 3.2. *(ϵ, ζ)-DP and fairness is impossible. [VG: need proof?]*

Corollary 3.3. $\Gamma(h) = |\Pr\{h(x) = 1 | y = 1, a = 0\} - \Pr\{h(x) = 1 | y = 1, a = 1\}|$ *has sensitivity* $\max(\frac{1}{Z_{10}}, \frac{1}{|Z_{11}|}, \frac{\gamma_{10}}{|Z_{10}|-1} + \frac{\gamma_{11}}{|Z_{11}|+1}, \frac{\gamma_{10}}{|Z_{10}|+1} + \frac{\gamma_{11}}{|Z_{11}|-1})$

[VG: Proof 1 is in the finite sample setting and Proof 2 is in the distributional setting, not sure which of the two to keep. Finite sample is easier/more applicable but the distributional is general/converted easily to finite. We're leaning towards keeping finite sample but aren't sure]

[DK: this proof seems like appendix material]

[DK: should we use the term "database" or "sample" for the finite sample setting?]

Proof. We examine two cases for neighboring databases Z, Z':

Case 1: The difference in databases is in an entry within a subgroup (i.e. within Z_{1a} and Z'_{1a} for $a \in \{0,1\}$). WLOG let $a = 0$. Let the differing entries be called $z \in Z_{10}$ and $z' \in Z'_{10}$. Then, we have

$$\gamma_{11}^{Z'} = \gamma_{11}^{Z}$$

but

$$\gamma_{10}^{Z}(h) = \frac{1}{|Z_{10}|}\sum_{Z_{10}} h(x,0)$$

.

$$\gamma_{10}^{Z'}(h) = \frac{1}{|Z_{10}|}(\sum_{z \in Z_{10} \cup Z'_{10}} h(x,0) + h(x',0)).$$

Then,

$$\begin{aligned}\Gamma^{Z'}(h) &= \gamma_{10}^{Z'}(h) - \gamma_{10}^{Z'}(h)\\ &= \gamma_{10}^{Z'}(h) - \gamma_{11}^{Z}(h)\\ &\le \Gamma^{Z}(h) + \frac{1}{|Z_{10}|}\end{aligned}$$

Case 2: The neighboring databases are different in that an entry is added to one subgroup and another entry is removed from the other subgroup. Thus WLOG $Z' = (Z_{11} \setminus \{z_1\}) \cup (Z_{10} \setminus \{z_0\})$. Let $\gamma_{11}^{Z}(h) < \gamma_{10}^{Z}(h)$ then

$$\gamma_{11}^{Z'}(h) = \frac{1}{|Z'_{1a}|}\sum_{Z'_{1a}} h(x)$$

$$= \frac{1}{|Z'_{11}|}(\sum_{Z_{11}} h(x) - h(x_1)) = \frac{1}{|Z_{11}|-1}(\sum_{Z_{11}} h(x) - h(x_1))$$

Similarly,

$$\gamma_{10}^{Z'}(h) = \frac{1}{|Z'_{10}|}\sum_{Z'_{10}} h(x)$$

$$= \frac{1}{|Z'_{10}|}(\sum_{Z_{10}} h(x) + h(z_0)) = \frac{1}{|Z_{10}|+1}(\sum_{Z_{10}} h(x) + h(z_0))$$

Hence with notation $\Gamma^Z = \Gamma^Z(h)$ for clarity,

$$|\Gamma^{Z'} - \Gamma^{Z}| = (\gamma_{10}^{Z'} - \gamma_{11}^{Z'}) - (\gamma_{10}^{Z} - \gamma_{11}^{Z})$$

$$= (\gamma_{10}^{Z'} - \gamma_{10}^{Z}) + (\gamma_{11}^{Z} - \gamma_{11}^{Z'})$$

Let $\zeta_{10} = \gamma_{10}^{Z'} - \gamma_{10}^{Z}$ and $\zeta_{11} = \gamma_{11}^{Z} - \gamma_{11}^{Z'}$

$$\zeta_{10} = \frac{1}{|Z_{10}|+1}(\sum_{Z_{10}} h(x) + h(z_0)) - \gamma_{10}^{Z}$$

$$\zeta_{11} = \gamma_{11}^{Z} - \frac{1}{|Z_{11}|-1}(\sum_{Z_{11}} h(x) - h(x_1))$$

[DK: insert rest of proof, easy algebra]

$$|\Gamma^{Z'} - \Gamma^{Z}| = \zeta_{10} + \zeta_{11} = \frac{\gamma_{10}}{|Z_{10}|-1} + \frac{\gamma_{11}}{|Z_{11}|+1}$$

□

Proof. Construct ζ-close neighboring databases D and D' such that

$$\Pr[D' = d_{10}] - \Pr[D = d_{10}] = \zeta$$
$$\Pr[D = d_{11}] - \Pr[D' = d_{11}] = \zeta$$

and assume that $d_0, d_1 \in \text{supp}(D')$ and $\in \text{supp}(D)$. Let $d_{10} \in D_{10}$ and $d_{11} \in D_{11}$ and WLOG $\gamma_{10}^{Z}(h) > \gamma_{11}^{Z}(h)$ then,

$$\begin{aligned}\Gamma^{D'}(h) &= \gamma_{10}^{D'}(h) - \gamma_{11}^{D'}(h)\\ &= \int_{D'_{10}} h(x)dx - \int_{D'_{11}} h(x)dx\\ &= \int_{D_{10}} h(x)dx + \frac{\zeta}{dD_{10}}h(x_{10}) - \int_{D'_{11}} h(x)dx - \frac{\zeta}{dD_{11}}h(x_{11})\\ &\le \Gamma^{D}(h) + \frac{\zeta}{dD_{10}} + \frac{\zeta}{dD_{11}}\end{aligned}$$

where dD_{1a} is the measure of the set D_{1a}, for $a \in \{0,1\}$. Hence,

$$\zeta\Gamma = \frac{\zeta}{dD_{10}} + \frac{\zeta}{dD_{11}}.$$

□

[DK: Q: more elegant/complete proof? right now it seems we are doing by cases. ie. ζ close changes $\gamma_{ya}(h)$ by at most ... therefore ..]

4 APPROXIMATE FAIRNESS WITH DIFFERENTIALLY PRIVACY

We now turn to the finite sample setting where we release a hypothesis that minimizes the training error. Our goal is now approximate fairness. We use the exponential mechanism in the same way as it is used in private PAC learning. Defining the sample as Z, $|Z| = n$ and Γ^Z as our in-sample fairness measure, we give the algorithm:

$\mathcal{A}^{\epsilon}$: Output hypothesis $h \in \mathcal{H}$ with probability proportional to

$$\exp(-\frac{\epsilon \cdot u(Z,h)}{2\Delta u}) \tag{1}$$

where

$$u(Z,h) = \Gamma^Z(h) + \ell^Z(h)$$

.

$$\Delta u(Z,h) = \Delta\ell + \Delta\Gamma \approx O(1/n + 1/|Z_{10}| + 1/|Z_{11}|)$$

$$\ell^Z(h) = \frac{1}{n}\sum_{(x,y)\in Z} \Pr[h(x) \neq y]$$

This algorithm is the exponential mechanism in McSherry and Talwar, and so it is differentially private.

Lemma 4.1. *The algorithm $\mathcal{A}_\epsilon$ is ϵ-differentially private.*

Note that except for when $|\mathcal{H}|$ is polynomial, the exponential mechanism does not necessarily yield a polynomial time algorithm.

Theorem 4.2. *(Generic private fair learner) For all $d \in \mathbb{N}$, any concept class C_d whose cardinality is at most $\exp(poly(d))$ is privately and approximately fairly agnostically learnable using $\mathcal{H}_d = C_d$. More precisely, the learner uses $n = ..$ labeled examples from D, where ϵ, α, and β are parameters of the private learning.* [DK: need to be sure of constants before calculating n. Will be poly for sure]

Proof. Let $\mathcal{A}_\epsilon$ be as defined above. The privacy condition is satisfied by Lemma.

Now we show that the utility condition is also satisfied. Let the event $E = \{\mathcal{A}_\epsilon = h \text{ with } u(h) > OPT + \alpha\}$. We need that $\Pr[E] \leq \beta$. We define the utility of h as

$$u(Z, h) = \Gamma^Z(h) + \ell^Z(h)$$

. [DK: introduce this before proof and have a discussion about it? OR discussion after proof about utility and implication of α accuracy/fairness whp.]

Recall that Z is the sample drawn i.i.d. from a distribution D. By Chernoff-Hoeffding bounds (insert appendix ref. see below proof for now),

$$\Pr[|u(Z, h) - u(D, h)| \geq \rho] \leq 4\exp(\frac{-\rho^2 n}{2})$$

for all hypotheses $h \in \mathcal{H}_d$. Hence,

$$\Pr[|u(Z, h) - u(D, h)| \geq \rho \text{ for some } h \in \mathcal{H}_d] \leq 4|\mathcal{H}_d|\exp(\frac{-\rho^2 n}{2})$$

[DK: need following proof? relatively similar to kasiviswanathan what can we learn privately, except we have different utility and sensitivity] Now we analyze $\mathcal{A}_\epsilon(Z)$ conditioned on the event that for all $h \in \mathcal{H}_d$, $|u(Z, h) - u(D, h)| < \rho$. For every $h \in \mathcal{H}_d$, $\Pr[\mathcal{A}_\epsilon(Z) = h]$ is

$$\frac{\exp(-\frac{\epsilon}{2\Delta u} \cdot u(Z, h))}{\sum_{h' \in \mathcal{H}_d} \exp(-\frac{\epsilon}{2\Delta u} \cdot u(Z, h'))} \leq \frac{\exp(-\frac{\epsilon}{2\Delta u} \cdot u(Z, h))}{\max_{h' \in \mathcal{H}_d} \exp(-\frac{\epsilon}{2\Delta u} \cdot u(Z, h'))}$$
$$= \exp(-\frac{\epsilon}{2\Delta u}(u(Z, h) - \min_{h' \in \mathcal{H}_d} u(Z, h')))$$
$$\leq \exp(-\frac{\epsilon}{2\Delta u}(u(Z, h) - (OPT + \rho)))$$

Hence the probability that $\mathcal{A}_\epsilon(Z)$ outputs a hypothesis $h \in \mathcal{H}_d$ such that $u(Z, h) > OPT + 2\rho$ is at most $|\mathcal{H}_d| \exp(-\frac{\epsilon \cdot \rho}{2\Delta u})$

Setting $\rho = \alpha/3$. If $u(D, h) \geq OPT + \alpha$ then $|u(D, h) - u(Z, h)| \geq \alpha/3$ or $u(Z, h) \geq OPT + 2\alpha/3$. Hence

$$\Pr[E] \leq |\mathcal{H}_d|(4\exp(\frac{-\alpha^2 n}{18}) + \exp(-\frac{\epsilon \cdot \alpha}{6\Delta u})) \leq \beta$$

.

Where the inequality holds for $n \geq$. □

Theorem (Real-valued Additive Chernoff-Hoeffding Bound). Let $X_1, ..., X_d$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $a \leq X_i \leq b$ for all i. Then for every $\rho > 0$,

$$Pr[|\frac{\sum_i X_i}{n} - \mu| > \rho] \leq 2\exp(\frac{-2\rho^2 n}{(b-a)^2})$$

[DK: pretty sure the constants are correct and the union bound is necessary] Chernoff bounds for $u(Z, h)$:

Let

$$X_i^a = n(\frac{\mathbb{1}_{z \in Z_{1a}} h(x)}{|Z_{1a}|} - \frac{\mathbb{1}_{z \in Z_{1\neg a}} h(x)}{|Z_{1\neg a}|}) + \Pr[h(x) \neq y]$$

Then

$$\max_{a \in \{0,1\}} \frac{1}{n} \sum_Z X_i^a$$
$$= \frac{1}{n} \max_{a \in \{0,1\}} \sum_Z (n(\frac{\mathbb{1}_{z \in Z_{1a}} h(x)}{|Z_{1a}|} - \frac{\mathbb{1}_{z \in Z_{1\neg a}} h(x)}{|Z_{1\neg a}|}) + \Pr[h(x) \neq y])$$
$$= |\sum_Z \frac{\mathbb{1}_{z \in Z_{10}} h(x)}{|Z_{10}|} - \frac{\mathbb{1}_{z \in Z_{11}} h(x)}{|Z_{11}|}| + \sum_Z \frac{\Pr[h(x) \neq y]}{n}$$
$$= |\frac{1}{|Z_{10}|} \sum_{z \in Z_{10}} h(x) - \frac{1}{|Z_{11}|} \sum_{z \in Z_{11}} h(x)| + \sum_Z \frac{\Pr[h(x) \neq y]}{n}$$
$$= |\gamma_{10}^Z - \gamma_{11}^Z| + \ell^Z(h)$$
$$= ||(\Gamma^Z(h), \ell^Z(h)||_1 = u(Z, h)$$

Hence

$$\Pr[|\frac{1}{n} \sum_Z X_i^a - \mathbb{E}[X_i^a]| > \rho] \leq 2\exp(\frac{-2\rho^2 n}{(2-(-1))^2})$$

By union bound (over the choice of $a \in \{0, 1\}$

$$\Pr[|u(Z, h) - u(D, h)| > \rho] \leq 4\exp(\frac{-2\rho^2 n}{9})$$

[DK: Note: can do extension into other definitions of fairness (with constant factor). Can also extend into a different norm for $u(Z, h)$ if we have a concentration of measure theorem for the different norm. There are also random matrix concentration bounds (can do concentration for all functions of the confusion matrix?)]

5 A POLYNOMIAL TIME ALGORITHM FOR APPROXIMATELY FAIR AND PRIVATE CLASSIFICATION

[DK: we've found a promising method: kamishima et al. working on proving DP and figuring out code. kamishima provides their original data in their repo too !]

REFERENCES

[1] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*. 265–284.

[2] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9 (Aug. 2014), 211–407. https://doi.org/10.1561/0400000042

[3] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29*. 3315–3323.

[4] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan. 2010. The Limits of Two-Party Differential Privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. 81–90. https://doi.org/10.1109/FOCS.2010.14

[5] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Proceedings of the 2017 Conference on Learning Theory (Proceedings of Machine Learning Research)*, Satyen Kale and Ohad Shamir (Eds.), Vol. 65. PMLR, Amsterdam, Netherlands, 1920–1953. http://proceedings.mlr.press/v65/woodworth17a.html