Privacy ∩ Fairness

Rachel Cummings, Varun Gupta, Dhamma Kimpara, Jamie Morgenstern

August 6, 2018

1 Preliminaries

2 Achieving exact fairness and differential privacy is impossible

We start with the the continuous setting (which can be translated to the discrete sample setting later). Let $\mathcal{X}$ be the data universe with elements $d_i = (y_i, x_i, a_i) \in \mathcal{X}$ where each entry consists respectively of the response, features, and protected attribute. Note that x_i may also contain a_i to be used by a hypothesis h. We define a neighboring database as defined in [insert DP paper here by Talwar]:

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

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

The probability is taken over the random coins of $\mathcal{A}$.

Definition 2 (Equal Opportunity). We say that a binary predictor h satisfies equal opportunity with respect to A and Y if (need explicit form of database

$$\Pr\{h(x) = 1 | a = 1, y = 1\} = \Pr\{h(x) = 1 | a = 0, y = 1\}$$

Definition 3 (Statistical distance and δ-closeness). Databases D and D' (distributions over $\mathcal{X}$) are δ-close if the statistical distance between their distributions is at most δ, i.e.,

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

Definition 4? For ease of notation we define

$$D_{ya} := \{d_i \in D | y_i = y, a_i = a\}$$

Definition 5 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 1 Let $\mathcal{H}$ be a non-trivial hypothesis class. 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 Let D, D' be δ-close and suppose that the fair (as in definition 2) hypotheses h is released by algorithm $\mathcal{A}$ on input of D. Let $d(\cdot)$ be the measure as defined by the distribution D. Pick $d_1, d_0 \in D_{11}$ such that $h(x_1) > h(x_0)$ and assume that $d_1, d_0 \in supp(D)$ and $supp(D')$. Let $\delta < \min_{i,D,D'}(d_D(d_i))$ D' be a database with δ more probability mass of d_1 and δ less mass of d_0. Then recalling that $d_i = (y_i, x_i, a_i)$:

$$\begin{aligned}\Pr_{D'}\{h(x) = 1 | a = 1, y = 1\} &= \int_{D_{11}} h(x)dx + \frac{\delta}{d(D_{11})}h(x_1) - \frac{\delta}{d(D_{11})}h(x_0) \\ &= \Pr_{D}\{h(x) = 1 | a = 1, y = 1\} + \frac{\delta}{d(D_{11})}(h(x_1) - h(x_0)) \\ &> \Pr_{D}\{h(x) = 1 | a = 1, y = 1\}\end{aligned}$$

where the last inequality is due to the assumption that $h(x_1) > h(x_0)$. Hence hypothesis h is unfair when applied to database D' and cannot be released by the exactly fair mechanism $\mathcal{A}$ on input of database D'. So

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

.

Which violates the definition of privacy in definition 1. □

Lemma 2 (ϵ, δ)-DP and fairness is impossible. Need proof

2.1 Defining approximate fairness

We now turn to the approximate fairness and define analogs of the definitions in the exact fairness setting.

Define $Z_{ya} := \{z_i \in Z | y_i = y, a_i = a\}$ and

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

Definition 6 (α-discrimination [srebro ref]). 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.

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

When $\alpha = 0$ we are in the exact fairness setting. Note that by [Srebro ref] the sample fairness measure Γ^Z converges to the population measure Γ when n is large. To achieve an approximate analog to equality of opportunity in definition 2. We define:

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

We now focus only on the equality of opportunity setting.

Corollary 1 $\Gamma(h) = |\Pr\{h(x) = 1|y = 1, a = 0\} - \Pr\{h(x) = 1|y = 1, a = 1\}|$ has sensitivity

Proof 1 We examine two cases:

Case 1: neighboring database Z' is a change in entry within a subgroup ie. within Z_{1a}. Let z_0 be replaced by z_1 then $\gamma^{Z'}_{1\neg a} = \gamma^Z_{1\neg a}$ but

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

Case 2: neighboring database is not a change in entry within a subgroup. Thus deleting an entry in Z_{1a} and add one to $Z_{1\neg a}$.

Proof 2

Construct δ-close neighboring databases D and D' such that

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

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 $\gamma^Z_{10}(h) > \gamma^Z_{11}(h)$ then,

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

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

$$\Delta\Gamma = \frac{\delta}{dD_{10}} + \frac{\delta}{dD_{11}}$$

[DK: Q: more elegant/complete proof? right now it seems we are doing by cases. ie. δ close changes $\gamma_{ya}(h)$ by at most ... therefore .. Q: should we do this proof in finite sample context or just translate this to the finite sample (with $\delta = 1/n$)]

3 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

$$u(Z,h) = ||(\Gamma^Z(h), \ell^Z(h)||_1$$

.

$$\Delta u(Z,h) = ||(\Delta\ell, \Delta\Gamma)||_1 \approx O(||1/n, 1/|Z_{10}| + 1/|Z_{11}|||_1)$$

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

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

Lemma 2 The algorithm $\mathcal{A}_\epsilon$ is ϵ-differentially private. Need proof?

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

[DK: refactor constants] **Theorem 1**(Generic private fair learner) For all $d \in \mathbb{N}$, any concept class $\mathcal{C}_d$ whose cardinality is at most $\exp(\text{poly}(d))$ is privately fairly agnostically learnable using $\mathcal{H}_d = \mathcal{C}_d$. More precisely, the learner uses $n = ..$ labeled examples from D, where ϵ, α, and β are parameters of the private learning.

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

$$u(Z,h) = ||(\Gamma^Z(h), \ell^Z(h)||_1$$

.

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] 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 $n \geq$. $\square$

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: needs checking, refactor constants] 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?)]

4 Approximately fair correction with differentially private linear programs

Do DP-LP

5 A polynomial time algorithm for approximately fair and private classification

Old stuff:

Our initial approach was to look at this algorithm that achieves equality of opportunity or equal odds by flipping (with some probability) the label of the original non-fair algorithm, $\hat{Y}$, depending on the output label and the class membership A. $\hat{Y}$ is trained on (X, Y), the features X and the true label Y. When we flip labels given by $\hat{Y}$ we assume that we are given full access to the distribution $(Y, A, \hat{Y})$.

Essentially, modifying $\hat{Y}$ we create a new classifier $\tilde{Y}$ that is fair by setting the probabilities $\Pr\{\tilde{Y} = 1 | \hat{Y} = \hat{y}, A = a\}$ for $\hat{y}, a \in \{0, 1\}$ appropriately. The "pipeline" with the available information at each stage laid out plainly is as follows:

$$(X, Y) \xrightarrow{\text{vanilla training}} (\hat{Y}, A, Y) \xrightarrow{\text{fairness}} \tilde{Y}$$

6 Privatizing Woodworth et. al. (2017)

(This is the follow-up to Eq. of Opp.)

6.1 Brief summary

This paper gives a two step algorithm in a learning model with finite samples and hypothesis classes. They split the training set into two parts. The first step is traditional ERM (with 0-1 loss) but with a sample based fairness constraint using half of the training set. The second step takes the outputted hypothesis from step 1 and does the same post-processing step as in the algorithm of Equality of Opportunity.

7 Privatizing first

With the finite sample setting and also the case that we cannot achieve exact fairness privately, this line of inquiry now seems to be fruitful. It would not take long to analyze the effect of first privatizing via synthetic data or adding noise to the training data (smallDB etc.) and then training a fair classifier.

8 Appendix

p-norm discussion from exp mech?