ISYE 8813: Algorithmic Foundations of Ethical ML **Fall 2021**

Lectures 18: Individual Fairness

Lecturer: Juba Ziani

This lecture is based mostly on [DHP+12].

In the previous lecture, we saw group-based definitions of fairness. Most of these definitions were statistical: i.e., they looked at probabilistic properties of different groups, and were trying to (approximately) equalize these asymptotic properties over groups. However, such definition of fairness rarely say anything about what happens at the individual level: these definitions do not look at what happens in a more fine-grained manner within each group. This is where individual fairness comes in: we are going to aim to guarantee that at the *person-by-person* level, people are treated fairly.

Example 1 (Statistical/Demographic Parity)**.** *An example from previous lectures is the following. Suppose we want to satisfy demographic parity, i.e. we give the same probability of good outcomes to our two groups. Suppose we are in a binary classification setting in which we are aiming to hire applicants for a job. One way to do so is to hire the best applicants in group* 1*, but the worst applicants in group* 2*. This is intrisically unfair*

1. *Across individuals in different groups: individuals in group* 1 *must find it unfair that individuals in group* 2 *that are much less qualified than them were hired.*
2. *Across individuals in the same group: qualified individuals in group* 2 *seems to be discriminated against in favor of unqualified individuals.*

We already knew that demographic parity was a "weak" definition of fairness, in the sense that it can be satisfied without using any information about agents' features and does not have to be statistically dependent on an agent's true label or qualification level. But issues where statistical notions of fairness do not satisfy fairness across smaller sub-groups or individuals also arise for the other, less "naive" definitions of group fairness that we have seen in class:

Example 2 (Calibration)**.** *Let us look for example at calibration. Imagine we have* 3 *individuals. Individuals* 1 *and* 2 *have low risks (probability of recidivism for example) of* 0.2*; individual* 3 *has a very high risk of* 0.9*. One way to obtain calibration is the following:*

- *Individuals* 1 *is assigned a score of* 0.2*. This is obviously calibrated.*
- *Individuals* 2 *and* 3 *are assigned the same, high score of* 0.55*. This is once against calibrated, because among the individuals to which we assigned this score (agents* 2 *and* 3*), the average score is* $\frac{0.2+0.9}{2} = 0.55$*.*

But one may argue here that even if calibration holds, the scores are very unfair across individuals: we consider agent 2 *to be high risk but agent* 1 *to be low risk, while those agents are statistically indistinguishable! I.e., generally, statistical/group definitions of fairness do not guarantee fairness at the level of a few individuals.*

1 Individual fairness: the definition

Here, we consider a classification setting, in which we use a classifier or mechanism M to make decisions about individuals. We allow for randomized classifiers, that are functions from the feature space V to the set of distribution over outcomes $\Delta(\mathcal{O})$ ($\mathcal{O}$ denotes the set of outcomes). I.e., $M: V \to \Delta(\mathcal{O})$. For a feature vector $x \in V$, $M(x)$ is the distribution that describes the probability of giving a positive outcome to/accepting x.

Here, the high-level idea behind fairness is going to be similar to before. In most of what we have seen so far, we want (at the group-level) people with similar true labels y to be statistically treated the same. Here, we refine this idea to hold at the individual level: two individuals that have similar feature vectors x and x' should be treated similarly.

To do so formally, we introduce two metrics D and d (we talk about "metrics" in the lose sense here – i.e. they do not necessarily have to be a proper distance function). D works in feature space V and measures the distance between two feature vectors x and x'. d works in decision/outcome space $\Delta(\mathcal{O})$ and measures the difference between outcomes. We will say that *individual* or *metric* (since we are using metrics across individuals to define it) fairness holds if the following Lipschitz condition holds:

Definition 3 (Lipschitz property)**.** *A mapping $M: V \to \Delta(\mathcal{O})$ satisfies the (D, d)-Lipschitz property if and only if for every $x,\ x' \in V$, we have*

$$D(M(x), M(x')) \leq d(x, x').$$

So, informally, if x and x' are close to each other, it should be that $D(M(x), M(x'))$ is also small, which means the distributions of outcomes over x and x' are close to each other.

Remark 1. *Once again, finding a Lipschitz classifier is very easy. For example, any constant classifier that assigns the same distribution $M(x)$ to all $x \in V$ satisfies $D(M(x), M(x')) = 0$. The difficulty, once again, is to find a fair classifier that has good utility or accuracy guarantees.*

2 Finding the best individually fair classifier

Suppose we are interested in obtaining a "good" fair classifier. Here, we mean "good" in the sense that the classifier minimizes some (expectation of a) loss function $L(x, o)$. This loss function is $L: V \times \mathcal{O} \to \mathbb{R}$: it takes as an input a feature vector $x \in V$ and a decision/outcome $o \in \mathcal{O}$, and outputs a real number that measures the quality/accuracy/utility of decision o on an individual defined by features x. The problem we want to solve is then given as follows:

$$\begin{aligned} OPT = \min_{\{\mu_x\}_{x \in V}} \quad & \mathbb{E}_{x \sim V}\left[\mathbb{E}_{o \sim \mu_x}\left[L(x, o)\right]\right] \\ \text{s.t.} \quad & D(\mu_x, \mu_{x'}) \leq d(x, x')\ \forall x, x' \in V \\ & \mu_x \in \Delta(\mathcal{O})\ \forall x \in V, \end{aligned}$$

where here μ_x is a shorthand for $M(x)$. Note that if $\mathbb{E}_{x,M(x)}[L(x,o)]$ is convex in μ_x, and $D(.,.)$ is jointly convex in μ_x, μ'_x, this defines a convex optimization problem: indeed, each μ_x can be described as a vector of weights w_x where $w_x(o)$ is the probability of picking outcome o. The Lipschitz constraint is convex, because $D(.,.)$ is convex, and $d(x,x')$ is a constant term. The constraint that $\mu_x \in \Delta(\mathcal{O})$ is just a collection of linear constraints: $\sum_o w_x(o) = 1$ and $w_x(o) \geq o$ for all $o \in \mathcal{O}$.

In fact,

$$\mathbb{E}_{x,M(x)} L(x,o) = \sum_{x \in V} \Pr[x] \sum_{o \in \mathcal{O}} \Pr[M(x) = o] \, L(x,o).$$

Here, the decision variables are the $\Pr[M(x) = o]$'s (remember that we are trying to design the distribution $M(x)$), and $\Pr[x]$ and $L(x,o)$ are pre-defined parameters of the problem that are constant in the decision variables. Hence, $\mathbb{E}_{x,M(x)} L(x,o)$ is linear in the decision variables, and the whole problem we are trying to solve is a linear optimization problem.

Note however that problem OPT can be intractable in practice, despite being convex/linear: if V is high-dimensional, then we have exponentially (in the dimension of the problem) many $M(x)$'s to design and exponentially many fairness constraints across feature pairs (x,x') to satisfy. It is also possible that O is high-dimensional or infinite, in which case specifying $M(x)$ for even a single x is intractable. We will see one way to deal with such issues in Section 4.

3 What metrics to use?

[DHP+12] proposes a few metrics that one may use, both for d and D. Let us start with D: remember that D is a metric on $\Delta(\mathcal{O})$, hence it measures the difference between *probability distributions*.

One of the first metric that comes in mind for D is the statistical distance/total variation distance. It is formally defined as follows:

Definition 4 (Total Variation Distance)**.** *The Total Variation Distance between two probability distributions P and Q on a finite domain $\mathcal{O}$ is given by*

$$D_{tv}(P,Q) = \frac{1}{2} \sum_{o \in \mathcal{O}} |P(o) - Q(o)| \, .$$

This metric plays well with the optimization program above. Indeed, note that for some constant C, the constraint

$$\sum_{o \in \mathcal{O}} |P(o) - Q(o)| \leq C$$

can be rewritten as a collection of linear constraints, as follows:

$$\sum_{o \in \mathcal{O}} \delta_o \leq C \quad \text{and} \quad -\delta_o \leq P(o) - Q(o) \leq \delta_o \; \forall o \in \mathcal{O}.$$

Another possible metric is the *relative ℓ_∞ metric*, defined as follows:

Definition 5 (Relative ℓ_∞ metric)**.** *The Relative ℓ_∞ metric between two probability distributions P and Q on a finite domain $\mathcal{O}$ is given by*

$$D_\infty(P,Q) = \sup_{o\in\mathcal{O}} \ln\left(\frac{\Pr[M(x)=o]}{\Pr[M(x')=o]}, \frac{\Pr[M(x')=o]}{\Pr[M(x)=o]}\right).$$

Once again, this is a metric that can be tractably used in our optimization program. Indeed, the constraint $D_\infty(P,Q) \le C$ for some constant C is equivalent to the fact that for all $o \in \mathcal{O}$,

$$P(o) \le \exp(C)Q(o) \quad \text{and} \quad Q(o) \le \exp(C)P(o).$$

4 A mechanism for fairness via differential privacy

A natural connection to differential privacy Note that such a definition of fairness is inspired by and has a natural connection to differential privacy (in fact, if you look at the authors of [DHP+12], several of them were working on differential privacy before they started looking at fairness!)

Suppose V is discrete. Set $x \triangle x'$ to be the number of entries that differ between agents x and x', and let $d(x,x') = \varepsilon \cdot x \triangle x'$. Then the (D_∞, d)-Lipschitz condition states that

$$\sup_{o\in\mathcal{O}} \ln\left(\frac{\Pr[M(x)=o]}{\Pr[M(x')=o]}, \frac{\Pr[M(x')=o]}{\Pr[M(x)=o]}\right) \le \varepsilon \cdot x \triangle x',$$

which is equivalent to pure differential privacy! Note that this is indeed an equivalence: when $x \triangle x' \le 1$ i.e. x and x' are neighboring, we immediately see that the Lipschitz condition implies differential privacy. Now, note that differential privacy implies the Lipschitz condition, as it is a basic property of DP that $x \triangle x' = k$ implies a logarithmic distributional shift of at most $k\varepsilon$.

An implication in terms of fair mechanisms One way to obtain a fair mechanism in terms of D_∞ is to simply solve optimization program OPT. Sometimes, however, doing so may be impractical, for example when the instance size is large (think many possible $x \in V$, and we have to define a mapping $M(x)$ for every single one of them!). This suggests an alternative approach: one can simply use the exponential mechanism!

Here, the exponential mechanism can be instanced as follows: the range of possible outcomes is $R = \mathcal{O}$. A database is simply an element $x \in V$. For each database x, and each outcome o, we let the utility function be the loss: i.e., $u(x,o) = -L(x,o)$. We then pick an outcome o when the database is x with probability proportional to $\exp\left(\frac{-\varepsilon \cdot L(x,o)}{2\Delta_L}\right)$, as we saw in our DP lectures (Δ_L being the sensitivity with respect to x). The following then holds:

- $M(x)$ is a proper distribution over $\mathcal{O}$. Further, M satisfies differential privacy, hence the (D_∞, d)-Lipschitz condition holds.

- The loss cannot be too high. Letting $L^*(x) = \inf_o L(x, o)$, we do have by the accuracy guarantees of the exponential mechanism that with probability $1 - \delta$,

$$L(x, o) \leq L^*(x) + 2\frac{\Delta_L}{\varepsilon} \ln(|\mathcal{O}|/\delta),$$

guaranteeing that the loss is not too big. This is in particular a good bound when $\mathcal{O}$ has finite size, and Δ_L is relatively small (which holds when similar individuals have similar outcomes).

A major advantage of using the exponential mechanism over solving optimization program OPT (from Section 2) directly is tractability. Indeed, imagine a situation in which V is high-dimensional. For example, $V = \{0, 1\}^d$. Solving Program OPT then is intractable: we have 2^d possible vectors, hence 2^{2d} pairs of feature vectors (x, x'), each one associated with a fairness constraint! Hence, we have exponentially many variables to design (the μ_x's) and fairness constraints to satisfy, which seems hopeless. However, the exponential mechanism can be defined without having to iterate over all (x, x') pairs – we obtain the fairness constraint from differential privacy, without having to explicitly write it in the mechanism we use! Rather, the learner/decision-maker can just, for each x he faces, output x according to distribution $\exp\left(\frac{-\varepsilon \cdot L(x,o)}{2\Delta_L}\right)$. If O has tractable size, this is an easy problem!

[DHP+12] develops bounds in the case when $\mathcal{O} = V$ (so you output another vector $y \in V$, for example one that is representative of your data $x \in V$) that can hold for very large $|V|$, so long as the doubling dimension of V is small. This is beyond the scope of the lecture, but feel free to look at Section 5 of the paper if you want to read into this more carefully.

5 Is individual fairness compatible with statistical fairness? The case of statistical parity

[DHP+12] study some conditions under which their definition of individual fairness naturally leads to statistical fairness. We will skip over the proofs here; more details are available in Section 3 of the paper. Section 4 is also related and possibly of interest, but we will not cover it: in cases in which it is not possible to have statistical fairness and individual fairness at the same time, how can we achieve a trade-off?

Here, they consider the binary setting in which the decision space is $O = \{0, 1\}$. We define the bias across groups S and T as

$$\text{bias}_{D,d}(S, T) = \max_{M \text{ is } (D,d)-\text{fair}} \mu_S(0) - \mu_T(0),$$

where $\mu_S(o) \triangleq \Pr[M(x) = o | x \in S]$. So, the bias here quantifies how bad the distributions of ¶$[M(x) = 0]$ can be across groups over mappings that satisfy our condition. The paper shows that this is related to what is called the *Earthmover* distance:

Definition 6 (Earthmover/Wasserstein distance)**.** *Let $\sigma : V \times V \to \mathbb{R}$ be a non-negative distance function. The σ-Earthmover distance between two distributions S and T is given by*

$$\begin{aligned}\sigma_{EM}(S,T) \triangleq \min_h \quad & \sum_{x,x' \in V} h(x,x')\sigma(x,x') \\ s.t. \quad & \sum_{x' \in V} h(x,x') = S(x) \\ & \sum_{x \in V} h(x,x') = T(x') \\ & h(x,x') \geq 0.\end{aligned}$$

This distance is the solution to a *transportation* problem: i.e., what is the minimum cost we incur to move probability mass around to go from distribution S to distribution T? Here, $\sigma(x,x')$ is the cost for moving one unit of probability mass on a point x to point x'. $h(x,x')$ tells us with what probability we move a point of the form x to x'. Hence $\sum_{x,x' \in V} h(x,x')\sigma(x,x')$ is indeed the total probability mass we move around. The two equality conditions just ensure that the initial mass over x is $S(x)$, and the resulting mass over x' (so after moving) is $T(x')$.

It turns out that with $D = D_t v$, we have the following nice result:

Lemma 7.

$$bias_{D_{tv},d}(S,T) \leq d_{EM}(S,T).$$

Further, if $d(x,x') \leq 1$ for all $x, x' \in V$, we have equality.

Therefore, the Earthmover distance characterizes exactly the bias between groups S and T! In turn, the lemma shows that the bigger the distance between the distributions S and T is, the less demographic parity is compatible with (D,d)-fairness. This makes sense: if two groups have exactly the same distribution of features x, then individual fairness is compatible with statistical parity: by treating two people with the same features but different groups the same way, we get the same distribution of outcomes.

However, if the two distributions are different, incompatibility start arising. Imagine here that the features are just in fact the true label of an agent $y \in \{0,1\}$. In group 1, 10 percent of agents have $y = 1$, while in group 2, all agents have $y = 1$. The following mapping is individually fair: $M(y) = 1$ if $y = 1$ and 0 otherwise. But it gives an outcome $o = 1$ to 10 percent of group 1, but the entirety of group 2. This just means that individual fairness is harder to satisfy across groups when groups have, for example, very different distributions of qualifications, if we also want demographic parity. [DHP+12] provides an example and a picture of such situations (see Figure 2 in Section 4).

6 Shortcomings of individual fairness

We will talk about the shortcomings of individual fairness in the next lecture. There, we will see how since both individual fairness and statistical fairness have shortcomings, we may

want to aim for definitions of fairness that are "in-between" and interpolate between these two extremes.

References

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.