Fairness in Prediction and Allocation[*]

Jamie Morgenstern Aaron Roth

September 7, 2021

1 Introduction

Many high stakes decisions that are now aided by machine learning can be viewed as *allocation* problems. We will often have some *good* (such as a job or a loan) or *bad* (such as incarceration) to distribute amongst a population. People on at least one side of the problem will have preferences over the other side that are based on qualities that are not directly observable (e.g. banks want to give loans to the creditworthy; courts want to incarcerate the guilty or those likely to repeat their offenses). The machine learning task is to make predictions about those qualities from observable attributes. It is natural that when we make high stakes decisions, we will be concerned about *unfairness* — the potential for the system as a whole (of which a machine learning algorithm or statistical model may play only a relatively small role) to disproportionately favor some people over others, perhaps for reasons more related to their demographics than features relevant for the task at hand. And we note at the outset that these considerations are not hypothetical: statistical models are currently being used to inform bail and parole decisions, hiring and compensation decisions, lending decisions, and an increasingly extensive collection of high stakes tasks—and there are now an enormous number of cases of systemic decision-making that would be called sexist or racist—at least if the decisions had been made by a human being.

But what should we make of such decision making when it is carried out by an algorithm that was derived by optimizing a facially neutral objective function, like classification error? The first thing we need to understand is why we might expect machine learning to exacerbate unfair decision making in the first place. After all, at its heart, machine learning usually corresponds to simple, principled optimization. Typically, the process of machine learning will start by gathering a dataset consisting of various measured *features* for each person. For example, in the recidivism prediction problem that is often used to inform whether prisoners should be released on parole, features might include

[*] This chapter will appear in: *Online and Matching-Based Market Design.* Federico Echenique, Nicole Immorlica and Vijay V. Vazirani, Editors. Cambridge University Press. © 2021

basic demographics (e.g. age, sex), together with information about criminal history, like the number of previous arrests and convictions for violent and non-violent offences. It is also necessary to specify something that we wish to predict, called the *label* — importantly, something that we can measure. For example, in a recidivism prediction setting, the goal is to predict whether inmates will commit a violent crime within (e.g.) 18 months of being released — but because this isn't directly observable, the label is often taken to be a proxy variable such as whether the individual was *arrested* for a violent crime. And bias of the sort we are trying to avoid can creep in to the final decision making rule via the data, and via the optimization process itself in both obvious and non-obvious ways.

First, lets consider the labels that we are trying to predict. In the criminal recidivism application, are arrests an equally good proxy for crime in all populations? Perhaps not: if police are more likely to stop black people than white people, all things being equal, then the effect of using arrest labels as proxies for unmeasured criminality will be that black people in the dataset will in aggregate appear to be a higher risk population than they would if we could measure the actual variable of interest. And of course there is no reason to expect that applying machine learning techniques will somehow remove these human biases which have crept into the dataset via the labelling process: machine learning at best replicates the patterns already in the data.

But the problem goes beyond that, and bias can creep into machine learning even if the labels are correctly recorded in the data. It is enough that two different populations are statistically different in the sense that features correlate differently with the label, depending on the population. Consider, for example, a college admissions problem in which we seek to identify talented students from their high school records. Suppose there are two populations, one of which attends well resourced suburban high schools, and the other of which attends poorly resourced city high schools. One feature that we may have available to us in a college application is how many AP science courses a student has taken. For students from the well resourced highschools, which offer many AP science courses, this might be an informative predictor of student talent. But amongst students from the poorly resourced high schools – that offer many fewer, or perhaps even no AP courses, this feature is much less predictive (because ordinary and talented students alike do not take AP courses, since none are available). It may be that the distribution of talent is the same in both highschools — and even that in isolation, talent is equally predictable within both populations — but predictable using different models. Yet if we insist on selecting a single model for both populations[1], optimizing for overall error will result in finding the model that better fits the majority population, simply because error on the majority population contributes more to overall error by virtue of their numbers. In this case (because for the wealthy students, taking no AP classes is a negative signal) this will result in penalizing students who have not taken AP

[1] As we must if we don't want to explicitly use group membership as a feature in our decision making process — something that is explicitly illegal in many settings such as lending and insurance.

classes, which will have the effect of reducing admission rate on the population from under-resourced schools.

1.1 What is "Fairness" in Classification?

We have discussed informally how "unfairness" might creep in to statistical models learned from data — but if we want to quantify it and think about how to eliminate it, we need to be much more precise. So what should "fairness" entail in a limited setting like binary classification? Here we will intentionally start with a narrowly specified version of the problem, in which things like the data distribution are taken as exogenously given properties of the world, and the only object of study is the statistical model that we train from the data itself. Later in the chapter we will expand the scope of our analysis to account for the second order effects of the choices that we make in choosing a classification technology. We will focus on the simplest possible setting, assuming away many of the difficulties that arise in more realistic settings. For example, we will assume that there are only two disjoint groups of interest, which we will denote by $g \in \{0, 1\}$. This could denote any binary distinction made by e.g biological sex, race, or class. This avoids (important) complications that arise when there are many — potentially intersecting — groups for whom we care about fairness. We will also implicitly assume that the labels recorded in the data are the true labels — i.e. that our data is not contaminated with the sort of proxy label bias we discussed above. We will find that even in this simplified setting, there are already thorny issues to deal with.

Definition 1 (Data Distribution)**.** We model individuals as being sampled from *data distributions* $\mathcal{D}_g$ which may depend on their group g. The distributions $\mathcal{D}_g$ have support over $X \times \{0, 1\}$, where X represents an abstract *feature space*, and Y represents some binary outcome of interest.

In the above, we imagine that the feature space X is the same for both populations and does not encode group membership g, which we handle separately (because we wish to study the question of whether decision making should condition on g or not). We write $\gamma \in [0, 1]$ to represent the fraction of the overall population that group 0 represents — and thus $1 - \gamma$ is the fraction of the overall population that group 1 represents. We write $\mathcal{D}$ to denote the distribution on the entire population of both groups, defined as follows. $\mathcal{D}$ is supported on $\{0, 1\} \times X \times \{0, 1\}$ corresponding to triples (g, x, y) of group membership, features, and labels. To sample from $\mathcal{D}$, we:

1. Let $g = 0$ with probability γ (otherwise $g = 1$)
2. Sample $(x, y) \sim \mathcal{D}_g$.
3. Output the triple (g, x, y).

The fact that the distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ may differ allows us to model the unavoidable fact that distinct populations will have different statistical properties (without modeling for now the source of those differences — such as e.g. the difference in high school resources as discussed above).

The prediction problem is to learn some rule $h : \{0,1\} \times X \to \{0,1\}$ to predict the unknown label y as a function of the observable features x — and possibly the group label g.

Definition 2 (Unconstrained Error Minimization)**.** The overall error rate of a hypothesis h is:

$$\mathrm{error}(h) = \Pr_{(g,x,y)\sim\mathcal{D}}[h(g,x) \neq y]$$

The error rate of a hypothesis h on a group g is:

$$\mathrm{error}_g(h) = \Pr_{(x,y)\sim\mathcal{D}_g}[h(g,x) \neq y]$$

We denote by h^* the *Bayes optimal classifier* — i.e. the classifier that martials all available statistical information to optimally predict the label:

$$h^*(g,x) = \begin{cases} 1, & \text{if } \Pr[y=1|x,g] \geq 1/2; \\ 0, & \text{otherwise.} \end{cases}$$

Observe that the Bayes optimal classifier is both error optimal overall, and on each group in isolation:

$$\mathrm{error}(h^*) \leq \mathrm{error}(h) \quad \mathrm{error}_g(h^*) \leq \mathrm{error}_g(h)$$

for all h and for all g.

Perhaps the most immediately obvious definition of fairness in classification is that our deployed classifier h should not make explicit use of group membership. This is a moral analogue of the goal of "anonymity" discussed as a fairness objective in Chapter **??**. Here we are not making decisions in a way that is independent of the individual entirely (since in our setting, individuals are not identical—they are distinguished by their features). However, it is asking that individuals who are identical with respect to their "relevant" features x should be treated the same way, independently of their group membership g:

Definition 3 (Group Independence)**.** A classifier h is group independent if it does not make decisions as a function of group membership. In other words, if for all $x \in X$:

$$h(0,x) = h(1,x)$$

This is an appealing requirement at first blush, but it is not obviously desirable in isolation (although we will return to it in a dynamic model in Section 3). Consider the following toy example:

Example 4. We have a majority population $g = 0$ with proportion $\gamma = 2/3$ and a minority population $g = 1$. There is a single binary feature: $X = \{0,1\}$. $\mathcal{D}_0$ is uniform over $\{(0,0),(1,1)\}$ (i.e. the label is perfectly correlated with the feature) and $\mathcal{D}_1$ is uniform over $\{(0,1),(1,0)\}$ (i.e. the label is perfectly anti-correlated with the feature).

The Bayes optimal classifier h^* is group dependent and has $\text{error}(h^*) = 0$. But the error optimal group independent classifier is $h(x) = x$ — i.e. it fits the majority population, and has $\text{error}_0(h) = 0$, $\text{error}_1(h) = 1$, and $\text{error}(h) = 1 - \gamma = 1/3$. Here (in any setting in which higher error on a population corresponds to harm), requiring group independence has harmed the minority population without changing how the classifier behaves on the majority population, and decreasing its overall performance.

As Example 4 demonstrates, attempting to enforce fairness by prohibiting a classifier from using certain inputs ("fairness through blindness") can backfire, and a more promising way forward is to approach fairness by enunciating what properties of the outputs are undesirable. Doing this correctly can be tricky, and is context dependent. In the following, we go through the exercise of articulating the rationale for several popular formalizations of "statistical fairness". These differ in spirit from the notion of "envy freeness" introduced as a measure of fairness in Chapter **??** in that we do not necessarily have the aim of giving every individual what they want: instead, these measures are concerned with how different measures of the *mistakes* made by the classifier are distributed across populations. Notions like envy freeness are not appropriate when the designer's goal is explicitly to distribute some *bad* like incarceration: the incarcerated will always prefer not to be incarcerated, but committing ahead of time to incarcerate nobody (or everybody — the only two deterministic envy free solutions) is likely at odds with society's objectives.

1.1.1 Thinking about Fairness Constraints

Any statistical estimation procedure will inevitably make errors, and depending on the setting, those errors can cause personal harms. For example, in a criminal justice application in which we are making decisions about incarceration, we may judge that the individuals who are harmed the most are those who should *not* have been incarcerated (say, because they are innocent) but are mistakenly incarcerated. These kinds of errors can be cast as *false positives.* In contrast, in settings in which we are allocating a good — say when making a hiring decision or admitting students to college — we may judge that the individuals who are most harmed are those who should have received the good (because they were qualified for the job, e.g.) but were mistakenly denied it. These kinds of errors can be cast as *false negatives.* A sensible and popular approach to fairness questions is to ask that the harms caused by the mistakes of a classifier not be disproportionately borne by one population. This approach motivates error rate balance constraints:

Definition 5 (Error Rate Balance)**.** A hypothesis h satisfies false positive rate balance if:

$$\Pr_{(x,y)\sim\mathcal{D}_0}[h(x) = 1|y = 0] = \Pr_{(x,y)\sim\mathcal{D}_1}[h(x) = 1|y = 0]$$

It satisfies false negative rate balance if:

$$\Pr_{(x,y)\sim\mathcal{D}_0}[h(x)=0|y=1] = \Pr_{(x,y)\sim\mathcal{D}_1}[h(x)=0|y=1]$$

If h satisfies both conditions we say that it satisfies error rate balance. These definitions have natural approximate relaxations — rather than requiring exact equality, we can require that the difference in false positive or false negative rates across populations not exceed some threshold ϵ.

Of course, not all statistical estimators are directly used to take action: a (currently) more common use case is that statistical estimators are used to inform some downstream decision — often made by a human being — that will involve many sources of information. In such cases, we cannot directly attribute harms that result from the eventual decisions to mistakes made by the classification technology, and so it is difficult to ask that the "harms" due to the mistakes in classification be equally borne by all populations. For these mid-stream statistical estimators, we might instead ask that the predictions they make be *equally informative* for both populations. In other words, the *meaning of the inference that we can draw about the true label from the prediction should be the same for both populations.*

Definition 6 (Informational Balance)**.** A hypothesis h satisfies positive informational balance if:

$$\Pr_{(x,y)\sim\mathcal{D}_0}[y=1|h(x)=1] = \Pr_{(x,y)\sim\mathcal{D}_1}[y=1|h(x)=1]$$

It satisfies negative informational balance if:

$$\Pr_{(x,y)\sim\mathcal{D}_0}[y=1|h(x)=0] = \Pr_{(x,y)\sim\mathcal{D}_1}[y=1|h(x)=0]$$

If h satisfies both conditions, we say that it satisfies informational balance. Just as with error rate balance, we can easily define an approximate relaxation parameterized by an error tolerance ϵ.

We could continue and come up with additional fairness desiderata, but this will be plenty for this chapter.

2 The Need to Choose

We have enunciated two distinct — and reasonable — notions of balance: error rate balance and informational balance. Which of these (if any) should we impose on our statistical models? Or perhaps there is no need to choose — both conditions are reasonable, so why not ask for both?

It turns out that doing so is simply impossible, under generic conditions on the prediction problem. An important parameter for understanding this issue will be the *base rate* in each population g, which is simply the proportion of the population that has a label of 1:

Definition 7 (Base Rate)**.** The base rate of population g is:

$$B_g = \Pr_{(x,y)\sim\mathcal{D}_g}[y=1]$$

The next observation we can make is that the statistical quantities used to define error rate balance and informational balance are both conditional probabilities that are Bayes duals of one another — that is, they are directly related to each other via Bayes' rule:

Claim 8 (Bayes' Rule)**.**

$$\Pr_{(x,y)\sim\mathcal{D}_g}[y=1|h(x)=0] = \frac{\Pr_{(x,y)\sim\mathcal{D}_g}[h(x)=0|y=1]\cdot\Pr_{(x,y)\sim\mathcal{D}_g}[y=1]}{\Pr_{(x,y)\sim\mathcal{D}_g}[h(x)=0]}$$

Let's give some short hand for false positive and false negative rates, the quantities that will be of interest to us:

$$\mathrm{FP}_g(h) = \Pr_{(x,y)\sim\mathcal{D}_g}[h(x)=1|y=0] \quad \mathrm{FN}_g(h) = \Pr_{(x,y)\sim\mathcal{D}_g}[h(x)=0|y=1]$$

Note that the quantity appearing in the denominator, $\Pr_{(x,y)\sim\mathcal{D}_g}[h(x)=0]$, can be expanded out by observing that there are two ways in which a classifier $h(x)$ can predict 0. Either the true label is 1, and the classifier makes an error (a false negative), or the true label is 0, and the classifier is correct — i.e. it did not make a false positive error. In other words:

$$\Pr_{(x,y)\sim\mathcal{D}_g}[h(x)=0] = B_g\cdot\mathrm{FN}_g(h) + (1-B_g)(1-\mathrm{FP}_g(h))$$

We can now rewrite the right hand side of Bayes rule as follows:

$$\Pr_{(x,y)\sim\mathcal{D}_g}[y=1|h(x)=0] = B_g\cdot\left(\frac{\mathrm{FN}_g(h)}{B_g\cdot\mathrm{FN}_g(h)+(1-B_g)(1-\mathrm{FP}_g(h))}\right) \quad (1)$$

Observe that if h satisfies informational balance, the left hand side of the equation is equal across groups, and therefore the right hand side must be as well:

$$B_0\cdot\left(\frac{\mathrm{FN}_0(h)}{B_0\cdot\mathrm{FN}_0(h)+(1-B_0)(1-\mathrm{FP}_0(h))}\right) = B_1\cdot\left(\frac{\mathrm{FN}_1(h)}{B_1\cdot\mathrm{FN}_1(h)+(1-B_1)(1-\mathrm{FP}_1(h))}\right)$$

Now suppose h also satisfies error rate balance (i.e. $\mathrm{FP}_0(h) = \mathrm{FP}_1(h)$ and $\mathrm{FN}_0(h) = \mathrm{FN}_1(h)$). When can it be the case that the above equality holds? By inspection, there are only two ways. It could be that the base rates are equal: $B_0 = B_1$. In this case, the left hand side is identical to the right hand side. Or, it could be that the classifier is perfect. Then, the two sides are equal even if the base rate is not, because $\mathrm{FN}_g(h) = 0$, and so both sides evaluate to 0. But these are the only two cases. These observations combine to give a basic impossibility result.

Theorem 9. *For any two groups on which the base rates are unequal ($B_0 \neq B_1$), then any hypothesis h that simultaneously achieves error rate balance and informational balance must be perfect — i.e. must be such that* $\mathrm{error}(h) = 0$.

This is an impossibility result because:

1. The first hypothesis of the theorem — that base rates are unequal — is true in almost every interesting problem, and

2. The conclusion of the theorem — that prediction is perfect — is unobtainable in almost every interesting problem.

The correct interpretation is therefore that we must almost always settle for a hypothesis that either fails to satisfy error rate balance or fails to satisfy informational balance — in other words, we are forced to choose amongst the fairness desiderata that we discussed in Section 1.1.1.

This basic fact is quite intuitive if we reflect on what Equation 1 is telling us. To compute the conditional probability $\Pr_{(x,y)\sim\mathcal{D}_g}[y = 1|h(x) = 0]$, we first start with our *prior* belief that $y = 1$, before we see the output of the classifier. But this is just the base rate B_g. Then, after we see the output of the classifier, we must update our prior belief to form our posterior belief, based on the strength of the evidence that we have observed. Equation 1 shows us that the proper way to do this is to multiply our prior belief by the Bayes factor: $\left(\frac{\mathrm{FN}_g(h)}{B_g \cdot \mathrm{FN}_g(h)+(1-B_g)(1-\mathrm{FP}_g(h))}\right)$. But the Bayes factor — i.e. the strength of the evidence — is determined by the false positive and negative rates of our classifier! Thus, the impossibility result is telling us no more than the following: If we have two groups, and we start with different prior beliefs about their labels (because the base rates differ), then if we are to have identical posterior beliefs about their labels after we see a given classifier output, it must be that the classifier provides evidence of different strength for both groups. Or, equivalently, if we have a classifier that provides evidence of the same strength for both groups, then if we started out with different prior beliefs about the groups, we will continue to have different posterior beliefs about the group after seeing the output of the classifier. This is why informational balance is necessarily at odds with error rate balance.

But don't panic — this is ok! Remember that our normative justification for error rate balance applied in settings in which the classification algorithm was actually making decisions itself, whereas our normative justification for informational balance applied in settings in which the algorithm was informing downstream decision making. But it does mean that we must be thoughtful when we design algorithms with a desire for "fairness" about how the algorithm is going to be used; different use cases call for different notions of fairness, and the safe approach of having every algorithm satisfy them all is impossible.

3 Fairness in a Dynamic Model

How should we proceed after observing the impossibility result from Section 2? The message seemed to be that we should be thoughtful and choose amongst different fairness constraints in different settings, but how precisely should we go about choosing? In this section we go through a case study of one reasonable method:

1. Model the upstream and downstream effects of choices made by selecting a particular classification technology,
2. Enunciate a societal goal that you can evaluate within this larger system, and
3. Study which constraints on the classification technology are helpful or harmful in achieving that goal.

3.1 A Toy Criminal Justice Model

We derive a simple model using the language of criminal justice, in part because this is the setting in which the conflict between of error rate balance and informational balance has been most fiercely debated. In this model, our "societal goal" will be to minimize overall crime. But the reader can map our simple model onto a lending or college admissions setting, in which the corresponding societal goal will correspond to minimizing default or maximizing preparation. We stress at the outset that this is a toy model that plainly fails to capture many important aspects of criminal justice. The point is to set up the simplest possible mathematical scenario that:

1. Allows us to consider all of the kinds of classification technologies that we have discussed so far (Bayes optimal classifiers, group independence, error rate balance, and informational balance), in a setting in which they are in tension because of Theorem 9, and
2. Allows us to shed light on the incentives engendered by imposing different fairness constraints, and how those interact with system wide goals.

With these aims in mind, we proceed.

The impossibility result in Theorem 9 begins with the premise that different populations have different *base rates* — i.e. different proportions of positive labels. But when data points correspond to people, as they do when deploying classification technologies in criminal justice settings, base rates are just aggregates over lots of individual decisions. To model how these individual decisions are made, we will think of individuals as rational decision makers, who make a binary decision (crime vs. no crime) by weighing their expected utility conditioned on both choices, and making the decision that corresponds to the higher expected utility. In this sense, we model people as being identical to one another. However, they differ in the opportunities available to them: we model individuals as having some *outside option value* (or *legal employment opportunity*) that is drawn from a distribution that may be specific to their group g, thereby modelling the fact that different populations may have different legal opportunities available to them (eventually making crime relatively more appealing to some people than others).

Definition 10 (Outside Option Distributions)**.** Each individual from group g has an *outside option value* v drawn independently from a real valued distribution $v \sim \mathcal{D}_g$, which may differ by group.

Individuals will take an action $a \in \{C, N\}$ ((C)rime and (N)o Crime). If they decide not to commit a crime, they obtain their outside option value. If they decide to commit a crime, they obtain value I. (We could equally well have I be drawn from a group dependent distribution, but for simplicity we let it be a fixed value here). They will also experience some penalty P if they are incarcerated, which will occur with some probability that depends on their decision and on the classification technology that society ends up deploying, which we will elaborate on shortly. As a function of their action a, we obtain a "signal" i.e. some noisy information about what action they took. We can view this as an abstraction of the "evidence" that a crime was committed: allowing the evidence to be noisy takes into account both that criminals can go free for lack of evidence, and that innocent people can end up being jailed because of misleading evidence. As we shall see, it is also from this noise that the risk of stereotyping based on group membership g arises.

Definition 11 (Signal Distributions)**.** An individual who takes action a generates a *signal* s drawn independently from a real valued distribution $s \sim \mathcal{Q}_a$.

For the sake of intuition, it is helpful to imagine that larger signals correspond to stronger evidence of guilt and vice versa. This would be the case if e.g. $\mathcal{Q}_C$ first order stochastically dominates $\mathcal{Q}_D$ — but we won't actually need to make this assumption. Note however that we *are* making a crucial assumption here: namely, that signals depend *only* on the action that an individual takes, and in particular are conditionally independent of their group given their action. This assumption need not hold in practice, if e.g. evidence is gathered by a method that itself encodes bias.

There will be some classification technology (an extremely reduced form representation of the criminal justice system generally) that for each individual, takes as input the signal s that they generated, and then makes an incarceration decision — possibly as a function of their group membership g.

Definition 12 (Incarceration Rule)**.** An incarceration rule $h : \mathbb{R} \times \{0, 1\} \rightarrow \{0, 1\}$ takes as input a signal s and a group membership g and outputs an incarceration decision $h(s, g)$, where $h(s, g) = 1$ corresponds to incarceration.

Note that an incarceration rule fixes a false positive rate and false negative rate across the two populations. For each group g:

$$\mathrm{FP}_g(h) = \Pr_{s \sim \mathcal{Q}_N}[h(s, g) = 1] \quad \mathrm{FN}_g(h) = \Pr_{s \sim \mathcal{Q}_C}[h(s, g) = 0]$$

Once an incarceration rule h is fixed, we can speak about the expected payoff of an agent's actions, who has outside option value v and is a member of group g. Such an agent's payoff for choosing $a = C$ is:

$$u(g, v, h, C) = I - P \cdot \Pr_{s \sim \mathcal{Q}_C}[h(s, g) = 1] = I - P \cdot (1 - \mathrm{FN}_g(h))$$

In other words, they immediately get payoff I for choosing to commit a crime, but then receive penalty P in the event that they are incarcerated, which occurs exactly when they are *not* a false negative.

Similarly, the agent's payoff for choosing $a = N$ is:

$$u(g, v, h, N) = v - P \cdot \Pr_{s \sim \mathcal{Q}_N}[h(s, g) = 1] = v - P \cdot \mathrm{FP}_g(h)$$

In other words, they immediately get payoff equal to their outside option value v when they do not commit a crime, and receive a penalty P in the event that they are incarcerated, which occurs exactly when they *are* a false positive.

Thus, in our model, an individual will commit a crime when $u(g, v, h, C) \geq u(g, v, h, N)$, which by rearranging the expressions occurs exactly when:

$$v \leq I + P(\mathrm{FP}_g(h) + \mathrm{FN}_g(h) - 1)$$

Finally, this allows us to bring the model full circle and compute the base rates $B_g(h)$ in each population, which in our model are a function of the deployed classification technology h. We have that the base rate in population g is:

$$B_g(h) = \Pr_{v \sim \mathcal{D}_g}[v \leq I + P(\mathrm{FP}_g(h) + \mathrm{FN}_g(h) - 1)]$$

which is just the CDF of the outside option distribution $\mathcal{D}_g$ evaluated at $I + P(\mathrm{FP}_g(h) + \mathrm{FN}_g(h) - 1)$.

We pause here to make a couple of observations.

1. First, because the outside option distributions $\mathcal{D}_0$ and $\mathcal{D}_1$ are not equal, in general, the base rates $B_0(h)$ and $B_1(h)$ will also not be equal. Thus we are in a setting in which the impossibility result from Theorem 9 applies.

2. Because signal distributions are conditionally independent of group membership conditional on actions, we can equalize false positive and false negative rates across groups simply by selecting an incarceration rule that ignores group membership — i.e. an h such that for all s, $h(s, 0) = h(s, 1)$.

3. On the other hand, from Bayes rule, we know that:

$$\Pr[a = C|s, g] = B_g(h) \cdot \left(\frac{\Pr_{\mathcal{Q}_C}[s]}{B_g(h) \cdot \Pr_{\mathcal{Q}_C}[s] + (1 - B_g(h)) \cdot \Pr_{\mathcal{Q}_N}[s]} \right)$$

   which depends on g, exactly because base rates will differ across groups. In other words, when base rates differ, group membership really is statistically informative information about whether an individual has committed

> a crime or not, because it affects prior beliefs and therefore posterior beliefs. Therefore, the incarceration rule h that *minimizes overall classification error* — i.e. that is most likely to incarcerate the guilty and release the innocent — will not be independent of group membership, and therefore will not equalize false positive and negative rates.

This is all to say that we have established a model in which the selection of a classifier feeds back into the decisions of individuals, which in turn affects base rates, which in turn affects what an informative classifier must do — but we have found ourselves where we started, in which attempts to equalize harm will lead to decisions that are differently informative across populations and vice versa.

In this model, however, we can enunciate different goals compared to what we could ask for in a static model. For example, because base rates now depend dynamically on our choice of incarceration rule h, we can ask what properties h should have if we wish to *minimize crime rates.* For example, we might ambitiously hope that some incarceration rule h^* would result in simultaneously minimizing crime rates across both populations, i.e.:

$$h^* \in \arg\min_h B_0(h) \quad h^* \in \arg\min_h B_1(h)$$

So how should we do this? And is it even possible to simultaneously minimize base rates across two different populations, with different outside option distributions? First, recall that the base rate $B_g(h)$ for population g is simply the CDF of the distribution $\mathcal{D}_g$ evaluated at $I + P(\mathrm{FP}_g(h) + \mathrm{FN}_g(h) - 1)$. We can therefore think about how to minimize the crime rate within each population without needing to understand much about the particulars of the distributions $\mathcal{D}_g$, because cumulative distribution functions are monotone. Therefore to minimize the base rate, we wish to minimize $I + P(\mathrm{FP}_g(h) + \mathrm{FN}_g(h) - 1)$, and in this expression, only two of the parameters are under our control, and neither of them depends on $\mathcal{D}_g$. Therefore we can find h^* that minimizes the crime rate $B_g(h)$ on population g by solving for:

$$h^*(\cdot, 0) = \arg\min_{h(\cdot,0)} \mathrm{FP}_0(h) + \mathrm{FN}_0(h) \quad h^*(\cdot, 1) = \arg\min_{h(\cdot,1)} \mathrm{FP}_1(h) + \mathrm{FN}_1(h)$$

Note that this is not the same thing as minimizing overall error, because we are not weighting false positive and false negative rates by base rates.

Because signal distributions depend only on the action a taken by an individual, and not (directly) on group membership, the two minimization problems above must have exactly the same solution. The result is that the optimal incarceration rule h^* must be group independent — i.e. it must be that for all s, $h^*(s, 0) = h^*(s, 1)$. Recall that this also implies that the optimal solution equalizes false positive and false negative rates across populations. Thus we have obtained the following theorem:

Theorem 13. *For any set of outside option distributions $\mathcal{D}_0, \mathcal{D}_1$ and for any set of signal distributions $\mathcal{Q}_C, \mathcal{Q}_N$, there is a classifier h^* that simultaneously* *minimizes crime rates across both groups:*

$$h^* \in \arg\min_h B_0(h) \quad h^* \in \arg\min_h B_1(h)$$

and h^* has the following properties:

1. *h^* is independent of group membership g (even though g is statistically informative). In other words, for every s:*

   $$h^*(s, 0) = h^*(s, 1)$$

2. *h^* equalizes false positive and false negative rates across groups:*

   $$\mathrm{FP}_0(h^*) = \mathrm{FP}_1(h^*) \quad \mathrm{FN}_0(h^*) = \mathrm{FN}_1(h^*)$$

   in other words it satisfies error rate balance.

3.2 Interpreting Theorem 13

As we have already observed, when base rates differ across groups (as they do in this setting), Bayes rule tells us that the best classifier, from the point of view of minimizing the number of mistaken predictions we will make, will make decisions as a function of group membership. Thus, the classifier h^* in our model that is best from the perspective of minimizing crime rates, is doing something that seems odd at first blush: it is *intentionally* committing not to make decisions as a function of an informative variable g, and is instead only making decisions as a function of s. How come?

The answer sheds some light on why we view stereotyping — i.e. using group membership to inform our decision making — as unfair in the first place. Although g is a statistically informative variable, it is immutable and not under the control of the individual. On the other hand, the signal s *is* under the control of the individual, via their choice of action. Hence, by basing our decisions on g, we are distorting individual incentives — essentially reducing the disincentive that individuals from populations with higher base rates have to not commit crimes in the first place. If we used the classifier optimized purely for predictive performance, we would convict individuals from the population with reduced access to legal opportunities based on evidence that we would view as insufficiently strong to convict individuals from the higher opportunity population. This in turn would increase the crime rate in the lower opportunity population by increasing the false positive rate, thereby distorting the incentives of the criminal justice system. In our model, minimizing crime rates is all about correctly setting the incentives of the criminal justice system — and the way to provide equal incentives to both populations is to commit to ignoring immutable individual characteristics. In other words, stereotyping or racial profiling (i.e. doing statistical inference based on immutable group characteristics) can be statistically justified in a static model, but in a dynamic model serves to distort incentives in a way that has pernicious effects in equilibrium.

We have also confirmed our intuition (within this model, of course) that it is error rate balance that matters *when our classifier is being used to make utility relevant decisions for people.* On the other hand, if our statistical model is being used merely to *inform* future decisions, then we likely want something more like informational balance. Here, the *signal* s plays the role of the statistical instrument used to inform future decisions. Theorem 13 crucially relied on the fact that the signal distributions $\mathcal{Q}_C, \mathcal{Q}_N$ were equally informative about agent actions, independently of the group g to which the agent belonged. This is what allowed equal treatment of *signals* to correctly incentivize agents across groups; if the signal distributions had not been equally informative for members of both groups, then the downstream conclusion would also have failed to hold. This suggests that if we find ourselves in the position of designing a statistical model that serves the role of the signal, that we should strive for (some analogue of) informational balance.

4 Preserving Information Before Decisions

In Section 1.1.1 we have argued on normative grounds that when we are *taking actions* in consequential domains in which our errors lead to personal harms, that we might want to deploy decision rules that optimize error subject to error rate balance constraints; in Section 3 we rigorously justified this in a particular dynamic model. But currently, most deployed statistical models are not used to directly take action themselves, but rather to inform some downstream decision making task. This is the case, e.g. with the criminal recidivism prediction tools that have obtained scrutiny within the algorithmic fairness literature: they are used to inform bail and parole decisions, but those decisions are ultimately made by a human judge with access to other information as well. When this is the case, we might perhaps be better served by asking that our statistical models preserve enough information about each group g so that it is possible to implement the error optimal decision rule, subject to some fairness constraint like error rate balance, downstream at the actual decision making process. What information is needed?

As we already observed, a sufficient statistic for computing the error optimal classifier (i.e. the "Bayes Optimal Classifier") on any distribution over points is the conditional label expectation: $f(x, g) = \mathbb{E}_{\mathcal{D}_g}[y|x]$. Here we briefly observe that this is also a sufficient statistic for computing the error optimal classifier subject to error rate balance constraints. Observe that we can write the error rate of a binary classifier as:

$$\text{error}(h) = \mathbb{E}[h(x) \cdot (1 - \mathbb{E}[y|x]) + (1 - h(x)) \cdot \mathbb{E}[y|x]]$$

Note that minimizing error(h) is equivalent to minimizing $\mathbb{E}[h(x) \cdot (1 - 2\mathbb{E}[y|x])]$. Thus we can describe the error optimal classifier subject to error rate balance constraints as the solution to:

$$\text{Minimize}_h \quad \mathbb{E}[h(x) \cdot (1 - 2\mathbb{E}[y|x])]$$

Such that:

$$\mathbb{E}_{\mathcal{D}_0}[h(x)|y=0] = \mathbb{E}_{\mathcal{D}_1}[h(x)|y=0] \tag{λ_1}$$

$$\mathbb{E}_{\mathcal{D}_0}[1-h(x)|y=1] = \mathbb{E}_{\mathcal{D}_1}[1-h(x)|y=1] \tag{λ_2}$$

This is a linear program over the set of distributions over hypotheses h. Let λ_1 and λ_2 be the optimal dual solution to this linear program. By Lagrangian duality, we have that the error optimal classifier subject to error rate balance constraints is a minimizer of:

$$\mathbb{E}\Big[h(x)\cdot\Big((1-2\mathbb{E}[y|x]) + \lambda_1(1-\mathbb{E}[y|x])(\mathbb{1}[g=0]-\mathbb{1}[g=1])$$

$$+\lambda_2(\mathbb{E}[y|x](\mathbb{1}[g=1]-\mathbb{1}[g=0]))\Big)\Big]$$

In other words, the optimal such classifier h^* must satisfy:

$$h^*(x,g) = \begin{cases} 1, & \text{If } g=0 \text{ and } (2+\lambda_1+\lambda_2)\mathbb{E}_{\mathcal{D}_0}[y|x] > 1+\lambda_1; \\ 0, & \text{If } g=0 \text{ and } (2+\lambda_1+\lambda_2)\mathbb{E}_{\mathcal{D}_0}[y|x] < 1+\lambda_1; \\ 1, & \text{If } g=1 \text{ and } (2-\lambda_1-\lambda_2)\mathbb{E}_{\mathcal{D}_1}[y|x] > 1-\lambda_1; \\ 0, & \text{If } g=1 \text{ and } (2-\lambda_1-\lambda_2)\mathbb{E}_{\mathcal{D}_1}[y|x] < 1-\lambda_1. \end{cases}$$

(If none of these conditions are satisfied, the optimal classifier may need to randomize). From this we learn that the optimal classifier h^* remains a thresholding on the conditional label expectation, even under error rate balance constraints. Therefore, to allow a downstream decision maker to deploy an optimal classifier subject to error rate balance, it continues to suffice that our statistical estimator f correctly encode the conditional label expectations: $f(x,g) = \mathbb{E}_{\mathcal{D}_g}[y|x]$.

On the other hand, it is not hard to see that if we must learn a *binary* model h (e.g. by thresholding $\mathbb{E}_{\mathcal{D}_g}[y|x]$ or via any other method), then in most cases we will have destroyed the information needed to implement the optimal classifier h^* satisfying error rate balance downstream. Similarly, modifying the statistical estimator $f(x,g)$ such that it deviates from $f(x,g) = \mathbb{E}_{\mathcal{D}_g}[y|x]$ will generically preclude us from being able to implement the optimal classifier h^* subject to error rate balance downstream. This suggests a general rule of thumb:

> Constraints like error rate balance should be applied at the very end of decision making pipelines, at the moment that actions are taken that have the potential to harm people. At intermediate stages of decision making pipelines, we should strive to capture as accurate statistical information about the population as possible (ideally $\mathbb{E}_{\mathcal{D}_g}[y|x]$) — because failing to do this harms not only accuracy, but also our ability to usefully impose fairness constraints downstream.

But how should we think about realizing this rule of thumb? In general, we cannot hope to learn $\mathbb{E}_{\mathcal{D}_g}[y|x]$ from data, because if X is a large feature space, then we will see each particular feature vector x only infrequently, and so we will have few to no samples of the conditional label distributions, conditional on

x. At best we can hope to learn some "good" proxy $\bar{f}(x, g)$. But what makes a good proxy? In the following for simplicity, we will assume that g is encoded in the feature vector x so that we may write simply $f(x)$ and $\mathbb{E}[y|x]$.

Suppose we are given a function $\bar{f}$ that purports to represent conditional label distributions: $\bar{f}(x) = \mathbb{E}[y|x]$. How can we attempt to falsify this assertion? Here is one family of tests. Consider any subset $S \subseteq X$ of the feature space. If $\bar{f}(x) = \mathbb{E}[y|x]$, then we will have:

$$\mathbb{E}[y|x \in S] = \mathbb{E}[\bar{f}(x)|x \in S].$$

If $\bar{f}$ fails to satisfy this condition on any set S, then this certifies that f does not correctly represent the conditional label distribution. Moreover, this is a condition that we can easily test (approximately) from data, because for any set S with sufficiently large probability in the underlying distribution $\mathcal{D}$, we can estimate conditional expectations accurately from sample quantities. Suppose we have some very large collection G of such sets S: $G \subseteq 2^X$. This would parameterize a suite of statistical tests aimed at falsifying the conjecture that $\bar{f}$ correctly encoded the conditional label distribution, and we could ask that we produce estimators $\bar{f}$ that pass every statistical test in this suite — i.e. that satisfy $\mathbb{E}[y|x \in S] = \mathbb{E}[\bar{f}(x)|x \in S]$ simultaneously for every $S \in G$. At the very least we can check from data if our classifiers satisfy these conditions. Can we find classifiers that satisfy them as well? We can phrase this as an optimization problem:

$$\min_{\bar{f}} \max_{S \in G} \left|\mathbb{E}[y|x \in S] - \mathbb{E}[\bar{f}(x)|x \in S]\right|$$

There is a solution $\bar{f}$ that obtains optimal objective value 0 — i.e. the true conditional label distribution — and so the only question is whether we can efficiently find some $\bar{f}$ that does well according to this objective.

Remarkably, it turns out that we can — although the details of how are beyond the scope of this chapter (but see the references for further reading). We call ϵ approximate solutions to this problem ϵ-*multiaccurate* estimators $\bar{f}$ with respect to the collection of groups G:

Definition 14. A statistical estimator $\bar{f} : X \to \mathbb{R}$ is ϵ-multiaccurate with respect to a collection of groups $G \in 2^X$ if for every $S \in G$:

$$\left|\mathbb{E}[y|x \in S] - \mathbb{E}[\bar{f}(x)|x \in S]\right| \leq \epsilon$$

There are efficient algorithms that learn approximately multiaccurate estimators for any collection of sets G with sufficiently large measure, from datasets that have size only polynomial in $1/\epsilon$ and $\log |G|$. The logarithmic dependence on $|G|$ means that we can learn estimators that pass an exponentially large number of "sanity check" statistical tests — one for each of the groups $S \in G$.

Multiaccuracy is one of a family of "informational balance" constraints that are sensible to ask mid-stream statistical estimators to satisfy. This family of constraints can be generalized by enlarging the set of statistical tests that we insist that our estimator $\bar{f}$ satisfy, each aimed at falsifying the conjecture that

$\bar{f}$ correctly encodes the conditional label expectations. In addition to enlarging the collection of sets G that define our tests, we can ask that $\bar{f}$ pass similar statistical tests over sets S that are defined not just by the features x, but also by our predictions $\bar{f}(x)$ — this leads to a notion of statistical balance called *multi-calibration*. We can also ask for statistical tests based not just on the expectations of label distributions, but based on variances and other higher moments. Or we can ask that our statistical estimators be indistinguishable from the true conditional label distribution with respect to the actions taken by some well defined set of downstream decision makers who are informed by our predictions. All of these measures can not just be checked, but also guaranteed by learning procedures that have only small amounts of data — logarithmic in $|G|$, which allows us to ask for consistency with respect to an exponentially large collection of sets G. If we consider the sets in G as themselves representing demographic groups g that we wish to protect, multiaccuracy give us a way to think about enforcing statistical fairness constraints over large collections of finely defined and potentially overlapping groups, which can save us from the need to anticipate ahead of time every group g for which we might desire (e.g.) error rate balance in some downstream classification task.

References and Further Reading

The importance of error rate balance and informational balance were dramatically brought to the public consciousness in a 2016 Propublica article (Angwin et al., 2016) investigating bias in the COMPAS recidivism prediction tool that was used to inform bail and parole decisions in a number of US jurisdictions. The impossibility result that we derive in Section 2 was originally proven in Chouldechova (2017). A similar result for real valued predictors was proven by Kleinberg et al. (2017). The dynamic model from Section 3 corresponds to the "baseline model" from Jung et al. (2020). In (Jung et al., 2020), the model is generalized in various ways, including to cover the case in which signal distributions *do* depend on group membership, and the case in which observation of signal distributions is mediated by a third party with its own incentives (e.g. the police). That the conditional label distribution is a sufficient statistic to the optimal classifier subject to error rate balance constraints has been observed by several authors, including Hardt et al. (2016) and Corbett-Davies et al. (2017). The idea of multiaccuracy and multicalibration was proposed by Hébert-Johnson et al. (2018). Jung et al. (2021) generalize the notion of multicalibration from means to variances and other higher moments of the label distribution. Gupta et al. (2021) further generalize this idea and define "multivalid" statistical estimators of different sorts, including multivalid prediction intervals which can obtain tight 95% coverage intervals over large numbers of intersecting demographic groups. Zhao et al. (2021) define a multicalibration like notion of consistency that is defined with respect to a class of downstream decision makers. It is also possible to define analogues of error rate balance with respect to large collections of overlapping groups G, just as multiaccuracy and multicalibration provide analogues of informational balance in this setting — see Kearns et al. (2018). For a fuller popular treatment of the issues discussed in this Chapter, see Kearns and Roth (2019). In this chapter, we have focused on statistical estimation problems — although these sometimes are related to allocation problems, we have ignored capacity or supply constraints. We refer the reader to Elzayn et al. (2019); Donahue and Kleinberg (2020); Sinclair et al. (2020); Finocchiaro et al. (2021) for discussion of fairness in allocation problems with capacity constraints. There are also other papers that study related fairness desiderata in game theoretic settings: we refer the reader to Kannan et al. (2017); Hu and Chen (2018); Milli et al. (2019); Hu et al. (2019); Liu et al. (2020); Kannan et al. (2021) for other work in this style.

References

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. 2016. Machine bias. *ProPublica, May*, **23**, 2016.

Chouldechova, Alexandra. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, **5**(2), 153–163.

Corbett-Davies, Sam, Pierson, Emma, Feller, Avi, Goel, Sharad, and Huq, Aziz. 2017. Algorithmic decision making and the cost of fairness. Pages 797–806 of: *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining.*

Donahue, Kate, and Kleinberg, Jon. 2020. Fairness and utilization in allocating resources with uncertain demand. Pages 658–668 of: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.*

Elzayn, Hadi, Jabbari, Shahin, Jung, Christopher, Kearns, Michael, Neel, Seth, Roth, Aaron, and Schutzman, Zachary. 2019. Fair algorithms for learning in allocation problems. Pages 170–179 of: *Proceedings of the Conference on Fairness, Accountability, and Transparency.*

Finocchiaro, Jessie, Maio, Roland, Monachou, Faidra, Patro, Gourab K, Raghavan, Manish, Stoica, Ana-Andreea, and Tsirtsis, Stratis. 2021. Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness. Pages 489–503 of: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.*

Gupta, Varun, Jung, Christopher, Noarov, Georgy, Pai, Mallesh M, and Roth, Aaron. 2021. Online Multivalid Learning: Means, Moments, and Prediction Intervals. *arXiv preprint arXiv:2101.01739.*

Hardt, Moritz, Price, Eric, and Srebro, Nathan. 2016. Equality of opportunity in supervised learning. Pages 3323–3331 of: *Proceedings of the 30th International Conference on Neural Information Processing Systems.*

Hébert-Johnson, Úrsula, Kim, Michael, Reingold, Omer, and Rothblum, Guy. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. Pages 1939–1948 of: *International Conference on Machine Learning.*

Hu, Lily, and Chen, Yiling. 2018. A Short-term Intervention for Long-term Fairness in the Labor Market. Lyon, France: IW3C2.

Hu, Lily, Immorlica, Nicole, and Vaughan, Jennifer Wortman. 2019. The disparate effects of strategic manipulation. Pages 259–268 of: *Proceedings of the Conference on Fairness, Accountability, and Transparency.*

Jung, Christopher, Kannan, Sampath, Lee, Changhwa, Pai, Mallesh M., Roth, Aaron, and Vohra, Rakesh. 2020. Fair Prediction with Endogenous Behavior. Pages 677–678 of: Biró, Péter, Hartline, Jason, Ostrovsky, Michael, and Procaccia, Ariel D. (eds), *EC '20: The 21st ACM Conference on Economics and Computation, Virtual Event, Hungary, July 13-17, 2020.* ACM.

Jung, Christopher, Lee, Changhwa, Pai, Mallesh M, Roth, Aaron, and Vohra, Rakesh. 2021. Moment Multicalibration for Uncertainty Estimation. In: *COLT'21: The 34th Annual Conference on Learning Theory.*

Kannan, Sampath, Kearns, Michael, Morgenstern, Jamie, Pai, Mallesh, Roth, Aaron, Vohra, Rakesh, and Wu, Zhiwei Steven. 2017. Fairness incentives for myopic agents. Pages 369–386 of: *Proceedings of the 2017 ACM Conference on Economics and Computation.*

Kannan, Sampath, Niu, Mingzi, Roth, Aaron, and Vohra, Rakesh. 2021. Best vs. All: Equity and Accuracy of Standardized Test Score Reporting. *arXiv preprint arXiv:2102.07809.*

Kearns, Michael, and Roth, Aaron. 2019. *The ethical algorithm: The science of socially aware algorithm design.* Oxford University Press.

Kearns, Michael, Neel, Seth, Roth, Aaron, and Wu, Zhiwei Steven. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. Pages 2564–2572 of: *International Conference on Machine Learning.*

Kleinberg, Jon M., Mullainathan, Sendhil, and Raghavan, Manish. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. Pages 43:1–43:23 of: Papadimitriou, Christos H. (ed), *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA.* LIPIcs, vol. 67. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Liu, Lydia T, Wilson, Ashia, Haghtalab, Nika, Kalai, Adam Tauman, Borgs, Christian, and Chayes, Jennifer. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. Pages 381–391 of: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.*

Milli, Smitha, Miller, John, Dragan, Anca D, and Hardt, Moritz. 2019. The social cost of strategic classification. Pages 230–239 of: *Proceedings of the Conference on Fairness, Accountability, and Transparency.*

Sinclair, Sean R, Jain, Gauri, Banerjee, Siddhartha, and Yu, Christina Lee. 2020. Sequential Fair Allocation of Limited Resources under Stochastic Demands. *arXiv preprint arXiv:2011.14382.*

Zhao, Shengjia, Kim, Michael P, Sahoo, Roshni, Ma, Tengyu, and Ermon, Stefano. 2021. Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibration. *arXiv preprint arXiv:2107.05719.*