Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Auditing and Achieving Counterfactual Fairness

Presented by: Alan Mishler

Accepted by: Department of Statistics & Data Science

Readers:

_____
Edward Kennedy, Advisor

_____
Alexandra Chouldechova

_____
Aaditya Ramdas

_____
Cosma Rohilla Shalizi

_____
Ilya Shpitser

_____
Larry Wasserman

Approved by the Committee on Graduate Degrees:

_____
Richard Scheines, Dean Date

Carnegie Mellon University

Auditing and Achieving Counterfactual Fairness

A Dissertation Submitted to the Graduate School
in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy
in
Statistics & Data Science
by

Alan Mishler
Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

August 2021

Acknowledgements

Thanks to my loved ones, family and friends, who supported me through this journey and made it worthwhile. Thanks to the professors and scholars who have inspired and guided me, especially my advisor, Edward Kennedy.

Chapter 1 was presented with Niccolò Dalmasso at the 2019 NeurIPS workshop *"Do the right thing": machine learning and causal inference for improved decision making.*

Chapter 2 was presented with Edward Kennedy and Alexandra Chouldechova at the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT).

Don't be afraid to be confused. Try to remain permanently confused.

— George Saunders, The New Mecca

"If you can see a thing whole," he said, "it seems that it's always beautiful. Planets, lives... But close up, a world's all dirt and rocks. And day to day, life's a hard job, you get tired, you lose the pattern. You need distance, interval. The way to see how beautiful earth is, is to see it from the moon."

— Ursula K. Le Guin, The Dispossessed

To be running breathlessly, but not yet arrived, is itself delightful, a suspended moment of living hope.

— Anne Carson, Eros the Bittersweet

Abstract

Machine learning is increasingly involved in high stakes decisions in domains such as healthcare, criminal justice, and consumer finance. In these settings, ML models often take the form of Risk Assessment Instruments (RAIs): given covariates such as demographic information and an individual's medical/criminal/financial history, the model predicts the likelihood of an adverse outcome, such as a dangerous medical event, recidivism, or default on a loan. Rather than rendering an automatic decision, the model produces a "risk score," which a decision maker may take into account when deciding whether to prescribe a medical treatment, release a defendant on bail, or issue a personal loan.

The proliferation of machine learning has raised concerns that learned models may be discriminatory with respect to sensitive features like race, sex, age, and socioeconomic status. These concerns have led to an explosion of methods in recent years for developing fair models and auditing the fairness of existing models. The most widely discussed fairness criteria impose constraints on the joint distribution of a sensitive feature, a predictor, and an outcome. These "observational" fairness criteria are inappropriate for RAIs, however. RAIs are not concerned with the *observable outcomes* in the training data ("Did patients of this type historically experience serious complications?"), which are themselves a product of historical treatment decisions. Rather, they are concerned with the *potential outcomes* associated with available treatment decisions ("*Would* patients of this type experience complications *if not treated?*"). Because treatments are not assigned at random—doctors naturally treat the patients they think are at high risk—these are distinct questions.

In this thesis, I consider counterfactual versions of common algorithmic fairness criteria, which are defined with respect to potential rather than observable outcomes. I develop methods to audit the fairness of existing predictors and build predictors which satisfy these fairness criteria.

In Chapter 1, I show how the use of observable rather than potential outcomes in algorithmic RAIs can lead to worse outcomes compared to before the RAI was trained. In Chapter 2, I develop a post-processing procedure that can render an existing binary predictor fair with respect to counterfactual equalized odds, while maximizing its counterfactual accuracy. This procedure yields predictors

whose excess risk and excess unfairness decay at $\sqrt{n}$ rates when nuisance parameters are estimated sufficiently fast. I also provide estimators of the counterfactual risk and error rates of a large class of (possibly randomized) binary classifiers. These estimators are $\sqrt{n}$-consistent and asymptotically normal under similar assumptions. I show that the post-processing procedure improves fairness on both simulated and real data, and that this does not necessarily incur a substantial decrease in accuracy.

In Chapter 3, I develop a flexible framework for building predictors that are fair and accurate with respect to either observable or counterfactual outcomes. Within this framework, I propose three methods: the first minimizes risk subject to fairness constraints, the second minimizes unfairness subject to risk constraints, and the third incorporates a set of fairness penalty parameters that allow users to efficiently build large sets of predictors that trace out different paths in fairness-accuracy space. These methods accommodate users who wish to improve the fairness of an existing model without sacrificing accuracy, or vice versa. They also allows users to explore the tradeoffs between fairness and accuracy and between different fairness criteria in their problem, and they provide flexibility in choosing a predictor with an appealing combination of risk and fairness properties. These predictors converge to oracle predictors at fast (up to $\sqrt{n}$) rates. This approach substantially improves both the fairness and accuracy of an existing commercial recidivism predictor, and it yields many predictors that perform comparably to or better than other fairness methods on an income prediction task, while allowing users much more flexibility in the final model form.

Chapter 4 briefly considers the (un)fairness of randomized vs. deterministic classifiers.

Contents

List of Tables

List of Figures

Chapter 1

When the Oracle Misleads: Modeling the Consequences of Using Observable Rather than Potential Outcomes in Risk Assessment Instruments

1.1 Introduction

Machine learning is increasingly widely used to support decision making in domains as diverse as healthcare, criminal justice, and consumer finance. In particular, predictive models are often used to estimate the risk of a negative outcome such as death, recidivism, or default on a loan (Kourou et al., 2015; Caruana et al., 2015; Colubri et al., 2016; Brennan et al., 2009; Khandani et al., 2010). Scores from these Risk Assessment Instruments (RAIs) are made available to decision makers, such as doctors, judges, or loan officers, who may take them into account when deciding whether or not to admit a patient to a hospital, release a defendant on bail, or issue a loan to an applicant.

When the decision maker's goal is to reduce the risk of the predicted outcome, they are naturally concerned with *potential outcomes*, the outcomes that would occur under each available decision. When these outcomes correspond to an intervention that actually takes place, they are observable; otherwise, they are counterfactual. (Many authors use "counterfactual outcomes" as a synonym

for potential outcomes.) RAIs are typically trained on observational data, in which outcomes are affected by historical decisions, and they are typically designed to predict exclusively observable outcomes. Hence, these RAIs can only be sensibly understood as predicting the risk of an outcome *under the historical decision process that generated the data*; they are not generally appropriate for helping decision makers decide among different courses of action.

Although RAIs based on potential outcomes have been proposed in the context of medicine (Schulam and Saria, 2017; Shalit et al., 2017) and recidivism prediction (Mishler, 2019), RAIs designed to predict observable outcomes are in widespread use. While many of the limitations of such RAIs have been acknowledged (Chen and Asch, 2017; Veale et al., 2018), and problems associated with particular RAIs have been investigated (Povyakalo et al., 2013; Lum and Isaac, 2016), there does not appear to be a general mathematical model that provides insight into how and why such RAIs can lead users astray.

In this work, we aim to fill this gap, showing how RAIs based on observable outcomes can lead to *worse* outcomes, i.e., more severe departures from an optimal treatment regime, than before the RAI was introduced. This has nothing to do with the quality of prediction; it can occur even when (1) the oracle predictor is available and (2) there is no unmeasured confounding. We describe several dangerous properties of these RAIs and illustrate their suboptimality with a simple example.

1.2 Setup: RAIs and optimal treatment regimes

We anchor the problem in the context of medicine, but the results generalize to any domain where estimated risk is used to drive decision making designed to mitigate that risk.

Suppose that at time $t = 0$ we have random variables drawn from a counterfactual distribution $(U, X, A, Y^0, Y^1, Y) \sim \mathbb{Q}_0$, where $U \in \mathbb{R}^{p'}$ is a set of unobserved confounders, $X \in \mathbb{R}^p$ is a set of observed covariates, $A \in \{0, 1\}$ is a binary treatment or intervention decision, and $Y \in \{0, 1\}$ is an outcome, with $Y = 1$ indicating an adverse event such as patient death. Y^0 and Y^1 denote the potential outcomes under treatment decisions $A = 0, 1$. Let $\mathbb{P}_0$ denote the marginal distribution of the observable vector (X, A, Y) at $t = 0$. We use $\mathbb{E}_t$ and $\mathbb{P}_t$ to denote expectations and probabilities at time t, but when these do not change over time we drop the subscript and use $\mathbb{E}$ and $\mathbb{P}$.

Now suppose that iid data drawn from $\mathbb{P}_0$ is used to construct a predictor $s(X)$ of Y given X. For example, suppose that $s(X) = \hat{\mathbb{E}}_0[Y|X]$. This predictor is made available to decision makers in the form of an RAI, as a "risk score," giving rise at time $t = 1$ to new distributions $(U, X, A, Y^0, Y^1, Y) \sim \mathbb{Q}_1$ and $(X, A, Y) \sim \mathbb{P}_1$. We make the following assumptions at all time points t:

1. $Y = AY^1 + (1 - A)Y^0$ *(Consistency)*

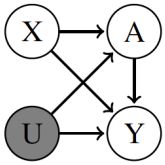


Figure 1.1: Causal graph at time $t = 0$, with unobserved confounders U.

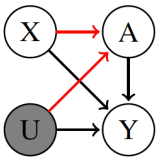


Figure 1.2: Causal graph at time $t = 1$, with possibly changed treatment decision process.

2. $\mathbb{P}_t[\mathbb{P}_t(0 < \pi_t(X) < 1)] = 1$ *(Positivity)*

3. $A \perp\!\!\!\perp Y^a | X, U$, for $a \in \{0, 1\}$ *(No confounders beyond X and U)*

4. $(U, X, Y^0, Y^1)_{\mathbb{Q}_t} \stackrel{d}{=} (U, X, Y^0, Y^1)_{\mathbb{Q}_{t+1}}$ *(Only the treatment and outcome change after the RAI is introduced.)*

5. $0 < \mathbb{P}_t(Y^1 < Y^0) < 1$ *(Treatment sometimes helps and sometimes hurts overall. For example, hospitalization can expose patients to dangers such as MRSA or medical errors.)*

Note that U is not observable by the researchers who construct $s(X)$, but it may include variables that are available to doctors at the time they render a treatment decision. That is, the treatment decision process may change in light of the new RAI, but the RAI does not otherwise affect patient outcomes or the distribution of covariates. Causal graphs representing the change in the decision process from time 0 to time 1 are given in Figures 1.1 and 1.2.

Given all possible treatment decision functions $\mathcal{D} = \{d : (X, U) \mapsto \{0, 1\}\}$, it is easy to show that the optimal treatment regime with respect to the expectation of Y is

$$d^{\text{opt}}(X, U) := \arg\min_{d \in \mathcal{D}} \mathbb{E}[Y^{d(X,U)}] = \mathbb{1}\{\mathbb{E}[Y^1 | X, U] < \mathbb{E}[Y^0 | X, U]\}. \tag{1.1}$$

where the expectations in this expression do not change over time, as a consequence of Assumption 4. Given that $s(X)$ is not designed as an estimator of d^{opt}, the questions of interest are:

1. When is $\mathbb{E}_1[Y] \leq \mathbb{E}_0[Y]$, as desired? That is, when does the RAI make things better, or at least not worse?

2. How far is $\mathbb{E}_1[Y]$ from $\mathbb{E}_1[Y^{d^{\text{opt}}}]$, the optimal outcome?

We are also interested in versions of these questions where the quantities are conditional on (X, U). For example, we wish to know when outcomes get better or worse differentially for patients from different demographic groups, which could cause the RAI to be considered unfair.

1.3 RAIs can make things worse

Let $\pi_t(X) = \mathbb{P}_t(A = 1|X, U)$ denote the treatment propensity at time t, with $\Gamma(X, U) := \pi_1(X, U) - \pi_0(X, U)$, and let $\mu^a(X, U) = \mathbb{E}[Y|X, U, A = a]$ denote the outcome regression functions, for $a \in \{0, 1\}$. μ^a does not change over time, per assumptions 3 and 4. We have:

$$\Delta := \mathbb{E}_1[Y] - \mathbb{E}_0[Y] = \mathbb{E}\left\{\Gamma(X, U)(\mu^1(X, U) - \mu^0(X, U))\right\} \tag{1.2}$$

(See the derivation in the Appendix). It is easy to see that Δ can be positive, meaning that more patients die after the introduction of the RAI, and that even if it is negative, outcomes could worsen for particular strata of (X, U). For example, consider a subpopulation for whom $\mu^1(X, U) < \mu^0(X, U)$ and $\Gamma(X, U) < 0$. These could be patients who historically benefited from hospitalization and were hospitalized at high rates, so that $\mathbb{E}_0[Y|X]$, their likelihood of death in the training data, is small. The apparent low risk could prompt doctors to reduce the rate at which they hospitalize these patients, causing death rates to rise. Of course, if Δ is positive, then $\mathbb{E}_1[Y] - \mathbb{E}[Y^{d^{\text{opt}}}]$ will be positive as well.

For ease of exposition, we now restrict our attention to a special case of the above scenario, wherein $U = \emptyset$, so that there is no unmeasured confounding, and $s(X) = \mathbb{E}[Y|X]$, so we have access to the MSE-minimal oracle predictor. We suppose that once the RAI is introduced, decisions are made deterministically according to a threshold rule $d(x) = \mathbb{1}\{s(X) \geq \theta\}$ for some $\theta \in [0, 1]$. That is, doctors hospitalize patients iff their estimated risk is at or above θ. We illustrate with a toy example.

1.3.1 Toy example

We assume a single covariate $X \sim \text{Unif}(0, 1)$, representing a marker of disease severity. We let both the treatment propensity and the risk of non-treatment increase in X, with $\pi_0(X) = \mathbb{E}[Y^0|X] = X$, and we let the risk of treatment be $\mathbb{E}[Y^1|X] = (0.7 - X)^2$. This represents a situation in which treatment is beneficial on average above a certain level of X but harmful otherwise.

Figure 1.3 (a) shows the two conditional expectations $\mathbb{E}[Y^0|X], \mathbb{E}[Y^1|X]$. The optimal treatment rule here is $d^{\text{opt}}(X) = \mathbb{1}(X \geq 0.22)$, indicated by the dashed line. The rule that is actually implemented at time $t = 1$ is $d(X) = \mathbb{1}\{\mathbb{E}_0[Y|X] \geq \theta\}$ for the chosen threshold θ. Figure 1.3 (b) shows the mean difference in outcomes Δ from time 0 to time 1 as a function of θ. (See the Appendix for a derivation.) This difference is around 1/3, regardless of the θ chosen, indicating that more patients die as a result of the RAI. (In this scenario, $s(X)$ is bounded in [0, 0.30], so we only show thresholds in this range.)

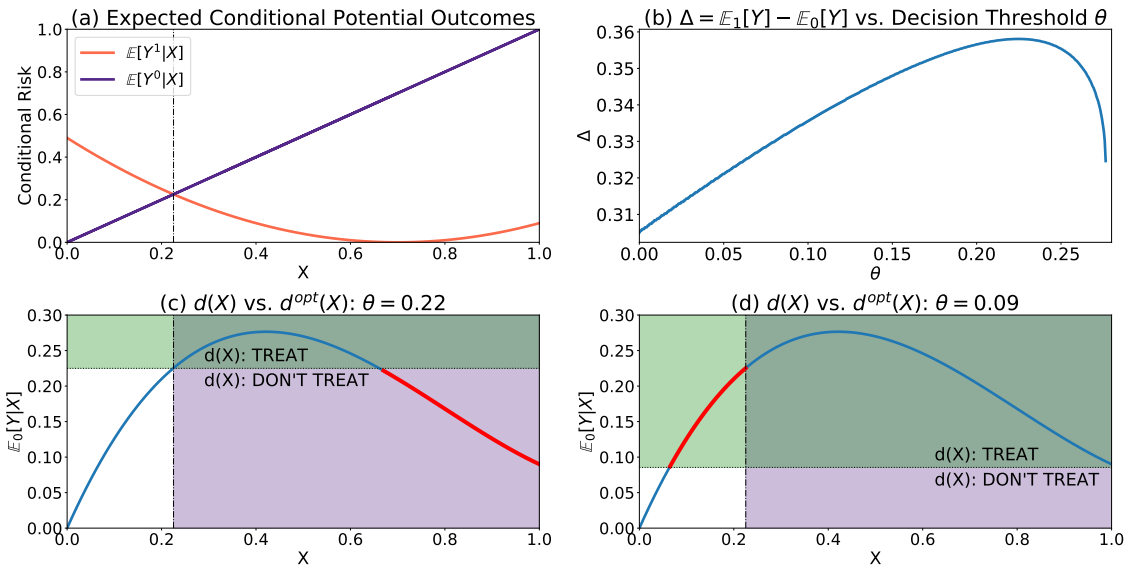


Figure 1.3: (a) Conditional expectations in the toy example. (b) Behavior of $\Delta = \mathbb{E}_1[Y] - \mathbb{E}_0[Y]$ as a function of the cutoff θ. (c) and (d) show groups treated at time 1 under $d(X)$ and $d^{\text{opt}}(X)$, for two possible values of θ. The optimal treatment group is $\{X > 0.22\}$, in purple. The group treated under $d(X)$ is $\{\mathbb{E}_0[Y|X] \geq \theta\}$, in green. Red lines indicate groups that are harmed under $d(X)$ as a result of not receiving or receiving treatment, respectively.

The reason that all values of θ lead to worse outcomes is that θ corresponds to a threshold for $s(X) = \mathbb{E}_0[Y|X]$ rather than a threshold for X. In Figure 1.3 (c) and (d), the vertical purple block represents the optimal treatment group $\{X \geq 0.22\}$, while the overlapping horizontal green block represents the group $\{\mathbb{E}_0[Y|X] \geq \theta\}$ that is actually treated under $d(X)$. Panel (c) shows the effect of choosing $\theta = 0.22$, the optimal threshold for X: we would fail to provide treatment to the group $\{X \geq 0.67\}$, indicated in red. This happens to be the group with the highest values of $\mathbb{E}[Y^0|X]$, i.e., the worst outcomes under no treatment. Conversely, Figure (d) shows the results of selecting the cutoff such that all those who would receive treatment under $d^{\text{opt}}(X)$ also receive treatment under $d(X)$: we wrongly treat the group $\mathbb{E}[Y^1|X] > \mathbb{E}[Y^0|X]$, again indicated in red.

These same problems can obviously arise in more complex scenarios, for example when $U \neq \emptyset$, when X is high dimensional, and when the relationship between X and the outcome is complex. In particular, we identify three properties of $s(X)$ that can give rise to these and other problems.

1.3.2 $s(X)$ doesn't map nicely to a quantity of interest like $\mathbb{E}[Y^0|X]$, $\mathbb{E}[Y^1|X]$, or $d^{\text{opt}}(X)$

Even though it is designed to predict outcomes under a historical treatment decision process, the RAI could usefully inform a new decision process if it bore some readily apprehensible relationship with a

potential outcome-based quantity of interest. For example, if $s(X)$ were monotonic in $d^{\text{opt}}(X)$, then doctors might be able to use $s(X)$ to get closer to $d^{\text{opt}}(X)$, even without an explicit awareness of this relationship. In general, however, the relationship between $s(X)$ and any potential outcome-based quantity can be arbitrarily complex.

1.3.3 Expertise can make things worse

The more skilled doctors are at time $t = 0$, the worse the system can get at time $t = 1$. As an extreme example, if doctors are already behaving according to the optimal policy at time $t = 0$, then necessarily, $\mathbb{E}_1[Y] \geq \mathbb{E}_0[Y]$. Alternatively, suppose that there are two medical systems $\mathbb{P}_0$ and $\mathbb{P}_0^*$ that are identical in their distribution of (X, U, Y^0, Y^1). Imagine that they're also identical in terms of A, except that in system $\mathbb{P}_0^*$ doctors are more skilled at identifying who needs to be hospitalized:

$$\mathbb{P}_0^*\big(A = 1 | d^{\text{opt}}(X, U) = 1\big) > \mathbb{P}_0\big(A = 1 | d^{\text{opt}}(X, U) = 1\big)$$

Then, under a threshold decision rule, we have that $\mathbb{E}_0^*[Y] < \mathbb{E}_0[Y]$ but $\mathbb{E}_1^*[Y] > \mathbb{E}_1[Y]$, so, perversely, people in system $\mathbb{P}_0^*$ are **better off** than people in system $\mathbb{P}$ at time 0 and **worse off** at time 1.

1.3.4 The procedure is unstable under iteration

Imagine that we iterate the process of gathering data from the system, developing a predictor, and implementing the threshold-based decision rule above. This seems like a plausible occurrence, in that as RAIs get integrated into more and more systems, necessarily any future data gathered from those systems will reflect the influence of those tools.

For time points $t = 1, 2, \ldots$, we have

$$\mathbb{E}_t[Y|X] = \mathbb{1}\{\mathbb{E}_{t-1}[Y|X] > \theta\}\mathbb{E}[Y^1|X] + (1 - \mathbb{1}\{\mathbb{E}_{t-1}[Y|X] > \theta\})\mathbb{E}[Y^0|X]$$

Suppose we have some X for which $\mathbb{E}_0[Y^1|X] < \theta, \mathbb{E}_0[Y^0|X] > t$ and $\mathbb{E}[Y_{(0)}|X] > \theta$. Then we'll have the situation depicted in Table 1.1, in which the treatment decision for this stratum just alternates at different time points. Ideally, as more and more data is gathered from a system, a decision procedure gets closer and closer to optimal. In this scenario, however, the treatment decision is the optimal one only at odd time points, while at even time points it's precisely the opposite.

Time t	Treatment decision	$\mathbb{E}[Y_t\|X]$	$\mathbb{E}_t[Y\|X]$ relative to θ
0	Treat with probability $\pi_0(X)$	$\mathbb{E}_0[Y\|X]$	$> \theta$
1	Treat all	$\mathbb{E}[Y^1\|X]$	$< \theta$
2	Treat none	$\mathbb{E}[Y^0\|X]$	$> \theta$
3	Treat all	$\mathbb{E}[Y^1\|X]$	$< \theta$
4	Treat none	$\mathbb{E}[Y^0\|X]$	$> \theta$
...			

Table 1.1: Treatment decisions and mean outcomes in stratum X over time, under the deterministic decision rule that treats patients at time t iff $\mathbb{E}_{t-1}[Y|X] > \theta$ for some threshold θ, and assuming that $\mathbb{E}_0[Y^1|X] < \theta, \mathbb{E}_0[Y^0|X] > t$ and $\mathbb{E}[Y_{(0)}|X] > \theta$.

1.4 Conclusion

Decision makers choosing among different courses of action are naturally interested in the risk associated with each option. RAIs are in widespread use in many domains, but they are typically designed to predict outcomes under the historical decision process that generated the training data, rather than predicting potential outcomes under the available courses of action. This makes them generally unsuitable for informing future treatment or intervention decisions that are designed to reduce risk. Although previous work has proposed using potential outcome-based predictors in certain contexts, there has been little formal modeling of the consequences of current practice. Here, we show how RAIs based on observable rather than potential outcomes can plausibly lead to worse outcomes overall or for specific demographic groups than before their introduction, making them potentially both dangerous and unfair.

1.A Derivations

1.A.1 Equation (1.2): difference in mean outcome from time 0 to time 1

Recall that we define

$$\pi_t(X, U) = \mathbb{E}_t[A|X, U]$$
$$\mu^a(X, U) = \mathbb{E}[Y^a|X, U]$$

for $t = 1, 2, \ldots$ and $a \in \{0, 1\}$. Recall also that, per assumption 4, the distribution of the covariates (X, U) doesn't change over time, so functions of (X, U) don't change either.

For any time point t, we have

$$\begin{aligned}
\mathbb{E}_t[Y] &= \mathbb{E}_t\left\{\mathbb{E}_t[AY^1 + (1-A)Y^0|X,U]\right\} \\
&= \mathbb{E}_t\left\{\mathbb{E}_t[A|X,U]\mathbb{E}_t[Y^1|X,U] + (1-\mathbb{E}_t[A|X,U])\mathbb{E}_t[Y^0|X,U]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}_t[A|X,U]\mathbb{E}[Y^1|X,U] + (1-\mathbb{E}_t[A|X,U])\mathbb{E}[Y^0|X,U]\right\} \\
&= \mathbb{E}\left\{\pi_t(X,U)\mu^1(X,U) + (1-\pi_t(X,U))\mu^0(X,U)\right\}
\end{aligned}$$

where the second equality follows because $A \perp\!\!\!\perp Y^a|X,U$ and the third equality follows from assumption 4. With $\Gamma(X,U) := \pi_1(X,U) - \pi_0(X,U)$, we have:

$$\begin{aligned}
\mathbb{E}_1[Y] - \mathbb{E}_0[Y] &= \mathbb{E}\left\{\pi_1(X,U) - \pi_0(X,U))(\mu^1(X,U) - \mu^0(X,U))\right\} \\
&= \mathbb{E}\left\{\Gamma(X,U)(\mu^1(X,U) - \mu^0(X,U))\right\}
\end{aligned}$$

1.A.2 Calculating Δ in the toy example

We have $U = \emptyset, X \sim \text{Unif}(0,1)$, $\pi_0(X) = \mathbb{E}[Y^0|X] = X$, $\mathbb{E}[Y^1|X] = (0.7-X)^2$, and $\pi_1(X) = \mathbb{1}\{\mathbb{E}_0[Y|X] \geq \theta\}$ for some chosen θ. Plugging these into the above yields

$$\begin{aligned}
\Delta(\theta) = \mathbb{E}_1[Y] - \mathbb{E}_0[Y] &= \int \big(\mathbb{1}\{\mathbb{E}_0[Y|X] \geq \theta\} - X\big)((0.7-X)^2 - X)d\mathbb{P}(X) \\
&= \int \big(\mathbb{1}\{\mathbb{E}_0[AY^1 + (1-A)Y^0|X] \geq \theta\} - X\big)\big((0.7-X)^2 - X\big)d\mathbb{P}(X) \\
&= \int \big(\mathbb{1}\{\pi_0(X)\mu^1(X) + (1-\pi_0(X))\mu^0(X) \geq \theta\} - X\big)\big((0.7-X)^2 - X\big)d\mathbb{P}(X) \\
&= \int \big(\mathbb{1}\{X(0.7-X)^2 + (1-X)X \geq \theta\} - X\big)\big((0.7-X)^2 - X\big)d\mathbb{P}(X)
\end{aligned}$$

where the second equality follows from the consistency assumption, and the third equality follows from the no unmeasured confounding assumption and assumption 4. This yields the curve in Figure 1.3(b).

Chapter 2

Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds

2.1 Introduction

Machine learning is increasingly involved in high stakes decisions in domains such as healthcare, criminal justice, and consumer finance. In these settings, ML models often take the form of Risk Assessment Instruments (RAIs): given covariates such as demographic information and an individual's medical/criminal/financial history, the model predicts the likelihood of an adverse outcome, such as a dangerous medical event, recidivism, or default on a loan. Rather than rendering an automatic decision, the model produces a "risk score," which a decision maker may take into account when deciding whether to prescribe a medical treatment, release a defendant on bail, or issue a personal loan.

The proliferation of machine learning has raised concerns that learned models may be discriminatory with respect to sensitive features like race, sex, age, and socioeconomic status. For example, there has been vigorous debate about whether a widely used recidivism prediction tool called COMPAS is biased against black defendants (Angwin et al., 2016; Angwin and Larson, 2016; Dieterich et al., 2016; Larson and Angwin, 2016; Lowenkamp et al., 2016). Concerns have also been

raised about risk assessments used to identify high risk medical patients (Obermeyer et al., 2019) and about common credit scoring algorithms such as FICO (Rice and Swesnik, 2012), among many others. Collectively, these types of algorithms directly impact a large and growing swath of the global population.

These concerns have led to an explosion of methods in recent years for developing fair models and auditing the fairness of existing models. The most widely discussed fairness criteria impose constraints on the joint distribution of a sensitive feature, an outcome, and a predictor. These "observational" fairness criteria are inappropriate for RAIs, however. RAIs are not concerned with the *observable outcomes* in the training data ("Did patients of this type historically experience serious complications?"), which are themselves a product of historical treatment decisions. Rather, they are concerned with the *potential outcomes* associated with available treatment decisions ("*Would* patients of this type experience complications *if not treated?*"). Because treatments are not assigned at random—doctors naturally treat the patients they think are at high risk—these are distinct questions.

Coston et al. (2020) showed how RAIs that are optimized to predict observable rather than potential outcomes systematically underestimate risk for units that have historically been receptive to treatment, leading to suboptimal treatment decisions. They further showed how evaluations of the performance and fairness properties of RAIs with respect to observable outcomes are misleading. They proposed that RAIs should instead target counterfactual versions of standard performance and fairness metrics. However, they left open the question of how to develop predictors that satisfy such fairness notions.

In this paper, we develop a method to generate predictors that satisfy the fairness criterion *approximate counterfactual equalized odds*. While many existing methods target observational fairness criteria (Kamiran and Calders, 2012; Hardt et al., 2016; Calmon et al., 2017; Zafar et al., 2017; Donini et al., 2018; Narasimhan, 2018; Kim et al., 2019) and various types of causally motivated fairness (Kilbertus et al., 2017; Kusner et al., 2017; Nabi and Shpitser, 2018; Nabi et al., 2019), no methods currently exist that target counterfactual versions of standard observable fairness criteria like equalized odds. Our method post-processes an arbitrary existing predictor, extending previous post-processing methods (Hardt et al., 2016) to the counterfactual setting.

Our contributions are as follows. We first define approximate counterfactual equalized odds (§2.2). After discussing related work (§2.3) and motivating the use of equalized odds over other candidate criteria (§2.4), we present a linear program that produces a loss-optimal post-processed predictor that satisfies this criterion (§2.5). We provide theoretical results that our post-processed predictor is consistent in a particular sense at rates that depend on certain nuisance parameters. We show that our method performs well on both simulated and real data (§2.6).

2.2 Notation and fairness definitions

A table listing all notational choices can be found in Appendix 2.F.

Let A, D, Y denote a sensitive feature, decision, and outcome, respectively. We consider the setting in which all three are binary, though most of the definitions below extend readily to continuous settings. We define the counterfactual quantities of interest via the potential outcomes framework of (Neyman, 1923; Holland, 1986; Rubin, 2005). Denote by Y^0, Y^1 the potential (equivalently, "counterfactual") outcomes $Y^{D=0}, Y^{D=1}$. Y_i^d is the outcome that would be observed for unit i if, possibly contrary to fact, the decision were set to $D_i = d$. We refer to the two levels of the sensitive feature A as the two "groups," and we use "treatment" and "intervention" synonymously with "decision." Let S be any random variable that takes values in $\{0, 1\}$.

In most RAI settings, one of the decision options is a natural baseline corresponding to "no intervention" ($D = 0$). Examples include the risk of recidivism if a defendant is released pretrial, or the risk of neglect or abuse if a child welfare call is not screened in for further investigation. Many or most RAIs do not generate a separate risk score for the outcome associated with intervention. In the case of child welfare, for example, call screeners must screen in any case in which a child is in apparent danger of neglect or abuse, regardless of the chances that a subsequent intervention will successfully prevent that neglect or abuse.

Denote the observational and counterfactual false positive rates of S for group a by $\text{FPR}(S, a) = \mathbb{P}(S = 1 \mid Y = 0, A = a)$ and $\text{cFPR}(S, a) = \mathbb{P}(S = 1 \mid Y^0 = 0, A = a)$. For example, $\text{cFPR}(S, 0)$ could represent the chance of being falsely labeled high-risk, among those black defendants who would not actually go on to recidivate if released pretrial, while $\text{cFPR}(S, 1)$ could represent the corresponding error rate for white defendants who would not recidivate if released pretrial. Let FNR, cFNR, TPR, and cTPR denote the corresponding observational and counterfactual false negative and true positive rates.

Definition 2.2.1. A predictor S satisfies *observational equalized odds* (oEO) with respect to A and Y if $S \perp\!\!\!\perp A \mid Y$. It satisfies *counterfactual equalized odds* (cEO) if $S \perp\!\!\!\perp A \mid Y^0$.

When A, Y, and S are all binary, equalized odds is equivalent to requiring that the corresponding false positive and false negative rates be equal for the two levels of A. Our post-processed predictor will target a relaxation of this criterion, defined below.

Definition 2.2.2. The *counterfactual error rate differences* for a predictor S are the differences Δ^+ and Δ^- in the cFPR and cFNR for the two groups $A = 0, A = 1$, defined as follows:

$$\Delta^+(S) = \text{cFPR}(S, 0) - \text{cFPR}(S, 1)$$
$$\Delta^-(S) = \text{cFNR}(S, 0) - \text{cFNR}(S, 1)$$

Definition 2.2.3. When A, Y, and S are all binary, S satisfies *approximate counterfactual equalized odds* with *fairness constraints* $\epsilon^+, \epsilon^- \in [0, 1]$ if

$$|\Delta^+(S)| \leq \epsilon^+$$
$$|\Delta^-(S)| \leq \epsilon^-$$

In general, a fairness-constrained predictor would not outperform an optimal unconstrained predictor, and in some cases, satisfying cEO exactly might degrade performance to the point that the RAI is no longer useful. This relaxation of cEO allows RAI designers to negotiate this tradeoff. This is similar in spirit to notions of approximate fairness that appear throughout the literature (Kearns et al., 2017; Donini et al., 2018; Menon and Williamson, 2018).

2.3 Related work

2.3.1 Observational and counterfactual fairness

Equalized odds is one of several popular fairness criteria that impose constraints on the joint distribution of (A, Y, S) (Barocas et al., 2018). These criteria appear under a variety of names. Equalized odds is known more generally as *separation*, a term which covers settings in which these variables are not necessarily binary. The other two popular criteria in this class are *independence* ($S \perp\!\!\!\perp A$) and *sufficiency* ($Y \perp\!\!\!\perp A \mid S$). Independence also manifests as *demographic parity*, *statistical parity*, and *group fairness*. Sufficiency is equivalent to *calibration* or *predictive parity* when all three variables are binary. Variants of all three criteria may be defined for example by conditioning on additional variables.

The counterfactual versions of these criteria simply replace Y with the potential outcome Y^0 that is of interest (Coston et al., 2020). Note that these definitions cannot accommodate more than one potential outcome, such as the vector (Y^0, Y^1), because only one of these outcomes is observed for each unit. This is the "fundamental problem of causal inference" (Holland, 1986).

Except in highly constrained, unrealistic conditions, these three criteria are pairwise unsatisfiable, regardless of whether they are defined with respect to Y or Y^0 (Kleinberg et al., 2017; Chouldechova, 2017; Barocas et al., 2018)*. We must therefore choose and justify which criterion we wish to target.

2.3.2 Other causal fairness criteria

The counterfactual fairness criteria just described consider potential outcomes with respect to a decision D. There is a distinct set of causally motivated fairness criteria that consider counterfactuals of the sensitive feature, or a proxy for the sensitive feature. They characterize a decision or prediction as fair if the sensitive feature or proxy does not "cause" the decision or prediction, either directly or along a prohibited pathway (Kilbertus et al., 2017; Kusner et al., 2017; Nabi and Shpitser, 2018; Zhang and Bareinboim, 2018; Nabi et al., 2019; Wang et al., 2019). There is some controversy over whether it is meaningful to discuss a counterfactual of a feature like race or gender (VanderWeele and Robinson, 2014; Glymour and Glymour, 2014; Hu and Kohler-Hausmann, 2020). Additionally, satisfying these metrics typically precludes use of most of the features that go into risk assessment, like prior history, which is not tenable in practice (Coston et al., 2020). Finally, it is not clear that counterfactuals of the sensitive feature are useful or appropriate to consider in the context of risk assessment. For example, in the child welfare setting, workers are compelled to screen in calls whenever a child is in danger of neglect or abuse. While it is important to ensure that risk is assessed accurately for different groups, it would be inappropriate to make screen-in decisions based on what a child's risk of neglect or abuse would be *if they had been of a different race* their whole life, even if such an assessment were possible.

2.3.3 Ways of achieving fairness

There are three broad approaches to developing fair models: (1) preprocessing the input data to remove bias (Kamiran and Calders, 2012; Calmon et al., 2017), (2) constraining the learning process (aka "in-processing") (Zafar et al., 2017; Donini et al., 2018; Narasimhan, 2018), and (3) post-processing a model to satisfy fairness constraints (Hardt et al., 2016; Kim et al., 2019).

Our approach belongs to class (3). We refer to the predictor that our method returns equivalently as a "post-processed" or "derived" predictor. Each approach has advantages and disadvantages. Many widely used RAIs are proprietary tools developed by for-profit companies, so they are not amenable to internal tinkering. Developing new, fair(er) RAIs would be costly and perhaps infeasible from a policy perspective. The advantage of post-processing in this setting is that it can be applied to models that are already in use. The predictor that our method returns requires access at runtime

*See (Imai and Jiang, 2020) for a set of sufficient conditions under which these unsatisfiability results disappear.

only to the sensitive feature and the output of the existing predictor, so in principle, it could easily be incorporated into existing risk assessment pipelines.

In particular, our approach extends the work of Hardt et al. (2016), who proposed a method to post-process binary predictors to satisfy observational equalized odds (oEO) while minimizing loss with respect to observable Y. Their post-processed predictor is the solution to a simple linear program. We adapt their method to the counterfactual setting, in which the fairness criterion is approximate cEO and the loss function is weighted classification error with respect to Y^0. Because Y^0 is not observable when $D \neq 0$, we require tools from causal inference to solve this problem. Hardt et al.'s analysis treats the joint distribution of (A, S, Y) as known and frames post-processing primarily as an optimization problem. We build on their results by not making this assumption and treating post-processing as a statistical estimation problem.

2.3.4 Why equalized odds?

When evaluating a predictive system, it seems natural to focus on its real-world impact rather than its outputs per se. One desirable property of a decision process is the avoidance of disparate impact. Disparate impact is a legal doctrine enshrined in U.S. law that prohibits practices which have an unjustifiable adverse impact on people who share a protected characteristic, regardless of discriminatory intent. By way of shorthand, we will say that if $D \not\perp\!\!\!\perp A \mid Y^0$, then the system exhibits discriminatory disparate impact[†]. In recidivism prediction, for example, this could mean that black defendants ($A = 0$) who would not recidivate if released ($Y^0 = 0$) are more likely to be detained pretrial ($D = 1$) than white defendants who would not recidivate if released ($A = 1, Y^0 = 0$).

In the context of RAIs, decision makers typically have wide latitude in how they interpret and act on the risk scores, so constraining the RAI does not enforce fairness with respect to their decisions. However, if decision makers, after the introduction of the RAI, make their decisions only on the basis of the RAI scores and other variables U which are independent of the RAI and A given Y^0, then counterfactual equalized odds will imply $D \perp\!\!\!\perp A \mid Y^0$. That is, let $D = f(S, U)$ represent the function f describing the decision process after the RAI S is introduced. If cEO is satisfied and $U \perp\!\!\!\perp (S, A) \mid Y^0$, then it follows that $D \perp\!\!\!\perp A \mid Y^0$. Even if $U \not\perp\!\!\!\perp (S, A) \mid Y^0$, it is easy to see that if the conditional independence statement *nearly* holds, or if f depends primarily on S rather than U, then discriminatory disparate impact can be small.

No such guarantees hold for predictors satisfying either independence or sufficiency. Chouldechova (2017) in particular showed how predictors which satisfy sufficiency (predictive parity) are likely to yield decisions such that $D \not\perp\!\!\!\perp A \mid Y$; these arguments are unchanged when we substitute Y^0

[†]Some authors use "disparate impact" to refer to the criterion $S \perp\!\!\!\perp A$, i.e. *independence* (Zafar et al., 2017).

for Y. Though there is no consensus about how to quantify fairness, this is at least one consideration in favor of equalized odds.

2.4 Motivating Example

Having motivated equalized odds over predictive parity or independence, we now motivate the use of counterfactual rather than observational equalized odds.

Consider a school district that assigns tutors to students who are believed to be at risk of academic failure. The school district wishes to develop a RAI, S, to better identify students who need tutors while ensuring that this resource is allocated fairly across two levels of the sensitive feature A. Let $D \in \{0, 1\}$ represent the decision to assign (1) or not assign (0) a tutor, and let $Y \in \{0, 1\}$ represent academic success (0) or failure (1).

A cEO predictor S satisfies $\mathbb{P}(S \mid Y^0, A) = \mathbb{P}(S \mid Y^0)$, while an oEO predictor S satisfies $\mathbb{P}(S \mid Y, A) = \mathbb{P}(S \mid Y)$. Divergence in these predictors is driven by the extent to which $Y \neq Y^0$ in the training data. In order to parameterize this divergence, we introduce the following definitions.

Definition 2.4.1. The *need rate* for group a is $\mathbb{P}(Y^0 = 1 \mid A = a)$, the probability that a student from group a would fail without a tutor.

Definition 2.4.2. The *opportunity rate* for group a is $\mathbb{P}(D = 1 \mid Y^0 = 1, A = a)$, the probability that a student in group a who needs a tutor receives one.

Definition 2.4.3. The *intervention strength* for group a is $\mathbb{P}(Y^1 = 0 \mid Y^0 = 1, A = a)$, the probability that a student in group a who would fail without a tutor would succeed with a tutor.

We simulate a simple data generating process in which we allow the intervention strength to vary, while constraining it to be equal for the two groups. We fix all other parts of the distribution. In particular, we set $\mathbb{P}(A = 1) = 0.7$, set the need rates to 0.4 and 0.2 for groups 0 and 1, and set the opportunity rates to 0.6 and 0.4. We set the probabilities that a tutor is assigned when it is not needed to $\mathbb{P}(D = 1 \mid Y^0 = 0, A = 0) = 0.3$ and $\mathbb{P}(D = 1 \mid Y^0 = 0, A = 1) = 0.2$. This represents a scenario in which the minority group has greater need, perhaps due to socioeconomic factors or prior educational opportunities, and also is likelier than the majority group to receive resources at baseline (prior to the development of the RAI). Finally, we set $\mathbb{P}(Y^1 = 0 \mid Y^0 = 0) = 1$, meaning that tutoring never *increases* the risk of failure.

We consider a hypothetical oEO predictor S with fixed false positive rate $\mathbb{P}(S = 1 \mid Y = 0, A) = \mathbb{P}(S = 1 \mid Y = 0) = 0.1$ and false negative rate $\mathbb{P}(S = 0 \mid Y = 1, A) = \mathbb{P}(S = 0 \mid Y) = 0.2$. We assume $S \perp\!\!\!\perp Y^0 \mid A, Y$, as would be the case for example when S is a high quality predictor of

Y. Figure 2.1 shows the cTPRs (counterfactual true positive rates) for this predictor as a function of intervention strength, relative to the baseline opportunity rates for the two groups. When the intervention has no effect (strength 0), the cTPRs are equal because $Y \equiv Y^0$, so the cTPR and TPR are identical. (Of course, a strength of 0 means the tutoring is worthless.) For all strength values > 0, the cTPR of the minority group is lower than for the majority group. The difference in error rates increases as intervention strength increases. A cEO predictor avoids this problem by design: the cTPRs for the two groups are constrained to be equal.

This example makes it clear that oEO predictors in general will not prevent discriminatory disparate impact, whereas, as discussed in section 2.3.4, counterfactual EO predictors have at least the potential to mitigate or avoid it.

This example also illustrates how oEO predictors can reduce rates of appropriate intervention. For example, suppose that decision makers, after the introduction of the RAI, set $D \equiv S$, i.e. they assign tutors precisely to students whom the RAI labels as high risk. Then, for any intervention strength > 0.5, the opportunity rate for the minority group decreases below baseline: the RAI *harms the minority group.*

2.5 An optimal fair derived predictor

Having motivated counterfactual equalized odds, we now develop a method to generate predictors which satisfy it.

2.5.1 Estimand

We expand our notation in order to fully describe our problem setting. Consider a random vector $Z = (A, X, D, S, Y) \sim \mathbb{P}$, where in addition to the binary sensitive feature A, decision D, and outcome Y, we have covariates $X \in \mathbb{R}^p$ and a previously trained binary predictor $S \in \{0, 1\}$. We require only that S is observable; we do not require access to its inputs or internal structure. S in practice could represent a RAI that is already in use, such as a recidivism prediction tool. The covariates X may or may not overlap with the inputs to S. Their role in the analysis is to render counterfactual quantities identifiable.

Our target is a derived predictor that satisfies approximate cEO. As in the case of observable equalized odds considered by Hardt et al. (2016), we achieve this by randomly flipping S with probabilities that depend only on S and A. Consider a column vector $\theta = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1}) \in$

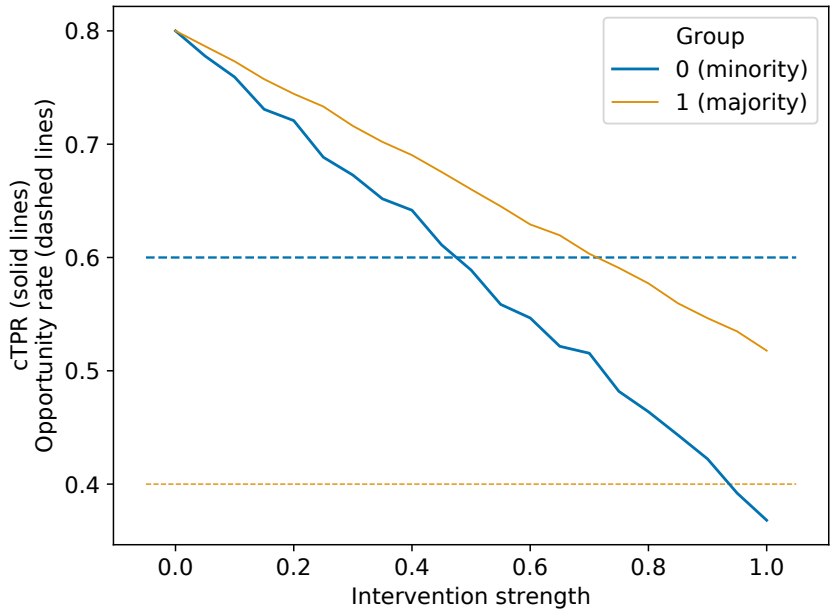


Figure 2.1: Counterfactual true positive rates (cTPRs; solid lines) for a RAI satisfying observational equalized odds (oEO), as a function of the intervention strength $\mathbb{P}(Y^1 = 0 \mid Y^0 = 1)$. Dashed lines indicate opportunity rates $\mathbb{P}(D = 1 \mid Y^0 = 0)$ prior to the development of the RAI. The more effective the tutoring (the higher the intervention strength), the worse the RAI is at identifying students who need it, and the greater the disparity in its performance between the minority and the majority group. When tutoring is more effective, the RAI may reduce the appropriate assignment of tutors below the baseline opportunity rates.

$[0,1]^4$. We define an associated derived predictor S_θ:

$$\begin{aligned} &S_\theta \sim \text{Bern}(\theta_{A,S}) \\ &\text{where } \theta_{A,S} = \sum_{a,s\in\{0,1\}} \mathbb{1}\{A=a, S=s\}\theta_{a,s} \end{aligned}$$

In other words, the $\theta_{a,0}$ parameters represent conditional probabilities that S flips, while the $\theta_{a,1}$ parameters represent conditional probabilities that S doesn't flip. Notice that for $\tilde{\theta} = (0,1,0,1)$, we have $S_{\tilde{\theta}} = S$: the derived predictor is equal to the input predictor.

Our target is a loss-optimal fair predictor S_{θ^*}, where the fairness criterion is approximate cEO. The loss function we consider is weighted classification error. For fixed θ, denote the loss[‡] by $\text{Loss}(S_\theta; w^+, w^-) = w^+\mathbb{P}(Y^0 = 0, S_\theta = 1) + w^-\mathbb{P}(Y^0 = 1, S_\theta = 0)$, where w^+, w^- are chosen by the user to capture the relative importance of false positives and false negatives. (We will generally suppress the dependence of Loss on w^+, w^-.) The estimand is

$$\begin{aligned} \theta^* \in \arg\min_\theta \ &\text{Loss}(S_\theta) \\ \text{subject to } \ &\theta \in [0,1]^4 \\ &|\Delta^+(S_\theta)| \le \epsilon^+ \\ &|\Delta^-(S_\theta)| \le \epsilon^- \end{aligned}$$

where Δ^+, Δ^- are given above in Definition 2.2.2, and the fairness constraints $\epsilon^+, \epsilon^- \in [0,1]$ are chosen by the user. Setting both these constraint parameters to 0 requires cEO to be satisfied exactly, while setting them to 1 allows S_{θ^*} to be arbitrarily unfair. Setting ϵ^- to 0 regardless of ϵ^+ forces S_{θ^*} to satisfy *counterfactual equal opportunity*; see Hardt et al. (2016) for the observational definition of this criterion.

Remark 1. The full vector Z is required only to estimate the parameter θ^* that defines the optimal fair derived predictor. Once θ^* has been estimated, the resulting derived predictor requires access at runtime only to the sensitive feature A and the input predictor S.

Since our estimands involve counterfactual quantities, distributional assumptions are required in order to equate them to observable quantities.

[‡]We refer to this quantity as "loss" instead of the conventional "risk" in order to avoid confusion between risk assessment and the error rate of a predictor.

2.5.2 Identification

In this subsection we show that the counterfactual error rates and loss can be identified under standard causal inference assumptions. All the quantities to be identified can be written in terms of the loss and the counterfactual error rates of the input predictor S. For ease of notation, we first define two nuisance parameters that appear in the estimand and associated estimators, namely the outcome regression and propensity score function. We generally suppress the arguments of these functions in subsequent usage for the sake of conciseness.

$$\mu_0(A, X, S) = \mathbb{E}[Y \mid A, X, S, D = 0]$$
$$\pi(A, X, S) = \mathbb{P}(D = 1 \mid A, X, S)$$

We make the following standard "no unmeasured confounding"-type causal inference assumptions:

A1. (Consistency)	$Y = DY^1 + (1 - D)Y^0$
A2. (Positivity)	$\exists \delta \in (0, 1)$ s.t. $\mathbb{P}(\pi(A, X, S) \leq 1 - \delta) = 1$
A3. (Ignorability)	$Y^0 \perp\!\!\!\perp D \mid A, X, S$

The consistency assumption means that the outcome observed for each individual is precisely the potential outcome corresponding to the treatment received. This implies that one person's treatment assignment does not affect another person's outcomes, meaning, for example, that an individual's recidivism behavior does not depend on whether other individuals are detained or released. The positivity or *overlap* assumption requires that within strata of (A, X, S) of measure > 0, individuals have some chance of receiving no intervention. Finally, the ignorability or *no unmeasured confounding* assumption requires that within strata of (A, X, S), the treatment D is essentially random with respect to Y^0. Satisfying ignorability assumptions typically requires collecting a rich enough set of deconfounding covariates. In the present case, even if X is low dimensional, the ignorability assumption is plausible if the input predictor S substantially drives decision making, or if it happens to be an accurate (if not necessarily fair) predictor of Y^0.

Before giving the identifying expressions for $\text{Loss}(S_\theta)$ and the error rate differences Δ^+, Δ^-, we give identifying expressions for the error rates of the input predictor S, which themselves appear in the expressions for Δ^+, Δ^-.

Proposition 1. Under assumptions A1-A3, the counterfactual error rates of the input predictor S are identified as follows:

$$\text{cFPR}(S,a) = \frac{\mathbb{E}[S(1-\mu_0)\mathbb{1}\{A=a\}]}{\mathbb{E}[(1-\mu_0)\mathbb{1}\{A=a\}]}$$
$$\text{cFNR}(S,a) = \frac{\mathbb{E}[(1-S)\mu_0\mathbb{1}\{A=a\}]}{\mathbb{E}[\mu_0\mathbb{1}\{A=a\}]}$$

Proofs of propositions are given in Appendix 2.A. We now define several quantities that appear in the identifying expressions for $\text{Loss}(S_\theta), \Delta^+$, and Δ^-:

$$\beta_{a,s} = \mathbb{E}\left[\mathbb{1}\{A=a, S=s\}\left(w^+ - (w^+ + w^-)\mu_0\right)\right], \text{ for } a,s \in \{0,1\}$$
$$\beta = (\beta_{0,0}, \beta_{0,1}, \beta_{1,0}, \beta_{1,1})$$
$$\beta^+ = (1 - \text{cFPR}(S,0),\ \text{cFPR}(S,0),\ \text{cFPR}(S,1) - 1,\ -\text{cFPR}(S,1))$$
$$\beta^- = (-\text{cFNR}(S,0),\ \text{cFNR}(S,0) - 1,\ \text{cFNR}(S,1),\ 1 - \text{cFNR}(S,1))$$

Proposition 2. Under assumptions A1-A3, the loss and error rates of the derived predictor S_θ are identified as:

$$\text{Loss}(S_\theta) = \theta^T\beta + w^-\mathbb{E}[\mu_0]$$
$$\Delta^+(S_\theta) = \theta^T\beta^+$$
$$\Delta^-(S_\theta) = \theta^T\beta^-$$

Since the term $w^-\mathbb{E}[\mu_0]$ in the loss is fixed, we can drop it without changing the minimizer of the loss. We can therefore rewrite the estimand as

$$\begin{aligned} \theta^* \in \arg\min_\theta \quad & \theta^T\beta \\ \text{subject to } & \theta \in [0,1]^4 \\ & |\theta^T\beta^+| \le \epsilon^+ \\ & |\theta^T\beta^-| \le \epsilon^- \end{aligned} \tag{2.1}$$

In other words, the optimal fair derived predictor is the solution to a linear program (LP). We refer to this as the "true LP" since it defines the estimand. We now define an estimator $\widehat{\theta}$ as the solution to an "estimated LP."

2.5.3 Estimation

An estimator for θ^* is derived by computing estimates $\widehat{\beta}, \widehat{\beta}^+, \widehat{\beta}^-$ of the true LP coefficients and then solving the resulting estimated LP:

$$\begin{aligned} \widehat{\theta} \in \arg\min_{\theta} \quad & \theta^T \widehat{\beta} \\ \text{subject to} \quad & \theta \in [0,1]^4 \\ & |\theta^T \widehat{\beta}^+| \leq \epsilon^+ \\ & |\theta^T \widehat{\beta}^-| \leq \epsilon^- \end{aligned}$$

Any solution $\widehat{\theta}$ suffices. How should β, β^+, β^- be estimated? Before proposing a specific set of estimators, we first show that $S_{\widehat{\theta}}$ approaches optimal behavior at rates that depend on the performance of these estimators[§]. We define two quantities of interest: the *loss gap* and the *excess unfairness*, and give accompanying theorems. Proofs of all theorems are given in Appendix 2.B.

Following standard usage, we say that an estimator $\widehat{\psi}$ of a parameter ψ is consistent at rate $f(n)$ for some real-valued function $f(n)$ if $\|\widehat{\psi} - \psi\| = O_{\mathbb{P}}(1/f(n))$ for a suitable norm $\|\cdot\|$. For example, if $f(n) = \sqrt{n}$, then we say $\widehat{\psi}$ converges at $\sqrt{n}$ rates. We say that an estimator converges *faster* than $f(n)$ if $\|\widehat{\psi} - \psi\| = o_{\mathbb{P}}(1/f(n))$. When $\psi \in \mathbb{R}^k$ for some k, the norm we are interested in is the Euclidean norm defined by $\|\psi\|^2 = \psi^T\psi$. When ψ is a function of the random variable Z, the relevant norm is the L_2 norm with respect to $\mathbb{P}$, i.e., $\|\psi\|^2 = \int \psi^2 d\mathbb{P}(z)$.

Definition 2.5.1. The *loss gap* is $\text{Loss}(S_{\widehat{\theta}}) - \text{Loss}(S_{\theta^*})$, the difference in loss between the derived predictor and the optimal derived predictor.

We use the term *loss gap* rather than *excess loss* to acknowledge that the loss of $S_{\widehat{\theta}}$ can be less than the loss of S_{θ^*}, if $\widehat{\theta}$ falls outside the true constraints. Of course, this can only occur if $S_{\widehat{\theta}}$ violates the true fairness constraints, which can happen because the constraints are estimated.

Theorem 2.1. *(Loss gap.) Suppose that $\widehat{\beta}, \widehat{\beta}^+$, and $\widehat{\beta}^-$ are all consistent at rate $f(n)$. Under Assumptions A1-A3:*

$$\text{Loss}(S_{\widehat{\theta}}) - \text{Loss}(S_{\theta^*}) = O_{\mathbb{P}}(1/f(n))$$

Definition 2.5.2. The *excess unfairness of $S_{\widehat{\theta}}$ in the* cFPR is

$$\text{UF}^+(S_\theta) := \max\{|\,\text{cFPR}(S_{\widehat{\theta}}, 0) - \text{cFPR}(S_{\widehat{\theta}}, 1)| - \epsilon^+, 0\},$$

[§]We ignore optimization error, since this is a function of the number of optimization iterations and can be made arbitrarily small (Boyd and Vandenberghe, 2004).

and the *excess unfairness of* $S_{\widehat{\theta}}$ *in the* cFNR is

$$\mathrm{UF}^{-}(S_\theta) := \max\{|\,\mathrm{cFNR}(S_{\widehat{\theta}}, 0) - \mathrm{cFNR}(S_{\widehat{\theta}}, 1)| - \epsilon^{-}, 0\}$$

Since the estimated constraints should fluctuate around the true constraints, it's possible for $S_{\widehat{\theta}}$ to have error rate differences that are smaller than $\epsilon^{+}, \epsilon^{-}$, which motivates bounding these quantities below by 0.

Theorem 2.2. *(Excess unfairness.) Suppose that* $\widehat{\beta}, \widehat{\beta}^{+}$, *and* $\widehat{\beta}^{-}$ *are all consistent at rate* $f(n)$. *Under assumptions A1-A3:*

$$\max\left\{\mathrm{UF}^{+}(S_{\widehat{\theta}}), \mathrm{UF}^{-}(S_{\widehat{\theta}})\right\} = O_{\mathbb{P}}(1/f(n))$$

Remark 2. (The behavior of $\widehat{\theta}$ vs. $S_{\widehat{\theta}}$). Without assumptions about how the loss and fairness of $S_{\widehat{\theta}}$ depend on $\widehat{\theta}$, there is no guarantee about the rate at which $\widehat{\theta}$ will approach θ^*. This is not a concern, however, since the object of interest is not θ^* per se but a predictor that behaves like S_{θ^*}.

Arguably the simplest estimators for $\beta, \beta^{+}, \beta^{-}$ involve plugging an estimate of the regression function μ_0 into the identifying expressions in Proposition 1 and then computing empirical means in place of expectations. We propose instead using doubly robust, influence function-based estimators, which yield faster rates of convergence than plugin estimators in general nonparametric settings (van der Vaart, 2002; Tsiatis, 2006).

For ease of notation, let

$$\phi = \frac{1-D}{1-\pi}(Y - \mu_0) + \mu_0$$

denote the uncentered efficient influence function for $\mathbb{E}(Y^0)$, and let $\widehat{\phi}$ denote an estimate constructed from estimates $\widehat{\mu}_0$ and $\widehat{\pi}$ (Bickel et al., 1993; Hahn, 1998; van der Laan and Robins, 2003; Kennedy, 2016). Both these nuisance functions can be estimated with arbitrary nonparametric learners. To minimize the use of indices, let $\mathbb{P}_n(f(Z)) = n^{-1}\sum_{i=1}^{n} f(Z_i)$ denote the sample average of any function of Z. The doubly robust estimators for individual coefficients are:

$$\widehat{\beta}_{a,s} = \mathbb{P}_n[\mathbb{1}\{A = a, S = s\}(w^{+} - (w^{+} + w^{-})\widehat{\phi})] \tag{2.2}$$

$$\widehat{\mathrm{cFPR}}(S, a) = \frac{\mathbb{P}_n[\mathbb{1}\{A = a\}S(1 - \widehat{\phi})]}{\mathbb{P}_n[\mathbb{1}\{A = a\}(1 - \widehat{\phi})]} \tag{2.3}$$

$$\widehat{\mathrm{cFNR}}(S, a) = \frac{\mathbb{P}_n[\mathbb{1}\{A = a\}(1 - S)\widehat{\phi}]}{\mathbb{P}_n[\mathbb{1}\{A = a\}\widehat{\phi}]} \tag{2.4}$$

These estimates are assembled into the corresponding vectors $\widehat{\beta}, \widehat{\beta}^+, \widehat{\beta}^-$.

In order to obtain optimal convergence rates, it is generally necessary to estimate the nuisance functions $\widehat{\mu}_0$ and $\widehat{\pi}$ on one sample and then compute the sample mean $\mathbb{P}_n$ on an independent sample conditional on those estimates. To obtain full sample size efficiency, one can swap the folds, repeat the procedure, and average the results, an approach that is popularly called cross-fitting (Bickel and Ritov, 1988; Robins et al., 2008; Zheng and van der Laan, Mark, 2010; Chernozhukov et al., 2018). A k-fold version of cross-fitting with $k > 2$ is also possible. If $\widehat{\mu}_0$ and $\widehat{\pi}$ are assumed to be sufficiently "well-behaved," i.e. if they belong to Donsker classes, then no such sample splitting is necessary. We prefer to avoid this assumption and utilize sample splitting. See Appendix 2.C for a schematic of the sample splitting procedure.

The next theorem captures the double robustness property: under this sample splitting procedure, the coefficient estimators converge at a rate determined by the product of rates for the nuisance parameter estimators. One additional mild assumption is required.

A4. (Bounded propensity estimator)

$$\exists \gamma \in (0,1) \text{ s.t. } \mathbb{P}(\widehat{\pi}(A,X,S) \leq 1-\gamma) = 1$$

Assumption A4 is the empirical analogue of the positivity assumption (A2). It can be trivially satisfied by truncating $\widehat{\pi}$ at $1-\delta$, the positivity threshold in assumption A2.

Theorem 2.3. *(Double robustness.) Suppose that* $\|\widehat{\mu}_0 - \mu_0\|\|\widehat{\pi} - \pi\| = O_{\mathbb{P}}(g(n))$ *for some function* $g(n)$*. Under assumptions A1-A4:*

$$\|\widehat{\beta} - \beta\| = O_{\mathbb{P}}\left(\max\left\{g(n), n^{-1/2}\right\}\right)$$

and the same result holds for $\|\widehat{\beta}^+ - \beta^+\|$ *and* $\|\widehat{\beta}^- - \beta^-\|$*.*

Corollary 2.3.1. *If* $\|\widehat{\mu}_0 - \mu_0\|\|\widehat{\pi} - \pi\| = O_{\mathbb{P}}(n^{-1/2})$*, then* $\|\widehat{\beta} - \beta\| = O_{\mathbb{P}}(n^{-1/2})$*, and likewise for* $\|\widehat{\beta}^+ - \beta^+\|$ *and* $\|\widehat{\beta}^- - \beta^-\|$*.*

The corollary shows that it is possible to obtain $\sqrt{n}$ convergence, the fastest rate attainable in general nonparametric settings, even when the nuisance parameters are estimated at slower than $\sqrt{n}$ rates. The condition of Corollary 2.3.1 can be satisfied under relatively weak and nonparametric smoothness or sparsity assumptions (Györfi et al., 2002; Raskutti et al., 2011). For example, let $d = p + 2$ be the dimension of (A, X, S). If μ_0 and π are in Hölder classes with smoothness index $s > d/2$, then there exist nonparametric estimators $\widehat{\mu}_0$ and $\widehat{\pi}$ such that $\|\widehat{\mu}_0 - \mu\| = o_{\mathbb{P}}(n^{-1/4})$ and

$\|\widehat{\pi} - \pi\| = o_{\mathbb{P}}(n^{-1/4})$, in which case the product of the rates would be $o_{\mathbb{P}}(n^{-1/2})$, which is faster than $O_{\mathbb{P}}(n^{-1/2})$.

By contrast, a plugin version of $\widehat{\beta}$ would converge at a rate of $\sqrt{n}$ or the rate for $\|\widehat{\mu}_0 - \mu_0\|$, whichever is slower, and likewise for plugin versions of $\widehat{\beta}^+$ and $\widehat{\beta}^-$. Since $\sqrt{n}$ rates are generally unattainable in nonparametric regression, this means that plugin estimators would converge at slower than $\sqrt{n}$ rates, which, per Theorems 2.1 and 2.2, would result in slower convergence in the loss gap and excess unfairness.

2.5.4 Estimating performance of the derived predictor

Once $\widehat{\theta}$ has been computed, it is of interest to check both $\text{Loss}(S_{\widehat{\theta}})$ and the error rate differences $\Delta^+(S_{\widehat{\theta}})$, $\Delta^-(S_{\widehat{\theta}})$ of the resulting derived predictor $S_{\widehat{\theta}}$, for example to understand the performance "cost" of fairness and to check whether the procedure successfully controlled the error rate differences.

These estimates should be computed on a test set that is independent of the sample used to estimate $\widehat{\theta}$. Within the test set, the same sample splitting considerations apply: unless they are assumed to belong to Donsker classes, the nuisance parameters $\widehat{\mu}_0$ and $\widehat{\pi}$ should be estimated on separate folds from the folds used to compute the relevant sample means $\mathbb{P}_n$.

Since the estimators below are conditional on a fixed $\widehat{\theta}$, they can in fact be applied to any fixed parameter value $\theta \in [0,1]^4$. We define two additional quantities of interest.

Definition 2.5.3. The *loss change* for a derived predictor S_θ relative to an input predictor S is $\Gamma(S_\theta) = \text{Loss}(S_\theta) - \text{Loss}(S)$.

We refer to a *loss change* rather than an *increase in loss* because it is possible for S_θ to have smaller loss that S. This is not a typical expectation: in fair prediction problems, the set of fair classifiers is necessarily smaller than the set of fair and unfair classifiers, so there is a fairness-accuracy tradeoff. In the RAI setting, however, since predictors are typically trained to predict observable outcomes, their performance may be arbitrarily bad with respect to the potential outcome Y^0. It is therefore not implausible than a derived fair predictor could have higher accuracy than the input predictor.

Definition 2.5.4. The *predictive change* for a derived predictor S_θ relative to an input predictor S is $\mathbb{P}(S_{\widehat{\theta}} \neq S)$.

The predictive change is the proportion of input predictions that the post-processed predictor flips, which gives a measure of the effect of post-processing.

Once again, we propose using doubly robust estimators. These estimators are essentially identical to the estimators of the LP coefficients used in the previous section. Here, however, we are interested

in properties of these estimators, rather than properties of our derived predictor $S_{\widehat{\theta}}$. In particular, we are interested in deriving confidence intervals, in addition to guaranteeing rates of convergence. The estimators are

$$\widehat{\text{Loss}}(S_\theta) = \theta^T \widehat{\beta} + w^- \mathbb{P}_n(\widehat{\phi}) \qquad \text{(loss)}$$

$$\widehat{\Gamma}(S_\theta) = (\theta - \tilde{\theta})^T \widehat{\beta} \qquad \text{(loss change)}$$

$$\widehat{\text{cFPR}}(S_\theta, a) = \theta_{a,0} \left(1 - \widehat{\text{cFPR}}(S, a)\right) + \theta_{a,1} \widehat{\text{cFPR}}(S, a) \qquad \text{(cFPR)}$$

$$\widehat{\text{cFNR}}(S_\theta, a) = (1 - \theta_{a,0}) \left(\widehat{\text{cFNR}}(S, a)\right) + \theta_{a,1} \left(1 - \widehat{\text{cFNR}}(S, a)\right) \qquad \text{(cFNR)}$$

$$\widehat{\Delta}^+(S_\theta) = \theta^T \widehat{\beta}^+ \qquad \text{(error rate difference in cFPR)}$$

$$\widehat{\Delta}^-(S_\theta) = \theta^T \widehat{\beta}^- \qquad \text{(error rates difference in cFNR)}$$

where recall $\tilde{\theta} = (0, 1, 0, 1)$, so that $S_{\tilde{\theta}} = S$ (i.e., the derived predictor is simply the input predictor). Note that the loss estimator adds back in the portion of the loss that doesn't depend on θ and that we consequently removed from the LP in (2.1).

The predictive change does not involve counterfactual quantities, so it can be straightforwardly estimated with a plugin estimator: $\widehat{\mathbb{P}}(S_{\widehat{\theta}} \neq S) = \mathbb{P}_n \left\{ \mathbb{P}\left(S_{\widehat{\theta}} \neq S | A, S\right) \right\} =$

$$\mathbb{P}_n \left\{ \sum_{a \in \{0,1\}} [\theta_{a,0} \mathbb{1}\{A = a, S = 0\} + (1 - \theta_{a,1}) \mathbb{1}\{A = a, S = 1\}] \right\}$$

Since this is a sample average, it is asymptotically normal, and confidence intervals can be derived via the central limit theorem. In order to obtain asymptotic normality for the remaining estimators, we require an additional rate assumption on the nuisance parameter estimators:

A5. (Nuisance estimator rates).

$$\|\widehat{\mu}_0 - \mu_0\| = o_{\mathbb{P}}(1),$$

$$\|\widehat{\pi} - \pi\| = o_{\mathbb{P}}(1),$$

$$\|\widehat{\mu}_0 - \mu_0\| \|\widehat{\pi} - \pi\| = o_{\mathbb{P}}(1/\sqrt{n})$$

Assumption A5 states that $\widehat{\mu}_0$ and $\widehat{\pi}$ are consistent for μ_0 and π, and that the product of their errors is smaller than $1/\sqrt{n}$. As described above, this assumption can be satisfied under sparsity or smoothness conditions on μ_0 and π.

Theorem 2.4. *(Asymptotic normality.) Fix $\theta \in [0,1]^4$. Under assumptions A1-A5:*

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\text{Loss}}(S_\theta) - \text{Loss}(S_\theta)\right) \rightsquigarrow N\left(0, \text{var}\left(f_\theta\right)\right) \\
&\sqrt{n}\left(\widehat{\Gamma} - \Gamma\right) \rightsquigarrow N\left(0, \text{var}\left(f_\theta - f_{\tilde{\theta}}\right)\right) \\
&\sqrt{n}\left(\widehat{\text{cFPR}}(S_\theta, a) - \text{cFPR}(S_\theta, a)\right) \rightsquigarrow N\left(0, \text{var}\left(g_a\right)\right) \\
&\sqrt{n}\left(\widehat{\text{cFNR}}(S_\theta, a) - \text{cFNR}(S_\theta, a)\right) \rightsquigarrow N\left(0, \text{var}\left(h_a\right)\right) \\
&\sqrt{n}\left(\widehat{\Delta}^+ - \Delta^+\right) \rightsquigarrow N\left(0, \text{var}(g_0 - g_1)\right) \\
&\sqrt{n}\left(\widehat{\Delta}^- - \Delta^-\right) \rightsquigarrow N\left(0, \text{var}(h_0 - h_1)\right)
\end{aligned}$$

where

$$f_\theta = (w^+ - (w^+ + w^-)\phi)\theta_{A,S} + w^-\phi$$

$$g_a = (\theta_{a,1} - \theta_{a,0})\frac{\mathbb{1}\{A = a\}(1-\phi)(S - \text{cFPR}(S,a))}{\mathbb{E}[\mathbb{1}\{A = a\}(1-\phi)]}$$

$$h_a = (\theta_{a,1} - \theta_{a,0})\frac{\mathbb{1}\{A = a\}\phi(S - \text{cFNR}(S,0))}{\mathbb{E}[\mathbb{1}\{A = a\}\phi]}$$

and recall

$$\theta_{A,S} = \sum_{a,s \in 0,1} \theta_{a,s}\mathbb{1}\{A = a, S = s\}$$

and the estimators $\widehat{\text{Loss}}$, $\widehat{\Gamma}, \widehat{\text{cFPR}}$, $\widehat{\text{cFNR}}$, $\widehat{\Delta}^+$, $\widehat{\Delta}^-$ attain the nonparametric efficiency bound, meaning that no other estimator has smaller asymptotic variance.

Corollary 2.4.1. *Given a consistent estimator for $\text{var}(f_\theta)$, an asymptotically valid 95% confidence interval for $\text{Loss}(S_\theta)$ is given by $\widehat{\text{Loss}}(S_\theta) \pm 1.96 \cdot \widehat{\text{var}}(f_\theta)/\sqrt{n}$. An asymptotically valid test of the hypothesis $\text{Loss}(S_\theta) = C$ for any C consists of evaluating whether C is in the confidence interval. Analogous results hold for Γ, $\text{cFPR}, \text{cFNR}, \Delta^+, \Delta^+$.*

Perhaps the most natural estimators for these variances are the sample variances of $\widehat{f}_\theta$, $\widehat{g}_a$, $\widehat{h}_a$,- $\widehat{g}_0 - \widehat{g}_1$, and $\widehat{h}_0 - \widehat{h}_1$, where these quantities are defined by the following:

$$\widehat{f}_\theta = (w^+ - (w^+ + w^-)\widehat{\phi})\theta_{A,S} + w^-\widehat{\phi}$$

$$\widehat{g}_a = (\theta_{a,1} - \theta_{a,0})\frac{\mathbb{1}\{A = a\}(1 - \widehat{\phi})(S - \widehat{\text{cFPR}}(S_\theta, a))}{\mathbb{P}_n[\mathbb{1}\{A = a\}(1 - \widehat{\phi})]^{-1}}$$

$$\widehat{h}_a = (\theta_{a,1} - \theta_{a,0})\frac{\mathbb{1}\{A = a\}\widehat{\phi}(S - \widehat{\text{cFNR}}(S_\theta, 0))}{\mathbb{P}_n[\mathbb{1}\{A = a\}\widehat{\phi}]^{-1}}$$

The quantities f_θ, g_a, and h_a are the efficient influence functions for the loss and error rates.

2.6 Results

There is no previous method designed to achieve counterfactual equalized odds or related fairness criteria that we can compare our method to. We instead compare our method to an approach that uses plugin estimators for the LP coefficients, in order to illustrate the advantages of the doubly robust estimators.

2.6.1 Simulations

We use one set of simulations to illustrate Theorems 2.1-2.3 and another set to explore fairness-performance tradeoffs. We use equal misclassification weights $w^+ = w^- = 1$, so that false positives and false negatives contribute equally to the loss. Simulations illustrating Theorem 2.4 can be found in Appendix 2.E.

Each estimation procedure was run 500 times for each sample size $n \in \{100, 200, 500, 1000, 5000, 20000\}$. Since μ_0 is known here, the "true" loss and fairness values were computed on a separate validation set of size 500,000, using plugin estimators with the true μ_0. These values showed negligible variation over many repetitions.

Setup

First, we define a *pre-RAI* data generating process. Using this data, we train a predictor S to predict observable outcomes Y, mirroring how RAIs are typically constructed in practice. We then define a *post-RAI* data generating process, which only differs in that the predictor S now affects the decisions D. This emulates the way RAIs are intended to work in practice; for example, a criminal defendant labeled high-risk ($S = 1$) be a RAI might be less likely to be released pre-trial ($D = 0$) than they would have been prior to the introduction of the RAI. The data generating process is

designed to meet assumptions A1-A3, with $\pi(D \mid A, X, S)$ upper bounded at 0.975. It is described fully in Appendix 2.D.

Theorems 2.1-2.3

To simulate the estimation of the LP coefficient vectors at a particular rate, we add random noise ϵ of magnitude $o_{\mathbb{P}}(1/n^{1/4})$ to the nuisance parameters μ_0 and π[¶]. As described above, in general nonparametric settings, regression functions cannot be estimated at $\sqrt{n}$ rates, but they can be estimated at $n^{-1/4}$ rates under relatively weak assumptions (van der Vaart, 2002).

Figure 2.2 shows $\text{Loss}(S_{\widehat{\theta}})$ and the excess unfairness values $\text{UF}^+(S_{\widehat{\theta}})$, $\text{UF}^-(S_{\widehat{\theta}})$ for the post-processed predictor $S_{\widehat{\theta}}$ with fairness constraints $\epsilon^+ = 0.10,\ \epsilon^- = 0.20$. As expected, when doubly robust estimators are used, the loss and excess unfairness values converge at $\sqrt{n}$ rates to $\text{Loss}(S_{\theta^*})$, the loss of the optimal derived predictor ,and 0, respectively. When plugin estimators are used, the rates are slower than $\sqrt{n}$.

Fairness-performance tradeoffs

Figure 2.3 shows the loss change $\Gamma(S_{\theta^*}) = \text{Loss}(S_{\theta^*}) - \text{Loss}(S)$ for each point in a grid of fairness constraints $\epsilon^+,\ \epsilon^-$. Here, S is the Bayes-optimal predictor of Y^0 in our data generating scenario, meaning $S(A, X) = \mathbb{E}[Y^0 \mid A, X]$. Since any derived predictor necessarily has greater loss than the Bayes-optimal predictor, we refer to the loss change here equivalently as the *performance cost*.

In the data generating process used in the previous section, the Bayes-optimal predictor has absolute error rate differences of only 0.05 (Δ^+) and 0.04 (Δ^-), which leaves little room to illustrate the potential cost of fairness. For these simulations, therefore, we alter the data generating process slightly. (See Appendix 2.D). This results in a Bayes-optimal predictor with absolute error rate differences of 0.23 (Δ^+) and 0.40 (Δ^-) and a loss of 0.24, which are plausible values for a real predictor.

As expected, when $\epsilon^+ \geq \Delta^+(S)$ or $\epsilon^- \geq \Delta^-(S)$, the performance cost is 0: the input predictor already falls satisfies the fairness constraints, so our method simply returns the input predictor. As the tolerances tighten towards 0, the performance declines, though never substantially. For $\epsilon^+ = \epsilon^- = 0$, when the derived predictor is constrained to satisfy exact cEO, the loss increases by 0.10, to 0.34. The different values for $\Delta^+(S)$ and $\Delta^-(S)$ in the input predictor are reflected in the differing costs of satisfying fairness along the two axes: the cost of controlling $\Delta^+(S_\theta)$ are lower than the costs of controlling $\Delta^-(S_\theta)$.

[¶]The noise is added on the logit scale to ensure that $\widehat{\mu}_0, \widehat{\pi}$ remain in $[0, 1]$, and $\widehat{\pi}$ is again truncated to 0.975.

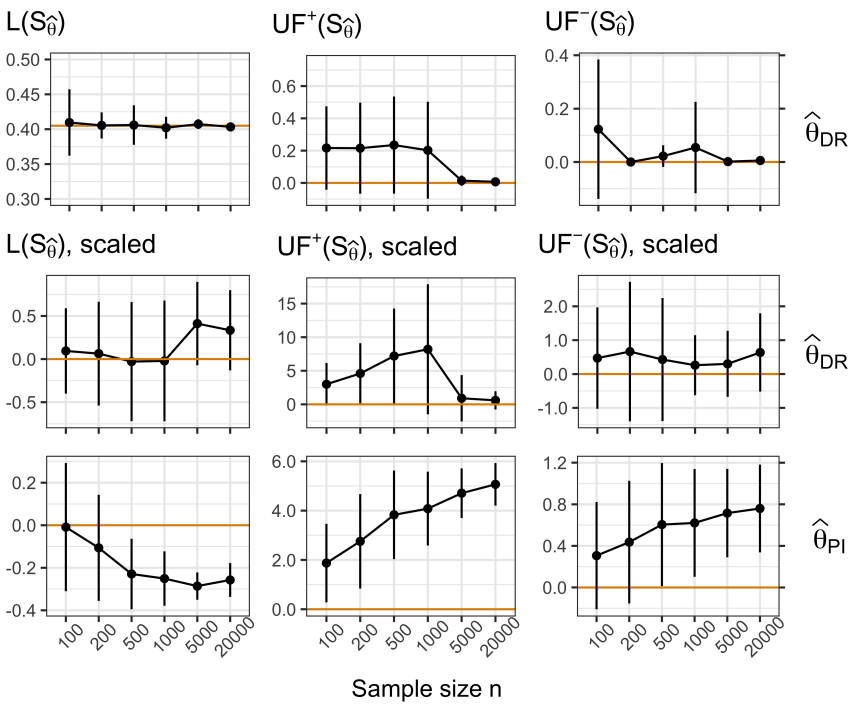


Figure 2.2: **(Illustration of Theorems 2.1-2.3)**. Loss $\mathrm{L}(S_{\widehat{\theta}})$ and excess unfairness values $\mathrm{UF}^+(S_{\widehat{\theta}}), \mathrm{UF}^-(S_{\widehat{\theta}})$ for the derived predictor $S_{\widehat{\theta}}$ for samples of size 100 to 20,000. Each vertical line represents a mean ± 1 sd over 500 simulations. Orange horizontal lines represent the loss of the optimal derived predictor S_{θ^*} (top left panel) or 0. The top row represents our doubly robust (DR) procedure and shows that the loss and excess unfairness converge to their target values. The bottom two rows represent values from the DR procedure or a plugin (PI) procedure, transformed by $\psi(S_{\widehat{\theta}}) \mapsto \sqrt{n}(\psi(S_{\widehat{\theta}}) - \psi(S_{\theta^*}))$, where ψ is Loss or UF^+ or UF^-, as appropriate. These rows illustrate that $\sqrt{n}$-convergence is only guaranteed for $\widehat{\theta}_{\mathrm{DR}}$: the scaled values for $\widehat{\theta}_{\mathrm{DR}}$ do not grow in n, while the scaled values for $\widehat{\theta}_{\mathrm{PI}}$ begin to diverge.

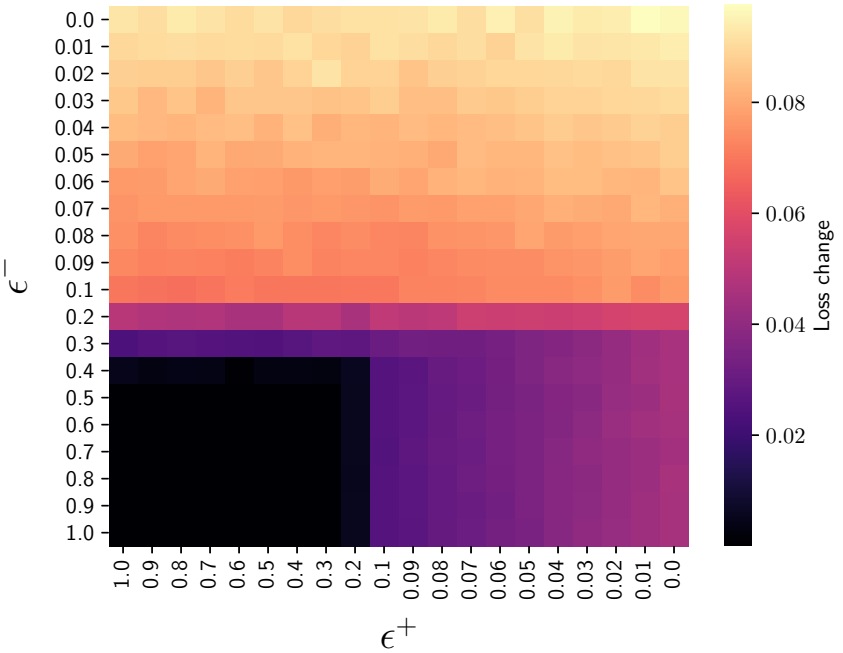


Figure 2.3: (Fairness-performance tradeoffs). Loss change $\Gamma(\theta^*) = \text{Loss}(S_{\theta^*}) - \text{Loss}(S)$ for the Bayes-optimal input predictor $S(A, X) = \mathbb{E}[Y^0 \mid A, X]$ and θ^* corresponding to different fairness constraints ϵ^+, ϵ^-. The black area represents fairness constraints that are looser than the error rate differences of the input predictor ($\Delta^+(S) = 0.24$, $\Delta^-(S) = 0.40$), which incur no performance cost. The highest performance cost (0.10) occurs when the error rates differences are both constrained to be 0, meaning the derived predictor S_{θ^*} satisfies cEO exactly.

Woodworth et al. (2017) showed that post-processing can result in predictors with poor performance, but it is unclear how likely this is to be a problem in practice. While the fairness-accuracy tradeoff naturally depends on the data generating process, our example illustrates that fairness can in some cases be achieved without substantial performance costs.

2.6.2 COMPAS data

We illustrate our method on the COMPAS recidivism dataset gathered by ProPublica (Angwin et al., 2016; Larson et al., 2016). COMPAS refers to a collection of tools designed to assess the risk of recidivism. The dataset comprises public arrest records, criminal records, and COMPAS RAI scores from Broward County, Florida, spanning 2013–2016. After filtering the data in the same manner as Larson et al. (2016) and restricting to defendants who are labeled African-American ($A = 0$) or Caucasian ($A = 1$), we are left with data for 5278 individuals (3175 African-American, 2103 Caucasian).

We utilize the COMPAS scores for general, as opposed to violent, recidivism. The scores are given in risk deciles. Since our method operates on a binary input predictor S, we follow ProPublica and set scores of 1-4 to $S = 0$ ("low risk") and scores of 5-10 to $S = 1$ ("high risk"). The outcome Y is recidivism within a two-year time period. (See Larson et al. (2016) for how recidivism is operationalized.) ProPublica's analysis focuses on the use of COMPAS to inform pretrial release decisions. The dataset includes dates in and out of jail but does not indicate whether defendants were released pretrial, so we set the treatment D to 0 if defendants left jail within three days of being arrested, and 1 otherwise. This yields 3645 released individuals (2158 African-American, 1487 Caucasian) and 1633 incarcerated individuals (1017 African-American, 616 Caucasian). Note that this threshold is somewhat arbitrary. Florida state law generally requires individuals to be brought before a judge for a bail hearing within 48 hours of arrest, but it may take time for individuals to post bail if they are required and able to do so.

The covariates X consist of gender (coded male or female), age (coded categorically for < 25, between 25 and 45, and > 45), the number of prior crimes, and charge degree (misdemeanor or felony). Without consulting with domain experts, it is difficult to assess the plausibility of the positivity and ignorability assumptions given these covariates. Hence we intend our analysis primarily to be illustrative of our method, and we resist drawing strong substantive conclusions about COMPAS.

We weight false positives and false negatives equally, i.e. we set $w^+ = w^- = 1$. We randomly split the data into training and test sets of equal size. For $\epsilon \in \{0, 0.01, 0.05, 0.10, 0.20, \ldots, 0.90, 1\}$, we set the fairness constraints to $\epsilon^+ = \epsilon^- = \epsilon$, compute the corresponding estimate $\widehat{\theta}$ on the training

set, and estimate properties of the post-processed predictor $S_{\widehat{\theta}}$ on the test set. We also estimate properties of the binarized COMPAS score S on the test set. We use random forests to estimate both the propensity scores $\widehat{\pi}$ and the outcome regression $\widehat{\mu}_0$. To reduce the variance of the estimates, we employ 5-fold cross-fitting: within the train set, we compute five estimates $\widehat{\theta}_j, j = 1, \ldots 5$, using four folds at a time to estimate the nuisance parameters and the held-out fold to compute $\widehat{\theta}_j$. Then $\widehat{\theta} := \frac{1}{5}\sum_j \widehat{\theta}_j$. We utilize the test set in an analogous fashion for the remaining estimators.

Table 2.1 contains estimates and confidence intervals for COMPAS and for the post-processed predictor corresponding to fairness constraints of $\epsilon^+ = \epsilon^- = 0.05$. The loss for COMPAS is 0.36, and the differences in the cFPR and cFNR are -0.24 and 0.16, respectively. The signs of these differences are consistent with what ProPublica found in their analysis with respect to observable Y: the false positive rates are higher for African-American defendants, while the false negative rates are higher for Caucasian defendants. The post-processing procedure successfully shrinks these differences to -0.05 and -0.03, which fall within the target range of $[-0.05, 0.05]$. This reduction corresponds to flipping 9% of the COMPAS scores, and it incurs an increase in risk of only 0.03.

The value of $\widehat{\theta}$ corresponding to $S_{\widehat{\theta}}$ here is $(0, 0.91, 0.23, 1)$. The 0 and the 1 indicate that $S_{\widehat{\theta}}$ does not change the COMPAS scores for African-American defendants who receive a "low-risk" score or Caucasian defendants who receive a "high-risk" score. The scores for high-risk African-American defendants are flipped to low-risk $1 - 0.91 = 9\%$ of the time, while the scores for low-risk Caucasian defendants are flipped to high-risk 23% of the time. This has the effect of increasing the false positive rate and decreasing the false negative rate for Caucasians, while moving the rates in the opposite directions for African-Americans.

Figure 2.4 shows the loss, error rate differences, and predictive change for fairness constraints ranging from 0 (requiring no gap in error rates) to 1 (imposing no fairness constraints). Each constraint induces an estimate $\widehat{\theta}$ and a corresponding post-processed predictor $S_{\widehat{\theta}}$. The estimated fairness gaps fall along or within the lines $y = \pm x$, indicating that each $S_{\widehat{\theta}}$ satisfies its target constraints. At the most stringent setting of 0, the loss for $S_{\widehat{\theta}}$ is approximately 0.40, which compares favorably with the estimated baseline loss of 0.36 for COMPAS. This $S_{\widehat{\theta}}$ flips slightly less than 20% of the scores.

For $\epsilon > 0.24$, the fairness constraints are essentially no longer active, since COMPAS itself satisfies these constraints. Indeed, as expected, the $\widehat{\theta}$ values for $\epsilon > 0.24$ are all essentially $[0, 1, 0, 1]$, meaning that $S_{\widehat{\theta}} = S$, and the estimated risk and fairness values all fall close to the estimated values for COMPAS. (There is still some variation in the estimated values due to randomness in the k-fold cross-fitting procedure.)

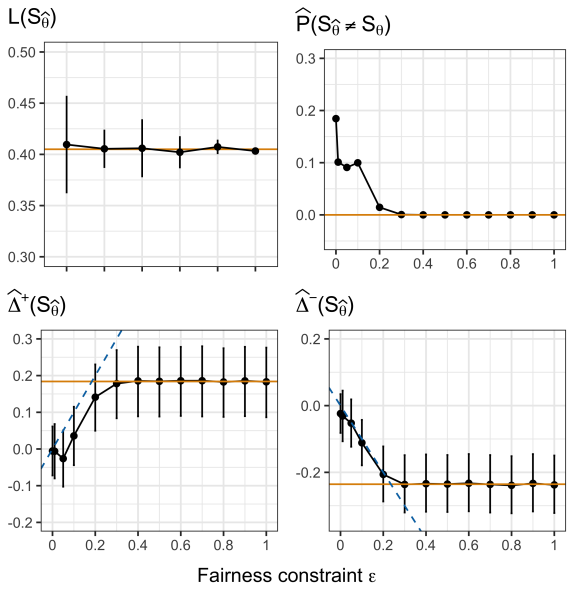


Figure 2.4: Convergence of the estimated loss $\widehat{\text{Loss}}(S_{\widehat{\theta}})$, predictive change $\widehat{\mathbb{P}}(S_{\widehat{\theta}} \neq S)$, and error rate differences $\widehat{\Delta}^+(S_{\widehat{\theta}}), \widehat{\Delta}^-(S_{\widehat{\theta}})$, for post-processed versions of the binarized COMPAS predictor. Fairness constraints are set to $\epsilon^+ = \epsilon^- = \epsilon$ over a range of values ϵ. Vertical lines are 95% CIs. Horizontal orange lines indicate the reference values for COMPAS, or 0 in the case of predictive change. The dashed blue lines $y = x$ and $y = -x$, mark the target fairness constraints.

These results illustrate that our approach performs as intended on a real dataset: if these data were indeed generated from a distribution satisfying the identifying assumptions, then our post-processed predictor would satisfy approximate counterfactual equalized odds while incurring little cost in performance.

2.6.3 Child welfare data

Cost-sensitive loss functions can drive $\widehat{\theta}$ to a trivial classifier that always predicts one class. We illustrate this phenomenon on a dataset representing calls to a child-welfare hotline in Allegheny County, Pennsylvania. The data comprises over 30,000 calls and contains over 1,000 features. The features describe allegations made in the call, assessments of risk made by hotline workers, and features pertaining to individuals associated with the call. Workers must decide whether to *screen in* a call, which means opening an investigation into the allegations. The baseline decision $D = 0$ is to screen out, meaning no investigation takes place. The outcome Y is re-referral to the hotline within a six month period. For further details about the child welfare setting and this dataset in particular, see Chouldechova et al. (2018) and Coston et al. (2020).

Unlike the COMPAS dataset, this dataset does not include a previously trained predictor. We therefore first build a predictor S that predicts Y^0, and then we post-process S. In this setting,

Table 2.1: Estimates and 95% confidence intervals for the loss Loss, loss change Γ, error rates cFPR and cFNR for groups 0 and 1, error rate differences Δ^+, Δ^-, and predictive change $\mathbb{P}(S_{\widehat{\theta}} \neq S)$ for the binarized COMPAS predictor S and the post-processed predictor $S_{\widehat{\theta}}$, with ϵ^+ and ϵ^- set to 0.05.

	S	$S_{\widehat{\theta}}$
$\widehat{\text{Loss}}(\cdot)$	0.36 (0.32, 0.41)	0.39 (0.35, 0.42)
$\widehat{\Gamma}(\cdot)$	–	0.03 (0.01, 0.04)
$\widehat{\text{cFPR}}(\cdot, 0)$	0.43 (0.36, 0.49)	0.39 (0.33, 0.45)
$\widehat{\text{cFPR}}(\cdot, 1)$	0.24 (0.18, 0.31)	0.42 (0.37, 0.47)
$\widehat{\text{cFNR}}(\cdot, 0)$	0.30 (0.25, 0.35)	0.36 (0.31, 0.40)
$\widehat{\text{cFNR}}(\cdot, 1)$	0.53 (0.46, 0.60)	0.41 (0.35, 0.46)
$\widehat{\Delta}^+(\cdot)$	-0.24 (-0.32, -0.15)	-0.05 (-0.12, 0.02)
$\widehat{\Delta}^-(\cdot)$	0.18 (0.09, 0.28)	-0.03 (-0.10, 0.05)
$\widehat{\mathbb{P}}(\cdot \neq S)$	–	0.09 (0.09, 0.09)

we have reason to believe that the identification assumptions in section 2.5.2 are plausible, once cases with the highest propensity for screen-in are removed; see Coston et al. (2020). (RAIs are not necessary or useful for cases that are already guaranteed to be screened in.) In order to accomplish this filtering, we first build a propensity score model using random forests on roughly one third of the data. The model appears well-calibrated, so we filter out the approximately 20% of the cases with estimated propensity scores greater than 0.99. Note that downstream results did not change substantially when these cases were left in.

We then train a classification random forest S to predict Y conditional on $A, X, D = 0$, using the same third of the data. Under the identifying assumptions, $Y|X, A, D = 0$ is equal in distribution to $Y^0|A, X$, so S is indeed an estimate of the target Y^0. Following recommended usage in this setting, we set the classification threshold to capture the top 25% riskiest cases (Chouldechova et al., 2018).

The predictor S has estimated error rate differences and 95% confidence intervals of $\widehat{\Delta}^+ = -0.02 \pm 0.01$ and $\widehat{\Delta}^- = 0.09 \pm 0.08$. It is unsurprising that these differences are small, given that rereferral rates are similar for Black (0.24) and White (0.27) cases. See Chouldechova (2017) for an examination of the relationship between base rates and error rates.

In order to have nontrivial (active) fairness constraints, we set $\epsilon^+ = \epsilon^- = 0.01$. Figure 2.5 shows the value of $\widehat{\theta}$ over a range of cost ratios w^+/w^- and w^-/w^+. When false positives are weighted more than 1.5 times as heavily as false negatives, post-processing returns classifiers that are very close to the simple majority classifier $S_{(0,0,0,0)} \equiv 0$. When false negatives are weighted more than 2 times as heavily as false positives, post-processing returns the simple minority classifier $S_{(1,1,1,1)} \equiv 1$. Since the input classifier is approximately fair, between those ranges, post-processing returns classifiers

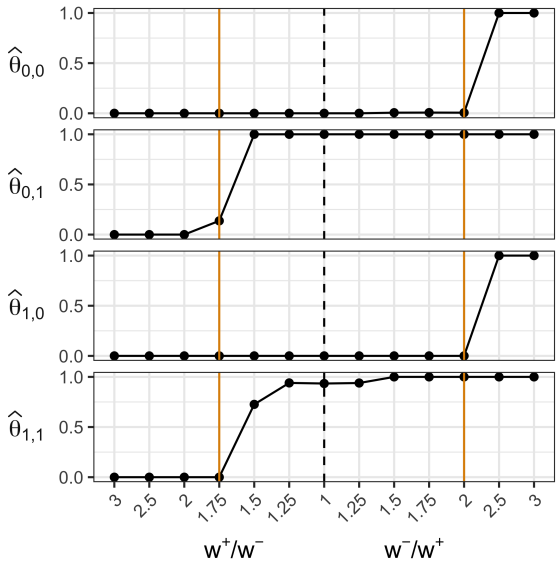


Figure 2.5: Cost-sensitive post-processing for the child welfare predictor over a range of cost ratios, with fairness constraints $\epsilon^+ = \epsilon^- = 0.01$. Each column represents a single $\widehat{\theta}$, with the four components $\theta_{a,s}$ for $a, s \in \{0, 1\}$, in rows. False positives are weighted between 1.25 and 3 times as heavily as false negatives to the left of the dashed line, and vice versa to the right. Extreme cost ratios push the post-processed classifier to a trivial classifier that always predicts 0 (to the left of the orange lines) or 1 (to the right). Between these, post-processing essentially returns the input predictor.

that are very close to the input classifier $S = S_{(0,1,0,1)}$, with only the fourth component $\widehat{\theta}_{1,1}$ deviating slightly from 1.

This behavior is expected. Note that a simple majority or minority classifier always satisfies counterfactual equalized odds, since the error rate differences are 0. Since the post-processed predictor only has access at runtime to two binary features, as either false positives or false negatives become sufficiently important, one of these simple classifiers will at some point become the lowest risk option. This is possible in principle when w^+ and w^- are equal, but it is guaranteed as their ratio grows.

Since this dataset did not include a pretrained predictor of Y^0, it would be preferable to adopt an in-processing approach, i.e. to train a predictor that satisfies the desired fairness constraints in a single stage, rather than training an unconstrained predictor and then post-processing it. We pursue this task in ongoing work.

2.7 Discussion and conclusion

In this paper we considered fairness in risk assessment instruments (RAIs), which are naturally concerned with potential outcomes rather than strictly observable outcomes. We defined the fairness criterion *approximate counterfactual equalized odds* (approximate cEO), which allows users to negotiate the tradeoff between fairness and performance. We argued that this fairness criterion is likelier than other candidate criteria to reduce discriminatory disparate impact, which we defined as $D \not\perp\!\!\!\perp A \mid Y^0$.

We presented a method to post-process an existing binary predictor to satisfy approximate cEO using doubly robust estimators, and we showed that our method has favorable convergence properties. Our rate results translate readily to the post-processing setting of (Hardt et al., 2016), in which the outcome of interest is the observable Y and the fairness criterion is (approximate) observational equalized odds.

Once it is constructed, the post-processed predictor requires access at runtime only to the sensitive feature and the input predictor, making it relatively feasible to implement on top of existing RAIs. A predictor trained from scratch would be constrained by the set of covariates available in deployment, whereas the post-processing approach allows researchers to devise a set of suitable deconfounding covariates and then collect an appropriate dataset on a one-time basis.

In closing, we note that from our perspective, notions of fairness in predictive systems ought to be subordinate to notions of fairness grounded in the actual decisions or events that those systems inform, and the impact that those decisions have on people's lives. Though little is currently known about how decision makers respond to RAIs, there is some evidence that judges do not have much faith in recidivism predictions and that RAIs can have little impact on decisions (Jonnson, 2018; Stevenson, 2018). As RAIs and the general public's understanding of how they function co-evolve, it is likely that the ways in which decision makers respond to them will evolve as well.

Nevertheless, it seems plausible that some fairness criteria for RAIs are likelier than others to lead to increased (un)fairness with respect to decisions and outcomes. While this is ultimately an empirical question, we believe that this kind of consideration ought to ground discussions of fairness in RAIs and predictive systems generally. As long as there are domains involving high stakes decisions that we do not wish to fully automate, RAIs will remain relevant, and so will the task of ensuring that they lead to a society that is more fair, not less.

2.A Proofs of propositions

For convenience, we restate our three assumptions.

1. (Consistency) $Y = DY^1 + (1-D)Y^0$
2. (Positivity) $\exists \delta \in (0,1) : \mathbb{P}(\pi(A,X,S) \leq 1-\delta) = 1$
3. (Ignorability) $Y^0 \perp\!\!\!\perp D \mid A, X, S$

Proof of Proposition 1 (Identification of error rates for input predictor S)

$$\begin{aligned}
\text{cFPR}(S,a) &= \mathbb{P}(S=1 \mid Y^0=0, A=a) \\
&= \frac{\mathbb{P}(S=1, Y^0=0, A=a)}{\mathbb{P}(Y^0=0, A=a)} \\
&= \frac{\mathbb{E}[S(1-Y^0)\mathbb{1}\{A=a\}]}{\mathbb{E}[(1-Y^0)\mathbb{1}\{A=a\}]} \\
&= \frac{\mathbb{E}[S(1-\mathbb{E}[Y^0 \mid A,X,S,D=0)]\mathbb{1}\{A=a\}}{\mathbb{E}[(1-\mathbb{E}[Y^0 \mid A,X,S,D=0])\mathbb{1}\{A=a\}]} \\
&= \frac{\mathbb{E}[S(1-\mu_0)\mathbb{1}\{A=a\}]}{\mathbb{E}[(1-\mu_0)\mathbb{1}\{A=a\}]}
\end{aligned}$$

$$\begin{aligned}
\text{cFNR}(S,a) &= \mathbb{P}(S=0 \mid Y^0=1, A=a) \\
&= \frac{\mathbb{P}(S=0, Y^0=1, A=a)}{\mathbb{P}(Y^0=1, A=a)} \\
&= \frac{\mathbb{E}[(1-S)Y^0\mathbb{1}\{A=a\}]}{\mathbb{E}[Y^0\mathbb{1}\{A=a\}]} \\
&= \frac{\mathbb{E}[(1-S)\mathbb{E}[Y^0 \mid A,X,S,D=0]\mathbb{1}\{A=a\}}{\mathbb{E}[\mathbb{E}[Y^0 \mid A,X,S,D=0]\mathbb{1}\{A=a\}]} \\
&= \frac{\mathbb{E}[(1-S)\mu_0\mathbb{1}\{A=a\}]}{\mathbb{E}[\mu_0\mathbb{1}\{A=a\}]}
\end{aligned}$$

The fourth equality in both derivations uses iterated expectation as well as positivity and ignorability, and the fifth equality uses consistency.

Proof of Proposition 2 (Identification of the loss and fairness constraints)

Considering just the first component of the loss, we have:

$$\begin{aligned}
(w^+)\mathbb{P}(S_\theta = 1, Y^0 = 0) &= (w^+)\mathbb{E}[S_\theta(1 - Y^0)] \\
&= (w^+)\mathbb{E}[\mathbb{E}[S_\theta(1 - Y^0)|A, S, X]] \\
&= (w^+)\mathbb{E}\left\{\mathbb{E}[S_\theta|A, S](1 - \mathbb{E}[Y^0|A, X, S])\right\} \\
&= (w^+)\mathbb{E}\left\{\theta_{A,S}(1 - \mathbb{E}[Y^0|A, X, S, D = 0])\right\} \\
&= (w^+)\mathbb{E}\left\{\theta_{A,S}(1 - \mu_0)\right\}
\end{aligned}$$

where the third equality uses that S_θ only depends on (A, S), the fourth uses the definition of $\theta_{A,S}$ and ignorability, and the fifth uses consistency. Similar reasoning shows that $(w^-)\mathbb{P}(S = 0, Y^0 = 1) = (w^-)\mathbb{E}\left\{(1 - \theta_{A,S})\mu_0\right\}$. Combining these, we have

$$\begin{aligned}
\text{Loss}(S_\theta) :&= w^+\mathbb{P}(S = 1, Y^0 = 0) + w^-\mathbb{P}(S = 0, Y^0 = 1) \\
&= \mathbb{E}[\theta_{A,S}(w^+ - (w^+ + w^-)\mu_0)] + (w^-)\mathbb{E}[\mu_0] \\
&= \theta^T\beta + (w^-)\mathbb{E}[\mu_0]
\end{aligned}$$

We turn now to the fairness constraints. The error rates of the derived predictor S_θ depend on the error rates on the input predictor S as follows. Beginning with $\text{cFPR}(S_\theta, a)$, we have:

$$\begin{aligned}
&\mathbb{P}(S_\theta = 1 \mid Y^0 = 0, A = a) = \\
&\quad \sum_{s\in\{0,1\}} \mathbb{P}(S_\theta = 1 \mid Y^0 = 0, A = a, S = s)\mathbb{P}(S = s \mid Y^0 = 0, A = a) \\
= &\quad \sum_{s\in\{0,1\}} \mathbb{P}(S_\theta = 1 \mid A = a, S = s)\mathbb{P}(S = s \mid Y^0 = 0, A = a) \\
= &\quad \theta_{a,0}(1 - \text{cFPR}(S, a)) + \theta_{a,1}\,\text{cFPR}(S, a)
\end{aligned}$$

where the first equality simply involves conditioning on S, and the second equality uses that $S_\theta \perp\!\!\!\perp Y^0 \mid A, S$. In other words, the false positive rate of S_θ depends only on θ and the false positive rate of the input predictor S. For the cFNR, by similar reasoning, we have:

$$\begin{aligned}
&\mathbb{P}(S_\theta = 0 \mid Y^0 = 1, A = a) = \\
&\quad 1 - \theta_{a,0}(\text{cFNR}(S, a)) + \theta_{a,1}(\text{cFNR}(S, a) - 1)
\end{aligned}$$

The identification statements in the proposition follow by simply substituting in the expressions for $\text{cFPR}(S,a), \text{cFNR}(S,a)$ from Proposition 1 and rearranging.

2.B Proofs of Theorems

2.B.1 Theorem 1 (Loss gap)

We first introduce a lemma used in the proof of the theorem. The lemma gives sufficient conditions under which the optimal value of an estimated convex program converges at a particular rate $f(n)$ to the optimal value of the target convex program. It is a adaptation of Theorem 3.5 in Shapiro (1991) that follows immediately from Theorems 2.1 and 3.4 in that same paper.

Lemma 2.4.1. *(Shapiro, 1991) Let Θ be a compact subset of $\mathbb{R}^k$. Let $C(\Theta)$ denote the set of continuous real-valued functions on Θ, with $\mathscr{L} = C(\Theta) \times \ldots \times C(\Theta)$ the r-dimensional Cartesian product. Let $\psi(\theta) = (\psi_0, \ldots, \psi_r) \in \mathscr{L}$ be a vector of convex functions. Consider the quantity α^* defined as the solution to the following convex optimization program:*

$$\begin{aligned} \alpha^* = \min_{\theta \in \Theta} \quad & \psi_0(\theta) \\ \textit{subject to } & \psi_j(\theta) \leq 0, \ j = 1, \ldots, r \end{aligned}$$

Assume that Slater's condition holds, so that there is some $\theta \in \Theta$ for which the inequalities are satisfied and non-affine inequalities are strictly satisfied, i.e. $\psi_j(\theta) < 0$ if ψ_j is non-affine. Now consider a sequence of approximating programs, for $n = 1, 2, \ldots$:

$$\begin{aligned} \widehat{\alpha}_n = \min_{\theta \in \Theta} \quad & \widehat{\psi}_{0n}(\theta) \\ \textit{subject to } & \widehat{\psi}_{jn}(\theta) \leq 0, \ j = 1, \ldots, r \end{aligned}$$

with $\widehat{\psi}_n(\theta) := \left(\widehat{\psi}_{0n}, \ldots, \widehat{\psi}_{rn}\right) \in \mathscr{L}$. Assume that $f(n)(\widehat{\psi}_n - \psi)$ converges in distribution to a random element $W \in \mathscr{L}$ for some real-valued function $f(n)$. Then:

$$f(n)(\widehat{\alpha}_n - \alpha_0) \rightsquigarrow L$$

for a particular random variable L. It follows that $\widehat{\alpha}_n - \alpha_0 = O_{\mathbb{P}}(1/f(n))$.

Proof of theorem

We expand the loss by introducing the term $\widehat{\beta}^T\widehat{\theta}$, which is the quantity that is minimized in the course of computing $\widehat{\theta}$. We proceed by splitting the loss into two terms and showing that each of those terms is $O_{\mathbb{P}}(1/f(n))$.

Proof. The loss gap can be expanded as follows:

$$\begin{aligned}\text{Loss}(S_{\widehat{\theta}}) - \text{Loss}(S_{\theta^*}) &= \beta^T\widehat{\theta} - \beta^T\theta^* \\ &= \underbrace{\left(\beta^T\widehat{\theta} - \widehat{\beta}^T\widehat{\theta}\right)}_{(1)} + \underbrace{\left(\widehat{\beta}^T\widehat{\theta} - \beta^T\theta^*\right)}_{(2)}\end{aligned}$$

For term (1), we have

$$\begin{aligned}\widehat{\theta}^T\left(\beta - \widehat{\beta}\right) &\leq \|\widehat{\theta}\|\|\beta - \widehat{\beta}\| \\ &\leq 2\|\beta - \widehat{\beta}\| \\ &= O_{\mathbb{P}}(1/f(n))\end{aligned}$$

where the first line uses Cauchy-Schwarz, the second line follows from the fact that $\widehat{\theta} \in [0,1]^4$, and the third line follows by assumption. For term (2), we rely on Lemma 2.4.1. Note that we can write

$$\begin{aligned}\text{Loss}(S_{\theta^*}) = \min_{\theta\in\Theta} \quad & \psi_0(\theta) \\ \text{subject to } & \psi_j(\theta) \leq 0,\ j = 1,\ldots,4 \\ \widehat{\text{Loss}}(S_{\widehat{\theta}}) = \min_{\theta\in\Theta} \quad & \widehat{\psi}_0(\theta) \\ \text{subject to } & \widehat{\psi}_j(\theta) \leq 0,\ j = 1,\ldots,4\end{aligned}$$

with $\Theta = [0,1]^4$, and $\psi(\theta) = (\psi_0(\theta),\ldots,\psi_4(\theta))$ defined by

$$\begin{aligned}\psi(\theta) &= (\text{Loss},\ \Delta^+ - \epsilon^+,\ -\Delta^+ - \epsilon^+,\ \Delta^- - \epsilon^-,\ -\Delta^- - \epsilon^-) \\ \widehat{\psi}(\theta) &= (\widehat{\text{Loss}},\ \widehat{\Delta}^+ - \epsilon^+,\ -\widehat{\Delta}^+ - \epsilon^+,\ \widehat{\Delta}^- - \epsilon^-,\ -\widehat{\Delta}^- - \epsilon^-)\end{aligned}$$

where for brevity we omit the argument S_θ to Loss and the error rate differences Δ^+, Δ^-. Since these are linear programs, Slater's condition is satisfied. (The LPs are always feasible, since $(0,0,0,0)$ and $(1,1,1,1)$ are always solutions.) By assumption, each of the estimators in $\widehat{\psi}(\theta)$ converges at rate $f(n)$, so $f(n)\left(\widehat{\psi}(\theta) - \psi(\theta)\right)$ converges to some (unknown) random variable. (We rule out

pathological cases in which this does not happen.) Per Lemma 2.4.1, it follows that $\widehat{\beta}^T\widehat{\theta} - \beta^T\theta^* = \widehat{\text{Loss}}(S_{\widehat{\theta}}) - \text{Loss}(S_{\theta^*}) = O_{\mathbb{P}}(1/f(n))$.

The sum of the two terms in the loss gap is therefore also $O_{\mathbb{P}}(1/f(n))$. □

2.B.2 Theorem 2 (Excess unfairness)

The proof relies on the following lemma, as well as the convergence of the estimated LP coefficient vectors $\widehat{\beta}^+, \widehat{\beta}^-$. When $\widehat{\beta}^+, \widehat{\beta}^-$ are close to β^+, β^-, the excess unfairness must be small for any $\theta \in \Theta = [0,1]^4$, including of course $\widehat{\theta}$.

Lemma 2.4.2. *Let ξ, W be constant vectors and $\widehat{\xi}_n, \widehat{W}_n$ be random vectors, with $\|\xi - \widehat{\xi}_n\| = O_{\mathbb{P}}(1/f(n))$ for some real-valued $f(n)$. If, for all $M > 0$, $\mathbb{P}(\|W - \widehat{W}_n\| > M) \leq \mathbb{P}(\|\xi - \widehat{\xi}_n\| > CM)$ for some constant C, then $\|W - \widehat{W}_n\| = O_{\mathbb{P}}(1/f(n))$.*

Proof. For any $\epsilon > 0$, there exists some $M_\epsilon > 0$ such that $\mathbb{P}(f(n)\|\xi - \widehat{\xi}_n\| > M_\epsilon) < \epsilon$ for all n large enough. Set $M = M_\epsilon/C$. Then $\mathbb{P}(f(n)\|W - \widehat{W}_n\| > M) < \epsilon$ for all n large enough. □

Proof of theorem

Proof. We have

$$
\begin{aligned}
&\mathbb{P}\left(\text{UF}^+(S_{\widehat{\theta}}) > \delta \text{ or } \text{UF}^-(S_{\widehat{\theta}}) > \delta\right) \\
&\leq \mathbb{P}\Big(|\theta^T\beta^+| - |\theta^T\widehat{\beta}^+| > \delta \text{ or } |\theta^T\beta^-| - |\theta^T\widehat{\beta}^-| > \delta \\
&\qquad\qquad \text{for some } \theta \in [0,1]^4\Big) \\
&\leq \mathbb{P}\Big(|\theta^T\beta^+ - \theta^T\widehat{\beta}^+| > \delta \text{ or } |\theta^T\beta^- - \theta^T\widehat{\beta}^-| > \delta \\
&\qquad\qquad \text{for some } \theta \in [0,1]^4\Big) \\
&\leq \mathbb{P}\Big(\|\theta\| \cdot \|\widehat{\beta^+} - \beta^+\| > \delta \text{ or } \|\theta\| \cdot \|\widehat{\beta^-} - \beta^-\| > \delta \\
&\qquad\qquad \text{for some } \theta \in [0,1]^4\Big)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(2\|\widehat{\beta}^+ - \beta^+\| > \delta \text{ or } 2\|\widehat{\beta}^- - \beta^-\| > \delta\right) \\
&\leq \mathbb{P}\left(2\|\widehat{\beta}^+ - \beta^+\| > \delta\right) + \mathbb{P}\left(2\|\widehat{\beta}^- - \beta^-\| > \delta\right) \\
&= \mathbb{P}\left(\|\widehat{\beta}^+ - \beta^+\| > \delta/2\right) + \mathbb{P}\left(\|\widehat{\beta}^- - \beta^-\| > \delta/2\right)
\end{aligned}
$$

where the third inequality uses Cauchy-Schwartz, the fourth uses that $\theta \in [0,1]^4 \implies \|\theta\| \leq 2$, and the fifth uses the union bound. (The norm here is the Euclidean norm.) The reasoning in the first

inequality is as follows: if $\text{UF}^+(S_{\widehat{\theta}}) > \delta$, then $|\widehat{\theta}^T\beta^+| - |\widehat{\theta}^T\widehat{\beta}^+| > \delta$, since $\widehat{\theta}^T\widehat{\beta}^+ \leq \epsilon^+$ by construction. A necessary condition, then, is that $|\theta^T\beta^+| - |\theta^T\widehat{\beta}^+| > \delta$ for some $\theta \in [0,1]^4$.

It follows from Lemma 2.4.2 that

$$\max\left\{\text{UF}^+(S_{\widehat{\theta}}), \text{UF}^-(S_{\widehat{\theta}})\right\} = O_{\mathbb{P}}(1/f(n))$$

□

2.B.3 Theorem 2.3 (Double robustness.)

In this proof and the proof of Theorem 2.4, for any function $f : \mathcal{Z} \mapsto \mathbb{R}$, we let $\mathbb{P}(f) = \int f(z)d\mathbb{P}(z)$ denote the expected value of the random variable $f(Z)$ conditional on the function f. For example, $\mathbb{P}(\widehat{\phi}) = \int \widehat{\phi}(z)d\mathbb{P}(z)$ is the expected value of $\widehat{\phi}(Z)$ conditional on $\widehat{\phi}$. We use the notation $a \lesssim b$ to denote that $a \leq Cb$ for some constant $C > 0$ that does not depend on n.

The proofs of each of these theorems utilize the following two lemmas.

Lemma 2.4.3. *Let W be a function of (at most) A, X, S such that $\|W\| \leq M < \infty$ for some M. Then, under assumption A2 (positivity),*

$$\mathbb{P}\left(W(\widehat{\phi} - \phi)\right) \lesssim \|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\|$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(W(\widehat{\phi} - \phi)\right) &= \mathbb{P}\left(W\left(\frac{1-D}{1-\widehat{\pi}}(Y - \widehat{\mu}_0) + \widehat{\mu}_0 - \frac{1-D}{1-\pi}(Y - \mu_0) - \mu_0\right)\right) \\
&= \mathbb{P}\left(W\left(\frac{1-D}{1-\widehat{\pi}}(\mu_0 - \widehat{\mu}_0) + \widehat{\mu}_0 - \frac{1-D}{1-\pi}(\mu_0 - \mu_0) - \mu_0\right)\right) \\
&= \mathbb{P}\left(W\left(\frac{1-\pi}{1-\widehat{\pi}}(\mu_0 - \widehat{\mu}_0) + \widehat{\mu}_0 - \mu_0\right)\right) \\
&= \mathbb{P}\left(W\left(\frac{(\mu_0 - \widehat{\mu}_0)(\widehat{\pi} - \pi)}{1-\pi}\right)\right) \\
&\leq \frac{1}{\delta}\mathbb{P}(W(\mu_0 - \widehat{\mu}_0)(\widehat{\pi} - \pi)) \\
&\leq \frac{1}{\delta}\|W\|\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\| \\
&\lesssim \|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\|
\end{aligned}$$

where the second and third lines use iterated expectation, conditioning on (A, X, S); the fifth line uses Assumption A2 (positivity); and the sixth line uses the Cauchy-Schwarz inequality. □

The next lemma is a restatement of Lemma 2 in Kennedy et al. (2020).

Lemma 2.4.4. *(Kennedy, 2020)*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi) = O_{\mathbb{P}}\left(\frac{\|\widehat{\phi} - \phi\|}{\sqrt{n}}\right)$$

It follows immediately from Lemma 2.4.3 that

$$(\mathbb{P}_n - \mathbb{P})(\widehat{\phi} - \phi) = O_{\mathbb{P}}\left(\frac{\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\|}{\sqrt{n}}\right)$$

Proof of the theorem

We assume throughout that $\|\widehat{\phi} - \phi\| = O_{\mathbb{P}}(1)$, meaning $\widehat{\phi}$ does not diverge from ϕ. Recall that in the statement of the theorem, $g(n)$ is the convergence rate of $\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\|$, i.e. $\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\| = O_{\mathbb{P}}(g(n))$.

Proof. Note that for a fixed length vector $v \in \mathbb{R}^k$:

$$\|\widehat{v} - v\| = O_{\mathbb{P}}(f(n)) \iff \widehat{v}_j - v_j = O_{\mathbb{P}}(f(n)),\ j = 1, \dots k$$

where the norm on the left is the Euclidean norm. It therefore suffices to show that the rate result in the theorem holds for each component of $\widehat{\beta}, \widehat{\beta}^+, \widehat{\beta}^-$.

It is straightforward to show that the identification results in Propositions 1 and 2 hold when μ_0 is replaced by ϕ. Starting with $\widehat{\beta}_{a,s}$, a component of $\widehat{\beta}$, we have the following, by simple addition and subtraction of measures:

$$\begin{aligned}\widehat{\beta}_{a,s} - \beta_{a,s} = &(\mathbb{P}_n - \mathbb{P})\left\{\mathbb{1}\{A = a, S = s\}(w^+ - (w^+ + w^-)\phi)\right\} \ + \\ &(\mathbb{P}_n - \mathbb{P})\left\{(w^- - w^+)(\widehat{\phi} - \phi)\right\} \ + \\ &\mathbb{P}\left\{(w^- - w^+)(\widehat{\phi} - \phi)\right\}\end{aligned}$$

The first term is $O_{\mathbb{P}}(1/\sqrt{n})$ by the central limit theorem. The second term is $O_{\mathbb{P}}(g(n)/\sqrt{n})$ by Lemmas 2.4.3 and 2.4.4. The third term is $O_{\mathbb{P}}(g(n))$ by Lemma 2.4.3. Thus

$$\widehat{\beta}_{a,s} - \beta_{a,s} = O_{\mathbb{P}}\left(\max\{n^{-1/2}, g(n)\}\right)$$

and the result therefore holds for $\|\widehat{\beta} - \beta\|$.

We now turn to $\widehat{\beta}^+$ and $\widehat{\beta}^-$. It suffices to show that the rate result holds for $\widehat{\text{cFPR}}(S, a)$ and $\widehat{\text{cFNR}}(S, a)$, $a \in \{0, 1\}$. For notational convenience, let $\gamma_a = (1-\phi)\mathbb{1}\{A = a\}$ and $\widehat{\gamma}_a = (1-\widehat{\phi})\mathbb{1}\{A =$

$a\}$. We have

$$\begin{aligned}
\widehat{\text{cFPR}}(S,a) - \text{cFPR}(S,a) &= \frac{\mathbb{P}_n[S\widehat{\gamma}_a]}{\mathbb{P}_n[\widehat{\gamma}_a]} - \frac{\mathbb{E}[S\gamma_a]}{\mathbb{E}[\gamma_a]} \\
&= \frac{\mathbb{P}_n[S\widehat{\gamma}_a]\mathbb{P}[\gamma_a] - \mathbb{P}[S\gamma_a]\mathbb{P}_n[\widehat{\gamma}_a]}{\mathbb{P}_n[\widehat{\gamma}_a]\mathbb{P}[\gamma_a]} \\
&= \frac{\mathbb{P}[\gamma_a]\Big(\mathbb{P}_n[S\widehat{\gamma}_a] - \mathbb{P}[S\gamma_a]\Big) - \mathbb{P}[S\gamma_a]\Big(\mathbb{P}_n[\widehat{\gamma}_a] - \mathbb{P}[\gamma_a]\Big)}{\mathbb{P}_n[\widehat{\gamma}_a]\mathbb{P}[\gamma_a]} \\
&= \mathbb{P}_n[\widehat{\gamma}_a]^{-1}\Big\{ \underbrace{(\mathbb{P}_n[S\widehat{\gamma}_a] - \mathbb{P}[S\gamma_a])}_{(1)} - \text{cFPR}(S,a)\underbrace{(\mathbb{P}_n[\widehat{\gamma}_a] - \mathbb{P}[\gamma_a])}_{(2)} \Big\} \qquad (2.5)
\end{aligned}$$

The two terms can be expanded as follows:

$$\begin{aligned}
(1) &= (\mathbb{P}_n - \mathbb{P})S\gamma_a + (\mathbb{P}_n - \mathbb{P})(S(\widehat{\gamma}_a - \gamma_a)) + \mathbb{P}(S(\widehat{\gamma}_a - \gamma_a)) \\
(2) &= (\mathbb{P}_n - \mathbb{P})\gamma_a + (\mathbb{P}_n - \mathbb{P})(\widehat{\gamma}_a - \gamma_a) + \mathbb{P}(\widehat{\gamma}_a - \gamma_a)
\end{aligned}$$

Once again, in both these expressions the first term is $O_{\mathbb{P}}(1/\sqrt{n})$ by the central limit theorem, the second term is $O_{\mathbb{P}}(g(n)/\sqrt{n})$ by Lemma 2.4.4, and the third term is $O_{\mathbb{P}}(g(n))$. Under assumption A4 (bounded propensity estimator), $\mathbb{P}_n[\widehat{\gamma}_a]^{-1}$ is bounded a.s., and $\text{cFPR}(S,a)$ is always bounded in $[0,1]$. Therefore, we can rewrite (2.5) as

$$(\mathbb{P}_n - \mathbb{P})\Big\{\mathbb{P}_n[\widehat{\gamma}_a]^{-1}\big(S - \text{cFPR}(S,a)\big)\gamma_a\Big\} + O_{\mathbb{P}}(g(n)) \qquad (2.6)$$

This expression is $O_{\mathbb{P}}(\max\{n^{-1/2}, g(n)\}$ and therefore so is $\|\widehat{\beta}^+ - \beta^+\|$. The result for $\widehat{\text{cFNR}}(S,a)$, and consequently for $\|\widehat{\beta}^- - \beta^-\|$, follows by identical reasoning, with γ_a redefined as $\phi\mathbb{1}\{A=a\}$ so that $\text{cFNR}(S,a) = \mathbb{E}[(1-S)\gamma_a]/\mathbb{E}[\gamma_a]$. □

2.B.4 Theorem 2.4 (asymptotic normality)

For ease of reference, we reiterate the following quantities defined in the theorem.

$$\begin{aligned}
f_\theta &= (w^+ - (w^+ + w^-)\phi)\theta_{A,S} + w^-\phi \\
g_a &= (\theta_{a,1} - \theta_{a,0})\frac{\mathbb{1}\{A=a\}(1-\phi)(S - \text{cFPR}(S,a))}{\mathbb{E}[\mathbb{1}\{A=a\}(1-\phi)]} \\
h_a &= (\theta_{a,1} - \theta_{a,0})\frac{\mathbb{1}\{A=a\}\phi(S - \text{cFNR}(S,0))}{\mathbb{E}[\mathbb{1}\{A=a\}\phi]}
\end{aligned}$$

Define $\widehat{f}_\theta,\ \widehat{g}_a,\ \widehat{h}_a$ analogously, substituting $\widehat{\phi},\ \widehat{\text{cFPR}},\ \widehat{\text{cFNR}}$ for ϕ, cFPR, cFNR.

We first prove the statements for the loss Loss and loss change Γ. Note that $\text{Loss}(S_\theta) = \mathbb{E}[f_\theta]$ and $\widehat{\text{Loss}}(S_\theta) = \mathbb{P}_n[\widehat{f}_\theta]$. By simple addition and subtraction of measures, we have

$$\widehat{\text{Loss}}(S_\theta) - \text{Loss}(S_\theta) = (\mathbb{P}_n - \mathbb{P})f_\theta + (\mathbb{P}_n - \mathbb{P})(\widehat{f}_\theta - f_\theta) + \mathbb{P}(\widehat{f}_\theta - f_\theta)$$

Under assumption A5 (nuisance parameter rates), the second term in this sum is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2.4.4, and the third term is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2.4.3. We can therefore write

$$\widehat{\text{Loss}}(S_\theta) - \text{Loss}(S_\theta) = (\mathbb{P}_n - \mathbb{P})f_\theta + o_{\mathbb{P}}(1/\sqrt{n})$$

By equivalent reasoning,

$$\widehat{\Gamma}(\theta, S) - \Gamma(\theta, S) = (\mathbb{P}_n - \mathbb{P})(f_\theta - f_{\tilde{\theta}}) + o_{\mathbb{P}}(1/\sqrt{n})$$

Therefore, by the central limit theorem,

$$\begin{aligned}
\sqrt{n}\left(\widehat{\text{Loss}}(S_\theta) - \text{Loss}(S_\theta)\right) &\rightsquigarrow N\left(0, \text{var}\left(f_\theta\right)\right) \\
\sqrt{n}\left(\widehat{\Gamma} - \Gamma\right) &\rightsquigarrow N\left(0, \text{var}\left(f_\theta - f_{\tilde{\theta}}\right)\right)
\end{aligned}$$

as claimed.

The reasoning for the fairness estimators is virtually identical. From equation (2.6) and Assumption A5 (nuisance parameter rates), we have

$$\widehat{\text{cFPR}}(S, a) - \text{cFPR}(S, a) = (\mathbb{P}_n - \mathbb{P})g_a + o_{\mathbb{P}}(1/\sqrt{n})$$

By equivalent reasoning,

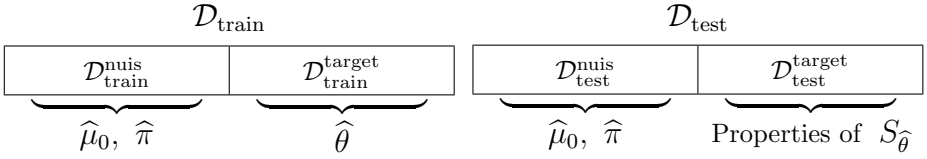
$$\begin{aligned}
\widehat{\text{cFNR}}(S, a) - \text{cFNR}(S, a) &= (\mathbb{P}_n - \mathbb{P})h_a + o_{\mathbb{P}}(1/\sqrt{n}) \\
\widehat{\text{cFPR}}(S, a) - \text{cFPR}(S, a) &= (\mathbb{P}_n - \mathbb{P})(g_0 - g_1) + o_{\mathbb{P}}(1/\sqrt{n}) \\
\widehat{\text{cFPR}}(S, a) - \text{cFPR}(S, a) &= (\mathbb{P}_n - \mathbb{P})(h_0 - h_1) + o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

The results follow immediately from the central limit theorem.

2.C Sample splitting

A training sample, $\mathcal{D}_{\text{train}}$, is used to estimate $\widehat{\theta}$, while a separate sample $\mathcal{D}_{\text{test}}$ is used to estimate the risk and fairness properties of the derived predictor $S_{\widehat{\theta}}$ conditional on $\widehat{\theta}$. Within each sample, separate folds are used to estimate the nuisance parameters μ_0 and $\widehat{\pi}$ and the target parameters.

The following schematic illustrates this procedure. k-fold cross fitting can be used within each sample to recover full sample size efficiency. For convenience, we assume that each of the four samples is of size n, though our results require only that each sample is $O(n)$.



2.D Simulations: data generating process

The data generating process used in section 2.6.1 to illustrate Theorems 1 and 2 is as follows, for data $Z = (A, X, S, D, Y^0, Y^1, Y)$.

$$
\begin{aligned}
\mathbb{P}(A = 1) &= 0.3 \\
X \mid A &\sim \mathrm{N}(A * (1, -0.8, 4, 2)^T, I_4) \\
\mathbb{P}_{\text{pre}}(D = 1 \mid A, X) &= \min\{0.975, \text{expit}((A, X)^T(0.2, -1, 1, -1, 1))\} \\
\mathbb{P}_{\text{post}}(D = 1 \mid A, X, S) &= \min\{0.975, \text{expit}((A, X, S)^T(0.2, -1, 1, -1, 1, 1))\} \\
\mathbb{P}(Y^0 = 1 \mid A, X) &= \text{expit}((A, X)^T(-5, 2, -3, 4, -5)) \\
\mathbb{P}(Y^1 = 1 \mid A, X) &= \text{expit}((A, X)^T(1, -2, 3, -4, 5)) \\
Y &= (1 - D)Y^0 + DY^1
\end{aligned}
$$

where I_4 denotes the 4×4 identity matrix and N denotes a Gaussian distribution. The predictor $S(A, X)$ is trained using random forests. The pre-RAI decision making process doesn't depend on S; the post-RAI process does.

For the simulations used in section 2.6.1 to illustrate fairness-performance tradeoffs, the distribution is identical except that $\mathbb{P}(Y^0 = 1 \mid A, X) = \text{expit}((A, X)^T(-4, 0.4, 0.6, 0.8, -1))$.

2.E Asymptotic normality of doubly robust estimators

To illustrate Theorem 2.4, an additional set of simulations was run using the same data generating process described above. First, θ was randomly set to $(0.74, 1.0, 0, 0.8)$. (Note that solutions to a linear program with a compact feasible set must occur at an extreme point of the set, so the presence of 0 and/or 1 in $\widehat{\theta}$ and θ^* is virtually guaranteed.) The "true" risk $\text{Loss}(S_\theta)$, risk change $\Gamma(S_\theta)$, and error rate differences $\Delta^+(S_\theta), \Delta^-(S_\theta)$ were again computed on a separate validation set of size 500,000, using plugin estimators with the true μ_0. For conciseness, we omit results for $\widehat{\text{cFPR}}$ and $\widehat{\text{cFNR}}$.

Figures 2.6 and 2.7 illustrate results for plugin vs. doubly robust estimators of these quantities, for samples of size 100 to 20,000. Each vertical line represents a mean ± 1 sd over 500 simulations. Orange horizontal lines represent the true parameter values (top rows in each figure) or 0. The top row represents shows that the doubly robust (DR) estimators converge to their target values. The bottom two rows represent values from the DR estimator or a plugin (PI) estimator, transformed by $\psi(S_{\widehat{\theta}}) \mapsto \sqrt{n}(\widehat{\psi} - \psi)$, where $\widehat{\psi}, \psi$ are the relevant estimator and parameter for that column. These rows illustrate that $\sqrt{n}$-convergence is only guaranteed for $\widehat{\theta}_{\text{DR}}$: the scaled values for $\widehat{\theta}_{\text{DR}}$ do not grow in n, while the scaled values for $\widehat{\theta}_{\text{SR}}$ begin to diverge (at least for $\widehat{\Gamma}$ and $\widehat{\Delta}^-$).

Table 2.2 contains coverage results of 95% confidence intervals for the error rates, error rate differences, loss, and loss change for the same arbitrary S_θ. The CIs were constructed using sample variances. To ensure that they did not exceed the bounds of the possible parameter values (i.e. $[0, 1]$ for the loss and error rates, $[-1, 1]$ for the error rate differences and loss change), the CIs were constructed using the Delta method, via the transformations $\widehat{\psi} \mapsto \text{logit}(\widehat{\psi})$, for $\widehat{\psi} \in \{\widehat{\text{cFPR}}, \widehat{\text{cFNR}}, \widehat{\text{Loss}}\}$, or $\widehat{\psi} \mapsto \text{logit}((\widehat{\psi} + 1)/2)$, for $\widehat{\psi} \in \{\widehat{\Delta}^+, \widehat{\Delta}^-, \widehat{\Gamma}\}$. Nominal coverage is achieved for various quantities at various sample sizes, but since the coverage guarantees are asymptotic, it is not surprising that it is not achieved everywhere. Interestingly, the median coverage rate in the table is 0.95.

A separate set of CIs was computed without using the Delta method; those results did not differ substantially and are therefore omitted here.

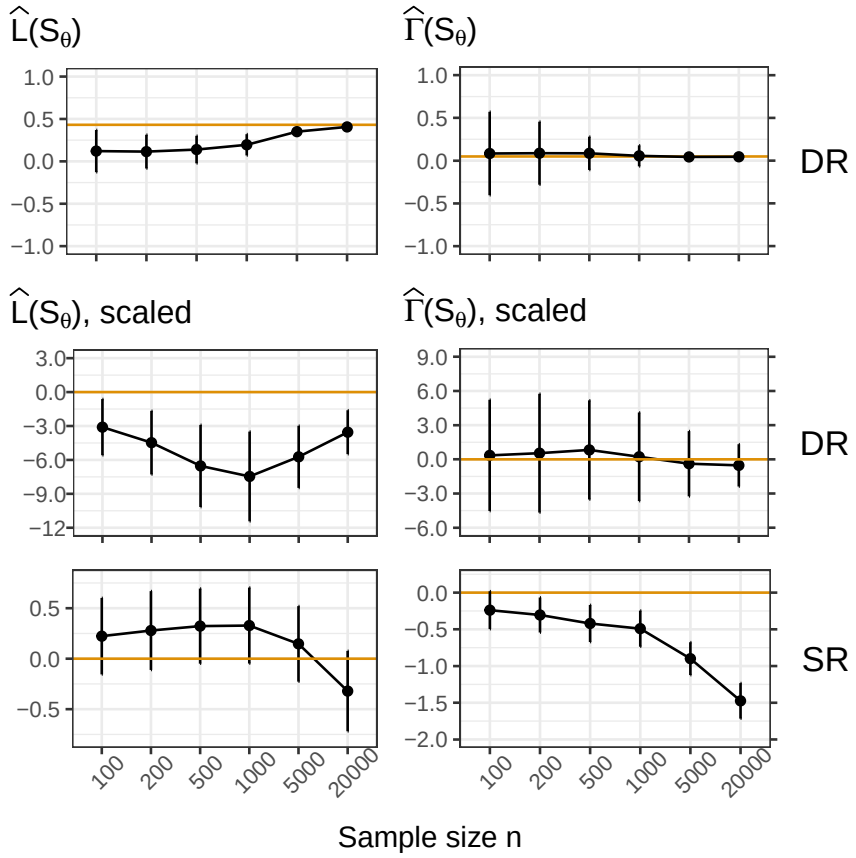


Figure 2.6: Doubly robust (DR) vs. plugin (PI) estimates of the loss and loss change for an arbitrary derived predictor S_θ, with $\theta = (0.74, 1.0, 0, 0.8)$, for samples of size 100 to 20,000.

	100	200	500	1000	5000	20000
$\widehat{\text{Loss}}(S_\theta)$	0.98	0.92	0.87	0.84	0.84	0.85
$\widehat{\Gamma}(S_\theta)$	1.00	0.99	0.93	0.94	0.71	0.78
$\widehat{\text{cFPR}}(S_\theta, 0)$	0.99	0.98	0.98	0.96	0.94	0.95
$\widehat{\text{cFPR}}(S_\theta, 1)$	0.90	0.89	0.93	0.95	0.96	0.57
$\widehat{\text{cFNR}}(S_\theta, 0)$	0.99	0.99	0.98	0.99	0.92	0.93
$\widehat{\text{cFNR}}(S_\theta, 1)$	0.99	0.99	0.99	1.00	0.98	0.71
$\widehat{\Delta}^+(S_\theta)$	0.98	0.98	0.97	0.99	0.97	0.92
$\widehat{\Delta}^-(S_\theta)$	0.99	1.00	0.99	0.99	0.94	0.94

Table 2.2: 95% CI coverage at sample sizes ranging from 100 to 20,000 for the loss, loss change, error rates, and error rate differences, for an arbitrary derived predictor S_θ with parameter $\theta = (0.74, 1.0, 0, 0.8)$. Coverage varies by estimator and sample size, though the median coverage is 95%.

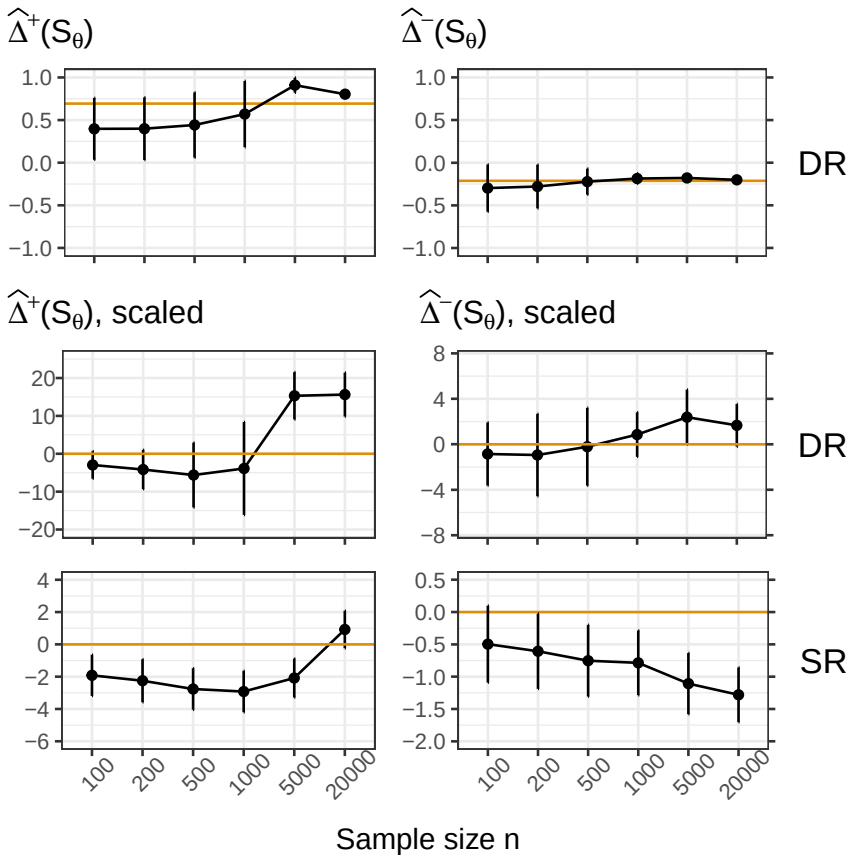


Figure 2.7: Doubly robust (DR) vs. plugin (PI) estimates of the error rate differences for an arbitrary derived predictor S_θ, with $\theta = (0.74, 1.0, 0, 0.8)$, for samples of size 100 to 20,000.

2.F Notation

Input data	
$Z = (A, X, D, S, Y) \sim \mathbb{P}$	Sensitive feature A, covariates X, decision (treatment, intervention) D, input predictor S, outcome Y
Derived predictor	
$S_\theta \sim \text{Bern}(\theta_{A,S})$	Predictor derived from S
$\theta_{a,s} = \mathbb{P}(S_\theta = 1 \mid A = a, S = s)$	Conditional probability that defines S_θ
$\theta_{A,S} = \sum_{a,s\in\{0,1\}} \mathbb{1}\{A = a, S = s\}\theta_{a,s}$	RV that takes value $\theta_{a,s}$ with probability $\mathbb{P}(A = a, S = s)$
$\theta = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})^T$	Optimization parameter
$\tilde{\theta} = (0, 1, 0, 1)$	The value such that $S_{\tilde{\theta}} = S$
Nuisance parameters	
$\pi = \pi(A, X, S) = \mathbb{P}(D = 1 \mid A, X, S)$	Propensity score for the decision
$\mu_0 = \mu_0(A, X, S, D) = \mathbb{E}[Y \mid A, X, S, D = 0]$	Outcome regression
$\phi = \frac{1-D}{1-\pi}(Y - \mu_0) + \mu_0$	Uncentered influence function for $E[Y^0]$
Loss parameters	
w^+, w^-	Weights on the false positive and false negative rates
$\beta_{a,s} = \mathbb{E}[\mathbb{1}\{A = a, S = s\}(w^+ - (w^+ + w^-)\mu_0)]$	A coefficient in the loss, for $a, s \in \{0, 1\}$
$\beta = (\beta_{0,0}, \beta_{0,1}, \beta_{1,0}, \beta_{1,1})^T$	Vector of loss coefficients
$\text{Loss}(S_\theta) = w^+\mathbb{P}(S_\theta = 1, Y^0 = 0) + w^-\mathbb{P}(S_\theta = 0, Y^0 = 1)$	Loss of S_θ, equivalent to $\theta^T\beta + w^-\mathbb{E}[\mu_0]$
$\Gamma(S_\theta)$	loss change $\text{Loss}(S_\theta) - \text{Loss}(S)$
Fairness parameters	
$\text{cFPR}(S_\theta, a) = \mathbb{P}(S_\theta = 1 \mid Y^0 = 0, A = a)$	Counterfactual FPR for S_θ for group a
$\text{cFNR}(S_\theta, a) = \mathbb{P}(S_\theta = 0 \mid Y^0 = 1, A = a)$	Counterfactual FNR for S_θ for group a
$\beta^+ = (1 - \text{cFPR}(S, 0), \text{cFPR}(S, 0), \text{cFPR}(S, 1) - 1, -\text{cFPR}(S, 1))$	Coefficients for the fairness constraints
$\beta^- = (-\text{cFNR}(S, 0), \text{cFNR}(S, 0) - 1, cFNR(S, 1), 1 - \text{cFNR}(S, 1))$	
$\Delta^+(S_\theta) = \theta^T\beta^+ = \text{cFPR}(S_\theta, 0) - \text{cFPR}(S_\theta, 1)$	Error rate differences of the predictor S_θ in the cFPR
$\Delta^-(S_\theta) = \theta^T\beta^- = \text{cFNR}(S_\theta, 0) - \text{cFNR}(S_\theta)$	Error rate differences of the predictor S_θ in the cFNR
ϵ^+, ϵ^-	Fairness constraints on Δ^+ and Δ^-
$\text{UF}^+(S_\theta) = \max(\mid \Delta^+(S_\theta) \mid - \epsilon^+, 0)$	Excess unfairness in the cFPR
$\text{UF}^-(S_\theta) = \max(\mid \Delta^-(S_\theta) \mid - \epsilon^-, 0)$	Excess unfairness in the cFNR
Optimal fair derived predictor	
$\arg\min_\theta \text{Loss}(S_\theta)$ s.t. $\lvert\Delta^+(S_\theta)\rvert \le \epsilon^+, \lvert\Delta^-(S_\theta)\rvert \le \epsilon^-$	Parameter defining the optimal fair derived predictor S_{θ^*}

Chapter 3

Least Squares for Observable and Counterfactual Fairness

3.1 Introduction

Classification and regression models are increasingly widely used to inform or render decisions in domains such as healthcare, criminal justice, education, hiring, and consumer finance. Given the high-stakes nature of such decisions, it is important to ensure that these models are both accurate, to maximize their overall benefits and minimize their overall harms; and fair, so that the benefits and harms do not accrue disproportionately to already (under)privileged groups. In recent years, there have been many well-publicized cases of algorithmic systems whose performance varies over sensitive features such as race and gender in ways that appear to harm marginalized populations (Angwin and Larson, 2016; Buolamwini and Gebru, 2018; Obermeyer et al., 2019).

In response to concerns such as these, the algorithmic fairness community has developed a wide array of methods for removing or minimizing unfairness in models. In some cases, the most accurate models under consideration do not satisfy a chosen fairness criterion, so there is a fairness-accuracy tradeoff (Friedler et al., 2019; Menon and Williamson, 2018; Zhao and Gordon, 2019). Many methods therefore aim to maximize predictive accuracy subject to a bound on some quantitative unfairness criterion (Zafar et al., 2017; Donini et al., 2018; Woodworth et al., 2017). Some methods adopt a complementary perspective, seeking to minimize unfairness subject to an accuracy constraint (Zafar et al., 2017; Coston et al., 2021). Many strict versions of fairness criteria are pairwise unsatisfiable in real-world settings, so there may also be fairness-fairness tradeoffs (Chouldechova, 2017; Kleinberg et al., 2017; Kim et al., 2020).

In many cases, however, the tradeoffs are small to nonexistent: model fairness can be increased with minimal loss of accuracy, or vice versa (Dutta et al., 2020; Coston et al., 2021; Rodolfa et al., 2021). Recently there has been growing interest in characterizing these tradeoffs both theoretically and empirically for specific problems and specific classes of models (Berk et al., 2017; Kim et al., 2020; Liu and Vicente, 2021). However, current methods for illuminating these tradeoffs are designed to handle *observable* accuracy and fairness criteria, i.e. criteria that depend on observable outcomes. They do not address *counterfactual* criteria, which depend on counterfactual outcomes and which are relevant to many settings in which algorithms are used to support decision making. Additionally, they do not readily accommodate users who wish to improve the fairness and/or accuracy of an existing benchmark model rather than exploring the fairness-accuracy space.

We propose *Least Squares for Fairness (LS-Fair)*, a simple and flexible framework that builds predictors as weighted combinations of basis functions that are chosen by the user. Within this framework, we develop three methods: (1) minimizing risk subject to fairness constraints, (2) minimizing unfairness subject to a risk constraint, and (3) efficiently generating a large class of unfairness-penalized predictors. The weights in method (3) have a closed-form expression that varies smoothly over a vector of unfairness penalty parameters, allowing users to trace out paths in fairness-accuracy space. It is computationally extremely fast to compute and evaluate thousands or even tens of thousands of models of this form. These methods accommodate users who wish to improve the fairness of an existing model without sacrificing accuracy, or vice versa, or who wish to understand fairness-accuracy and fairness-fairness tradeoffs in their problem.

Our contributions are as follows. We develop a simple, flexible, and computationally efficient framework that enables users to build and evaluate a large set of models that represent different fairness-accuracy and fairness-fairness tradeoffs (Sections 3.3–3.5). This framework allows users to target specific fairness and accuracy constraints as well as to explore the fairness-accuracy space. Our approach accommodates a range of both observable and counterfactual performance and fairness criteria. It collapses the distinction between in-processing and post-processing, enabling users to combine previously trained and newly trained predictors. We analyze the convergence rate of our estimators and provide a finite-sample performance guarantee. We illustrate our method on simulated data (Section 3.6) and on real data. Our method substantially improves both the fairness and accuracy of the COMPAS recidivism predictor (Section 3.7), and it yields many predictors that perform comparably to or better than other fairness methods on an income prediction task, while allowing users much more flexibility in the final model form (Section 3.8).

3.2 Background and Related Work

We use the terms "predictor" and "model" interchangeably to refer to any mapping from covariates to outputs that is intended to estimate an unknown quantity, whether that quantity is an unobserved label or an as-yet unrealized outcome. We use "accuracy" or "risk" to refer to any measure that tracks how well a predictor estimates the target quantity, such as mean-squared error or 0-1 error, and "performance" to refer to a model's joint accuracy and fairness characteristics.

3.2.1 Ways of achieving fairness

The fairness literature generally distinguishes three approaches for developing fair predictors. *Pre-processing* approaches transform the input data to remove bias (Calmon et al., 2017; Feldman et al., 2015; Kamiran and Calders, 2012; Zemel, 2013). *In-processing* or *in-training* approaches enforce fairness via constraints or regularization terms during the learning process (Donini et al., 2018; Kamishima et al., 2012; Woodworth et al., 2017; Zafar et al., 2017). *Post-processing* approaches learn functions to map the outputs of existing predictors to new outputs (Hardt et al., 2016; Pleiss et al., 2017; Kim et al., 2019). Our approach enables users to combine previously existing predictors with newly trained predictors or other basis functions, essentially collapsing the distinction between in-processing and post-processing.

3.2.2 Observational and Counterfactual Fairness

Many popular fairness criteria place restrictions on the joint distribution of predictions, outcomes, and a sensitive feature. For example, the criterion of *independence*, also known as *statistical parity* or *demographic parity*, requires that the predictions be independent of the sensitive feature (Calders et al., 2009; Barocas et al., 2018), while *separation* or *equalized odds* requires that they be independent conditional on the outcome (Hardt et al., 2016). Criteria that are sensitive to the outcome may be defined with respect to observable or potential (aka "counterfactual") outcomes. Counterfactual versions of these criteria are appropriate in risk assessment settings, i.e. settings in which the model is meant to estimate the risk of an adverse outcome absent an intervention (Coston et al., 2020). In these settings, the potential outcomes of interest are the outcomes that would occur if, possibly counterfactually, a decision variable were set to some baseline level (Neyman, 1923; Holland, 1986). Examples arise in areas such as healthcare, where doctors must predict who would develop complications without further treatment; criminal justice, where judges must predict who would recidivate if released pretrial; and consumer finance, where banks must predict who would default if issued a loan.

A distinct set of causal fairness criteria consider counterfactuals with respect to the sensitive feature rather than with respect to a decision variable (Kilbertus et al., 2017; Kusner et al., 2017; Nabi and Shpitser, 2018; Zhang and Bareinboim, 2018; Nabi et al., 2019; Wang et al., 2019). These criteria consider questions like "what would the risk prediction be if the defendant had been of a different race their whole life?" rather than "what would the outcome be if this person were released pretrial?" We do not consider these criteria here; see (Mishler et al., 2021), Section 3.2 for further discussion.

Most of the existing fairness literature is concerned with observable fairness criteria and accuracy measures. To our knowledge, only two papers have developed methods to satisfy the type of counterfactual criteria described above. Mishler et al. (2021) developed a post-processing method that maximizes accuracy while satisfying (approximate) counterfactual equalized odds or related fairness criteria. Their approach takes as input a binary classifier and outputs a randomized binary classifier. In contrast, our method applies to both classification and regression and it combines in-processing and post-processing. Coston et al. (2021) developed a method to minimize various unfairness measures subject to an accuracy constraint. They considered both the observable setting and a *selective labels* setting, when the outcome of interest is observed only for a (non-representative) subset of the population. Although the terminology differs, this is essentially equivalent to the counterfactual setting that we consider. Their methods outputs a single classification or regression model, and it is relatively computationally intensive. In contrast to both these methods, our framework can handle any of the following: (1) minimizing risk subject to fairness constraints, (2) minimizing unfairness subject to an accuracy constraint, and (3) efficiently producing a large set of models that vary in their risk and fairness properties. Our methods apply to a large class of observable and counterfactual accuracy and fairness criteria.

3.2.3 Fairness-accuracy and fairness-fairness tradeoffs

Within a candidate set of models, the most accurate model and the most fair model may not be the same model, in which case there is a fairness-accuracy tradeoff. The shape of this tradeoff depends on the model set, the accuracy and fairness criteria, and the distribution of the data (Dutta et al., 2020). While some papers emphasize the unavoidable existence of such tradeoffs (Corbett-Davies et al., 2017; Menon and Williamson, 2018; Woodworth et al., 2017; Zhao and Gordon, 2019), other papers have found that in practical settings they are sometimes so small as to be irrelevant; that is, relative to a baseline model, it may be possible to substantially improve a given fairness criterion with little to no decrement in accuracy, or vice versa (Coston et al., 2021; Rodolfa et al., 2021).

Different fairness criteria may also trade off with one another. In their strictest form, many fairness criteria are mutually unsatisfiable in real-world conditions (Chouldechova, 2017; Kleinberg et al., 2017). In practice, many methods make use of continuous-valued relaxations of these criteria, which may be more or less simultaneously satisfiable, to a degree that again depends on the modeling choices and data distribution.

Recent work aims to characterize fairness-accuracy and fairness-fairness tradeoffs both theoretically (Dutta et al., 2020; Kim et al., 2020) and empirically (Berk et al., 2017; Liu and Vicente, 2021). Like Berk et al. (2017), our penalized predictor method uses fairness regularization terms to trace out different paths in fairness-accuracy space; however, their results consider observable accuracy and fairness measures, whereas ours encompass both observable and counterfactual measures. We also consider a class of fairness criteria that yield closed-form solutions, which avoids the need for iterative optimization, and we provide theoretical guarantees for our methods.

In some cases, users have clear accuracy or fairness constraints that they wish their models to satisfy. These constraints might derive from moral, legal, or business considerations. For example, a business might wish to ensure that a hiring algorithm generates positive recommendations for roughly equal percentages of male and female applicants in order to avoid potential disparate impact. Conversely, a business might wish to improve the fairness of an existing model without sacrificing accuracy (profit). We provide an explicit correspondence between our constrained and penalized predictors and show how the set of penalized models can be "seeded" with models that target specific fairness or accuracy constraints.

Our method also makes it easy for users and auditors to understand whether a model in use could be made more fair without a substantial loss of accuracy, or vice versa. This is useful both for improving model performance and for understanding whether a particular level of unfairness can be justified as a type of "business necessity," or whether fairness can be improved without compromising accuracy (Coston et al., 2021).

3.3 Setup and Estimands

Our data is of the form $Z = (A, X, S, D, Y) \sim \mathbb{P}$, for sensitive feature $A \in \{0, 1\}$, additional covariates $X \in \mathcal{X}$, previously trained predictor(s) $S \in \mathcal{S}$, decision $D \in \mathcal{D}$, and outcome or label $Y \in [\ell_y, u_y]$ with known bounds ℓ_y, u_y. If no previously trained predictors are available, then we have $S = \emptyset$. We denote by Y_i^0 the potential outcome $Y_i^{D=0}$, that is, the outcome or label that would be observed for individual i if, possibly contrary to fact, the decision were set to $D_i = 0$. For example, Y^0 could indicate whether an individual would recidivate if released pretrial. We assume that Y^0 also lies in $[\ell_y, u_y]$. In "pure prediction" settings (Kleinberg et al., 2015) where we are

interested in observational rather than counterfactual fairness, we may have $D = \emptyset$. We assume that $Z \subseteq \mathcal{Z} \subset \mathbb{R}^p$ for compact $\mathcal{Z}$.

Let $W = (A, X, S) \in \mathcal{W}$ represent the collected covariates. We let $\widetilde{Y}$ denote either Y and Y^0 as appropriate, since we are interested in both observational and counterfactual fairness and accuracy measures. We refer to $\widetilde{Y} = Y$ as the *observable* setting and $\widetilde{Y} = Y^0$ as the *counterfactual setting*. Broadly speaking, we seek functions of the form $f : \mathcal{W} \mapsto [\ell_y, u_y]$ that are both accurate and fair. Our goals are (1) to enable users to target specific fairness or accuracy constraints, and (2) to trace out the fairness and accuracy properties of a large set of models, both in order to understand setting-specific fairness-accuracy and fairness-fairness tradeoffs and in order to maximize the user's ability to choose a desirable model.

Remark 3 (Additional notation)**.** We let $\|\cdot\|$ denote an appropriate L_2 norm. That is, for any random variable $f(Z)$ taking values in $\mathbb{R}$, $\|f(Z)\| = (\int (f(Z))^2 d\mathbb{P}(Z))^{1/2}$ denotes the $L_2(\mathbb{P})$ norm, while for a non-random vector $v \in \mathbb{R}^k$, $\|v\| = (\sum_{j=1}^k v_j^2)^{1/2}$ denotes the Euclidean L_2 norm. For a random vector $f(Z)$ taking values in $\mathbb{R}^k$, $\|f(Z)\| = (\sum_{j=1}^k \|f(Z)\|^2)^{1/2}$.

3.3.1 Accuracy and fairness measures

The accuracy measure we consider is the MSE, $\mathbb{E}[(f(W) - \widetilde{Y})^2]$. We consider (un)fairness measures that can be expressed in the form

$$\text{unfairness}(f(W)) = |\mathbb{E}[g(W, \widetilde{Y}) f(W)]| \tag{3.1}$$

where $g(W, \widetilde{Y})$ is a bounded *fairness function* that depends only on W and $\widetilde{Y}$. This accommodates a broad range of measures, including measures described by the following proposition. All proofs are given in the Appendix.

Proposition 3. Let $\alpha_0, \alpha_1 \in \mathbb{R}$ and let h_0, h_1 be mappings from $\{0,1\} \times \widetilde{Y}$ to $\{0,1\}$. Let

$$g(W, \widetilde{Y}) = \alpha_0 \frac{h_0(A, \widetilde{Y})}{\mathbb{E}[h_0(A, \widetilde{Y})]} - \alpha_1 \frac{h_1(A, \widetilde{Y})}{\mathbb{E}[h_1(A, \widetilde{Y})]} \tag{3.2}$$

Then

$$\Big|\mathbb{E}[g(W, \widetilde{Y}) f(W)]\Big| = \Big|\alpha_0 \mathbb{E}[f(W) | h_0(A, \widetilde{Y}) = 1] - \alpha_1 \mathbb{E}[f(W) | h_1(A, \widetilde{Y}) = 1]\Big| \tag{3.3}$$

That is, (3.1) is compatible with any fairness measure that can be expressed as a (weighted) difference of average predictions conditioned on events that are a function of the sensitive feature and the outcome. We focus in this paper on the following measures, which we refer to equivalently

as disparities. We first express each in a canonical form, and then we identify the corresponding functions fairness function $g(W, \widetilde{Y})$.

Definition 3.3.1. The *rate difference (rate-diff)* measure is

$$|\mathbb{E}[f(W)|A=0] - \mathbb{E}[f(W)|A=1]| \tag{3.4}$$

with fairness function

$$g^{\text{rate}} = \frac{1-A}{\mathbb{E}[1-A]} - \frac{A}{\mathbb{E}[A]} \tag{3.5}$$

Definition 3.3.2. For $\widetilde{Y} \in \{0,1\}$, the *generalized False Positive Rate Difference (FPR-diff)* measure is

$$\left|\mathbb{E}[f(W)|A=0,\widetilde{Y}=0] - \mathbb{E}[f(W)|A=1,\widetilde{Y}=0]\right| \tag{3.6}$$

with fairness function

$$g^{\text{FPR}} = \frac{(1-\widetilde{Y})(1-A)}{\mathbb{E}[(1-\widetilde{Y})(1-A)]} - \frac{(1-\widetilde{Y})A}{\mathbb{E}[(1-\widetilde{Y})A]} \tag{3.7}$$

Definition 3.3.3. For $\widetilde{Y} \in \{0,1\}$, the *generalized False Negative Rate Difference (FNR-diff)* measure is

$$\left|\mathbb{E}[1-f(W)|A=0,\widetilde{Y}=1] - \mathbb{E}[1-f(W)|A=1,\widetilde{Y}=1]\right| \tag{3.8}$$

with fairness function

$$g^{\text{FNR}} = \frac{\widetilde{Y}A}{\mathbb{E}[\widetilde{Y}A]} - \frac{(1-\widetilde{Y})(1-A)}{\mathbb{E}[\widetilde{Y}(1-A)]} \tag{3.9}$$

These definitions are closely related to common fairness criteria. The criterion of *independence*, also known as statistical parity or demographic parity, requires predictions $f(W)$ to be independent of the sensitive feature A. The rate-diff measure provides a measure of violations of this criterion (Calders and Verwer, 2010). Equal opportunity requires the false negative rates to be equal across the two groups, while equalized odds requires both the false positive and the false negative rates to be equal (Hardt et al., 2016). FNR-diff therefore measures violations of equal opportunity, while FPR-diff and FNR-diff together measure violations of equalized odds.

For continuous-valued predictors, it may be challenging or impossible to attain full (conditional) independence. Hence it is common to focus only on average conditional predictions (Corbett-Davies et al., 2017).

3.3.2 Predictor classes

We consider predictors that lie in the linear span of a set of basis functions $b = b(W) = (b_1(W), \dots b_k(W))$, where each function $b_j(W)$ maps from $\mathcal{W}$ to $\mathbb{R}$. That is, for given b we seek predictors in the set $\mathcal{F}_b$, where

$$\mathcal{F}_b = \{b^T\beta : \beta \in \mathbb{R}^k\} \tag{3.10}$$

We refer to these as "aggregated" predictors (Tsybakov, 2003). The vector b is determined by the user. It can include for example previously trained predictors, newly trained predictors, or arbitrary orthogonal basis functions such as trigonometric functions or polynomials. For the majority of this paper, we consider a regime in which $k < n$, where n is the sample size, since this simplifies estimation. Our approach makes it easy for users to search across a range of different bases b. In our asymptotic analyses, k is allowed to grow with n, but we generally assume that the basis is eventually fixed, i.e. the set $\mathcal{F}_b$ does not change after some n. In practice, users might wish to use bases of dimension $k \geq n$, such as spline bases or kernel basis functions. We consider these and other possibilities in Appendix 3.E.

Depending on b, the set $\mathcal{F}_b$ may be relatively rich. For example, b could be a truncated orthonormal basis of the space $L_2(\mathcal{W})$, in which case $\mathcal{F}_b$ could approximate L_2 to a degree chosen by the user.

Assumption 1. Uniformly in n, the eigenvalues of $\mathbb{E}[bb^T]$ are bounded above and away from 0.

This assumption asserts that the basis functions $b_1(W), \dots b_k(W)$ are not too collinear. It means that $\mathbb{E}[bb^T]$ is always positive semi-definite. In a regime in which the basis is eventually fixed, this assumption simply requires that the basis functions are never perfectly collinear.

Assumption 2. Uniformly in n, $\sup_{w \in \mathcal{W}} \|b(w)\| < \infty$, where $\|b(w)\|$ is the Euclidean L_2 norm.

If the dimension of b does not grow to infinity, then this assumption simply requires the norm $\|b(w)\|$ to be finite over the covariate space.

3.3.3 Estimands

We first define two estimands that are solutions to constrained least squares problems. These estimands represent users who have clear target fairness or accuracy constraints. We then show that these estimands can be equivalently expressed via penalized least squares problems that admit closed-form solutions. These solutions are indexed by an unfairness penalty parameter; by varying this parameter, we may trace out curves in the accuracy-fairness space over $\mathcal{F}_b$.

Suppose that there are t fairness measures that can be expressed via fairness functions $g_j, j = 1, \ldots, t$. For a given k-dimensional basis b, define the risk-minimization (*risk-min*) parameter β_r^* and the unfairness-minimization (*unfair-min*) parameter β_u^* as follows:

$$\beta_r^* = \underset{\beta \in \mathbb{R}^k}{\arg\min}\, \mathbb{E}[(b^T\beta - \widetilde{Y})^2] \tag{3.11}$$

$$\text{subject to } \; (\mathbb{E}[g_j b^T \beta])^2 \leq \epsilon_j^2, \quad j = 1, \ldots t \tag{3.12}$$

$$\beta_u^* = \underset{\beta \in \mathbb{R}^k}{\arg\min} \sum_{j=1}^{t} \alpha_j (\mathbb{E}[g_j b^T \beta])^2 \tag{3.13}$$

$$\text{subject to } \mathbb{E}[(b^T\beta - \widetilde{Y})^2] \leq \epsilon \tag{3.14}$$

for user-chosen constraints $\epsilon_j \geq 0$, $\epsilon > 0$, and weights α_j. That is, β_r^* indexes the most accurate predictor in $\mathcal{F}_b$ among those that satisfy t specified fairness constraints, and β_u^* indexes the most fair predictor among that satisfy a specified risk constraint. We constrain $\epsilon > 0$ because otherwise we'd be insisting on a risk-free predictor, which is generally impossible; we allow $\epsilon_j = 0$ because for many fairness criteria it is possible to achieve an exactly fair predictor. For example, a constant predictor will always have a rate difference, FPR difference, and FNR difference of 0.

The risk-minimization problem is always feasible, since the predictor defined by $\beta = 0$ always satisfies the fairness constraints. Under Assumption 1, β_r^* is unique, since the objective is strictly convex. The unfairness-minimization problem may be infeasible, if there is no predictor in $\mathcal{F}_b$ whose risk is less than or equal to ϵ. This may not be an issue in practice, if ϵ represents (an estimate of) the risk of an existing benchmark model. With slight modifications, all our subsequent results would carry through if this constraint were explicitly expressed with respect to a benchmark model; for the sake of simplicity, however, we leave it in this form. If the unfairness-minimization problem is feasible and $\sum_{j=1}^{t} \alpha_j \mathbb{E}[g_j b]\mathbb{E}[g_j b]^T$ is positive definite, then β_u^* is unique, since the objective is strictly convex.

Note that the fairness constraints in the risk-minimization problem can be equivalently written in linear form, as $\left|\mathbb{E}[g_j f(W)]\right| \leq \epsilon_j$. We choose the squared form for notational consistency with the penalized form, which is defined as follows.

For any $\lambda = (\lambda_1, \ldots, \lambda_r)$, with all $\lambda_j \geq 0$, define the penalized-minimization (*penalized-min*) estimand β^*_λ as

$$\beta^*_\lambda = \underset{\beta \in \mathbb{R}^k}{\arg\min}\, \mathbb{E}[(b^T\beta - \widetilde{Y})^2] + \sum_{j=1}^{t} \lambda_j \left(\mathbb{E}[g_j b^T \beta]\right)^2 \tag{3.15}$$

This can be written in an equivalent closed form:

$$\beta^*_\lambda = \left(\mathbb{E}[bb^T] + \sum_{j=1}^{t} \lambda_j \mathbb{E}[g_j b]\mathbb{E}[g_j b]^T\right)^{-1} \mathbb{E}[\widetilde{Y} b] \tag{3.16}$$

Under Assumption 1, by Weyl's inequality the matrix inverse always exists, since each matrix $\mathbb{E}[g_j b]\mathbb{E}[g_j b]^T$ is positive semi-definite.

We now establish a correspondence between the constrained and penalized forms. Let $\mathcal{I}$ denote the set of active fairness constraints at β^*_r, that is, $\mathcal{I} = \{j \in \{1, \ldots, t\} : (\mathbb{E}[g_j b^T \beta^*_r])^2 = \epsilon_j^2\}$.

Assumption 3. The set of vectors $\{\mathbb{E}[g_j b]\mathbb{E}[g_j b]^T \beta^*_r : j \in \mathcal{I}\}$ is linearly independent.

Assumption 3, which is expected to hold for any basis and set of fairness definitions that would be used in practice, is the Linear Independence Constraint Qualification, which yields a mapping from the constrained to the penalized forms (Wachsmuth, 2013).

Proposition 4. Under Assumptions 1 and 3, for any β^*_r there exists a unique $\lambda \in \mathbb{R}^t_{0+}$ such that $\beta^*_\lambda = \beta^*_r$.

We will utilize the penalized form to efficiently construct a large set of predictors that vary in their accuracy and fairness properties. We will see that we can exploit an empirical analogue of Proposition 4 to "seed" this set with models that target specified fairness constraints.

Proposition 5. Fix $\lambda \in \mathbb{R}^t_{0+}$. Under Assumption 1, $\beta^*_\lambda = \beta^*_r$ with fairness constraints $\epsilon_j^2 = (\mathbb{E}[g_j b^T \beta^*_\lambda])^2$.

Proposition 5 expresses the converse direction of the relationship between β^*_r and β^*_λ. This will facilitate interpretation of the corresponding penalized estimators.

An analogous penalized form can be written that corresponds to β^*_u, with results that match Propositions 4 and 5. For estimation purposes, however, we will only be interested in β^*_λ, so we do not develop that here.

Remark 4. Any $\beta \in \mathbb{R}^k$ indexes a predictor $b^T\beta \in \mathcal{F}_b$. Since Y^0 and Y are bounded in $[\ell_y, u_y]$, however, the resulting predictor will be defined by

$$f_\beta = [b^T\beta]_{\ell_y}^{u_y} = \begin{cases} \ell_y, & b^T\beta < -\ell_y \\ b^T\beta, & b^T\beta \in [\ell_y, u_y] \\ u_y, & b^T\widehat{\beta} > u_y \end{cases} \tag{3.17}$$

assuming that the bounds ℓ_y, u_y are known.

Our final estimands consist of the risk and (un)fairness properties of any fixed predictor f_β:

$$\text{Risk}(f_\beta) = \mathbb{E}[(f_\beta - \widetilde{Y})^2] \tag{3.18}$$

$$\text{UF}_j(f_\beta) = \mathbb{E}[g_j f_\beta], \quad j = 1, \dots t \tag{3.19}$$

In particular, once we have computed some estimate $\widehat{\beta}$ of β_r^*, β_u^*, or β_λ^*, it is of interest to estimate the risk and fairness of the resulting predictor $f_{\widehat{\beta}}$.

3.4 Identification

When $\widetilde{Y} = Y^0$, i.e. when the risk and/or fairness functions are defined with respect to potential rather than observable outcomes, we require assumptions in order to identify these quantities in terms of the observed data. For ease of notation, we first define three nuisance parameters that appear in the estimands and associated estimators.

$$\pi = \pi(W) = \mathbb{P}(D = 1 \mid W)$$
$$\mu_0 = \mu_0(W) = \mathbb{E}[Y \mid W, D = 0]$$
$$\nu_0 = \nu_0(W) = \mathbb{E}[Y^2 \mid W, D = 0]$$

$\pi(W)$ is the propensity score, while μ_0 and ν_0 are regressions with respect to the observed outcome and the squared observed outcome. In a classification setting with $Y \in \{0, 1\}$, we have $Y^2 = Y$, so $\nu_0 = \mu_0$. We make the following standard "no unmeasured confounding"-type causal inference assumptions:

Assumption 4 (Consistency)**.** $Y = DY^1 + (1 - D)Y^0$

Assumption 5 (Positivity)**.** $\exists \delta \in (0, 1)$ s.t. $\mathbb{P}(\pi(W) \leq 1 - \delta) = 1$

Assumption 6 (Ignorability)**.** $Y^0 \perp\!\!\!\perp D \mid W$

We also define the following for convenience:

$$\phi = \phi(Z) = \frac{1-D}{1-\pi}(Y - \mu_0) + \mu_0 \tag{3.20}$$

$$\underline{\phi} = \underline{\phi}(Z) = \frac{1-D}{1-\pi}(Y^2 - \nu_0) + \nu_0 \tag{3.21}$$

Under the identifying assumptions, these are the uncentered influence functions for $\mathbb{E}[Y^0]$ and $\mathbb{E}[(Y^0)^2]$, respectively.

Proposition 6. Under the identifying assumptions, the counterfactual risk, FPR-diff, and FNR-diff for any function $f : \mathcal{W} \mapsto \mathbb{R}$ are identified as follows. We provide two expressions for the risk, which are used in the *risk-min* and *unfair-min* estimators, respectively. We do not include the rate-diff, since this does not involve outcomes and is therefore trivially identified.

$$\mathbb{E}[(f - Y^0)^2] = \mathbb{E}[(f - \mu_0)^2] + \text{var}(Y^0) \tag{3.22}$$

$$= \mathbb{E}[f^2 - 2f\mu_0 + \nu_0] \tag{3.23}$$

$$\mathbb{E}[g^{\text{cFPR}} f(W)] = \mathbb{E}\left[\left\{\frac{(1-\mu_0)(1-A)}{\mathbb{E}[(1-\mu_0)(1-A)]} - \frac{(1-\mu_0)A}{\mathbb{E}[(1-\mu_0)A]}\right\} f(W)\right] \tag{3.24}$$

$$\mathbb{E}[g^{\text{cFNR}} f(W)] = \mathbb{E}\left[\left\{\frac{\mu_0 A}{\mathbb{E}[\mu_0 A]} - \frac{(1-\mu_0(1-A)}{\mathbb{E}[\mu_0(1-A)]}\right\} f(W)\right] \tag{3.25}$$

These expressions also hold if μ_0 is replaced with ϕ and ν_0 is replaced with $\underline{\phi}$.

Remark 5. Expressions (3.22) and (3.23) show that when estimating the *risk-min* parameter β_r^*, we can either minimize an estimate of $\mathbb{E}[(f-\mu_0)^2]$ or an estimate of $\mathbb{E}[f^2 - 2f\mu_0]$; the terms $\text{var}(Y^0)$ and ν_0 are constant with respect to f and so drop out of the minimization. The nuisance parameter ν_0 will only be required when we wish to estimate the actual risk of a given predictor, as well as when estimating the *unfair-min* parameter $\widehat{\beta}_u$, since that involves a constraint on the actual risk. Note that ν_0 would also not be required to solve *unfair-min* if the accuracy constraint were defined with respect to an existing benchmark model, since the two ν_0 terms in the constraint would cancel out.

3.5 Estimation

We require a training set $\mathcal{D}_{\text{train}}$, which is used to construct estimates $\widehat{\beta}$, and a test set $\mathcal{D}_{\text{test}}$, which is used to estimate the risk and fairness values of the resulting predictor(s) $f_{\widehat{\beta}}$. If the user wishes to train new basis predictors, then an additional dataset $\mathcal{D}_{\text{learn}}$ is also required. This is not needed

if the user is only aggregating arbitrary basis functions, like trigonometric functions, or previously existing predictors.

In order to obtain fast rates for our estimators, in the counterfactual setting we split $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ into separate folds for estimating the nuisance parameters and the target parameters. The sample splitting scheme is shown in Figure 3.1. For simplicity, we illustrate a single split, but in practice cross-fitting can be used within each dataset.

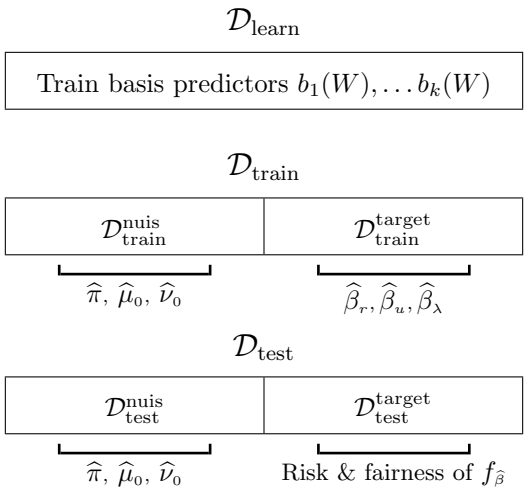


Figure 3.1: Sample splitting scheme. $\mathcal{D}_{\text{learn}}$ is not needed if the basis functions already exist. Splitting $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ is only required in the counterfactual setting, since there are no nuisance parameters in the observable setting. In practice, cross-fitting may be used within both $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$.

We solve empirical versions of the identified minimization problems that define the estimands. Let $\widehat{\phi}, \widehat{\underline{\phi}}$ denote estimates of ϕ and $\underline{\phi}$ constructed from estimates $\widehat{\pi}$, $\widehat{\mu}_0$, $\widehat{\nu}_0$.

For any fixed function $f : \mathcal{Z} \mapsto \mathbb{R}$, let $\mathbb{P}_n(f(Z)) = n^{-1} \sum_{i=1}^n f(Z)$ and $\mathbb{P}(f) = \int f d\mathbb{P}(Z)$ denote the sample and true expectations of f, so that for example $\mathbb{P}(\phi) = \mathbb{E}[\phi]$ while $\mathbb{P}(\widehat{\phi}) = \mathbb{E}[\widehat{\phi}|\mathcal{D}_{\text{train}}]$ or $\mathbb{E}[\widehat{\phi}|\mathcal{D}_{\text{test}}]$ is the expected value of $\widehat{\phi}(Z)$ once the relevant nuisance function estimate $\widehat{\phi}$ has been constructed. That is, $\mathbb{P}(\widehat{\phi})$ is a random variable that depends on the nuisance data, while $\mathbb{P}_n(\widehat{\phi})$ is a random variable that depends on both the nuisance and target data. We will rely on context to make it clear whether $\mathcal{D}_{\text{train}}$ or $\mathcal{D}_{\text{test}}$ is under consideration, with explicit clarification where necessary.

For notational convenience, let $\widehat{g}_j$ with no arguments denote $g_j(W, Y)$ in the observable setting and $g_j(W, \widehat{\phi})$ in the counterfactual setting. That is $\widehat{g}_j = g_j$ in the observable setting, since there is no nuisance quantity to estimate, but $\widehat{g}_j \neq g_j$ in the counterfactual setting. The occasional use of $\widehat{g}_j$ for both settings allows us to concisely state certain conditions and results.

Assumption 7 (Bounded propensity estimator)**.** $\exists \gamma \in (0, 1)$ s.t. $\mathbb{P}(\widehat{\pi}(A, X, S) \leq 1 - \gamma) = 1$

Assumption 7 is the empirical analogue of the positivity assumption (5). It can be trivially satisfied by truncating $\widehat{\pi}$ at $1-\delta$, the positivity threshold in Assumption 5.

Assumption 8 (Consistent nuisance estimators)**.** $\|\widehat{\pi}-\pi\| = o_{\mathbb{P}}(1)$ and $\|\widehat{\mu}_0-\mu_0\| = o_{\mathbb{P}}(1)$ and $\|\widehat{\nu}_0-\nu_0\| = o_{\mathbb{P}}(1)$.

This assumption is reasonable if nonparametric methods are used to construct the nuisance parameter estimates. With slight procedural modifications, this assumption can be relaxed to instead require consistency in the influence function estimators $\widehat{\phi}$ and $\widehat{\underline{\phi}}$; for simplicity, we do not address this.

3.5.1 Constrained least squares

The *risk-min* and *unfair-min* estimators $\widehat{\beta}_r$ and $\widehat{\beta}_u$ are defined for the observable and counterfactual settings in Tables 3.1 and 3.2.

Observable ($\widetilde{Y}=Y$)	Counterfactual ($\widetilde{Y}=Y^0$)
$\widehat{\beta}_r = \underset{\beta\in\mathbb{R}^k}{\arg\min}\, \mathbb{P}_n[(b^T\beta - Y)^2]$ s.t. $\left(\mathbb{P}_n[g_j(W,Y)b^T\beta]\right)^2 \le \epsilon_j^2,\ j=1,\dots t$	$\widehat{\beta}_r = \underset{\beta\in\mathbb{R}^k}{\arg\min}\, \mathbb{P}_n[(b^T\beta - \widehat{\phi})^2]$ s.t. $\left(\mathbb{P}_n[g_j(W,\widehat{\phi})b^T\beta]\right)^2 \le \epsilon_j^2,\ j=1,\dots t$

Table 3.1: Definition of the *unfair-min* estimator $\widehat{\beta}_r$ in the observable and counterfactual settings.

Observable ($\widetilde{Y}=Y$)	Counterfactual ($\widetilde{Y}=Y^0$)
$\widehat{\beta}_u = \underset{\beta\in\mathbb{R}^k}{\arg\min} \sum_{j=1}^{t} \alpha_j \left(\mathbb{P}_n[g_j(W,Y)b^T\beta]\right)^2$ s.t. $\mathbb{P}_n\left[(b^T\beta - Y)^2\right] \le \epsilon^2$	$\widehat{\beta}_u = \underset{\beta\in\mathbb{R}^k}{\arg\min} \sum_{j=1}^{t} \alpha_j \left(\mathbb{P}_n[g_j(W,\widehat{\phi})b^T\beta]\right)^2$ s.t. $\mathbb{P}_n\left[(b^T\beta)^2 - 2(b^T\beta)\widehat{\phi} + \widehat{\underline{\phi}}\right] \le \epsilon^2$

Table 3.2: Definition of the *risk-min* estimator $\widehat{\beta}_u$ in the observable and counterfactual settings.

As with the corresponding estimands, the optimization problem that defines $\widehat{\beta}_r$ is always feasible, while the problem that defines $\widehat{\beta}_u$ may not be, if there is no predictor in $\mathcal{F}_b$ with estimated risk less than or equal to ϵ. If $\mathbb{P}_n[bb^T]$ is positive definite, then $\widehat{\beta}_r$ is unique, since the objective is strictly

convex. Under Assumption 1, this will hold with probability approaching 1 in n, or with probability 1 if, say, one of the covariates in W is continuously distributed. We next consider the excess risk and the excess unfairness for the constrained predictors. In the counterfactual setting, we require assumptions on the rate at which the nuisance parameters are estimated.

Assumption 9 (Nuisance parameter rates)**.**

$$\|\widehat{\pi} - \pi\|\|\widehat{\mu}_0 - \mu_0\| = o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.26}$$

$$\|\widehat{\pi} - \pi\|\|\widehat{\nu}_0 - \nu_0\| = o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.27}$$

Definition 3.5.1. The *excess risk* for $\widehat{\beta}_r$ and $\widehat{\beta}_u$ is defined as:

$$\mathbb{P}[(b^T\widehat{\beta}_r - \widetilde{Y})^2] - \mathbb{P}[(b^T\beta_r^* - \widetilde{Y})^2] \tag{risk-min}$$

$$\mathbb{P}[(b^T\widehat{\beta}_u - \widetilde{Y})^2] - \epsilon^2 \tag{unfair-min}$$

Theorem 3.1 (Excess risk in the constrained setting)**.** *Under Assumptions 1–2 for the observable setting, and Assumptions 1–2 and 4–9 for the counterfactual setting:*

$$\mathbb{P}[(b^T\widehat{\beta}_r - \widetilde{Y})^2] - \mathbb{P}[(b^T\beta_r^* - \widetilde{Y})^2] = O_{\mathbb{P}}(1/\sqrt{n}) \tag{risk-min}$$

$$\mathbb{P}[(b^T\widehat{\beta}_u - \widetilde{Y})^2] - \epsilon^2 = O_{\mathbb{P}}(1/\sqrt{n}) \tag{unfair-min}$$

Definition 3.5.2. The *excess unfairness* for $\widehat{\beta}_r$ and $\widehat{\beta}_u$ is defined as:

$$\max_{j=1,\ldots,t}\left\{\left((\mathbb{P}[g_j b^T\widehat{\beta}_r])^2 - \epsilon_j^2\right)_+\right\} \tag{risk-min}$$

$$\sum_{j=1}^{t}\alpha_j\left\{(\mathbb{P}[g_j b^T\widehat{\beta}_u])^2 - (\mathbb{P}[g_j b^T\beta_u^*])^2\right\} \tag{unfair-min}$$

where $(x)_+ = \max\{x, 0\}$ denotes the positive part function.

Theorem 3.2 (Excess unfairness in the constrained setting)**.** *Under Assumptions 1–2 for the observable setting, and Assumptions 1–2 and 4–9 for the counterfactual setting:*

$$\max_{j=1,\ldots,t}\left\{\left((\mathbb{P}[g_j b^T\widehat{\beta}_r])^2 - \epsilon_j^2\right)_+\right\} = O_{\mathbb{P}}(1/\sqrt{n}) \tag{risk-min}$$

$$\sum_{j=1}^{t}\alpha_j\left\{(\mathbb{P}[g_j b^T\widehat{\beta}_u])^2 - (\mathbb{P}[g_j b^T\beta_u^*])^2\right\} = O_{\mathbb{P}}(1/\sqrt{n}) \tag{unfair-min}$$

As described in Section 3.3, the basis is allowed to grow with n, as long as it is eventually fixed. These results show that if a user has specific fairness or risk constraints in mind, in the observable setting, they can generate a predictor in an arbitrarily rich linear space that is asymptotically guaranteed to meet these constraints, while minimizing the corresponding risk or unfairness. In the counterfactual setting, they can do the same thing as long as the nuisance parameters are estimated at fast enough rates.

Of course, any particular estimates $\widehat{\beta}_r, \widehat{\beta}_u$ may violate their target risk and fairness constraints by arbitrary amounts. Suppose that $\widehat{\beta}_r$ was evaluated on the test set, and one of its estimated unfairness values was found to exceed the constraint ϵ_j by an unacceptable amount. To remedy this, the user could lower the value of ϵ_j and compute a new $\widehat{\beta}_r$ under this more stringent constraint. They could repeat this process until they found a $\widehat{\beta}_r$ with acceptable estimated fairness. Since $\widehat{\beta}_r$ is the solution to a quadratic program, however, this is computationally costly, and there is no guarantee that additional searching will yield improvements. A predictor that is more fair with respect to one fairness constraint may be more *unfair* with respect to other constraints, or may incur unacceptable additional risk. Ideally, the user might wish to treat the fairness constraints as tuning parameters, selecting a large set of constraint vectors $(\epsilon_1, \ldots, \epsilon_t) \in \mathbb{R}^t_{0+}$, computing $\widehat{\beta}_r$ for each vector, and comparing the risk and fairness properties of all the resulting predictors. This is computationally costly, however.

In the next section, we use the closed-form penalized estimators to accomplish something equivalent to this, with trivial computational cost.

3.5.2 Penalized least squares

In both the observable and counterfactual settings, the estimator $\widehat{\beta}_\lambda$ takes the following equivalent forms, which mirror the two expressions given for β^*_λ:

$$\begin{aligned}
\widehat{\beta}_\lambda &= \underset{\beta \in \mathbb{R}^k}{\arg\min}\, \mathbb{P}_n[(b^T\beta - Y)^2] + \sum_{j=1}^{t} \lambda_j \left(\mathbb{P}_n[g_j b^T \beta]\right)^2 && \text{(Observable)} \\
&= \left(\mathbb{P}_n(bb^T) + \sum_{j=1}^{t} \lambda_j \mathbb{P}_n(g_j b)\mathbb{P}_n(g_j b)^T\right)^{-1} \mathbb{P}_n(bY) \\
\widehat{\beta}_\lambda &= \underset{\beta \in \mathbb{R}^k}{\arg\min}\, \mathbb{P}_n[(b^T\beta - \widehat{\phi})^2] + \sum_{j=1}^{t} \lambda_j \left(\mathbb{P}_n[\widehat{g}_j b^T \beta]\right)^2 && \text{(Counterfactual)} \\
&= \left(\mathbb{P}_n(bb^T) + \sum_{j=1}^{t} \lambda_j \mathbb{P}_n(\widehat{g}_j b)\mathbb{P}_n(\widehat{g}_j b)^T\right)^{-1} \mathbb{P}_n(b\widehat{\phi})
\end{aligned}$$

assuming that the relevant matrix inverse exists. A sufficient condition for it to exist is that $\mathbb{P}_n[bb^T]$ is positive definite, which, as discussed above, will happen with probability 1 or approaching 1 under Assumption 1.

The procedure we propose is given in Figure 3.2. The user first chooses a large set of vectors Λ_n, which we assume may depend on sample size. They then compute the solution set $\widehat{\mathcal{B}}_n = \{\widehat{\beta}_\lambda : \lambda \in \Lambda_n\}$, estimate the risk and fairness properties of each $f_\beta : \beta \in \widehat{\mathcal{B}}_n$, and select a predictor with a favorable performance profile.

1. Pick a large set of vectors $\Lambda_n \subset \mathbb{R}^t_{0+}$.
2. Compute the solution set $\widehat{\mathcal{B}}_n = \{\widehat{\beta}_\lambda : \lambda \in \Lambda_n\}$.
3. Compute the estimated risk and fairness properties of each $f_\beta : \beta \in \widehat{\mathcal{B}}_n$.
4. Select a predictor f_β with favorable risk and fairness properties.

Figure 3.2: Estimation procedure in the penalized setting.

Propositions 4 and 5 established a correspondence between the constrained and penalized estimands, so each $\widehat{\beta}_\lambda$ may be regarded either as an estimate of the penalized-minimizer β^*_λ or as an estimate of some risk-minimizer β^*_r. The value of the penalized perspective is that Step 2 in this procedure can be carried out extremely efficiently. Since each matrix $\mathbb{P}_n(\widehat{g}_j b)\mathbb{P}_n(\widehat{g}_j b)^T$ has rank 1, the overall matrix inverse can be computed by computing $\mathbb{P}_n(bb^T)^{-1}$ and then applying a series of simple algebraic operations, per the Sherman-Morrison update formula. This is expressed in the following proposition.

Proposition 7. Let

$$\overline{\lambda}_j = (\lambda_1, \ldots, \lambda_j), \text{ so that } \overline{\lambda}_t = \lambda \tag{3.28}$$

$$m_j = \mathbb{P}_n(\widehat{g}_j b) \tag{3.29}$$

$$\widehat{\mathbf{Q}}_0 = \mathbb{P}_n(bb^T)^{-1} \tag{3.30}$$

$$\widehat{\mathbf{Q}}_1(\lambda_1) = \widehat{\mathbf{Q}}_0 - \frac{\lambda_1 \widehat{\mathbf{Q}}_0 m_1 m_1^T \widehat{\mathbf{Q}}_0}{1 + \lambda_1 m_1^T \widehat{\mathbf{Q}}_0 m_1} \tag{3.31}$$

$$\widehat{\mathbf{Q}}_j(\overline{\lambda}_j) = \widehat{\mathbf{Q}}_{j-1}(\overline{\lambda}_{j-1}) - \frac{\lambda_j \widehat{\mathbf{Q}}_{j-1}(\overline{\lambda}_{j-1}) m_j m_j^T \widehat{\mathbf{Q}}_{j-1}(\overline{\lambda}_{j-1})}{1 + \lambda_j m_j^T \widehat{\mathbf{Q}}_{j-1}(\overline{\lambda}_{j-1}) m_j}, \quad \text{for } j = 2, \ldots, t \tag{3.32}$$

Then

$$\widehat{\beta}_\lambda = \begin{cases} \widehat{\mathbf{Q}}_t(\lambda_t)^{-1}\mathbb{P}_n(b\widehat{\phi}) & \text{(Counterfactual)} \\ \widehat{\mathbf{Q}}_t(\lambda_t)^{-1}\mathbb{P}_n(bY) & \text{(Observable)} \end{cases} \tag{3.33}$$

Proposition 7 says that to compute the set $\widehat{\mathcal{B}}_n$ requires only a single matrix inversion, to compute $\widehat{\mathbf{Q}}_0$. Each vector m_j also only needs to be computed once. The remaining operations are algebraic. In particular, define an *algebraic update* as the computation of a matrix $\widehat{\mathbf{Q}}_j(\overline{\lambda}_j)$ for some $j \in \{1, \ldots, t\}$, conditional on the quantity $\widehat{\mathbf{Q}}_{j-1}$ having already been computed. We have the following corollary.

Corollary 3.2.1.

Since $\widehat{\mathbf{Q}}_0$ is a $k \times k$ matrix and each m_j is a vector of length k, if b is a relatively small basis, then $\widehat{\mathbf{Q}}_0$ will be fast to compute, and all the algebraic updates will be fast. In our simulations and real data analyses, we show that we can get good results with a very small number of basis functions (e.g. 4 to 6), which yield extremely fast computations.

How should Λ_n be chosen in Step 1? Since $\Lambda_n \subset \mathbb{R}^t_{0+}$, one simple possibility is to take a one-dimensional grid of points between 0 and some arbitrary large number and then construct the $t-$dimensional Cartesian product. Since $\widehat{\beta}_\lambda$ is smooth in λ, and since the risk and fairness measures are smooth in $\widehat{\beta}$, we can expect that such a grid will enable us to move smoothly around the fairness-accuracy space, and that we won't be missing desirable predictors that lie in between the grid points*.

Another possibility is to "seed" Λ_n with values that correspond to a particular $\widehat{\beta}_r$. That is, fix some constraints ϵ_j and compute the corresponding risk-min estimator $\widehat{\beta}_r$. Since the constraints are affine, Slater's condition holds, so the duality gap is 0. The associated λ can therefore be computed by solving the dual of the problem that defines $\widehat{\beta}_r$, and Λ_n can then be constructed as a grid around this λ.

The constraints ϵ_j that define $\widehat{\beta}_r$ are arguably easier to reason about than the penalties λ_j that define $\widehat{\beta}_\lambda$. This "seeding" approach provides a way to ensure that the set $\{\widehat{\beta}_\lambda : \lambda \in \Lambda_n\}$ includes estimators that in some sense target reasonable constraints, particularly for users with specific constraints in mind. This approach requires solving just a single constrained optimization problem, to establish a point of reference in fairness-accuracy space.

The correspondence between the constrained estimand β^*_r and the penalized estimand β^*_λ holds also between $\widehat{\beta}_r$ and $\widehat{\beta}_\lambda$, as expressed in the next two propositions, which are empirical versions of

*The movement won't be entirely smooth if predictions are truncated to lie in $[\ell_y, u_y]$.

Propositions 4 and 5. Let $\mathcal{I}_n$ denote the set of active fairness constraints at $\widehat{\beta}_r$, that is, $\mathcal{I}_n = \{j \in \{1, \ldots, t\} : (\mathbb{P}_n[\widehat{g}_j b^T \widehat{\beta}_r])^2 = \epsilon_j^2\}$, where $\widehat{g}_j$ denotes $g_j(W, Y)$ or $g_j(W, \widehat{\phi})$ as appropriate.

Assumption 10. The set of vectors $\{\mathbb{P}_n[\widehat{g}_j b]\mathbb{P}_n[\widehat{g}_j b]^T \widehat{\beta}_r : j \in \mathcal{I}_n\}$ is linearly independent.

Assumption 10 is expected to hold with probability 1 for any realistic combination of data generating process, basis, and fairness functions.

Proposition 8. If $\widehat{\beta}_r$ exists, then under assumption 10, there exists a unique $\lambda \in \mathbb{R}^t_{0+}$ such that $\widehat{\beta}_\lambda = \widehat{\beta}_r$.

Proposition 9. Fix $\lambda \in \mathbb{R}^t_{0+}$. If $\widehat{\beta}_\lambda$ exists, then $\widehat{\beta}_\lambda = \widehat{\beta}_r$ with fairness constraint value $\epsilon_j^2 = (\mathbb{P}_n[\widehat{g}_j b^T \widehat{\beta}_\lambda])^2$.

The procedure we have described allows users to efficiently construct and evaluate a very large set of models that fall in different points in the fairness-accuracy space. In sections 3.6, 3.7, and 3.8, we show that this procedure enables us to find high-performing models in both observable and counterfactual settings, with simulated and real data. With minimal searching over possible bases, we are able to find models that substantially outperform existing models and methods with respect to both fairness and accuracy.

Remark 6. Since β^*_λ is constructed as a penalized equivalent of β^*_r, the seeding approach to constructing Λ_n that we have described allows users to target particular fairness constraints but not particular risk constraints. It is straightforward to develop an analogous procedure around a penalized version of β^*_u that allows users to seed Λ_n with estimators that target particular risk constraints. In practice, it is not likely to matter much, since the construction of $\widehat{\beta}_\lambda$ should allow users to flexibly explore the fairness-accuracy space and find an estimator that accommodates their desired constraints, if one exists in the span of the chosen basis.

Remark 7. An even simpler and plausibly just as effective alternative to computing the collection $\{\widehat{\beta}_\lambda : \lambda \in \Lambda_n\}$ is to simply define an arbitrary set $\mathcal{B} \subset \mathbb{R}^k$, perhaps constrained to lie in the simplex or in an L_1 box around the origin. That is, the user could simply evaluate arbitrary sets of basis weights to see if any of them yields a reasonable predictor. This set could be similarly constructed as a grid around a particular $\widehat{\beta}_r$ or $\widehat{\beta}_u$, if users have specific fairness or accuracy constraints they wish to target.

Remark 8. $\widehat{\beta}_\lambda$ resembles a ridge regression estimator. In ridge regression and other regularized estimators, however, the penalty tuning parameter λ is expected to go to 0 as $n \to \infty$. In our setting, λ serves to enforce fairness rather than to modulate the variance-bias tradeoff, so there is

no reason for it to shrink with n. Without unfairness penalties, the predictor won't automatically get more fair as the data gets larger.

Theoretical results: excess risk and unfairness

We now develop theoretical guarantees for our procedure. Let $h(n)$ denote the rate at which the product $\|\widehat{\pi} - \pi\|\|\widehat{\mu}_0 - \mu_0\|$ grows or converges, and let $\underline{h}(n)$ denote the rate for $\|\widehat{\pi} - \pi\|\|\widehat{\nu}_0 - \nu_0\|$. That is,

$$\|\widehat{\pi} - \pi\|\|\widehat{\mu}_0 - \mu_0\| = O_{\mathbb{P}}(h(n)) \tag{3.34}$$

$$\|\widehat{\pi} - \pi\|\|\widehat{\nu}_0 - \nu_0\| = O_{\mathbb{P}}(\underline{h}(n)) \tag{3.35}$$

Under ideal conditions, Assumption 9 will hold, so that the product of nuisance parameter errors decay faster than $1/\sqrt{n}$, but the subsequent results do not require this.

Assumption 11 (Compact superset Λ)**.** For all n, $\Lambda_n \subseteq \Lambda \subset \mathbb{R}^t$ for some compact set Λ.

Definition 3.5.3. For any $\lambda \in \mathbb{R}^t_{0+}$, the *excess risk* for $\widehat{\beta}_\lambda$ is

$$\mathbb{P}[(b^T\widehat{\beta}_\lambda - \widetilde{Y})^2] - \mathbb{P}[(b^T\beta^*_\lambda - \widetilde{Y})^2] \tag{3.36}$$

Theorem 3.3 (Uniform rate for excess risk in the penalized setting)**.** *Under Assumptions 1–2 for the observable setting; and Assumptions 1–2, 4–9, and 11 for the counterfactual setting:*

$$\sup_{\lambda\in\Lambda}\left\{\mathbb{P}\left[\left(b^T\widehat{\beta}_\lambda - \widetilde{Y}\right)^2\right] - \mathbb{P}\left[\left(b^T\beta^*_\lambda - \widetilde{Y}\right)^2\right]\right\} = O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}(h(n)) \tag{3.37}$$

In other words, the excess risk goes to 0 uniformly at $\sqrt{1/n}$ or the nuisance rate $h(n)$, whichever is slower. We have a similar result for the excess unfairness, which is defined as follows.

Definition 3.5.4. For any $\lambda \in \mathbb{R}^t_{0+}$, the *excess unfairness* for $\widehat{\beta}_\lambda$ is

$$\left\{\max_{j\in 1,\ldots,t}\left(\mathbb{P}\left[g_j b^T\widehat{\beta}_\lambda\right] - \mathbb{P}\left[g_j b^T\beta^*_\lambda\right]\right)\right\} \tag{3.38}$$

We have defined excess unfairness as the max over j, but it makes little difference if we define it instead as the sum over j. Note that here we haven't used the squared unfairness.

Theorem 3.4 (Uniform rate for excess unfairness in the penalized setting)**.** *Under Assumptions 1–2 for the observable setting; and Assumptions 1–2, 4–9, and 11 for the counterfactual setting:*

$$\sup_{\lambda\in\Lambda}\left\{\max_{j\in 1,\ldots,t}\left(\mathbb{P}\left[g_j b^T\widehat{\beta}_\lambda\right]-\mathbb{P}\left[g_j b^T\beta^*_\lambda\right]\right)\right\}=O_{\mathbb{P}}(\sqrt{1/n})+O_{\mathbb{P}}(h(n)) \tag{3.39}$$

Remark 9. We can obtain similar theoretical results in a regime in which k is allowed to grow to ∞, if we require that $\sup_{w\in\mathcal{W}}\|b(w)\|=O(\sqrt{k})$ and that $k\log(k)/n\to 0$. The first requirement is a stronger version of Assumption 2, while the second insists that k not grow too fast in n. Under these additional requirements, we attain a rate of $O_{\mathbb{P}}(\sqrt{k/n})+O_{\mathbb{P}}(\sqrt{k}\cdot h(n))$ in Theorems 3.3 and 3.3. These results extend the results of Belloni et al. (2015) to a setting with nuisance parameters and penalty terms. As illustrated in that paper, these requirements are weak enough to allow the basis to asymptotically span rich function spaces such as the space of square integrable functions.

3.5.3 Risk and unfairness of a fixed predictor

The risk and unfairness of a fixed predictor f_β are estimated as

$$\widehat{\text{Risk}}(f_\beta)=\begin{cases}\mathbb{P}_n[(f_\beta-Y)^2] & \text{(Observable)}\\ \mathbb{P}_n[f_\beta^2-2f_\beta\widehat{\phi}+\widehat{\underline{\phi}}] & \text{(Counterfactual)}\end{cases} \tag{3.40}$$

$$\widehat{\text{UF}}_j(f_\beta)=\begin{cases}\mathbb{P}_n[g_j(W,Y)f_\beta] & \text{(Observable)}\\ \mathbb{P}_n[g_j(W,\widehat{\phi})f_\beta] & \text{(Counterfactual)}\end{cases} \tag{3.41}$$

for $j=1,\ldots t$.

Theorem 3.5 (Asymptotic normality of risk and unfairness estimators)**.** *Consider fairness functions $g_j\in\{g^{rate},g^{FPR},g^{FNR}\}$. Under Assumptions 1–2 for the observable setting; and Assumptions 1–2, 4–9, and 11 for the counterfactual setting:*

$$\sqrt{n}\left(\widehat{\text{Risk}}(f_\beta)-\text{Risk}(f_\beta)\right)\xrightarrow{d}\begin{cases}N\left(0,\text{var}((f_\beta-Y)^2)\right) & \textit{(Observable)}\\ N\left(0,\text{var}(f_\beta^2-2f_\beta\phi+\underline{\phi})\right) & \textit{(Counterfactual)}\end{cases} \tag{3.42}$$

$$\sqrt{n}\left(\widehat{\text{UF}}_j(f_\beta)-\text{UF}_j(f_\beta)\right)\xrightarrow{d}\begin{cases}N\left(0,\text{var}(g_j(W,Y)f_\beta)\right) & \textit{(Observable)}\\ N\left(0,\text{var}\left(\mathbb{P}(\gamma_0)^{-1}\eta_0-\mathbb{P}(\gamma_1)^{-1}\eta_1\right)\right) & \textit{(Counterfactual)}\end{cases} \tag{3.43}$$

where, for $a \in \{0, 1\}$,

$$\gamma_a = \begin{cases} (1-\phi)\mathbb{1}\{A = a\} & \textit{(for } g^{FPR}\textit{)} \\ \phi\mathbb{1}\{A = a\} & \textit{(for } g^{FNR}\textit{)} \end{cases} \tag{3.44}$$

$$\eta_a = \gamma_a \left(f_\beta - \frac{\mathbb{P}[\gamma_a f_\beta]}{\mathbb{P}[\gamma_a]} \right) \tag{3.45}$$

3.6 Simulations

All computations in this and subsequent sections were carried out on a 2013 MacBook Pro with a 2.4 GHz dual-core processor and 8GB of RAM.

3.6.1 Data-generating process

We illustrate the penalized-min procedure with respect to potential outcomes Y^0. As in a real data setting, each estimator $\widehat{\beta}_\lambda$ is constructed using only observable data, but unlike in a real data setting, we use the known values of Y^0 to evaluate the resulting predictors. The data generating process is as follows, for data $Z = (A, X, D, Y^0, Y^1, Y)$.

$$\begin{aligned} \mathbb{P}(A = 1) &= 0.3 \\ X \mid A &\sim N\left(A * (1, -0.8, 4, 2)^T, I_4\right) \\ \mathbb{P}(D = 1 \mid A, X) &= \min\{0.975,\ \text{expit}((A, X)^T(0.2, -1, 1, -1, 1))\} \\ \mathbb{P}(Y^0 = 1 \mid A, X, D) &= \text{expit}((A, X)^T(-5, 2, -3, 4, -5)) \\ \mathbb{P}(Y^1 = 1 \mid A, X, D) &= \text{expit}((A, X)^T(1, -2, 3, -4, 5)) \\ Y &= (1 - D)Y^0 + DY^1 \end{aligned}$$

where I_4 denotes the 4×4 identity matrix. $A = 1$ represents the minority group. There are no previously trained predictors; i.e. $S = \emptyset$, so the collected covariates consist of $W = (A, X)$. This data generating process satisfies Assumptions 4–6: the last line expresses the consistency assumption; the propensity score $\pi(A, X) = \mathbb{P}(D = 1 \mid A, X)$ is upper bounded at 0.975 to satisfy positivity; and $Y^a \perp\!\!\!\perp D|W$ for $a \in \{0, 1\}$, satisfying ignorability.

The two groups $A = 0$ and $A = 1$ differ in the distribution of covariates (Figure 3.3) and decisions and outcomes (Table 3.3). The minority group experiences a positive decision ($D = 1$) 18% of the time, while the majority group experiences it 50% of the time. Outcomes Y^0 and Y^0 are higher for the minority group, with a larger disparity for potential outcomes than for observable outcomes.

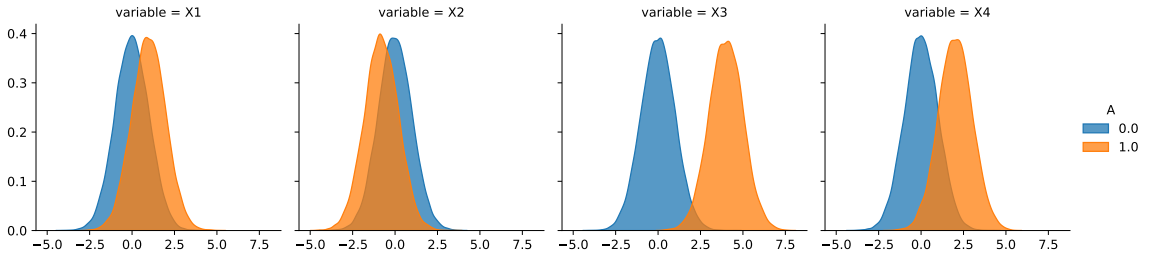


Figure 3.3: Conditional covariate distributions for the two groups $A = 0$ and $A = 1$ in the simulated data. Curves are kernel density estimates.

A	$\mathbb{E}[D\|A]$	$\mathbb{E}[Y^0\|A]$	$\mathbb{E}[Y\|A]$
0	0.50	0.50	0.67
1	0.18	0.76	0.71

Table 3.3: Distribution of decisions and outcomes for groups $A = 0$ and $A = 1$ in the simulated data.

As a reference point for our method, in Table 3.4 we compute the performance of the counterfactual Bayes-optimal predictor $f(A, X) = \mathbb{E}[Y^0|A, X]$, which is defined in the data-generating process. All measures are computed using the known values of Y^0 in a dataset of size 50,000. We include both MSE, with is the measure directly targeted by our method, as well as area under the curve (AUC). The Bayes-optimal predictor is highly accurate, with an MSE of 0.05 and an AUC of 0.98. The MSE of 0.05 is a lower bound on the risk achievable by any predictor. Unsurprisingly, given the difference in the distribution of outcomes across the two groups, the Bayes-optimal predictor has a large rate disparity $|\mathbb{E}[f|A = 0] - \mathbb{E}[f|A = 1]|$. The differences in generalized false positive and false negative rates, however, are relatively small.

Model	MSE	AUC	rate-diff	FPR-diff	FNR-diff
Bayes-optimal	0.05	0.98	0.26	0.07	0.05

Table 3.4: Risk and fairness measures with respect to Y^0 for the Bayes-optimal predictor $\mathbb{E}[Y^0|A, X]$ in the simulated data. The predictor is highly accurate, with low MSE and high AUC. It has a relatively large rate disparity but small disparities in generalized false positive and false negative rates.

3.6.2 Base predictors and nuisance models

We now investigate the performance of our method in a counterfactual setting. We randomly sample three iid datasets of size n = 1000, representing $\mathcal{D}_{\text{learn}}, \mathcal{D}^{\text{nuis}}_{\text{train}}$, and $\mathcal{D}^{\text{target}}_{\text{train}}$. We train four base predictors on $\mathcal{D}_{\text{train}}$, with A, X as covariates and Y as the outcome. We train only on data in

which $D = 0$: under the ignorability assumption, $\mathbb{E}[Y|A, X, D = 0] = \mathbb{E}[Y^0|A, X]$, so this results in predictors which are designed to estimate Y^0. The predictors consist of a random forest, a gradient boosted (GB) classifier, a Gaussian Naive Bayes model, and a ridge regression, all chosen for convenience and ease of computation. (In practice, a logistic regression would be a natural choice for a base predictor. Since the actual regression function $\mathbb{E}[Y^0|A, X]$ is logistic, however, we do not use this model in order to avoid making the problem too easy, and to simulate a real data setting in which it is unlikely that the true regression function is known up to a finite dimensional parameter.) In addition to these predictors, we include a mean predictor, which always predicts the value $\mathbb{P}_n(Y|D = 0)$. This plays essentially the same role as an intercept in ordinary linear regression.

We use random forest classifiers to estimate the propensity and outcome models $\widehat{\pi}$ and $\widehat{\mu}_0$ on $\mathcal{D}^{\text{nuis}}_{\text{train}}$. All models were trained with their default tuning parameters using the scikit-learn library in Python. Predictors $\widehat{\beta}_\lambda$ are computed using $\mathcal{D}^{\text{target}}_{\text{train}}$.

After a set of model coefficients $\widehat{\mathcal{B}}_n$ is computed, we estimate the risk of fairness properties of every $f_\beta : \beta \in \widehat{\mathcal{B}}_n$. In order to understand the true range of risk and fairness values that our method produces, we use a large test set $\mathcal{D}_{\text{test}}$ of size 10,000, and in place of $\widehat{\phi}$, the nuisance quantity that would be required in a real data setting, we use the known values of Y^0 to compute the risk and fairness estimates. (For comparison purposes, estimates were also computed using μ_0 instead of Y^0; the results were virtually identical.) Since the true Y^0 is used, there is no need to split $\mathcal{D}_{\text{test}}$ into $\mathcal{D}^{\text{nuis}}_{\text{test}}$ and $\mathcal{D}^{\text{target}}_{\text{test}}$.

Table 3.5 shows the performance of the base predictors as well as the ordinary (unpenalized) least squares (OLS) solution, i.e. the predictor f_{β_λ} with $\lambda = 0$. The OLS predictor is the (estimated) MSE-minimal aggregation of the five base predictors, computed without regard for fairness. The OLS weights are $[-0.27, 0.09, 0.40, -0.11, 0.94]$; each base predictor appears to make a nontrivial contribution. The MSE of the base predictors ranges from 0.08 to 0.27. The four non-constant predictors improve substantially on the mean predictor with respect to both MSE and AUC. The mean predictor necessarily has a value of 0 for all three disparity measures, while the disparity values of the other base predictors vary between 0.10 and 0.59. As expected, the OLS predictor has lower MSE than any of the base predictors. The performance of the OLS predictor is similar to the performance of the Bayes-optimal predictor in Table 3.5: both have a small MSE, a relatively large rate disparity, and relatively small error rate disparities.

Model	MSE	AUC	rate-diff	FPR-diff	FNR-diff
Mean	0.27	0.50	0.00	0.00	0.00
Random Forest	0.09	0.95	0.28	0.21	0.11
GB Classifier	0.08	0.96	0.31	0.22	0.10
Naive Bayes	0.17	0.84	0.52	0.59	0.40
Ridge	0.09	0.98	0.22	0.09	0.10
OLS	0.07	0.98	0.26	0.10	0.08

Table 3.5: Performance of the five base predictors and the ordinary least squares (OLS) predictor in the simulated data. The OLS weights are $[-0.27, 0.09, 0.40, -0.11, 0.94]$. The OLS predictor substantially improves on the MSE of the base predictors. The performance profile of the OLS predictor is close to the profile of the Bayes-optimal predictor in Table 3.4.

3.6.3 Results: one fairness penalty

We now compute a set of fairness-penalized models, applying a single fairness penalty at a time. Let

$$\Lambda_{n,1} = \{0, 0.001, 0.01, 1, 10, 20, 50, 100, 500, 1000, 2000\}. \tag{3.46}$$

For each $\lambda \in \Lambda_{n,1}$, we compute $\widehat{\beta}_\lambda$ for each fairness function $g \in \{g^{\text{rate}}, g^{\text{FPR}}, g^{\text{FNR}}\}$. The value $\lambda = 0$ corresponds in each case to the OLS solution, so this yields a total of $(|\Lambda_{n,1}| - 1) * 3 + 1 = 31$ models.

The risk and fairness values for each model are plotted in Figure 3.4. The disparity corresponding to the targeted constraint is represented by a solid line, while the other two disparities and the MSE are represented by dashed lines. We emphasize that the values in this figure are computed on $\mathcal{D}_{\text{test}}$, after the predictors $\widehat{\beta}_\lambda$ are computed on $\mathcal{D}_{\text{train}}$.

As expected, as λ increases, the targeted disparity of the resulting predictor generally decreases. The decrease is monotonic, except at one point: $\lambda = 1$ for FPR-diff, which may be a result of sampling noise. The rate difference decreases from 0.26 to 0.04. The FPR difference decreases from 0.10 to 0, then remains at 0.01. The FNR difference decreases from 0.08 to 0.04. When the rate-diff is penalized, the decrease in the target disparity is accompanied by a slight increase in MSE, from 0.07 to 0.09, as well as small increases in FPR-diff and FNR-diff. When FPR-diff is penalized, all three disparities fall together, while the increase in MSE is miniscule, from 0.067 to 0.071. The same is true when FNR-diff is targeted: the MSE increases from 0.067 to 0.069.

These results illustrate (1) that the penalty term successfully controls the target disparity, (2) that an increase in fairness need not come at the cost of a substantial decrease in accuracy, and (3) that a decrease in one disparity need not produce an increase in other disparities. In the second two panels, the penalized predictors are uniformly more fair than the OLS predictor, with essentially no

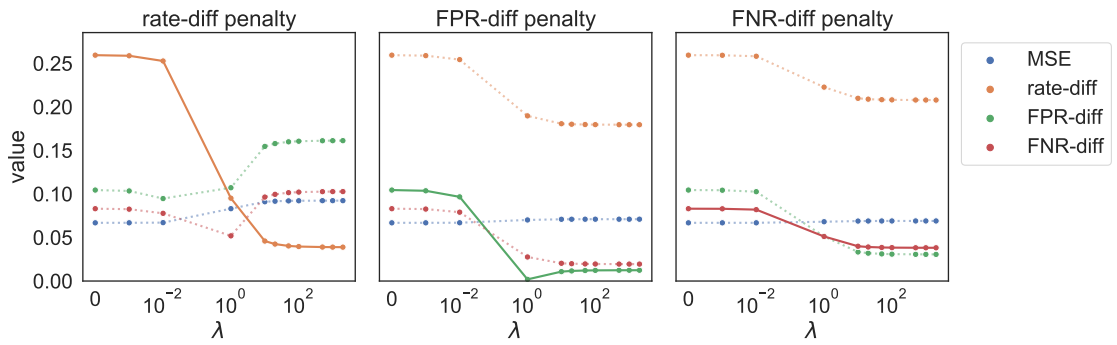


Figure 3.4: Risk and fairness for models subject to one of three penalties, in the simulated data. The x-axis represents the fairness penalty coefficient λ. The y-axis represents the MSE and the disparity values of the resulting predictor $\widehat{\beta}_\lambda$, computed on an independent test set of size 10,000 using the known values of Y^0. The leftmost point ($\lambda = 0$) in each panel corresponds to the OLS solution. Solid lines indicate the metric that is penalized in training.

change in accuracy. Additionally, in these two panels, even though only one disparity was penalized at a time, all three disparities decreased as λ increased.

3.6.4 Results: multiple fairness penalties

We now apply all three fairness penalties simultaneously. Define $\Lambda = \Lambda_{n,1} \times \Lambda_{n,1} \times \Lambda_{n,1} \subset \mathbb{R}^3_{0+}$. The collection $\widehat{\mathcal{B}}_n = \{\widehat{\beta}_\lambda : \lambda \in \Lambda\}$ now contains $|\Lambda_{n,1}|^3 = 1331$ models. We use the same base predictors and nuisance predictors as in the previous section. The process of training the predictors, computing $\widehat{\mathcal{B}}_n$, and estimating the risk and fairness for each predictor, took less than 10 seconds.

Figure 3.5 plots each of the three disparities against MSE, for each of the 1331 predictors, as well as the base predictors and the OLS predictor. As expected, the OLS predictor has the smallest MSE. Fewer than 1331 dots are visible in each panel, due to the fact that many of the models substantially overlap in fairness-accuracy space. Nevertheless, the models span a wide range of performance profiles. For all three disparities, models exist that take the disparity to 0, with relatively small increase in MSE relative to the OLS predictor. For rate-diff and FNR-diff, these models notably do not appear in Figure 3.4, where the lowest value for these two disparities are 0.04. We only discover these models by applying multiple penalties simultaneously. All the aggregated predictors are substantially more accurate than the mean predictor, which has disparities of 0 but the highest MSE of any model.

Figure 3.6 plots the same 1331 models with respect to each pair of disparities, with color indicating MSE. This figure illustrates the interplay of three metrics at once, and is of interest to users who wish to control two disparities simultaneously. For example, users who wish to target

(counterfactual) equalized odds would be interested in the bottom panel that plots FNR-diff and FPR-diff.

These views once again reveal a wide range of model behavior. Unsurprisingly, many of the highest MSE models are close to the origin, but the relationship between MSE and distance to the origin is far from monotonic. In all three panels, there is a line of models stretching from the OLS predictor that represent improvements in both disparities with minimal increase in MSE. In the bottom panel, for example, there are models with FPR-diff close to 0, FNR-diff under 0.05, and MSE under 0.10. These models approximately satisfy equalized odds, and they represent an increase in MSE of less than 0.03 relative to the OLS predictor.

In order to examine this more precisely, Table 3.6 shows the performance of the models with the minimum L2 distance from the origin in the fairness-accuracy space defined by MSE as well as one to three disparities. For example, the 'MSE + rate-diff' row represents the model with the smallest L_2 norm in the (MSE, rate-diff) vector. FPR-diff and FNR-diff can be minimized, singly or jointly, with no increase in MSE relative to the OLS predictor. Rate-diff can be substantially reduced with relatively small increase in MSE. Perhaps surprisingly, all three disparities can be jointly minimized, to 0.06 (rate-diff), 0.03 (FPR-diff), and 0.02 (FNR-diff), with only a 0.06 increase in MSE and a 0.02 decrease in AUC relative to the unpenalized OLS predictor.

	MSE	AUC	rate-diff	FPR-diff	FNR-diff
MSE + rate-diff	0.09	0.95	0.04	0.16	0.10
MSE + FPR-diff	0.07	0.98	0.19	0.00	0.03
MSE + FNR-diff	0.07	0.98	0.18	0.01	0.02
MSE + rate-diff + FPR-diff	0.12	0.90	0.04	0.05	0.09
MSE + rate-diff + FNR-diff	0.10	0.95	0.04	0.16	0.08
MSE + FPR-diff + FNR-diff	0.07	0.98	0.18	0.01	0.02
MSE + rate-diff + FPR-diff + FNR-diff	0.13	0.96	0.06	0.03	0.02
OLS	0.07	0.98	0.26	0.10	0.08

Table 3.6: Performance of the models that minimize the Euclidean norm of MSE and one or more disparities, in the simulated data. The OLS predictor is included again for reference. All three disparities can be minimized, singly or jointly, with no impact or a small impact on MSE.

3.7 Results: Recidivism risk prediction

We next illustrate our method on the COMPAS dataset gathered by ProPublica (Angwin and Larson, 2016; Angwin et al., 2016). The dataset comprises public arrest records, criminal records, and COMPAS scores from a single county in Florida, spanning 2013–2016. COMPAS is a collection of tools developed by the company Equivant (formerly Northpointe) designed to assess the risk of

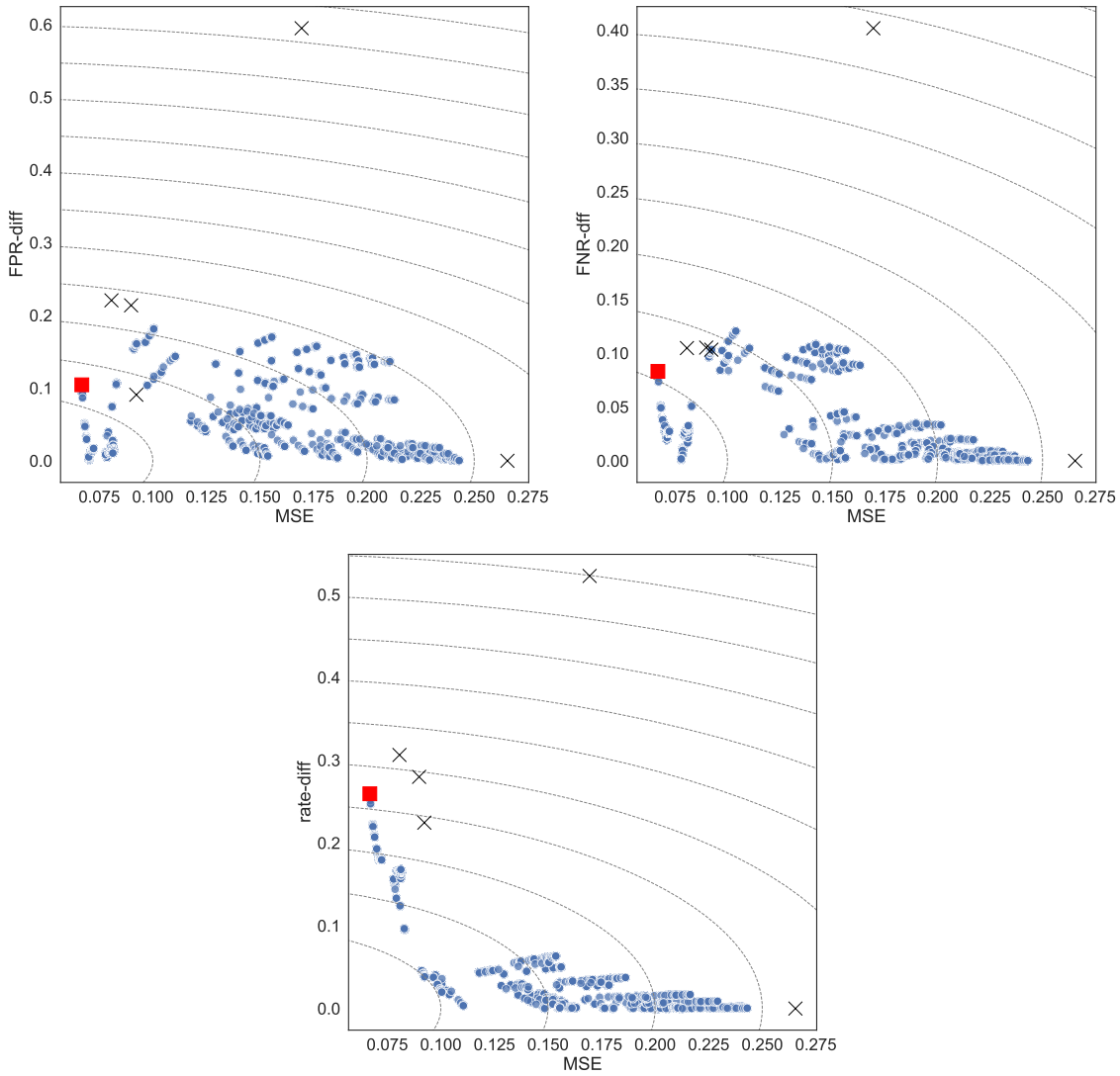


Figure 3.5: Disparity and MSE values for each of 1331 models in the simulated data. Black 'X's represent the base predictors, with the mean predictor at the bottom right of each panel. The red square is the OLS predictor, and the blue dots are the penalized predictors. Radius lines indicate distance from the origin. Despite substantial overlap, the predictors span a wide range of fairness and accuracy values. For each disparity, many models exist which take that disparity to 0, at a small cost in MSE relative to the OLS predictor.

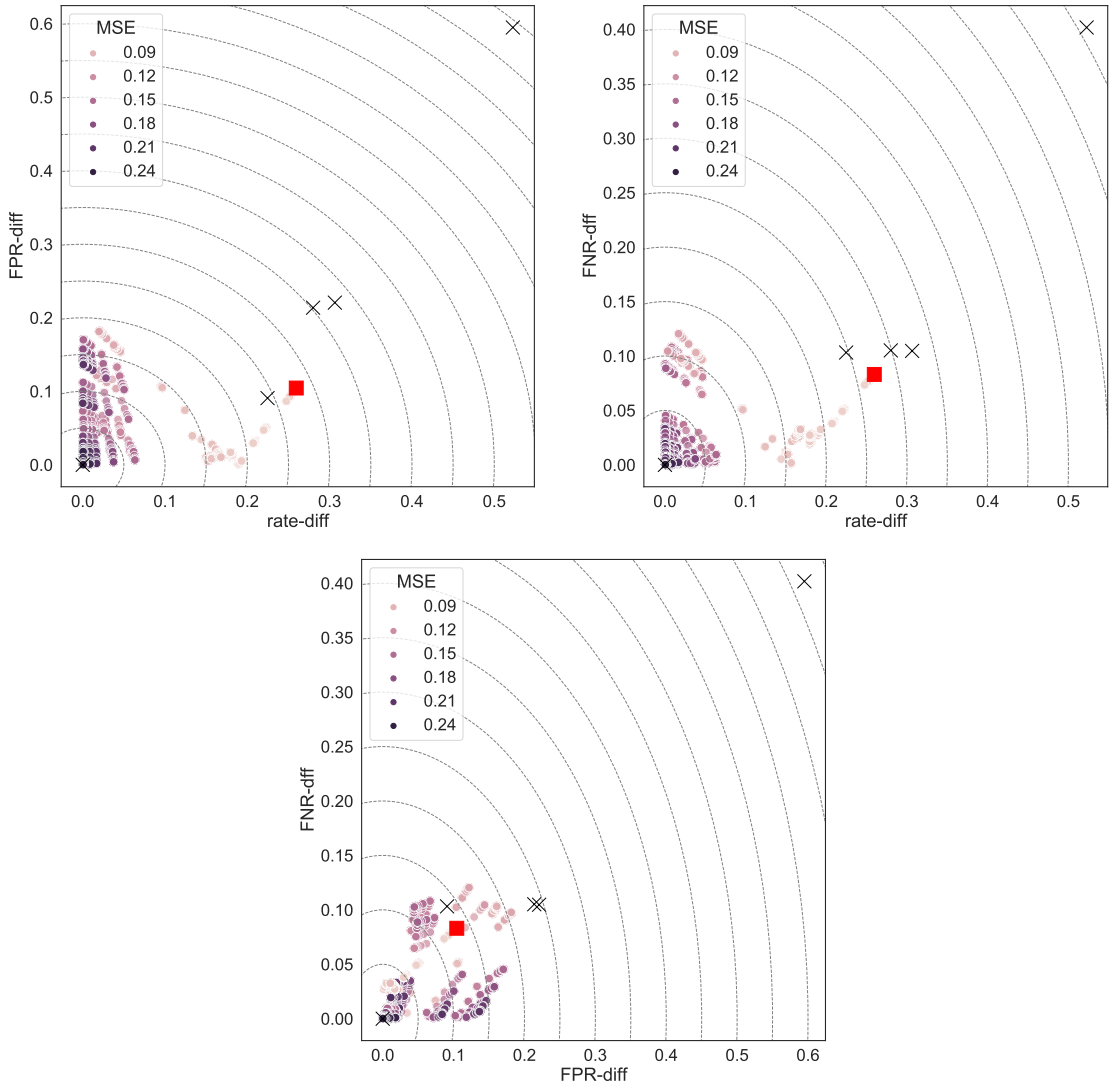


Figure 3.6: Pairs of disparity values and MSE values for each of 1331 models in the simulated data. Black 'X's represent the base predictors, with the mean predictor at the origin in each panel. The red square is the OLS predictor, and the dots are the penalized predictors. Radius lines indicate distance from the origin. Each pair of disparities can be jointly decreased with minimal increase in MSE relative to the OLS predictor.

recidivism. We utilize the COMPAS scores for general, as opposed to violent, recidivism. The scores consist of risk deciles, coded 1-10, which we normalize to the range $[0.1, 1]$. COMPAS takes as input up to 137 features (Northpointe, 2015; Rudin et al., 2020), which are unavailable in this dataest. We utilize just three features as covariates: an indicator for defendant age greater than 45, an indicator for defendant age less than 25, and the number of prior arrests, ranging from 0 to 29. Previous work has found that models trained using just these covariates perform similarly to COMPAS (Angelino et al., 2018).

The sensitive feature is race, restricted to defendants who are coded African-American ($n = 3175$) or Caucasian ($n = 2013$). The decision variable D represent pretrial release, with $D = 0$ if defendants are released and $D = 1$ if they are detained. The outcome of interest Y^0 is rearrest within two years, should a defendant be released pretrial. Since it difficult to assess the plausibility of the positivity and ignorability assumptions without consulting with domain experts, we conducted analyses in both the counterfactual and observable setting. The results and conclusions were largely the same, so we only include the counterfactual results below.

We split the data into five datasets, each with approximately 1040 rows: $\mathcal{D}_{\text{learn}}$, $\mathcal{D}^{\text{nuis}}_{\text{train}}$, $\mathcal{D}^{\text{target}}_{\text{train}}$, $\mathcal{D}^{\text{nuis}}_{\text{test}}$, and $\mathcal{D}^{\text{target}}_{\text{test}}$. As base predictors, we used the four model types from the previous section as well as a logistic regression. We used random forest classifiers for the nuisance predictors in both the training and test data. Table 3.7 gives the estimated performance of the five base predictors, COMPAS, and the OLS predictor, which spans COMPAS and the base predictors. Previous work found differences in a binarzed version of COMPAS for both observable (Angwin and Larson, 2016) and counterfactual (Mishler, 2019) false positive vs false negative rates for African-American vs. Caucasian defendants. Those differences appear here in the generalized error rates. COMPAS also has a large rate disparity. Perhaps surprisingly, the base predictors all yield smaller disparities than COMPAS, even though they generally also have smaller MSE.

We compute aggregated models using the same sets of penalty vectors $\Lambda_{n,1}$ and Λ as in the previous section. Figure 3.7 shows the result of applying one penalty at a time. Once again, the targeted disparity can be decreased with a minimal cost in MSE. Here, all three disparities appear to rise or fall together. Figure 3.8 shows disparities and MSE values for all 1331 models. Most models fall within a narrow range of MSE values that also includes COMPAS, so the primary value of aggregation here is in reducing disparities. Nearly all the aggregated models improve on COMPAS in terms of both risk and fairness. The top row of Figure 3.8 shows that all three disparities can be individually reduced to 0 with minimal cost in MSE relative to the OLS predictor, and the bottom row shows that these improvements also extend over pairs of disparities.

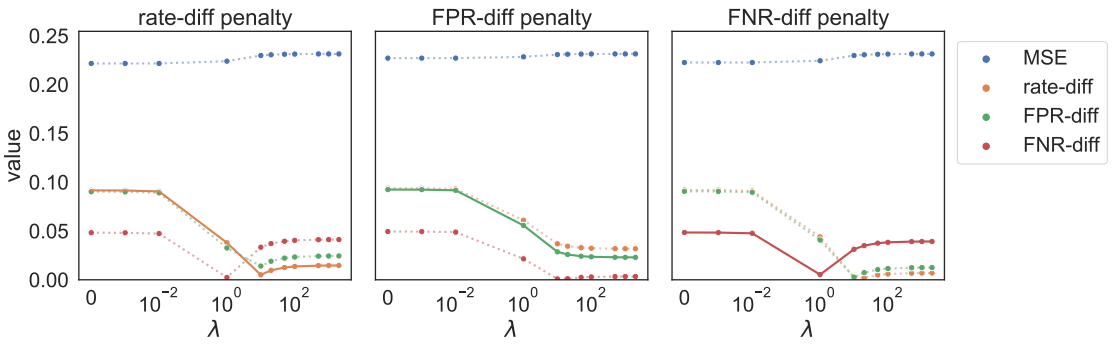


Figure 3.7: Risk and fairness for models subject to one of three penalties, in the COMPAS data. The x-axis represents the fairness penalty coefficient λ. The y-axis represents the MSE and the disparity values of the resulting predictor $\widehat{\beta}_\lambda$, computed on an independent test set of size 10,000 using the known values of Y^0. The leftmost point ($\lambda = 0$) in each panel corresponds to the OLS solution. Solid lines indicate the metric that is penalized in training.

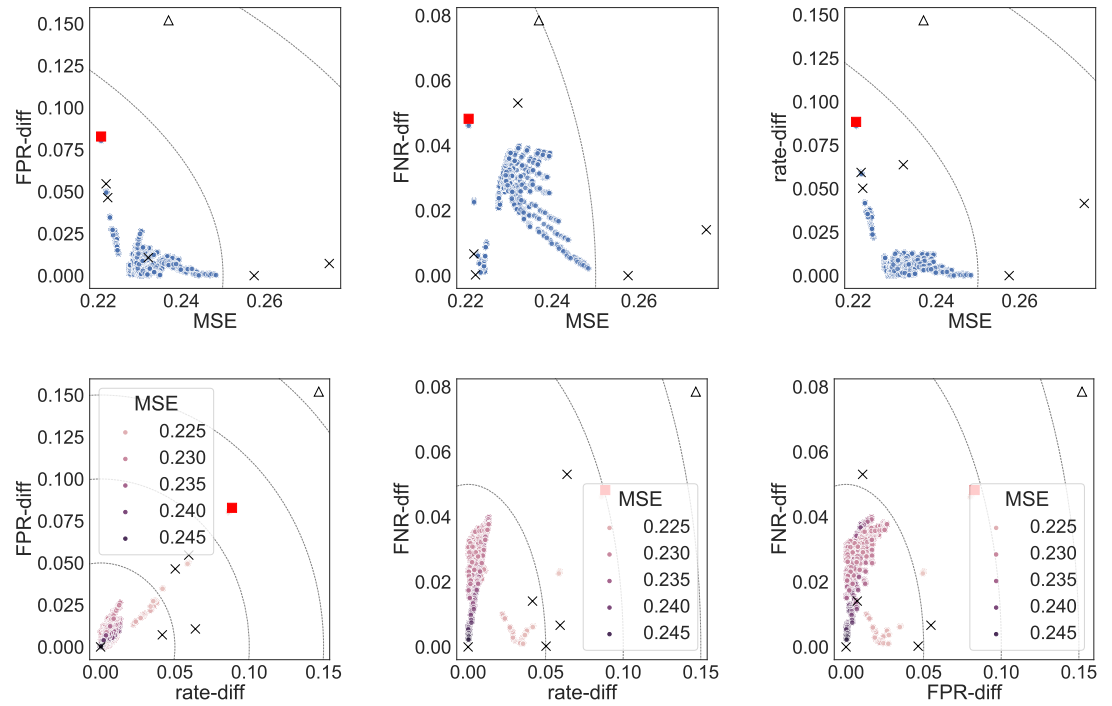


Figure 3.8: Disparity against MSE (top row) or pairs of disparities colored by MSE (bottom row), for each of 1331 models in the COMPAS data. The black triangle represents COMPAS, the black 'X's represent the other base predictors, the red square is the OLS predictor, and the blue dots are the penalized predictors. Radius lines indicate distance from the origin. Most of the models improve on COMPAS in terms of both MSE and the relevant disparity. For each disparity, many models exist which take that disparity to 0, at a small cost in MSE relative to the OLS model.

	MSE	rate-diff	FPR-diff	FNR-diff
Mean	0.26	0.00	0.00	0.00
Random Forest	0.28	0.05	0.03	0.02
Logistic	0.22	0.06	0.06	0.00
GB Classifier	0.23	0.06	0.01	0.05
Ridge	0.22	0.05	0.05	0.00
COMPAS	0.24	0.15	0.15	0.08
OLS	0.22	0.09	0.09	0.05

Table 3.7: Estimated performance of the five base predictors, COMPAS, and the OLS predictor in the COMPAS dataset. The OLS weights are $[0.39, 0.12, 0.79, 0.10, -1.08, 0.93]$. The OLS predictor does not perform substantially better than the base predictors. COMPAS has different false positive and false negative rates for African-American vs Caucasian defendants, as well as a rate disparity. The base predictors all have smaller disparities than COMPAS, and generally smaller MSE.

Table 3.8 gives the performance of the models that are closest to the origin in various fairness-accuracy subspaces. All three disparities can be minimized, jointly or in any combination, while improving on the MSE relative to COMPAS.

Since the MSE of the mean predictor is only 0.26, and the MSEs of the models in Table 3.8 fall in the range 0.22–0.23, these results raise the possibility that these are close to the trivial mean predictor. Figure 3.9 rules this out however; it shows histograms of the model predictions and calibration curves for the "Best" model (the model that minimizes MSE + rate-diff + FPR-diff + FNR-diff) vs. COMPAS. Calibration curves were computed on the subset of the data for which $D = 0$, i.e. for which $Y = Y^0$ under Assumption 4. Although the Best model does not produce predictions lower than 0.28 (vs. 0.10 for COMPAS), the predictions are far from concentrated around a single point. The Best model appears to be at least as well calibrated as COMPAS.

Model	MSE	rate-diff	FPR-diff	FNR-diff
MSE (OLS)	0.22	0.09	0.09	0.05
MSE + rate-diff	0.23	0.02	0.02	0.01
MSE + FPR-diff	0.22	0.03	0.03	0.00
MSE + FNR-diff	0.22	0.04	0.04	0.01
MSE + rate-diff + FPR-diff	0.23	0.02	0.02	0.01
MSE + rate-diff + FNR-diff	0.22	0.03	0.03	0.00
MSE + FPR-diff + FNR-diff	0.22	0.03	0.03	0.00
MSE + rate-diff + FPR-diff + FNR-diff	0.23	0.02	0.02	0.01
COMPAS	0.24	0.15	0.15	0.08

Table 3.8: Performance of the models that minimize the Euclidean norm of MSE and zero to three disparities, in the COMPAS data. The OLS predictor is included again for reference. All three disparities can be minimized, singly or jointly, with no impact or a small impact on MSE.

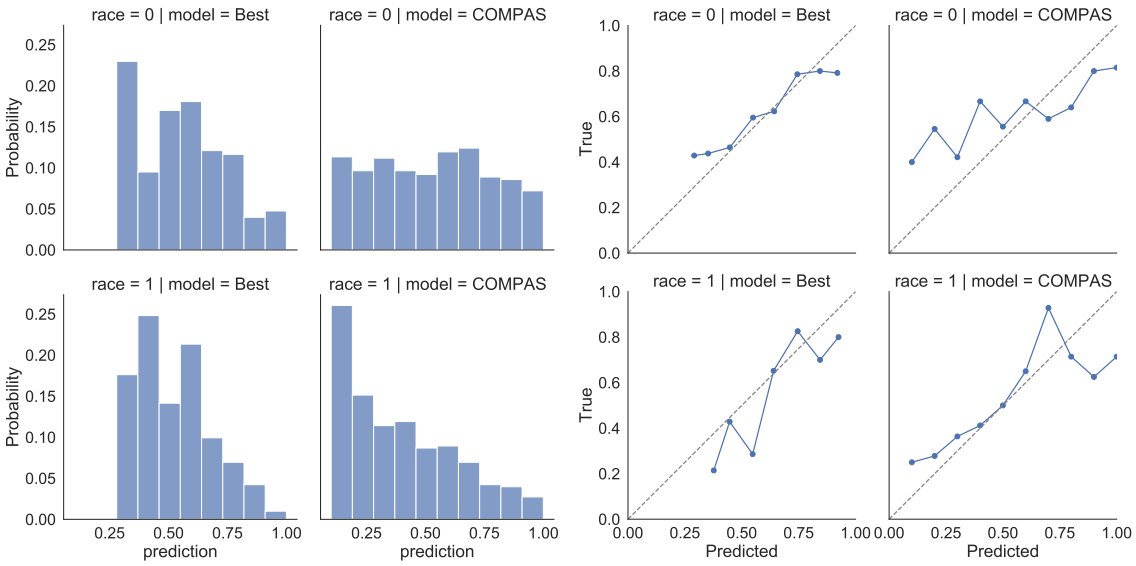


Figure 3.9: Histograms of predictions and calibration curves for COMPAS and the "Best" model, which jointly minimizes the Euclidean norm of the MSE and the three disparities. Values were computed $\mathcal{D}_{\text{test}}^{\text{target}}$; calibration was assessed on the subset of the data with $D = 0$ ($n = 351$). The Best model predictions range from 0.28 to 0.97, vs. 0.10 to 1.0 for COMPAS. The Best model appears to be at least as well calibrated as COMPAS.

3.8 Results: Income prediction

Finally, we apply our method in the observable setting, using the Adult dataset (Dua and Graff, 2017). This dataset comprises demographic variables derived from the 1994 U.S. Census. We consider sex as a sensitive feature, coded 0 or 1, and we utilize as covariates a set of indicator variables indicating age by decade, and a set of indicator variables indicating the number of years of education. The classification task is to predict whether an individual's income is over \$50K/year, for example for the purpose of deciding whether to issue a loan.

We randomly split the data into four datasets: $\mathcal{D}_{\text{learn}}$ and $\mathcal{D}_{\text{train}}$, consisting of 14,653 and 14,652 rows; and $\mathcal{D}_{\text{test}}$ and $\mathcal{D}_{\text{validate}}$, consisting of 9,768 and 9,769 rows. $\mathcal{D}_{\text{validate}}$ is used to compare the performance of the selected best models to the fair predictors.

This dataset does not contain any previously trained predictors. We use the same five base predictor types as in the COMPAS analysis. Additionally, we use $\mathcal{D}_{\text{train}}$ to train three "fair" models with other fairness methods: *adversarial debiasing* (Zhang et al., 2018), *reductions* (Agarwal et al., 2018), and a *meta-algorithm* (Celis et al., 2020). All models were trained using the Python library aif360, a set of tools that provide access to a range of fairness methods via a consistent interface (Bellamy et al., 2018). The three chosen methods yield binary classifiers, and they are all designed to minimize *rate-diff* in an observable setting with a binary outcome. Since, for a binary classifier

$\widehat{Y}$, MSE is equal to classification error $\mathbb{P}(\widehat{Y} \neq Y)$, we may regard these three predictors as the result of methods that seek to minimize MSE among specific classes of binary predictors.

We construct aggregated predictors using the five base models ("base5") or using the five base models and the three fair predictors ("base8"). Table 3.9 gives the performance of the base predictors, the fair predictors, and the two OLS predictors. Compared to the base predictors, two of the three fairness methods result in a substantially lower rate-diff, which is the disparity they aim to minimize. The base predictors and the OLS predictors have lower MSE, higher AUC, and higher disparities compared to these two fair predictors.

Model	MSE	AUC	rate-diff	FPR-diff	FNR-diff
Mean	0.18	0.50	0.00	0.00	0.00
Random Forest	0.14	0.81	0.19	0.14	0.24
Logistic	0.14	0.81	0.20	0.14	0.25
GB Classifier	0.14	0.82	0.19	0.13	0.24
Ridge	0.14	0.81	0.17	0.13	0.16
Adversarial	0.21	0.67	0.07	0.00	0.03
Reductions	0.22	0.62	0.01	0.03	0.07
Meta	0.30	0.68	0.18	0.26	0.26
OLS - base5	0.14	0.82	0.19	0.13	0.24
OLS - base8	0.14	0.81	0.20	0.14	0.25

Table 3.9: Estimated performance in the Adult dataset of five base predictors, three "fair" predictors, and the two OLS predictors, which aggregate only the five base predictors or all eight predictors. The OLS weights are $[0.03, 0.29, 0.65, 0.03, 0.01]$, for base5, and $[-0.01, 0.26, 0.56, 0.18, 0.04, -0.02, -0.03, 0]$, for base8. Only two of the three fairness methods successfully control their targeted disparity, rate-diff. The Meta predictor has a rate-diff which is comparable to the base predictors which are trained without regard to fairness. The OLS predictors perform comparably to the base predictors.

We compute aggregated models using the same sets of penalty vectors $\Lambda_{n,1}$ and Λ as in the previous two sections. Figure 3.10 shows the result of applying one penalty at a time. Once again, the targeted disparity can be decreased, though here there is a consistent tradeoff between fairness and MSE. As in the COMPAS data, all three disparities appear to rise or fall together. Figure 3.8 shows disparities and MSE values for all 1331 models. Fairness-accuracy tradeoffs are most evident for rate-diff and FPR-diff in the top row of this figure, where only the five base predictors are used. In the bottom row, where the fair predictors are included as basis functions, the tradeoffs essentially disappear: all three disparities can be reduced to 0 with virtually no cost in MSE. Plots of two disparities colored by MSE are included in Appendix 3.F.

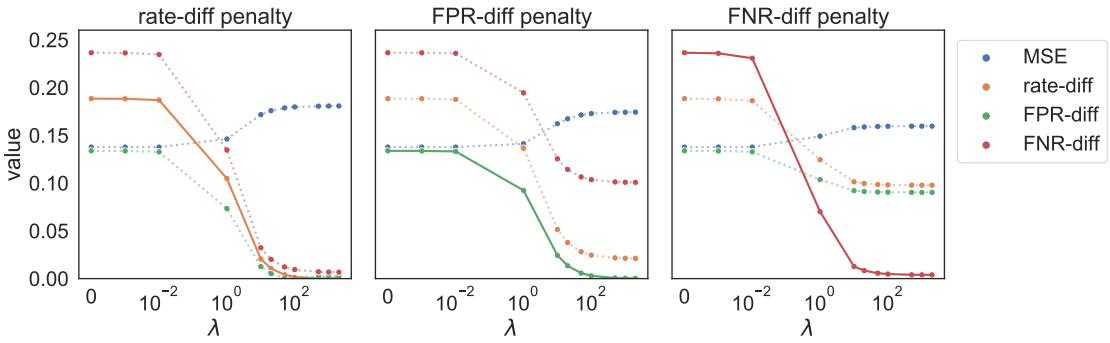


Figure 3.10: Risk and fairness for models subject to one of three penalties, in the Adult data, using five base predictors. The x-axis represents the fairness penalty coefficient λ. The y-axis represents the MSE and the disparity values of the resulting predictor $\widehat{\beta}_\lambda$, computed on an independent test set of size 10,000 using the known values of Y^0. The leftmost point ($\lambda = 0$) in each panel corresponds to the OLS solution. Solid lines indicate the metric that is penalized in training.

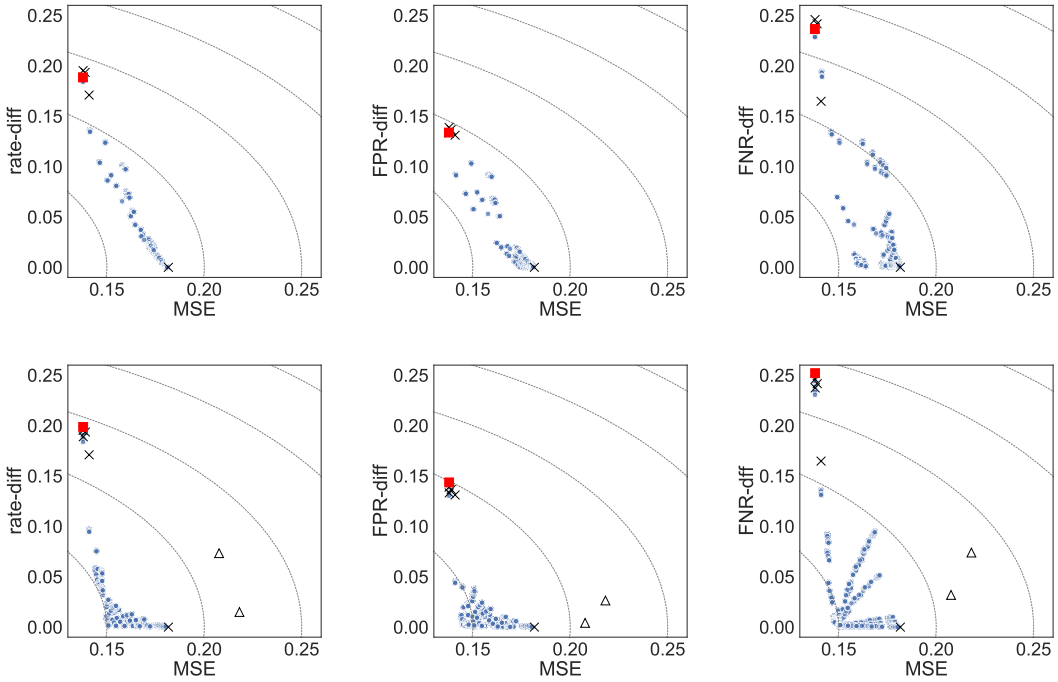


Figure 3.11: Disparity against MSE for models based on the five base predictors (top row) or the five base predictors plus the three fair predictors, for each of 1331 models in the Adult data. Black 'X's represent the base predictors, the red square is the OLS predictor, and the blue dots are the penalized predictors. Radius lines indicate distance from the origin. For the sake of legibility, the Meta predictor, which has an MSE of 0.30, is excluded. The row exhibits small but clear fairness-accuracy tradeoffs for rate-diff and FPR-diff. The bottom row shows that with the inclusion of the fair predictors, each disparity can be taken to 0 with almost no cost in MSE relative to the OLS predictor.

3.8.1 Model validation

Using the performance estimates from the test data, we again select the same seven models that minimize the distance from the origin in various accuracy-fairness subspaces, for both the base5 and base8 models. Table 3.10 shows the performance estimates, as well as the three fair predictors, on the validation data. The estimates on the test and validation data differed by no more than approximately 0.005.

Both the base5 and base8 penalized predictors are substantially more fair than the OLS models, while incurring very small increases in MSE. The high AUC values confirm that these are accurate predictors. All the penalized predictors have small disparities compared to the OLS predictors; this reiterates the result observed in Figure 3.10, in which targeting any single disparity tends to reduce all three. Explicitly minimizing multiple disparities simultaneously is also not necessarily more costly in terms of performance than minimizing a single disparity.

The penalized predictors have substantially lower MSE and higher AUC than the fair predictors, and they are in many cases more fair. The fair predictors achieve values of 0.08, 0.01, and 0.17 for rate-diff the disparity they aim to minimize. The base5 predictors that include rate-diff in their criteria achieve rate-diffs of 0.04, 0.04, 0.06, and 0.02. The corresponding base8 predictors achieve rate-diffs of 0.01, 0.03, 0.01, and 0.03.

The base5 results show that our method yields predictors that perform comparably to or better than existing fairness methods. The base8 results highlight the flexibility of our approach: multiple predictors can be aggregated, regardless of whether or not they are trained with fairness properties in mind, with different weights to target different disparities. In this case, including the fair predictors in the aggregation improves both accuracy and fairness.

Each of the fair prediction methods contains tuning parameters that can be adjusted to return different predictors, as well as settings that allow them to target different fairness constraints, such as equalized odds. However, each method can only target a single fairness constraint at once. Additionally, these methods take substantial time to run. The Meta method took roughly 5 seconds, the Reductions method ran in approximately 15 seconds, and the Adversarial method, which relies on neural nets, took roughly a minute. By contrast, we were able to train the base predictors and compute and evaluate 1331 penalized models in 25 seconds.

The three fair predictors are binary by construction, whereas our method returns continuous predictors. Of course, these continuous predictors can be treated as binary, either by thresholding the output (for a deterministic classifier) or by treating the output as a probability and sampling from a corresponding Bernoulli distribution (for a randomized classifier). It is fast to compute estimates of the accuracy and fairness values from either of these two binarized classifiers and choose one which

minimizes the criteria of interest. For example, the base5 models include a model which, when thresholded at 0.5 to yield a deterministic binary classifier, achieves a classification error of 0.24 and disparities of 0 (to two digits). This has very slightly higher classification error than the Adversarial and Reductions predictors, but it exactly achieves equalized odds and demographic parity.

	Model	MSE	AUC	rate-diff	FPR-diff	FNR-diff
base5	MSE (OLS-base5)	0.14	0.82	0.19	0.13	0.24
	MSE + rate-diff	0.16	0.73	0.04	0.02	0.10
	MSE + FPR-diff	0.15	0.80	0.09	0.06	0.13
	MSE + FNR-diff	0.16	0.75	0.10	0.09	0.01
	MSE + rate-diff + FPR-diff	0.16	0.73	0.04	0.02	0.10
	MSE + rate-diff + FNR-diff	0.16	0.75	0.06	0.05	0.01
	MSE + FPR-diff + FNR-diff	0.16	0.75	0.06	0.05	0.01
	MSE + rate-diff + FPR-diff + FNR-diff	0.17	0.73	0.02	0.02	0.00
base8	MSE (OLS-base8)	0.14	0.81	0.20	0.14	0.25
	MSE + rate-diff	0.15	0.79	0.01	0.03	0.02
	MSE + FPR-diff	0.14	0.79	0.06	0.01	0.10
	MSE + FNR-diff	0.15	0.79	0.05	0.01	0.01
	MSE + rate-diff + FPR-diff	0.15	0.79	0.03	0.00	0.01
	MSE + rate-diff + FNR-diff	0.15	0.79	0.01	0.03	0.01
	MSE + FPR-diff + FNR-diff	0.15	0.79	0.04	0.00	0.01
	MSE + rate-diff + FPR-diff + FNR-diff	0.15	0.79	0.03	0.01	0.01
fair	Adversarial	0.21	0.67	0.08	0.01	0.04
	Reductions	0.22	0.62	0.01	0.03	0.06
	Meta	0.30	0.68	0.17	0.26	0.24

Table 3.10: Performance of the models that minimize the Euclidean norm of MSE and zero to three disparities, in the Adult data. Models are selected on the test data and evaluated on the validation data. The three fair predictors are included for reference. The base5 predictors aggregate the five base predictors, and the base8 predictors aggregate over all eight predictors. The "MSE" rows represent the OLS predictors. All three disparities can be minimized, singly or jointly, with only a small impact on MSE, and a small impact on AUC. Aggregated predictors are more accurate than the fair predictors, and have comparable or smaller values of rate-diff, the disparity that the fair predictors aim to minimize.

3.9 Conclusion

We developed a least squares framework for constructing fair predictors. This framework is extremely flexible, allowing users to combine arbitrary sets of predictors, including previously trained predictors and newly trained ones, regardless of whether they are designed to satisfy fairness constraints or not. Our framework can accommodate a wide range of disparities and allows users to minimize multiple disparities simultaneously. Our framework also accommodates both observable and counterfactual outcomes.

Within this framework, we developed three methods. The first two "constrained" methods allow users to minimize mean squared error subject to explicit fairness constraints, or minimize unfairness subject to an explicit constraint on the mean squared error. The third "penalized" method allows users to efficiently construct large sets of predictors and evaluate their risk and fairness properties. The penalized method enables users to explore fairness-accuracy and fairness-fairness tradeoffs in their problem setting, and it enables them to find a model with a favorable risk and fairness profile. Our results show that in many cases, disparities can be substantially reduced with no tangible increase in error relative to the unpenalized least squares solution.

Although our penalized approach is designed to minimize mean squared error and to penalize certain classes of disparities, the resulting models can naturally be evaluated with respect to any accuracy or fairness metric. For example, users might wish to consider only binary classifiers, so they may wish to evaluate classification error on thresholded versions of the penalized models. The penalized approach provides a principled way to explore various fairness-accuracy spaces, even if the fairness and/or accuracy metrics of interest aren't explicitly represented in the penalized expression.

Finally, the efficiency of our penalized method relies on the particular closed form of the parameterized predictors, which arises as a result of the mean squared error and the squared fairness terms. However, any quadratic function that involves a positive definite matrix has a closed form solution. This form could be preserved under different accuracy metrics and fairness terms by, for example, adding a regularization term $\beta^T M \beta$ for some positive-definite matrix M. This suggests that our approach could be adapted to explicitly target other accuracy and/or fairness metrics.

3.A Proof preliminaries

For convenience, we collect all the assumptions that appear in the paper:

1. For all n, $\mathbb{E}[bb^T]$ is positive definite. (PSD)
2. Uniformly in n, $\sup_{w \in \mathcal{W}} \|b(w)\| < \infty$ (Bound on basis)
3. The set $\{\mathbb{E}[g_j b]\mathbb{E}[g_j b]^T \beta_r^* : j \in \mathcal{I}\}$ is linearly independent. (LICQ - estimands)
4. $Y = DY^1 + (1-D)Y^0$ (Consistency)
5. $\exists \delta \in (0,1)$ s.t. $\mathbb{P}(\pi(W) \leq 1-\delta) = 1$ (Positivity)
6. $Y^0 \perp\!\!\!\perp D \mid W$ (Ignorability)
7. $\exists \gamma \in (0,1)$ s.t. $\mathbb{P}(\widehat{\pi}(A,X,S) \leq 1-\gamma) = 1$ (Bounded propensity estimator)
8. $\|\widehat{\phi} - \phi\| = o_{\mathbb{P}}(1)$, and $\|\widehat{\underline{\phi}} - \underline{\phi}\| = o_{\mathbb{P}}(1)$ (Consistent nuisance estimators)
9. $\|\widehat{\pi} - \pi\|\|\widehat{\mu}_0 - \mu_0\| = o_{\mathbb{P}}(1/\sqrt{n})$ (Nuisance parameter rates)

   $\|\widehat{\pi} - \pi\|\|\widehat{\nu}_0 - \nu_0\| = o_{\mathbb{P}}(1/\sqrt{n})$
10. The set $\{\mathbb{P}_n[\widehat{g}_j b]\mathbb{P}_n[\widehat{g}_j b]^T \widehat{\beta}_r : j \in \mathcal{I}_n\}$ is linearly independent. (LICQ - empirical)
11. $\Lambda_n \subseteq \Lambda \subset \mathbb{R}^t$ for some compact Λ. (Compact Λ)

Recall that for any function $f : \mathcal{Z} \mapsto \mathbb{R}$, we defined $\mathbb{P}_n(f) = n^{-1}\sum_{j=1}^n f = \int f d\mathbb{P}_n(Z)$ and $\mathbb{P}(f) = \int f d\mathbb{P}(Z)$ as the sample and true expectations of f, so that for example $\mathbb{P}(\widehat{\phi}) = \mathbb{E}[\widehat{\phi}|\mathcal{D}_{\text{train}}]$ or $\mathbb{E}[\widehat{\phi}|\mathcal{D}_{\text{test}}]$ is the expected value of $\widehat{\phi}(Z)$ once the relevant nuisance function estimate $\widehat{\phi}$ has been constructed.

We state several lemmas that are used in the proofs for the constrained and penalized settings. The first is a restatement of Lemma 2 in Kennedy et al. (2020).

Lemma 3.5.1. *(Kennedy, 2020) Let $\widehat{f} : \mathcal{Z} \mapsto \mathbb{R}$ be a function estimated on a nuisance dataset $\mathcal{D}^{nuis}$, and let $f : \mathcal{Z} \mapsto \mathbb{R}$ be another function. Assume* $\text{var}(\widehat{f} - f|\mathcal{D}^{nuis}) < \infty$. *Then*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|}{\sqrt{n}}\right)$$

Lemma 3.5.2 (Double robustness)**.** *Let $f : \mathcal{W} \mapsto \mathbb{R}^p$ for any p be a function with $\|f(W)\| \leq M < \infty$ for some M. Under Assumption 5 (positivity),*

$$\|f(W)(\widehat{\phi} - \phi)\| = O_{\mathbb{P}}\left(\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\|\right) \tag{3.47}$$

$$\|f(W)(\widehat{\underline{\phi}} - \underline{\phi})\| = O_{\mathbb{P}}\left(\|\nu_0 - \widehat{\nu}_0\|\|\widehat{\pi} - \pi\|\right) \tag{3.48}$$

It follows immediately that

$$\mathbb{P}\left(f(W)(\widehat{\phi} - \phi)\right) = O_{\mathbb{P}}\left(\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\|\right)$$
$$\mathbb{P}\left(f(W)(\widehat{\underline{\phi}} - \underline{\phi})\right) = O_{\mathbb{P}}\left(\|\nu_0 - \widehat{\nu}_0\|\|\widehat{\pi} - \pi\|\right)$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(f(W)(\widehat{\phi} - \phi)\right) &= \mathbb{P}\left(f(W)\left(\frac{1-D}{1-\widehat{\pi}}(Y - \widehat{\mu}_0) + \widehat{\mu}_0 - \frac{1-D}{1-\pi}(Y - \mu_0) - \mu_0\right)\right) \\
&= \mathbb{P}\left(f(W)\left(\frac{1-D}{1-\widehat{\pi}}(\mu_0 - \widehat{\mu}_0) + \widehat{\mu}_0 - \frac{1-D}{1-\pi}(\mu_0 - \mu_0) - \mu_0\right)\right) \\
&= \mathbb{P}\left(f(W)\left(\frac{1-\pi}{1-\widehat{\pi}}(\mu_0 - \widehat{\mu}_0) + \widehat{\mu}_0 - \mu_0\right)\right) \\
&= \mathbb{P}\left(f(W)\left(\frac{(\mu_0 - \widehat{\mu}_0)(\widehat{\pi} - \pi)}{1-\pi}\right)\right) \\
&\leq \frac{1}{\delta}\mathbb{P}(f(W)(\mu_0 - \widehat{\mu}_0)(\widehat{\pi} - \pi)) \\
&\leq \frac{1}{\delta}\|f(W)\|\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\| \\
&= O_{\mathbb{P}}\left(\|\mu_0 - \widehat{\mu}_0\|\|\widehat{\pi} - \pi\|\right)
\end{aligned}$$

where the second and third lines use iterated expectation, conditioning on W; the fifth line uses Assumption 5 (positivity); and the sixth line uses the Cauchy-Schwarz inequality. □

3.B Proof of Theorem 3.5

We prove this theorem first, since the result will be used in the proofs of the other theorems. In the observable setting, the theorem follows immediately from the central limit theorem, so the subsequent derivations focus on the counterfactual setting.

3.B.1 Asymptotic normality of the risk estimator

$$\mathbb{P}_n[f_\beta^2 - (2b^T\beta)\widehat{\phi} + \widehat{\underline{\phi}}] - \mathbb{P}[f_\beta^2 - (2b^T\beta)\phi + \underline{\phi}] = \tag{3.49}$$

$$(\mathbb{P}_n - \mathbb{P})\left\{f_\beta^2 - 2f_\beta\phi + \underline{\phi}\right\} + (\mathbb{P}_n - \mathbb{P})\left\{2f_\beta(\phi - \widehat{\phi}) + (\widehat{\underline{\phi}} - \underline{\phi})\right\} + \mathbb{P}\left\{2f_\beta(\phi - \widehat{\phi}) + (\widehat{\underline{\phi}} - \underline{\phi})\right\} \tag{3.50}$$

The second term is $O_{\mathbb{P}}(\|\widehat{\mu}_0 - \mu_0\|\|\widehat{\pi} - \pi\|/\sqrt{n}) = o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.1, Lemma 3.5.2, and Assumption 9. The third term is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.2 and Assumption 9. We therefore have

$$\widehat{\text{Risk}}(f_\beta) - \text{Risk}(f_\beta) = (\mathbb{P}_n - \mathbb{P})\left\{f_\beta^2 - 2f_\beta\phi + \underline{\phi}\right\} + o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.51}$$

and the result follows by the central limit theorem.

3.B.2 Asymptotic normality of the unfairness estimators

Since g^{rate} does not depend on the outcome, we have $\widehat{g}^{\text{ind}} = g^{\text{rate}}$, and the result follows immediately from the central limit theorem. We now prove the result for g^{FPR} in the counterfactual setting. We have

$$\mathbb{P}_n(\widehat{g}_j f_\beta) - \mathbb{P}(g_j f_\beta) = \left\{\frac{\mathbb{P}_n[\widehat{\gamma}_0 f_\beta]}{\mathbb{P}_n[\widehat{\gamma}_0]} - \frac{\mathbb{P}_n[\widehat{\gamma}_1 f_\beta]}{\mathbb{P}_n[\widehat{\gamma}_1]}\right\} - \left\{\frac{\mathbb{P}[\gamma_0 f_\beta]}{\mathbb{P}[\gamma_0]} - \frac{\mathbb{P}[\gamma_1 f_\beta]}{\mathbb{P}[\gamma_1]}\right\} \tag{3.52}$$

Considering just the $\widehat{\gamma}_0$ and γ_0 terms, we have

$$\begin{aligned}
&\frac{\mathbb{P}_n[\widehat{\gamma}_0 f_\beta]}{\mathbb{P}_n[\widehat{\gamma}_0]} - \frac{\mathbb{P}[\gamma_0 f_\beta]}{\mathbb{P}[\gamma_0]} = \frac{\mathbb{P}_n[\widehat{\gamma}_0 f_\beta]\mathbb{P}[\gamma_0] - \mathbb{P}[\gamma_0]\mathbb{P}_n[\widehat{\gamma}_0]}{\mathbb{P}_n[\widehat{\gamma}_0]\mathbb{P}[\gamma_0]} \\
&= \frac{\mathbb{P}[\gamma_0]\Big(\mathbb{P}_n[\widehat{\gamma}_0 f_\beta] - \mathbb{P}[\gamma_0 f_\beta]\Big) - \mathbb{P}[\gamma_0 f_\beta]\Big(\mathbb{P}_n[\widehat{\gamma}_0] - \mathbb{P}[\gamma_0]\Big)}{\mathbb{P}_n[\widehat{\gamma}_0]\mathbb{P}[\gamma_0]} \\
&= \mathbb{P}_n[\widehat{\gamma}_0]^{-1}\Big\{\underbrace{(\mathbb{P}_n[\widehat{\gamma}_0 f_\beta] - \mathbb{P}[\gamma_0 f_\beta])}_{(1)} - \frac{\mathbb{P}[\gamma_0 f_\beta]}{\mathbb{P}[\gamma_0]}\underbrace{(\mathbb{P}_n[\widehat{\gamma}_0] - \mathbb{P}[\gamma_0])}_{(2)}\Big\}
\end{aligned} \tag{3.53}$$

Terms (1) and (2) in (3.53) can be expanded as follows:

$$\begin{aligned}
(1) &= (\mathbb{P}_n - \mathbb{P})\gamma_0 f_\beta + (\mathbb{P}_n - \mathbb{P})((\widehat{\gamma}_0 - \gamma_0)f_\beta) + \mathbb{P}((\widehat{\gamma}_0 - \gamma_0)f_\beta) \\
(2) &= (\mathbb{P}_n - \mathbb{P})\gamma_0 + (\mathbb{P}_n - \mathbb{P})(\widehat{\gamma}_0 - \gamma_0) + \mathbb{P}(\widehat{\gamma}_0 - \gamma_0)
\end{aligned}$$

In both these expressions, the second term is $O_{\mathbb{P}}(\|\widehat{\phi} - \phi\|/\sqrt{n}) = o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.1 and Assumption 9, and the third term is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.2 and Assumption 9. Under Assumption 7, $\mathbb{P}_n[\widehat{\gamma}_0]^{-1}$ is bounded, while $\mathbb{P}[\gamma_0 f_\beta]/\mathbb{P}[\gamma_0]$ is bounded under Assumption 5. Therefore,

we can rewrite (3.52) as

$$\mathbb{P}_n[\widehat{\gamma}_0]^{-1}(\mathbb{P}_n - \mathbb{P}) \left\{ \gamma_0 \left(f_\beta - \frac{\mathbb{P}[\gamma_0 f_\beta]}{\mathbb{P}[\gamma_0]} \right) \right\} + o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.54}$$

$$= \mathbb{P}_n[\widehat{\gamma}_0]^{-1}(\mathbb{P}_n - \mathbb{P})\eta_0 + o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.55}$$

We can therefore rewrite $\mathbb{P}_n(\widehat{g}_j f_\beta) - \mathbb{P}(g_j f_\beta)$ as

$$\mathbb{P}_n[\widehat{\gamma}_0]^{-1}(\mathbb{P}_n - \mathbb{P})\eta_0 + \mathbb{P}_n[\widehat{\gamma}_1]^{-1}(\mathbb{P}_n - \mathbb{P})\eta_1 + o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.56}$$

Note that the analysis of term (2) in (3.53) yields that $\mathbb{P}_n[\widehat{\gamma}_0] - \mathbb{P}[\gamma_0] = o_{\mathbb{P}}(1)$. Applying the central limit theorem to the vector (η_0, η_1), followed the continuous mapping theorem, Slutsky's theorem, and the delta method, we have

$$\sqrt{n}\left(\mathbb{P}_n(\widehat{g}_j f_\beta) - \mathbb{P}(g_j f_\beta)\right) \xrightarrow{d} N\left(0, \operatorname{var}\left(\mathbb{P}(\gamma_0)^{-1}\eta_0 - \mathbb{P}(\gamma_1)^{-1}\eta_1\right)\right) \tag{3.57}$$

as desired. The result for g^{FNR} follows by identical reasoning.

3.C Proofs for the constrained setting

We state two additional lemmas that are only used in the constrained setting. The first lemma gives sufficient conditions under which the optimal value of an estimated convex problem converges at a particular rate to the optimal value of the target convex program. It is a adaptation of Theorem 3.5 in Shapiro (1991) that follows immediately from Theorems 2.1 and 3.4 in that same paper.

Lemma 3.5.3. *(Shapiro, 1991) Let Θ be a compact subset of $\mathbb{R}^k$. Let $C(\Theta)$ denote the set of continuous real-valued functions on Θ, with $\mathcal{L} = C(\Theta) \times \ldots \times C(\Theta)$ the r-dimensional Cartesian product. Let $\psi(\theta) = (\psi_0, \ldots, \psi_r) \in \mathcal{L}$ be a vector of convex functions. Consider the quantity α^* defined as the solution to the following convex optimization program:*

$$\alpha^* = \min_{\theta \in \Theta} \quad \psi_0(\theta)$$
$$\text{subject to } \psi_j(\theta) \leq 0, \ j = 1, \ldots, r$$

Assume that Slater's condition holds, so that there is some $\theta \in \Theta$ for which the inequalities are satisfied and non-affine inequalities are strictly satisfied, i.e. $\psi_j(\theta) < 0$ if ψ_j is non-affine. Now

consider a sequence of approximating programs, for $n = 1, 2, \ldots$:

$$\begin{aligned} \widehat{\alpha}_n = \min_{\theta \in \Theta} \quad & \widehat{\psi}_{0n}(\theta) \\ \textit{subject to } & \widehat{\psi}_{jn}(\theta) \leq 0, \ j = 1, \ldots, r \end{aligned}$$

with $\widehat{\psi}_n(\theta) := \left(\widehat{\psi}_{0n}, \ldots, \widehat{\psi}_{rn}\right) \in \mathcal{L}$. Assume that $f(n)(\widehat{\psi}_n - \psi)$ converges in distribution to a random element $W \in \mathcal{L}$ for some real-valued function $f(n)$. Then:

$$f(n)(\widehat{\alpha}_n - \alpha_0) \rightsquigarrow L$$

for a particular random variable L. It follows that $\widehat{\alpha}_n - \alpha_0 = O_{\mathbb{P}}(1/f(n))$.

3.C.1 Intermediate result

The next lemma applies Lemma 3.5.3 to the *risk-min* and *unfair-min* settings. For analytical purposes, we suppose that for each k, the quantities $\beta_r^*, \widehat{\beta}_r, \beta_u^*, \widehat{\beta}_u$ are constrained to lie in some (arbitrarily large) compact set $\Theta_k \subseteq \mathbb{R}^k$. Since $k \not\to \infty$, ultimately Θ_k is fixed to some set Θ. For example, Θ could be given by box constraints defined by the largest and smallest numbers the machine can represent. Since this is a device for asymptotic analysis, we do not express it in the actual optimization. Under Assumption 2, it follows that $b^T\beta$ is uniformly bounded in Θ. (Recall that in practice, the output of any predictor will be truncated to lie in $[\ell_y, u_y]$.)

Per Proposition 6 and Remark 5, we can write the objective function for the *risk-min* parameter β_r^* equivalently as $\mathbb{P}[(b^T\beta)^2 - 2(b^T\beta)\phi]$ in the counterfactual setting (or $\mathbb{P}[(b^T\beta)^2 - 2(b^T\beta)Y]$ in the observable setting), since the term $\mathbb{P}[\phi^2]$ (or $\mathbb{P}[Y^2]$) drops out of the minimization. We utilize this form for analysis.

Denote by $\psi_0, \ldots \psi_{t+1}$ and $\widehat{\psi}_0, \ldots \widehat{\psi}_{t+1}$ the population and empirical risk and unfairness functions, each of which is a mapping from Θ to $\mathbb{R}$. For the counterfactual setting, these are given by

$$\begin{aligned} &\psi_0(\beta) = \mathbb{P}[(b^T\beta)^2 - 2(b^T\beta)\phi] && \widehat{\psi}_0 = \mathbb{P}_n[(b^T\beta)^2 - 2(b^T\beta)\phi)] \\ &\psi_j(\beta) = (\mathbb{P}[g_j b^T\beta)])^2 && \widehat{\psi}_j(\beta) = (\mathbb{P}_n[\widehat{g}_j b^T\beta])^2, \qquad j = 1, \ldots t \\ &\psi_{t+1}(\beta) = \mathbb{P}[(b^T\beta)^2 - (2b^T\beta)\phi + \underline{\phi}] && \psi_{t+1}(\beta) = \mathbb{P}_n[(b^T\beta)^2 - (2b^T\beta)\widehat{\phi} + \widehat{\underline{\phi}}] \end{aligned} \tag{3.58}$$

The observable setting substitutes Y for ϕ, Y^2 for $\underline{\phi}$, and g_j for $\widehat{g}_j$. Let $\mathcal{C}(\Theta)$ denote the set of continuous real-valued functions on Θ, with $\mathcal{L}(\Theta) = \mathcal{C}(\Theta) \times \ldots \times \mathcal{C}(\Theta)$ the Cartesian product (with suitable dimension). Let $\psi_\bullet, \widehat{\psi}_\bullet : \Theta \mapsto \mathcal{L}(\Theta)$ be the vectors of functions that define the population

and empirical optimization problem, for $\bullet \in \{r, u\}$ That is, for *risk-min*, define

$$\psi_r = \big(\psi_0(\beta),\ \psi_1(\beta), \ldots, \psi_t(\beta)\big)^T \tag{3.59}$$

$$\widehat{\psi}_r = \Big(\widehat{\psi}_0(\beta),\ \widehat{\psi}_1(\beta), \ldots, \widehat{\psi}_t(\beta)\Big)^T \tag{3.60}$$

and for *unfair-min*, define

$$\psi_u = \left(\sum_{j=1}^{t} \alpha_j \psi_j(\beta),\ \psi_{t+1}(\beta)\right)^T \tag{3.61}$$

$$\widehat{\psi}_u = \left(\sum_{j=1}^{t} \alpha_j \widehat{\psi}_j(\beta),\ \widehat{\psi}_{t+1}(\beta)\right)^T \tag{3.62}$$

The first element in each of ψ_r and ψ_u is the objective function, and the remaining elements are the constraint functions.

Lemma 3.5.4 (Convergence rates of estimated functions)**.** *Under Assumptions 1 and 2 for the observable setting, and Assumptions 1, 4–6 for the counterfactual setting, there exist random elements C_r, C_u taking values in the appropriate space $\mathcal{L}(\Theta)$ such that*

$$\sqrt{n}(\widehat{\psi}_r - \psi_r) \xrightarrow{d} C_r \tag{3.63}$$

$$\sqrt{n}(\widehat{\psi}_u - \psi_u) \xrightarrow{d} C_u \tag{3.64}$$

where the convergence is in L_2 norm.

Proof. We will utilize the fact that the class $\{b^T\beta : \beta \in \Theta\}$ is $\mathbb{P}$-Donsker, since $b^T\beta$ is parametric and Lipschitz in β under Assumption 2.

In the observable setting, we have

$$\widehat{\psi}_r - \psi_r = (\mathbb{P}_n - \mathbb{P})\left((b^T\beta)^2 - 2(b^T\beta)Y,\ g_1 b^T\beta,\ \ldots\ g_t b^T\beta\right)^T \tag{3.65}$$

so that the result follows immediately from the central limit theorem and the Donsker condition. We now turn to the counterfactual setting. First, consider the objective function ψ_0.

$$\widehat{\psi}_0(\beta) - \psi_0(\beta) = \mathbb{P}_n\left\{(b^T\beta)^2 - 2(b^T\beta)\widehat{\phi}\right\} - \mathbb{P}\left\{(b^T\beta)^2 - 2(b^T\beta)\phi\right\} \tag{3.66}$$

$$= (\mathbb{P}_n - \mathbb{P})\left\{(b^T\beta)^2\right\} - \left\{\mathbb{P}_n(b^T\beta\widehat{\phi}) - \mathbb{P}(b^T\beta\phi)\right\} \tag{3.67}$$

$$= (\mathbb{P}_n - \mathbb{P})\left\{(b^T\beta)^2 - \phi\right\} + (\mathbb{P}_n - \mathbb{P})(2(b^T\beta)(\phi - \widehat{\phi})) + \mathbb{P}(2(b^T\beta)(\phi - \widehat{\phi})) \tag{3.68}$$

The second term is $O_{\mathbb{P}}(\|\widehat{\mu}_0 - \mu_0\|\|\widehat{\pi} - \pi\|/\sqrt{n}) = o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.1, Lemma 3.5.2, and Assumption 9. The third term is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.2 and Assumption 9. We therefore have

$$\widehat{\psi}_0(\beta) - \psi_0(\beta) = (\mathbb{P}_n - \mathbb{P})\left\{(b^T\beta)^2 - \phi\right\} + o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.69}$$

We now consider the unfairness functions $\psi_j, j = 1, \dots t$. We have

$$\widehat{\psi}_j(\beta) - \psi_j(\beta) = \left\{\mathbb{P}_n(\widehat{g}_j b^T\beta) + \mathbb{P}(g_j b^T\beta)\right\}\left\{\mathbb{P}_n(\widehat{g}_j b^T\beta) - \mathbb{P}(g_j b^T\beta)\right\} \tag{3.70}$$

$$= \left\{\mathbb{P}_n(\widehat{g}_j b^T\beta) + \mathbb{P}(g_j b^T\beta)\right\}\left(\mathbb{P}_n(\widehat{\gamma}_0)^{-1},\ \mathbb{P}_n(\widehat{\gamma}_1)^{-1}\right)(\mathbb{P}_n - \mathbb{P})\begin{pmatrix}\eta_0\\ \eta_1\end{pmatrix} + o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.71}$$

where the second line follows the derivation in Section 3.B.2, coupled with the fact that $\mathbb{P}_n(\widehat{g}_j b^T\beta) + \mathbb{P}(g_j b^T\beta) = o_{\mathbb{P}}(1)$. Finally, the analysis of ψ_{t+1} is already given in Section 3.B.1:

$$\widehat{\psi}_{t+1} - \psi_{t+1} = (\mathbb{P}_n - \mathbb{P})\left((b^T\beta)^2 - 2(b^T\beta)\phi + \underline{\phi}\right) + o_{\mathbb{P}}(1/\sqrt{n}) \tag{3.72}$$

Suppose we have a single fairness function g_j. Combining (3.69), (3.71), and (3.72), we have shown that $\widehat{\psi}_r - \psi_r$ can be written as

$$\widehat{\psi}_r - \psi_r = M(\mathbb{P}_n - \mathbb{P})\begin{pmatrix}(b^T\beta)^2 - \phi\\ \eta_0\\ \eta_1\end{pmatrix} + \begin{pmatrix}o_{\mathbb{P}}(1/\sqrt{n})\\ o_{\mathbb{P}}(1/\sqrt{n})\end{pmatrix}, \text{ where} \tag{3.73}$$

$$M = \begin{bmatrix}1 & 0 & 0\\ 0 & \left\{\mathbb{P}_n(\widehat{g}_j b^T\beta) + \mathbb{P}(g_j b^T\beta)\right\}\mathbb{P}_n(\widehat{\gamma}_0)^{-1} & -\left\{\mathbb{P}_n(\widehat{g}_j b^T\beta) + \mathbb{P}(g_j b^T\beta)\right\}\mathbb{P}_n(\widehat{\gamma}_1)^{-1}\end{bmatrix} \tag{3.74}$$

Applying the central limit theorem, Slutsky's theorem, the continuous mapping theorem, and the delta method, we have that $\sqrt{n}(\widehat{\psi}_r(\beta) - \psi_r(\beta))$ converges to a normal distribution for any fixed β. Under the Donsker condition, this convergence is uniform over β, and $\sqrt{n}(\widehat{\psi}_r - \psi_r)$ converges to a Gaussian process. Equivalent reasoning applies in the case of multiple fairness functions, and to $\sqrt{n}(\widehat{\psi}_u - \psi_u)$. □

We now prove Theorems 3.1 and 3.2. The two proofs proceed along similar lines. We will again utilize the fact that $\{b^T\beta : \beta \in \Theta\}$ is $\mathbb{P}$-Donsker, so that the empirical process $\{\sqrt{n}(\mathbb{P}_n - \mathbb{P})(b^T\beta) : \beta \in \Theta\}$ converges to a Gaussian process.

3.C.2 Proof of Theorem 3.1 (Excess risk in the constrained setting)

Proof. We consider the *risk-min* problem first. We expand the excess risk by adding and subtracting the objective function at the solution $\widehat{\beta}_r$:

$$\mathbb{P}\left[(b^T\widehat{\beta}_r)^2 - 2(b^T\widehat{\beta}_r)\widehat{\phi}\right] - \mathbb{P}\left[(b^T\beta_r^*)^2 - 2(b^T\beta_r^*)\phi\right] \tag{3.75}$$

$$= \mathbb{P}\left[(b^T\widehat{\beta}_r)^2 - 2(b^T\widehat{\beta}_r)\widehat{\phi}\right] - \mathbb{P}_n\left[(b^T\widehat{\beta}_r)^2 - 2(b^T\widehat{\beta}_r)\widehat{\phi}\right] + \tag{3.76}$$

$$\mathbb{P}_n\left[(b^T\widehat{\beta}_r)^2 - 2(b^T\widehat{\beta}_r)\widehat{\phi}\right] - \mathbb{P}\left[(b^T\beta_r^*)^2 - 2(b^T\beta_r^*)\phi\right] \tag{3.77}$$

The second term is $O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.4 and Shapiro's theorem. The first term is just $\psi_0(\widehat{\beta}_r) - \widehat{\psi}_0(\widehat{\beta}_r)$, which is $O_{\mathbb{P}}(1/\sqrt{n})$ by (3.69) in the proof of Lemma 3.5.4 coupled with the Donsker condition. Hence, the excess risk is $O_{\mathbb{P}}(1/\sqrt{n})$, as claimed.

We now turn to the *unfair-min* problem. The excess risk is

$$\mathbb{P}[(b^T\widehat{\beta}_u)^2 - 2(b^T\widehat{\beta}_u)\phi + \underline{\phi})] - \epsilon^2 \tag{3.78}$$

$$\leq \mathbb{P}[(b^T\widehat{\beta}_u)^2 - 2(b^T\widehat{\beta}_u)\phi + \underline{\phi})] - \mathbb{P}_n[(b^T\widehat{\beta}_u)^2 - 2(b^T\widehat{\beta}_u)\widehat{\phi} + \widehat{\underline{\phi}})] \tag{3.79}$$

$$= -(\mathbb{P}_n - \mathbb{P})\left[(b^T\widehat{\beta}_u)^2 + (2b^T\widehat{\beta}_u)\phi + \underline{\phi}\right] + \tag{3.80}$$

$$(\mathbb{P}_n - \mathbb{P})\left[(2b^T\widehat{\beta}_u)(\widehat{\phi} - \phi) + (\widehat{\underline{\phi}} - \underline{\phi})\right] + \tag{3.81}$$

$$\mathbb{P}\left[(2b^T\widehat{\beta}_u)(\widehat{\phi} - \phi) + (\widehat{\underline{\phi}} - \underline{\phi})\right] \tag{3.82}$$

The first term is $O_{\mathbb{P}}(1/\sqrt{n})$ by the central limit theorem and the Donsker condition. The second term is $O_{\mathbb{P}}(\|\widehat{\phi} - \phi\|/\sqrt{n}) = o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.1, Lemma 3.5.2, and Assumption 8. The last term is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.2. The excess risk is therefore $O_{\mathbb{P}}(1/\sqrt{n})$, as claimed. □

3.C.3 Proof of Theorem 3.2 (Excess unfairness in the constrained setting)

Proof. We consider the *unfair-min* problem first. We expand the excess unfairness by adding and subtracting the objective function at the solution $\widehat{\beta}_u$:

$$\sum_{j=1}^{t} \alpha_j (\mathbb{P}[g_j b^T\widehat{\beta}_u])^2 - \sum_{j=1}^{t} \alpha_j (\mathbb{P}[g_j b^T\beta_u^*])^2 \tag{3.83}$$

$$= \sum_{j=1}^{t} \alpha_j \left\{(\mathbb{P}[g_j b^T\widehat{\beta}_u])^2 - (\mathbb{P}_n[\widehat{g}_j b^T\widehat{\beta}_u])^2\right\} + \sum_{j=1}^{t} \alpha_j \left\{(\mathbb{P}_n[\widehat{g}_j b^T\widehat{\beta}_u])^2 - (\mathbb{P}[g_j b^T\beta_u^*])^2\right\} \tag{3.84}$$

Again, the second term is $O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.4 and Shapiro's theorem. The first term is equal to $\sum_{j=1}^{t} \alpha_j(\psi_j(\widehat{\beta}) - \widehat{\psi}_j(\widehat{\beta}))$, which is $O_{\mathbb{P}}(1/\sqrt{n})$ by (3.71) in the proof of Lemma 3.5.4 coupled with the Donsker condition. The excess unfairness is therefore $O_{\mathbb{P}}(1/\sqrt{n})$, as claimed.

We now turn to the *risk-min* problem. The excess unfairness for constraint j is

$$(\mathbb{P}[g_j b^T \widehat{\beta}_u])^2 - \epsilon^2 \tag{3.85}$$

$$\leq (\mathbb{P}[g_j b^T \widehat{\beta}_u])^2 - (\mathbb{P}_n[\widehat{g}_j b^T \widehat{\beta}_u])^2 \tag{3.86}$$

$$= O_{\mathbb{P}}(1/\sqrt{n}) \tag{3.87}$$

where the last line simply uses the analysis for term (1) from (3.84). The excess unfairness is therefore $O_{\mathbb{P}}(1/\sqrt{n})$, as claimed. □

3.D Proofs for the penalized setting

Throughout this section, let

$$\widehat{\mathbf{Q}}_\lambda = \mathbb{P}_n(bb^T) + \sum_{j=1}^{t} \lambda_j \mathbb{P}_n(\widehat{g}_j b)\mathbb{P}_n(\widehat{g}_j b)^T \tag{3.88}$$

$$\mathbf{Q}_\lambda = \mathbb{P}(bb^T) + \sum_{j=1}^{t} \lambda_j \mathbb{P}(g_j b)\mathbb{P}(g_j b)^T \tag{3.89}$$

so that

$$\beta_\lambda = \mathbf{Q}_\lambda^{-1} \mathbb{P}(b\widetilde{Y}) \tag{3.90}$$

$$\widehat{\beta}_\lambda = \begin{cases} \widehat{\mathbf{Q}}_\lambda^{-1} \mathbb{P}_n(bY) & \text{(Observable)} \\ \widehat{\mathbf{Q}}_\lambda^{-1} \mathbb{P}_n(b\widehat{\phi}) & \text{(Counterfactual)} \end{cases} \tag{3.91}$$

Under the assumptions of Theorems 3.3 and 3.4, we prove several preliminary results that are used in the theorem proofs.

Lemma 3.5.5 (Bounded output)**.** $\sup_{\lambda \in \Lambda} \|\beta_\lambda^*\| < \infty$, *and* $\sup_{w \in \mathcal{W}, \lambda \in \Lambda} |b(w)^T \beta_\lambda^*| < \infty$.

Proof. The eigenvalues of $\lambda_j \mathbb{P}(bg_j)\mathbb{P}(bg_j)^T$ are 0, with multiplicity $k-1$, and $\lambda_j \mathbb{P}(bg_j)^T \mathbb{P}(bg_j)$, which is uniformly bounded above due to Assumptions 2 and 11 and the fact that g_j is bounded. By Weyl's inequality, this means that the eigenvalues of $\mathbf{Q}_\lambda$ are uniformly bounded above and away from 0. Since $\widetilde{Y}$ is also bounded, the claims follow by Cauchy-Schwarz. □

Lemma 3.5.6. *(Vector norm with nuisance parameter). Let $\eta(Z)$ be a bounded function (namely, ϕ or g_j). Then*

$$\|\mathbb{P}_n(b\widehat{\eta}) - \mathbb{P}(b\eta)\| = O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}(\|\mathbb{P}(\widehat{\eta} - \eta)\|) \tag{3.92}$$

Proof. First, note that

$$\begin{aligned}
\|\mathbb{P}(b(\widehat{\eta} - \eta))\| &\leq \|\sqrt{\mathbb{P}(b^2)\mathbb{P}((\widehat{\eta} - \eta)^2)}\| && \text{(Cauchy-Schwarz)} \\
&\leq \|\sqrt{\mathbb{P}(b)^2(\mathbb{P}(\widehat{\eta} - \eta))^2}\| && \text{(Jensen's)} \\
&= \|\mathbb{P}(b)\mathbb{P}(\widehat{\eta} - \eta)\| && (3.93) \\
&= \|\mathbb{P}(b)\| \cdot \|\mathbb{P}(\widehat{\eta} - \eta)\| && (3.94) \\
&= O_{\mathbb{P}}\left(\|\mathbb{P}(\widehat{\eta} - \eta)\|\right) && \text{(Assumption 2)}
\end{aligned}$$

Now, expanding 3.92, we have

$$\begin{aligned}
\mathbb{P}_n(b\widehat{\eta}) - \mathbb{P}(b\eta) &= (\mathbb{P}_n - \mathbb{P})(b\eta) + (\mathbb{P}_n - \mathbb{P})(b\widehat{\eta} - b\eta) + \mathbb{P}(b\widehat{\eta} - b\eta) && (3.95) \\
\implies \|\mathbb{P}_n(b\eta) - \mathbb{P}(b\eta)\| &= O_{\mathbb{P}}(\sqrt{1/n}) + o_{\mathbb{P}}(\sqrt{1/n}) + \|\widehat{\eta} - \eta\| && (3.96) \\
&= O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}(\|\mathbb{P}(\widehat{\eta} - \eta)\|) && (3.97)
\end{aligned}$$

□

Lemma 3.5.7. *(Bounded norm for $\widehat{\mathbf{Q}}_\lambda^{-1}$).*

$$\mathbb{P}(\|\widehat{\mathbf{Q}}_\lambda^{-1}\|) \leq C \to 1 \text{ for some constant } C \tag{3.98}$$

Proof. This follows from Assumption 1 plus the consistency of $\widehat{\mathbf{Q}}_\lambda$ for $\mathbf{Q}_\lambda$. It follows that $\|\widehat{\mathbf{Q}}_\lambda^{-1}\| = O_{\mathbb{P}}(1)$. □

Lemma 3.5.8. *Fix a $\lambda \in \Lambda$. Then*

$$\|\widehat{\beta}_\lambda - \beta_\lambda^*\| = \begin{cases} O_{\mathbb{P}}(\sqrt{1/n}) & \textit{(Observable)} \\ O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}\left(h(n)\right) & \textit{(Counterfactual)} \end{cases} \tag{3.99}$$

Proof. In the observable setting, we have

$$\widehat{\beta}_\lambda - \beta^*_\lambda = (\widehat{\mathbf{Q}}_\lambda^{-1} - \mathbf{Q}_\lambda^{-1})\mathbb{P}(bY) \tag{3.100}$$

$$= \widehat{\mathbf{Q}}_\lambda^{-1}(\mathbf{Q}_\lambda - \widehat{\mathbf{Q}}_\lambda)\mathbf{Q}^{-1}\mathbb{P}(bY) \tag{3.101}$$

$$= \widehat{\mathbf{Q}}_\lambda^{-1}(\mathbf{Q}_\lambda - \widehat{\mathbf{Q}}_\lambda)\beta^*_\lambda \tag{3.102}$$

$$= \widehat{\mathbf{Q}}_\lambda^{-1}\Big\{(\mathbb{P}_n - \mathbb{P})(bb^T\beta^*_\lambda) + \sum_{j=1}^{t}\left[\lambda_j(\mathbb{P}_n - \mathbb{P})(bg_j)\mathbb{P}(g_jb^T\beta^*_\lambda) + \mathbb{P}_n(bg_j)(\mathbb{P}_n - \mathbb{P})(g_jb^T\beta^*_\lambda)\right]\Big\} \tag{3.103}$$

The norm of each term in the braces is $O_{\mathbb{P}}(1/\sqrt{n})$ by the central limit theorem. By Lemma 3.5.7, $\widehat{\mathbf{Q}}_\lambda^{-1}$ doesn't contribute to the rate, so $\|\widehat{\beta}_\lambda - \beta^*_\lambda\| = O_{\mathbb{P}}(1/\sqrt{n})$ as claimed.

In the counterfactual setting, we have

$$\widehat{\beta}_\lambda - \beta^*_\lambda = \widehat{\mathbf{Q}}_\lambda^{-1}\mathbb{P}_n(b\widehat{\phi}) - \mathbf{Q}_\lambda^{-1}\mathbb{P}(b\phi) \tag{3.104}$$

$$= (\widehat{\mathbf{Q}}_\lambda^{-1} - \mathbf{Q}_\lambda^{-1})\mathbb{P}(b\phi) + \widehat{\mathbf{Q}}_\lambda^{-1}(\mathbb{P}_n(b\widehat{\phi}) - \mathbb{P}(b\phi)) \tag{3.105}$$

$$= \widehat{\mathbf{Q}}_\lambda^{-1}(\mathbf{Q}_\lambda - \widehat{\mathbf{Q}}_\lambda)Q^{-1}\mathbb{P}(b\phi) + \widehat{\mathbf{Q}}_\lambda^{-1}(\mathbb{P}_n(b\widehat{\phi}) - \mathbb{P}(b\phi)) \tag{3.106}$$

$$= \underbrace{\widehat{\mathbf{Q}}_\lambda^{-1}(\mathbf{Q}_\lambda - \widehat{\mathbf{Q}}_\lambda)\beta^*_\lambda}_{(1)} + \underbrace{\widehat{\mathbf{Q}}_\lambda^{-1}(\mathbb{P}_n(b\widehat{\phi}) - \mathbb{P}(b\phi))}_{(2)} \tag{3.107}$$

The norm of term (2) in (3.107) is $O_{\mathbb{P}}(1/\sqrt{n}) + O_{\mathbb{P}}(h(n))$ by Lemma 3.5.2 and Lemma 3.5.7. For term (1), ignoring the leading $\widehat{\mathbf{Q}}_\lambda^{-1}$ for now, we have

$$(\mathbf{Q}_\lambda - \widehat{\mathbf{Q}}_\lambda)\beta^*_\lambda = \underbrace{(\mathbb{P}_n - \mathbb{P})(bb^T\beta^*_\lambda)}_{(a)} + \tag{3.108}$$

$$\sum_{j=1}^{t}\lambda_j\big[\underbrace{(\mathbb{P}_n(b\widehat{g}_j) - \mathbb{P}(bg_j))\mathbb{P}(g_jb^T\beta^*_\lambda)}_{(b)} + \underbrace{\mathbb{P}_n(b\widehat{g}_j)(\mathbb{P}_n(\widehat{g}_jb^T\beta^*_\lambda) - \mathbb{P}(gb^T\beta^*_\lambda)}_{(c)}\big] \tag{3.109}$$

The norm of term (a) is $O_{\mathbb{P}}\left(\sqrt{1/n}\right)$ by the central limit theorem. Terms (b) and (c) decompose as follows:

$$(b) = \mathbb{P}(g_jb^T\beta^*_\lambda)\left\{(\mathbb{P}_n - \mathbb{P})(bg_j) + (\mathbb{P}_n - \mathbb{P})(b(\widehat{g}_j - g_j)) + \mathbb{P}(b(\widehat{g}_j - g_j))\right\} \tag{3.110}$$

$$(c) = \mathbb{P}_n(b\widehat{g}_j)\left\{(\mathbb{P}_n - \mathbb{P})(g_jb^T\beta^*_\lambda) + (\mathbb{P}_n - \mathbb{P})(\widehat{g}_j - g_j)(b^T\beta^*_\lambda) + \mathbb{P}((\widehat{g}_j - g_j)b^T\beta^*_\lambda)\right\} \tag{3.111}$$

The norms of the first term in braces in each of these two expressions is $O_{\mathbb{P}}(1/\sqrt{n})$ by the central limit theorem. The norm of the second term is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 3.5.1 and Assumption 8. The norm of the third term is $O_{\mathbb{P}}(1/\sqrt{n}) + O_{\mathbb{P}}(h(n))$ by Lemma 3.5.2. Using Lemma 3.5.7, the consistency

of $\mathbb{P}_n(b\widehat{g}_j)$ for $\mathbb{P}(bg_j)$, and the boundedness of $\mathbb{P}(g_j b^T \beta^*_\lambda)$, we have

$$\|\widehat{\beta}_\lambda - \beta^*_\lambda\| = O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}\left(h(n)\right) \tag{3.112}$$

as claimed. □

We now prove the two theorems. We will use the fact that under Assumption $\widehat{\beta}_\lambda - \beta^*_\lambda$ is Lipschitz in λ, and Λ is compact, so the set $\{\widehat{\beta}_\lambda - \beta^*_\lambda : \lambda \in \Lambda\}$ is Donsker.

3.D.1 Proof of Theorem 3.3 (Excess risk in the penalized setting)

Fix a $\lambda \in \Lambda$. We have

$$\mathbb{P}\left[\left(b^T\widehat{\beta}_\lambda - \widetilde{Y}\right)^2\right] - \mathbb{P}\left[\left(b^T\beta^*_\lambda - \widetilde{Y}\right)^2\right] = \|b^T\widehat{\beta}_\lambda - \widetilde{Y}\|^2 - \|b^T\beta^*_\lambda - \widetilde{Y}\|^2 \tag{3.113}$$

$$= \left(\|b^T\widehat{\beta}_\lambda - \widetilde{Y}\| - \|b^T\beta^*_\lambda - \widetilde{Y}\|\right)\left(\|b^T\widehat{\beta}_\lambda - \widetilde{Y}\| + \|b^T\beta^*_\lambda - \widetilde{Y}\|\right) \tag{3.114}$$

Since $\widehat{\beta}_\lambda$ is consistent for β^*_λ, the second factor is $O_{\mathbb{P}}(1)$, so we can just consider the first factor.

$$\|b^T\widehat{\beta}_\lambda - \widetilde{Y}\| - \|b^T\beta^*_\lambda - \widetilde{Y}\| \leq \|b^T\widehat{\beta}_\lambda - b^T\beta^*_\lambda\| \tag{3.115}$$

$$= O_{\mathbb{P}}\left(\|\widehat{\beta}_\lambda - \beta^*_\lambda\|\right) \tag{3.116}$$

$$= O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}(h(n)) \tag{3.117}$$

where the first line uses the reverse triangle inequality, the second line uses Assumption 2, and the third line uses Lemma 3.5.8. Under the Donsker condition, the convergence is uniform over Λ:

$$\sup_{\lambda\in\Lambda}\left\{\mathbb{P}\left[\left(b^T\widehat{\beta}_\lambda - \widetilde{Y}\right)^2\right] - \mathbb{P}\left[\left(b^T\beta^*_\lambda - \widetilde{Y}\right)^2\right]\right\} = O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}(h(n)) \tag{3.118}$$

3.D.2 Proof of Theorem 3.4 (Excess unfairness in the penalized setting)

Fix a $\lambda \in \Lambda$. The excess unfairness for g_j is

$$\mathbb{P}\left[g_j b^T \widehat{\beta}_\lambda\right] - \mathbb{P}\left[g_j b^T \beta^*_\lambda\right] = \mathbb{P}[g_j b^T(\widehat{\beta}_\lambda - \beta^*_\lambda)] \tag{3.119}$$

$$\leq \mathbb{P}[|g_j b^T(\widehat{\beta}_\lambda - \beta^*_\lambda)|] \tag{3.120}$$

$$= \sqrt{(\mathbb{P}[|g_j b^T(\widehat{\beta}_\lambda - \beta^*_\lambda)|])^2} \tag{3.121}$$

$$\leq \sqrt{\mathbb{P}[(g_j b^T(\widehat{\beta}_\lambda - \beta^*_\lambda))^2]} \tag{3.122}$$

$$= O_{\mathbb{P}}\left(\sqrt{\mathbb{P}[(\widehat{\beta}_\lambda - \beta^*_\lambda)^2]}\right) \tag{3.123}$$

$$= \|\widehat{\beta}_\lambda - \beta^*_\lambda\| \tag{3.124}$$

$$= O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}(h(n)) \tag{3.125}$$

where the last line uses Lemma 3.5.8. Under the Donsker condition, the convergence is uniform over Λ:

$$\sup_{\lambda\in\Lambda}\left\{\max_{j\in 1,\ldots,t}\left(\mathbb{P}\left[g_j b^T \widehat{\beta}_\lambda\right] - \mathbb{P}\left[g_j b^T \beta^*_\lambda\right]\right)\right\} = O_{\mathbb{P}}(\sqrt{1/n}) + O_{\mathbb{P}}(h(n)) \tag{3.126}$$

3.E Bases with dimension $k \geq n$

We can generalize our estimators slightly to accommodate case where where $k \geq n$, meaning the dimension of the basis is greater than the sample size, as is the case for example with smoothing splines or RKHSs. We simply add an appropriate penalty matrix term $\lambda_0 \beta^T \mathbf{K} \beta$ to the penalized estimator expression or to the objective function for the constrained estimators, where $\mathbf{K}$ is a $k \times k$ smoothing matrix. In the former case, for example, the estimator $\widehat{\beta}_\lambda$ becomes

$$\arg\min_{\beta\in\mathbb{R}^k} \mathbb{P}_n[(b^T\beta - \widehat{\phi})^2] + \lambda_0 \beta^T \mathbf{K} \beta + \sum_{j=1}^{t} \lambda_j (\mathbb{P}_n[\widehat{g}_j b^T \beta])^2 \tag{3.127}$$

$$= \left(\mathbb{P}_n(bb^T) + \lambda_0 \mathbf{K} + \sum_{j=1}^{t} \lambda_j \mathbb{P}_n(\widehat{g}_j b)\mathbb{P}_n(\widehat{g}_j b)^T\right)^{-1} \mathbb{P}_n(b\widehat{\phi}) \tag{3.128}$$

For instance, in a smoothing spline setting, b represents a spline basis, and $\mathbf{K}_{ij} = \int_{\mathcal{W}} b_i''(w) b_j''(w) dw$. In an RKHS, we'd have $b_i = \sum_{j=1}^{n} k(\cdot, w_j)$ and $\mathbf{K}_{ij} = k(w_i, w_j)$. In a ridge setting we'd have $\mathbf{K} = I$. The penalty term ensures the invertibility of the large matrix in (3.128), and it preserves the fast computability of a large set of solutions $\widehat{\beta}_\lambda$.

This penalty term may also be useful to prevent overfitting even in cases where $k < n$, if k is close to n or if the basis is very expressive.

3.F Additional Plots for the Adult data

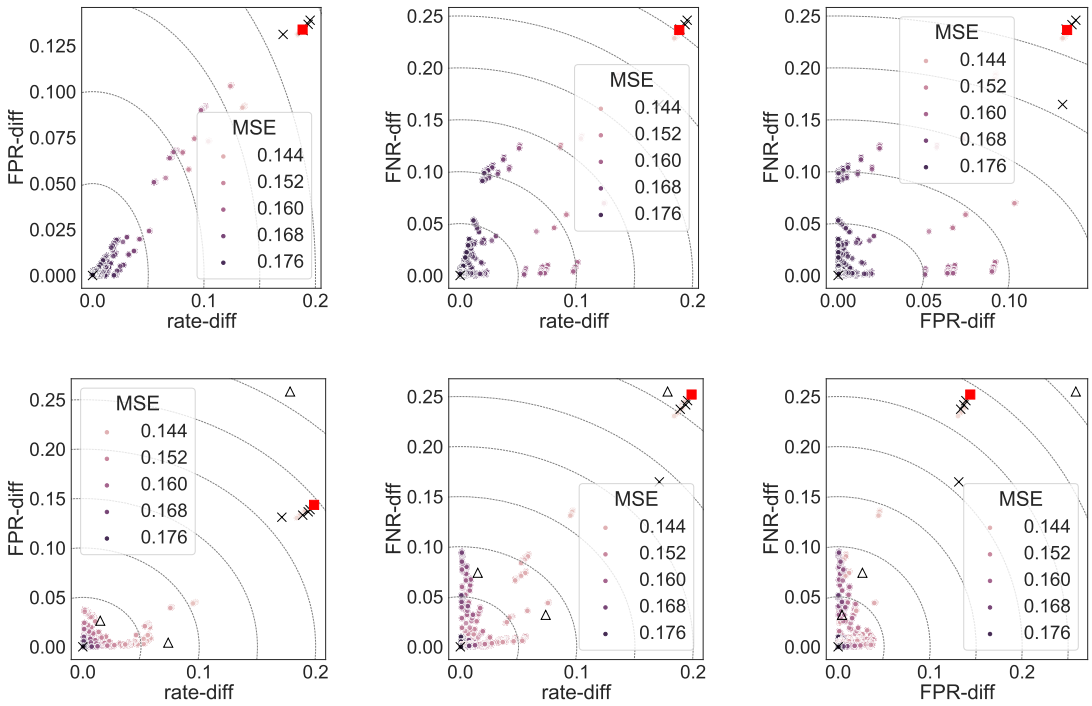


Figure 3.12: Pairs of disparities colored by MSE for the base5 predictors (top row) and base8 predictors (bottom row) for the Adult data. Black 'X's represent the base predictors, the red square is the OLS predictor, and the blue dots are the penalized predictors. Radius lines indicate distance from the origin. Inclusion of the three fair predictors, in the base8 models, improves the tradeoffs relative to the base5 models.

Chapter 4

On the fairness of randomized vs. deterministic classifiers

It is common in the fairness literature to consider *randomized classifiers*, which (implicitly or explicitly) map each individual to a distribution of outcomes rather than to a single outcome (Dwork et al., 2011; Hardt et al., 2016; Agarwal et al., 2018). An outcome for an individual in a particular instance may then be understood as a draw from their distribution. This means, for example, that a defendant who is run through the same classifier twice may be labeled high-risk in one instance and low-risk in the other.

There are at least two reasons for considering randomized classifiers, both practical rather than ethical. The first is that they constitute a larger set of models to search over, since deterministic classifiers are special cases of randomized classifiers with degenerate output distributions. This affords a greater opportunity to find a fair model with nontrivial accuracy. The second reason is that the randomization parameters induce smoothness in various optimization routines. For example, randomization in Chapter 2 translates into a smooth search over a unit hypercube; without randomization, it would not be possible in this post-processing setup to achieve counterfactual equalized odds.

It may seem that this type of randomness is unfair in itself, irrespective of how the classifier behaves in expectation. Consider a post-processed recidivism predictor that randomly flips the outputs of COMPAS. Imagine that a binary COMPAS score is computed and then an administrator reaches into an urn for a marble whose color determines whether the score will be flipped before it

is delivered to the judge.* This procedure seems to violate an intuition that justice should not be random (although see Harcourt (2008) for a contrary point of view).

We argue, however, that this randomness is not categorically different from other types of randomness that are inherent in prediction, and that the more fundamental question is about how to quantify (un)fairness in the first place.

There are numerous sources of randomness or variation that ultimately affect an individual's risk prediction, including the choice of covariates to measure, the samples that are collected for training and testing, the choice of model type (e.g. logistic regression, ridge regression, neural net), and even the model training procedure. For example, the training of a random forest typically involves random bootstrap sampling for each tree and random subsampling of features at each node. All these factors affect the outputs of the resulting predictor, regardless of whether the predictor is deterministic or random. An individual who is labeled low-risk by a given deterministic predictor may have been labeled high-risk by a different deterministic predictor that would have been produced by a slight variation in the training sample.

Even conditional on the data, predictors with a particular performance profile are not generally unique, as the results of Chapter 3 show. This phenomenon has been observed outside the context of fairness, where it has been dubbed the "Rashomon effect" (Breiman, 2001; Fisher et al., 2019). The set of correctly and incorrectly labeled individuals may vary substantially across a set of models that are otherwise equivalent in their (finite sample) accuracy and fairness properties. Which model among this set is surfaced and selected may come down to arbitrary factors like which parts of a parameter space the model builders choose to examine.

We conjecture that most people would say that these sources of randomness are not intrinsically unfair, or at least that they're less unfair than the randomness involved in drawing a marble to determine the final classifier output. Assuming this is true, we further conjecture that the difference in intuition stems from the accessibility of relevant counterfactuals. In the marble drawing case, the scenario decomposes into two stages: first a risk prediction is generated, and then it is randomly flipped, or not. It is easy to imagine the first stage as representing a kind of ground truth from which the randomized predictor then arbitrarily deviates. Furthermore, the effect of the marble drawing on each individual is independent, so it is easy to imagine a counterfactual scenario under which an individual's outcome would have been different: all else being held constant, the administrator would simply have had to draw the other color marble. By contrast, it is harder to concretely visualize how a difference in the training sample might have changed the prediction for a given individual.

The two stage decomposition in the marble drawing scenario is an artifact of the framing, however, which doesn't represent how predictions are generated in practice. Regardless of whether

*This visualization is due to Cosma Shalizi.

a predictor is deterministic or randomized, a judge will only see one score per defendant. The possibility that the defendant would have received a different score remains counterfactual in both cases, even if that counterfactual possibility is easier to conceptualize with the randomized predictor.

Additionally, COMPAS does not represent the ground truth. If the randomized classifier had the same accuracy as COMPAS, then it would be just as likely to flip an incorrect prediction to a correct one as the other way around, in which case it is less obvious whether one predictor should be labeled more fair than the other. If it were more accurate than COMPAS (a possibility discussed in Chapter 3, under Definition 2.5.3), it would be more likely to flip an incorrect prediction to a correct one, in which case it could plausibly be labeled more fair than COMPAS.

If the randomness of a randomized classifier is not categorically different from the randomness of a deterministic classifier, then what else might make the marble drawing scenario unfair? Perhaps it is unfair only if it degrades the accuracy of the "original" predictions. More generally, perhaps it is unfair to choose a predictor that is less accurate than another available predictor, insofar as more people are worse off under the less accurate predictor. Indeed, some authors have asserted that any reduction in predictive accuracy in order to satisfy criteria like equalized odds is both suboptimal, with respect to utility, *and* unfair, with respect to economic notions of unfairness like taste-based discrimination and legal notions like disparate impact[†] (Corbett-Davies et al., 2017; Corbett-Davies and Goel, 2018). These and other authors show how standard fairness constraints can hurt the utility of the groups they're intended to protect, which leads them to question the use of these constraints altogether (Corbett-Davies and Goel, 2018; Hu and Chen, 2020). In place of these constraints, Corbett-Davies and Goel (2018) advocate estimating risk as accurately as possible and focusing fairness efforts on downstream policy interventions, while Rudin et al. (2020) emphasize predictor interpretability and transparency as forms of procedural fairness.

Any time a classifier makes an error–a false positive or false negative–it may lead to harm, regardless of the source of the error, e.g. whether the classifier is deterministic or randomized. Given the impossibility of perfect predictors, we are faced with the task of choosing among imperfect ones, which will end up incorrectly classifying different individuals. While we remain agnostic about how best to do this in general, we argue that the distinction between randomized and deterministic predictors is a second order distinction, and that the more fundamental question is about how to operationalize fairness in the first place. There is the subject of much debate in the algorithmic fairness community, and there is no consensus at present. Arguably, the majority of work in this area is concerned with distributional notions of fairness, which are formalized as functionals on

[†]Although many algorithmic fairness researchers ground their work in legal notions of discrimination like disparate treatment and disparate impact, it is not entirely clear how these terms map onto quantitative fairness criteria in the eyes of the law. Some researchers use variants like *impact disparity* in order to explicitly distinguish the algorithmic criteria from their motivating legal concepts (Lipton et al., 2018).

the joint distribution of covariates, predictions, and outcomes. These criteria are insensitive to predictions at the individual level, and they are "static" in that they do not take into account how predictions are generated. However, fairness in the colloquial sense seems to be a rich and multidimensional construct, and algorithmic fairness will ultimately need to reflect this richness. Distributional fairness criteria will likely have a role to play, alongside other notions of fairness that emphasize how predictors are developed, how they are put to use, how transparent they are, and the relationships between those with the power to develop and deploy them and those who are subject to their predictions.

Bibliography

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 60–69. PMLR. 85, 105

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18:1–78. 82

Angwin, J. and Larson, J. (2016). Bias in criminal risk scores is mathematically inevitable, researchers say. *ProPublica*. 9, 53, 79, 82

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica*. 9, 31, 79

Barocas, S., Hardt, M., and Narayanan, A. (2018). *Fairness and Machine Learning. http://www.fairmlbook.org*. 12, 13, 55

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *Arxiv preprint, arXiv:1810.01943 [cs]*. 85

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *arXiv:1212.0442 [econ, stat]*. 73

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *Arxiv preprint, arXiv:1706.02409 [cs, stat]*. 54, 57

Bickel, P. J., Ritov, C. A. K. Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, Baltimore. 22

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives : Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, 50(3):381–393. 23

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press. 21

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231. 106

Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior*, 36(1):21–40. 1

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR. 53

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE. 55

Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292. 59

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, NIPS 2017, pages 3992–4001. Curran Associates, Inc. 10, 13, 55

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA. ACM. 1

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2020). Classification with fairness constraints: A meta-algorithm with provable guarantees. *Arxiv preprint, arXiv:1806.06055 [cs, stat]*. 85

Chen, J. H. and Asch, S. M. (2017). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *The New England journal of medicine*, 376(26):2507–2509. 2

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68. 23

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163. 13, 14, 34, 53, 57

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 134–148, New York, NY, USA. PMLR. 33, 34

Colubri, A., Silver, T., Fradet, T., Retzepi, K., Fry, B., and Sabeti, P. (2016). Transforming Clinical Data into Actionable Prognosis Models: Machine-Learning Framework and Field-Deployable App to Predict Outcome of Ebola Patients. *PLOS Neglected Tropical Diseases*, 10(3):e0004549. 1

Corbett-Davies, S. and Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *Arxiv preprint, arxiv:1808.00023 [cs]*. 107

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Arxiv preprint, arxiv:1701.08230 [cs]*. 56, 60, 107

Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA. Association for Computing Machinery. 10, 12, 13, 33, 34, 55

Coston, A., Rambachan, A., and Chouldechova, A. (2021). Characterizing fairness over the set of good models under selective labels. *Arxiv preprint, arXiv:2101.00352 [cs, stat]*. 53, 54, 56, 57

Dieterich, W., Mendoza, C., and Brennan, T. (2016). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Technical report, Northpointe. 9

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, NeurIPS 2018, pages 2791–2801. Curran Associates, Inc. 10, 12, 13, 53, 55

Dua, D. and Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml. 85

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., and Varshney, K. R. (2020). Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning*, page 11. 54, 56, 57

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2011). Fairness Through Awareness. *arXiv:1104.3913 [cs]*. arXiv: 1104.3913. 105

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA. Association for Computing Machinery. 55

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81. 106

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 329–338. ACM Press. 53

Glymour, C. and Glymour, M. R. (2014). Commentary: Race and sex are causes. *Epidemiology*, 25(4):488–490. 13

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer. 23

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331. 22

Harcourt, B. E. (2008). *Against Prediction*. University of Chicago Press. 106

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, NIPS 2016, pages 3315–3323. Curran Associates, Inc. 10, 13, 14, 16, 18, 36, 55, 59, 105

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):968. 11, 12, 55

Hu, L. and Chen, Y. (2020). Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 535–545, New York, NY, USA. Association for Computing Machinery. 107

Hu, L. and Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 513, New York, NY, USA. Association for Computing Machinery. 13

Imai, K. and Jiang, Z. (2020). Principal fairness for human and algorithmic decision-making. *Arxiv preprint, arxiv:2006.01770 [cs.CY]*. 13

Jonnson, M. (2018). The influence of risk assessment evidence on judicial sentencing decisions. Master's thesis, Simon Fraser University. 36

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33. 10, 13, 55

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In Flach, P. A., De Bie, T., and Cristianini, N., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524, pages 35–50. Springer Berlin Heidelberg. Series Title: Lecture Notes in Computer Science. 55

Kearns, M., Roth, A., and Wu, Z. S. (2017). Meritocratic fairness for cross-population selection. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1828–1836. PMLR. 12

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In He, H., Wu, P., and Chen, D.-G. D., editors, *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer. 22

Kennedy, E. H., Balakrishnan, S., and G'Sell, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030. 42, 91

Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787. 1

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, NIPS 2017, pages 656–666. Curran Associates, Inc. 10, 13, 56

Kim, J. S., Chen, J., and Talwalkar, A. (2020). FACT: A diagnostic for group fairness trade-offs. In *Proceedings of the 37th International Conference on Machine Learning*, page 11. 53, 54, 57

Kim, M. P., Ghorbani, A., and Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247–254, New York, NY, USA. Association for Computing Machinery. 10, 13, 55

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, 105(5):491–495. 57

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 13, 53, 57

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17. 1

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, NIPS 2017, pages 4066–4076. Curran Associates, Inc. 10, 13, 56

Larson, J. and Angwin, J. (2016). Technical response to northpointe. *ProPublica*. 9

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analayzed the COMPAS Recidivism Algorithm. *ProPublica*. 31

Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ml's impact disparity require treatment disparity? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. 107

Liu, S. and Vicente, L. N. (2021). Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Arxiv preprint, arXiv:2008.01132 [cs, stat]*. 54, 57

Lowenkamp, A. W., Kristin, F., and T., B. C. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks". *Federal Probation*, 80(2):38–46. 9

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19. 2

Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118, New York, NY, USA. PMLR. 12, 53, 56

Mishler, A. (2019). Modeling Risk and Achieving Algorithmic Fairness Using Potential Outcomes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19*, pages 555–556, Honolulu, HI, USA. ACM Press. 2, 82

Mishler, A., Kennedy, E. H., and Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400. ACM. 56

Nabi, R., Malinsky, D., and Shpitser, I. (2019). Learning optimal fair policies. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4674–4682. PMLR. 10, 13, 56

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1931–1940. Association for the Advancement of Artificial Intelligence. 10, 13, 56

Narasimhan, H. (2018). Learning with complex loss functions and constraints. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1646–1654. PMLR. 10, 13

Neyman, J. (1923). Justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. Excerpts english translation (Reprinted). *Statistical Science*, 5:463–472. 11, 55

Northpointe (2015). Practitioners guide to COMPAS core. 82

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453. 10, 53

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On Fairness and Calibration. *Arxiv preprint, arXiv:1709.02012 [cs, stat]*. 55

Povyakalo, A. A., Alberdi, E., Strigini, L., and Ayton, P. (2013). How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography. *Medical Decision Making*, 33(1):98–107. 2

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994. 23

Rice, L. and Swesnik, D. (2012). Discriminatory effects of credit scoring on communities of color. *Suffolk University Law Review*, 46:935. 10

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In Nolan, D. and Speed, T., editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, Institute of Mathematical Statistics Collections, pages 335–421. Institute of Mathematical Statistics, Beachwood, Ohio, USA. 23

Rodolfa, K. T., Lamba, H., and Ghani, R. (2021). Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. *Arxiv preprint, arXiv:2012.02972 [cs]*. 54, 56

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331. 11

Rudin, C., Wang, C., and Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1). 82, 107

Schulam, P. and Saria, S. (2017). Reliable Decision Support using Counterfactual Models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 1697–1708. Curran Associates, Inc. 2

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, page 10. 2

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186. 94

Stevenson, M. (2018). Assessing risk assessment in action. *Minnesota Law Review*, 103(1):83. 36

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York, NY. 22

Tsybakov, A. B. (2003). Optimal rates of aggregation. In Schölkopf, B. and Warmuth, M. K., editors, *Learning Theory and Kernel Machines*, volume 2777, pages 303–313. Springer Berlin Heidelberg. Series Title: Lecture Notes in Computer Science. 60

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, New York, NY. 22

van der Vaart, A. (2002). Semiparametric statistics. In Bernard, P., editor, *Lectures on probability theory and statistics*, number 1781 in Lecture notes in mathematics, Berlin. Springer. 22, 28

VanderWeele, T. J. and Robinson, W. R. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4):473–484. 13

Veale, M., Van Kleek, M., and Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14, Montreal QC, Canada. ACM Press. 2

Wachsmuth, G. (2013). On LICQ and the uniqueness of lagrange multipliers. *Operations Research Letters*, 41(1):78–80. 62

Wang, Y., Sridhar, D., and Blei, D. M. (2019). Equal opportunity and affirmative action via counterfactual predictions. *Arxiv preprint, arxiv:1905.10870 [stat]*. 13, 56

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953, Amsterdam, Netherlands. PMLR. 28, 53, 55, 56

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. 10, 13, 14, 53, 55

Zemel, R. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, page 9. JMLR: W&CP. 55

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM. 85

Zhang, J. and Bareinboim, E. (2018). Fairness in decision-making – the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 2037–2045. Association for the Advancement of Artificial Intelligence. 13, 56

Zhao, H. and Gordon, G. (2019). Inherent tradeoffs in learning fair representations. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 53, 56

Zheng, W. and van der Laan, Mark (2010). Asymptotic Theory for Cross-valiyeard Targeted Maximum Likelihood Estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 273. 23