Provable Detection of Propagating Sampling Bias in Prediction Models

Pavan Ravishankar [1], Qingyu Mo [1], Edward McFowland III [2], Daniel B. Neill [1]

[1]Machine Learning for Good Laboratory, New York University
[2]Harvard Business School
pr2248@nyu.edu, qm348@nyu.edu, emcfowland@hbs.edu, daniel.neill@nyu.edu

Abstract

With an increased focus on incorporating fairness in machine learning models, it becomes imperative not only to assess and mitigate bias at each stage of the machine learning pipeline but also to understand the downstream impacts of bias across stages. Here we consider a general, but realistic, scenario in which a predictive model is learned from (potentially biased) training data, and model predictions are assessed post-hoc for fairness by some auditing method. We provide a theoretical analysis of how a specific form of data bias, differential sampling bias, propagates from the data stage to the prediction stage. Unlike prior work, we evaluate the downstream impacts of data biases quantitatively rather than qualitatively and prove theoretical guarantees for detection. Under reasonable assumptions, we quantify how the amount of bias in the model predictions varies as a function of the amount of differential sampling bias in the data, and at what point this bias becomes provably detectable by the auditor. Through experiments on two criminal justice datasets– the well-known COMPAS dataset and historical data from NYPD's stop and frisk policy– we demonstrate that the theoretical results hold in practice even when our assumptions are relaxed.

Introduction

Machine learning models are being used in numerous applications such as healthcare (De Fauw et al. 2018), online advertising (Perlich et al. 2014), and finance (Malekipirbazari and Aksakalli 2015). Due to its increased proliferation, there is a rising concern in the machine learning community to deploy fair machine learning models (Barocas, Hardt, and Narayanan 2017; Mehrabi et al. 2021). Since decision-making in machine learning comprises of various stages such as the data stage, modeling stage, and prediction stage (Suresh and Guttag 2019), it becomes imperative to look at the fairness problem across stages, rather than limiting the discussion to a single stage. For instance, the **data stage** could be biased due to members of a subgroup being systematically selected with a higher or a lower probability than others (Medical-Dictionary 2016), also known as sample selection bias. Such biases could propagate to the **prediction stage**, and the resulting biases in prediction could be compounded by other sources such as model misspecification (Gajane and Pechenizkiy 2017). However, it is unclear precisely how and to what extent the data bias would affect the predictions, and when the resulting prediction biases would be detectable by some auditing approach. Such biases, once detected and precisely characterized, could then be corrected, e.g., by resampling to de-bias the data.

In this paper, we analyze the propagation of **differential sampling bias** from the data stage to the prediction stage. Differential sampling bias is a form of sample selection bias in which some subpopulation S is sampled non-uniformly, such that the distribution of an outcome variable Y given predictor variables X in the sampled data for S differs from the true (population) distribution of Y given X for S. [1]

This bias can arise in many different circumstances. For example, in criminal justice, both the organizational biases of police departments (e.g., a policy of conducting large numbers of pedestrian stops in predominantly minority neighborhoods) and the perceptual biases of individual police officers (e.g., higher likelihood of stopping and frisking Black individuals) led to much higher proportions of Black individuals being arrested for marijuana possession, despite similar rates of use in the population as a whole (Edwards et al. 2020). In our analysis of NYPD stop and frisk data, we consider the race of the stopped individual as our outcome variable, and observe that Pr(race = "Black") is significantly increased as compared to a "less biased" alternative policing strategy.

Differential sampling bias can also result from **concept shift**: a model meant for prediction of outcome variable Y in one setting is learned using data from a different setting where the relationship between Y and the predictor variables X differs. For example, if criminal justice data from one jurisdiction is used to predict a defendant's risk of reoffending in a different jurisdiction, or if historical data is used and reoffending patterns have changed over time, the training data will exhibit differential sampling bias: the proportion of reoffenders for certain demographics may be higher or lower in the training data as compared to the true probabilities for the jurisdiction and time period of interest. In our experimental analysis of the COMPAS dataset, we inject simulated differential sampling bias (assuming concept shift) by weighted resampling of the training data.

[1]Note that differential sampling bias would *not* be present if subpopulation S was under- or over-sampled but the distribution of Y given X for S remained unchanged. We do not address other forms of sample selection bias here.

Figure 1: Problem setup. Step 1: Differential sampling bias Δ is induced into subgroup S^T. Step 2: A classification model is trained on the biased data to predict the probability of belonging to class $Y = 1$. Step 3: Predicted probabilities of belonging to $Y = 1$, given by $\{\widetilde{p}_i\}$, are made on the test data. Step 4: Bias Scan finds the most biased subgroup S^* and its log-likelihood ratio score F^* based on the predictions.

Here we introduce the first formal analysis of how differential sampling bias induced in the data stage (i.e., biased training data) propagates through the modeling and prediction stages, leading to significant biases in prediction. These propagated biases can then be detected by an *auditor* that compares the model predictions with the observed outcomes.

Our **problem setup** is shown in Figure 1: first, in the *data stage*, we assume initially unbiased training and test data records drawn i.i.d. from some joint probability distribution $f_{X,Y}(x, y)$ of the predictor variables X and binary outcome variable Y. Then differential sampling bias Δ is injected into the "true" subgroup S^T for the training data only. Without loss of generality, we define Y such that the differential sampling bias increases the probability $\widetilde{\mathbb{P}}(Y = 1|X)$, thus over-sampling records with $Y = 1$ in subgroup S^T. We parameterize the multiplicative increase in the odds of $Y = 1$ by $\Delta > 1$. Second, in the *modeling stage*, a classification model is trained using the biased data. Third, in the *prediction stage*, the classifier makes predictions $\widetilde{p}_i$ (the estimated probability that $Y = 1$ for each data record) for the test data. Fourth, Bias Scan (Zhang and Neill 2016) is used to assess whether the predictions $\widetilde{p}_i$ are systematically biased as compared to the test outcomes y_i for any intersectional subgroup.

Given this problem setup, we present theoretical and empirical results showing (a) the amount of bias that propagates from the data stage to the prediction stage, as measured by the log-likelihood ratio (LLR) score found by Bias Scan; and (b) when the bias will exceed a threshold for significance, assuming a fixed false positive rate α, thus enabling detection by Bias Scan. Our specific **contributions** are as follows:

1. We define and quantify the differential sampling bias Δ induced into subgroup S^T in the binary outcome Y.
2. We derive a new closed-form expression for the LLR score of Bias Scan, used to audit a consistent classifier trained on large data with differential sampling bias.
3. We present a new asymptotic result for the null distribution of the Bias Scan score, which leads to a threshold score $h(\alpha)$ for detection at a fixed false positive rate α.
4. We demonstrate detection with full asymptotic power, $\mathbb{P}_{H_1}(\text{Reject } H_0) \to 1$, as the data size becomes large.
5. Using the threshold $h(\alpha)$, we find the minimum amount of bias Δ that needs to be induced in subgroup S^T for it to be provably detectable in the finite sample case.
6. We evaluate our theoretical results empirically on two different criminal justice datasets. On the well-known COMPAS dataset, we compare the empirical and theoretical relationships between the Bias Scan score F^* and the amount of injected bias Δ, across two different classification models and two types of bias injection (marginal and intersectional). We also analyze historical data from the NYPD's "stop-question-frisk" (SQF) policy, estimating the amount of differential sampling bias Δ in the data as compared to a "less biased" alternative policing strategy.
7. For both datasets, we observe that the empirical relationship between the propagated bias in predictions (as measured by the Bias Scan score F^*) and the differential sampling bias in data (as measured by Δ) corresponds well to the theoretical values. We also confirm that, if enough bias is present in the data stage, then the affected subgroup is detectable by the auditor in the prediction stage with high accuracy. These two conclusions demonstrate the validity of the theoretical assumptions and provide reasoning when theoretical and empirical results differ.

Related Work

Stage-specific notions of fairness and bias: The machine learning community has typically centered the fairness problem in either the data stage or the prediction stage (Barocas, Hardt, and Narayanan 2017). In the data stage, various attempts have been made to detect and mitigate data biases. For example, Zemel et al. (2013), Madras et al. (2018), and Song et al. (2019) attempt to de-bias data by learning fair representations. Silvia et al. (2020), Oneto and Chiappa (2020), and Ravishankar, Malviya, and Ravindran (2021) discuss causal notions of fairness such as path-specific fairness, and use them to detect and mitigate unfairness in the data generation process. Similarly, many approaches have been proposed to address biases in the prediction stage: Berk et al. (2021) state multiple fairness definitions such as demographic parity and calibration based on model predictions; Corbett-Davies and Goel (2018) discuss the limitations of these fairness definitions; Kleinberg, Mullainathan, and Raghavan (2016) and Chouldechova (2017) prove that, except in special cases, these definitions are incompatible; Zadrozny (2004) proposes a framework to correct bias in model predictions; and Pedreschi, Ruggieri, and Turini (2009) propose novel measures of discrimination to correct discriminatory patterns. None of the aforementioned works have analyzed how bias *propagates* downstream, across different stages of the pipeline.

Bias propagation pipelines: Suresh and Guttag (2019) discuss the bias problem holistically, rather than centering it to a particular stage, by laying out a framework comprising of biases originating at different stages of the pipeline. Similarly, an opinion article by Hooker (2021) proposes that bias should be viewed and analyzed as an aggregation of the biases arising in different stages. However, neither of these works provide any formal, quantitative analysis of how bias propagates between stages. Rambachan and Roth (2019) quantitatively analyze how selection bias propagates from the data stage to the prediction stage. However, the study makes a strong assumption about the form of the selection process, and does not discuss whether the propagated bias is detectable or how it can be detected in the prediction stage.

Frameworks for detection of intersectional biases: Several recent approaches have been proposed to detect biases affecting a subpopulation defined along multiple data dimensions (Zhang and Neill 2016; Kearns et al. 2018). Here we apply Bias Scan (Zhang and Neill 2016) to assess models learned from biased data, detecting intersectional subgroups where the model predictions $\widetilde{p}_i$ most significantly overestimate $\mathbb{P}(Y = 1 \mid X = x_i)$. Bias Scan builds on previous univariate and multivariate subset scan approaches (Neill 2012; Neill, McFowland III, and Zheng 2013). Additionally, McFowland III, Somanchi, and Neill (2018) use a similar multidimensional scan framework to discover the subgroups that are most significantly affected by a treatment in a randomized experiment, and provide statistical guarantees on detection. However, all of these approaches focus on a single pipeline stage (predictions or outcomes), while our work examines the propagation of data biases into model predictions.

Preliminaries

Notations

Assume test data $D = \{(x_i, y_i)\}$ drawn i.i.d. from joint probability distribution $f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y|x)$ and training data $\widetilde{D} = \{(\widetilde{x}_i, \widetilde{y}_i)\}$ drawn i.i.d. from joint probability distribution $\widetilde{f}_{X,Y}(x, y) = \widetilde{f}_X(x)\widetilde{f}_{Y|X}(y|x)$. Here Y is a binary outcome variable, and thus we can write $f_{Y|X}(y|x)$ and $\widetilde{f}_{Y|X}(y|x)$ as the probabilities $\mathbb{P}(Y = y \mid X = x)$ and $\widetilde{\mathbb{P}}(Y = y \mid X = x)$, $y \in \{0, 1\}$, for test and training data respectively. Let $p_i = \mathbb{P}(Y = 1 \mid X = x_i)$ be the true probability that $Y = 1$ for test record $s_i = (x_i, y_i)$, and let $\widetilde{p}_i$ and $\hat{p}_i$ be the estimated probabilities that $Y = 1$ for test record s_i from classification models learned from training data with and without differential sampling bias. Note that $\widetilde{p}_i = \hat{p}_i$ when $\Delta = 1$ (under the null hypothesis of no bias).

We assume that X consists of a set of discrete-valued[2] predictor variables $\{X_1, \ldots, X_Q\}$ and that each variable X_i takes on a set of values V_i. An intersectional **subgroup** S is defined as a subset of the Cartesian product $V = V_1 \times \ldots \times V_Q$. A **rectangular** subgroup is one that can be represented as the Cartesian product of subsets of attribute values, $S = S_1 \times \ldots \times S_Q$, for $S_i \subseteq V_i$. For example, if X_1 = Gender, V_1 = {Male, Female}, X_2 = Race, and V_2 = {Black, White, Other}, then {Male, Female} × {Black, White} = {(Male, Black), (Female, Black), (Male, White), (Female, White)} is a rectangular subgroup, while {(Male, Black), (Female, White)} is non-rectangular. Let $rect(X)$ denote the set of all rectangular subgroups of X. Finally, for test dataset D, we associate with any given subgroup S the subset of matching data records $D_S = \{(x_i, y_i)\} \subseteq D : x_i \in S$.

Bias Scan

Bias Scan (Zhang and Neill 2016) is a multi-dimensional subset scanning algorithm used to detect intersectional subgroups for which a classifier's probabilistic predictions $\widetilde{p}_i$ of a binary outcome y_i are significantly biased as compared to the observed outcomes y_i. More precisely, Bias Scan searches for the rectangular subgroup S^* which maximizes a Bernoulli log-likelihood ratio (LLR) scan statistic,[3]

$$S^* = \underset{S \in rect(X)}{\arg\max}\ F(S),$$

with corresponding LLR score $F^* = F(S^*)$.

To obtain the score function for a given subgroup S, Bias Scan computes the generalized log-likelihood ratio $F(S) = \max_{\widetilde{q}} \log \frac{P(D \mid H_1(S,\widetilde{q}))}{P(D \mid H_0)}$, assuming the following hypotheses:

$$H_0: \quad \text{odds}(y_i) = \frac{\widetilde{p}_i}{1-\widetilde{p}_i},\ \forall s_i \in D.$$

$$H_1(S, \widetilde{q}): \quad \text{odds}(y_i) = \frac{\widetilde{q}\,\widetilde{p}_i}{1-\widetilde{p}_i},\ \forall s_i \in D_S,$$

$$\text{odds}(y_i) = \frac{\widetilde{p}_i}{1-\widetilde{p}_i},\ \forall s_i \in D \setminus D_S.$$

Here we detect biases where the probabilities $\widetilde{p}_i$ are overestimated, and thus $0 < \widetilde{q} < 1$. As derived in the Technical Appendix, the resulting log-likelihood ratio score $F(S)$ is

$$F(S) = \max_{0<q<1} \left(\sum_{s_i \in D_S} y_i \log q - \sum_{s_i \in D_S} \log(1 - p_i + q\, p_i) \right) \tag{1}$$

The Bias Scan algorithm for optimizing $F(S)$ over rectangular subgroups is provided in the Technical Appendix.

Differential Sampling Bias

In this section, we quantify differential sampling bias for a subgroup S, as follows:

Definition 1. *A subgroup S exhibits **differential sampling bias** $\Delta > 1$ towards the outcome $Y = 1$ if, for all $x \in S$,*

$$\widetilde{\mathbb{P}}(Y = 1 | X = x) = \frac{\Delta \mathbb{P}(Y{=}1|X{=}x)}{\Delta \mathbb{P}(Y{=}1|X{=}x) + \mathbb{P}(Y{=}0|X{=}x)} \tag{2}$$

For example, differential sampling bias could be injected into unbiased training data by re-drawing data elements $\{(\widetilde{x}_i, \widetilde{y}_i)\}$, for $\widetilde{x}_i \in S$, with replacement, with sampling weights $w_i = \Delta$ for $\widetilde{y}_i = 1$ and $w_i = 1$ for $\widetilde{y}_i = 0$. We use this approach to inject bias into the COMPAS dataset in our experiments below.

Theoretical Results

In this section, we derive theoretical results to understand the propagated effects of differential sampling bias and to provide statistical guarantees for detectability.

More precisely, we prove four main theorems. Given the problem setup described above and the assumptions listed below, **Theorem 1** provides an asymptotic closed-form formulation of the Bias Scan log-likelihood ratio (LLR)

[2] Sensitive covariates (e.g. race, ethnicity, and gender) are usually discrete. Continuous covariates can be discretized as a preprocessing step, using the observed covariate distribution or domain knowledge.

[3] We assume that the bias is injected into a rectangular subgroup, a common formulation (e.g., used in decision trees), as it is representative of a cohesive and interpretable subpopulation.

score $F(S^T)$ of the injected subgroup S^T as a function of the amount of differential sampling bias Δ. If S^T is a rectangular subgroup, $S^T \in rect(X)$, this score is a lower bound on the overall Bias Scan score $F^* = \max_{S \in rect(X)} F(S)$. **Theorem 2** provides an upper bound for the null distribution of F^* (i.e., assuming no bias is present), enabling us to compute a threshold score for detection. Finally, **Theorems 3 and 4** combine these results to show asymptotic detection with full power for any $\Delta > 1$ as the sizes of the training and test data go to infinity, as well as computing the minimum amount of bias Δ needed for detection in finite test data.

These Theorems rely on **three key assumptions**:
(A1) *Consistency* of the classifier used in the prediction stage, for learning the conditional distribution $\widetilde{f}_{Y|X}$.
(A2) *Full support* of the biased training data: $support(\widetilde{f}_X) \supseteq support(f_X)$, and $support(f_X) \cap S \neq \emptyset$.
(A3) *Positivity*: $0 < \mathbb{P}(Y = 1 \mid X = x) < 1, \ \forall x$.
Given these assumptions, we first derive the relationship between the amount of differential sampling bias Δ injected into subgroup S, and the Bias Scan score $F(S)$ as follows,

Theorem 1. *Assume that a classifier is trained on data $\widetilde{D}$ with differential sampling bias $\Delta > 1$ for subgroup S and makes predictions $\widetilde{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$. If Bias Scan is used to assess bias in $\widetilde{p}_i$ as compared to y_i, then under assumptions (A1)-(A3), as the number of training data records $|\widetilde{D}| \to \infty$, the Bias Scan score $F(S)$ of subgroup S converges to:*

$$F(S) \to F_{old}(S) - \sum_{s_i \in D_S} y_i \log \Delta + \sum_{s_i \in D_S} \log(\Delta p_i + 1 - p_i),$$

if $\Delta > \hat{q}_{MLE}$, and $F(S) \to 0$ otherwise, where $\hat{q}_{MLE}$ is the maximum likelihood estimate of $\widetilde{q}$ for Bias Scan assuming no differential sampling bias ($\Delta = 1$), satisfying

$$\sum_{s_i \in D_S} y_i = \sum_{s_i \in D_S} \frac{\hat{q}_{MLE}\, p_i}{\hat{q}_{MLE}\, p_i + 1 - p_i}, \text{ and}$$

$$F_{old}(S) = \sum_{s_i \in D_S} y_i \log \hat{q}_{MLE} - \sum_{s_i \in D_S} \log(1 - p_i + \hat{q}_{MLE} p_i)$$

is the Bias Scan score of subgroup S assuming no differential sampling bias ($\Delta = 1$).

The proof of Theorem 1 is provided in the Technical Appendix. Critically, under assumptions (A1)-(A3), as $|\widetilde{D}| \to \infty$, we have $\widetilde{p}_i \to \widetilde{\mathbb{P}}(Y = 1 \mid X = x_i) = \frac{\Delta p_i}{\Delta p_i + 1 - p_i}$ for all $s_i \in D_S$, and the corresponding predicted probabilities with no differential sampling bias, $\hat{p}_i \to \mathbb{P}(Y = 1 \mid X = x_i) = p_i$. We then show that the maximum likelihood estimate (MLE) of $\widetilde{q}$ for Bias Scan is $\hat{q}_{MLE}/\Delta$, where $\hat{q}_{MLE}$ is the corresponding MLE with no differential sampling bias. Finally, we plug in the expressions for $\widetilde{p}_i$, $\hat{p}_i$, and $\widetilde{q}_{MLE}$, and simplify.

Corollary 1. *Under the conditions of Theorem 1, as the number of test data records $|D| \to \infty$, the normalized Bias Scan score $F(S)/|D|$ of subgroup S converges to:*

$$\frac{F(S)}{|D|} \to \mathbb{P}(x \in S)\mathbb{E}_{s_i \in D_S}[\log(\Delta p_i + 1 - p_i) - p_i \log \Delta],$$

an increasing function of Δ.

Next, we provide statistical guarantees for the detection of bias. To do so, we first consider the distribution of the Bias Scan score $F^* = \max_{S \in rect(X)} F(S)$ under the null hypothesis of no bias, H_0. For a given false positive rate α, we find a score threshold $h(\alpha)$ such that $\mathbb{P}_{H_0}(F^* > h(\alpha)) \leq \alpha$.

To do so, we make the additional assumption:
(A4) The number of unique covariate profiles in the test data, M, is large enough so that Gaussian approximations hold (e.g., $M > 30$) but finite (i.e., M remains constant as the number of test data records $|D| \to \infty$).

Then we can show the following:

Theorem 2. *Assume that a classifier is trained on unbiased training data $\widetilde{D}$ and makes predictions $\hat{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$, and Bias Scan is used to assess bias in $\hat{p}_i$ as compared to y_i. Let $F^* = \max_{S \in rect(X)} F(S)$ be the Bias Scan score, maximized over all rectangular subgroups S. Then under assumptions (A1)-(A4), as the number of training data records $|\widetilde{D}| \to \infty$ and the number of test data records $|D| \to \infty$, for a given Type-I error rate $\alpha > 0$, there exists a critical value $h(\alpha)$ and constants $k_1 \approx 0.202$, $k_2 \approx 0.523$ such that*

$$\mathbb{P}(F^* > h(\alpha)) \leq \alpha, \text{ where}$$

$$h(\alpha) = k_1 M + k_2 \Phi^{-1}(1 - \alpha)\sqrt{M}, \tag{3}$$

and Φ is the Gaussian cdf.

Critically, $h(\alpha)$ does not depend on the number of test data records $|D|$, but only on the number of unique covariate profiles in the test data M. Now, we prove that under the presence of bias Δ, $h(\alpha)$ serves as a threshold for rejecting the null hypothesis of no bias with full asymptotic power.

Theorem 3. *Assume that a classifier is trained on data $\widetilde{D}$ with differential sampling bias $\Delta > 1$ for rectangular subgroup S^T and makes predictions $\widetilde{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$, and Bias Scan is used to assess bias in $\widetilde{p}_i$ as compared to y_i. Let $F^* = \max_{S \in rect(X)} F(S)$ be the Bias Scan score, and let $h(\alpha)$ be the score threshold for detection at a fixed Type-I error rate of α, as given in Equation (3). Then for any $\alpha > 0$ and $\Delta > 1$, under assumptions (A1)-(A4), as the number of training data records $|\widetilde{D}| \to \infty$ and the number of test data records $|D| \to \infty$, $\mathbb{P}(F^* > h(\alpha)) \to 1$.*

We now find the minimum bias that needs to be induced into subgroup S to be detectable for a given Type-I error rate.

Theorem 4. *Assume that a classifier is trained on data $\widetilde{D}$ with differential sampling bias $\Delta > 1$ for rectangular subgroup S^T and makes predictions $\widetilde{p}_i$ for unbiased test data $D = \{(x_i, y_i)\}$, and Bias Scan is used to assess bias in $\widetilde{p}_i$ as compared to y_i. Let $F^* = \max_{S \in rect(X)} F(S)$ be the Bias Scan score, and let $h(\alpha)$ be the score threshold for detection at a fixed Type-I error rate of α, as given in Equation (3). Further, assume D_{S^T} is fixed, with finite size $|D_{S^T}|$ and $\left(\sum_{s_i \in D_{S^T}} y_i\right) < |D_{S^T}|$. Then for any $\alpha > 0$, under assumptions (A1)-(A4), as the number of training data*

records $|\widetilde{D}| \to \infty$*, there exists* $\Delta_{thresh} \geq 1$ *such that, if* $\Delta > \Delta_{thresh}$*, then* $\mathbb{P}(F^* > h(\alpha)) \to 1$*, where*

$$\Delta_{thresh} = \max(1, Q^{-1}(h(\alpha) - F_{old}(S^T))),$$

$$Q(\Delta) = \sum_{s_i \in D_{S^T}} (\log(\Delta p_i + 1 - p_i) - y_i \log \Delta),$$

and $F_{old}(S^T)$ *is the Bias Scan score of subgroup* S^T *assuming no differential sampling bias (*$\Delta = 1$*).*

Proofs of Theorems 1-4 are provided in the Appendix.

Experiments

We perform experiments on two criminal justice datasets to validate our theoretical results: semi-synthetic predictions of recidivism risk derived from the well-known COMPAS dataset, and real-world "stop, question and frisk" (SQF) data from the New York Police Department (NYPD).

Experiments on COMPAS/ProPublica Data

COMPAS is a commercial decision-support algorithm which has been applied in many jurisdictions to estimate a defendant's probability of reoffending, with impacts on criminal justice outcomes such as bail, sentencing, and parole. COMPAS gained notoriety when investigative journalists from ProPublica published a study arguing that COMPAS was racially biased against Black defendants (Angwin et al. 2016). The public dataset compiled by ProPublica[4], including COMPAS risk predictions for 7,214 defendants in Broward County, Florida, from 2013-2014, and a two-year follow-up to record which defendants were rearrested, has been studied by numerous algorithmic bias researchers (Barenstein 2019).

While most of these analyses focus on assessing biases in the COMPAS risk predictions (Chouldechova 2017; Kleinberg et al. 2018), we instead utilize this dataset to learn predictive models for the binary outcome (rearrest within two years) as a function of five categorical predictor variables[5], and use these models to study how differential sampling bias in the data propagates to the model predictions.

To do so, we consider differential sampling biases $\Delta \in \{1, 1.25, 1.5, \ldots, 10\}$ injected into one of two rectangular subgroups. Letting X_1 = Gender, X_2 = Race, and V_j = the set of all possible values for attribute X_j, we consider the subgroups $S^T = \{\text{Female}\} \times V_2 \times \ldots \times V_5$ and $S^T = \{\text{Female}\} \times \{\text{Caucasian}\} \times V_3 \times \ldots \times V_5$. The first subgroup represents a *marginal bias* against females (since we are oversampling females who reoffended, as compared to females who did not reoffend, by a factor of Δ in the training data, thus leading to an overestimate of their reoffending risk), while the second subgroup represents an *intersectional bias* against white females. We also consider two different classifiers, random forest and logistic regression, and average results over 100 trials for each combination of classifier, injected subgroup S^T, and amount of bias Δ.

[4]https://github.com/propublica/compas-analysis/compas-scores-two-years.csv

[5]Predictors include gender, race, charge degree, age < 25, and number of prior offenses ("none", "1 to 5", or "more than 5").

For each trial, we randomly partition the data into 80% training and 20% testing data. If $\Delta > 1$, then differential sampling bias Δ is injected into subset S^T for the training data $\widetilde{D}$, resampling data records $(\widetilde{x}_i, \widetilde{y}_i) \in \widetilde{D}_{S^T}$ with replacement (where records with $\widetilde{y}_i = 1$ have weight Δ and records with $\widetilde{y}_i = 0$ have weight 1), and leaving the test data D and the rest of the training data unchanged. The classifier is trained on the biased training data, and used to make predictions $\widetilde{p}_i$ on the unbiased test data. Then Bias Scan is used to assess whether these predictions are biased, reporting the highest scoring subgroup $S^* = \arg\max_{S \in rect(X)} F(S)$ and its score $F^* = F(S^*)$. We then compare the values of the Bias Scan score F^*, the score of the injected subgroup $F(S^T)$ (calculated by equation (1)), and the theoretical score of S^T, which we denote as $F_{theo}(S^T)$. The value of $F_{theo}(S^T)$ is computed using only the *unbiased* training and test data, as defined in Theorem 1: $F_{theo}(S^T) = F_{old}(S^T) - \sum_{s_i \in D_{S^T}} y_i \log \Delta + \sum_{s_i \in D_{S^T}} \log(\Delta p_i + 1 - p_i)$, if $\Delta > \hat{q}_{MLE}$, and $F_{theo}(S^T) = 0$ otherwise. We also compute the overlap (Jaccard coefficient) between the injected subset of test data records D_{S^T} and the detected subset D_{S^*}:

$$\text{overlap} = \frac{|D_{S_T} \cap D_{S^*}|}{|D_{S_T} \cup D_{S^*}|}.$$

Finally, we use Theorems 2 and 4 to estimate the critical value $h(\alpha)$ and the corresponding threshold value Δ_{thresh}, for which we expect $\mathbb{P}(F^* > h(\alpha)) \to 1$ when $\Delta > \Delta_{thresh}$.

Given these values for each amount of bias Δ (averaged over the 100 trials, for a given classifier and a given injected subgroup S^T), we form two plots: one comparing F^*, $F(S^T)$, and $F_{theo}(S^T)$ as a function of Δ, and one showing overlap between D_{S^T} and D_{S^*} as a function of Δ, as compared to Δ_{thresh}.

If assumptions (A1)-(A4) hold, as the size of the training data grows to infinity, we expect perfect overlap between the curves for $F_{theo}(S^T)$ and $F(S^T)$ by Thm. 1. As Δ becomes large compared to Δ_{thresh}, we expect $S^* \approx S^T$, and thus $F^* \approx F(S^T)$ and overlap ≈ 1, while for small Δ, we expect $F^* > F(S^T)$ and overlap $\ll 1$. We now examine whether these expectations are met for the finite, real-world COMPAS dataset, for each classifier and each injected subgroup S^T.

Experimental results For the logistic regression classifier learned from training data injected with marginal differential sampling bias (Figure 2), we observe near-perfect overlap between the observed score $F(S^T)$ and theoretical score $F_{theo}(S^T)$ for the injected subgroup S^T, suggesting the validity of our theoretical results above. As expected, the Bias Scan score $F^* \approx F(S^T)$ and overlap ≈ 1 for $\Delta > \Delta_{thresh}$, while $F^* > F(S^T)$ and overlap $\ll 1$ for small Δ. For the random forest classifier learned from training data injected with marginal differential sampling bias (Figure 3), we see similar results, but with $F(S^T)$ slightly greater than $F_{theo}(S^T)$ for large Δ. This is likely due to data sparsity: the combination of finite training data and high bias may lead to few or no training data records with $\widetilde{y}_i = 0$ for some covariate profiles in the injected subgroup, leading to inaccurate estimation of $\widetilde{\mathbb{P}}(Y = 1 \mid X)$. This pattern is repeated for the random forest classifier learned from training data injected

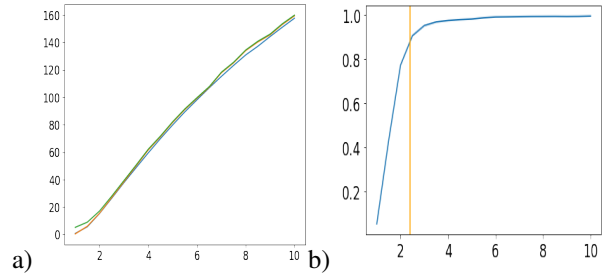

Figure 2: Logistic regression classifier with marginal bias. (a) Scores F^* (green), $F(S^T)$ (orange), and $F_{theo}(S^T)$ (blue) vs. Δ. (b) Overlap vs. Δ, as compared to Δ_{thresh}.

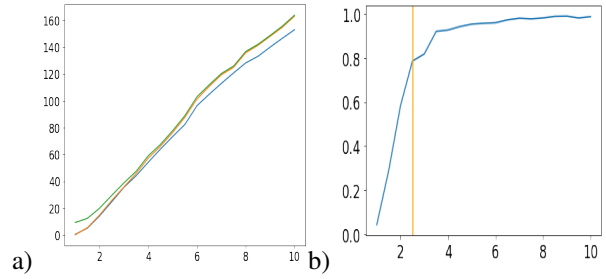

Figure 3: Random forest classifier with marginal bias. (a) Scores F^* (green), $F(S^T)$ (orange), and $F_{theo}(S^T)$ (blue) vs. Δ. (b) Overlap vs. Δ, as compared to Δ_{thresh}.

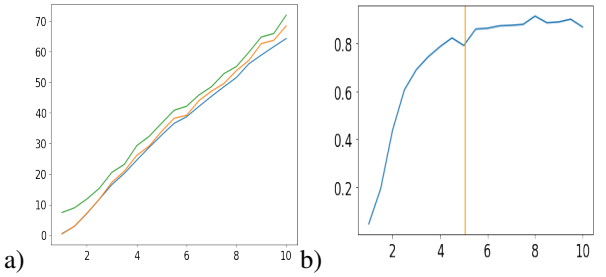

Figure 4: Random forest classifier with intersectional bias. (a) Scores F^* (green), $F(S^T)$ (orange), and $F_{theo}(S^T)$ (blue) vs. Δ. (b) Overlap vs. Δ, as compared to Δ_{thresh}.

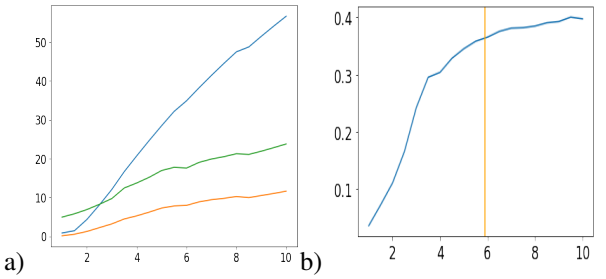

Figure 5: Logistic regression classifier with intersectional bias. (a) Scores F^* (green), $F(S^T)$ (orange), and $F_{theo}(S^T)$ (blue) vs. Δ. (b) Overlap vs. Δ, as compared to Δ_{thresh}.

with intersectional differential sampling bias (Figure 4), with a larger gap between $F(S^T)$ and $F_{theo}(S^T)$, most likely due to the smaller amount of training data in S^T. Similarly, the smaller amount of test data in S^T leads to some noise in the detected subgroup, resulting in overlap ≈ 0.9 rather than 1, and thus $F^* = \max_{S \in rect(X)} F(S) > F(S^T)$. Nevertheless, these results suggest that the theoretical values of $F_{theo}(S^T)$ and Δ_{thresh} are good approximations even for finite data.

For the logistic regression classifier learned from training data injected with intersectional differential sampling bias (Figure 5), however, we see a very different picture: as Δ increases, the Bias Scan score F^* and the score of the injected subgroup $F(S^T)$ are both much smaller than the theoretical score $F_{theo}(S^T)$, and the overlap between S^* and S^T plateaus around 0.4 even for large Δ. This is because assumption (A1) is violated: the logistic regression model is misspecified and cannot learn the intersectional bias against white females, instead learning separate (and much smaller) marginal biases against all females and all white individuals via the learned model coefficients on these terms. When an interaction term for white females is manually added to the logistic regression model specification (Figure 6), we observe that this additional term resolves the problem, and we again have a near-perfect match between the theoretical and observed scores for the injected subgroup S^T.

Experiments on NYPD Stop and Frisk Data

The New York Police Department (NYPD) has long been plagued with accusations of racially discriminatory policing practices related to its "stop, question, and frisk" (SQF) policies. Gelman, Fagan, and Kiss (2007) found that persons of color "were stopped more frequently than whites, even after controlling for precinct variability and race-specific estimates of crime participation". Goel, Rao, and Shroff (2016) concluded that Black and Hispanic individuals were disproportionately impacted by "low hit rate" stops, where the officer suspected the stopped individual of criminal possession of a weapon (CPW) but the *ex ante* probability of recovering a weapon was low. Here we assess racial bias in NYPD policing practices by analyzing five years of SQF data during the peak of the stop and frisk policy, prior to a 2013 court ruling (Floyd v. City of New York) that NYPD stop-and-frisk tactics were unconstitutionally targeting New Yorkers of color.

Thus our dataset consists of 760,489 pedestrian stops (made by NYPD officers for suspected CPW) from 2008-2012, downloaded from the city's web site[6]. Following Goel, Rao, and Shroff (2016), we first fit a logistic regression model to predict the probability that each stopped individual was found to have a weapon, using location ("housing", "transit", or "neither"), precinct, and 18 binary variables describing the circumstances of the stop[7] as predictors. Stops with *ex*

[6] www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

[7] These circumstances include suspicious object, fits description, casing, acting as lookout, suspicious clothing, drug transaction, furtive movements, actions of violent crime, suspicious bulge, witness report, ongoing investigation, proximity to crime scene, evasive response, associating with criminals, changed direction, high crime area, time of day, and sights and sounds of criminal activity.

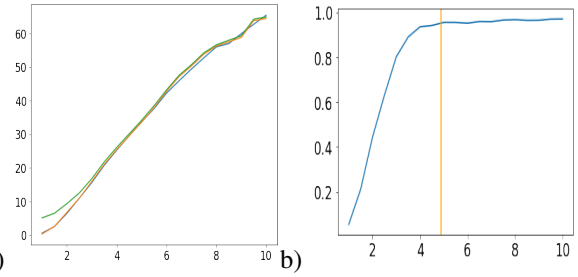

Figure 6: Logistic regression classifier (including interaction term) with intersectional bias. (a) Scores F^* (green), $F(S^T)$ (orange), and $F_{theo}(S^T)$ (blue) vs. Δ. (b) Overlap vs. Δ, as compared to Δ_{thresh}.

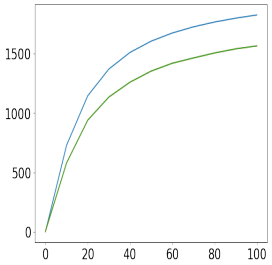

Figure 7: Random forest classifier with heterogeneous bias (SQF data). Scores F^* (green), $F(S^T)$ (orange), and $F_{theo}(S^T)$ (blue) vs. proportion of low probability stops k.

ante probability of recovering a weapon at least 0.1 were marked as "high probability". If only high probability stops were conducted, 4.8% of stops would have been made, 46% of weapons would have been recovered, and the proportion of stopped individuals who were neither Black nor Hispanic would have more than doubled, from 9% to 23%.

Next we create a new dataset with the demographics of each stopped individual (borough, sex, race, and age decile, all of which were excluded from the predictive model above), and whether each was a high or low probability stop. We then assess racial bias by considering the race of the stopped individual as the outcome variable, and comparing the original, biased policing data to an alternative, "less biased"[8] policing practice in which only high probability stops were made.

More precisely, we perform the following steps, for each value of $k \in \{0, 10, \ldots, 100\}$: (1) split the data into equal-sized training and test sets; (2) remove all low probability stops from the test data; (3) remove $(100 - k)\%$ of the low probability stops from the training data; (4) learn a random forest classifier from the training data to estimate the probability that Race = Black for each stopped individual, conditional on the other demographic features; (5) use the learned model to predict the probability $\widetilde{p}_i$ that Race = Black for each stop in the test data; and (6) run Bias Scan on the predicted probabilities $\widetilde{p}_i$ and observed outcomes $y_i = \mathbf{1}\{\text{Race} = \text{Black}\}$ to identify the highest-scoring subgroup S^* and its score $F^* = F(S^*)$. Here $k = 0$ corresponds to drawing the training data from the same, "less biased" distribution of stops as the test data, and $k = 100$ corresponds to drawing the training data from the original, "biased" distribution of stops.

Thus, for $k > 0$, this process can be thought of as injecting differential sampling bias, increasing the odds that Race = Black by some factor $\Delta > 1$, as compared to the alternative policing practice of only making high probability stops. However, this scenario poses several new challenges for our theoretical analysis: we do not know the injected subgroup S^T or the amount of bias Δ, and in fact the bias may be heterogeneous (different Δ for different covariate profiles). Thus we make several simplifying assumptions. First, when auditing predictions from the model learned from the most biased training data ($k = 100$), Bias Scan identifies a large, high-scoring subgroup S^* consisting of individuals with Gender $\in \{\text{Male}, \text{Female}\}$, Age < 70, and Borough $\in \{\text{Manhattan}, \text{Brooklyn}, \text{Queens}, \text{Staten Island}\}$ (excluding the Bronx). We assume that this S^* is the true injected subgroup S^T. Second, we assume that Δ is constant over subgroup S^T, and thus compute the odds ratio $\Delta = \frac{p_k(1-p_0)}{(1-p_k)p_0}$, where p_k is the proportion of Black individuals in subgroup S^T of the training dataset for a given value of k. Thus we have amounts of differential sampling bias ranging from $\Delta = 1$ for $k = 0$ to $\Delta = 2.675$ for $k = 100$. We then use these values of Δ along with the "less biased" training and test data ($k = 0$) to plot $F_{theo}(S^T)$ as a function of k, and compare these theoretical values to the Bias Scan score F^* and the subgroup score $F(S^T)$. In Figure 7, we observe that $F^* = F(S^T)$ except when $k = 0$, i.e., the same subgroup S^* is detected for all $k > 0$. Additionally, we see that $F_{theo}(S^T)$ is a relatively good approximation for $F(S^T)$, with $F(S^T)$ consistently about 16% lower than $F_{theo}(S^T)$ across all values of k. This difference can be explained by our approximation of the heterogeneous bias Δ_x, for covariate profiles $x \in S^T$, by estimating a single, constant Δ value.

Conclusion

It is critical both to analyze the downstream impacts of biases as they propagate through the learning pipeline, and to create new analytical tools to detect and mitigate propagating biases. With this work, we take a step toward these goals by quantifying how a particular data bias, differential sampling bias, propagates into biased model predictions, and providing theoretical guarantees for detection of the propagated biases. We validate our theoretical results through experiments on real-world criminal justice data where our assumptions are relaxed. In future work, we plan to extend our theoretical analysis of propagating biases to other types of data bias (e.g., measurement bias) as well as biases in other pipeline stages. We are particularly interested in analyzing when model predictions are impacted by multiple, interacting biases, which we believe is often the case in complex, real-world settings.

[8]We refer to the high probability stop data as "less biased" rather than "unbiased" because it still contains biases based on which neighborhoods the NYPD officers chose to patrol, but eliminates the many low probability stops which predominantly and unfairly target racial minorities.

Acknowledgments

This work was partially supported by the National Science Foundation Program on Fairness in Artificial Intelligence in Collaboration with Amazon, grant IIS-2040898. We gratefully acknowledge input from Prof. Ravi Shroff for designing experiments on the NYPD Stop and Frisk Data.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. In *Ethics of Data and Analytics*, 254–264. Auerbach Publications.

Barenstein, M. 2019. ProPublica's COMPAS Data Revisited. *arXiv preprint arXiv:1906.04711*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *NeurIPS Tutorials*, 1: 2.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1): 3–44.

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

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*.

De Fauw, J.; Ledsam, J. R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9): 1342–1350.

Edwards, E.; Greytak, E.; Madubuonwu, B.; Sanchez, T.; Beiers, S.; Resing, C.; Fernandez, P.; and Sagiv, G. 2020. A Tale of Two Countries: Racially Targeted Arrests in the Era of Marijuana Reform. *ACLU Research Report*.

Gajane, P.; and Pechenizkiy, M. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Gelman, A.; Fagan, J.; and Kiss, A. 2007. An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias. *J. Amer. Statist. Assoc.*, 102: 813–823.

Goel, S.; Rao, J. M.; and Shroff, R. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, 10(1): 365–394.

Hooker, S. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4): 100241.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In *Proc. 35th Intl. Conf. on Machine Learning*, volume 80, 2564–2572. PMLR.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, 22–27.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *Intl. Conf. on Machine Learning*, 3384–3393. PMLR.

Malekipirbazari, M.; and Aksakalli, V. 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10): 4621–4631.

McFowland III, E.; Somanchi, S.; and Neill, D. B. 2018. Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. *arXiv preprint arXiv:1803.09159*.

MedicalDictionary. 2016. Sampling Bias. *Medical Dictionary*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6): 1–35.

Neill, D. B. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2): 337–360.

Neill, D. B.; McFowland III, E.; and Zheng, H. 2013. Fast subset scan for multivariate event detection. *Stat. Med.*, 32(13): 2185–2208.

Oneto, L.; and Chiappa, S. 2020. Fairness in machine learning. In *Recent Trends in Learning From Data*, 155–196. Springer.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2009. Measuring discrimination in socially-sensitive decision records. In *Proc. SIAM Intl. Conf. on Data Mining*, 581–592. SIAM.

Perlich, C.; Dalessandro, B.; Raeder, T.; Stitelman, O.; and Provost, F. 2014. Machine learning for targeted display advertising: transfer learning in action. *Machine learning*, 95(1): 103–127.

Rambachan, A.; and Roth, J. 2019. Bias in, bias out? Evaluating the folk wisdom. *arXiv preprint arXiv:1909.08518*.

Ravishankar, P.; Malviya, P.; and Ravindran, B. 2021. A causal approach for unfair edge prioritization and discrimination removal. In *Asian Conference on Machine Learning*, 518–533. PMLR.

Silvia, C.; Ray, J.; Tom, S.; Aldo, P.; Heinrich, J.; and John, A. 2020. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3633–3640.

Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2164–2173. PMLR.

Suresh, H.; and Guttag, J. V. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, 114.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *Intl. Conf. on Machine Learning*, 325–333. PMLR.

Zhang, Z.; and Neill, D. B. 2016. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*.