CPSC 464: Analyzing Police Bias and Fairness in Polya Urn Based Predictive Policing

Jaron Hsu, Suba Ramesh, Daniel Liu, Vivian Vasquez
Professor: Nisheeth K. Vishnoi

December 9, 2022

Abstract

Predictive policing algorithms are being used more frequently to determine how police are distributed across a city to best prevent crime. Furthermore, the disproportionate arrests and incarceration of Black individuals in the United States elucidate the need for assurance that the predictive policing algorithms being utilized do not exacerbate issues of racial bias that have plagued policing in America for centuries. Building off of Ensign et al. [9], and in response to the American Civil Liberty Union's recommendations for how to encourage systemic change in mitigating the harm that policing has on oppressed groups in America, we did the following. First, we modified the urn model designed by Ensign et al. [9] to include the effects of police bias against minority groups, which will skew the model away from the true crime rate. Then, we attempted to remedy this bias with an unfairness penalty as described by Mohler et al.[14] to ensure equal policing of different groups. The conflicting behavior of the urn model and the fairness penalty demonstrated that it is difficult to have both fairness and accuracy in predictive policing. We plot how bias can impact the accuracy of the model and the effects of the fairness penalty on different starting configurations.

Contents

1 Introduction

High level description of the problem. According to the Federal Bureau of Prisons, as of October 2022, Black individuals represent 38.4 percent of incarcerated individuals nationally despite being only a mere 13.6 percent of the entire United States population [10]. Since the founding of the Black Lives Matter movement in response to the acquittal of Trayvon Martin's murderer, increased scrutiny has been placed on the big data analytical methods that police departments deploy in predicting crime and in allocating police surveillance resources. Proponents of predictive policing posit that big data analytics offers a neutral alternative to nontechnical policing methods highly influenced by the individual racial biases of police officers. Others apply a more critical lens to the widespread implementation of predictive policing algorithms with concerns that these algorithms may increase the risk of false positives (predicting that crime is occurring within a population when it is not happening), concentrating amongst Black populations and other minoritized racial groups under the guise of increased objectivity. This further obscures the resulting social inequalities inaccurate predictive policing algorithms can produce. While the prison abolitionist movement claims that eliminating prisons and police departments altogether and replacing these institutions with rehabilitative systems is the preferred approach toward increasing racial justice in the United States, police reform activists support a vision in which targeted, pragmatic changes are implemented in America's police forces in the short term. As academics, we operate under the belief that in the interim, before either prison abolition and/or widespread prison reforms are executed by social scientists and grassroots activists, it is incumbent upon academics to identify ways that predictive policing algorithms can both be more accurate and exhibit less bias, with specific attention to racial bias.

Motivation to study the problem. Granted that we have a vested interest as social scientists in both mitigating poverty (increased mass incarceration in the last decade is linked with national poverty according to Cato Institute [16]) and in avoiding disproportionate incarceration of Black individuals in the United States, our project aims to ameliorate the issues mentioned above. Personal interest in the project emanates from one of our group member's lived experiences observing family members living in predominantly Black and Latinx low-income areas facing increased scrutiny and surveillance by the Los Angeles Police Department, which is one of the police agencies at the forefront of data analytics.

The problem of racially biased policing, given its longstanding history, is not devoid of potential solutions already existing within data and non-data-oriented spaces.

Some have proposed mandatory implicit bias training for police as a potential mitigator of issues of over-policing and racial disparities in arrest rates. Here we define implicit bias training as training that is meant to help individuals, in this context, police officers, to confront and minimize their inner unconscious biases against those from minority racial backgrounds. While the widespread implementation of training, in theory, is well-intentioned, the University of Chicago's Crime Lab's survey of the efficacy of this work, "Implicit Bias Training for Police," [1] speaks to the need for other solutions to racial biases in policing.

The paper delineates how implicit bias training, which is not limited to, but includes practices of perspective taking, counter stereotypic stimuli (exposure to situations in which outgroup members perform differently than anticipated), or evaluative conditioning (fostering positive association between outgroup members and positive stimuli) tends to yield no results within 24 hours

that the intervention is administered. The actors who generally are proponents of this solution are police organizations themselves, such as the International Association of Chiefs of Police (IACP). The IACP's Center for Police Research and Policy released the most positive, yet vague results in support of implicit bias training as a solution for over-policing of Black communities. Their most prominent publication on implicit bias training touting that 58% of New York City police officers had self-reported "using" their implicit bias training a month after the training occurred. While an elevated understanding of implicit bias is certainly commendable, it does not directly address the issue of over-policing we seek to mitigate in our work [1]. Further, as the American Civil Liberties Union states in their article "How Do We End Racism In Policing?," [6] incremental internal police reforms have failed to gain traction as a viable solution for racially biased over-policing. This is a result of its requirement to "invest more money in police departments," which is in direct opposition to the tactics that nonprofit and social organizations committed to racial justice have dedicated their primary advocacy efforts towards.

Nonprofit legal and social organizations such as the American Civil Liberties Union, the Brennan Center for Justice, as well as the Black Lives Matter Global Network Foundation Inc., have come to a consensus on their proposed solution to over-policing: the simultaneous defunding of the police and redirection of police funds to social services. The American Civil Liberty Union clearly explains how the mechanism of divestment and reinvestment of police department funds will address over-policing. First, they propose ending enforcement of minor offenses that criminalize minor behaviors (for instance, marijuana possession and distribution). Other measures proposed include ending police presence in schools as a means of protecting minority students, as well as developing more hotlines as alternatives to police responses to mental health crises. Lastly, the measures proposed by the ACLU include banning pretextual stops and consent searches (often avenues for police officers exhibiting racial bias to engage in violence against Black individuals) and implementing legal constraints on when physical violence is admissible by police officers. For the purposes of our technical analysis of PredPol, we will put aside the notion of complete police abolitionism. We will operate under the assumption in this text that the state is morally justified in using policing as a mechanism of deterring individuals from violating the law and by extension, producing societal harm.

While functioning within the general belief that police abolition is the key towards racial justice, civil rights organizations like the American Civil Liberties Union have adopted a stance on predictive policing that, while still ultimately advocating against the usage of predictive policing mechanisms, they also posit that simultaneous efforts with regards to predictive policing should be undertaken before their abolition occurs. The ACLU explained their stance on predictive policing in "Statement of Concern About Predictive Policing," [15] in which they identify predictive policing as a positive feedback loop issue. Algorithms use biased historical data collected by law enforcement to further reinforce historical over-policing, which then generates additional biased input data for the algorithm.

The proposed solution to issues created by predicting policing is to increase transparency about how predictive algorithms are utilized and to ensure the racial fairness of predictive policing algorithms before implementation. Our motivation for this project is to explore the latter solution.

Most related prior works.

1. Ensign et al.[9] demonstrate the issues with runaway feedback loops inherent in policing algorithms such as PredPol. They use an urn theory reinforcement model, where colored balls

are drawn from an urn, and the proportion of each color within the urn represents the current crime rate prediction. The goal of the urn model is to have the proportion of each colored ball in the urn converge on the true crime rate in each area. In a simulated environment, the true crime rate can be set and the accuracy of the algorithm can be measured.

Each police officer draws and replaces a colored ball from the urn to determine which area to patrol. Based on whether the police officer then discovers a crime in that area, an additional ball of the area's color is added to the urn, increasing the likelihood of patrolling that area on the next day. The researchers show that, because the discovery of crime in an area relies on the police officer's presence in an area, a runaway feedback loop forms. This causes the predicted crime rate to fail to converge to the true crime rate. However, by using an modified ball replacement strategy, where the replacement of balls in the urns is diminished if the probability of patrolling in an area is high, the urn reinforcement model is able to successfully converge on true crime rates.

After demonstrating the presence of feedback loops and a functional solution in the urn model, Ensign et al. focus on improving PredPol itself. A black box approach is used, where the inner mechanics of PredPol are not known or modified. Instead, only the data inputs to the algorithm are filtered. Data regarding "discovered" incidents (crimes that are observed and logged by patrolling police officers, rather than reported through 911) has a lower likelihood of being included in the next day's crime rate predictions if the probability of the district being patrolled (due to high predicted crime rate from the day before) is high. Prior to modification, PredPol's performance was poor, with large spread, high inaccuracy, and a tendency (as a result of the feedback loop) to predict almost all crime coming from Top1 (the district with slightly more crime than Top2 and Random). As a result of their changes, PredPol's performance showed significant improvement, especially in lowering the mean percent error. However, the spread of the results was still large, fluctuating significantly above and below the true crime rate, demonstrating additional unreliability in the PredPol algorithm.

Ensign et al. note a few limitations of their research. First, the research focuses on a model with only two areas. However, Ensign et al. hypothesize that their findings will apply to models with multiple areas as well. More importantly, the assumption is made that the true crime rate corresponds to the discovery and report rates of crimes. However, crime rate statistics in the real world are often distorted and biased based on the types of crime and demographics of the area.

2. Mohler, a co-founder of PredPol, mentions in his paper Mohler et al. [14] that it is possible that biased arrests are amplified through self-excitation in predictive policing algorithms. They introduce an algorithm which adds a penalty to term to the likelihood that the amount of police patrol received by M demographic groups is proportional to their representation in the population. They do claim that the patrol rates can match that of the demographics so it is "fair" but this leads to less crime rate reduction and lowered algorithm accuracy compared to a baseline of professional police analysts.

   They define the amount of patrol a particular group receives per individual per day. Using this we can define a fairness metric which compares the patrol amount between groups. They use a Hawkes process model of crime upon which they add this fairness metric. Compared to the

baseline Hawkes process model, the more fair model policed all groups equally proportionate to their share of the population while the original model policed black and hispanic minorities at double the rate.

There are a few assumptions which the writers of this paper make in order to simplify the problem. First, they choose the top K hotspots and assume that they should all receive the same amount of policing. This may not be accurate because even though they are hotspots there can be variation in their crime rates and demographic. This model also assumes that fair policing is proportionally equivalent policing between all groups. What happens when one group has a higher crime rate? Their newer, more fair model does demonstrate a loss of accuracy. Is this the cause of the loss of accuracy? What happens if the policing is proportional but the officers are inherently biased? This could be another source of loss in accuracy because we believe that it is a fair algorithm but a population will still have disproportionate reporting. We hope to analyze these assumptions and how the model's performance changes when we introduce bias or non-uniform crime rates.

3. This paper by Chapman et al. [8] is one of the first papers which showed that predictive policing algorithms such as PredPol can create positive feedback loops which incorrectly represent the true crime distribution of an area. They analyze the models on 3 different simulated datasets: uniform-random data, uniform-biased with randomized hot spots, and a distribution sampled from real crime data in Kent (a county in the UK). After creating a map of crime, the algorithm will send police officers to areas dependent on the amount of crime reported to the system. Then, new crime reports are generated from the police placed in these locations. Crime is again randomly placed on the map, and this process is repeated for over a month of these "days."

   Their experiments found that in both entirely random crime data and randomized data with hotspots, the PredPol algorithm clustered the crime into hotspots which did not really exist. This demonstrated the self-reinforcing property of the PredPol algorithm since the model predicts high crime rates in areas which previously had high crime reporting rates even if the true rates are uniform. This shows that the algorithm is inherently unfair because the reporting rate is not proportional to the true rate of crime before any biases or real hotspots are introduced.

   In their model, many assumptions are made on the distribution and human behavior. The distribution is uniform so there is no space for bias towards one group or another which plays a huge role in unfair policing. Additionally, the algorithm places officers based off of algorithmic recommendations rather than human intelligence which is also unrealistic compared to the real world since an algorithm will always try to optimize perfectly. Therefore, it does not explore places outside of these hotspots, but this may not be true in the real world. Since this model assumes that there is no bias, we can analyze the change in convergence rate (the rate at which the clustering becomes clear) if there are different groups who are impacted by varying amounts of bias.

Contributions Our novel contributions are primarily built upon the work of Ensign et al.[9] in "Runaway Feedback Loops in Predictive Policing," but integrate additional ideas from Mohler et al.[14]

- The urn reinforcement learning model developed by Ensign et al. [9] assumes that crime discovery and reporting rates correspond to true crime rates in each area (Assumption 3.3, Ensign et al.) [9]. Our first novel contribution is to add the effects of police bias to the model. Since Ensign et al.[9] do not consider race, we will add a parameter measuring the proportion of a certain oppressed minority to the urn model. To quantify police biases, we will add an augmenting multiplier to the police discovery rates, making it more or less likely for a police officer on patrol in an area to discover a crime, based on the racial makeup of the area. We expect that including these values will cause a decrease in algorithm accuracy and fairness.

- To correct for racial biases in the model, we will integrate an unfairness penalty such as the one tested in Mohler et al. [14] This will penalize the model for producing policing patterns which disproportionately affect one group over another.

- The adversarial nature of accuracy in the urn model and the defintion of fairness leads to theoretical bounds on the accuracy and fairness of the optimal police allocation. This proof is based off of the research in the paper by Celis et al. [7]. This will show that it is impossible for this type of predictive policing algorithm to be both fair and accurate.

- We evaluate the model's performance using simulated datasets to test special cases as well as real world data from Oakland, California based off of the work by Lum and Isaac [12].

2 Concurrent Works

We also have had the opportunity to glean insights into how peers, particularly Group 3 in CPSC 464, address the issue of inefficient, racially biased allocation of police resources and runaway feedback loops in their paper "Enforcing Fairness in Predictive Policing by Penalizing Feedback Loops." In their paper, the authors discuss ways to address the problem of racially biased allocation of police resources and feedback loops in predictive policing. They propose using the "Rooney Rule" as a conceptual tool. The "Rooney Rule" requires that a selection panel consider at least one candidate from an underrepresented group when evaluating potential candidates. This is intended to prevent implicit bias from affecting the selection process. The authors apply this principle to the PredPol algorithm, specifically concerning the Epidemic Type Aftershock Sequence (ETAS) model that uses a Poisson distribution to describe the predicted likelihood of events based on past and nearby occurrences where policing areas. Further, they propose an intervention that will incentivize officers to be disseminated to under-patrolled areas in the same way that panel members are exposed to underrepresented groups. Their specific suggested intervention proposes using the existing PredPol ETAS algorithm as a proxy for the data on whether a crime was discovered "organically" or as a result of the algorithm. The broad intuition for the intervention is that for a given grid, the background rate based on historical crime data is found. This information is used to calculate the likelihood of the accuracy of the PredPol ETAS model in predicting if the crime occurred on the grid at the time that crime that the model trained it to occur. After implementing the proposed intervention, they found that their contribution distributed the predictions across the grid elements more evenly.

Our works are similar in that they explore the notion of building upon a fairness penalty, thus deriving large theoretical insights from Mohler. We also both ground our work in the Ensign's work on how urn models illuminate how positive feedback loops function. One limitation of their work is

that they chose to use real-world data from the LAPD, which is afflicted with existing police biases. A possible next step would be to examine the results of their project with the less-biased data from a public health survey, as detailed by Lum and Isaac. Further, in how we contextualize our work, we emphasize more how various law enforcement agencies and community organizing groups are pushing more toward police abolition than police reform, while they contextualize their work with a comprehensive history of both predictive policing generally and of PredPol.

3 Other Related Works

This article by Lum et al. [12] discusses the implications and issues with existing policing algorithms. The researchers use a simulated "synthetic" environment of Oakland, California, with data from the U.S. Census and the National Survey on Drug Use and Health (NSDUH) to set up their model. The researchers chose the NSDUH data to determine the true rates, since this data is collected using best-practice population sampling techniques, and people are less likely to hide their drug use habits from a public health survey compared to from police officers. The police data source is based only on arrest data. The researchers used the data to test PredPol, and found additional evidence of the positive feedback loops also discussed in other papers. PredPol would feed off the racial biases in historical data and increase police activity in overpoliced areas, while marketing itself as an objective, race-neutral system.

Although drug use rates are somewhat uniform across Oakland, certain areas, particularly non-white low-income areas, are vastly overrepresented in drug arrests. Lum and Isaac show that the use of predictive policing based on historical drug crime data results in increased over-policing of historically over-policed communities.

4 Preliminaries / Problem Statement

Polya Urn The most fundamental probability theorem we will be utilizing comes from the work by Ensign et al. [9]: the theory of urns. The theory of urns is the most simple stochastic model of reinforcement in probability theory that on a very basic level is an idealized way of modeling real life problems as if they involved drawing balls out of an urn. In their paper, they explain that a generalized Pólya urn model containing balls of two colors from which we draw one ball from the urn and note its color. Given which color ball was drawn from the urn, the contents of the urn are altered. As stated in Pemantle's survey of urn theory, the most basic urn models include choosing a ball of one color (for instance, red or black) at random and putting the ball back in the urn with one extra ball of that color. Here we will think of the long term probability of seeing a red ball as the long term estimate of crime in a particular area (let's call an example area, Area A). While the specific mathematical notation of how the urn model functions is lengthy and can be found in the Ensign paper, it is important to note that the urn models do capture some of the key elements of the model that PredPol uses: namely, how PredPol updates its model based on discovered and reported incidents. The main framework of our project uses generalized Pólya urns to model policing feedback loops. In the urn model, colored balls are drawn from urns and replaced using the result from the previous draw. In particular, the urn replacement matrix of interest is

$$\begin{pmatrix} w_d d_A + w_r r_A & w_r r_B \\ w_r r_A & w_d d_B + w_r r_B \end{pmatrix}$$

where d_A, d_B are the rates at which police discover incidents in neighborhoods A and B, and r_A, r_B are the rates of reported incidents in A and B. Here, w_d and w_r are weights such that $w_d + w_r = 1$ so that the authors can control the ratio of discovered to reported incidents. In the paper, and in our framework, we will only consider $w_d = w_r = \frac{1}{2}$, i.e. that discovered and reported incidents are equally weighted. Using the urns, the researchers show that runaway feedback loops cause the predicted crime rate to fail to converge to the true crime rate. However, by using an improved ball replacement strategy, where the replacement of balls in the urns is diminished based on the probability of policing, the urn reinforcement model is able to successfully converge on true crime rates.

Statistical Rate Mohler et al. [14] utilized a penalty in their loss function which was attempted to have equal policing per person per race per day. This is similar to statistical rate defined as

$$SR(f, S) = \frac{\min_l Pr_S[f=1|Z=l]}{\max_l Pr_S[f=1\|Z=l]}$$

where f is the classifier, S is the data, Z is a protected identity, and l is a certain identity in Z. This is equivalent to the penalty described by Mohler et al. within a 2 race model which we are using. When race is the protected identity and there exist only 2 of them, then one must be the minimum and the other must be the maximum in the statistical rate. In the context of using an allocator of police officers instead of a classifier within a set, we can equate $Pr_S[f = 1|Z = l]$ as the expected amount of policing an individual in S will experience given that they are of race l under the allocation by f.

Adversarial Perturbations and Fairness In Celis et al. [7] they discuss fair classification with adversarial perturbations. In the case that there is an adversary perturbing the data with a rate η, we would like to find a classifier f which has maximum error ϵ and a fairness greater than some desired fairness threshold τ. The fairness is determined by statistical rate which was defined above. On a high level, in their paper, they find that there is a guarantee that there is an optimization program with parameters $\eta \in [0, 1], \tau \in [0, 1]$, class F, and perturbed data $\hat{S}$ such that the optimal solution f^o satisfies $\epsilon \leq 2\eta$ and $SR(f^o, S) \geq \tau - O(\eta/\lambda)$. They also claim that it is information-theoretically impossible that $\epsilon < \eta$ and $SR(f^o, S) \geq \tau - o(\eta/\lambda)$. Here λ is a constant such that the true classifier f^* which minimizes error and satisfies the fairness constraint classifies λ proportion of the samples given $Z = l$ positively. They claim that this can be approximated empirically to $\lambda_l = Pr_{\hat{D}}[Z = l]$ in empirical distribution $\hat{D}$ of $\hat{S}$.

5 Model and Methodology

Race and Bias We implemented race and bias to the model by assigning each neighborhood A and B to have a proportion of each race in the area. We implemented our model for 2 different races but we believe this can be extrapolated to more races in the future with similar results. We divide these racial groups such that one is subjected to racial bias by the police, while the other is not.

The discovery rate of neighborhood A, d_A, is evaluated with the following equation:

$$d'_A = (1 + \beta) * (prop\_x_A * \lambda_A) + prop\_y_A * \lambda_A$$

where $prop\_x_A$ and $prop\_y_A$ are the population proportions of race x and y in neighborhood A, λ_A is the true crime rate of the area, and β is the bias of the police officers. The bias β is defined as the percentage by which the police are more likely to discover a crime by race x. The discovery rate for neighborhood B d_B was defined symmetrically using its corresponding rates. The police bias is constant between the two neighborhoods.

We also followed the assumption by Ensign et al. that the reported crime rate will be equivalent to the true crime rate of the area. Therefore, we chose that $r_A = \lambda_A$. This means that the final update matrix used for our urn model is

$$\begin{pmatrix} d'_A + r_A & r_B \\ r_A & d'_B + r_B \end{pmatrix}$$

Fairness Penalty Discovering the true crime rates and optimally allocating policing by race may lead to negative societal effects due to reinforcement of biases.

Since the true crime rates of each neighborhood are not known, we use Mohler's version of policing fairness, where each person of each race experiences the same level of policing. This relies on the assumption that each race is equally prone to committing crimes, and has the added benefit of avoiding negative stereotypes and societal impacts due to the reinforcement of preexisting biases. Let p_X be the amount of policing that an individual of race X faces in neighborhood A or B. We would like each individual of each race to face an equal amount of policing so that the police allocation is fair. First we calculate the policing of race X per person as follows:

$$p_X = (urn\_prob_A * prop\_x_A + urn\_prob_B * prop\_x_B)/(population_X)$$

and similarly evaluate p_Y, the policing of race Y per person. Then we take the proportion $\frac{\max(p_X, p_Y)}{\min(p_X, p_Y)}$ as our urn multiplier. In the case that race X is being over policed, then we choose the neighborhood which has a lower proportion of race X and multiply that neighborhood's urn by the urn augmenting multiplier, and vice versa in the case of race Y. To regulate the impact of this multiplier, we set a constant exponent as the fairness penalty power $\in [0, 1]$ before multiplying.

6 Theoretical Results

Effect of Bias With the addition of bias, the crime rate which the model converges to is inaccurate. The original model proposed by Ensign et al. uses a matrix where both d_A and r_A are equivalent to λ_A which is the true crime rate (Assumption 3.3, Ensign et al.).

$$\begin{pmatrix} \lambda_A & w_r\lambda_B \\ w_r\lambda_A & \lambda_B \end{pmatrix}$$

This update matrix will allocate police according to the true crime rates of the area. This is because as the number of updates, each representing a day, grows large, the urns will be updated proportional to the true crime rates so the initial ball counts will become insignificant. In our model we added bias and we perturbed the original update matrix accordingly so that our model will not converge to the true crime rates.

$$\begin{pmatrix} d'_A + r_A & r_B \\ r_A & d'_B + r_B \end{pmatrix}$$

Over a large number of days, this model will approach an allocation of police proportional to the biased discovery rate. Every time that a police officer goes to neighborhood A, they will observe a biased d'_A amount of crime instead of the true λ_A amount. This is similar for neighborhood B. Therefore, addition of bias into this model will cause the model to converge to an inaccurate proportion. It is important to note that the impact of this biased discovery rate is diluted by the weighting of the reporting rate which is equivalent to the true crime rate. Here we are weighting discovery and reporting rate equally.

Fairness The fairness penalty assumes that the crime rate between races is equivalent. Therefore, the allocation of police officers which maximizes fairness is an allocation which is proportional to the populations of neighborhood A and B. This metric is independent of the true crime rate of each area. In the case where the populations of A and B are equal, we would have a 50 / 50 allocation of officers between the neighborhoods. Since this version of fairness strives for equal policing of populations regardless of true crime rates, there are situations where we cannot have both fairness and accuracy together: one comes at the cost of the other. The fairness penalty will encourage the urns to converge to values which are proportional to the populations of neighborhood A and B while the update matrix causes it to converge towards the crime rates (albeit biased).

In order to perfectly apply a penalty to the urns to cause them to converge back to the true crime rate, we must have information about the true crime rate as an observer within the neighborhood. The biased update matrix must be changed to the true crime rate matrix used by Ensign et al. to converge to the true crime rate. As a police officer using the biased urn model, they are unable to discover crime at a rate different than that of their biased rate. They draw balls from the urn which determines which neighborhood they go to and then they will observe a biased amount of crime in that neighborhood. In order to reorient their discovery rate to return to the true rate, they must know the difference between their biased rate and the true rate. This is not possible without assumptions such as $\lambda_A = r_A = d_A$.

Bounds on Error and Fairness Referencing the work of Celis et al. [7] we can draw similarities between the police allocations in the urns and the classification in the paper. The perturbation by the adversary is the racial bias of the police. The previously defined $Pr[f^* = 1|Z = l]$ from the paper can in this case be interpreted as the probability of a person being policed given that they are a member of a certain race l.

To calculate λ, we will use the example shown in (Assumption 1, Celis et al.). In this situation

$$\min_{l \in [p]} Pr_D[Z = l] = \min(prop_x, prop_y)$$

where $prop_x$ is the proportion of x in the total population. We also know that

$$\forall l \in [p] Pr[f^* = 1|Z = l] \geq \min(policing_x, policing_y)$$

because for all races, the probability that they will be policed is equivalent to the policing of their race which we calculated in the fairness penalty. This means that

$$\lambda \geq \min(prop_x, prop_y) * \min(policing_x, policing_y)$$

To calculate η we want to find the fraction of samples which were changed, this would be equivalent to the proportion of crimes which were discovered due to bias instead of true crime rate. This means that $true = policing_A * \lambda_A + policing_B * \lambda_B$ and $biased = policing_A * d'_A + policing_B * d'_B$. Then, $\eta = \frac{biased - true}{true}$. Using these definitions of λ and η we can find guarantees about our ability to balance fairness and accuracy proportional to police bias.

7 Empirical Results

7.1 Setup

In each simulation, an urn is seeded with an initial ball distribution, representing the two neighborhoods being modeled. Additionally, the demographic information of the two neighborhoods is set. Finally, True crime rates, police bias, and fairness penalty power are varied and effects are observed. Each model is simulated 1000 times, with 1000 simulated days in each trial.

7.2 Results

We tested our model with real world data from the work of Lum and Isaac in their paper [12], as well as with artificial data. The performance of our simulations is measured using the final urn composition. Specifically, the percent error between the proportion of neighborhood A balls in the urn and the proportion of true crime rate of neighborhood A.

7.2.1 Real World Simulations

Scenario 1 In the first scenario, we recreate the Top1 vs Top2 experiment of Ensign et al. [9], in which the two districts experiencing the most crime in Oakland, California are compared. We only modify it only to add the corresponding population sizes and racial makeups, and do not add police bias or statistical rate power. Although there is some minor trial to trial variation, the resultant percent error after 1000 trials of 1000 days is very accurate, within a percentage point of the true value.

Pop_A	Pop_B	$\text{Prop}_{X,A}$	$\text{Prop}_{X,B}$	λ_A	λ_B	Police Bias	Stat Rate Power	**Percent Error**
3153	3063	0.84	0.64	3.68511	2.81646	0	0	-0.519

Scenario 2 In the second scenario, modify Ensign's experiment and activate the racial bias of the police force. In 250 jurisdictions across the U.S., ABC News found in FBI data from 2015 to 2018 that black people were 10 times more likely to be arrested compared to white people [2]. This is added to the model as 9, representing a 900% increase in police discovery rate. The impact of police bias is immediately visible. The algorithm's estimate of crime rate proportion in neighborhood A is more than 10% too high. Obviously, this has many real world impacts on the vulnerable populations of the area.

Pop_A	Pop_B	$\text{Prop}_{X,A}$	$\text{Prop}_{X,B}$	λ_A	λ_B	Police Bias	Stat Rate Power	**Percent Error**
3153	3063	0.84	0.64	3.68511	2.81646	9	0	10.425

Scenario 3 In the third scenario, Ensign's experiment is further modified by adding the statistical rate power correction. The correction is able to reduce the over policing of race X, allowing the model to reach an accurate conclusion, with less than one percent error.

Pop_A	Pop_B	$\text{Prop}_{X,A}$	$\text{Prop}_{X,B}$	λ_A	λ_B	Police Bias	Stat Rate Power	**Percent Error**
3153	3063	0.84	0.64	3.68511	2.81646	9	0.034	-0.107

Scenario 4 In the fourth scenario, we remove police racial bias in order to show the potential negative impact of the statistical rate modification, where the modification's desire to police each person of each race equally causes accuracy to diminish as a result. We see that error rate returns to around five percent. This is because the statistical rate modification is now re-balancing a racial injustice that was removed from the model.

Pop_A	Pop_B	$\text{Prop}_{X,A}$	$\text{Prop}_{X,B}$	λ_A	λ_B	Police Bias	Stat Rate Power	**Percent Error**
3153	3063	0.84	0.64	3.68511	2.81646	0	0.034	-5.016

7.2.2 Accuracy vs Police Bias

In Graph 1, we can see that percent error increases in a logarithmic manner as police bias increases. With these arbitrary input conditions, the percent error line seems to plateau around 14% error.

Pop_A	Pop_B	$\text{Prop}_{X,A}$	$\text{Prop}_{X,B}$	λ_A	λ_B	Police Bias	Stat Rate Power	**Percent Error**
1000	1000	0.75	0.25	3	3	X	0.034	Y

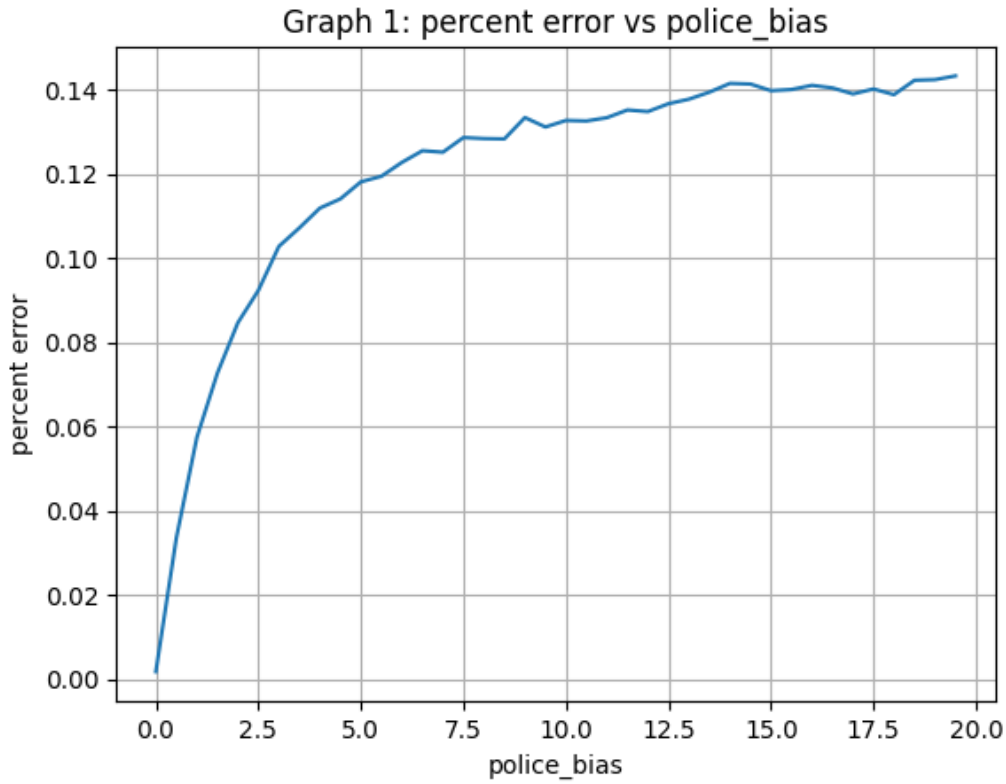

Figure 1: yay

7.2.3 Accuracy vs Fairness Penalty

Accuracy Improves then Worsens When varying the statistical rate power, we see the same behavior as before, where increasing the corrective power causes the percent error trend toward zero, then overshoot the ideal point and over correct the racial bias. This is because the predicted crime proportion of neighborhood A is driven towards the exact population proportion of neighborhood A, resulting in a percent error of a $(\text{Prop}_{X,A} - \lambda_A)$ / λ_A

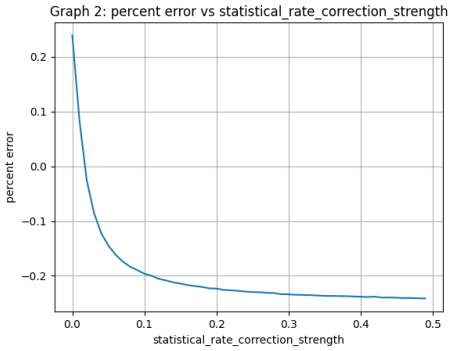

Figure 2: % error starts above zero and falls below zero
$\lambda_A > \text{Prop}_{X,A}$

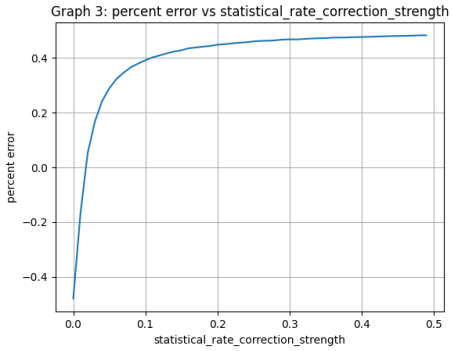

Figure 3: % error starts below zero and rises above zero
$\lambda_A < \text{Prop}_{X,A}$

Accuracy Improves In the next case, the accuracy is increased with increased statistical rate correction strength. However, the percent error lines fall short of reaching zero, and instead plateau at the $(\text{Prop}_{X,A} - \lambda_A)$ / λ_A point.

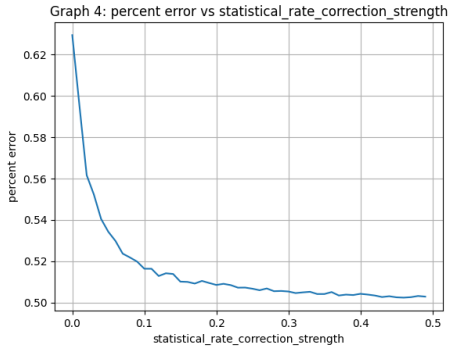

Figure 4: % error starts above zero and falls but not below zero
$\lambda_A < \text{Prop}_{X,A}$

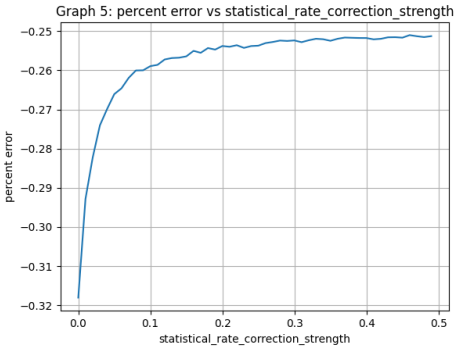

Figure 5: % error starts below zero and rises but not above zero
$\lambda_A > \text{Prop}_{X,A}$

Accuracy Improves to Near Zero Percent Error In this case, accuracy is increased with statistical rate to an asymptote at zero percent error. This is due to setting the population proportions equal to the true crime rate proportions. This results in $(\text{Prop}_{X,A} - \lambda_A) \ / \ \lambda_A$ evaluating to zero.

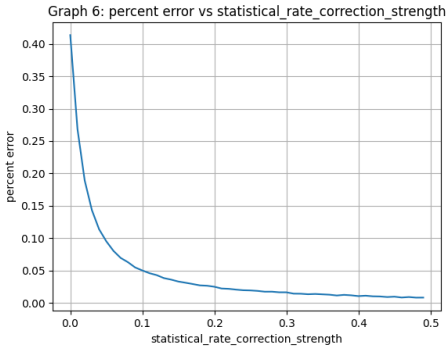

Figure 6: % error starts above zero and converges on zero
$\lambda_A == \text{Prop}_{X,A}$

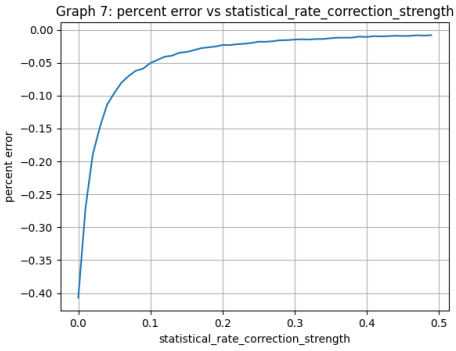

Figure 7: % error starts below zero and converges on zero
$\lambda_A == \text{Prop}_{X,A}$

Accuracy Worsens In this last case, the accuracy starts off poorly, and only gets worse as statistical rate correction strength is increased. This is because the value that $(\text{Prop}_{X,A} - \lambda_A) \ / \ \lambda_A$ converges on is even less accurate than the initial estimate.

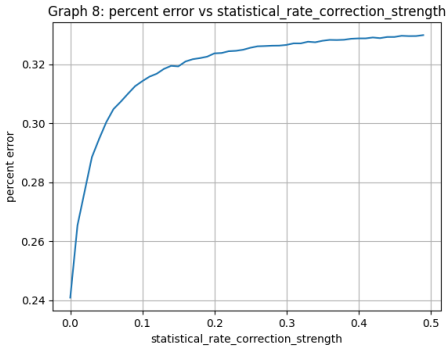

Figure 8: % error starts above zero and goes even higher
$\lambda_A < \text{Prop}_{X,A}$

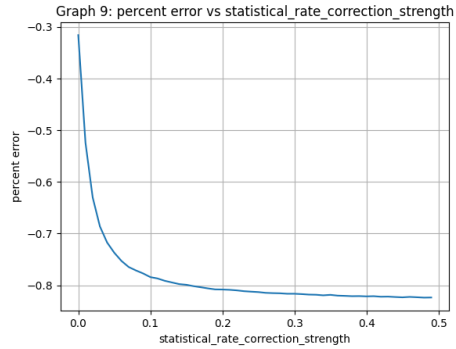

Figure 9: % error starts below zero and goes even lower
$\lambda_A > \text{Prop}_{X,A}$

8 Conclusion, Limitations, and Future Work

Conclusion From our research, we have found that the urn reinforcement learning model developed by Ensign et al. is compatible with a police bias modifier. The modifier is able to significantly impact the accuracy and fairness of the model, leading the race X citizens of neighborhood A or B to suffer from unjust over policing. The addition of a fairness metric and correction, which causes the model to tend toward equal policing for each person of each race, has benefits and detriments, because of the tradeoff between fairness and accuracy. When the fairness correction counteracts the effects of police bias, the model becomes significantly more fair and more accurate. However, it is difficult to estimate the appropriate power level of the correction, due to the hidden nature of true crime rates and police bias.

Limitations Similar to Ensign et al. [9] we make the assumption that the crime reporting rate is equivalent to the true crime rate. This leads to a dilution of the bias, and is not accurate because in the real world, distrust in the police can lead to skewed reporting rates. Another limitation is that we are unable to converge to the true crime rate under bias because we do not allow our model to observe the true crime rate when there is a bias in the system. However, this limitation is due to the realism of our model because in reality true rates are not available.

Future Works It would be interesting to solve for the optimal allocations of the police officers given the results from Celis et al. [7] and the connections drawn to that paper. Also, finding a different way to represent the fairness of a model like this could provide more robust solutions, but may not be possible without introducing simulations with more minority feature data or violating the idea of fairness.

When considering how our novel contributions will be received, we anticipate the following. There are a few key actors in the landscape of predictive policing that are worthy of mention: members of the communities being policed, law enforcement, and policymakers.

Community Members Using the community response to the Los Angeles Police Department's now-defunct Operation LASER Program (which utilized the PredPol geographic hotspot algorithm), we can infer that our amendments to the PredPol algorithm will not sufficiently appease public disapproval of predictive policing algorithms. The Los Angeles Police Department's Operation LASER Program was originally intended to extract criminal "offenders" from the Los Angeles community with laser, data-driven precision. The Stop LAPD Spying Coalition [5], founded in 2011 in response to the LASER Program's implementation, offers insights into how community members who are directly affected by predictive policing react to promises to amend, as opposed to abolish, predictive policing algorithms. Operation LASER was abolished in 2019 following vast public outcry and a subsequent internal audit that found that the LASER program prioritized "precision" and "extraction" [11] of community members (i.e., arrests) rather than functioning as a crime deterrent. Following the abolition of Operation LASER, the LAPD introduced an alternative: "data-informed community-focused policing (DICFP)," which promised to build trust between community members and police while continuing to use data to prevent crime. The Stop LAPD Spying Coalition responded with the publication of zines and articles mocking community policing's efforts and calling for the complete abolition of data-driven policing, stating that "community

policing is never friendly." [5] Given that very little information is publicly available on the development of DICFP beyond its announcement and its uncanny resemblance to hotspot geographic location-based predictive policing algorithms, it is possible that coalitions and community groups similar to Stop LAPD Spying Coalition would react more favorably to our model's assumption of equal crime rates between individuals of different racial groups as an interim, practical solution to mitigation potential racial biases in DICFP before their abolitionist aims come to fruition.

Law Enforcement Law enforcement's potential response to our recommendation would likely also be met with pushback given that our model sacrifices accuracy of predicting "true crime rates" (given, the true crime rates are based upon data collected by police who may exhibit their own personal racial biases) to maximize fairness. As NYPD Commissioner Bratton stated on predictive policing, the draw of predictive policing techniques lies in its increased "accurate and reliable probability modeling"[4]; surely, sacrificing accuracy in favor of algorithmic fairness would be met with pushback by individuals such as William Bratton. However, as Sarah Brayne states in her chapter on police pushback and responses to predictive policing in her book Predict and Surveil[3], law enforcement officials that are not in managerial roles and who are lower in the police hierarchy seem to be generally resistant to technological advancements, like predictive policing, that could open them up to managerial surveillance mechanisms. Our contribution, which would likely necessitate more scrutiny in how crime data is collected, would likely yield similar aversions on behalf of police officers. could function as a way of safeguarding against police racial biases in instances of search and seizure and could therefore make police more apt to defend their presence in predominantly Black and Latinx areas as backed by statistical evidence and algorithmic fairness. While this possibility could make police officers more likely to view our modifications to the algorithm unfavorably, simultaneously it could serve to diffuse responsibility from police officers when they search someone without reasonable cause under the guise of our model's improved algorithmic fairness. Discussions of reasonable cause segue us into our next demographic of consideration, that is, lawmakers.

Court Opinions When thinking about predictive policing algorithms and the law generally [13], a useful starting point is a definition of the Fourth Amendment of the Constitution [18], which predicts citizens from unreasonable searches and seizures. In a recent 2020 case taken before the United States Court of Appeals, Curry v. United States, predictive policing is discussed at length, offering us intuition into responses by lawmakers to our modifications to the algorithm. While the specifics of the case are less relevant to our project, commentary that locates the potential legal and civil rights limitations of predictive policing are useful to us. Judge Rodger Gregory [17] warns that in the case discussed, predictive policing techniques led officers to "ignore the assistance" of community members who signaled where a potential suspect was heading and that hotspot based predictive policing algorithms "fails as a matter of law" because they treat people who live in "violent crime locales" as second class citizens by making them more vulnerable to Fourth Amendment searches. This comment is promising in thinking about our contributions in this paper; we could infer from this comment that Gregory would be less likely to condemn modifications like ours, that, in explicitly considering race and algorithmic fairness, exhibit an air of self awareness of the limitations of predictive policing generally and that are in line with adopting predictive policing as a check against police overreach rather than an enabler of Constitutional violations.

Legislative Recommendations A specific policy recommendation we would offer to legislators is to mandate that police departments calculate and report various racial fairness metrics alongside the weekly CompStat reports. Although this would not directly implement a measure to increase justice and fairness, the increased transparency would help lead the public to demand institutional change when significant bias is quantified and identified.

Another policy recommendation to consider is to mandate that predictive policing algorithms are made public and open-source, in order to ensure a greater degree of transparency. Police watchdogs and investigative journalists will be better equipped to hold law enforcement agencies accountable.

References

[1] URL: https://urbanlabs.uchicago.edu/attachments/a11adfec96ff6054bc4146c1d366bdf26861fcc7/store/35ceee1c8a33feebad18b35aa80f7c55c435ce0f7f9e56d6cbee40b6bf27/Implicit+Bias+Training+for+Police.pdf.

[2] URL: https://abcnews.go.com/US/abc-news-analysis-police-arrests-nationwide-reveals-stark/story?id=71188546.

[3] URL: https://www.researchgate.net/publication/346349647_Police_Pushback_When_the_Watcher_Becomes_the_WatchedWhen_the_Watcher_Becomes_the_Watched.

[4] eugenefalik on January 29 and Steven on May 17. *Op-ed: Why NYPD's 'Predictive policing' should scare you*. July 2017. URL: https://citylimits.org/2015/01/29/why-nypds-predictive-policing-should-scare-you/.

[5] *About Us*. June 2022. URL: https://stoplapdspying.org/about/.

[6] Aclu. *How do we end racism in policing?: News amp; commentary*. Oct. 2021. URL: https://www.aclu.org/news/criminal-law-reform/how-do-we-end-racism-in-policing.

[7] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. *Fair Classification with Adversarial Perturbations*. 2021. DOI: 10.48550/ARXIV.2106.05964. URL: https://arxiv.org/abs/2106.05964.

[8] Adriane Chapman et al. "A Data-driven analysis of the interplay between Criminological theory and predictive policing algorithms". In: *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (2022). DOI: https://doi.org/10.1145/3531146.3533071.

[9] Danielle Ensign et al. "Runaway Feedback Loops in Predictive Policing". In: (2017). eprint: arXiv:1706.09847.

[10] *Federal Bureau of Prisons*. URL: https://www.bop.gov/about/statistics/statistics_inmate_race.jsp.

[11] *LAPD ends another data-driven crime program touted to target violent offenders*. Apr. 2019. URL: https://www.latimes.com/local/lanow/la-me-laser-lapd-crime-data-program-20190412-story.html.

[12] Kristian Lum and William Isaac. "To predict and serve?" In: *Significance* 13.5 (2016), pp. 14–19. DOI: 10.1111/j.1740-9713.2016.00960.x.

[13] Margo McGehee. *Predictive Policing Technology: Fourth Amendment and Public Policy Concerns*. July 2022. URL: https://uclawreview.org/2021/02/17/predictive-policing-technology-fourth-amendment-and-public-policy-concerns/.

[14] George Mohler et al. "A Penalized Likelihood Method for Balancing Accuracy and Fairness in Predictive Policing". In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Oct. 2018. DOI: 10.1109/smc.2018.00421. URL: https://doi.org/10.1109/smc.2018.00421.

[15] *Statement of concern about predictive policing by ACLU and 16 civil rights privacy, racial justice, and technology organizations*. URL: https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice.

[16] Michael D. Tanner. "Poverty and Criminal Justice Reform". In: *Cato's Project on Poverty and Inequality in California: Final Report* (2021). DOI: https://doi.org/10.36009/WP.20211021.

[17] *United States v. Curry, no. 18-4233 (4th cir. 2019)*. URL: https://law.justia.com/cases/federal/appellate-courts/ca4/18-4233/18-4233-2019-09-05.html.

[18] *What does the Fourth Amendment Mean?* URL: https://www.uscourts.gov/about-federal-courts/educational-resources/about-educational-outreach/activity-resources/what-does-0.