# Faking Fairness via Stealthily Biased Sampling

**Kazuto Fukuchi,**[*1,3] **Satoshi Hara,**[2] **Takanori Maehara**[3]
[1]University of Tsukuba, [2]Osaka University, [3]RIKEN Center for Advanced Intelligence Project
[3]{kazuto.fukuchi, takanori.maehara}@riken.jp, [2]satohara@ar.sanken.osaka-u.ac.jp

**Abstract**

Auditing fairness of decision-makers is now in high demand. To respond to this social demand, several fairness auditing tools have been developed. The focus of this study is to raise an awareness of the risk of malicious decision-makers who fake fairness by abusing the auditing tools and thereby deceiving the social communities. The question is whether such a fraud of the decision-maker is detectable so that the society can avoid the risk of fake fairness. In this study, we answer this question *negatively*. We specifically put our focus on a situation where the decision-maker publishes a benchmark dataset as the evidence of his/her fairness and attempts to deceive a person who uses an auditing tool that computes a fairness metric. To assess the (un)detectability of the fraud, we explicitly construct an algorithm, the stealthily biased sampling, that can deliberately construct an evil benchmark dataset via subsampling. We show that the fraud made by the stealthily based sampling is indeed difficult to detect both theoretically and empirically.

## 1 Introduction

**Background** Machine learning models are being increasingly used in individuals' consequential decisions such as loan, insurance, and employment. In such applications, the models are required to be *fair* in the sense that their outputs should be irrelevant to the individuals' sensitive feature such as gender, race, and religion (Pedreshi, Ruggieri, and Turini 2008). Several efforts have been devoted to establishing mathematical formulation of fairness (Dwork et al. 2012; Hardt et al. 2016; Dwork and Ilvento 2018) and to propose algorithms that meet the fairness criteria (Bolukbasi et al. 2016; Feldman et al. 2015; Joseph et al. 2016).

With increasing attention to fairness, social communities now require to audit systems that incorporate machine learning algorithms to prevent unfair decisions. For example, a 2014 White House Report (Podesta et al. 2014) mentioned "[t]he increasing use of algorithms to make eligibility decisions must be carefully monitored for potential discriminatory outcomes for disadvantaged groups, even absent discriminatory intent". A similar statement also appeared in a 2016 White House Report (Munoz, Smith, and Patil 2016).

To respond to the above social request, several fairness auditing tools have been developed (Adebayo 2016; Bellamy et al. 2018; Saleiro et al. 2018). These tools help the decision-maker to investigate the fairness of their system by, e.g., computing several fairness metrics (Saleiro et al. 2018), measuring the significance of the system inputs (Adebayo 2016), and identifying minority groups with unfair treatments (Bellamy et al. 2018). If the decision-maker found unfairness in their systems, she/he can then fix the systems by inspecting the causes of unfairness.

These auditing tools are also useful for promoting fairness of the decision-maker's system to the social communities. For promoting fairness of the system, the decision-maker publishes the outputs of the auditing tools. If the outputs suggest no unfairness in the system, the fact can be seen as an evidence of the system's fairness. The decision-maker thus can appeal fairness of their system by publishing the fact to earn the trust of the social communities.

**Risk of Fake Fairness** The focus of this study is to raise awareness of the potential risk of malicious decision-makers who fake fairness. If the decision-maker is malicious, he may control the auditing tools' results so that his system looks fair for the social communities even if the system is indeed unfair. Such a risk is avoidable if the social communities can detect the decision-maker's fraud. Therefore, the question is whether such a fraud is detectable. In this study, we answer this question *negatively*. That is, the fraud is very difficult to detect in practice, which indicates that the society is now facing a potential risk of fake fairness. In what follows, we refer to a person who attempts to detect the decision-maker's fraud as a *detector*.

If the decision-maker only publishes the auditing tools' outputs, the detectability of the decision-maker's fraud is considerably low. That is, the malicious decision-maker may modify the auditing tools' outputs arbitrary, whereas the detector has no way to certify whether or not the outputs are modified. This concludes that the decision-maker who only publishes the auditing tools' outputs might be untrustable.

The decision-maker should publish more information

[*]The authors are listed in alphabetical order

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

about their system in addition to the auditing tools' outputs to acquire the social communities' trust. However, because the system's information usually involves some confidential information, the decision-maker wants to prove fairness by publishing minimal information about their system.

In this study, we investigate a decision-maker who attempts to prove the fairness of their system by constructing a *benchmark dataset*. That is, the decision-maker publishes a subset of his dataset with his decisions as minimal information for proving the fairness of the system. Given the benchmark dataset with the decisions, the detector can confirm the fairness of the system by using the auditing tools. In particular, we focus our attention on an auditing tool that computes a fairness metric. With this setup, we assess the detectability of the decision-maker's fraud.

**Biased Sampling Attack** With the setup above, we consider a type of decision-maker's attacking algorithm, *biased sampling attack*. In the biased sampling attack, an attacker has a dataset $D$ obtained from an underlying distribution $P$. Here, the dataset $D$ involves the decisions made by the decision-makers' system, which are possibly unfair. The attacker deliberately selects a subset $Z \subseteq D$ as the benchmark dataset so that the value of the fairness metric for $Z$ is within a fair level. Then, the detector who employs an auditing tool that computes the fairness metric cannot detect unfairness of the decision-maker's system.

The simplest method of the biased sampling attack might be *case-control sampling* (Mantel and Haenszel 1959). If the sensitive information is gender (man or woman) and the decision is binary (positive or negative), this method classifies the dataset into four classes: (man, positive), (woman, positive), (man, negative), and (woman, negative). Then, it samples the desired numbers of points from the classes. By controlling the number of points in each class appropriately, it produces a fair subset $Z$.

Fortunately, the fraud of the case-control sampling could be detected as follows. The detector compares the distribution of the benchmark dataset $Z$ with her prior knowledge (e.g., distributions of ages or zip-codes). Then, because the case-control samples involve a bias from the original distribution, the detector may discover some unnatural thing, which indicates the decision-maker's fraud in the data-revealing process.

To hide the fraud, the malicious decision-maker will select fair subset $Z$ whose distribution looks similar to that of $D$. We refer to such a subset as *stealthily biased subset* and the problem of sampling such a subset as *stealthily biased sampling*. Intuitively, the problem is formulated as follows. The mathematical formulation of the problem is given in Section 3.

**Problem 1** (Stealthily biased sampling problem (informal))**.** Given a possibly unfair dataset $D$ obtained from an underlying distribution $P$, sample subset $Z \subseteq D$ such that (i) $Z$ is fair in terms of some fairness criteria, and (ii) the distinguishing of the distribution of $Z$ from $P$ is difficult.

**Our Contributions** In this study, we develop an algorithm for the stealthily biased sampling problem and demonstrate its difficulty of detection.

First, we formulate the stealthily biased sampling problem as a *Wasserstein distance minimization problem*. We show that this problem is reduced to the *minimum-cost flow problem* and solved it in polynomial time. (Section 3)

Second, we show the difficulty of the detection of the proposed algorithm. We introduce an ideal detector who can access the underlying distribution $P$ and compares the distribution of $Z$ and $P$ by a statistical test. The ideal detector has full information to perform the previously-mentioned fraud detection procedure, and any realistic detector cannot have such access. Therefore, if the ideal detector cannot detect the fraud, we can conclude that any realistic detector either cannot detect the fraud. We prove that the Wasserstein distance is an upper-bound of the *advantage*, which is a distinguishability measure used in the cryptographic theory (Goldreich 2009), with respect to the Kolmogorov–Smirnov statistical test (KS test) (Massey Jr 1951) (Theorem 4). This means that the proposed algorithm is hard to detect even if the ideal detector uses the KS test. (Section 4)

Finally, through synthetic and real-world data experiments, we show that the decision-maker can indeed pretend to be fair by using the stealthily biased sampling. Specifically, we demonstrate that the detector cannot detect the fraud of the decision-maker. In the experiments, we investigate detectability against a detector who can access an independent observation from $P$ but cannot $P$. This detector is also ideal but more practical than the detector introduced in Section 4. The experimental results thus show more practical detectability than the theoretically analyzed one. (Section 5)

## 2 Preliminaries

**Wasserstein Distance** Let $V$ be a finite set, and $\mu, \nu \colon V \to \mathbb{R}_{\geq 0}$ be measures on $V$. A measure $\pi$ on $V \times V$ is a *coupling measure* of $\mu$ and $\nu$ if $\mu_i = \sum_{j \in V} \pi_{ij}$, and $\nu_j = \sum_{i \in V} \pi_{ij}$, which is denoted by $\pi \in \Delta(\mu, \nu)$. Let $(\mathcal{X}, d)$ be a metric space, i.e., $d \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite, symmetric, and satisfies the triangle inequality. Suppose that each $i \in V$ has feature $x_i \in \mathcal{X}$ on the metric space. Then, the *Wasserstein distance between $\mu$ and $\nu$*, denoted by $W(\mu, \nu)$, is defined by the optimal value of the following optimization problem (Vaserstein 1969):

$$\min \sum_{i,j \in V} d(x_i, x_j)\pi_{ij}, \text{ s.t. } \pi \in \Delta(\mu, \nu). \tag{2.1}$$

The Wasserstein distance is computed in polynomial time by reducing to the minimum-cost flow problem or using the Sinkhorn iteration (Peyré and Cuturi 2019).

**Minimum-Cost Flow** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph, where $\mathcal{V}$ is the vertices, $\mathcal{E}$ is the edges, $c \colon \mathcal{E} \to \mathbb{R}$ is the capacity, and $a \colon \mathcal{E} \to \mathbb{R}$ is the cost. The minimum-cost flow

problem is given by

$$\begin{aligned} \min \quad & \sum_{e\in\mathcal{E}} a(e)f(e) \\ \text{s.t.} \quad & 0 \le f(e) \le c(e),\ e \in \mathcal{E}, \\ & \sum_{e\in\delta^+(u)} f(e) - \sum_{e\in\delta^-(u)} f(e) = \begin{cases} 0, & u \in \mathcal{V}\setminus\{s,t\}, \\ d, & u = s, \\ -d, & u = t, \end{cases} \end{aligned} \tag{2.2}$$

where $\delta^+(u) = \{(u,v) \in \mathcal{E}\}$ and $\delta^-(v) = \{(u,v) \in \mathcal{E}\}$ are the outgoing edges from $u$ and the incoming edges to $v$, respectively. $d \ge 0$ is the required amount of the flow. This problem is solvable in $\tilde{O}(\mathcal{E}\sqrt{\mathcal{V}})$ time in theory (Lee and Sidford 2013), where $\tilde{O}$ suppresses $\log$ factors. The practical evaluation of the minimum-cost flow algorithms are given in the study by (Kovács 2015).

## 3 Algorithm for Stealthily Biased Sampling

We formulate the stealthily biased sampling problem as a Wasserstein distance minimization problem. The difficulty of detecting the stealthily biased sampling is studied in Section 4. Here, we present a formulation for "categorical biasing," which controls the number of points in each category.

**Problem Formulation** Let $\mathcal{X}$ be a metric space for the feature space and $\mathcal{Y}$ be a finite set representing the outcome of the decisions. An entry of $x \in \mathcal{X}$ corresponds to a sensitive information; let $\mathcal{S}$ be a finite set representing the class of sensitive information, and let $s\colon \mathcal{X} \to \mathcal{S}$ be the mapping that extracts the sensitive information from the feature.

The dataset is given by $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where $x_i \in \mathcal{X}$ is the feature of the $i$-th point and $y_i \in \mathcal{Y}$ is the decision of the $i$-th point. For simplicity, we write $i \in D$ for $(x_i, y_i) \in D$.

Let $\nu$ be the uniform measure on $D$, whose expected number of points is $K$, i.e., $\nu_i = \frac{K}{N}, (i \in D)$. This is our reference distribution, i.e., if the decision-maker is not cheating, he will disclose subset $Z \subseteq D$ following this distribution, i.e., $\mathbb{P}(i \in S) = \nu_i$, where $\mathbb{P}$ denotes the probability.

However, as the decision-maker wants to show that the output is fair, he constructs another distribution $\mu$. Similar to the case-control sampling discussed in Section 1, we classify the dataset into bins $\mathcal{S} \times \mathcal{Y}$, and control the expected number of points sampled from each bin. Let $k\colon \mathcal{S} \times \mathcal{Y} \to \mathbb{Z}$ be the number of points of the bins, where $K = \sum_{s\in\mathcal{S}, y\in\mathcal{Y}} k(s,y) \le |D|$. Then, $\mu$ satisfies the requirement if

$$\sum_{(x_i,y_i)\in D: s(x_i)=s, y_i=y} \mu_i = k(s,y), \quad (s \in \mathcal{S}, y \in \mathcal{Y}). \tag{3.1}$$

We denote by $\mu \in P(k)$ if $\mu$ satisfies the above constraint. Note that by choosing $k$ appropriately, we can show that $Z$ is fair, thus meeting the first requirement in Problem 1.

To meet the second requirement in Problem 1, the decision-maker must determine distribution $\mu$ such that $\mu$ is indistinguishable from reference distribution $\nu$. Here, we propose to measure the indistinguishability by using the Wasserstein distance. Then, the stealthily biased sampling problem is mathematically formulated as follows.

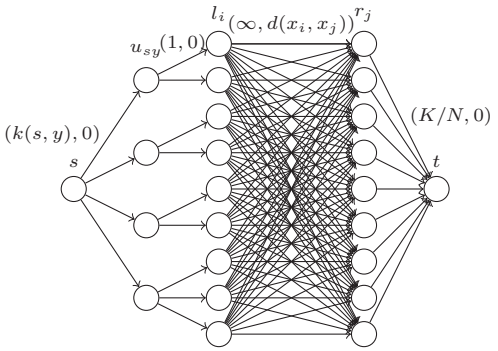



Figure 3.1: Flow network for biased sampling. $(c, a)$ on the edge if it has capacity $c$ and cost $a$.

**Problem 2** (Stealthily biased sampling problem (formal))**.** $\min W(\mu,\nu)$, s.t. $\mu \in P(k)$.

By substituting the definition of the Wasserstein distance into Problem 2, we obtain

$$\min \sum_{i,j\in D} d(x_i, x_j)\pi_{ij}, \text{s.t. } \pi \in \Delta(\mu,\nu), \mu \in P(k). \tag{3.2}$$

As the objective function is linear in $\pi$ and both $\Delta(\mu,\nu)$ and $P(k)$ are polytopes, Problem 3.2 is a linear programming problem, hence is solved in a polynomial time (Grötschel, Lovász, and Schrijver 1981).

**Efficient Algorithm** To establish an efficient algorithm for the stealthily biased sampling problem, we reduce the problem to a minimum-cost flow problem.

We construct the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in Figure 3.1 with capacity $c$ and cost $a$. Vertices $\mathcal{V}$ consist of the following five classes: (i) supersource $s$, (ii) case-controlling vertices $u_{sy}$ for all $s \in \mathcal{S}$ and $y \in \mathcal{Y}$, (iii) left vertices $l_i$ for all $i \in D$, (iv) right vertices $r_j$ for all $j \in D$, and (v) supersink $t$. Edges $\mathcal{E}$ consist of the following four classes: (i') $(s, u_{sy})$ for all $s \in \mathcal{S}$ and $y \in \mathcal{Y}$, whose cost is 0 and capacity is $k(s,y)$, (ii') $(u_{sy}, l_i)$ for all $i \in D$ with $s(x_i) = s$ and $y_i = y$, whose cost is 0 and capacity is one, (iii') $(l_i, r_j)$ for all $i \in D$ and $j \in D$, whose cost is $d(x_i, x_j)$ and capacity is $\infty$, and (iv') $(r_j, t)$ for all $j \in D$, whose cost is 0 and capacity is $K/N$.

By setting the flow amount to $K$, the solution to the above instance gives the solution to the stealthily biased sampling problem, where $\pi_{ij}$ is the flow across edge $(l_i, r_j)$, and $\mu_i$ is the flow across edge $(u_{s(x_i)y_i}, l_i)$. As $|\mathcal{V}| = O(|D|)$ and $|\mathcal{E}| = O(|D|^2)$, the problem is solvable in $\tilde{O}(|D|^{2.5})$ time (Lee and Sidford 2013).

## 4 Stealthiness of Sampling

We theoretically confirm that the stealthily biased sampling is difficult to detect. Recall that the decision-maker's purpose is to make distribution $\mu$ indistinguishable from the

uniform distribution $\nu$. To measure the indistinguishability, we introduce *advantage*, which is used in cryptographic theory (Goldreich 2009).

Let $\nu^K$ be $K$ product distribution of a sample drawn from the uniform probability distribution, and let $\mu^K$ be $K$ product distribution of a sample generated by our stealthily biased sampling algorithm. To define the advantage, let us consider the following game in which a detector attempts to distinguish $\mu^K$ and $\nu^K$: (1). Flip an unbiased coin. (2). If a head outcome is achieved, the decision-maker reveals $D \sim \mu^K$ to the detector; otherwise, the decision-maker reveals $D \sim \nu^K$ to the detector. (3). The detector estimates the side of the flipped coin. If the probability that the detector estimates the outcome of the unbiased coin correctly is near $1/2$, the detector cannot distinguish whether the obtained samples are biased.

Let $H$ be a random variable such that $\mathbb{P}(H = 1) = \mathbb{P}(H = 0) = 1/2$, which represents the flipped unbiased coin. The detector's estimation algorithm is a mapping $\Phi$ from $D$ to $\{0, 1\}$, where the output is 1 if the detector expects that the samples are drawn from $\nu^K$; otherwise, the output is 0. The probability that the detector detects bias correctly is obtained as $\mathbb{P}(\Phi(D) = H)$, where the randomness comes from flipped coin $H$ and dataset $D$. Then, the advantage is defined as follows:

$$\mathrm{Adv}(\Phi; \mu^K, \nu^K) = \left|\mathbb{P}(\Phi(D) = H) - \tfrac{1}{2}\right|. \qquad (4.1)$$

A smaller Adv value implies that biased distribution $\mu^K$ is more difficult to distinguish from $\nu^K$ against a detector with the test algorithm $\Phi$.

**Stealthiness against Kolmogorov–Smirnov Test** To assess the difficulty of detecting the stealthily biased sampling, we consider an ideal detector who can access the underlying distribution $\nu$. Here, we analyze the advantage when the ideal detector who uses the KS test.

The KS test is a goodness-of-fit test for real-valued samples. Let $F_\nu$ be the cumulative distribution function of distribution $\nu$, and let $F_K$ be the cumulative distribution function of the empirical measure of the obtained samples. Then, the KS statistic is defined as $\mathrm{KS}(D; \nu) = \sup_x |F_K(x) - F_\nu(x)|$. The KS test is rejected if $\mathrm{KS}(D; \nu)$ is larger than an appropriate threshold.

Let us consider the detector's algorithms based on the KS statistic. We formally define a detector's algorithm that returns 1 if the KS statistic is larger than threshold $\tau$ as $\Phi_{\mathrm{KS},\tau}(D) = \mathbb{I}(\mathrm{KS}(D; \nu) > \tau)$, where $\mathbb{I}$ is the indicator function.

We analyze the advantage against $\Phi_{\mathrm{KS},\tau}$ under a flatness assumption on sample distribution $\nu$. For $x \in \mathcal{X}$, let $B_\epsilon(x)$ be the $\epsilon$-ball centered at $x$. Then, the flatness assumption is defined as follows:

**Assumption 3.** There exist constants $s, C > 0$ such that for any $\epsilon > 0$, $\sup_{x \in \mathcal{X}} \nu(B_\epsilon(x)) \leq (\epsilon/C)^s$.

Many natural distributions on a real line satisfy Assumption 3. For example, the one-dimensional normal distribution satisfies Assumption 3 with $s = 1$ and $C = \sqrt{2/\pi}$.

Under the flatness assumption on $\nu$, we reveal an upper bound on the advantage against the KS test in the categorical biasing setting. Let $M$ be the number of pair types of decision and sensitive attribute. Let $\kappa$ and $\kappa'$ be the distribution over pairs of decision and sensitive attribute on the sample distribution and biased distribution. Then, we reveal the following theorem.

**Theorem 4.** Let $W(\mu^K, \nu^K)$ be the Wasserstein distance equipped with the distance $d(D, D') = \min_{i=1,\dots,K} d(x_i, x'_i)$ for $D = \{x_1, ..., x_K\}$ and $D' = \{x'_1, ..., x'_K\}$. Under Assumption 3, for threshold $\tau \geq (C/K)^{1/s}/2$, we have

$$\mathrm{Adv}(\Phi_{\mathrm{KS},\tau}; \mu^K, \nu^K) \leq K^{1/s} W(\mu^K, \nu^K)/C^{1/s} + 4K! \left((1+\mathrm{TV}(\kappa,\kappa'))/K\right)^K, \qquad (4.2)$$

where $s$ and $C$ are the constants from Assumption 3, $\mathrm{TV}(\kappa, \kappa') = \sum_i^M |\kappa_i - \kappa'_i|/2$ is the total variation distance.

The proof of this theorem can be found in the supplementary material. Since $1 + \mathrm{TV}(\kappa, \kappa') < e$ and $K! \sim (K/e)^K$, the second term in (4.2) is $o(1)$ and is dominated by the first term. Because the stealthily biased sampling minimizes the Wasserstein distance (i.e., the first term of (4.2)), it also minimizes the upper-bound of the advantage. This implies that the stealthily biased sampling is difficult to detect for the ideal detector. Consequently, for any realistic detector who has less information than the ideal one, it is even more difficult to detect the stealthily biased sampling.

## 5 Experiments

In this section, we show that the stealthily biased sampling is indeed difficult to detect, through experiments on synthetic data and two real-world data (COMPAS and Adult).[1] In the experiments, we adopted the *demographic parity* (DP) (Calders, Kamiran, and Pechenizkiy 2009) as the fairness metric for auditing. Here, let $s \in \{0, 1\}$ be a sensitive feature and $y \in \{0, 1\}$ be a decision. The DP is then defined as $\mathrm{DP} = |\mathbb{P}(y = 1 \mid s = 1) - \mathbb{P}(y = 1 \mid s = 0)|$. A large DP indicates that the decision is unfair because the decision-maker favors providing positive decisions to one group over the other group.

**Summary of the Results** Before moving to each experiment, we summarize the main results here. In the experiments, we investigated detectability of the decision-makers' fraud against an ideal detector who can access an independent observation $D'$ from the underlying distribution $P$. In all the experiments, we verified the following three points.

**R1.** Both the stealthily biased and case-control sampling could reduce the DP of the sampled set $Z$.

**R2.** The stealthily biased sampling was more resistant against the detector's fraud detection compared to the case-control sampling. Specifically, the stealthily biased sampling marked low scores of the fraud detection criteria for a wide range of the experimental settings.

[1]The codes can be found at https://github.com/sato9hara/stealthily-biased-sampling

**R3.** In all the experiments, the decision-makers successfully pretended to be fair. They could select a subset $Z$ with small DPs and small fraud detection criteria.

**Implementation** We used Python 3 for data processing. In all the experiments, we used the squared Euclidean distance $d(x_i, x_j) = \|x_i - x_j\|^2$ as the metric in Wasserstein distance. To solve the minimum-cost flow problem (3.2), we used the network simplex method implemented in LEMON Graph Library.[2] With LEMON, the problem could be solved in a few seconds for the datasets with the size $N$ up to a few thousand. For the Adult dataset, we used a bootstrap-type estimator to improve the computational scalability (see the full version (Fukuchi, Hara, and Maehara 2019) for the detail).

## 5.1 Synthetic Example

**Example 5** (Loan check)**.** Consider a decision-maker who decides to lend money ($y = 1$) or not ($y = 0$) based on the applicants sensitive feature $s \in \{0, 1\}$ (e.g., gender) and a $d$-dimensional feature vector $x \in [0, 1]^d$, where first feature $x_1$ is an income. Here, we model the criteria of the decision-maker as

$$y = \mathbb{I}(x_1 + bs > 0.5), \tag{5.1}$$

where $b \geq 0$ is a constant. Note that this decision-maker is unfair if $b \neq 0$.

To pretend to be a fair, for a set of individual's feature, sensitive feature, and the decision $D = \{(x_i, s_i, y_i)\}_{i=1}^N$, the decision-maker selects subset $Z \subseteq D$ as evidence that the decisions are fair. We solve this problem by using both the stealthily biased and case-control sampling.

**Data** We set the underlying data distribution $P$ as follows. We sampled sensitive feature $s$ with $\mathbb{P}(s = 1) = 0.5$, and sampled feature vector $x$ in a uniformly random manner over $[0, 1]^d$ with $d = 1$.[3] Decision $y$ is made by following the criteria (5.1). We sampled dataset $D$ with $N = 1,000$ observations from the underlying distribution $P$. We set the parameters in the criteria (5.1) to be $b = 0.2$. Thus, the DP of the decision-maker is $0.2$.

**Attacker** To reduce the DP through sampling, the sampled set needs to satisfy $\mathbb{P}(y = 1 \mid s = 1) \approx \mathbb{P}(y = 1 \mid s = 0) \approx \alpha$ for a predetermined ratio of positive decisions $\alpha \in [0, 1]$. The expected number of sampling in each bin $(s, y) \in \{0, 1\} \times \{0, 1\}$ is then determined by $k(s, y) = \lceil 0.5K\alpha^y(1 - \alpha)^{1-y} \rceil$ (recall that $\mathbb{P}(s) = 0.5, \forall s \in \{0, 1\}$).

**Detector** As a detector, we adopted the Kolmogorov–Smirnov two-sample test. The detector has an independent observation $D' = \{(x'_j, s'_j)\}_{j=1}^{200}$ as a referential dataset sampled from underlying distribution $P$. Here, we note that the detector has no access to decision $y$ for $D'$ because the decision criteria (5.1) is not disclosed. Given $Z$, the detector applies the Kolmogorov–Smirnov two-sample test to detect whether the distribution of $S$ is different from that of referential set $D'$. Here, we consider the strongest detector: we assume that she knows that only income $x_1$ is used in $x$ for the decision. We denote the distribution of income $x_1$ in $Z$ and $D'$ by $\mathbb{P}_S(x_1)$ and $\mathbb{P}_{D'}(x_1)$, respectively. The detector can then use the Kolmogorov–Smirnov two-sample test[4] in three ways: (i) test $\mathbb{P}_S(x_1) = \mathbb{P}_{D'}(x_1)$, (ii) test $\mathbb{P}_S(x_1 \mid s = 1) = \mathbb{P}_{D'}(x_1 \mid s = 1)$, and (iii) test $\mathbb{P}_S(x_1 \mid s = 0) = \mathbb{P}_{D'}(x_1 \mid s = 0)$. In the experiment, we set the significance level of the test to be $0.05$.

**Result** We selected a subset $Z \subseteq D$ with size $|Z| = 200$ using both the stealthily biased and case-control sampling. We repeated the experiment 100 times, and summarized the results in Figure 5.1, for several different ratios of positive decisions $\alpha$. As we summarized earlier, three key observations R1, R2, and R3 can be found in the figures.

**R1.** Figure 5.1a shows that both the stealthily biased and case-control sampling successfully reduced DP to less than $0.1$ through sampling the subset $Z$. We note that no significant differences were observed in DPs between the two sampling methods.

**R2.** Figures 5.1b, 5.1c, and 5.1d show that the stealthily biased sampling was more resistant to the Kolmogorov–Smirnov test, compared to the case-control sampling. Specifically, the stealthily biased sampling attained a small rejection rate in a wide range of $\alpha$ in the sampling process.

**R3.** By using the stealthily biased sampling, the decision-maker successfully pretended to be fair. By setting $\alpha$ in the sampling to be $0.6$, none of the tests could confidently reject that disclosed dataset $Z$ is different from the referential dataset $D'$. For $\alpha = 0.6$, the rejection rates of all the three tests were kept around 0.05, which is exactly the same as the significance level. These results indicate that the detector cannot detect the fraud made by the stealthily biased sampling: the DP of $Z$ is small, and its distribution is sufficiently natural so that the statistical test cannot reject it. The case-control sampling showed higher rejection rates in tests of $\mathbb{P}(x \mid s = 1)$ and $\mathbb{P}(x \mid s = 0)$, and thus was outperformed by the stealthily biased sampling.

Lastly, we note that the stochastic decision-maker can be far more evil than the deterministic decision-maker considered in this section. See the full version (Fukuchi, Hara, and Maehara 2019) for the detail.

## 5.2 Real-World Data: COMPAS

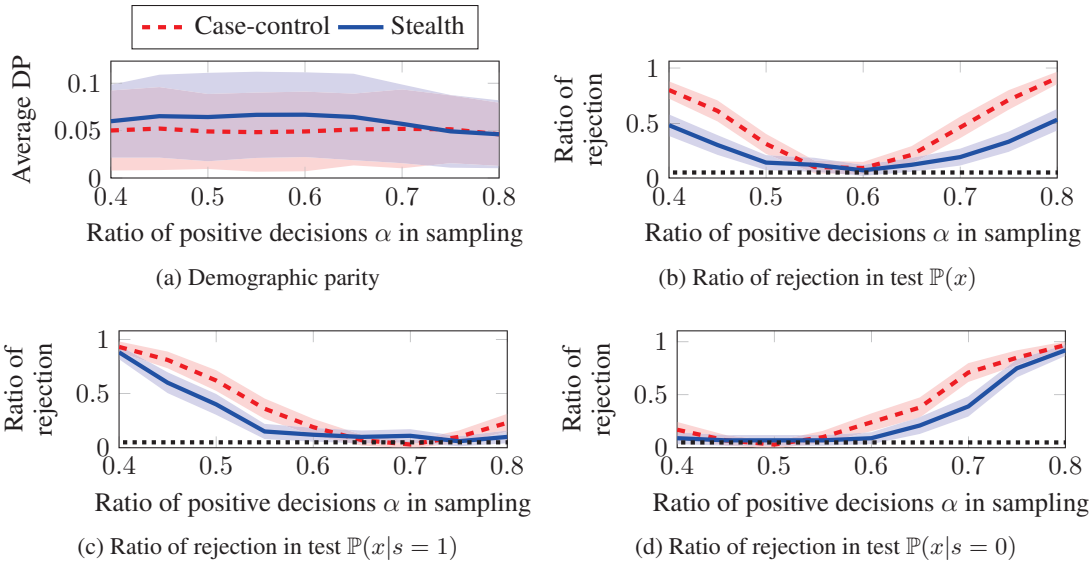
For the first real-world data experiment, we focus on the COMPAS dataset (Angwin et al. 2016).[5] The COMPAS dataset contains several defendant's records obtained from the Broward County Sheriff's Office in Florida. Each defendant is scored his or her risk of recidivism using a software called COMPAS. ProPublica (Angwin et al. 2016) revealed that the COMPAS risk score is discriminative: it tends to score white defendants with low scores while scoring black defendants with high scores.

[2] https://lemon.cs.elte.hu/trac/lemon

[3] Results for higher dimensional settings were almost the same as $d = 1$. See the full version (Fukuchi, Hara, and Maehara 2019).

[4] In practice, the detector does not know that $x_1$ is a key feature. Thus, the detector needs to use the two-sample test for multi-dimensional data. However, in our preliminary experiments, we found that multi-dimensional tests have very low detection powers. Therefore, we used an advantageous setting for the detector.

[5] https://github.com/propublica/compas-analysis




(a) Demographic parity

(b) Ratio of rejection in test $\mathbb{P}(x)$

(c) Ratio of rejection in test $\mathbb{P}(x|s=1)$

(d) Ratio of rejection in test $\mathbb{P}(x|s=0)$

Figure 5.1: Results for the decision-maker with $a = 0.2$. The shaded regions in (a) denotes the average DP $\pm$ std. The shaded regions in (b)–(d) denote 95% confidence intervals. The dotted line in (b)–(d) denotes the significance level 0.05.

Because Florida had strong open-records laws, the entire COMPAS dataset was made public, and the bias in the COMPAS risk score was revealed. Here, we consider a virtual scenario that the decision-maker was aware of the bias in the risk score, and he wants to pretend to be fair by hiding the bias. To attain this goal, the decision-maker discloses a subset of the COMPAS dataset as evidence that the COMPAS risk score is fair.

**Data** We used the same data preprocessing following the analysis of ProPublica (Angwin et al. 2016), which results in eight features $x \in \mathbb{R}^8$ of each defendant, with race as sensitive attribute $s \in \{0(\text{black}), 1(\text{white})\}$, and the decision $y \in \{0(\text{low-risk}), 1(\text{middle/high-risk})\}$. The preprocessed data includes $5,278$ records, which we randomly held out $1,278$ records as the referential dataset $D'$ for the detector. From the remaining $4,000$ records $D$, we sampled $2,000$ records as $Z$ using both the stealthily biased and case-control sampling. To reduce the DP in the sampling, we required the sampled set to satisfy $\mathbb{P}(y = 1 \mid s = 1) \approx \mathbb{P}(y = 1 \mid s = 0) \approx \alpha$ for some $\alpha \in [0, 1]$.

**Detector** The detector tries to detect the bias in the disclosed dataset $Z$ by comparing its distribution with the referential dataset $D'$. In the experiment, we adopted the Wasserstein distance (WD) as the detector's detection criteria.[6] If the WD between $Z$ and $D'$ is sufficiently large, the detector can detect the bias in $Z$, and thus the fraud of the decision-maker is revealed.

**Result** We repeated the experiment 100 times by randomly changing the data splitting, and summarized the results in Figure 5.2.[7] As the baseline without any biased sampling, we computed DP and the WD for randomly sampled records from $D$, which are denoted as *Baseline* in the figures. The figures show the clear success of the stealthily biased sampling, as we summarized in R1, R2, and R3. In Figure 5.2(a), with the stealthily biased sampling, the DPs of $Z$ have reduced significantly (R1). In Figures 5.2(b), the WDs between $Z$ and $D'$ were sufficiently small for $\alpha = 0.6$ so that they are completely indistinguishable from the baselines (R3). The case-control sampling had higher WDs, and it was thus easier for the detector to detect (R2).

## 5.3 Real-World Data: Adult

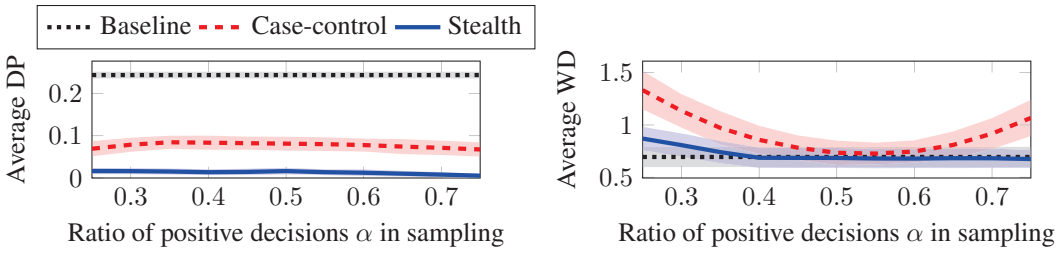
As the second real-world data experiment, we used the Adult dataset (Dheeru and Karra Taniskidou 2017). The Adult dataset contains 48,842 records with several individual's features and their labels (high-income or low-income). The dataset is known to include gender bias: in the dataset, while 30% of the male have high-income, only 10% of the female have high-income. The DP of the dataset is therefore 0.2. If we naively train a classifier using the dataset, the resulting classifier inherits the bias and becomes discriminative, i.e., the classifier favors to classify males as high-income. The goal of this experiment is to show that as if the biased classifier is fair by disclosing a part of the dataset with classifier's decision.

**Data & Classifier** In the data preprocessing, we converted categorical features to numerical features.[8] We randomly split 10,000 records for the training set, 20,000 records for the test set, and the remaining 18,842 records for the referential set $D'$ for the detector. In the experiment, we first train a classifier using the training set. As a classifier, we used logistic regression and random forest with 100 trees. We la-

[6] In COMPAS and Adult experiments, we did not adopt the multi-dimensional two-sample tests because they were too weak.

[7] Here, we measured the WD on $\mathbb{P}(x)$. The WD on $\mathbb{P}(x \mid s = 1)$ and $\mathbb{P}(x \mid s = 0)$ can be found in the full version.

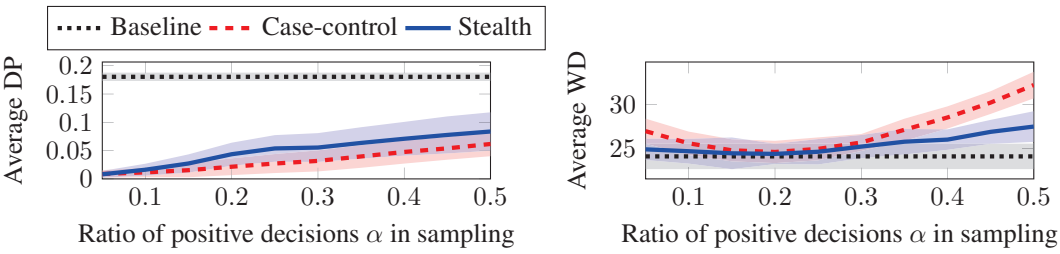
[8] We used the implementation used in https://www.kaggle.com/kost13/us-income-logistic-regression/notebook




(a) Demographic parity

(b) Wasserstein distance in $\mathbb{P}(x)$

Figure 5.2: Results for the COMPAS dataset: The shaded regions in (a) denotes the average DP $\pm$ std. The shaded regions in (b) denote the average WD $\pm$ std.




(a) Demographic parity

(b) Wasserstein Distance in $\mathbb{P}(x)$

Figure 5.3: Results for the Adult dataset: The shaded regions in (a) denotes the average DP $\pm$ std. The shaded regions in (b) denote the average WD $\pm$ std.

beled all the records in the test set using the trained classifier and obtained the dataset $D$ with the classifier's decision. We then sample the subset $Z \subseteq D$ with size $|Z| = 2,000$ using both the stealthily biased and case-control sampling. To reduce the DP in the sampling, we required the sampled set to satisfy $\mathbb{P}(y = 1 \mid s = 1) \approx \mathbb{P}(y = 1 \mid s = 0) \approx \alpha$ for a predetermined ratio of positive decisions $\alpha \in [0, 1]$.

**Detector** We adopted the same detector as the COMPAS data experiment, who refers to the WD as the bias detection metric.

**Result** We repeated the experiment 100 times by randomly changing the data splitting, and summarized the resultsf for logistic regression in Figure 5.3.[9] As the baseline, we computed the DP and the WD for randomly sampled $2,000$ sampled records from $D$, which is denoted as *Baseline* in the figure. Similar to the results of COMPAS, the figures again show the clear success of the stealthily biased sampling (R1, R2, and R3).

## 6 Conclusion

We assessed the risk of malicious decision-makers who try to deceive auditing tools, by investigating the detectability of the decision-maker's fraud. We specifically put our focus on an auditing tool that computes a fairness metric. To assess the (un)detectability of the fraud, we considered the biased sampling attack, where the decision-maker publishes a benchmark dataset as the evidence of his or her fairness. In this study, we demonstrated the undetectability by explicitly constructing an algorithm, the stealthily based sampling, that can deliberately construct a fair benchmark dataset. To derive the algorithm, we formulated the sampling problem as a Wasserstein distance minimization, which we reduced to a minimum-cost flow problem for efficient computation. We then showed that the fraud made by the stealthily based sampling is indeed difficult to detect both theoretically and empirically.

A recent study of (Aïvodji et al. 2019) has shown that malicious decision-makers can rationalize their unfair decisions by generating seemingly fair explanations, which indicates that an explanation will not be effective for certifying fairnesses. Our results indicate that passing the auditing tools will not be sufficient as the evidence of the fairness as well. Assessing the validity of other auditing tools and mechanisms against malicious decision-makers would be essential.

Lastly, in this study, we revealed the difficulty of detecting decision-maker's fraud. While auditing tools are getting popular, we will need additional social mechanisms that certify the reported results of these tools. We hope that our study opens up new research directions for practical social mechanisms that can certify fairnesses.

**Acknowledgments.** We would like to thank Sébastien Gambs and Ulrich Aïvodji for their helpful comments. Kazuto Fukuchi is supported by JSPS KAKENHI Grant Number JP19H04164. Satoshi Hara is supported by JSPS KAKENHI Grant Number JP18K18106.

[9] Here, we measured the WD on $\mathbb{P}(x)$. The WD on $\mathbb{P}(x \mid s = 1)$ and $\mathbb{P}(x \mid s = 0)$ can be found in the full version (Fukuchi, Hara, and Maehara 2019). The results for random forest can be found also in the full version (Fukuchi, Hara, and Maehara 2019).

## References


Adebayo, J. A. 2016. FairML: Toolbox for diagnosing bias in predictive modeling. *Master's thesis, Massachusetts Institute of Technology*.

Aïvodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; and Tapp, A. 2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, 161–170.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *ProPublica, May* 23.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357.

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

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

Dwork, C., and Ilvento, C. 2018. Individual fairness under composition. In *Proceedings of Fairness, Accountability, Transparency in Machine Learning*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

Fukuchi, K.; Hara, S.; and Maehara, T. 2019. Faking fairness via stealthily biased sampling. *arXiv preprint arXiv:1901.08291*.

Goldreich, O. 2009. *Foundations of cryptography: volume 2, basic applications*. Cambridge university press.

Grötschel, M.; Lovász, L.; and Schrijver, A. 1981. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica* 1(2):169–197.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 325–333.

Kovács, P. 2015. Minimum-cost flow algorithms: an experimental evaluation. *Optimization Methods and Software* 30(1):94–127.

Lee, Y. T., and Sidford, A. 2013. Path Finding II: An $\tilde{O}(m\sqrt{n})$ algorithm for the minimum cost flow problem. *arXiv preprint arXiv:1312.6713*.

Mantel, N., and Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22(4):719–748.

Massey Jr, F. J. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association* 46(253):68–78.

Munoz, C.; Smith, M.; and Patil, D. J. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. Technical report, Executive Office of the President, The White House.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568. ACM.

Peyré, G., and Cuturi, M. 2019. Computational optimal transport. *Foundations and Trends® in Machine Learning* 11(5-6):355–607.

Podesta, J.; Pritzker, P.; Moniz, E. J.; Holdern, J.; and Zients, J. 2014. Big data - seizing opportunities, preserving values. Technical report, Executive Office of the President, The White House.

Saleiro, P.; Kuester, B.; Stevens, A.; Anisfeld, A.; Hinkson, L.; London, J.; and Ghani, R. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Vaserstein, L. N. 1969. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii* 5(3):64–72.