EXPOSING THE ILLUSION OF FAIRNESS: AUDITING VULNERABILITIES TO DISTRIBUTIONAL MANIPULATION ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Proving the compliance of AI algorithms has become an important challenge with the growing deployment of such algorithms for real-life applications. Inspecting possible biased behaviors is mandatory to satisfy the constraints of the regulations of the EU Artificial Intelligence's Act. Regulation-driven audits increasingly rely on global fairness metrics, with Disparate Impact being the most widely used. Yet such global measures depend highly on the distribution of the sample on which the measures are computed. We investigate first how to manipulate data samples to artificially satisfy fairness criteria, creating minimally perturbed datasets that remain statistically indistinguishable from the original distribution while satisfying prescribed fairness constraints. Then we study how to detect such manipulation. Our analysis (i) introduces mathematically sound methods for modifying empirical distributions under fairness constraints using entropic or optimal transport projections, (ii) examines how an auditee could potentially circumvent fairness inspections, and (iii) offers recommendations to help auditors detect such data manipulations. These results are validated through experiments on classical tabular datasets in bias detection. The code is available at `https://anonymous.4open.science/r/Inspection-76D6/`.

1 INTRODUCTION

Fairness auditing has emerged as a critical practice to ensure that machine learning models comply with ethical and legal standards by not exhibiting discriminatory bias (Barocas et al., 2019; Besse et al., 2022; Oneto & Chiappa, 2020; Wang et al., 2022). High-profile investigative audits, such as the ProPublica analysis of the COMPAS recidivism risk tool, have exposed significant biases against certain demographic groups Angwin et al. (2016). These findings underscored the societal harms of unverified AI systems and prompted calls for regular fairness audits by independent parties Raji et al. (2020). In response, regulators have begun instituting fairness compliance requirements. For instance, the EU's proposed AI Act mandates bias monitoring, and in the U.S., the Disparate Impact DI doctrine (the "80% rule") is used to quantify indirect discrimination in algorithms Feldman et al. (2015). This doctrine requires that the selection rate for a protected group be at least 80% of that of the most favored group. Consequently, the demographic parity metric (also known as statistical parity) has become a standard global fairness criterion which has inspired many mitigation methods, for example, in Hardt et al. (2016b); Gouic et al. (2020); Chzhen et al. (2020). A small demographic parity gap or a ratio above 0.8 is expected for fairness under this rule, along with other metrics such as equalized odds and predictive parity. These metrics provide quantifiable targets for auditors and have been integrated into various auditing toolkits Bellamy et al. (2018); Bird et al. (2020).

We consider an auditing framework in which the auditee submits a sub-sample of their dataset to a regulatory authority, either by providing the algorithm's outputs on that sample or by sharing the sample itself along with API access to the model, allowing the auditor to compute outputs independently. The supervisory body is then responsible for verifying that the submitted sample is sufficiently representative (in terms of distributional distances) of the auditee's complete dataset. This framework is particularly relevant for high-risk systems, where rigorous oversight is required. The supervisory authority may be internal (e.g., a general inspection body) or external, such as the Cour des Comptes for public administration or the ACPR for banking supervision.

Ensuring the good faith of the auditee is critical, as exemplified by the Volkswagen emissions scandal Jacobs & Kalbers (2019).Similar concerns arise in machine learning, where inconsistencies in model behavior during auditing have been documented, such as in the case of Facebook's models discussed in Bourrée et al. (2025). In summary, the auditing framework similar to the one proposed in Fukuchi et al. (2020) involves three distinct entities:

1. **The audited entity**, which provides a subset of its data and, in the context of the audit, grants access to run its algorithm on this data.
2. **The auditor**, who applies standardized procedures to assess whether the dataset and corresponding model outputs satisfy a prescribed fairness criterion, in this case, the DI.
3. **The supervisory authority**, a higher-level body that oversees the integrity of the entire auditing process. It ensures that the audited submits a dataset that is representative of the full underlying data distribution, and that the auditor adheres to accepted auditing protocols.

In this work, our objective is to support supervisory authorities by identifying potential strategies that audited entities might use to circumvent fairness audits, and by providing tools to detect such attempts. Building on the notion of *manipulation-proof* introduced in Yan & Zhang (2022), we show how a dataset that initially violates a fairness criterion, such as Disparate Impact, can be minimally altered to appear compliant, with limited distributional shift as measured by the Kullback–Leibler (KL) divergence or the Wasserstein distance. By systematically analyzing these plausible manipulations, our aim is to raise awareness of audit vulnerabilities and to equip oversight bodies with methods to detect suspicious modifications, thereby strengthening the reliability and robustness of fairness auditing processes. Our contributions are the following:

- We introduce an entropic projection under constraint tool to a new field that is fairness application and auditing, we also build upon this tool to enable constraints on DI.
- We provide mathematical foundations for Wasserstein projection under constraint and implement its application, as well as other Wasserstein-minimizing algorithm, to control distribution shift under fairness constraint.
- We assess whether–and to what extent– a sample from the projected distribution can significantly increase, without being detected by distributional-based statistical tests, the Disparate Impact from 7 unfair tabular datasets. These tests would be used by the supervisory authority to assert the representativeness of the sample.

2 Related Works

Bias mitigation. Achieving fairness under these metrics has prompted extensive research. One line of work proposes *pre-processing* techniques that alter the training data to remove bias. Kamiran & Calders (2009) introduced a "data massaging" approach, flipping class labels of a few selected instances to reduce discrimination. Feldman et al. (2015) proposed repairing datasets by adjusting feature values to remove disparate impact, while Calmon et al. (2017) framed the problem as a convex optimization for probabilistic data transformation. More recently, Celis et al. (2019) introduced a maximum entropy approach to learn fair distributions under statistical constraints. Gordaliza et al. (2019); Del Barrio et al. (2019) or Chakraborty et al. (2024) applied optimal transport (OT) to modify datasets by respectively reweighting the training dataset or reducing the relationship between the sensitive attribute and the covariates, ensuring fairness criteria are met while minimizing divergence from the original distribution. On the other hand, post-processing methods modify the outputs of the model to enforce fairness and let practitioners retrofit fairness to black-box systems. These methods modify the attributes of the individuals to modify global fairness measures such as statistical parity or Equality of opportunity or odds. In Hardt et al. (2016b), authors solves a linear program to find flipping probabilities that equalize FPR and FNR. OT method can also be used to move the outputs towards the Wasserstein barycenter. This post-processing method is proven to be optimal with respect to the accuracy of this model, as discussed in Jiang et al. (2020), Gouic et al. (2020) or Chzhen et al. (2020).

Fair-washing. Ironically, these same tools can be misused to *fake* fairness. A growing body of work highlights how auditees may deliberately manipulate data or outputs to deceive auditors, a

phenomenon called *fair-washing* (Aivodji et al., 2019). One major vulnerability is that global fairness metrics measure the impact of a sensitive attribute on the decision or on the loss of the model. Yet estimating these probabilities require observations from a test sample. Hence these fairness measures depend heavily on the audited sample. Fukuchi et al. (2020) proposed so-called *stealthily biased sampling*, where biased decision-makers curate benchmark datasets that appear fair but mask discrimination in the full data. Their method guarantees that the audited sample passes fairness checks , while remaining close to the biased distribution in a way that is hard to detect.

Another form of fair-washing involves model output manipulation. Aivodji et al. (2019) showed that interpretable proxy models can be trained to mimic the behavior of a black-box model but appear much fairer. These surrogates can be presented as evidence of fairness, while the actual deployed model remains biased. Le Merrer & Trédan (2020) discusses altering decisions during audits, for instance, temporarily approving more minority applicants to artificially satisfy demographic parity. Case studies have confirmed discrepancies between model behavior shown to auditors and that experienced by users Raji et al. (2020), reinforcing the idea that audits based solely on observed data or queries can be manipulated. We note that even explainability tools are vulnerable to exploitation. Slack et al. (2020) and Anders et al. (2020) demonstrated attacks on the explainable methods LIME Ribeiro et al. (2016) and SHAP Lundberg & Lee (2017), generating biased models that appear fair by masking the influence of sensitive attributes. Shamsabadi et al. (2023) examined the theoretical limits of fair-washing detection, showing that under certain conditions, audit evasion may be provably undetectable. Other approaches leverage *external consistency*. Garcia-Borruey et al. (2023) proposed two-source audits, comparing outputs across APIs and user-facing systems to identify inconsistencies indicative of manipulation. Bourrée et al. (2025) suggested using prior knowledge or independent ground-truth data to detect implausible distributions. They provide bounds on the extent of bias an auditee can inject without detection.

In summary, fairness auditing is undergoing an arms race between auditees' capacity to fake compliance and auditors' ability to detect manipulation. Our contribution formalize entropic and optimal transport (OT)-based data transformation methods to simulate audit circumvention and analyze their detectability, offering guidance for designing more resilient auditing frameworks and oversight mechanisms.

3 METHODS

3.1 METHODOLOGY

Statistical parity property ensures that the decision of the algorithm does not depend on the sensitive attribute. In our work we use the well-known Disparate Impact, defined for a model $\hat{Y} = f(X)$ by the ratio $DI(f, Q_n) := \dfrac{\mathbb{P}(\hat{Y} = 1 \mid S = 0)}{\mathbb{P}(\hat{Y} = 1 \mid S = 1)}$. This quantity is equal to 1 when no probabilistic relationship exists between the outcome of the model and the sensitive variable, which implies a strict independence in the case where $f(X)$ is a two-class classification model. Hence, several norms or regulations impose that a model should have its disparate impact greater than a given level t, often set to $t = 0.8$ as chosen originally by EEOC et al. (1978).

In this part, we propose a methodology that enables stakeholders to evade an audit based on the application of a fairness criterion, the Disparate Impact. Our method aims to construct a dataset whose distribution is close to the distribution of the original data, while ensuring that the fairness measure is above a threshold, as required by the regulations. Let $(E, \mathcal{B}(E))$ be a measurable space. Denote by $\mathcal{P}(E)$ and $\mathcal{M}(E)$, respectively the space of probability measures on E and the space of finite measures in E. Consider a distance d in E. In reality, given an empirical distribution $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$ where Z_i is an i.i.d. sample of a random variable with value in E, the construction of a falsely compliant dataset is modeled as finding the solution to the optimization problem : $\operatorname{argmin}_{P \in \mathcal{P}(E), DI(f,P) \geqslant t} d(P, Q_n)$. In the following, we will consider two different distances: in Section 3.2, one is related to the similarity for probabilistic inference (KL information), while in Section 3.3, the other distance captures geometric information between distances (Monge-Kantorovitch a.k.a. Wasserstein distance). Consequently, fair-washing amounts to modify the initial distribution of the data by providing a fake but plausible distribution Q_t in order to achieve that $DI(f, Q_t) = t$ or $DI(f, Q_t) \geqslant t$.

3.2 USING ENTROPIC PROJECTION TO FAKE FAIRNESS

Entropic distributional projection. Set Q a probability measure on E. If P is another probability measure on $(E, \mathcal{B}(E))$, then the KL information is $D_{\mathrm{KL}}(P\|Q) = \int_E \log \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}P$, if $P \ll Q$ and $\log \frac{\mathrm{d}P}{\mathrm{d}Q} \in L^1(P)$, and $+\infty$ otherwise. For any resulting dimension $k \geqslant 1$, let $\Phi : Z = (X, S, \hat{Y}, Y) \in E \mapsto \Phi(X, S, \hat{Y}, Y) \in \mathbb{R}^k$ be a measurable function representing the shape of the stress deformation on the whole input. Note that our results are stated for a generic function Φ of all variables $Z = (X, S, \hat{Y}, Y)$. This includes the case of functions depending only on X, (X, Y) or $(X, \hat{Y})$. We set for two vectors $x, y \in \mathbb{R}^k$ the scalar product as $\langle x, y\rangle = x^\top y$. The problem can be stated as follows: given the distribution Q_n, our aim is to construct a distribution close to Q_n but satisfying a constraint expressed through the mean of the chosen function Φ. Actually, for $t \in \mathbb{R}^k$, we aim at finding a new distribution Q_t satisfying the constraint $\int_E \Phi(x)\,\mathrm{d}Q_t(x) = t$ and being the closest possible to the initial empirical distribution Q_n in the sense of KL divergence, *i.e.* with $D_{\mathrm{KL}}(Q_t\|Q_n)$ as small as possible. The following theorem, whose proof can be found in Bachoc et al. (2023), characterizes the distribution solution Q_t.

Theorem 3.1. *Let $t \in \mathbb{R}^k$ and $\Phi : E \to \mathbb{R}^k$ be measurable. Assume that t can be written as a convex combination of $\Phi(X_1, \hat{Y}_1, Y_1), \ldots, \Phi(X_n, \hat{Y}_n, Y_n)$, with positive weights. Assume also that the empirical covariance matrix $\mathbb{E}_{Q_n}(\Phi\Phi^\top) - \mathbb{E}_{Q_n}(\Phi)\mathbb{E}_{Q_n}(\Phi^\top)$ is invertible.*

Let $\mathcal{D}_{\Phi,t}$ be the set of all probability measures P on E such that $\int_E \Phi(x)\,\mathrm{d}P(x) = t$. For a vector $\xi \in \mathbb{R}^k$, let $Z(\xi) := \frac{1}{n}\sum_{i=1}^n e^{\langle \Phi(X_i, \hat{Y}_i, Y_i), \xi\rangle}$. Define now $\xi(t)$ as the unique minimizer of the strictly convex function $H(\xi) := \log Z(\xi) - \langle \xi, t\rangle$. Then, $Q_t := \operatorname{arginf}_{P \in \mathcal{D}_{\Phi,t}} D_{\mathrm{KL}}(P\|Q_n)$ (1)

exists and is unique. It can also be computed as $Q_t = \frac{1}{n}\sum_{i=1}^n \lambda_i^{(t)} \delta_{X_i, \hat{Y}_i, Y_i}$,

with, for $i \in \{1, \ldots, n\}$, $\lambda_i^{(t)} = \exp\Big(\langle \xi(t), \Phi(X_i, \hat{Y}_i, Y_i)\rangle - \log Z(\xi(t))\Big)$.

Faking Statistical Parity using Entropic Projection. Let t_{init} such that $DI(f, Q_n) = t_{\text{init}}$. We aim at building a distribution Q_t such that $DI(f, Q_n) = t_{\text{new}} \geqslant t_{\text{init}}$ for a given t_{new}. Define the fairness improvement $\Delta_{DI} := DI(f, Q_t) - DI(f, Q_n)$. Note that $DI(f, Q_n) = \frac{\lambda_0/n_0}{\lambda_1/n_1}$ where for $i \in \{0, 1\}$, $n_s = |\{i = 1, \ldots, n | S_i = s\}$ and $\lambda_s = |\{i = 1, \ldots, n | \hat{Y}_i = 1 \wedge S_i = s\}|$. Note also that $\lambda_0 = \sum_{i=1}^n \hat{Y}_i(1 - S_i)$ and $\lambda_1 = \sum_{i=1}^n \hat{Y}_i S_i$. Hence modifying the DI can be achieved applying Theorem 3.1 for $Z = (S, \hat{Y})$ and selecting the function

$$\Phi(s, f(x)) = \begin{pmatrix} (1-s)f(x) \\ sf(x) \\ s \\ 1-s \end{pmatrix} \text{ and } m = \begin{pmatrix} \lambda_0 + \delta_0 \\ \lambda_1 - \delta_1 \\ n_1 \\ n_0 \end{pmatrix} \tag{2}$$

Our purpose is to improve the perceived fairness of the model. Accordingly, we only consider increasing the numerator $+\delta_0 \geqslant 0$ and decreasing the denominator $-\delta_1 \leqslant 0$.

Proposition 3.2 (KL-fair washing method)**.** *Finding a solution Q_t such that $D_{\mathrm{KL}}(Q_t\|Q_n)$ is minimum and $DI(f, Q_t) = DI(f, Q_0) + \Delta_{DI}$ is achieved by finding the solution to equation 1 with Φ defined as in equation 2 and with the two possible choices of parameters:*

- *Balanced case : set $\delta_0 = \delta_1$ and $\delta_1 = \dfrac{\lambda_1}{1 + \frac{n_1}{n_0 \Delta_{DI}}(1 + \frac{\lambda_0}{\lambda_1})}$*
- *Proportional case : set $\dfrac{\delta_0}{n_0} = \dfrac{\delta_1}{n_1}$ and $\delta_1 = \dfrac{\lambda_1}{1 + \frac{1}{\Delta_{DI}}\big(1 + \frac{n_1 \lambda_0}{n_0 \lambda_1}\big)}$*

Remark 3.1. *The balanced case corresponds to modifying the individuals from both classes equally, while the proportional one adjusts the amount of modification in proportion to the classes sizes.*

If the target value t_{new} is chosen according to the balanced case or the proportional case, we refer respectively in the Experimental section to the method as `Entropic_balanced` and `Entropic_proportional`.

3.3 FAIR-WASHING USING OPTIMAL TRANSPORT.

Monge Kantorovich (MK) Projection For two distributions P and Q_n over $E \subset \mathbb{R}^d$ a compact subset, endowed with the norm $\|.\|$, recall that their 2 Monge-Kantorovich, a.k.a. Wasserstein distance, is defined as:

$$W_2^2(P, Q_n) = \min_{\pi \in \Pi(P,Q_n)} \int_{x \in E, y \in E} \| x - y \|^2 \, d\pi(x, y), \tag{3}$$

where $\Pi(P, Q_n)$ denotes the set of distributions on $E \times E$ with marginals P and Q_n. We will write $T_\sharp Q = Q \circ T^{-1}$ to denote the push-forward of a measure by the transport map. As in Section 3.2, consider for a given $k \geqslant 1$, a continuous function $\Phi \colon E \to \mathbb{R}^k$ representing the constraints. For fixed $t \in \mathbb{R}^k$, the set $\mathcal{D}_{\Phi,t} = \{P \in \mathcal{M}(E) \mid \int_E \Phi(x) dP(x) = t\}$ is closed for the weak convergence and convex, since it is linear in P. The function $P \mapsto W_2^2(P, Q_n)$ is convex as it is the supremum of linear functionals by Kantorovich duality (see Santambrogio (2015)), therefore the following projection problem $\operatorname{arginf}_{P \in \mathcal{D}_{\Phi,t}} W_2^2(P, Q_n)$ is well-defined.

Theorem 3.3. *Consider $Q_n = \frac{1}{n} \sum_{i=1}^n \delta_{Z_i}$. Then Q_t is a solution to $\operatorname{arginf}_{P \in \mathcal{D}_{\Phi,t}} W_2^2(P, Q_n)$ if, and only if, it is defined as $Q_t = T_{\lambda^\star} \sharp Q_n = \dfrac{1}{n} \sum_{i=1}^n \delta_{T_{\lambda^\star}(Z_i)}$, where T_λ is defined as*

$$T_\lambda(Z_i) \in \arg\min_{x \in E} \|x - Z_i\|^2 - \langle \lambda, \Phi(x) \rangle \tag{4}$$

and $\lambda^\star$ satisfies $t = \frac{1}{n} \sum_{i=n}^n \Phi(T_{\lambda^\star}(Z_i))$.

Remark 3.2. *The previous result stated for the empirical distribution is valid for any distribution Q. The constraint can be modified to include the condition $\mathcal{D}_{\Phi,t} = \{P \in \mathcal{M}(E) \mid \int_E \Phi(x) dP(x) \geqslant t\}$. This is detailed in Proposition D.1 in the Appendix.*

Faking Statistical Parity using MK projection. The objective is to construct a fake dataset drawn from a distribution Q_t defined as the solution to $Q_t = \operatorname{arginf}_{P \in \mathcal{D}_{\text{DI},t}} W(Q_n, P)$ with $\mathcal{D}_{\text{DI},t} = \{P \in \mathcal{P}(E), DI(P) \geqslant t\}$. Following the framework of the previous section, for a fixed λ, we set $\Phi(x, s, f(x), y) = f(x)$. Then the constraint on the Disparate Impact $DI(Q) \geqslant t$, can be reformulated with the double inequality $\int_{x \in E | s=0} f(x) dQ(x) \geqslant t_0 = t + \delta_0$ and $\int_{x \in E | s=1} f(x) dQ(x) \leqslant t_1 = t - \delta_1$ with $\frac{t_0}{t_1} \geqslant t$. As such, we divide the dataset for each $s \in \{0, 1\}$, and consider $Q_t = \pi Q_{t,1} + (1 - \pi) Q_{t,0}$ with $\pi = \mathbb{P}_{Q_n}(S = 1)$ and $Q_{t,s} := \operatorname{arginf}_{P \in \mathcal{D}_{\text{DI},t_s}} W(Q_{n,s}, P)$ with the conditional distributions $Q_{n,s} := Q_n(\cdot \mid S = s)$, $\mathcal{D}_{\text{DI},t_0} = \{P \in \mathcal{P}(E), DI(P) \geqslant t_0\}$ and $\mathcal{D}_{\text{DI},t_1} = \{P \in \mathcal{P}(E), DI(P) \leqslant t_1\}$.

Following Theorem 3.3, we compute for all $x = Z_i$, the solution $T_\lambda(x)$ of the minimization problem w.r.t x: $\mathcal{L}(x, \lambda) = \|Z_i - x\|_2^2 + \langle \lambda, t - f(x) \rangle$. (5)
This minimization does not have a closed form in general, but it can be achieved using a gradient descent using a learning step of η and computing $x^t = x^{t-1} - \eta \frac{\partial \mathcal{L}}{\partial z}(x^t)$ with $\frac{\partial \mathcal{L}}{\partial x} = 2(z - x) - \langle \lambda, \nabla_x f(x) \rangle$. We point out that this method requires knowledge of the gradients of the classifier, which will be estimated at each step of the method. To complete the method's explanation, we need to clarify how to choose t_0 and t_1 and λ:

1. The choice of t_0 and t_1 is explained in Section 3.2: balanced or proportional case.
2. λ is a constraint regulation coefficient, meaning that the bigger λ is, the more the optimization solution will take into account the constraint $\int_{x \in E | S=s} f(x) dQ(x) \leqslant t_s$. And consequently, the bigger λ is, the farther the solution will be from the original distribution. Therefore, we start by solving equation 5 with a low λ, and we increase it until the constraint is respected.

This method creates new individuals without any constraint on the covariates X, this might be an issue as this implies no restriction of types (discrete variable staying discrete, i.e., age $= 1.002$) or of bounds (age $= -1$). Thereby, we created a variant of this method that constrains the achievable covariates: we transport, variable per variable, each covariates toward the nearest (for the L_2 norm) achieved value in the dataset; we call this variant the 1D-transport variant.

Remark 3.3. *Note that we chose to modify the output of the model, $f(x) \in \{0, 1\}$. For practical purposes, to know when the convergence is attained, we look at the logits of the neural network instead of the binary values: after a sigmoid, $f(x) \in [0, 1]$. We could therefore apply our constraint*

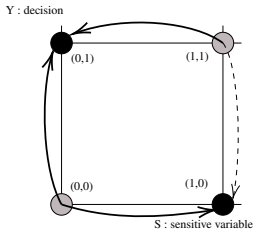


Figure 1: Admissible modifications on $\tau : \{0,1\}^2 \mapsto \{0,1\}^2$ increasing Disparate Impact

Algorithm 1 `Replace(`S`,`$\hat{Y}$`)` algorithm

$Z^j = (Z_1, \cdots, Z_n), Z_i = (S_i, \hat{Y}_i), t \in]0,1[$
$\tau_i := (\tau, j)$ such as $\tau \in \mathcal{A}, i \in 1, \cdots, n$
while $\text{DI}(Z^j) < t$ **do**
  $\tau_{i_0} \in \text{argmax } \text{DI}(\tau_{i_0}(Z^j)) - \text{DI}(Z^j)$
with $\tau_{i_0}(Z^j) := (Z_i, \cdots, \tau(Z_{i_0}), \cdots, Z_n)$
  $Z^j \leftarrow Z^{j+1} = \tau_{i_0}(Z^j)$
end while
return Z^j

on the logits, which highlights the ability to use these methods for non-binary tasks, for instance, in regression settings. The constraints are imposed separately on the squared Wasserstein distances $W_2^2(Q_{n,1}, Q_{t,1})$ and $W_2^2(Q_{n,0}, Q_{t,0})$. The inequality $W_2^2(\pi Q_{n,1} + (1-\pi)Q_{n,0}, \pi Q_{t,1} + (1-\pi)Q_{t,0}) \leqslant \pi W_2^2(Q_{n,1}, Q_{t,1}) + (1-\pi)W_2^2(Q_{n,0}, Q_{t,0})$ provides an upper bound on the overall distance between the two samples. The proof of this result is deferred to Appendix E.4.

To summary, we had introduced four methods: (1) `Grad_balanced`, and (2) `Grad_proportional`, which differ based on the gradient constraints satisfying $\delta_0 = \delta_1$ or $\frac{\delta_0}{n_0} = \frac{\delta_1}{n_1}$; and (3) `Grad_balanced_1D-transport`, and (4) `Grad_proportional_1D-transport`, which apply the corresponding 1D-transport variant of each method.

Faking Statistical Parity using sensitive attributes replacement. For this method, we consider that the auditor does not have access to the model f and only request the outcome of the algorithm $\hat{Y}$, without computing it from the observations $f(X)$. Hence, faking fairness can be achieved by manipulating only the outcomes and sensitive attributes associated with each individual. Let $Z_i = (S_i, \hat{Y}_i)$ and $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{S_i,\hat{Y}_i}$. Consider the optimization problem : $\text{arginf}_{P \in \mathcal{D}_{\text{DI},t}} W_2^2(Q_n, P)$ with $\mathcal{D}_{\text{DI},t} = \{P \in \mathcal{P}(E), DI(Q_t) \geqslant t\}$. A solution can be achieved as follows. Note that $\hat{Y} \in \{0,1\}$ and $S \in \{0,1\}$, thus we only have 4 possible values for the points. Each individual with characteristic $Z_i \in \{0,1\}^2$ can be modified to the individual $\tau(Z_i) = (\tau_S(S_i), \tau_{\hat{Y}}(\hat{Y}_i)) \in \{0,1\}^2$. We first point out that not all solutions improve the disparate impact and we can restrict ourselves to a set of admissible changes $\tau \in \mathcal{A}$ as pointed in Fig. 1, with more details explaining why are in Section H.1 in the Appendix. Then iteratively we approximate the exact solution by an iterative method starting from $Z = (Z_1, \cdots, Z_n)$ and testing every possible modification $Z^j = (Z_1, \cdots, Z_n)$ maximizing the DI at each step j. The method based on this algorithm is denoted by `Replace(`S`,`$\hat{Y}$`)` in our experiments.

Faking Statistical Parity using constrained matching. In the previous case, the observations X_i are not taken into account. A natural variant consists in combining this minimization scheme and adding a discrete displacement on the variables X. Namely, we define a matching algorithm using $Z = (X, S, \hat{Y})$ and $\tau(Z_i) = Z_k$, with $k \in \{1, \cdots, n\}$. We use the same proceedings as Alg. 1 with the newly defined τ, but at every iteration j of the while loop we maximize for every candidate τ_{i_0}: $\dfrac{\text{DI}(\tau_{i_0}(Z^j)) - \text{DI}((Z^j))}{\|\tau_{i_0}(Z^j) - Z^j\|}$.

In our experiments, we refer to the method based on this algorithm as $M_{W(X,S,\hat{Y})}$.

Remark 3.4. *This algorithm transports individuals towards others ($\tau(Z_i) = Z_k$), therefore, contrary to its counterpart, it can be used in any type of audit (with or without access to the model).*

3.4 Method detection: Statistical tests

We outline below potential strategies a supervisory authority could adopt to assess whether the auditee conducted compliance tests using a sample drawn from the original data distribution. The auditee presents a sample $\mathcal{D}_{n,t}$, drawn from a distribution $Q_{n,t}$. To verify the authenticity of this sample, the authority must be granted access to the full dataset upon request. This access enables the

Table 1: Dataset presentation, sensitive variable (S) associated, and original Disparate Impact (DI)

	Adult	INC	TRA	MOB	BAF	EMP	PUC
S chosen	Sex	Sex	Sex	Age	Age	Disability	Disability
DI	0.30	0.67	0.69	0.45	0.35	0.30	0.32

authority to infer the ground-truth distribution and determine whether the submitted data has been manipulated or follows the initial distribution Q_n. To assess representativeness, the authority must rely on statistical testing. Two main categories of tests are available. The first includes hypothesis tests that evaluate, at a chosen confidence level, whether the distribution of the submitted sample $\mathcal{D}_{n,t}$ is statistically similar to the original distribution Q_n. In their study Fukuchi et al. (2020), the authors apply a Kolmogorov–Smirnov (KS) test for one-dimensional data ($X \in \mathbb{R}^1$), and a test based on the Wasserstein distance for higher-dimensional settings ($X \in \mathbb{R}^k$, with $k > 1$). In our framework, we apply both the KS test and the Wasserstein test on the conditional distribution $\hat{Y} \mid S$.

The second approach evaluates whether the sample $\mathcal{D}_{n,t}$ could plausibly result from a random draw from the original distribution Q_n, by measuring a divergence or distance metric d. The idea is to test whether the observed value $d(\mathcal{D}_{n,t}, Q_n)$ lies within the $(1-\alpha/2)$ confidence interval of $d(Q_{n,t}, Q_n^\sigma)$, where Q_n^σ represents a reference sample drawn from the original distribution. For the distance metric d, we considered several options, including the Maximum Mean Discrepancy (MMD) Gretton et al. (2012), the Wasserstein distance, and the Kullback–Leibler (KL) divergence.

Extension to non tabular data: The method we develop is originally meant to handle tabular data but we could use it directly on images or text flattened as vectors. Yet, using as previously the L^2 distance between individuals, might not be the natural way to capture semantic similarity between images or token distributions. A way to circumvent this issue is to represent the images in another space, where the regular distances would have semantic meanings. The construction of such a space has already seen numerous works using PCA projections or latent spaces of AE, VAE or CNN classifiers. We present such results on the CelebA dataset Liu et al. (2015) in Section C of the Appendix.

4 Experiments

4.1 Experimental settings

Datasets. We use 7 benchmarking datasets : Adult Census Income dataset where Y is whether an individual's income is above 50k (Adult) Becker & Kohavi (1996). We also use 5 benchmark datasets from Ding et al. (2021) which records information about the USA's population, including income (INC), mobility (MOB), employment (EMP), travel time to work (TRA) and public system coverage (PUC). We also include the Bank Account Fraud generated dataset (BAF) from Jesus et al. (2022). We refer to Table 1 for the sensitive variable and the original Disparate Impact (DI) of each dataset.

Neural network predictions. As we are working only with tabular data, we provide a $\hat{Y}$ with a multilayer perceptron (MLP) neural network f ending with a sigmoid activation function ($f(x) \in [0,1]$). While having the best prediction accuracy was not the goal of experiments, we still achieve reasonable accuracy learning with the *ScheduleFree* optimizer Defazio et al. (2024). We defined the logit threshold based ground truth mean : $l_{th} := \min_{l \in]0,1[} |\mathbb{E}(\hat{Y}_l) - \mathbb{E}(Y)|$ with $\hat{Y}_l = \{f(x) > l | x \in \mathcal{D}\}$. This was especially necessary for the BAF dataset, where the learning task is basically an anomaly detection task, and $\mathbb{E}(Y) \approx 0.01$.

4.2 Results

Fairness cost: distribution shift per method. Fig 2 illustrates the comparative performance of each method across different distance metrics (D_{KL}, W). Specifically, these metrics quantify the extent of distributional change, and help assess each method's ability to evade detection by the statistical tests. We also provide complementary results on simulated data and the computation time and memory cost of each method in the Appendix (see Sec F). The smallest surface area in the radar chart is archived by $M_{W(X,S,\hat{Y})}$, hence, given the results, this method appears to be the most suitable method

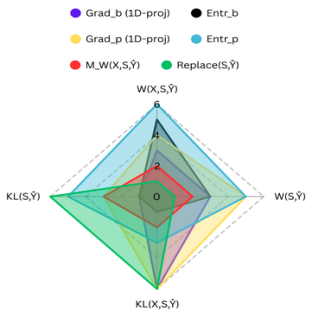


Figure 2: Radar graph ranking the optimization result depending on the fair-washing method. This graph shows why $M_{W(X,S,\hat{Y})}$ is the most promising method.

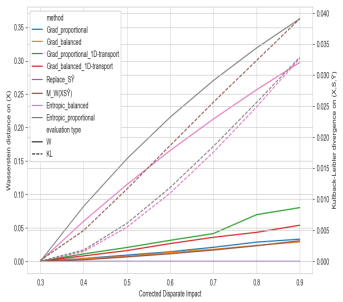


Figure 3: Line plot showing the trade-off between fairness correction and distribution shift on the Adult dataset. Wasserstein Distance on the individual's characteristics X, and global KL divergence depending on the fairness correction per method.

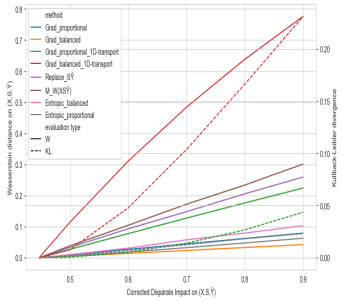


Figure 4: Line plot showing the trade-off between fairness correction and distribution shift on the Mobility dataset. Global Wasserstein distance ($(X, S, \hat{Y})$, and global Kullback-Leibler divergence depending on the fairness correction per method.

Table 2: Results of the 7 tests independently; for each unbiasing method (DI=0.8) and datasets. Sampling is stopped as soon as one sample satisfies the $\mathcal{H}_0$ hypothesis, or after 30 tries if none do. The symbol $-$ means the method was undetected by the test for both sampling sizes of 10% and 20% ($\mathcal{H}_0$ accepted), ○ means that only the 20% sampling size was undetected ($\mathcal{H}_0$ accepted for 20% and rejected for 10%); and ◎ means that the method was detected at both 10% and 20% sample sizes ($\mathcal{H}_0$ rejected). Positional and color coding indicate which test each result corresponds to, in the following order and color scheme: KL$(X, S, \hat{Y})$, KL$(S, \hat{Y})$, $W(X, S, \hat{Y})$, $W(S, \hat{Y})$, K-S$(\hat{Y})$, MMD$(X, S, \hat{Y})$, MMD$(S, \hat{Y})$. `Grad_proportional` (Grad_p) and `Grad_balanced` (Grad_b) have been merged with their 1D counterpart due to identical test results.

	Methods					
Dataset	Grad_p(1D-t)	Grad_b(1D-t)	Rep $(S, \hat{Y})$	$M_{W(X,S,\hat{Y})}$	Entropic_b	Entropic_p
ADULT	◎○−◎◎−◎	◎○−◎◎−◎	◎○−◎◎−◎	○○−◎◎○◎	−○−◎◎−◎	−○−◎◎−◎
EMP	◎◎○◎◎◎◎	◎◎○◎◎○◎	◎○−◎◎−◎	◎○−◎◎○◎	○◎○◎◎○◎	○○○◎◎−◎
INC	◎−−−−−−	◎−−−−−−	◎−−−−−−	−−−−−−−	−−−−−−−	−−−−−−−
MOB	◎◎○◎◎◎◎	◎○−◎◎−◎	◎○−◎◎−◎	◎○−◎◎○◎	○◎○◎◎○◎	○○○◎◎◎◎
PUC	◎◎○◎◎◎◎	◎◎○◎◎○◎	◎◎○◎◎○◎	◎○○◎◎◎◎	○◎○◎◎○◎	○○○◎◎◎◎
TRA	◎−−−−−−	◎−−−−−−	◎○−−−−−	−−−−−−−	−−−−−−−	−−−−−−−
BAF	◎○−−−−−	◎○−−−−−	◎−−−−−−	−−−−−−−	−−−−−−−	−○−−−−−

for someone seeking to disguise their dataset, as it significantly improves the DI while preserving a distribution close to the original data.

Fraud detection through distributional shifts. In the case of a compliant audit, is it possible to detect a fraud attempt based on the provided sample and the original dataset? Based on our previous methods we create a falsely compliant sample from several Q_t (with DI$(Q_t) \geqslant 0.8$). More precisely, we randomly sample from theses Q_t, and apply the seven distinct statistical tests explained in Section 3.4, which verify the null hypothesis $\mathcal{H}_0$; which is that the original and sampled manipulated distributions are equivalent. Table 2 reports the results based on a modified sample of 10% or 20% and the original dataset. Additional details on the probability of passing the tests are provided in Section K.1 in the Appendix. Methods modifying individual characteristics (`Grad` methods) are easily detected (rejection of $\mathcal{H}_0$) regardless of the sampling size. The fair-washing done by the $M_{W(X,S,\hat{Y})}$ and Entropic-based methods is undetected for the INC, TRA and BAF datasets. For the TRA and INC datasets, the DIs of the original data were close to that of the modified data (see Table 3), implying that the required modifications were minimal and therefore difficult to detect. For the BAF dataset, we remind that $\mathbb{E}(Y) \approx 0.01$, as a result, only limited modifications were also needed in this case.

Table 3: Highest undetected achievable Disparate Impact for each dataset, sample size (S Size) and fair-washing method. The symbol – indicates that some methods couldn't reach a DI improvement. To emphasize the best method to use in order to deceive the auditor, we put the DI achieved in bold when one or two overperformed the others.

Dataset	Original	S size (%)	Grad_p	Grad_b	Grad_p 1D	Grad_b 1D	Rep $(S, \hat{Y})$	Entr_b	Entr_p	$M_{W(X,S,\hat{Y})}$
ADULT	0.30	10	0.47	0.43	0.49	0.44	0.50	**0.54**	0.52	**0.54**
		20	0.39	0.40	0.38	0.39	0.41	**0.42**	0.41	**0.42**
EMP	0.30	10	–	–	–	–	–	0.36	0.36	**0.37**
		20	–	–	–	–	–	0.34	**0.36**	0.35
INC	0.67	10	0.75	–	–	–	0.88	0.94	**0.95**	0.93
		20	–	–	–	–	0.83	0.83	**0.84**	**0.84**
MOB	0.45	10	0.53	0.51	–	0.50	**0.53**	0.52	–	0.52
		20	–	–	–	0.48	0.50	0.50	–	0.50
PUC	0.32	10	–	–	–	–	–	0.33	**0.35**	**0.35**
		20	–	–	–	–	–	–	–	–
TRA	0.69	10	0.72	0.79	0.77	0.73	0.80	0.83	0.84	**0.85**
		20	–	–	–	–	0.77	0.79	0.79	**0.80**
BAF	0.35	10	–	–	–	–	–	1	1	1
		20	–	–	–	–	–	0.77	**0.80**	0.79

Trade-off: DI improvement vs distribution shift. Fig. 3 and Fig. 4 illustrates the trade-off between fairness correction and distribution shift on the Adult and Mobility datasets by the Wasserstein distance and KL divergence between the full original and modified distributions. `Replace` $(S, \hat{Y})$, $M_{W(X,S,\hat{Y})}$ and `Grad` variant methods preserve the structure of the input space and are better alternatives to the entropic projection method. We recall that since `Replace` only modifies S and $\hat{Y}$, it naturally leads to the lower difference between distributions.

Fairest undetected sample. When considering Table 2 results, which presents the inability to cheat the supervisory authority in the Adult, EMP, MOB and PUC datasets with $DI(Q_t) \geqslant 0.8$, we study based on the previously mentioned trade-off the highest undetected fairness correction we are able to achieve. Table 3 presents the highest Disparate Impact (DI) values not rejected by the combination of seven statistical tests, selected from 100 random samples comprising 10% and 20% of the modified dataset, respectively. The four `Grad` variants and `Replace` methods failed to secretly increase the DI on three common datasets (EMP, PUC, and BAF), primarily due to consistent rejection by the KL based tests (100% rejection rate). This highlights the advantage of `Entropic` methods and the matching approach $M_{W(X,S,\hat{Y})}$, which avoids scenarios where $D_{\mathrm{KL}}(Q_t \| Q_n) = +\infty$, and thus performs more robustly. Overall, $M_{W(X,S,\hat{Y})}$ consistently achieves the best results. The results also underscore the influence of sample size: for example, in the PUC dataset, improvements were possible with 10% samples but not with 20%, indicating that as the sample size increases (relative to the original dataset), it becomes more challenging to increase DI without detection, we further highlight this point in Fig 8 in the Appendix.

5 Conclusion and Perspectives

This work presents a comprehensive study of methods designed to manipulate data in order to satisfy the Disparate Impact (DI) criterion. We provide a theoretical analysis demonstrating why these methods can minimize the distance between the original and modified data distributions while satisfying fairness constraints. Our findings are supported by experiments on both simulated and real benchmark datasets. Our results show that with the recursive Wasserstein-minimizing matching method, $M_{W(X,S,\hat{Y})}$, an auditee can very likely increase fraudulently the Disparate Impact without being detected by rigorous statistical tests. Hence supervisory authorities should be aware of the possibility that datasets may have been intentionally manipulated. Their countermeasure is first to use multiple statistical tests combining different geometrical properties of the distributions as shown in Table 3.We showed that a second option is to increase the sample size required from the auditee. Our study focused on tabular data, but the approach extends to text and images when applied to higher-level representations (descriptors), as is common in evaluating generative models Heusel et al. (2018). We provided preliminary results in this direction and leave a more comprehensive exploration to future work.

Concluding Ethical Statement

This work explores the potential for malicious actors to manipulate dataset samples in order to falsely appear compliant with fairness regulations, specifically with respect to Disparate Impact, with possible extensions to other various fairness metrics. Our primary objective is to expose and analyze these vulnerabilities. We believe that research into adversarial strategies is essential to improving the robustness and reliability of fairness auditing procedures. By providing detailed methods for faking compliance, alongside statistical tests for detection, our intent is to support supervisory authorities and auditors in developing more resilient oversight mechanisms. We emphasize that our findings are not intended to be used as tools for deceptive practices. To this end, we have deliberately omitted full implementation details that would lower the barrier to misuse, and we have focused our analysis on defensive strategies available to regulators. Moreover, the public release of our code is designed to assist the research community in building stronger auditing tools, not to enable audit circumvention. We encourage regulators and institutions to develop governance frameworks that anticipate such adversarial behavior and recommend routine adoption of statistical tests to verify the representativeness of audit samples.

Reproducibility statement

The algorithms corresponding to each proposed fair-washing method are detailed in the paper. For the simplified versions of the Replace (S, Y) and $M_{W(X,S,Y)}$ methods, we refer to Alg. 1 in the main paper. The full, non-simplified version is provided in Alg. 2 in the Appendix. Additionally, the Wasserstein-based gradient optimization fair-washing methods are described in Alg. 3, also in the Appendix.

Our experiments including our simulated dataset, the publicly available datasets we use Becker & Kohavi (1996); Ding et al. (2021); Jesus et al. (2022); Liu et al. (2015) and the code to reproduce exactly every result shown in this paper thanks to (1) seed setting and (2) intermediary results registered are available at `https://anonymous.4open.science/r/Inspection-76D6/`.

Our Github repository is structured as such:

- Data: datasets folder (with mostly csv files)
- Pre-processing: Jupyter notebooks.
- Src: python functions which includes our fair-washing methods.
- Project: Network training and inference, fairness evaluation, fair-washing and fraud detection using statistical tests.
- Result: Final and intermediary results (csv, npy, json files).

Github repository limits at 50Mo, hence we uploaded the rest (csv and numpy matrices) on Google Drive. It will be made available as soon as the double peer-review process ends.

REFERENCES

Ulrich Aivodji, Paolo Alesi, Sébastien Gambs, Kévin Huguenin, Salvador R Martínez, and Matthieu Roy. Fairwashing: the risk of rationalization. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 160–166, 2019.

Cedric Anders, Solon Barocas, Jonathan Zittrain, Pushmeet Kohli, and Manuel Gomez-Rodriguez. Fairwashing explanations with optimal transport. https://arxiv.org/abs/2006.05241, 2020. arXiv:2006.05241.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

François Bachoc, Fabrice Gamboa, Max Halford, Jean-Michel Loubes, and Laurent Risser. Explaining machine learning models using entropic variable projection. *Information and Inference: A Journal of the IMA*, 12(3):1686–1715, 05 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad010. URL https://doi.org/10.1093/imaiai/iaad010.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://fairmlbook.org.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, and et al. Ai fairness 360 open source toolkit, 2018. https://aif360.mybluemix.net.

Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 76(2):188–198, 2022.

Sarah Bird, Drew Dimmery, and Kush R. Varshney Walker. Fairlearn: A toolkit for assessing and improving fairness in ai, 2020. https://fairlearn.org.

Jade Garcia Bourrée, Augustin Godinot, Martijn De Vos, Milos Vujasinovic, Sayan Biswas, Gilles Tredan, Erwan Le Merrer, and Anne-Marie Kermarrec. Robust ml auditing using prior knowledge, 2025. URL https://arxiv.org/abs/2505.04796.

Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Fair distributions from biased samples. In *International Conference on Machine Learning (ICML)*, pp. 1365–1374. PMLR, 2019.

Sayantan Chakraborty, Steve Oudot, and Nicolas Vayatis. Constrained reweighting of distributions: an optimal transport approach. https://arxiv.org/abs/2310.12447, 2024. arXiv:2310.12447.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33: 7321–7331, 2020.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled, 2024.

Eustasio Del Barrio, Paula Gordaliza, and Jean-Michel Loubes. A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8(4):817–849, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=bYi_2708mKK.

EEOC et al. (1978). Section 4d, uniform guidelines on employee selection procedures, 1978.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Kazuto Fukuchi, Satoshi Hara, and Takanori Maehara. Faking fairness via stealthily biased sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 412–419, 2020.

Diego Garcia-Borruey, Patrick Loiseau, and Erwan Le Merrer. Mitigating fairwashing using two-source audits. https://arxiv.org/abs/2305.13883, 2023. arXiv:2305.13883.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pp. 2357–2365. PMLR, 2019.

Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016a. URL http://arxiv.org/abs/1610.02413.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3315–3323, 2016b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500.

Daniel Jacobs and Lawrence P. Kalbers. Volkswagen emissions scandal, 2019. URL https://www.cpajournal.com/2019/07/22/9187/.

Sérgio Jesus, José Pombal, Duarte Alves, André Cruz, Pedro Saleiro, Rita P. Ribeiro, João Gama, and Pedro Bizarro. Turning the tables: Biased, imbalanced, dynamic tabular datasets for ml evaluation. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=bYi_2708mKK.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 862–872. PMLR, 22–25 Jul 2020.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, pp. 1–6. IEEE, 2009.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929, 2012. doi: 10.1109/ICDM.2012.45.

Erwan Le Merrer and Gilles Trédan. Faking fairness via model manipulation. In *Companion Proceedings of the Web Conference 2020*, pp. 661–665, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6360–6369. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lohaus20a.html.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Luca Oneto and Silvia Chiappa. Fairness in machine learning. In *Recent trends in learning from data: Tutorials from the INNS big data and deep learning conference (INNSBDDL2019)*, pp. 155–196. Springer, 2020.

J. Peypouquet. *Convex Optimization in Normed Spaces*. Springer Cham, 2015. URL https://api.semanticscholar.org/CorpusID:124131607.

Inioluwa Deborah Raji, Andrew Smart, Rebecca White, and et al. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, pp. 33–44. ACM, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs and Modeling*. Birkhäuser Cham, 2015. URL https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf.

Azade Shamsabadi, Amal Douzal-Chouakria, Giuseppe Contissa, and Bruno Lepri. Is fairwashing provably undetectable? https://arxiv.org/abs/2303.10427, 2023. arXiv:2303.10427.

Dylan Slack, Sam Hilgard, Emily Jia, Sorelle Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. https://arxiv.org/abs/1911.02508, 2020. arXiv:1911.02508.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567.

Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022. ISSN 2731-5843. doi: 10.1007/s44176-022-00006-z. URL https://doi.org/10.1007/s44176-022-00006-z.

Yudai Yamamoto and Satoshi Hara. Fast stealthily biased sampling using sliced wasserstein distance. In Vu Nguyen and Hsuan-Tien Lin (eds.), *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pp. 873–888. PMLR, 05–08 Dec 2025.

Tom Yan and Chicheng Zhang. Active fairness auditing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24929–24962. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yan22c.html.

Appendix

A OTHER FAIRNESS METRICS

In our paper, we focused on the Disparate Impact (DI) fairness metric, as it is one of the most widely used metrics. While this choice is justified, it is natural to wonder whether our results are specific to this metric or whether they are metric-agnostic.

Our fair-washing method could have been implemented to minimize the distribution shift while being constrained to other global fairness metrics as long as we can write them as an integrable function or a combination of integrable functions. This condition is not very restrictive in our case. In fact, it only excludes the individual fairness metric, whereas most global fairness metrics can still be expressed in the required form.

To prove this point, we decided to implement our best-performing method, the $M_{W(X,S,\hat{Y})}$ for the Equality of Odds (EoO) metric :

$$\text{EoO} = |\mathbb{P}(\hat{Y} = 1|S = 1 \wedge Y = 1) - \mathbb{P}(\hat{Y} = 1|S = 0 \wedge Y = 1)|$$

Note that similarly to the Disparate Impact, which is the multiplicative counterpart of the Disparate Parity, we could have taken the multiplicative definition of the EoO. However, we choose the additive definition because the multiplicative case is trivial for us, as we could have directly applied our DI-fair-washing method on the $Q_{n,Y=1}$.

The only difference to the matching method $M_{W(X,S,\hat{Y})}$ going from DI constraint to EoO constraint is, following the notation of Section 3.3, iteratively from maximizing the left part of Eq. 6 to its right part.

$$\frac{\text{DI}(\tau_{i_0}(Z^j)) - \text{DI}((Z^j))}{\|\tau_{i_0}(Z^j) - Z^j\|} \rightarrow -\frac{\text{EoO}(\tau_{i_0}(Z^j)) - \text{EoO}((Z^j))}{\|\tau_{i_0}(Z^j) - Z^j\|} \tag{6}$$

The minus sign comes from the difference between the fairness metric : an independence toward the sensitive variable S for the Disparate Impact implies $DI = 1$, we therefore try to maximize the DI. On the other hand, independence for the EoO implies $EoO = 0$, leading us to minimize this criterion (i.e., maximizing minus the criteria). We illustrate this capacity in Fig. 5.

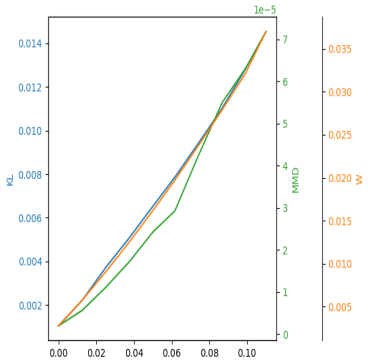


Figure 5: Distribution shift of Wasserstein distance, KL divergence and MMD, when constraining the Equality of Odds (EoO) metric on the Adult dataset using the $M_{W(X,S,Y)}$ method minimizing the EoO.

B COMPARISON WITH OTHER FAIR-WASHING METHODS

B.1 BIAS MITIGATION METHODS AS FAIR-WASHING METHODS

As discussed in the introduction, bias mitigation methods can also be misused to artificially improve fairness metrics, thereby creating the illusion of fairness while concealing underlying biases. Several such methods have been proposed in the literature, including the approach presented in Bourrée et al. (2025), that we reference here, such as ROC Mitigation Kamiran et al. (2012), Optimal Label Transport (OT-L) Jiang et al. (2020), Linear Relaxation (LinR) Lohaus et al. (2020), and Threshold Manipulation (ThreshOpt) Hardt et al. (2016a).

We directly compare these approaches to the `Replace` $(X, S, \hat{Y})$ method. This method, like the others mentioned, modifies the decision-making process based on the sensitive attribute S, for example by applying different decision thresholds conditioned on S. This common dependency on S makes these methods similar in spirit to `Replace` $(X, S, \hat{Y})$, as the fairness outcome is explicitly linked to sensitive attributes. However, a key distinction is that methods like ROC Mitigation or OT-L typically yield reproducible, model-dependent outcomes, while `Replace` $(X, S, \hat{Y})$, as explained in Section 3.3, cannot be applied in audits where auditors have direct access to the model and its decision thresholds.

Nevertheless, from the perspective of a supervisory authority, all these methods share a fundamental limitation. Since supervisory audits in our framework employ statistical tests based on the Kullback-Leibler (KL) divergence, these methods are easily detectable. Specifically, because they generate *new* synthetic individuals, the KL divergence between the original and manipulated distributions satisfies $\mathrm{KL}(Q_n, Q_t) = +\infty$.

One might ask whether the Wasserstein distance provides a more suitable metric. However, since these methods do not modify the covariates X (Gouic et al. (2020) projects individuals towards their Wasserstein barycenter only to change the network's output), the global Wasserstein distance remains unchanged, i.e., $W(Q_{n_X}, Q_{t_X}) = 0$. Regarding the Wasserstein distance between $(S, \hat{Y})$, $W(Q_{n_{(S,\hat{Y})}}, Q_{t_{(S,\hat{Y})}})$, the choice of method has minimal impact, since the distance is computed within the finite set of categorical bins $S, \hat{Y} \in \{0, 1\}^2$.

B.2 COMPARISON WITH FAIRNESS MANIPULATION VIA THE STEALTHILY BIASED SAMPLING (`SBS`) METHOD

Explaination of the `SBS` method The method designed by Fukuchi et al. (2020) minimizes the distribution shift measured by $W(X)$ under a fairness constraint on the Disparate Parity (DP):

$$\mathrm{DP} = |\mathbb{P}(\hat{Y} = 1|S = 1) - \mathbb{P}(\hat{Y} = 1|S = 0)|$$

Notably, the method does not allow specifying a target threshold t for the fairness criterion. Instead, the authors designed their sampling procedure to produce a perfectly fair dataset, such that $\mathrm{DP} = 0$ in expectation. The only tunable hyperparameter is the common acceptance rate for positive outcomes, denoted by $\alpha := \mathbb{P}(\hat{Y} = 1|S = 1) = \mathbb{P}(\hat{Y} = 1|S = 0)$.

This lack of flexibility in selecting a targeted DP complicates direct comparisons. As demonstrated in our paper, achieving fairness solely to pass compliance checks (i.e., fair-washing) often remains detectable by our statistical tests. To evaluate robustness, we progressively relax the fairness constraint until samples evade detection. Such adaptive calibration is not feasible with their approach.

One practical advantage of their method is that it outputs individual sampling probabilities rather than a fixed dataset, similar to our `Entropic` approach. This allows us to resample and generate distributions with varying degrees of fairness.

Their reported results were obtained via grid search over α values, as illustrated in Fig. 9. Consequently, to benchmark their method, we either had to identify the optimal α minimizing distribution shift or evaluate performance across all tested α values.

This method's high computational cost, already acknowledged by the authors in a subsequent paper Yamamoto & Hara (2025), is a notable limitation. Due to these computational constraints, we applied

their method exclusively on the Adult dataset, as experiments on larger datasets failed to complete within reasonable time frames.

Technical insecurities When following the installation instructions from their GitHub page, we encountered a problem. Indeed, the CMake version the authors used was 2.8 which is no longer supported with CMake (oldest version supported is 3.5) ; when changing in the CMake file the minimum version to 3.5, we had encountered another error with their (CMake_policy(set CMP0048 OLD)) which is no longer supported as well, we change it to the new version and ended up with a warning but could continue from there.

When we use the command make, we had the warning that *"ISO C++17 does not allow 'register' storage class specifier"*, and other warnings with "this statement may fall through' associated with if statement or case statement.

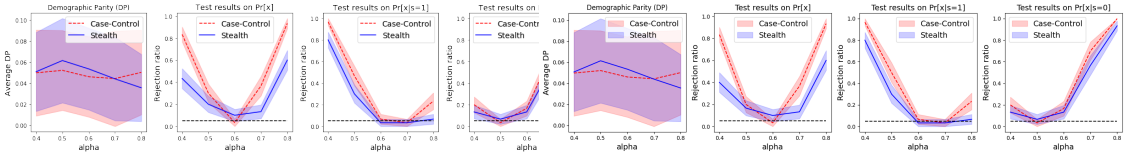
However, when we used 'make' for each of the *stealth-sampling* and *wasserstein* files, it went without any issues or warning, hence, we attempted to replicate their experiments; the provided random seed should have ensured identical results.

The authors did not use a requirement file, or specify which version of libraries to use. Hence, some errors within their code appear, mainly discrepancies between the former and new behavior of numpy. We modified the code for it to work with the latest version of the different libraries, while keeping the exact intended functioning.

	Version	Accuracy	DP	WD on Pr[x]	WD on Pr[x—s=1]	WD on Pr[x—s=0]
Baseline	Old	0.851	0.1824	22.1638	25.6454	35.0421
	New	0.85115	–	–	–	–
Case-control	Old	NaN	0.0250	23.9060	22.5855	37.9543
	New	NaN	0.0243	23.2002	23.3179	37.8548
Stealth	Old	NaN	0.0712	23.6396	24.2404	36.1657
	New	NaN	0.0708	24.1415	25.1640	35.5028

Table 4: Old and new results of the SBS method on the Adult dataset, '–' means that the new version has **exactly** the same result as the old one.

As you can see in Table 4, we observed a slight mismatch between the value they obtained and the value we obtained running exactly their code. This might indicate that because of the warnings we mentioned above, the performance was affected; or it could simply also be a different behavior from the newer version of libraries, and our result might actually be more representative. As you can see on Fig. 6, the aggregated results are alike. The authors would have obtained the same results and thus produced the same paper, we choose to use this implementation instead of "their" Sliced Wasserstein Distance method Yamamoto & Hara (2025), which compromises slightly the results for a significant boost in computational time.



(a) Original version

(b) Latest version

Figure 6: Original (Old) and Latest (New) results for their synthetic datasets, the experiments were done with 30 runs for several different α.

Result of the SBS method in our audit In this section, we evaluate on the Adult dataset: (1) the distribution shift incurred when creating a fair distribution ($\text{DI}(Q_t) > 0.8$) and (2) the maximum achievable DI without detection by our statistical tests. The first comparison (1) is well-aligned with the purpose of the method, which targets compliance. However, the second (2) inherently disadvantages their approach, as it was not designed to trade off fairness against detectability.

Dataset	Unbiasing Methods								
	SBS	Grad_p	Grad_b	Grad_p_1D	Grad_b_1D	Rep $(S, \hat{Y})$	$M_{W(X,S,\hat{Y})}$	Entr_b	Entr_p
$W(X, S, \hat{Y})$	0.91	0.10	0.08	0.13	0.09	**0.05**	**0.06**	0.28	0.35
$W(S, \hat{Y})$	**0.00**	0.09	0.08	0.09	0.08	0.05	0.05	0.08	0.09
KL$(X, S, \hat{Y})$	0.73	∞	∞	∞	∞	∞	0.03	**0.02**	0.03
KL$(S, \hat{Y})$	0.73	0.02	0.02	0.02	0.02	0.03	0.03	0.02	0.03

Table 5: Metric result of the fair-washing method (DI$(Q_t) \geqslant 0.8$), cost calculated on the projected dataset (or projected distribution for Entropic and SBS) on the Adult dataset.

Regarding $d(Q_n, Q_t)$, their method, like our Entropic approaches, produces sampling probabilities rather than a direct sample. This yielded strong performance on $W(S, \hat{Y})$, but it underperformed on KL$(S, \hat{Y})$ and did not stand out on $W(X, S, \hat{Y})$. For KL$(X, S, \hat{Y})$, it was less competitive, though it notably avoided divergence to infinity, making it one of the more globally competitive Wasserstein-based methods.

Dataset	S size (%)	SBS	Grad_p	Grad_b	Grad_p 1D	Grad_b 1D	Rep $(S, \hat{Y})$	Entr_b	Entr_p	$M_{W(X,S,\hat{Y})}$
ADULT	10	0.47	0.47	0.43	0.49	0.44	0.50	**0.54**	0.52	**0.54**
	20	–	0.39	0.40	0.38	0.39	0.41	**0.42**	0.41	**0.42**

Table 6: Highest undetected achievable Disparate Impact for the Adult dataset, for each sample size (S Size) and fair-washing method. The symbol – indicates that some methods couldn't reach a DI improvement. To emphasize the best method to use in order to deceive the auditor, we put the DI achieved in bold when one or two over-performed the others. We remind that the original DI of our Adult dataset is 0.30.

Conversely, as shown in Table 6, the method is not suitable for maximizing fairness without detection. To assess this, we computed 500 samples from each tuple sample size, α and observed whether the samples passed our statistical tests. Across $500 * 10 * 2$ samples (10 α and 2 sample size), 5 samples passed the 7 statistical tests, they were all for $\alpha = 0.25$ and sample size of 10% (instead of 20%).

C EXTENSION TO OTHER DATA TYPE

The method we develop is originally meant to handle tabular data. However we propose some natural direction to extend this work to text or images. The distances used to evaluate Wasserstein distance or the Maximum Mean Discrepancy (MMD) relies on the inherent informative information between individual within the input space, in another word they rely on the fact that the distance between individual is proportional to their semantic similarity. This hypothesis is always verified on tabular data (with the L_2 distance, for instance), but it might not be on images or token distributions. We will first evaluate our method based on $W(X)$ or MMD(X) to detect fraud attempt and expect the method to achieve a lower efficiency because $d(Q_t, Q_n)$ is hardly related to the semantic meaning of the images. Thus a fair-washing manipulation might not change this distance distribution.

Hence we embed the images in another space, where the regular distances have semantic meanings. The construction of such a space has already seen numerous works, including Principal Component Analysis, using the latent space of Auto-encoder or Variational Auto-Encoder, or using the latent space of Convolutional Neural Network classifiers. Using such space, which we call descriptor D, have become common practice after the introduction of the Fréchet Inception Distance (FID)Heusel et al. (2018). We define the function E such as

$$\begin{aligned} E\colon\ & \mathbb{R}^N \to \mathbb{R}^m \qquad\qquad N, m \in \mathbb{N}, N \gg m \\ & X \mapsto E(X) = D \end{aligned}$$

and set $E(Q) := \{E(X) | X \in Q\}$ if Q is a distribution. We choose in the following latent features given by the CNN classifier.

C.1 EXPERIMENTAL SETTINGS

We audited the CelebA dataset Liu et al. (2015): predicting the attractiveness, with the sensitive attribute being having heavy makeup. We note here that we choose this sensitive attribute instead of others for mainly two reasons:

1. Low DI : 0.4 on the whole dataset
2. Representativeness : Similarly to what we saw for the BAF dataset having a low probability of $\mathbb{P}(Y = 1) = 0.01$, If the sensitive variable was too rare, then detecting a modification on X would be impossible (for tabular or non-tabular data)

Note also that the variable *young* would have been another viable candidate.

The fair-washing method used in those experiments is the Wasserstein-based matching method $M_{W(X,S,Y)}$. We fine-tune 3 CNN models: an InceptionV3 Szegedy et al. (2015), a ResNet18 and a ResNet101 He et al. (2015) on CelebA and select part of the test set to audit, on this subset we observe respectively a DI of 0.34, 0.35 and 0.35. The malicious auditee, aware that the statistical tests on the covariates X might not be on the pixels of the images, but on the descriptor of the images, could minimize $d(E(Q_n), E(Q_t))$ instead of $d(Q_n, Q_t)$. Therefore, we consider 6 different fair-washing scenarios given (1) the network choice amongst ResNet18, InceptionV3 and ResNet101 which implies different descriptors' space and (2) if the auditee optimized the Wasserstein-based matching method on the pixel's space or on the latent space of those models. We first investigate the use of statistical tests directly of the pixels' space. Secondly, for each of the above scenario, we use statistical tests based of the latent space of the CNN. We remind here that (1) in term of complexity, the CCN are ranked as follow: ResNet18 (11 million parameters) $<$ InceptionV3 (27 million parameters) $<$ ResNet101 (44 million parameters), (2) the latent space of the CNN is the space at the hidden layer before the last linear layer, for the three models above's latent space share the same dimension size of 1000.

C.2 TOPICS OF INTEREST AND ANSWERS

In this subsection, we present key questions of interest and provide direct answers. These answers are supported by selected (cherry-picked) results for illustrative purposes, we refer to Table 7 for complete results on the descriptors analysis results. We address the following questions:

1. Are the statistical tests presented in our paper for tabular data relevant for non-tabular data?
   - Yes. We consider several data-agnostic statistical tests, such as $W(S, Y)$, MMD(S, Y), KL divergence, and the KS test. These tests remain applicable regardless of data type.
   - What about tests based on $W(X, S, Y)$ or MMD(X, S, Y), which rely on $W(X)$ or MMD(X) where X is high-dimensional? Tests conducted directly in pixel space can be useful. For example, fair-washing attempts with a sampling size of 20% were detected using tests applied directly to raw pixel values. In contrast, with a sampling size of 10%, and fair-washing to a Disparate Impact (DI) of 0.95, the null hypothesis $\mathcal{H}_0$ was not rejected, this indicates that these manipulations may go undetected at smaller sampling sizes.
2. Are statistical tests based on learned descriptors more effective?
   - Yes. Descriptor-based tests can detect fair-washing even with smaller samples. For instance, fraud attempts were identified using only 10% of the CelebA dataset (noting that detection becomes harder with smaller samples).
   - However, the auditee could potentially optimize their manipulation based on the descriptor used by the auditor, rendering these tests ineffective again at the 10% sample size.
3. Is the choice of the descriptor impactful ?
   - Yes, the results do depends on both the auditee choice of descriptor as well as the descriptor used in the statistical tests.
   - Yes for the manipulation: it was harder to detect manipulation based on the ResNet18 descriptors, and the easiest was the manipulation based on the ResNet101 descriptors. For instance, for the sampling size of 20%, even while testing using the ResNet101 descriptors, for a fair-washing at $DI = 0.55$, it was undetected when optimizing using ResNet18 descriptors but detected when optimizing using ResNet101's one (by "detected", we mean that across multiple samples, 50 in this case, none accepted the null hypothesis $\mathcal{H}_0$, i.e., the hypothesis that the sample and original distributions are the same.).
   - Yes for the fraud detection: Statistical tests based on the ResNet18 was more easily fooled by manipulation. To support this claim, we refer to for example to the last three columns of the Table 7 where with optimization on the image pixels, for a 10% sampling size, no fair-washing method was detected even with $DI = 0.95$.
4. Is there a difference using statistical tests based on the latent space the auditee's fair-washing method optimized on?
   - No, our results are not conclusive enough to answer this question positively. For InceptionV3 and ResNet101, we did not observe a significant difference.
   - That being said, in our experiments, with a 20% sampling size and fair-washing to $DI = 0.60$, only the auditee optimizing on the same descriptors (ResNet18) was able to generate an undetected sample when the test was based on those same descriptors.
   - Importantly, in practice, it is unlikely that the supervisory authority would use the same descriptors as the auditee. Even if the authority had full access to the auditee's network (which is rare, since this would go beyond API access), they may deliberately avoid using the same descriptors to prevent optimization-based circumvention.

In conclusion, on non-tabular modalities, running statistical tests directly on raw signals (in our cases pixels) is not useless, but tests in a learned descriptor space are markedly more sensitive. The choice of descriptor is critical: tests based on higher-capacity, semantically rich encoders (e.g., ResNet101) are substantially more robust to manipulations. We therefore recommend that supervisory authorities apply statistical tests both on the raw data and in a high-quality descriptor space. For text datasets, though not evaluated here, a natural first descriptor we would recommend is the CLS embedding from a BERT-style model Devlin et al. (2019), we leave this for a further work.

Descriptors	Size (%)	Fair-washing minimization objective					
		18	101	v3	18 pixels	101 pixels	v3 pixels
ResNet18	10	$\geqslant 0.95$	$\geqslant 0.95$	$\geqslant 0.95$	$\geqslant 0.95$	$\geqslant 0.95$	$\geqslant 0.95$
	20	$[0.6 - 0.7[$	$[0.5 - 0.55[$	$[0.5 - 0.55[$	$[0.4 - 0.5[$	$[0.4 - 0.5[$	$[0.4 - 0.5[$
Inceptionv3	10	$\geqslant 0.95$	$\geqslant 0.95$	$[0.8 - 0.95[$	$\geqslant 0.95$	$[0.8 - 0.95[$	$[0.8 - 0.95[$
	20	$[0.55 - 0.6[$	$[0.5 - 0.55[$	$[0.5 - 0.55[$	$[0.4 - 0.5[$	$[0.4 - 0.5[$	$[0.4 - 0.5[$
ResNet101	10	$\geqslant 0.95$	$\geqslant 0.95$	$\geqslant 0.95$	$\geqslant 0.95$	$\geqslant 0.95$	$[0.7 - 0.8[$
	20	$[0.55 - 0.6[$	$[0.5 - 0.55[$	$[0.55 - 0.6[$	$[0.4 - 0.5[$	$[0.4 - 0.5[$	$[0.4 - 0.5[$

Table 7: Highest DI without being detected for the CelebA Dataset using the matching fair-washing method based on different minimization objective testing on the descriptors which are the latent space of the different models, for sample size of 10% and 20%. The different scenarios are the following: 18, 101 and v3 are respectively a ResNet18, a ResNet101 and an Inceptionv3 optimized on their latent space ; the 18 pixels, 101 pixels and v3 pixels are the methods optimized on the pixel space (even if they have the same objective, they are different because the prediction of each network might be different).

D AUXILIARY RESULTS

Proposition D.1. *Consider the following minimization problem*

$$\min W_2^2(P, Q_n) \text{ such that } \int_E \Phi(x) dP(x) \geqslant t. \tag{7}$$

Then Q_t is optimal for equation 7 if, and only if, it is defined as the push-forward

$$Q_t = T_{\lambda^\star \#} Q_n$$

where $T_\lambda(y) \in \arg\min_x \left\{ \|x - y\|^2 - \lambda^T \Phi(x) \right\}$ and and then $\lambda^\star \in \mathbb{R}^k_{\geqslant 0}$ solves

- $\int_E \Phi(T_{\lambda^\star}(x)) dQ(x) \geqslant t$,
- *and* $\langle \lambda^\star, t - t_{\lambda^\star} \rangle = 0$

E PROOFS

E.1 PROOF OF PROPOSITION 3.2

Proof. Theorem 3.1 implies the existence of a distribution Q_t such that

$$DI(f, Q_t) = \frac{\lambda_0 + \delta_0}{\lambda_1 - \delta_1} \frac{n_1}{n_0} = t_1.$$

We have

$$\Delta_{DI} = \frac{n_1}{n_0} \left(\frac{\lambda_1 \delta_0 + \lambda_0 \delta_1)}{(\lambda_1 - \delta_1)\lambda_1} \right)$$

Among all possible solutions, we privilege the two solutions described in the Proposition. Knowing the new DI desired, we can obtain a set of solution for δ_0 and δ_1. □

E.2 Proof of Theorem 3.3

Proof. First, notice that the definition of T_λ implies

$$\begin{aligned}
W_2^2(Q_n, T_{\lambda\#}Q_n) &\leqslant \int_E \|T_\lambda(y) - y\|^2 dQ_n(y) \\
&= \int_E \|T_\lambda(y) - y\|^2 dQ_n(y) + \frac{1}{n}\left(\sum_{i=1}^n \lambda^\top \Phi(T_\lambda(Z_i)) - \sum_{i=1}^n \lambda^\top \Phi(T_\lambda(Z_i))\right) \\
&= \int_E \|T_\lambda(y) - y\|^2 - \lambda^\top \Phi(T_\lambda(y)) dQ_n(y) + \int_E \lambda^\top \Phi(y) dT_{\lambda\#}Q_n(y) \\
&= \int_E \inf_x \left\{\|x - y\|^2 - \lambda^\top \Phi(x)\right\} dQ_n(y) + \int_E \lambda^\top \Phi(y) dT_{\lambda\#}Q_n(y) \\
&= \int_E (\lambda^\top \Phi)^c(y) dQ_n(y) + \int_E \lambda^\top \Phi(y) dT_{\lambda\#}Q_n(y).
\end{aligned}$$

Strong duality of the Kantorovich problem, see Santambrogio (2015), guarantees that this inequality is indeed an equality. Since our equality constraint is linear, a necessary and sufficient condition for P^* to be a minimizer, see Peypouquet (2015), is finding Lagrange multipliers $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ such that

$$\begin{aligned}
&\sum_{i=1}^k \lambda_i \nabla g_i(P^*) \in \partial f(P^*) \quad \text{(extremality condition)} \\
&g(P^*) = 0 \quad \text{(feasibility)}
\end{aligned}$$

where $g(P) = \int_E \Phi(x) dP(x) - t$ and $f(P) = W_2^2(P, Q_n)$. The subgradient of f is given, see Proposition 7.17 in Santambrogio (2015), by the set of Kantorovich potentials between P^* and Q:

$$\partial f(P^*) = \left\{\phi \in C(E) \mid \int \phi dP^* + \int \phi^c dQ = W_2^2(P^*, Q)\right\}. \tag{8}$$

Our computations above prove the extremality condition for $P^* = T_{\lambda^\star\#}Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{T_{\lambda^\star}(Z_i)}$ since $\nabla g_i(P) = \int \Phi dP$. The feasibility condition for the empirical measure Q_n is to find $\lambda^\star$ such that

$$t = \int_E \Phi(y) dT_{\lambda^\star\#}Q_n(y) = \frac{1}{n}\sum_{i=1}^n \Phi(T_{\lambda^\star}(Z_i)). \tag{9}$$

□

E.3 Proof of Proposition D.1

Proof. Let g be the continuous function $g(P) = t - \int_E \Phi(x) dP(x)$ and $f(P) = W_2^2(P, Q)$. The set $\{P \in \mathcal{M}(E) \mid \int_E \Phi(x) dP(x) \geqslant t\} = g^{-1}([0, \infty))$ is closed for the weak convergence as $[0, \infty)$ is closed. Then the projection problem is well-definedd. Before applying the Lagrange multiplier theorem, we must verify Slater's condition. By continuity of Φ_i and compacity of E we can consider, for $i = 1, \dots, k$, $x_0^i \in E$ such that $\Phi_i(x_0^i) = \min_{x \in E} \Phi_i(x)$. Take $\alpha \in \mathbb{R}$ such that $\max_{1 \leqslant i \leqslant k} t_i / \Phi_i(x_0^i) < \alpha$. Then $\bar{P} = \alpha \delta_{x_0^i}$ satisfies $g_i(\bar{P}) < 0$ for $i = 1, \dots, k$. The Lagrange multipliers theorem guarantees that P^* is optimal for 7 if, and only if, there exists $\lambda_1, \dots, \lambda_k \geqslant 0$ such that

$$\begin{aligned}
&\sum_{i=1}^k \lambda_i \nabla g_i(P^*) \in \partial f(P^*) \quad \text{(extremality condition)} \\
&g(P^*) \leqslant 0 \quad \text{and } \lambda_i g_i(P^*) = 0 \text{ for all } i = 1, 2, \dots, k \text{ (feasibility)}
\end{aligned}$$

The proof of the extremality condition is completely analogous to the proof of Theorem 3.3, replacing Q_n by Q. To conclude, we need to find $\lambda^\star \in \mathbb{R}^k_{\geqslant 0}$ such that the feasibility condition is satisfied:

$$t \leqslant \int_E \Phi(T_{\lambda^\star}(x)) dQ(x) \text{ and } \lambda^{\star\top}\left(t - \int_E \Phi(T_{\lambda^\star}(x)) dQ(x)\right) = 0. \tag{10}$$

□

E.4 JOINT CONVEXITY OF THE WASSERSTEIN DISTANCE UNDER MIXTURE-PRESERVING COUPLING

Let Q_n and Q_t be probability distributions over $\mathcal{X} \times \{0, 1\}$, where $X \in \mathcal{X}$ denotes the data and $S \in \{0, 1\}$ is a binary group attribute.

For each $s \in \{0, 1\}$, define the conditional distributions:

$$Q_{n,s} := Q_n(\cdot \mid S = s), \quad Q_{t,s} := Q_t(\cdot \mid S = s),$$

and let $\pi := Q_n(S = 1) \in [0, 1]$. Then, define the marginal (mixture) distributions over $\mathcal{X}$ as:

$$\mu := \pi Q_{n,1} + (1 - \pi)Q_{n,0}, \quad \nu := \pi Q_{t,1} + (1 - \pi)Q_{t,0}.$$

We prove the inequality:

$$W_2^2(\mu, \nu) \leqslant \pi W_2^2(Q_{n,1}, Q_{t,1}) + (1 - \pi)W_2^2(Q_{n,0}, Q_{t,0}).$$

Proof. Let $\gamma_1 \in \Pi(Q_{n,1}, Q_{t,1})$ and $\gamma_0 \in \Pi(Q_{n,0}, Q_{t,0})$ be couplings between the corresponding conditionals. Define the coupling:

$$\gamma := \pi\gamma_1 + (1 - \pi)\gamma_0.$$

Then $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, and its marginals are:

$$\gamma^X = \pi Q_{n,1} + (1 - \pi)Q_{n,0} = \mu, \quad \gamma^Y = \pi Q_{t,1} + (1 - \pi)Q_{t,0} = \nu.$$

Thus, $\gamma \in \Pi(\mu, \nu)$ is a valid coupling between μ and ν.

Now compute the transport cost under γ:

$$\int_{\mathcal{X}\times\mathcal{X}} d(x, y)^2 \, d\gamma(x, y) = \pi \int d(x, y)^2 \, d\gamma_1(x, y) + (1 - \pi) \int d(x, y)^2 \, d\gamma_0(x, y),$$

(Because the distance is an integrable function, we can use the linearity of the Lebesgue integral with respect to measures)

$$= \pi W_2^2(Q_{n,1}, Q_{t,1}) + (1 - \pi)W_2^2(Q_{n,0}, Q_{t,0}).$$

Since $W_2^2(\mu, \nu)$ is the infimum of such costs over all couplings in $\Pi(\mu, \nu)$, we obtain:

$$W_2^2(\mu, \nu) \leqslant \pi W_2^2(Q_{n,1}, Q_{t,1}) + (1 - \pi)W_2^2(Q_{n,0}, Q_{t,0}).$$

□

F RESULTS WITH SIMULATED DATASET

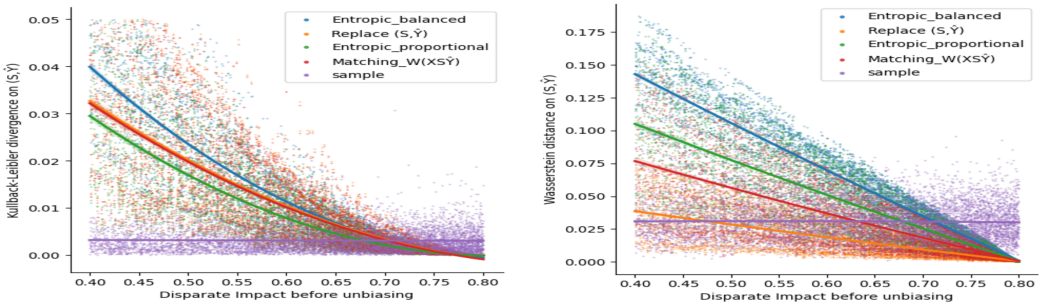


Figure 7: Logistic regression plots showing how the distance (left : $KL_{(S,\hat{Y})}$ and right: $W_{(S,\hat{Y})}$) between the original and 20% of manipulated datasets varies with the initial Disparate Impact for each unbiasing method, with the manipulated dataset having DI = 0.8. The sample's results in the legend represent values from random samples from the original distribution Q_n.

We create a simulated dataset to cover all possible cases where $S \in \{0,1\}$ and $\hat{Y} \in \{0,1\}$. The simulation parameters control $\mathbb{E}(S)$, $\mathbb{E}(\hat{Y} \mid S=0)$, and $\mathbb{E}(\hat{Y} \mid S=1)$, allowing us to represent a wide range of scenarios. Fig 7 presents two logistic regression graphs illustrating how the distance between the complete original and 20% of manipulated data evolves from the initial Disparate Impact (before debiasing), to reach a DI=0.8, for our different correction methods. Methods with the highest KL and Wasserstein distance implies a high risk of being detected by a statistical test on the distribution. The lower the initial DI, the greater the change required to reach an acceptable DI (making fraud detection more likely). When the original DI is $\geqslant 0.55$, the methods `Entropic`, `Replace` and `Matching` are equivalent in terms of KL divergence. Regarding the Wasserstein distance, they become equivalent for original DI values $\geqslant 0.65$. Since the `Sample` method does not modify the original data, it preserves the distributional distances (KL and Wasserstein), and can be used as a reference: when the logistic regression score of a method is lower than that of `Sample`, we can infer that the modified dataset would not be detected as significantly different from the original according to these criteria. Among all methods, $M_{W(X,S,\hat{Y})}$ with an original DI$\in [0.45, 0.70]$ achieves the best trade-off between KL divergence and Wasserstein distance, reaching the required DI while keeping the modified distribution close to the original.

G Further studies on the impact of the sample size

In our conclusion, we recommended strongly to the referring authorities, that in order to prevent undetectable fraud, with appropriate statistical tests, requiring a bigger sample size is one of the single most important point. To further support this claim, we provide in this section a study on the sample size impact on the Adult dataset.

Using the best performing fair-washing method ($M_{W(X,S,Y)}$, `Entropic_balanced` and `Entropic_proportional`), we observe on Fig. 8 the highest DI achievable without being detected by our 7 statistical tests depending on the sample size required.

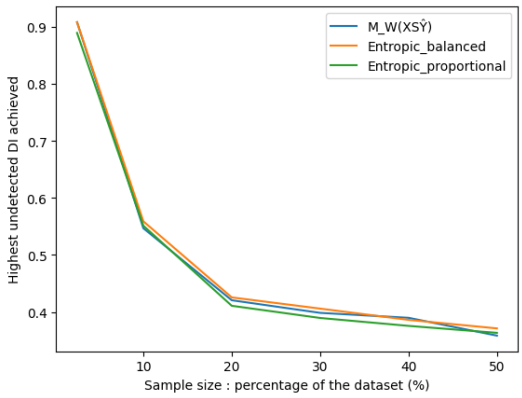


Figure 8: Highest undetected DI achieved without being detected in the Adult dataset by different fair-washing methods depending on the sample size.

H More information on the methods

H.1 Replacing key attributes and Wasserstein-minimizing sampling

In this section, we precise how Disparate Impact (DI) can be increased using methods based on optimal transport. We can exchange between the 4 bins of points :$(Y=0,S=0)$, $(Y=0,S=1)$, $(Y=1,S=0)$ and $(Y=1,S=1)$, thus $4(4-1)=12$ possible alterations. Due to the definition of DI, we can exclude the path from $(Y=1,S=0)$ to $(Y=0,S=0)$ and the path from $(Y=1,S=1)$ to $(Y=0,S=1)$ as it would decrease the DI, bringing the total to 10 possible transports.

Moreover, If we consider the Wasserstein cost only on $(S, \hat{Y})$), once again based on its definition, because it is more advantageous to have (Y=1,S=0) points instead of (Y=0, S=0), similarly for (Y=0, S=1) points instead of (Y=1, S=1), we only have 6 worthy alterations to consider instead of 10. Indeed, for instance, it would be suboptimal to transport a (Y=1, S=1) point to (Y=0, S=0) as moving it to (Y=1, S=0) would result in a higher DI with a lesser effort (with the cost is calculated only on $(S, \hat{Y})$)).

Furthermore, transport between the bins $(Y = 1, S = 0)$ and $(Y = 0, S = 1)$ (i.e., between the two back points in Fig. 1) can also be excluded from the optimal solution. Indeed, transport from $(Y = 1, S = 0)$ to $(Y = 0, S = 1)$ would both reduce the number of favorable outcomes (decreasing the numerator of the DI) and increase the number of unfavorable outcomes for the protected group (increasing the denominator), thus leading to a lower DI. In contrast, if the bin $(Y = 0, S = 1)$ move from $(Y = 0, S = 0)$, the DI is improved, as only the denominator increases while the numerator remains unchanged

To summarize, if the cost is calculated on $(S, \hat{Y})$), then we theoretically only have 4 moves to consider, the arrows in Fig. 1. The arrow from (Y=0, S=0) toward (Y=0, S=1) is doted for the following reason: in practice, for the simulated dataset presented in Section F, this transport was never optimal meaning not the one which increases the DI the most compared to other transports. The most rewarding one was usually from the transport from (Y=0,S=0) to (Y=1,S=0). This leads us to write the Alg. 2 which is the less concise version of Alg. 1. A notable difference between the two is that Alg. 2 has a `speed` parameter which express a trade-off performance rapidity as explained in Section H.3.

Termination analysis At every iteration of the while loop, the DI is strictly increasing, moreover, the number of iterations is limited by the number of points $|\{X|Y = 1, S = 1\}|$ and $|\{X|Y = 0, S = 0\}|$. In the extreme case where no transport would be possible (either of these sets is empty if $|\{X|Y = 1, S = 1\}| = 0$ or $|\{X|Y = 0, S = 0\}| = 0$) the algorithm could attempt to increase DI indefinitely (towards $+\inf$). This ensures that the algorithm necessarily terminates.

Objective analysis The condition of the while loop is precisely aligned with the objective of our problem. Consequently, exiting the loop implies that a solution has been found. Finally, a more challenging question concerns the optimality of the solution returned by the algorithm. We leave this question open and do not provide a formal guarantee of optimality.

Algorithm 2 `Replace(`S`,`$\hat{Y}$`)` non simplified algorithm

speed $\in \mathbb{N}^*; 0 <$ threshold < 1
b $= [|\{X|Y = 1, S = 1\}|, |\{X|Y = 0, S = 1\}|, |\{X|Y = 1, S = 0\}|, |\{X|Y = 0, S = 0\}|]$
DI $=$ DI_fct(b)
swap_possible = $\{Y_0S_0, Y_1S_1\} \to Y_0S_1, \{Y_0S_0\} \to Y_1S_1$
dic_swap_translation $= \{Y_0S_0 \to Y_1S_0 : [0, 0, 1, -1], Y_0S_0 \to Y_0S_1 : [0, 1, 0, -1], Y_1S_1 \to Y_1S_0 : [-1, 0, 1, 0]\}$
dic_swap_number $= Y_0S_0 \to Y_1S_0 : 0, Y_0S_0 \to Y_0S_1 : 0, Y_1S_1 \to Y_1S_0 : 0$
DI$_n = [0, 0, 0]$; Matrix_$b = M_{(3,4)}(0)$
while DI ¡ threshold **do**
    $i = 0$
    for swap ∈ swap_possible **do**:
        b_n = b + dic_swap_translation[swap] ▷ $Y_0S_0 \to Y_1S_0$ translation
        Matrix_$b[i, :]$ = copy(b_n) ▷ We keep in memory the bins
        DI$_n[i]$ = DI_fct(b_n)
        $i = i + 1$
    end for
    $j =$ argmax(DI$_n$)
    dic_swap_done[swap_possible[j]] = dic_swap_done[swap_possible[j]] + speed ▷ More information on speed discussed in next subsection
    $b = b +$ speed $*$ (Matrix_$b[j, :] - b$)
    DI $=$ DI_fct(b) ▷ Equal to DI$_n[j]$ only if speed $= 1$
end while
return dic_swap_number

H.2 WASSERSTEIN GRADIENT GUIDED METHOD

Algorithm 3 Fair-washing using Monge Kantorovich constrained projection algorithm `Grad`

Require: Neural network f, data Z_0, sensitive attribute $S \in \{0,1\}^n$, prediction threshold τ, desired DI threshold t, learning rate η, constraint weight λ, delta type ($\in$ {`balanced`, `proportional`}

Ensure: Updated samples Z minimizing $\|Z - Z_0\|^2$ while satisfying $\mathrm{DI}(f(Z), S) \geqslant t$

Compute: $\hat{Y} \leftarrow \mathbb{I}[f(Z_0) > \tau]$ ▷ where $\mathbb{I}$ is the indicator function

Compute $P_0 = \mathbb{E}[\hat{Y} \mid S = 0]$, $P_1 = \mathbb{E}[\hat{Y} \mid S = 1]$, $n_1 = \mathbb{E}[S = 1]$, $n_0 = \mathbb{E}[S = 0]$

Compute δ_s according to delta type ▷ Done following Prop.3.2

Set new target rates:

$$\tilde{P}_1 = P_1 - \delta_1/n_1, \quad \tilde{P}_0 = P_0 + \delta_0/n_0$$

for $s \in \{0,1\}$ **do**

  Initialize $\lambda^{(s)} \leftarrow \lambda$

  while $(\mathbb{E}[\hat{Y}^{(s)}] < \tilde{P}_s) \vee (\mathbb{E}[\hat{Y}^{(s)}] > \tilde{P}_s \wedge s = 1)$ **do**

    Initialize $Z_i^{(s)} \leftarrow Z_0^{(s)}$, $\eta_i \leftarrow \eta$ ▷ where $Z_0^{(s)}$ is the subset of inputs with $S = s$

    for $i \in 1, \cdots, 10$ **do**

      Iteration of gradient step:

$$\nabla = 2(Z_i^{(s)} - Z_0^{(s)}) + \lambda^{(s)} \cdot \nabla_Z f(Z_i^{(s)}) \cdot d_s$$

where $d_s = \begin{cases} +1 & \text{if } s = 0 \\ -1 & \text{if } s = 1 \end{cases}$ ▷ Gradient choice following Thm. 3.3

$$Z_i^{(s)} \leftarrow Z_i^{(s)} - \eta_i \cdot \nabla$$

      Recompute predictions $\hat{Y}_i^{(s)} = \mathbb{I}[f(Z_i^{(s)}) > \tau]$

      $\eta_i \leftarrow \eta_i / 1.2$ ▷ Planning strategies could improve the performance, 1.2 was what we founded worked the best in practice (following the choice of the coefficient multiplying $\lambda^{(s)}$)

      if `1D-transport variant` **then**

        Project each feature of $Z_i^{(s)}$ to its closest achievable value

      end if

      if $\mathbb{E}[\hat{Y}_i^{(s)}] < \tilde{P}_s$ (or $> \tilde{P}_s$ for $s = 1$) **then**

        Break Exit for and while loop

      end if

    end for

    Update : $\lambda^{(s)} \leftarrow 1.2 \times \lambda^{(s)}$ ▷ The solution of the optimization problem with this λ is not within the constrained space (or we did not converge towards it fast enought at least); hence we increase the λ progressively. Note that the 1.2 was what we founded worked best in practice (trade-off between precision with lower value and fast computation), however further tuning would be relevant.

  end while

  Compute perturbation $T^{(s)} = Z^{(s)} - Z_s$

end for

Assemble final perturbation T such that:

$$T_i = \begin{cases} T_i^{(0)} & \text{if } S_i = 0 \text{ and } \hat{Y}_i = 0 \\ T_i^{(1)} & \text{if } S_i = 1 \text{ and } \hat{Y}_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

return $Z = Z_0 + T$

Alg. 3, which is a simplified version of the true algorithm (code available on our Github[1]), explains the main ideas being :

[1] https://anonymous.4open.science/r/Inspection-76D6/

1. We find the target probabilities for each subgroup of s
2. We treat each $Q_{n,s} := Q_n(\cdot \mid S = s)$ separately
3. The gradient steps stem from Theorem 3.3
4. We start with a small constraint weight and increase it progressively

The elements present in our code but which we did not include in Alg. 3 for visibility are mostly computational optimizations. For instance, we did not compute the gradient on neither the points whose network decision we would not modify Z_i (i.e., with if $S_i = 0$ and $\hat{Y}_i = 1$) nor on points $Z_i^{(s)}$ whose $\hat{Y}_i^{(s)}$ are already modified. We also kept streakily only the minimum number of modification necessary: some gradient step would change the network decision of multiple points at the same time and without this process our result would not be tight regarding $\tilde{P}_1, \tilde{P}_0$ (as we would have changed individuals' outcome more than necessary) and thus overachieving $\mathrm{DI}(f(Z), S) \geqslant t$ which is not beneficial in our use case where we highlighted the trade-off between fairness correction and distribution shift.

H.3 Costs of the methods, solutions and tests

Methods	Summary	Solution
$M_{W(X,S,\hat{Y})}$	3–10 minutes	Trade-off possible
`Replace(`S`,`$\hat{Y}$`)`	$\leqslant$ 2 minutes	Trade-off possible
`Entropic_b` / `Entropic_p`	$\leqslant$ 1 minutes	
`Grad_p` / `Grad_b`	3–15 minutes, depends on λ and NN architecture	Trade-off possible
`Grad_p(1D-t)` / `Grad_b(1D-t)`	3–20 minutes, depends on λ and NN architecture	Trade-off possible

Table 8: Time cost analysis of the methods, note that every estimation depends on the original dataset, its Disparate Impact and the DI constraint. Time estimation given for a dataset size of 20k individuals.

Sample size	Test performed	Average testing time (second)
500	$\mathrm{DI}(S) \geqslant \mathrm{DI}(Q_t)$	0.00
	$KL(S, \hat{Y})$	0.00
	$KL(X, S, \hat{Y})$	0.78
	$W(S, \hat{Y})$	0.29
	$W(X, S, \hat{Y})$	0.48
1000	$\mathrm{DI}(S) \geqslant \mathrm{DI}(Q_t)$	0.00
	$KL(S, \hat{Y})$	0.00
	$KL(X, S, \hat{Y})$	0.85
	$W(S, \hat{Y}) + W(X, S, \hat{Y})$	1.57
2000	$\mathrm{DI}(S) \geqslant \mathrm{DI}(Q_t)$	0.00
	$KL(S, \hat{Y})$	0.02
	$KL(X, S, \hat{Y})$	3.13
	$W(S, \hat{Y})$	4.93
	$W(X, S, \hat{Y})$	15.51
4000	$\mathrm{DI}(S) \geqslant \mathrm{DI}(Q_t)$	0.00
	$KL(S, \hat{Y})$	0.02
	$KL(X, S, \hat{Y})$	3.34
	$W(S, \hat{Y})$	9.58
	$W(X, S, \hat{Y})$	32.30

Table 9: Time analysis done during our Highest undetected achievable DI per datasets and methods .

Time In Table 8, we wrote Trade-off possible for the methods which might take a more than a day to run with millions of individuals. The methods $M_{W(X,S,\hat{Y})}$ and `Replace(`S`,`$\hat{Y}$`)` evaluate at each step amongst 3 or 4 possibilities which is the optimal to take, we can only evaluate once for more step at the same time for both methods, this becomes a trade-off between speed and precision, this is what we mean by trade-off possible for those methods. Moreover, we can also think about a trade-off about the number of transport mapping to consider, as explained in the Section. H.2.

For the `Grad` variant methods, we do not anticipate any changes to the model architecture. However, if inference from the neural network is computationally expensive, the overall cost of the method will also be high. Developing an efficient solution to this issue remains an open challenge. However, with tabular data model's number of parameters tends to be controllable, and thus in our experiments the reason of such a long time compute time (relative to the number of individual) was because we optimized for the λ parameter. We remind that to have the best results we start with a very small λ which we progressively increase ; we thus can simply initialize the algorithm with a bigger λ to save computing time, another speed precision trade-off.

The results in Table 9 were obtained through the following procedure. For each sample, we recorded: (1) the total execution time of the testing pipeline, and (2) the reason the pipeline stopped. Since each sample must pass all five tests, the pipeline halts as soon as one test is failed. Based on our prior expectations regarding the relative runtime of the tests. To isolate the runtime of each individual test, we subtracted the mean runtime of the preceding tests from the total time observed at the stopping point. The results show that while all tests are fast for small sample sizes (e.g., 500 samples), the tests based on Wasserstein distances (in particular $W(X, S, \hat{Y})$) are the most time-consuming.

Methods	Summary	Solution
$M_{W(X,S,\hat{Y})}$	$N \times N$ distance matrix	
`Replace(`S`,`$\hat{Y}$`)`	Negligible	Trade-off possible
`Entropic_b` / `Entropic_p`	Negligible	
`Grad(b/p)` / `(1D)`	NN gradient to compute on at worse on N ind	Batch approach

Table 10: Memory cost analysis of the methods for a $N \times J$ dataset.

Memory We consider only the `Grad` variant methods to potentially pose memory-related issues. Although it would be natural to adapt these methods to operate in a batch-wise manner, we did not implement such an approach in our current work.

I Optimization result values

I.1 Wasserstein distance

	Unbiasing Methods							
Dataset	Grad_p	Grad_b	Grad_p_1D	Grad_b_1D	Rep $(S, \hat{Y})$	$M_{W(X,S,\hat{Y})}$	Entr_b	Entr_p
ADULT	0.10	0.08	0.13	0.09	**0.05**	**0.06**	0.28	0.35
EMP	0.18	0.10	0.18	0.10	**0.06**	0.08	0.22	0.37
INC	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
MOB	0.21	0.06	0.23	0.08	**0.03**	0.05	0.18	0.64
PUC	0.21	**0.13**	0.22	0.14	**0.12**	0.16	0.33	0.48
TRA	0.02	0.02	0.02	0.02	**0.01**	**0.01**	0.03	0.03
BAF	0.01	**0.00**	0.02	0.01	**0.00**	0.01	0.02	0.05

Table 11: Wasserstein distance manipulation cost of the fair-washing methods (DI$(Q_t) \geqslant 0.8$), cost calculated on the projected dataset : $W(Q_n, Q_t)$ with the original dataset Q_n and $Q_t = f(Q_n)$ with f the fair-washing method

	Unbiasing Methods							
Dataset	Grad_p	Grad_b	Grad_p_1D	Grad_b_1D	Rep $(S, \hat{Y})$	$M_{W(X,S,\hat{Y})}$	Entr_b	Entr_p
ADULT	0.09	0.08	0.09	0.08	**0.05**	**0.05**	0.08	0.09
EMP	0.18	0.10	0.18	0.10	**0.06**	**0.06**	0.10	0.18
INC	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
MOB	0.18	0.06	0.18	0.06	**0.03**	**0.03**	0.06	0.18
PUC	0.21	0.13	0.21	0.13	0.12	**0.10**	0.13	0.21
TRA	0.02	0.02	0.02	0.02	**0.01**	**0.01**	0.02	0.02
BAF	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00

Table 12: Wasserstein distance manipulation cost of the fair-washing methods(DI$(Q_t) \geqslant 0.8$), cost calculated on the projected dataset : $W(Q_{n,(S,\hat{Y})}, Q_{t,(S,\hat{Y})})$ with the original dataset Q_n and $Q_t = f(Q_n)$ with f the fair-washing method

I.2 Kullback-Leibler divergence

	Unbiasing Methods							
Dataset	Grad_p	Grad_b	Grad_p_1D	Grad_b_1D	Rep $(S, \hat{Y})$	$M_{W(X,S,\hat{Y})}$	Entr_b	Entr_p
ADULT	∞	∞	∞	∞	∞	0.03	**0.02**	0.03
EMP	∞	∞	∞	∞	∞	**0.04**	**0.04**	0.07
INC	∞	∞	∞	∞	∞	0.00	0.00	0.00
MOB	∞	∞	∞	∞	∞	**0.02**	0.03	0.17
PUC	∞	∞	∞	∞	∞	0.06	0.07	0.10
TRA	∞	∞	∞	∞	∞	0.00	0.00	0.00
BAF	∞	∞	∞	∞	∞	0.00	0.00	0.00

Table 13: KL divergence manipulation cost of the fair-washing methods (DI$(Q_t) \geqslant 0.8$), cost calculated on the projected dataset : KL(Q_n, Q_t) with the original dataset Q_n and $Q_t = f(Q_n)$ with f the fair-washing method

	Unbiasing Methods							
Dataset	Grad_p	Grad_b	Grad_p_1D	Grad_b_1D	Rep $(S, \hat{Y})$	$M_{W(X,S,\hat{Y})}$	Entr_b	Entr_p
ADULT	0.02	0.02	0.02	0.02	0.03	0.03	0.02	0.03
EMP	0.06	**0.03**	0.06	**0.03**	0.04	0.04	0.04	0.07
INC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MOB	0.12	**0.03**	0.12	0.03	**0.02**	**0.02**	**0.03**	0.17
PUC	0.09	**0.06**	0.09	**0.06**	0.08	**0.07**	**0.07**	0.10
TRA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BAF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 14: KL divergence manipulation cost of the fair-washing methods (DI$(Q_t) \geqslant 0.8$), cost calculated on the projected dataset : KL$(Q_{n,(S,\hat{Y})}, Q_{t,(S,\hat{Y})})$ with the original dataset Q_n and $Q_t = f(Q_n)$ with f the fair-washing method

J Ablation study on the MMD test

We report the results in Table 15, which presents the highest Disparate Impact (DI) achieved by samples that remained undetected by the statistical tests based on KL divergence, Wasserstein distance, and the Kolmogorov–Smirnov (KS) test. This table corresponds to Table 3, but excludes the two tests based on the MMD distance.

From Table 15, we observe that excluding the MMD tests had negligible impact on detection outcomes. The only notable difference arises in the BAF dataset with a 20% sampling rate, where the achieved DI is slightly higher. We also point out that, due to the inherent randomness in sampling (100 random samples are drawn for each combination of dataset, fair-washing method, sample size, and target DI(Q_t)), we occasionally found samples that passed all seven tests and exhibited marginally higher DI than those evaluated with only five tests. These cases are indicated by a '+' symbol in parentheses in Table 15.

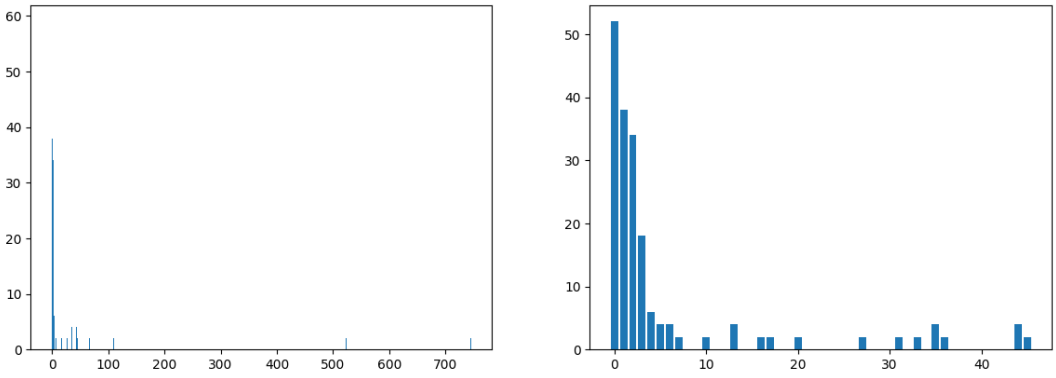
Dataset	Original	S size (%)	Rep $(S, \hat{Y})$	Entr_b	Entr_p	$M_{W(X,S,\hat{Y})}$
ADULT	0.30	10	0.45(-0.05)	0.53 (-0.01)	0.55 (+0.03)	**0.54** (+0.01)
		20	0.38(-0.03)	**0.43**(+0.01)	0.42(+0.01)	**0.43**(+0.01)
EMP	0.30	10	–	0.38(+0.03)	**0.39**(+0.03)	**0.39**(+0.02)
		20	–	**0.36**(+0.02)	0.35(-0.01)	**0.36**(+0.01)
INC	0.67	10	0.88	0.95(+0.01)	0.95(+0.01)	0.95(+0.02)
		20	0.83	0.84(+0.01)	0.84	0.84
MOB	0.45	10	0.54(+0.01)	0.53(+0.01)	0.51	**0.55**(+0.02)
		20	0.48(-0.01)	**0.50**	0.49	**0.50**
PUC	0.32	10	–	**0.36**(+0.03)	**0.36**(+0.01)	0.35
		20	–	–	–	–
TRA	0.69	10	0.76(-0.03)	0.84(+0.01)	0.84	0.84
		20	0.71(-0.06)	0.80(+0.01)	0.80(+0.01)	**0.81**
BAF	0.35	10	–	1	1	1
		20	–	0.83(+0.06)	0.84(+0.05)	**0.85**(+0.06)

Table 15: Highest undetected (without the MMD-based statistical tests) achievable Disparate Impact for each dataset, sample size (S Size) and fair-washing method. The symbol – indicates that some methods couldn't reach a DI improvement. To emphasize the best method to use in order to deceive the auditor, we put the DI achieved in bold when one or two over-performed the others. The number in parentheses are here to indicate the difference between those results and the results obtained with the MMD-based tests (Result without MMD - Result with).

K FRAUD DETECTION

K.1 DISTRIBUTION OF TRIES BEFORE ACCEPTANCE OF $\mathcal{H}_0$

As shown in Fig 9, taking only 30 or 50 samples instead of 1000 gives us respectively 73% or 78% accuracy for the tests. This is arguably not that high, however knowing that we would have needed the combination of 5 statistical tests to accept our sample in our use case, it still gives us a good approximation to whether the test have a chance to be accepted (as it is harder to be accepted by the combination of 5 tests than the individuals one).



(a) Complete distribution

(b) Zoom on under 50

Figure 9: Distribution of number of tries to find an accepted sample for $\mathcal{H}_0$ for the statistical test KS or KL$(S, \hat{Y})$ with a maximum of 1000 tries per configuration (method, dataset, test) for all datasets.

K.2 HIGHEST UNDETECTED ACHIEVABLE DISPARATE IMPACT PROBABILITIES AND ADDITIONAL GRAPH

We add details to Table 3 results, particularly its stability towards the number of sampling tries.

- We would have had 95%, 97% and 98% of similar results if we tried respectively 10, 20 and 30 samples compared to 100. (we had respectively 41, 24 and 19 scenarios which have us different results over the span of 896 combinations).
- In configurations where a fairer falsely compliant sample was found, it was generally around the 11^{th} sample, while the median was equal to 4. (See Fig. 11)

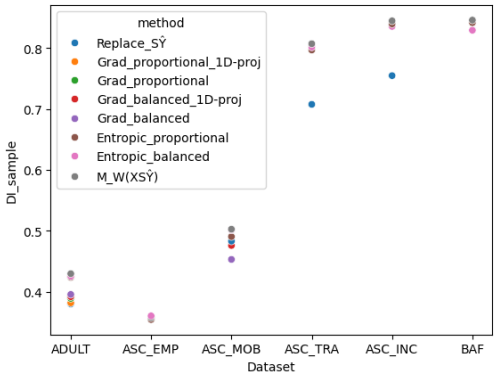


Figure 10: Highest achieved DI for all Datasets and methods (when they improve the original DI), with sample size of 20% of the dataset.

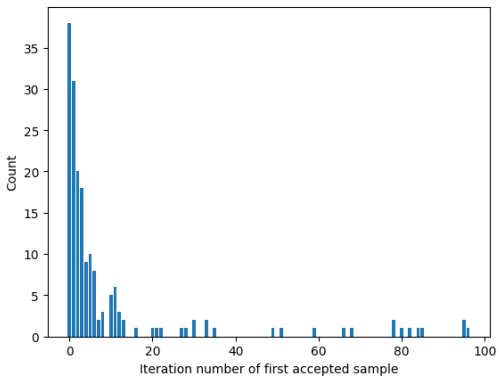


Figure 11: Distribution of the number of sample tried before first accepted one for all datasets.