Commun. Optim. Theory 2025 (2025) 48 https://doi.org/10.23952/cot.2025.48

MATHRES Communications in Optimization Theory
Available online at http://cot.mathres.org

OVERCOMING REPRESENTATION BIAS IN FAIRNESS-AWARE DATA REPAIR USING OPTIMAL TRANSPORT

ABIGAIL LANGBRIDGE[1], ANTHONY QUINN[1,2,*], ROBERT SHORTEN[1]

[1]Dyson School of Design Engineering, Imperial College London, UK
[2]Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

Abstract. Optimal transport (OT) has an important role in transforming data distributions in a manner which engenders fairness. Typically, the OT operators are learnt from the unfair attribute-labelled data, and then used for their repair. Two significant limitations of this approach are as follows: (i) the OT operators for underrepresented subgroups are poorly learnt (i.e. they are susceptible to representation bias); and (ii) these OT repairs cannot be effected on identically distributed but out-of-sample (i.e. archival) data. In this paper, we address both of these problems by adopting a Bayesian nonparametric stopping rule for learning each attribute-labelled component of the data distribution. The induced OT-optimal quantization operators can then be used to repair the archival data. We formulate a novel definition of the fair distributional target, along with quantifiers that allow us to trade fairness against damage in the transformed data. These are used to reveal excellent performance of our representation-bias-tolerant scheme in simulated and benchmark data sets.

Keywords. AI fairness; Bayesian nonparametrics; Data repair; Optimal transport; Representation bias; Stopping rule.

2020 MSC. 49Q22, 62G05, 62P25.

1. INTRODUCTION

Representation bias is a key issue in machine learning, with many classic datasets being biased towards majority groups such as Americans, white people, or men depending on the context [25, 35]. These biases – when left unaddressed – can limit trust in model outcomes, undermine efforts towards fairness correction, and exacerbate existing socioeconomic inequalities [33]. By designing a fair transformation which directly addresses the issue of representation bias, we can ensure fairness across all population subgroups.

The problem of generalisation is of critical importance in machine learning [26]. A model's ability to perform well on new, unseen data after being trained on a specific dataset is crucial to facilitate deployment. Over- and under-fitting are common, which lead to underlying causal

*Corresponding author.
E-mail address: a.quinn@imperial.ac.uk (A. Quinn).
Received August 8, 2024; Accepted October 26, 2024.

relationships in data not being captured. This can be as a result of poor modelling choices [26], or as a function of the size and composition of training data which can significantly effect the performance of machine learning models [12, 28, 34].

One area where good generalisation is essential is in AI Fairness (AIF). Poor generalization often leads to biased models that perform well on certain segments of the population (typically those overrepresented in the training data) but poorly on others [1, 25]. This can result in unfair treatment of underrepresented groups and exacerbate societal biases present in the underlying data.

There are a number of well-established methods for fairness correction which aim to remove or reduce the level of bias in models through changes to the model itself [5, 36] or the data [7, 10, 31]. A contemporary review of works in AIF is presented in [1]. While these data repair methods are popular since they don't restrict downstream model choice, they often require access to the entire (finite and static) dataset in order to conduct their repair - i.e. they are not designed to generalise. The previous works [3, 23] overcome this by learning a repair on a small training data set which can be applied to large archival or online data, analogous to generalisation in machine learning.

However, these repairs are highly sensitive to representation biases in the training data where the generalisation performance on under-represented classes is poorer due to incomplete learning of the distribution of the underlying data generating process. Representation bias is common in fairness settings, since disadvantaged classes of individuals have historically been denied access to opportunities. For example, in the Adult Income dataset [24], non-white people have on average a lower level of education than the white people in the dataset, which corresponds to lower predicted salaries. The presence of representation bias in data can inherently limit the fairness metrics which are attainable, even after repair [19]. Additionally, the presence of multiple behaviour modes of the data-generating process – due to intersectionality [13] – and the response of further segmenting training data (i.e. considering the outcomes of *non-white women* rather than just women in the Adult Income example above) exacerbates these issues with statistical power, repair validity and learning [16].

In this work, we propose a data-driven method for overcoming representation bias in fairness repair, extending the repair method initially proposed in [23] to improve robustness to unbalanced data and improve the generalisation of the method.

Our notational conventions are as follows: random variables (rvs) and their realizations, e.g. x, are not distinguished notationally; column vectors, e.g. $\mathbf{x}$, are bold-faced; sets are denoted via blackboard symbols, $\mathbb{R}, \mathbb{N}$, etc.; and functions via sans-serif letters, $\mathsf{T}(\cdot)$, $\mathsf{Q}(\cdot)$, etc. F is reserved for an unspecified (i.e. wildcard) probability measure with respect to a particular algebra, and may also be used to refer to its density, $\mathsf{F}(\cdot)$, or probability mass function, $\Pr[\cdot]$, with respect to the appropriate reference measure (Lebesgue or counting, respectively), as will be evident from the context. $\hat{\mathsf{F}}_n$ denotes the empirical estimate of F based on a random sample of size $n \geq 1$. Finally, the Dirichlet nonparametric process prior is denoted by $\mathscr{D}$.

The paper is laid out as follows: in Section 2, we focus on modelling and introduce our (un)fairness criterion. Sections 3 and 4 detail our proposal for a data-driven fairness correction scheme, with experimental evidence supporting our method in Section 5. Section 6 contains a discussion of these results, and we conclude in Section 7.

2. AI Unfairness and Representation Bias

Consider a sequential observational experiment, in which continuous-valued, d-dimensional data (feature vectors), $x_i \in \mathbb{R}^d$, $i \in \mathbb{N}$, are made available, each labelled with Bernoulli (i.e. binary, without loss of generality) attributes, $u_i \in \{1,0\}$ (an *unprotected* attribute) and $s_i \in \{1,0\}$ (a *protected*, i.e. sensitive attribute). Examples of these kinds of attributes are provided in Section 5.3.2. In this paper, we will gather a fully labelled *research data set*, $\mathbf{z}_n$, of n data. These fulfil the role of the training data in machine learning. They comprise $n_{u,s}$ data for each respective (u,s)-subgroup[1] of data, $\mathbf{z}_{u,s}$; i.e.:

$$\begin{aligned} z_i &\equiv \{x_i, u_i, s_i\}, \ \ i \in \{1,\dots,n\} \\ \mathbf{z}_{u,s} &\equiv \{z_i : (u_i,s_i) = (u,s)\} \ \ \#\mathbf{z}_{u,s} \equiv n_{u,s} \qquad (2.1) \\ \mathbf{z} &\equiv \{\mathbf{z}_{u,s} : (u,s) \in \{1,0\} \times \{1,0\}\} \\ n &\equiv \sum_{u,s} n_{u,s}. \qquad (2.2) \end{aligned}$$

The $n_{u,s} \geq 1$ are called the stopping numbers and will be learned from the data (Section 3). For notational ease, we may suppress the indexing variable, so that $z \equiv z_i$. When convenient, we will also denote the labelled data, z, via $x_{u,s}$.

The causal probability model [23] relating the attributes (i.e. labels) to the features (i.e. data) is (Figure 1a)

$$\mathsf{F}(x,u,s) \equiv \mathsf{F}(x|u,s)\overbrace{\Pr[u|s]\Pr[s]}^{p_{u,s}}. \qquad (2.3)$$

We enforce the usual machine learning assumption of static labelled data, i.e. the feature vector field, $\mathbf{x}$, is conditionally independent and identically distributed (ciid) given the observed labels, $\mathbf{u}$ and $\mathbf{s}$:

$$\mathsf{F}(\mathbf{x}|\mathbf{u},\mathbf{s}) \equiv \prod_{i=1}^{n} \mathsf{F}(x_i|u_i,s_i). \qquad (2.4)$$

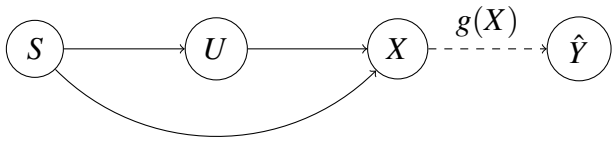
Here, the $\mathsf{F}_{u,s} \equiv \mathsf{F}(x|u,s)$ are the u,s-class-conditional observation models (pdfs or pmfs) for the features (continuous or discrete), which must be learnt (generally nonparametrically) via the respective (u,s)-segmented research data (sub)sets, $\mathbf{z}_{u,s}$. As we accumulate the labelled observations, z_i, we will quench (i.e. stop) learning each of these models only when we satisfy a nonparametric stopping criterion (Section 3) for each component, reaching the data-driven stopping numbers, $n_{u,s}$ (2.1), respectively.

2.1. Societal and AI Unfairness.

Two possible models for the static labelled observations, $\mathsf{F}(z_i) \equiv \mathsf{F}(z)$, are illustrated in Figure 1. Figure 1a demonstrates the presence of AI unfairness, driven by the direct edge (i.e., learning pathway) between s and x, and societal unfairness, driven by the edge between s and u. The details of these definitions are provided in [23]. In order to alleviate AI unfairness, we must break the link between s and x, enforcing independence of the two variables *conditional on* u. In Figure 1b, the conditional independence criterion $(x \perp\!\!\!\perp s)|u$ is satisfied.

This conditional definition of fairness, introduced in [17] and extended in our previous work [23], distinguishes between discrimination which is *illegal* (i.e. based on the sensitive attribute

[1] In this paper, the segmentation of $\mathbf{z}$ (2.2) by u will yield u-indexed *groups*, $\mathbf{z}_u$, each of which is further segmented into (u,s)-indexed *subgroups*, $\mathbf{z}_{u,s}$.



(A) Societal and AI Unfairness

(B) Societal Unfairness Alone

FIGURE 1. Comparison of the causal graphs of different sources of unfairness under our conditional independence model.

S), and that which is *explainable* (i.e. driven by some other influential variable which is crucially not protected). This conditional definition is less stringent than unconditional independence, which means that data needs to be repaired less than under an unconditional scheme, while still satisfying legal definitions of fairness [20].

In common with [23], the principal aim of this paper is statically to transform the labelled research data, $z \equiv x_{u,s}$ (2.1), via (u,s)-indexed (surjective) transformations, $\mathsf{T}_{u,s}$, in order to switch off these unfairness signals in the data, i.e. these lingering dependences of x on s, given each state of u (Fig. 1(a)). Under the ciid and labelled assumptions, these transformations can then be applied to remove AI unfairness from out-of-sample data (e.g. archival data from the same generative process, and, therefore, manifesting AI unfairness). In summary:

$$x_{u,s} \xrightarrow{\mathsf{T}_{u,s}} x'_{u,s} \equiv x'_u \;\; \text{s.t.} \;\; x'_u \sim \mathsf{F}(x'|u,s) \equiv \mathsf{F}(x'|u). \tag{2.5}$$

Here—by construction of the (u,s)-indexed transformations, $\mathsf{T}_{u,s}$—the image, x'_u, is conditionally independent of its sensitive attribute, s, given its unprotected attribute, u. In Section 4, we will use methods of optimal transport (OT) to design the $\mathsf{T}_{u,s}$, $u \in \{1,0\}$.

2.2. **Representation Bias.** Since the class-conditional distributions, $\mathsf{F}_{u,s}$, are unknown (being nonparametric processes in the general case), they must first be inferred from the respective training data sets (i.e. subgroups), $\mathbf{z}_{u,s}$ (2.1), typically via the empirical estimates,

$$\hat{\mathsf{F}}_{\delta,n_{u,s}} \equiv \frac{1}{n_{u,s}} \sum_i \delta_{x_{u,s,i}}, \tag{2.6}$$

where δ_x denotes the Dirac or Kronecker measure at x, as appropriate. In many AI fairness (AIF) settings, u and s are multivariate, i.e. many attributes are posited. This, of course, induces an exponentially increasing number of (u,s)-indexed subgroups (a manifestation of the curse-of-dimensionality, known as *intersectionality* in the AIF literature [13]). Over-segmentation of the n research data may yield u,s-indexed batches which are too small to ensure that learning of all the components, $\mathsf{F}_{u,s}$, is complete, an issue we call *dilution*. Even if intersectionality is avoided (as in the current paper, where u and s are assumed to be *scalar* Bernoulli r.v.s), dilution remains an issue if any of the (predictive) subgroup probabilities, $p_{u,s} \equiv \Pr[u,s]$ (2.3), are inherently small ($\ll \frac{1}{4}$). A central concern of this paper is to decouple the $n_{u,s}$ from the $p_{u,s}$ by implementing a Bayesian sequential decision-making scheme (i.e. stopping rule, Section 3)

which ensures that research data continue to be accumulated until learning of all $\mathsf{F}_{u,s}$ is complete (in the sense defined in Section 3).

Definition 2.1 (Representation bias)**.** The (u,s)-subgroup suffers representation bias in a set of n ciid data if

$$p_{u,s}n < n_{u,s},$$

where $n_{u,s}$ is the stopping number for learning $x_i \overset{\text{ciid}}{\sim} \mathsf{F}_{u,s}$. □

Many current methods for quantifying representation bias rely on approximately equalizing the representation of subgroups (i.e. enforcing $\frac{\Pr[s=0|u]}{\Pr[s=1|u]} \geq \tau_u \; \forall \; u \in \{1,0\}$, where $\tau_u \in (0,1)$ is some threshold (lower bound) of *allowable bias* in the representation of the $s=0$ subgroup[2] in the uth group [4]). Alternative definitions rely on the notion of *coverage*, where there must be a minimum number of samples for each (u,s)-subgroup irrespective of the numbers for all other subgroups [33]. A limitation of these approaches is that the thresholds must be set *a priori*.

Via Definition 2.1, the research data size, $n \equiv \sum_{u,s} n_{u,s}$ (2.2), ensures that

- the subgroup sizes, $n_{u,s}$, are a (stochastic) function of both the class probabilities, $p_{u,s}$, *and* the class models, $\mathsf{F}_{u,s}$ (2.3);
- learning of each $\mathsf{F}_{u,s}$ is completed, in the sense of the stopping rule to be explained in Section 3; i.e. representation bias is avoided in *every* (u,s)-subgroup, including those with intrinsically low $p_{u,s}$.

The proposed stopping rule will achieve *minimally sufficient* representation of each subgroup (as will be seen in Equation 3.5). This satisfies another objective: to minimize the computational complexity of designing and applying the OT-based repairs, $\mathsf{T}_{u,s}$ (2.5). Indeed, the availability of labelled research data—particularly those labelled by protected attributes, s—often requires expensive data-gathering procedures, legislative protection (such as that provided by the recent AI Act of the European Union [9]), and the adoption of arduous privacy-protecting measures. We now outline our scheme for sequential design of the stopping numbers, $n_{u,s}$.

3. Bayesian Learning of the Sub-Group Models, $\mathsf{F}_{u,s}$

For the present, consider any one of the (four) interleaved u,s-indexed learning processes, in each of which we sequentially observe $x_k \overset{\text{ciid}}{\sim} \mathsf{F}(x_k|u,s) \equiv \mathsf{F}_{u,s}$, being the (four) components of the mixture model in (2.3). For notational convenience in this section, we suppress the u,s subscripts, and have used $k \in \{1,2,\ldots\}$ to index the sub-sequence of u,s-indexed features. We also assume that $x_k \in \mathbb{R}$, i.e. that $d=1$ and that the feature is continuous. All of these assumptions can be readily relaxed if required.

Since F is an unknown distribution (here, a pdf), we model it as a nonparametric process, equipped with an appropriate Bayesian nonparametric (BNP) process prior, $\mathscr{F}$ [15], yielding the hierarchy,

$$x \sim \mathsf{F} \;\text{ and }\; \mathsf{F} \sim \mathscr{F}. \tag{3.1}$$

Specifically, we adopt the Dirichlet nonparametric process prior (DPP) [11], $\mathsf{F} \equiv \mathscr{D}(\hat{\mathsf{F}}_0, \nu_0)$, where $\hat{\mathsf{F}}_0 \equiv \mathsf{E}_{\mathscr{D}}[\mathsf{F}]$ is the prior expected distribution of x and $\nu_0 \in \mathbb{R}^+$ is the prior degrees-of-freedom (d.o.f.) parameter. Both of these parameters must be specified *a priori*. Since our

[2]We adopt the common convention of labelling the disadvantaged (sub)group with a zero (0).

uncertainty about F is a monotonically decreasing function of ν_0 [11], a minimally informative choice is a diffuse $\hat{\mathsf{F}}_0$ (such as a quasi-uniform distribution on $\mathbb{R}$, and $\nu_0 \downarrow 0^+$ [30].

The appropriate BNP stopping rule for learning the mixture components, F, is provided in [30]. It emerges via two key properties of the DPP, $\mathscr{D}$ [11]:

(i) $\mathscr{F}$ is the conjugate distribution [2] for learning F via ciid sampling:

$$\begin{aligned} \mathsf{F}|\{x_i,\dots,x_k\} &\sim \mathscr{D}(\hat{F}_k,\nu_k) \equiv \mathscr{D}_k, \\ \hat{F}_k &= (\nu_0+k)^{-1}\left(\nu_0\hat{F}_0+\sum_{j=1}^{k}\delta(x_j)\right), \\ \nu_k &= \nu_0+k, \end{aligned} \tag{3.2}$$

confirming the concentration of $\mathscr{D}$ around its mean distribution, $\hat{F}_k$, as $k\uparrow\infty$. Note that $\hat{F}_k$ is the Bayesian generalization of the classical empirical estimate of F,

$$\begin{aligned} \hat{F}_{\delta,k} &\equiv \frac{1}{k}\sum_{j=1}^{k}\delta(x_j), \\ \alpha_k &\equiv \frac{k}{\nu_0+k}, \\ \hat{F}_k &= (1-\alpha_k)\hat{F}_0+\alpha_k\hat{F}_{\delta,k}, \end{aligned} \tag{3.3}$$

i.e. the $\alpha_k\in(0,1)$-weighted mixture of the prior estimate, $\hat{F}_0$, of the unknown distribution and the empirical estimate, $\hat{F}_{\delta,k}$, based on the sequential realizations, $\mathbf{x}_k$.

(ii) F induces an unknown multinomial distribution, $\mathbf{p}\in\Delta$, on any finite, measurable partition of the codomain, $\mathbb{R}$, of $x_k\equiv x$. In the case where $\mathsf{F}\sim\mathscr{D}$, then the distribution induced on the simplex, Δ, is the Dirichlet distribution, $\mathbf{p}\in\mathsf{D}$.

The finite measurable partition in (ii) can be thought of as a *quantization*, $\mathsf{Q}_{\mathbb{V}}$, of $x\overset{\text{ciid}}{\sim}\mathsf{F}$:

$$x\xrightarrow{\mathsf{Q}_{\mathbb{V}}}x'\equiv\mathsf{Q}_{\mathbb{V}}(x),$$

where $\mathbb{V}$ is the pre-defined set of vertices of the quantizer. Here—as in [30]—$\mathbb{V}$ is sequentially defined as the ordered observations, $x_{(j)}$, themselves:

$$\begin{aligned} \mathbb{V}_0 &\equiv \{-\infty,+\infty\} \\ \mathbb{V}_k\equiv\{-\infty,\sigma_1,\dots,\sigma_k,+\infty\} &\equiv \mathbb{V}_{k-1}\cup\{x_k\},\ \ k=1,2,\dots, \\ \sigma_j &\equiv x_{(j)}. \end{aligned} \tag{3.4}$$

Note that the $x_k\overset{\text{ciid}}{\sim}\mathsf{F}$ are a.s. distinct in the assumed case where F is a continuous measure. In the discrete case, $\mathbb{V}_k$ assembles all distinct observed states of x (i.e. no repetitions), and so $\#\mathbb{V}_k-2\leq k$ in general.

This sequential, data-driven (re)quantization process, $\mathbb{V}_k$, $k=1,2,\dots$, constitutes a *partition refinement schedule*, in the sense that the number, $\#\mathbb{V}_k-2$, of ordered vertices, σ_j, is an increasing function of the number, k, of ciid observations, $\mathbf{x}_k\equiv\{x_1,\dots,x_k\}$. As $k\uparrow\infty$, knowledge of the underlying (typically continuous) unknown distribution, $\mathsf{F}\sim\mathscr{D}_k$ (3.2), concentrates around its sequential expected value, $\hat{\mathsf{F}}_k$ (i.e. the prior-empirical weighted mixture (3.3)). Learning can be quenched under a suitable stopping rule (criterion). In [30], this is constructed via the sequence of Kullback-Leibler divergences (KLDs) [22], $\mathsf{KLD}[\mathsf{D}_k||\mathsf{D}_{k-1}]$, $k\geq 2$, i.e. via the sequence of

Dirichlet distributions, D_k (such that $\mathbf{p}_k \sim \mathsf{D}_k$), induced via the proposed data-driven partition refinement schedule (3.4)[3]. The stopping rule is therefore defined via a threshold, $\varepsilon > 0$:

$$\hat{n} \equiv \min_{k \in \mathbb{N}^+} \{k : \mathsf{KLD}[\mathsf{D}_k || \mathsf{D}_{k-1}] < \varepsilon\}. \tag{3.5}$$

In practice, a smoothed version of the KLD sequence (3.5) is adopted. The details are provided in [30], along with the implied algorithm.

Recall that this BNP learning and stopping procedure is applied, in parallel, to all four components, $\mathsf{F}_{u,s}$, of the attribute-structured mixture model (2.3), yielding attribute-dependent stopping numbers, $\hat{n}_{u,s}$, $(u,s) \in \{1,0\} \times \{1,0\}$. The stopping thresholds can be chosen attribute-wise, i.e. $\varepsilon_{u,s}$. These provide a set of operating conditions which trade off sample complexity against accuracy of the BNP learning process for the $\mathsf{F}_{u,s}$ (3.1).

4. Data-Driven Fairness Correction

In this section, we propose an adaptation of the distributional approach for data repair in [23] to overcome representation bias using the nonparametrically learnt subgroup models, $\mathsf{F}_{u,s}$ (Section 3). This is a fundamentally different paradigm from other approaches to data repair [7, 8, 10], which do not comment on the selection of training data and hence representation bias may be amplified through the repair process.

4.1. Learning the Repair Operation.

More formally, our labelled and ciid research (i.e. training) data set, $\mathbf{z}$, is defined according to the quenching (i.e. stopping) of learning (3.5) in each (u,s)-subgroup (2.2). We consider the set of centroids of each interior cell, i.e. the arithmetic means of the interior vertices $\sigma \in \mathbb{V}_{u,s}$:

$$q_{u,s,j} \equiv \frac{(\sigma_j + \sigma_{j+1})}{2}, 1 \leq j \leq \hat{n}_{u,s} - 1, \tag{4.1}$$

where $q_{u,s,j} \in \mathbb{Q}_{u,s}$ is the centroid of the j^{th} cell.

Under the proposed partition refinement schedule (3.4), the quantized conditionals, $\mu_{u,s}$, are uniform on the ordered set $\mathbb{Q}_{u,s}$, such that

$$\mu_{u,s} \equiv \frac{1}{\hat{n}_{u,s} - 1} \sum_j \delta_{q_{u,s,j}}. \tag{4.2}$$

Note that this approach is distinct from the kernel density approximation method in [23], since we consider a uniform $\mu_{u,s}$ over a support defined by the observations $\mathbf{x}_{u,s}$ (4.1). Following [23], we employ techniques from optimal transport (OT) to design our repair. For a detailed treatment of OT theory, and further background related to OT for fairness, see [29] and [5, 7, 10] respectively.

The Wasserstein distance from $\mu_{u,0}$ to $\mu_{u,1}$ with respect to the cost function $\mathsf{C}(q_{j_0}, q_{j_1}) \equiv \mathsf{L}_p^p$, the latter being the p^{th} power of the L^p norm (with $p \equiv 2$ in this work):

$$\mathsf{W}_p^p(\mu_{u,0}, \mu_{u,1}) \equiv \min_{\pi \in \Pi(\mu_{u,0}, \mu_{u,1})} \sum_{q_{j_1}} \sum_{q_{j_0}} \mathsf{C}(q_{j_0}, q_{j_1}) \pi(q_{j_0}, q_{j_1}).$$

[3]The (a.s. in the continuous case) sequential splitting of mass in the multinomial process, $\mathbf{p}_k$, is a type of stick-breaking process [32] for representation of the underlying DPP (3.2). Importantly, the splitting probabilities are data-driven here, being controlled by the relationship between x_k and $\mathbf{x}_{k-1}$.

Here, $\Pi(\cdot)$ is the set of joint pmfs with marginals $\mu_{u,0}$ and $\mu_{u,1}$, defined on the product space, $\mathbb{Q}_{u,0} \times \mathbb{Q}_{u,1}$.

The t-indexed Wasserstein barycentres of $\mu_{u,0}$ and $\mu_{u,1}$ are

$$\nu_{t,u} \equiv \min_{\nu} \left\{ (1-t)\mathsf{W}_2^2(\mu_{u,0}, \nu) + t\mathsf{W}_2^2(\mu_{u,1}, \nu) \right\}, t \in [0,1], \tag{4.3}$$

defining the geodesic between the two empirical marginals.

We consider the centre of the geodesic, $\nu_u \equiv \nu_{0.5,u}$, as our fair u-conditional target design, since it is s-invariant by definition, and equally incurs damage—in a sense to be defined in Section 4.3—to both s-subgroups simultaneously. We compute ν_u via the $(\hat{n}_{u,0} - 1) \times (\hat{n}_{u,1} - 1)$ OT plan between the (u,s)-conditional quantized distributions, $\mu_{u,s}$:

$$\pi_u^* \equiv \underset{\pi \in \Pi\left(\mu_{u,0}, \mu_{u,1}\right)}{\operatorname{argmin}} \sum_{q_{j_1}} \sum_{q_{j_0}} \mathsf{C}\left(q_{j_0}, q_{j_1}\right) \pi\left(q_{j_0}, q_{j_1}\right). \tag{4.4}$$

Now, consider a labelled draw $x_{u,s} \overset{\text{ciid}}{\sim} \mathsf{F}_{u,s}$. Our task is to repair this datum by driving it to the corresponding point $x' \sim \nu_u$ on the barycentre (2.5). To achieve this, we define a stochastic operator $\mathsf{T}_{u,s}$ with two sources of randomness as follows:

(i) Denoting the round-down (i.e. truncation) state of x in $\mathbb{Q}_{u,s}$ by $q_{j_s} \equiv \lfloor x \rfloor$, we design a Bernoulli trial $b \sim \mathscr{B}(p)$ where $p \equiv \frac{x - q_{j_s}}{q_{j_s+1} - q_{j_s}}$. Then, $j_s \leftarrow j_s + b$ defines the appropriate row (where $s = 0$) or column (where $s = 1$) of the transport plan π_u^*.

(ii) Let $\pi^*_{u;j_0,:}$ be the $j_0{}^{\text{th}}$ row (or $\pi^*_{u;:,j_1}$ the $j_1{}^{\text{th}}$ column) of π_u^*. Depending on the value of $s \in \{1,0\}$, this row (or column) defines the probability vector of a multinomial trial. For $s = 0$, $\Pr[j_1 | u, s = 0, x] \propto \pi^*_{u;j_0,j_1}$, yielding the (randomly) repaired state, q_{j_1}. Correspondingly, for $s = 1$, j_1 provides the column index into π_u^*, yielding the multinomial for realizing the (randomly) repaired state, q_{j_0}.

Finally, the OT repair (2.5), mapping to the geodesic centre ($t = 0.5$) (4.3), is

$$x_{u,s} \to x'_u \equiv \mathsf{T}_{u,s}(x_{u,s}) \equiv 0.5 q_{j_0} + 0.5 q_{j_1} \tag{4.5}$$

4.2. **Evaluating (Un)Fairness.** Clearly, there are different degrees of fairness (Section 2.1) associated with the pre- ($\mathbf{x}$) and post-repair ($\mathbf{x}'$) data. We would like to quantify this fairness. To this end, we adapt the KLD-based fairness metric proposed by [23]. Unlike other widely-adopted fairness metrics [14, 27, 36], this measure does not depend on classification outcome (and therefore is model-invariant), rather it is a function of the complete distribution (Equation 2.3).

As proposed in [23], we evaluate the s-dependence of the u-conditional distributions using the u-expectation of the symmetrized Kullback–Leibler Divergence (KLD), $E(\mathbf{x})$:

$$\begin{aligned} E_u(\mathbf{x}) &\equiv \frac{1}{2}\mathsf{D}\left[\mathsf{F}_{u,0} \,||\, \mathsf{F}_{u,1}\right] + \frac{1}{2}\mathsf{D}\left[\mathsf{F}_{u,1} \,||\, \mathsf{F}_{u,0}\right], \\ E(\mathbf{x}) &= \sum_u \Pr[u]\, E_u(\mathbf{x}). \end{aligned} \tag{4.6}$$

The smaller this quantity, the fairer the distribution in the sense of breaking the conditional dependence between x and s, given u (Figure 1). We calibrate the unfairness after repair, $E(\mathbf{x}')$,

against the unfairness of the original data, yielding the summary metric, $\hat{E}$:

$$\hat{E} \equiv \frac{E(\mathbf{x}')}{E(\mathbf{x})} \tag{4.7}$$

For a competent repair, $\hat{E}$ should be less than 1, with $\hat{E} = 1$ corresponding to no improvement over the original data, and $\hat{E} = 0$ indicating total s-invariance (fairness).

4.3. **Data Damage.** Unfair data are being transformed in order to satisfy notions of fairness. A trivial way to ensure fairness according to (4.6) would be to enforce $x'_{u,s} \equiv \mathsf{const.}\ \forall(u,s) \in \{1,0\} \times \{1,0\}$, However, this destroys all predictive information in the features. We introduce a novel metric to evaluate this notion of *data damage*.

Following principles of information projection [6] and fully probabilistic design (FPD) [18], we evaluate the KLD from the repaired (fair) data distribution to the original (unfair) data distribution, in order to measure the degree to which data have been damaged by the OT repair operation (4.5).

Definition 4.1.

$$D_{u,s} \equiv \mathsf{D}\left[\mathsf{F}_{u,s} \,||\, \mathsf{F}'_{u,s}\right], \tag{4.8}$$

where $\mathsf{F}'_{u,s} \equiv \mathsf{F}(x'|u,s)$.

A summary quantifier of the damage incurred in the repaired data distribution, $\mathsf{F}(x')$, the (u,s)-expected value of $D_{u,s}$ is evaluated:

$$D = \sum_u \Pr[u] \sum_s \Pr[s|u]\ D_{u,s}. \tag{4.9}$$

Once again, smaller is better: $D \downarrow 0$ represents data which have been damaged less.

5. Experiments

In this section, we detail a number of experiments designed to validate the key proposals of this work. In Section 5.1, we explore the behaviour of the stopping rule proposed in Section 3 with quantized and mixture data in order to assess it in realistic fairness problems. Then, in Section 5.2, we evaluate the performance of our repair method with respect to (4.7) and (4.9) under varying mixture probabilities, $Pr[u,s]$, to simulate the impact of representation bias on our repair. Section 5.3 contains the results of benchmarking against two SOTA data repair approaches: geometric repair [10, 7] and distributional repair [23].

Throughout the experiments, unless otherwise specified, we set the prior confidence $\nu_0 \equiv 0.001$, the stopping criterion $\varepsilon \equiv 0.01$, we take the uniform prior $\hat{F}_0 \equiv \mathscr{U}(x_{\min}, x_{\max})$, and we assume scalar data (features), i.e. $d \equiv 1$.

5.1. **Learning non-Gaussian Models.**

5.1.1. *Multinomial Models.* For quantized data, the probability of coincident observations is non-zero. To evaluate the performance of the stopping rule in this setting, we construct categorical distributions with q categories, where the categorical probability is based on some reference measure X. We set $X \equiv \mathcal{N}(0,1^2)$, and define $p_i \equiv F_X(x_{i+1}) - F_X(x_i) \forall i \leq q$ where F_X is the cumulative distribution function for X and the support is defined $x \in \mathbb{Q}, \mathbb{Q} \equiv \{-5.0, \cdots, -5.0 + \frac{10i}{q}, \cdots, 5.0\}, |\mathbb{Q}| \equiv q+1$.

We evaluate the evolution of the log-KLD (LKLD) until stopping at $n \equiv \hat{n}$ (3.5) for five categorical models with $q \in \{5, 10, 50, 500, 5000\}$. As is evident from Figure 2, learning is quenched sooner for coarsely quantized data with low q. This suggests the stopping rule is able to exploit the information encoded in the coincident nature of samples to stop sooner.

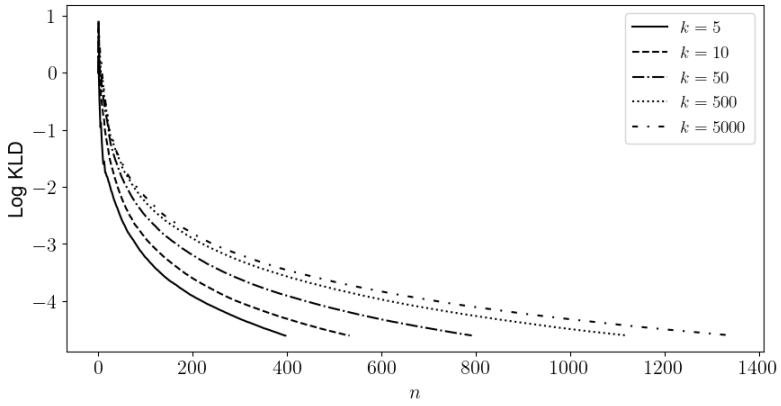


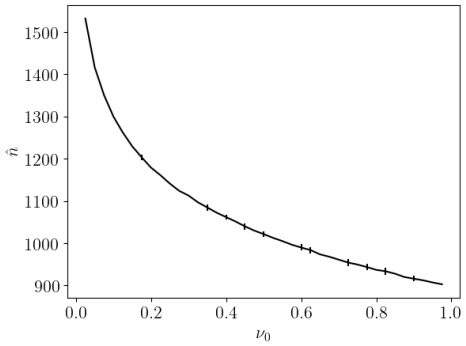
FIGURE 2. LKLD until stopping for observations from categorical variables with q states.

5.1.2. *Gaussian Mixture Models.* We further validate the stopping rule's convergence on Gaussian mixture model (GMM) data with a minority component, where $\mathsf{F} \equiv \sum_{i=1}^{2} \alpha_i \mathcal{N}(m_i, \sigma_i^2)$ and $\alpha \equiv [0.8, 0.2]$, $\mathbf{m} \equiv [-1, -5]$, $\sigma \equiv [1, 0.5]$. This model simulates the presence of intersectionality [13], such that within (u,s)-conditional subgroups a mixture of components remains. In this section, we conduct three experiments: (i) we evaluate the effect of the prior confidence ν_0 on stopping, (ii) the effect of a non-uniform prior, and (iii) we validate the behaviour of the stopping rule over 500 Monte-Carlo simulations.

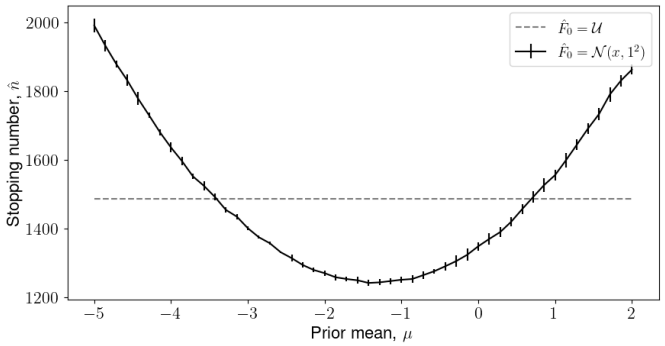
Figure 3A demonstrates that as $\nu_0 \to 1$, stopping occurs sooner since the prior has more influence over stopping. Where we are uncertain about the prior, a low $\nu_0 \approx 0.001$ ensures that stopping is data-driven.

The theoretical results in Section 3 are validated by our empirical findings in Figure 3B. The uniform prior is conservative (in the sense that it is minimally informative), with the Gaussian prior $\hat{F}_0 \equiv \mathcal{N}(m, 1^2)$ yielding higher stopping numbers, $\hat{n}$, when $x \not\sim \hat{F}_0$, i.e. when there is mismatch between the prior and the data generating process F (2.3).

Figure 4 shows that the stopping numbers, $\hat{n}$, are independent of the individual random data realisations, with the stopping rule converging to the underlying mean and variance when learning is quenched. This is consistent with the results reported in [30] for Student's t-distribution and supports the application of the stopping rule to the mixture-type data typical in machine learning applications.



(A) Stopping numbers, $\hat{n}$, for varying prior weights, ν_0 (and $\hat{\mathsf{F}}_0$ uniform).



(B) Stopping numbers, $\hat{n}$, for different prior means, $\hat{F}_0 \equiv \mathcal{N}(m, 1^2)$, compared to the uniform prior.

FIGURE 3

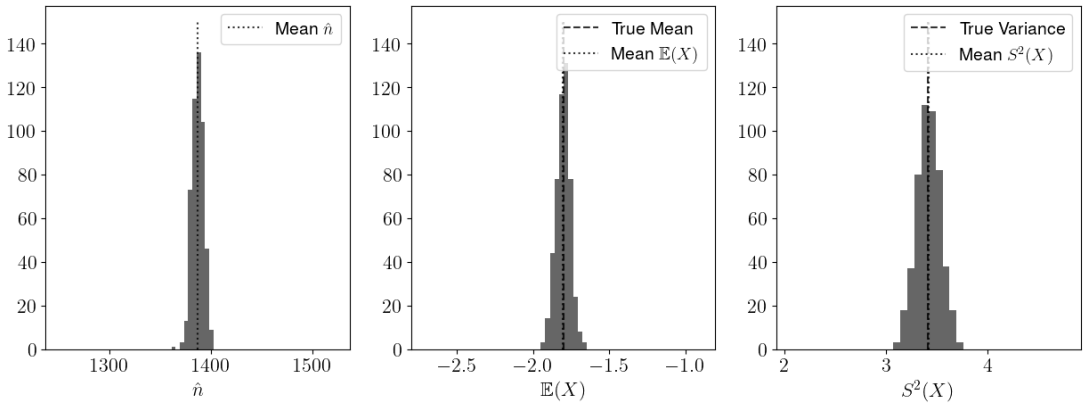


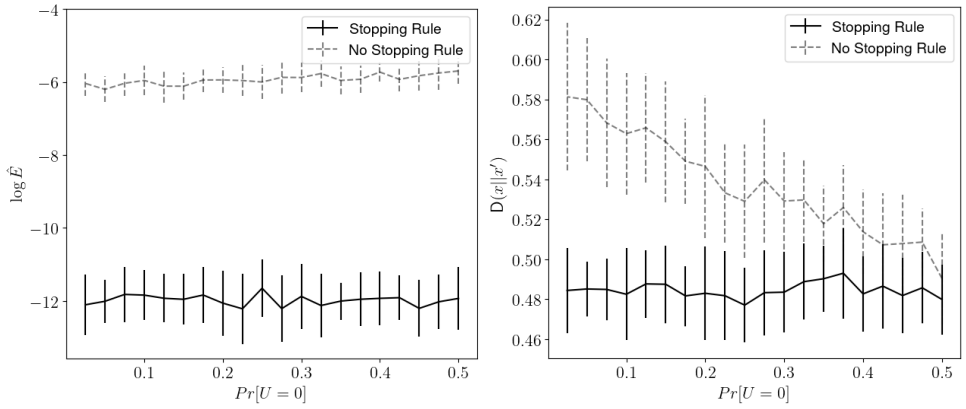
FIGURE 4. Stopping numbers, $\hat{n}$, sample mean, $\mathbb{E}(X)$, and sample variance, $S^2(X)$, for GMM data over 500 Monte-Carlo simulations.

5.2. **Fairness Repair Under Representation Bias.** To evaluate the method's performance with varying levels of representation bias, we consider a setting with Gaussian components $F(x,u,s)$ as given in Table 1. The mixture weights, $p_{u,s}$, are chosen subject to the constraint that $\Pr[s|u] \equiv 0.5$, $u \in \{1,0\}$. The u-weight, $\Pr[U=0]$, is then varied in $(0,0.5]$ to simulate balanced and unbalanced u classes.

u	s	$\mathsf{F}(x\|u,s)$
0	0	$\mathcal{N}(-1.0, 1^2)$
0	1	$\mathcal{N}(1.0, 1.2^2)$
1	0	$\mathcal{N}(-0.5, 1.2^2)$
1	1	$\mathcal{N}(1.5, 0.8^2)$

TABLE 1. Model components (2.3) in the representation bias simulation. $p_{u,s}$ is varied throughout the experiment.

Figure 5 demonstrates a key strength of our approach: even with significant representation bias in the data-generating process, where $Pr[U=0] \equiv 0.025$, our data-driven approach is able to repair data reliably. Further, repair damage is also shown to be representation bias-invariant, with no more damage to under-represented components than when representational parity is achieved at $Pr[U=0]=0.5$. A modified approach without the stopping rule, where $n_{u,s} \equiv p_{u,s}\hat{n}$, is unable to achieve comparable levels of repair even for $Pr[U=0] \approx 0.5$ (being the case in which representation bias is minimal) due to incomplete learning of subgroup models for some (u,s).



(A) S-invariance. (B) Data damage.

FIGURE 5. $\log \hat{E}$ and damage (Definition 4.1) for data with simulated representation bias $Pr[U=0] \in (0, 0.5]$. Results are reported $\pm$ standard deviation over 20 Monte-Carlo simulations.

5.3. **Benchmarking on Simulated and Real Data.** In this section, we compare our method to two state-of-the-art approaches: the distributional repair proposed in [23], and the geometric repair proposed by [10] and extended by [7]. Neither of these approaches deals explicitly with representation bias, and hence we evaluate these approaches on an alternative dataset, $\tilde{\mathbf{z}}_{u,s}$, defined as (2.2) with $n_{u,s} \equiv p_{u,s}\hat{n}$ (i.e. retaining the representation bias in u and s of the overall model F in a sample of equal size to the total stopping numbers).

5.3.1. *Simulated GMM Data with Intersectionality.* In order to simulate the impact of intersectionality and representation bias on different repair methods, we design a simulated mixture model dataset with $p_{u,s}$ imbalance parameterised as in Table 2. Note the bias in both the mixing probabilities $p_{u,s}$ and the mixture weights in $\mathsf{F}(x|s,u)$.

As Table 3 shows, our approach is able to outperform both geometric [10, 7] and distributional [23] approaches with respect to the s-invariance metric $\hat{E}$ (4.7) for both on- and off-sample repair. This is due to the complete learning of all $\mathsf{F}_{u,s}$ components, ensuring that underrepresented u,s subgroups are not neglected in the repair operation.

While distributional repair leads to the less damage than our approach, this is likely due to the higher remaining s-conditional dependence after distributional repair. The trade-off between the amount of repair and the predictive utility of data is explored in detail in [7].

u	s	$p_{u,s}$	$\mathsf{F}(x\|s,u)$
0	0	0.18	$0.8\mathscr{N}(-1.0,1^2)+0.2\mathscr{N}(-5.0,0.5^2)$
0	1	0.12	$0.6\mathscr{N}(1.0,1.2^2)+0.4\mathscr{N}(-1.75,0.5^2)$
1	0	0.42	$0.5\mathscr{N}(-1.0,1^2)+0.5\mathscr{N}(3.5,1.2^2)$
1	1	0.28	$0.1\mathscr{N}(-2.0,0.8^2)+0.9\mathscr{N}(5.0,1.5^2)$

TABLE 2. Model Components for simulated intersectionality study.

Method	On-Sample $\log\hat{E}$	On-Sample Damage	Off-Sample $\log\hat{E}$	Off-Sample Damage
Geometric [10, 7]	-8.365 ± 0.172	0.394 ± 0.020	-	-
Distributional [23]	-4.491 ± 0.120	$\mathbf{0.307\pm0.016}$	-4.229 ± 0.232	$\mathbf{0.325\pm0.016}$
Our Approach	$\mathbf{-13.550\pm0.782}$	0.399 ± 0.018	$\mathbf{-5.385\pm0.375}$	0.431 ± 0.020

TABLE 3. Repair metrics $\hat{E}$ and D for Geometric, Distributional and Stick-Breaking Repairs on a simulated GMM dataset over 500 Monte-Carlo trials. Note that $\hat{E}$ is reported on the log scale for clarity. Geometric repair cannot repair off-sample data.

5.3.2. *Adult Income Data.* We further validate our repair method on the Adult Income dataset [24]. The dataset consists of information about $n_0 \equiv 48,882$ individuals, including their age, sex, race, education level, and annual income. We consider sex as the protected attribute, with $S=1$ corresponding to the males and $S=0$ to the non-males. Our unprotected, explanatory attribute is education level, with $U=0$ corresponding to an education level of high-school graduate or below. In this data, both representation bias (due to differing access to higher education between men and women) and intersectionality (due to the impact of race on education) are present. Each subgroup's prior probability is given in Table 4.

u	s	$Pr[u,s]$
0	0	0.146
0	1	0.310
1	0	0.187
1	1	0.357

TABLE 4. Mixture weights for the Adult income data, stratified by sex and education level.

Table 5 summarises the s-conditional dependence remaining in the data after repair with (i) geometric repair and (ii) our method, and demonstrates that for on-sample repairs our method is able to match or outperform geometric repair, and is able to reduce the s-dependence present in unseen data by at least three times.

Feature	Geometric Repair On-Sample $\hat{E}$	Geometric Repair Off-Sample $\hat{E}$	Our Approach On-Sample $\hat{E}$	Our Approach Off-Sample $\hat{E}$
Age	**0.0031**	-	0.0041	**0.1374**
Capital Gain	0.2222	-	**0.1366**	**0.2103**
Capital Loss	*3.1058*	-	**0.5638**	**0.3018**

TABLE 5. $\hat{E}$ for Geometric and Stick-Breaking Repairs on the Adult dataset. Note that Geometric repair cannot repair off-sample data, and that $\hat{E} > 1$ corresponds to making data *less* fair with respect to S.

6. DISCUSSION

The results presented in Section 5 demonstrate the performance of our approach on a number of diverse models, F. The experiments in Section 5.1 support the Bayesian nonparametric theory of Section 3 and results presented in [30], demonstrating that the stopping rule is appropriate for a number of data-generating processes, including discretised and mixture models. In this Section, we further validate that – while the stopping rule is not prior-invariant – the *cost* of an inaccurate prior is not incomplete learning but rather a higher stopping number $\hat{n}$ (3.5). The uniform prior, $\hat{F}_0 \equiv \mathscr{U}$, is shown to be a powerful choice when there is large uncertainty about F, such as the fairness setting when we have no knowledge about the data-generating process *a priori*. Empirical guidance for the selection of the operating conditions of our method are given in Section 5.1, and applied through Sections 5.2 and 5.3.

The key advantage of our approach under representation bias is demonstrated in Section 5.2, where our data-driven stick-breaking approach is able to achieve representation-invariant repairs. Similarly, the results in Section 5.3 demonstrate that our approach performs better than the SOTA on a simulated GMM dataset designed to mimic the relationship between representation bias and intersectionality in real data.

It remains to calibrate our damage metric (Definition 4.1) on the unrepaired data as in our repair metric $\hat{E}$ (4.7), perhaps via the cross-entropy of the repaired data relative to unrepaired data.

One interesting avenue for future work is regarding exploiting the early stopping on some u,s subgroups to begin conducting repairs (4.5) sooner for samples $x_{u,s}$ from these quenched processes. Especially under data-streaming paradigms with large volumes of data, this approach has the potential to smooth the computational overhead of the mode-change from learning to repair. Additionally, it remains to adapt this method to non-stationary distributions, i.e. those with some drift in the underlying data-generating process, since the stationarity assumption in this work is particularly restrictive. Existing notions of forgetting [21], specifically the work by [30] on Bayesian nonparametrics for non-stationary processes, could be adapted to match our data repair paradigm and generalise our approach to a much broader class of $\mathsf{F}_{u,s}$.

7. CONCLUSION

In this work, we have presented a novel method for fairness correction which exploits a data-driven stopping rule to alleviate representation bias in data by ensuring all u,s conditional distributions are learned completely through a nonparametric stopping rule.

We present a comprehensive study of the performance of our Bayesian approach to learning the underlying distribution of sequential data under different operating conditions and when applied to diverse data modalities to simulate real world fairness correction. Further, we demonstrate that our repair approach is able to repair data even with severe representation bias, where the minority class is observed fewer than one time in 20, yielding repair quality and damage which are invariant to this bias. We also compare our method to the SOTA geometric and distributional repair operations on simulated GMM data and the Adult Income dataset.

Our method consistently outperforms the SOTA where representation bias is present, demonstrating the versatility of this data-driven approach. With the recent deployment of the AI Act [9], this is a key step towards the generalisability of fairness tools for machine learning applications.

Authors' Note

Uri Rothblum was a regular visitor to Professor Robert Shorten at the Hamilton Institute in Ireland. He was a scholar and a gentleman and is missed by all his friends, colleagues and collaborators.

Acknowledgements

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101070568. This work was also supported by Innovate UK under the Horizon Europe Guarantee, UKRI Reference Number: 10040569 (Human-Compatible Artificial Intelligence with Guarantees (AutoFair)).

REFERENCES

[1] S. Barocas, M. Hardt, and A. Narayanan. Fairness and machine learning: Limitations and opportunities. MIT Press, 2023.

[2] J. Bernardo and A. Smith. Bayesian Theory. Wiley, 2010.

[3] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. Advances in neural information processing systems, 30, 2017.

[4] L. E. Celis, V. Keswani, and N. Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In: International Conference on Machine Learning, pages 1349–1359. PMLR, 2020.

[5] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with Wasserstein barycenters. Advances in Neural Information Processing Systems, 33 (2020) 7321-7331.

[6] T. M. Cover. Elements of Ioformation Theory. John Wiley & Sons, 1999.

[7] E. Del Barrio, G. Fabrice, P. Gordaliza, and J.-M. Loubes. Obtaining fairness using optimal transport theory. In: International conference on machine learning, pages 2357–2365. PMLR, 2019.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In :Proceedings of the 3rd innovations in theoretical computer science conference, pages 214–226, 2012.

[9] European Commission. Laying down harmonised rules on artificial intelligence (artificial intelligence act), 2021.

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In :proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pages 259–268, 2015.

[11] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. The annals of Statistics, 1 (1973) 209-230.

[12] G. Foody, M. McCulloch, and W. Yates. The effect of training set size and composition on artificial neural network classification. International Journal of Remote Sensing, 16 (1995) 1707–1723.

[13] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan. An intersectional definition of fairness. In : 2020 IEEE 36th International Conference on Data Engineering (ICDE), pages 1918–1921. IEEE, 2020.

[14] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In :Proceedings of the conference on fairness, accountability, and transparency, pages 329–338, 2019.
[15] S. Ghosal and A. van der Vaart. Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press, Cambridge, 2017.
[16] A. Ghosh, L. Genuit, and M. Reagan. Characterizing intersectional group fairness with worst-case comparisons. In : Artificial Intelligence Diversity, Belonging, Equity, and Inclusion, pages 22–34. PMLR, 2021.
[17] F. Kamiran, I. Žliobaitė, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and information systems, 35 (2013) 613–644.
[18] M. Kárnỳ and T. Kroupa. Axiomatisation of fully probabilistic design. Information Sciences, 186 (2012) 105–113.
[19] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807, 2016.
[20] L. K. Koumeri, M. Legast, Y. Yousefi, K. Vanhoof, A. Legay, and C. Schommer. Compatibility of fairness metrics with eu non-discrimination laws: Demographic parity & conditional demographic disparity. arXiv preprint arXiv:2306.08394, 2023.
[21] R. Kulhavỳ and M. B. Zarrop. On a general concept of forgetting. International Journal of Control, 58 (1993) 905–924.
[22] S. Kullback and R. A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, 22 (1951) 79–86.
[23] A. Langbridge, A. Quinn, and R. Shorten. Optimal transport for fairness: Archival data repair using small research data sets. In: Proceedings of the 40th International Conference on Data Engineering Workshops. IEEE, 2024.
[24] M. Lichman. Adult dataset, UCI Machine Learning Repository, 2013.
[25] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54 (2021) 1–35.
[26] M. Mohri, A. Rostamizadeh, and A. Talwalkar. Foundations of machine learning. MIT press, 2018.
[27] A. Narayanan. 21 fairness definitions and their politics. Tutorial presented at the Conf. on Fairness, Accountability, and Transparency, 2018.
[28] T. Oates and D. Jensen. The effects of training set size on decision tree complexity. In: Sixth International Workshop on Artificial Intelligence and Statistics, pages 379–390. PMLR, 1997.
[29] G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning, 11 (2019) 355–607.
[30] A. Quinn and M. Kárnỳ. Learning for non-stationary Dirichlett processes. International Journal of Adaptive Control and Signal Processing, 21 (2007) 827–855.
[31] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In: Proceedings of the 2019 International Conference on Management of Data, pages 793–810, 2019.
[32] J. Sethuraman. A constructive definition of Dirichlet priors. Statistica Sinica, pages 639–650, 1994.
[33] N. Shahbazi, Y. Lin, A. Asudeh, and H. Jagadish. Representation bias in data: a survey on identification and resolution techniques. ACM Computing Surveys, 55 (2023) 1–39.
[34] S. Shalev-Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In: Proceedings of the 25th international conference on machine learning, pages 928–935, 2008.
[35] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536, 2017.
[36] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web, pages 1171–1180, 2017.