Data preprocessing to mitigate bias: A maximum entropy based approach

L. Elisa Celis [1] **Vijay Keswani** [1] **Nisheeth K. Vishnoi** [2]

Abstract

Data containing human or social attributes may over- or under-represent groups with respect to salient social attributes such as gender or race, which can lead to biases in downstream applications. This paper presents an algorithmic framework that can be used as a data preprocessing method towards mitigating such bias. Unlike prior work, it can efficiently learn distributions over large domains, controllably adjust the representation rates of protected groups and achieve target fairness metrics such as statistical parity, yet remains close to the empirical distribution induced by the given dataset. Our approach leverages the principle of maximum entropy – amongst all distributions satisfying a given set of constraints, we should choose the one closest in KL-divergence to a given prior. While maximum entropy distributions can succinctly encode distributions over large domains, they can be difficult to compute. Our main contribution is an instantiation of this framework for our set of constraints and priors, which encode our bias mitigation goals, and that runs in time polynomial in the *dimension* of the data. Empirically, we observe that samples from the learned distribution have desired representation rates and statistical rates, and when used for training a classifier incurs only a slight loss in accuracy while maintaining fairness properties.

1. Introduction

Datasets often under- or over-represent social groups defined by salient attributes such as gender and race, and can be a significant source of bias leading to discrimination in the machine learning applications that use this data (O'Neil, 2016; Calders & Žliobaitė, 2013; Kay et al., 2015). Methods to debias data strive to ensure that either 1) the representation of salient social groups in the data is consistent with ground truth (King & Zeng, 2001; Chawla et al., 2002; Zelaya, 2019), or 2) the outcomes (where applicable) across salient social groups are fair (Calders et al., 2009; Kamiran & Calders, 2012; Wang et al., 2019; Calmon et al., 2017; Xu et al., 2018; Feldman et al., 2015; Gordaliza et al., 2019). The goal of this paper is to learn a distribution that corrects for *representation* and *outcome* fairness but also remains as *close* as possible to the original distribution from which the dataset was drawn. Such a distribution allows us to generate new pseudo-data that can be used in downstream applications which is both true to the original dataset yet mitigates the biases it contains; this has the additional benefit of not requiring the original data to be released when there are privacy concerns. Learning this distribution in time polynomial in the size of the dataset and dimension of the domain (as opposed to the size of the domain, which is exponential in the number of attributes and class labels) is crucial in order for the method to be scalable. Further, attaining provable guarantees on the efficiency and desired fairness properties is an important concern. Hence, the question arises:
Can we develop methods to learn accurate distributions that do not suffer from biases, can be computed efficiently over large domains, and come with theoretical guarantees?

Our contributions. We propose a framework based on the maximum entropy principle which asserts that among all distributions satisfying observed constraints one should choose the distribution that is "maximally non-committal" with regard to the missing information. It has its origins in the works of Boltzmann, Gibbs and Jaynes (Gibbs, 1902; Jaynes, 1957a;b) and it is widely used in learning (Dudik, 2007; Singh & Vishnoi, 2014). Typically, it is used to learn probabilistic models of data from samples by finding the distribution over the domain that minimizes the KL-divergence with respect to a "prior" distribution, and whose expectation matches the empirical average obtained from the samples.

Our framework leverages two properties of max-entropy distributions: 1) any entropy maximizing distribution can be succinctly represented with a small (proportional to the dimension of the data) number of parameters (a consequence of duality) and, 2) the prior and expectation vector provides simple and interpretable "knobs" with which to control the statistical properties of the learned distribution.

[1]Department of Statistics and Data Science, Yale University, USA [2]Department of Computer Science, Yale University, USA. Correspondence to: L. Elisa Celis <elisa.celis@yale.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Table 1. **Comparison of our paper with related work:** The first two rows denote the fairness metrics that can be controlled by each approach (see Definitions 2.1 and 2.2). The last two rows denote whether the approach has the ability to sample from the entire domain, and whether it has a succinct representation. We compare our performance against these methods empirically in Section 5.

Properties	(Kamiran & Calders, 2012)	(King & Zeng, 2001)	(Calmon et al., 2017)	This paper
- Statistical Rate	✓(only for $\tau = 1$)	✗	✓(only for $\tau = 1$)	✓
- Representation Rate	✗	✓	✗	✓
- Entire domain	✗	✗	✓	✓
- Succinct representation	✓	✓	✗	✓

We show that by appropriately setting the prior distribution and the expectation vector, we can provably enforce constraints on the fairness of the resulting max-entropy distribution, as measured by the representation rate (the ratio of the probability assigned to the under-represented group and the probability assigned to the over-represented group - Definition 2.1) and statistical rate (the ratio of the probability of belonging to a particular class given individual is in the under-represented group and the probability of belonging to the same class given individual is in the over-represented group - Definition 2.2); see Theorem 4.5. However, existing algorithms to compute max-entropy distributions depend on the existence of fast oracles to evaluate the dual objective function and bounds on the magnitude of the optimal (dual) parameters (Singh & Vishnoi, 2014; Straszak & Vishnoi, 2019). Our main technical contribution addresses these problems by showing the existence of an efficient and scalable algorithm for gradient and Hessian oracles for our setting and a bound on the magnitude of the optimal parameters that is polynomial in the dimension. This leads to algorithms for computing the max-entropy distribution that runs in time polynomial in the size of the dataset and dimension of the domain (Theorem 4.4). Thus, our preprocessing framework for debiasing data comes with a provably fast algorithm.

Empirically, we evaluate the fairness and accuracy of the distributions generated by applying our framework to the Adult and COMPAS datasets, with gender as the protected attribute. Unlike prior work, the distributions obtained using the above parameters perform well for *both* representational and outcome-dependent fairness metrics. We further show that classifiers trained on samples from our distributions achieve high fairness (as measured by the classifier's statistical rate) with minimal loss to accuracy. Both with regard to the learned distributions and the classifiers trained on the de-biased data, our approach either matches or surpasses the performance of other state-of-the-art approaches across both fairness and accuracy metrics. Further, it is efficient on datasets with large domains (e.g., approx 10^{11} for the large COMPAS dataset), for which some other approaches are infeasible with regard to runtime.

Related work. Prior work on this problem falls, roughly, into two categories: 1) those that try to modify the dataset either by reassigning the protected attributes or reweighting the existing datapoints (Calders et al., 2009; Kamiran & Calders, 2012; Wang et al., 2019; King & Zeng, 2001), or 2) those that try to learn a distribution satisfying given constraints defined by the target fairness metric on the entire domain (Calmon et al., 2017).

The first set of methods often leads to efficient algorithms, but are unable to generate points from the domain that are not in the given dataset; hence, the classifiers trained on the re-weighted dataset may not generalize well (Chawla, 2009). Unlike the re-labeling/re-weighting approach of (Calders et al., 2009; Kamiran & Calders, 2009; 2012; King & Zeng, 2001) or the repair methods of (Gordaliza et al., 2019; Wang et al., 2019; Feldman et al., 2015; Zemel et al., 2013), we instead aim to learn a debiased version of the underlying distribution of the dataset across the entire domain. The second approach also aims to learn a debiased distribution on the entire domain. E.g., (Calmon et al., 2017) presents an optimization-based approach to learning a distribution that is close to the empirical distribution induced by the samples subject to fairness constraints. However, as their optimization problem has a variable for each point in the domain, the running time of their algorithm is at least the size of the domain, which is exponential in the dimension of the data, and hence often infeasible for large datasets. Since the max-entropy distribution can be efficiently represented using the dual parameters, our framework does not suffer from the enumeration problem of (Calders et al., 2009) and the inefficiency for large domains as in (Calmon et al., 2017). See Table 1 for a summary of the properties of our framework with key related prior work. Other preprocessing methods include selecting a subset of data that satisfies specified fairness constraints such as representation rate without attempting to model the distribution (Celis et al., 2016; 2018).

GAN-based approaches towards mitigating bias (Mariani et al., 2018; Sattigeri et al., 2019; Xu et al., 2018) are inherently designed to simulate continuous distributions and are neither optimized for discrete domains that we consider in this paper nor are prevalently used for social data and bench-

mark datasets for fairness in ML. While (Choi et al., 2017; Xu et al., 2018) suggest methods to round the final samples to the discrete domain, it is not clear whether such rounding procedures preserve the distribution for larger domains.

While our framework is based on preprocessing the dataset, bias in downstream classification tasks can also be addressed by modifying the classifier itself. Prior work in this direction fall into two categories: inprocessing methods that change the objective function optimized during training to include fairness constraints (Celis et al., 2019; Zhang et al., 2018), and post-processing methods that modify the outcome of the existing machine learning models by changing the decision boundary (Kamiran et al., 2012; Hardt et al., 2016).

2. Preliminaries

Dataset & Domain. We consider data from a discrete domain $\Omega := \Omega_1 \times \cdots \times \Omega_d = \{0,1\}^d$, i.e., each attribute Ω_i is binary.[1] The convex hull of Ω is denoted by $\mathrm{conv}(\Omega) = [0,1]^d$ and the size of the domain Ω is 2^d, i.e., exponential in the dimension d. We let the set (not multiset) $\mathcal{S} \subseteq \Omega$, along with a frequency $n_\alpha \geq 1$ for each point $\alpha \in \mathcal{S}$, denote a dataset consisting of $N = \sum_{\alpha \in \mathcal{S}} n_\alpha$ distinct points. We consider the attributes of Ω, indexed by the set $[d] := \{1, \ldots, d\}$, as partitioned into three index sets where 1) I_z denotes the indices of protected attributes, 2) I_y denotes the set of outcomes or class labels considered for fairness metric evaluation, and 3) I_x denotes the remaining attributes. We denote the corresponding sub-domains by $\mathcal{X} := \times_{i \in I_x} \Omega_i$, $\mathcal{Y} := \times_{i \in I_y} \Omega_i$, and $\mathcal{Z} := \times_{i \in I_z} \Omega_i$.

Fairness metrics. We consider the following two common fairness metrics; the first is "representational" (also known as "outcome independent") and depends only on the protected attributes and not on the class label, and the second one is an "outcome dependent" and depends on both the protected attribute and the class label.

Definition 2.1 (Representation rate). *For $\tau \in (0,1]$, a distribution $p : \Omega \to [0,1]$ is said to have representation rate τ with respect to a protected attribute $\ell \in I_z$ if for all $z_i, z_j \in \Omega_\ell$, we have*

$$\frac{p[Z = z_i]}{p[Z = z_j]} \geq \tau,$$

where Z is distributed according to the marginal of p restricted to Ω_ℓ.

[1] Our results can be extended to domains with discrete or categorical attributes by encoding an attribute of size k as binary using one-hot encodings: i.e., replace the cell with $e \in \{0,1\}^k$ where for a value $j \in [k]$ we set $e = \{e_1, \ldots, e_k\}$ with $e_j = 1$ and $e_\ell = 0$ for all $\ell \neq k$. To handle continuous features, one can apply discretization to reduce a continuous feature to a non-binary discrete feature. However, there is a natural tradeoff between domain size and correctness. We refer the reader to the survey (Kotsiantis & Kanellopoulos, 2006) for research on discretization techniques.

Definition 2.2 (Statistical rate). *For $\tau \in (0,1]$, a distribution $p : \Omega \to [0,1]$ is said to have statistical rate τ with respect to a protected attribute $\ell \in I_z$ and a class label $y \in \mathcal{Y}$ if for all $z_i, z_j \in \Omega_\ell$, we have*

$$\frac{p[Y = y \mid Z = z_i]}{p[Y = y \mid Z = z_j]} \geq \tau,$$

where Y is the random variable when p is restricted to $\mathcal{Y}$ and Z when p is restricted to Ω_ℓ.

We also refer to the statistical rate when the outcome labels are instead obtained using a classifier $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$. The classifier is said to have statistical rate τ if for all $z_i, z_j \in \Omega_\ell$, we have $\frac{\mathbb{P}[f(\alpha)=y | Z=z_i]}{\mathbb{P}[f(\alpha)=y | Z=z_j]} \geq \tau$, where the probability is over the empirical distribution of the test data.

In the definitions above, $\tau = 1$ can be thought of as "perfect" fairness and is referred to as representation parity and statistical parity respectively. In practice, however, these perfect measures of fairness are often relaxed: a popular example is the "80% rule" in US labor law (Biddle, 2006) to address disparate impact in employment, which corresponds to $\tau = 0.8$. The exact value of τ desired is context-dependent and will vary by application and domain.

The reweighting approach to debiasing data. A weight $w(\alpha)$ is assigned to each data point $\alpha \in \mathcal{S}$ such that $w(\alpha) \geq 0$, and $\sum_{\alpha \in \mathcal{S}} w(\alpha) = 1$. I.e., a probability distribution over samples is computed. These weights are carefully chosen in order to satisfy the desired fairness metrics, such as statistical parity (Kamiran & Calders, 2012) or representation parity (King & Zeng, 2001).

The optimization approach to debiasing data. The goal of learning a debiased probability distribution over the entire domain is formulated as a constrained optimization problem over the space $\mathcal{P}$ of all probability distributions over Ω (and not just $\mathcal{S}$). A prior distribution q is chosen that is usually supported on $\mathcal{S}$, a distance measure D is chosen to compare two probability distributions, and a function $J : \mathcal{P} \to \mathbb{R}^s$ that encodes the fairness criteria on the distribution is given. The goal is to find the solution to the following optimization problem: $\min_{p \in \mathcal{P}} D(p,q)$ s.t. $J(p) = 0$. For instance, (Calmon et al., 2017) use the total variation (TV) distance as the distance function and encode the fairness criteria as a linear constraint on the distribution.

The maximum entropy framework. Given $\Omega \subseteq \mathbb{R}^d$, a *prior* distribution $q : \Omega \to [0,1]$ and a *marginal* vector $\theta \in \mathrm{conv}(\Omega)$, the maximum entropy distribution $p^\star : \Omega \to [0,1]$ is the maximizer of the following convex program,

$$\sup_{p \in \mathbb{R}^{|\Omega|}_{\geq 0}} \sum_{\alpha \in \Omega} p(\alpha) \log \frac{q(\alpha)}{p(\alpha)}, \qquad \text{(primal-MaxEnt)}$$

$$\text{s.t.} \sum_{\alpha \in \Omega} \alpha p(\alpha) = \theta \quad \text{and} \quad \sum_{\alpha \in \Omega} p(\alpha) = 1.$$

The objective can be viewed as minimizing the KL-divergence with respect to the prior q. To make this program well defined, if $q(\alpha) = 0$, one has to restrict $p(\alpha) = 0$ and define $\log \frac{0}{0} = 1$. The maximum entropy framework is traditionally used to learn a distribution over Ω by setting $\theta := \frac{1}{N} \sum_{\alpha \in \mathcal{S}} \alpha \cdot n_\alpha$ and q to be the uniform distribution over Ω. This maximizes entropy while satisfying the constraint that the marginal is the same as the empirical marginal. It is supported over the entire domain Ω (as q is also supported on all of Ω) and, as argued in the literature (Dudik, 2007; Singh & Vishnoi, 2014), is information-theoretically the "least constraining" choice on the distribution that can explain the statistics of $\mathcal{S}$. Later we consider other choices for q that take $\mathcal{S}$ and our fairness goals into account and are also supported over the entire domain Ω.

Computationally, the number of variables in (primal-MaxEnt) is equal to the size of the domain and, hence does not seem scalable. However, a key property of this optimization problem is that it suffices to solve the dual (see below) that only has d variables (i.e., the dimension of the domain and not the size of the domain):

$$\inf_{\lambda \in \mathbb{R}^d} h_{\theta,q}(\lambda) := \log \left(\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \alpha - \theta, \lambda \rangle} \right), \tag{dual-MaxEnt}$$

where the function $h_{\theta,q} : \mathbb{R}^d \to \mathbb{R}$ is referred to as the dual max-entropy objective. For the objectives of the primal and dual to be equal (i.e., for strong duality to hold), one needs that θ lie in the "relative interior" of $\mathrm{conv}(\Omega)$; see (Singh & Vishnoi, 2014). In the case $\mathrm{conv}(\Omega) = [0,1]^d$, this simply means that $0 < \theta_i < 1$ for all $1 \le i \le d$. This is satisfied if for each attribute Ω_i there is at least one point in the set $\mathcal{S}$ that takes value 0 and at least one point that takes value 1.

Strong duality also implies that, if $\lambda^\star$ is a minimizer of $h_{\theta,q}$, then $p^\star$ can be computed as

$$p^\star(\alpha) = \frac{q(\alpha) e^{\langle \lambda^\star, \alpha \rangle}}{\sum_{\beta \in \Omega} q(\beta) e^{\langle \lambda^\star, \beta \rangle}};$$

see (Dudik, 2007; Singh & Vishnoi, 2014). Thus, the distribution $p^\star$ can be represented only using d numbers $\lambda_i^\star$ for $1 \le i \le d$. However, note that as some θ_i go close to an integral value or some $q(\alpha) \to 0$, these optimal dual variables might tend to infinity. Further, given a λ, computing $h_{\theta,q}$ requires computing a summation over the entire domain Ω – even in the simplest setting when q is the uniform distribution on Ω – that can a priori take time proportional to $|\Omega| = 2^d$. Hence, even though the dual optimization problem is convex and has a small number of variables (d), to obtain a polynomial (in d) time algorithm to solve it, we need both an algorithm that evaluate the dual function $h_{\theta,q}$ (a summation over the entire domain Ω) and its gradient efficiently at a given point λ, and (roughly) a bound on $\|\lambda^\star\|_2$ that is polynomial in d.

3. Our framework

Our approach for preprocessing data uses the maximum entropy framework and combines both the reweighting and optimization approaches. Recall that the maximum entropy framework requires the specification of the marginal vector θ and a prior distribution q. We use q and θ to enforce our goals of controlling representation and statistical rates as defined in Definitions 2.1 and 2.2, while at the same time ensuring that the learned distribution has support all of Ω and is efficiently computable in the dimension of Ω. Another advantage of computing the max-entropy distribution (as opposed to simply using the prior q) is that it pushes the prior towards the empirical distribution of the raw dataset, while maintaining the fairness properties of the prior. This leads to a distribution which is close to the empirical distribution and has fairness guarantees.

Prior distributions. Let u denote the uniform distribution on Ω: $u(\alpha) := \frac{1}{|\Omega|}$ for all $\alpha \in \Omega$. Note that the uniform distribution satisfies statistical rate with $\tau = 1$. We also use a reweighting algorithm (Algorithm 1) to compute a distribution w supported on $\mathcal{S}$. Our algorithm is inspired by the work of (Kamiran & Calders, 2012) and, for any given $\tau \in (0,1]$, Algorithm 1 can ensure that w satisfies the τ-statistical rate property; see Theorem 4.1. We introduce a parameter $C \in [0,1]$ that allows us to interpolate between w and u and define:

$$q_C^w := C \cdot u + (1 - C) \cdot w. \tag{1}$$

A desirable property of q_C^w, that we show is true, is that the dual objective function h_{θ,q_C^w} and its gradient are computable in time polynomial in N, d and the number of bits needed to represent θ for any weight vector w supported on $\mathcal{S}$; see Lemma 4.3. Further, we show that, if w has τ-statistical rate, then for any $C \in [0,1]$, the distribution q_C^w also has τ-statistical rate; see Theorem 4.1.

Thus, the family of priors we consider present no computational bottleneck over exponential-sized domains. Moreover, by choosing the parameter C, our framework allows the user to control how close they would like the learned distribution to be to the empirical distribution induced by $\mathcal{S}$. Finally, using appropriate weights w which encode the desired statistical rate, one can aim to ensure that the optimal distribution to the max-entropy program is also close to satisfying statistical parity (Theorem 4.5).

Marginal vectors. The simplest choice for the marginal vector θ is the marginal of the empirical distribution $\frac{1}{N} \sum_{\alpha \in \mathcal{S}} n_\alpha \cdot \alpha$. However, in our framework, the user can select any vector θ. In particular, to control the representation rate of the learned distribution with respect to a protected attribute ℓ, we can choose to set it differently. For instance, if $\Omega_\ell = \{0,1\}$ and we would like that in learned distribution the probability of this attribute being 1 is 0.5, it

Algorithm 1 Re-weighting algorithm to assign weights to samples for the prior distribution

Input: Dataset $\mathcal{S} := \{(X_\alpha, Y_\alpha, Z_\alpha)\}_{\alpha \in \mathcal{S}} \subseteq \mathcal{X} \times \mathcal{Y} \times \Omega_\ell$, frequency list $\{n_\alpha\}_{\alpha \in \mathcal{S}}$ and parameter $\tau \in (0, 1]$
for $y \in \mathcal{Y}$ **do**
  $c(y) \leftarrow \sum_{\alpha \in \mathcal{S}} \mathbf{1}(Y_\alpha = y) \cdot n_\alpha$
  $c(y, 0) \leftarrow \frac{1}{\tau} \cdot \sum_{\alpha \in \mathcal{S}} \mathbf{1}(Y_\alpha = y, Z_\alpha = 0) \cdot n_\alpha$
  $c(y, 1) \leftarrow \sum_{\alpha \in \mathcal{S}} \mathbf{1}(Y_\alpha = y, Z_\alpha = 1) \cdot n_\alpha$
end for
$w \leftarrow \mathbf{0}$
for $\alpha \in \mathcal{S}$ **do**
  $w(\alpha) \leftarrow n_\alpha \cdot {}^{c(Y_\alpha)}/_{c(Y_\alpha, Z_\alpha)}$
end for
$W \leftarrow \sum_{\alpha \in S} w(\alpha)$
return $\{{}^{w(\alpha)}/_{W}\}_{\alpha \in S}$

suffices to set $\theta_\ell = 0.5$. This follows immediately from the constraint imposed in the max-entropy framework. Once we fix a choice of θ and q, we need to solve the dual of the max-entropy program and we discuss this in the next section. The dual optimal $\lambda^\star$ can then be used to sample from the distribution $p^\star$ in a standard manner; see Section B in the supplementary material.

4. Theoretical results

Throughout this section we assume that we are given $C \in [0, 1]$, $\mathcal{S} \subseteq \Omega$ and the frequency of elements in $\mathcal{S}$, $\{n_\alpha\}_{\alpha \in \mathcal{S}}$.

The reweighting algorithm and its properties. We start by showing that there is an efficient algorithm to compute the weights w discussed in the previous section.

Theorem 4.1 (Guarantees on the reweighting algorithm). *Given the dataset $\mathcal{S}$, frequencies $\{n_\alpha\}_{\alpha \in \mathcal{S}}$ and a $\tau \in [0, 1]$, Algorithm 1 outputs a probability distribution $w : \mathcal{S} \to [0, 1]$ such that*

1. *The algorithm runs in time linear in N.*
2. *q_C^w, defined in Eq. (1) using w, satisfies τ-statistical rate, i.e, for any $y \in \mathcal{Y}$ and for all $z_1, z_2 \in \Omega_\ell$,*

$$\frac{q_C^w(Y = y \mid Z = z_1)}{q_C^w(Y = y \mid Z = z_2)} \geq \tau.$$

The proof of this theorem uses the fact that q_C^w is a convex combination of uniform distribution, which has statistical rate 1, and weights from Algorithm 1, which by construction satisfy statistical rate τ; it is presented in Section A in the supplementary material.

Computability of maximum entropy distributions. Since the prior distribution q_C^w is not uniform in general, the optimal distribution $p^\star$ is not a product distribution. Thus, as noted earlier, the number of variables in (primal-MaxEnt) is $|\Omega| = 2^d$, i.e., exponential in d, and standard methods from convex programming to directly solve primal-MaxEnt do not lead to efficient algorithms. Instead, we focus on computing (dual-MaxEnt). Towards this, we appeal to the general algorithmic framework of (Singh & Vishnoi, 2014; Straszak & Vishnoi, 2019). To use their framework, we need to provide (1) a bound on $\|\lambda^\star\|_2$ and (2) an efficient algorithm (polynomial in d) to evaluate the dual objective $h_{\theta,q}$ and its gradient. Towards (1), we prove the following.

Lemma 4.2 (Bound on the optimal dual solution). *Suppose θ is such that there is an $\eta > 0$ for which we have $\eta < \theta_i < 1 - \eta$ for all $i \in [d]$. Then, the optimal dual solution corresponding to such a θ and q_C^w satisfies*

$$\|\lambda^\star\|_2 \leq \frac{d}{\eta} \log \frac{1}{C}.$$

The proof uses a result from (Singh & Vishnoi, 2014) and appears in Section B in the supplementary material. We note that, for our applications, we can show that the assumption on θ follows from an assumption on the "non-redundancy" of the data set. Using recent results of (Straszak & Vishnoi, 2019), we can get around this assumption and we omit the details from this version of the paper.

Towards (2), we show that q_C^w has the property that not only can one evaluate h_{θ, q_C^w}, but also its gradient (and Hessian).

Lemma 4.3 (Oracles for the dual objective function). *There is an algorithm that, given a reweighted distribution $w : \mathcal{S} \to (0, 1]$, values $\theta, \lambda \in \mathbb{R}^d$, and distribution $q = q_C^w$, computes $h_{\theta,q}(\lambda)$, $\nabla h_{\theta,q}(\lambda)$, and $\nabla^2 h_{\theta,q}(\lambda)$ in time polynomial in N, d and the bit complexities of all the numbers involved: $w(\alpha)$ for $\alpha \in \mathcal{S}$, and e^{λ_i}, θ_i for $1 \leq i \leq d$.*

The proof of this lemma, along with the algorithm, appears in Section B.3 in the supplementary material. It uses the fact that q_C^w is a convex combination of uniform distribution (for which efficient oracles can be constructed) and a weighted distribution supported only on $\mathcal{S}$, and can be generalized to any prior q that similarly satisfies these properties.

Thus, as a direct corollary to Theorem 2.8 in the arxiv version of (Singh & Vishnoi, 2014) we obtain the following.

Theorem 4.4 (Efficient algorithm for max-entropy distributions). *There is an algorithm that, given a reweighted distribution $w : \mathcal{S} \to [0, 1]$, a $\theta \in [\eta, 1 - \eta]^d$, and an $\varepsilon > 0$, computes a λ° such that*

$$h_{\theta,q}(\lambda^\circ) \leq h_{\theta,q}(\lambda^\star) + \varepsilon.$$

Here $\lambda^\star$ is an optimal solution to the dual of the max-entropy convex program for $q := q_C^w$ and θ. The running time of the algorithm is polynomial in d, $\frac{1}{\eta}$, $\frac{1}{\varepsilon}$ and the number of bits needed to represent θ and w.

Fairness guarantees. Given a marginal vector θ that has representation rate τ, we can bound the statistical rate and representation rate of the the max-entropy distribution obtained using q_C^w and θ.

Theorem 4.5 (Fairness guarantees). *Given the dataset $\mathcal{S}$, protected attribute $\ell \in I_z$, class label $y \in \mathcal{Y}$ and parameters $\tau, C \in [0,1]$, let $w : S \to [0,1]$ be the reweighted distribution obtained from Algorithm 1. Suppose θ is a vector that satisfies $\frac{1}{2} \leq \theta_\ell \leq \frac{1}{1+\tau}$. The max-entropy distribution $p^\star$ corresponding to the prior distribution q_C^w and expected value θ has statistical rate at least τ' with respect to ℓ and y, where $\tau' = \tau - \frac{4\delta\cdot(1+\tau)}{C+4\delta}$, and $\delta = \max_{z\in\Omega_\ell} |p^\star(Y = y, Z = z) - q_C^w(Y = y, Z = z)|$; here Y is the random variable when the distribution is restricted to $\mathcal{Y}$ and Z is the random variable when the distribution is restricted to Ω_ℓ.*

The condition on θ, when simplified, implies that $(1-\theta_\ell)/\theta_\ell \geq \tau$ and $\theta_\ell/(1-\theta_\ell) \geq 1$, i.e., the marginal probability of $Z = 0$ is atleast τ times the marginal probability of $Z = 1$. This directly implies that the representation rate of $p^\star$ is at least τ. As we control the statistical rate using the prior q_C^w, the statistical rate of $p^\star$ depends on the distance between q_C^w and $p^\star$. The proof of Theorem 4.5 is provided in Section C in the supplementary material.

Remark 4.6. *Two natural choices for θ that satisfy the conditions of Theorem 4.5 are the following:*

1. *The reweighted vector $\theta^w := \sum_{\alpha\in\mathcal{S}} w(\alpha)\cdot\alpha$, where w is the weight distribution obtained using Algorithm 1; since w has representation rate τ, it can be seen that $\theta_\ell^w = 1/(1+\tau)$.*
2. *The vector θ^b that is the mean of the dataset $\mathcal{S}$ for all non-protected attributes and class labels, and is balanced across the values of any protected attribute. I.e.,*

$$\theta^b := \left(\sum_{\alpha\in\mathcal{S}} \frac{n_\alpha}{N} X_\alpha, \sum_{\alpha\in\mathcal{S}} \frac{n_\alpha}{N} Y_\alpha, \frac{1}{2}\right).$$

5. Empirical analysis

Our approach, as described above, is flexible and can be used for a variety of applications. [2] In this section we show its efficacy as compared with other state-of-the-art data debiasing approaches, in particular reweighting methods by (Kamiran & Calders, 2012; King & Zeng, 2001) and an optimization method by (Calmon et al., 2017). We consider two applications and three different domain sizes: The **COMPAS** criminal defense dataset using two versions of the data with differently sized domains, and the **Adult** financial dataset. With regard to fairness, we compare the statistical rate and representation rate of the de-biased datasets as well as the statistical rate of a classifier trained on the de-biased data. With regard to accuracy, we report both the divergence of the de-biased dataset from the raw data, as well as the resulting classifier accuracy. We find that our methods perform at least as well as if not better than existing approaches across all fairness metrics; in particular, ours are the only approaches that can attain a good representation rate while, simultaneously, attaining good statistical rate both with regard to the data and the classifier. Further, the loss as compared to the classifier accuracy when trained on raw data is minimal, even when the KL divergence between our distribution and the empirical distribution is large as compared to other methods. Finally, we report the runtime of finding the de-biased distributions, and find that our method scales well even for large domains of size $\sim 10^{11}$.

5.1. Setup for empirical analysis

Datasets. We consider two benchmark datasets from the fairness in machine learning literature.[3]

(a) The **COMPAS** dataset (Angwin et al., 2016; Larson et al., 2016) contains information on criminal defendants at the time of trial (including criminal history, age, sex, and race), along with post-trail instances of recidivism (coded as any kind of re-arrest). We use two versions of this dataset: the **small** version has a domain of size 144, and contains sex, race, age, priors count, and charge degree as features, and uses a binary marker of recidivism within two years as the label. We separately consider race (preprocessed as binary with values "Caucasian" vs "Not-Caucasian") and gender (which is coded as binary) as protected attributes. The **large** dataset has a domain of size approximately 1.4×10^{11} and consists of 19 attributes, 6 different racial categories and additional features such as the type of prior and juvenile prior counts.

(b) The **Adult** dataset (Dheeru & Karra Taniskidou, 2017) contains demographic information of individuals along with a binary label of whether their annual income is greater than \$50k, and has a domain of size 504. The demographic attributes include race, sex, age and years of education. We take gender (which is coded as binary) as the protected attribute.

Using our approach. We consider the prior distribution q_C^w, which assigns weights returned by Algorithm 1 for input $\mathcal{S}$ and $\tau = 1$ and $C = 0.5$.[4] Further, we consider the two different choices for the expectation vector as defined in Remark 4.6, namely: (1) The weighted mean of the samples θ^w using the weights w as obtained from Algorithm 1, and (2) the empirical expectation vector with the marginal of the protected attribute modified to ensure equal representation of both groups θ^b. In this case, since the protected attribute

[2] The code for our framework is available at https://github.com/vijaykeswani/Fair-Max-Entropy-Distributions.

[3] The details of both datasets, including a description of features are presented in Sections D and E of the supplementary material.

[4] This choice for C is arbitrary; we evaluate performance as a function of C in Section D of the supplementary material.

is binary we set $\theta^b_\ell = 1/2$. [5]

Baselines and metrics. We compare against the raw data, simply taking the prior q^w_C defined above, a reweighting method (Kamiran & Calders, 2012) for statistical parity, a reweighting method (King & Zeng, 2001) for representation parity, and an optimized preprocessing method (Calmon et al., 2017). We consider the distributions themselves in addition to classifiers trained on simulated datasets drawn from these distributions, and evaluate them with respect to well-studied metrics of fairness and accuracy.

For fairness metrics, we report the statistical rate (see Definition 2.2), i.e., the ratio between the probability of observing a favorable outcome given unprivileged group membership and the probability of observing a favorable outcome given privileged group membership. Note that this can be evaluated both with regard to the instantiation of the outcome variable in the simulated data, and with regard to the outcome predicted by the classifier; we report both. We also report the representation rate (see Definition 2.1) of the simulated data; for gender this corresponds to the ratio between fraction of women and men in the simulated datasets, while for race this corresponds to the ratio between fraction of Caucasian and Non-Caucasian individuals in the simulated datasets. For all fairness metrics, larger values, closer to 1, are considered to be "more fair".

We report the classifier accuracy when trained on the synthetic data. Further, we aim to capture the distance between the de-biased distribution and the distribution induced by the empirical samples. For the Adult dataset and small COMPAS dataset we report the KL-divergence.[6] For the large COMPAS dataset, the KL-divergence is not appropriate as most of the domain is not represented in the data. We instead consider the covariance matrix of the output dataset and the raw dataset and report the Frobenius norm of the difference of these matrices. In either case, lower values suggest the synthetic data better resembles the original dataset. Lastly, we report the runtime (in seconds) of each approach.

Implementation details. We perform 5-fold cross-validation for every dataset, i.e., we divide each dataset into five partitions. First, we select and combine four partitions into a training dataset and use this dataset to construct the distributions. Then we sample 10,000 elements from each distribution and train the classifier on this simulated dataset. We then evaluate our metrics on this simulated dataset and classifier (where the classifier accuracy and statistical rate is measured over the test set, i.e., the fifth partition of the original dataset). This sampling process is repeated 100 times for each distribution. We repeat this process 5 times for each dataset, once for each fold. We report the mean across all (500) repetitions and folds. Within each fold, the standard error across repetitions is low, less than 0.01 for all datasets and methods. Hence, for each fold, we compute the mean of metrics across the 100 repetitions and then report the standard deviation of this quantity across folds.

We use a decision tree classifier with gini information criterion as the splitting rule. A Gaussian naive Bayes classifier gives similar results. Further details are presented in Section D of the supplementary material. In the computation of the max-entropy distribution, we use a second-order algorithm inspired from works of (Allen Zhu et al., 2017; Cohen et al., 2017) that is also provably polynomial time in the parameters above and turns out to be slightly faster in practice. We present the details in Section F of the supplementary material. The machine specifications are a 1.8Ghz Intel Core i5 processor with 8GB memory.

5.2. Empirical results

The empirical results comparing our max-entropy approach against the state-of-the-art are reported in Table 2. The performance of using just the prior q^w_C is also reported in the table. For all datasets, the statistical rate of our max-entropy distributions is at least 0.97, which is higher than that of the raw data and higher or comparable to other approaches, including those specifically designed to optimize statistical parity (Calmon et al., 2017; Kamiran & Calders, 2012). Additionally, the representation rate of our max-entropy distributions is at least 0.97, which is higher than that of the raw data and higher or similar to other approaches, including those specifically designed to optimize the representation rate (King & Zeng, 2001). Recall that both fairness metrics can be at most 1, so this suggests the synthetic data our distributions produce have a near-equal fraction of individuals from both groups of protected attribute values (women/men or Caucasian/Not-Caucasian) *and* the probability of observing a favorable outcome is almost equally likely for individuals from both groups.

Note that Theorem 4.5 gives a bound on the statistical rate τ'. While this bound can be strong, the statistical rates we observe empirically are even better. E.g., for the small COMPAS dataset with gender as the protected attribute, by plugging in the value of δ for prior q^w_C and expected vector θ^w, we get that $\tau' = 0.85$ (i.e., satisfying the 80% rule), but we observe that empirically it is even higher (0.98). However, the bound may not always be strong. E.g., or the Adult dataset, we only get $\tau' = 0.23$. In this case, the distance between the prior q^w_C and max-entropy distribution $p^\star$ is large hence the bound on the statistical rate of $p^\star$, derived using q^w_C, is less accurate. Still, the statistical rate of

[5] In Section D of the supplementary material we evaluate the performance using alternate priors and expectation vectors such as q^d_C and θ_{d} which correspond to the raw data.

[6] For this to be well-defined, if a point does not appear in the dataset, before calculating KL-divergence, we assign it a very small non-zero probability ($\sim 10^{-7}$).

Table 2. **Empirical results.** Our max-entropy distributions use prior q_C^w for $C = 0.5$ and expected value θ^w or θ^b (as defined in Remark 4.6). "SR" denotes statistical rate. We report the mean across all folds and repetitions, with the standard deviation across folds in parentheses. For each measurement and dataset, the results that are not statistically distinguishable at p-value $= 0.05$ from the best result across all baselines and approaches are given in bold. Note that the approach is infeasible for larger domains, such as the large version of COMPAS datasets, and hence we do not present the results of (Calmon et al., 2017) on that dataset. The results in this table are represented graphically in Figure 7 in the Supplementary File.

				This paper			Baselines		
			Raw Data	Prior q_C^w	Max-Entropy with q_C^w, θ^w	Max-Entropy with q_C^w, θ^b	(Calmon et al., 2017)	(Kamiran & Calders, 2012)	(King & Zeng, 2001)
Adult	gender	Fairness	Data SR	0.36 (0)	**0.97** (0.02)	**0.98** (0.02)	**0.98** (0.02)	0.96 (0.01)	**0.97** (0.02)	0.36 (0)
			Representation Rate	0.49 (0)	**0.97** (0.01)	**0.97** (0.02)	**0.99** (0.01)	0.49 (0.01)	0.49 (0.01)	**0.98** (0)
			Classifier SR	0.36 (0)	**0.96** (0.03)	**0.95** (0.02)	**0.96** (0.01)	**0.97** (0.01)	0.85 (0.03)	0.36 (0)
		Accuracy	KL-divergence w.r.t raw data	**0** (0)	1.23 (0.03)	0.24 (0.01)	0.24 (0.01)	0.16 (0)	0.22 (0.01)	0.08 (0)
			Classifier Accuracy	**0.80** (0)	0.75 (0.01)	0.77 (0.02)	0.76 (0.01)	0.77 (0.01)	0.78 (0.01)	**0.80** (0)
			Runtime	-	0.73s	10s	10s	62s	0.16s	0.57s
COMPAS (small)	gender	Fairness	Data SR	0.73 (0.02)	**0.98** (0.01)	**0.98** (0.02)	**0.99** (0.01)	0.87 (0.02)	**0.98** (0.02)	0.73 (0.03)
			Representation Rate	0.24 (0.01)	**0.97** (0.02)	**0.98** (0.01)	**0.98** (0.02)	0.24 (0.01)	0.24 (0.01)	**0.98** (0)
			Classifier SR	0.72 (0.01)	**0.96** (0.02)	**0.95** (0.02)	**0.96** (0.02)	**0.93** (0.04)	**0.93** (0.03)	0.72 (0.01)
		Accuracy	KL-divergence w.r.t raw data	**0** (0)	0.57 (0.03)	0.35 (0.01)	0.37 (0.02)	0.02 (0)	0.14 (0.02)	0.24 (0)
			Classifier Accuracy	**0.66** (0.01)	**0.65** (0.01)	0.64 (0.01)	**0.65** (0.02)	**0.66** (0.01)	**0.66** (0.01)	**0.66** (0.01)
			Runtime	-	0.06s	2.5s	2.6s	25s	0.04s	0.10s
	race	Fairness	Data SR	0.76 (0.01)	**0.98** (0.01)	**0.98** (0.01)	**0.99** (0.01)	0.93 (0.01)	**0.98** (0.01)	0.76 (0.01)
			Representation Rate	0.66 (0.01)	**0.99** (0.01)	**0.99** (0.01)	**0.99** (0.01)	0.74 (0.02)	0.67 (0.02)	**0.99** (0)
			Classifier SR	0.75 (0.02)	**0.95** (0.03)	**0.96** (0.01)	**0.94** (0.03)	0.85 (0.09)	**0.96** (0.03)	0.75 (0.02)
		Accuracy	KL-divergence w.r.t raw data	**0** (0)	0.36 (0.02)	0.13 (0.01)	0.13 (0.01)	0.02 (0.01)	0.02 (0)	0.03 (0)
			Classifier Accuracy	**0.66** (0.01)	**0.64** (0.02)	**0.65** (0.02)	**0.65** (0.01)	0.58 (0.02)	**0.65** (0.01)	**0.66** (0.01)
			Runtime	-	0.06s	2.5s	2.6s	25s	0.04s	0.10s
COMPAS (large)	gender	Fairness	Data SR	0.71 (0.02)	0.97 (0.01)	**0.98** (0.01)	**0.97** (0.02)	-	**0.99** (0.01)	0.71 (0.02)
			Representation Rate	0.26 (0.01)	0.96 (0.01)	**0.98** (0.01)	**0.98** (0.01)	-	0.26 (0.01)	**0.98** (0)
			Classifier SR	0.73 (0.06)	**0.89** (0.02)	**0.88** (0.02)	**0.85** (0.06)	-	0.79 (0.01)	0.73 (0.03)
		Accuracy	Covariance matrix difference norm	**0** (0)	4.64 (0.26)	3.20 (0.44)	5.18 (0.84)	-	4.89 (0.04)	0.16 (0.01)
			Classifier Accuracy	**0.65** (0.01)	0.63 (0.01)	0.63 (0.01)	0.63 (0.01)	-	0.62 (0.02)	0.63 (0.01)
			Runtime	-	35s	40s	40s	-	0.25s	2s
	race	Fairness	Data SR	0.73 (0.03)	**0.98** (0.02)	**0.98** (0.02)	**0.97** (0.02)	-	**0.99** (0)	0.72 (0.03)
			Representation Rate	0.06 (0)	**0.99** (0.01)	**0.99** (0.01)	**0.99** (0.01)	-	0.01 (0.01)	**0.98** (0)
			Classifier SR	0.72 (0.01)	**0.89** (0.06)	**0.91** (0.06)	**0.91** (0.05)	-	**0.85** (0.11)	0.71 (0.13)
		Accuracy	Covariance matrix difference norm	**0.01** (0)	1.94 (0.25)	1.93 (0.24)	1.87 (0.26)	-	0.88 (0.14)	0.36 (0.01)
			Classifier Accuracy	**0.66** (0.01)	0.64 (0.01)	0.64 (0.01)	0.63 (0.01)	-	0.41 (0.08)	0.64 (0.01)
			Runtime	-	35s	40s	40s	-	0.25s	2s

max-entropy distribution is observed to be 0.97, suggesting that perhaps stronger fairness guarantees can be derived.

The statistical rate of the classifiers trained on the synthetic data generated by our max-entropy approach is comparable or better than that from other methods, and significantly better than the statistical rate of the classifier trained on the raw data. Hence, as desired, our approach leads to improved fairness in downstream applications. This is despite the fact that the KL-divergence of the max-entropy distributions from the empirical distribution on the dataset is high compared to most other approaches. Still, we note that the difference between the max-entropy distributions and the empirical distribution tends to be smaller than the difference between the prior q_C^w and the empirical distribution (as measured by KL divergence and the covariance matrix difference as discussed above). This suggests that, as expected, the max-entropy optimization helps push the re-weighted distribution towards the empirical distribution and highlights the benefit of using a hybrid approach of reweighting and optimization.

For the COMPAS datasets, the raw data has the highest accuracy and the average loss in accuracy when using the datasets generated from max-entropy distributions is at most 0.03. This is comparable to the loss in accuracy when using datasets from other baseline algorithms. In fact, for the small version of COMPAS dataset, the accuracy of the classifier trained on datasets from the max-entropy distribution using marginal θ^b is statistically similar to the accuracy of the classifier trained on the raw dataset. For the Adult dataset, (King & Zeng, 2001) achieves the same classifier accuracy as the raw dataset. As the Adult dataset is relatively more gender-balanced than COMPAS datasets and outcomes are not considered, (King & Zeng, 2001) do not need to modify the dataset significantly to achieve a high representation rate (indeed its KL-divergence from the empirical distribution of the raw data is the smallest). In comparison, all other methods that aim to satisfy statistical parity (max-entropy approach, (Calmon et al., 2017; Kamiran & Calders, 2012)) suffer a similar (but minimal) loss in accuracy of at most 0.03.

With respect to runtime, since (Kamiran & Calders, 2012), (King & Zeng, 2001) and prior q_C^w are simple re-weighting approaches and do not look at features other than class labels and protected attribute, it is not surprising that they have the best processing time. Amongst the generative models, the max-entropy optimization using our algorithm is significantly faster than the optimization framework of (Calmon et al., 2017). In fact, the algorithm of (Calmon et al., 2017) is infeasible for larger domains, such as the large COMPAS dataset, and hence we are not able present the results of their algorithm on that dataset.

6. Conclusion, limitations, and future work

We present a novel optimization framework that can be used as a data preprocessing method towards mitigating bias. It works by applying the maximum entropy framework to modified inputs (i.e., the expected vector and prior distribution) which are carefully designed to improve certain fairness metrics. Using this approach we can learn distributions over large domains, controllably adjust the representation rate or statistical rate of protected groups, yet remains close to the empirical distribution induced by the given dataset. Further, we show that we can compute the modified distribution in time polynomial in the *dimension* of the data. Empirically, we observe that samples from the learned distribution have desired representation rates and statistical rates, and when used for training a classifier incurs only a slight loss in accuracy while significantly improving its fairness.

Importantly, our pre-processing approach is also useful in settings where group information is not present at runtime or is legally prohibited from being used in classification (Edwards & Veale, 2017), and hence we only have access to protected group status it in the training set. Further, our method has an added privacy advantage of obscuring information about individuals in the original dataset, since the result of our algorithm is a distribution over all points in the domain rather than a reweighting of the actual dataset.

An important extension would be to modify our approach to improve fairness metrics across intersectional types. Given multiple protected attributes, one could pool them together to form a larger categorical protected attribute that captures intersectional groups, allowing our approach to be used directly. However, improving fairness metrics across multiple protected attributes *independently* seems to require additional ideas.

Achieving "fairness" in general is an imprecise and context-specific goal. The choice of fairness metric depends on the application, data, and impact on the stakeholders of the decisions made, and is beyond the scope of this work. However, our approach is not specific to statistical rate or representation rate and can be extended to other fairness metrics by appropriately selecting the prior distribution and expectation vector for our max-entropy framework.

Acknowledgements

This research was supported in part by NSF CCF-1908347 and an AWS MLRA Award. We thank Ozan Yildiz for initial discussions on algorithms for max-entropy optimization.

References

Allen Zhu, Z., Li, Y., Oliveira, R., and Wigderson, A. Much faster algorithms for matrix scaling. In *FOCS'17: Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, 2017.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. COMPAS recidivism risk score data and analysis, 2016. URL https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data\/--and-analysis.

Biddle, D. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

Calders, T. and Žliobaitė, I. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, pp. 43–57. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pp. 13–18. IEEE, 2009.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Celis, L. E., Deshpande, A., Kathuria, T., and Vishnoi, N. K. How to be fair and diverse? In *Fairness, Accountability, and Transparency in Machine Learning*, 2016.

Celis, L. E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., and Vishnoi, N. Fair and diverse DPP-based data summarization. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 716–725, Stockholmsmassan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 319–328. ACM, 2019.

Chawla, N. V. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pp. 875–886. Springer, 2009.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the Machine Learning for Health Care Conference, MLHC 2017, Boston, Massachusetts, USA, 18-19 August 2017*, pp. 286–305, 2017.

Cohen, M. B., Madry, A., Tsipras, D., and Vladu, A. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *FOCS'17: Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, 2017.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2017.

Dudik, M. Maximum entropy density estimation and modeling geographic distributions of species, 2007.

Edwards, L. and Veale, M. Slave to the algorithm? why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16:18, 2017.

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

Gibbs, J. W. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics*. C. Scribner's sons, 1902.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Jean-Michel, L. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pp. 2357–2365, 2019.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016.

Jaynes, E. T. Information theory and statistical mechanics. *Physical Review*, 106:620–630, May 1957a. doi: 10.1103/PhysRev.106.620.

Jaynes, E. T. Information theory and statistical mechanics. II. *Physical Review*, 108:171–190, October 1957b. doi: 10.1103/PhysRev.108.171.

Kamiran, F. and Calders, T. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication, 2009. IC4 2009.*, pp. 1–6. IEEE, 2009.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pp. 924–929, 2012. doi: 10.1109/ICDM.2012.45.

Kay, M., Matuszek, C., and Munson, S. A. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 3819–3828. ACM, 2015. ISBN 978-1-4503-3145-6.

King, G. and Zeng, L. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

Kotsiantis, S. and Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1): 47–58, 2006.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown/Archetype, 2016. ISBN 9780553418828.

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., and Varshney, K. R. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Singh, M. and Vishnoi, N. K. Entropy, optimization and counting. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 50–59. ACM, 2014.

Straszak, D. and Vishnoi, N. K. Maximum entropy distributions: Bit complexity and stability. In *Conference on Learning Theory*, pp. 2861–2891, 2019.

Wang, H., Ustun, B., and Calmon, F. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pp. 6618–6627, 2019.

Xu, D., Yuan, S., Zhang, L., and Wu, X. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

Zelaya, C. V. G. Towards explaining the effects of data preprocessing on machine learning. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 2086–2090. IEEE, 2019.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340. ACM, 2018.