

Fair Distributions from Biased Samples: A Maximum Entropy Optimization Framework

L. Elisa Celis¹, Vijay Keswani², Ozan Yildiz², and Nisheeth K. Vishnoi¹

¹Yale University

²École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

June 6, 2019

Abstract

One reason for the emergence of bias in AI systems is biased data – datasets that may not be true representations of the underlying distributions – and may over or under-represent groups with respect to protected attributes such as gender or race. We consider the problem of correcting such biases and learning distributions that are “fair”, with respect to measures such as proportional representation and statistical parity, from the given samples. Our approach is based on a novel formulation of the problem of learning a fair distribution as a maximum entropy optimization problem with a given expectation vector and a prior distribution. Technically, our main contributions are: (1) a new second-order method to compute the (dual of the) maximum entropy distribution over an exponentially-sized discrete domain that turns out to be faster than previous methods, and (2) methods to construct prior distributions and expectation vectors that provably guarantee that the learned distributions satisfy a wide class of fairness criteria. Our results also come with quantitative bounds on the total variation distance between the empirical distribution obtained from the samples and the learned fair distribution. Our experimental results include testing our approach on the COMPAS dataset and showing that the fair distributions not only improve disparate impact values but when used to train classifiers only incur a small loss of accuracy.

Contents

1	Introduction	3
2	Preliminaries	4
3	Our framework	6
4	Theoretical results	7
5	Empirical results	9
5.1	Results	11
6	Limitations and future work	12
7	Second order box-constrained Newton’s method	12
7.1	Bounding the number of iterations of Algorithm 1	12
7.2	Solving the inner optimization problem	15
8	Proof of second-order robustness: Proof of Lemma 4.2	15
9	Fast computation of max-entropy distribution: Proof of Theorem 4.1	17
9.1	Computability of the gradient and hessian of dual program	17
9.2	Runtime of primal path following algorithm	18
9.3	Proof of runtime of max-entropy distribution computation	20
10	Price of fairness	21
11	Fairness guarantees - Proof of Theorem 4.5	23
12	Using uniform weights in the prior distribution	25
13	Other experiments on smaller COMPAS dataset	26
14	Experiments on larger COMPAS dataset	27
14.1	Evaluating the disparate impact and accuracy of generated dataset	28
14.2	Evaluating the disparate impact and accuracy of classifier trained on generated dataset	29
14.3	Comparison with other algorithms	29

1 Introduction

Datasets that under or over represent certain populations can be an important source of bias and discrimination in ML and AI systems that use such datasets for training or end-use [20, 5, 17]. The class of pre-processing methods to control bias in AI systems aim to “correct” for these sampling biases across protected attributes such as gender and race. The general goal therein is to either ensure that the representation of protected populations in the datasets is consistent with the ground truth, or in the case of supervised learning (when the data points come with additional class labels) to ensure that the dataset satisfies statistical parity-like or group fairness conditions. A more general goal, and the focus of this paper, is to learn a distribution that does not suffer from these biases while being as “close” as possible to the distribution from which the given samples were drawn.

Prior approaches towards these goals include modification of protected attributes [14], weight assignment to samples [4], probabilistic transformation of the dataset [6], and new sample generation via generative adversarial networks (GAN) [19, 21, 24]. The approach of [4, 14] involves either removing sensitive attribute information or relabelling/re-weighting samples to ensure they are fair. While the resulting dataset satisfies the fairness constraints, this method cannot be used in scenarios where the classifier is insensitive to noise or does not accept weighted samples. [4] also suggest a randomized transformation of the samples based on the solution of an optimization problem they formulate to learn a fair distribution using the samples. Their method assumes the availability of the prior distribution or a large number of samples. It also involves enumeration of probabilities for the points in the domain and, hence, can be infeasible for larger domains. The GAN-based approaches [19, 21, 24] are inherently designed to simulate continuous distributions and are not optimized for discrete domains that we consider in this paper. While [8, 24] 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. A recent work of [6] presents an optimization-based approach to the problem of learning a distribution that is close to the empirical distribution induced by the samples subject to group and individual fairness constraints. While their approach is theoretically founded and general, it is not feasible for large domains as the running time of the distribution is at least the size of the domain, which can be exponential in the dimension of the data.

We propose a novel approach to learning a “fair” distribution that relies on the maximum entropy optimization framework. The maximum entropy framework is widely used to learn a probabilistic model of data from samples by finding the distribution over the domain that minimizes the KL-divergence with respect to a *prior* that is consistent with the statistics, such as the *expectation vector*, of the observed samples [10]. When the prior distribution is uniform, the distribution that minimizes KL-divergence maximizes entropy; hence the name. A remarkable feature of this framework is that by passing to the dual optimization problem, one can represent optimal distributions over exponentially-sized (in dimension of the data) domains using a small number of variables [22]. Leveraging on the recent theoretical developments on the computational aspects of this problem [22, 23] along with new algorithmic and structural results about maximum entropy distributions, we show that our framework addresses multiple problems faced in prior work – 1) it can generate unseen data points, 2) it comes with “knobs”, in the form of the expectation vector and the prior distribution, that allows one to control and provably guarantee the extent of representation and statistical parity fairness, 3) it supports discrete distributions, and 4) it comes with a new provably efficient second-order algorithm. We also provide quantitative bounds on the total variation distance between the empirical distribution obtained from the samples and the learned fair distribution (Section 10).

Unlike the re-labeling/re-weighting approach of [14], we do not modify the original dataset but rather try to learn the underlying distribution of the dataset with modified marginals of the

distribution to ensure fairness. Since the max-entropy distribution can be efficiently represented using the dual program, our framework does not suffer from the enumeration problem of [4] and the inefficiency for large domains as in [6]. Finally, we do not need to employ any kind of rounding procedures used in [8, 24] and are guaranteed that the final distribution is restricted to the domain we choose.

We evaluate the fairness and accuracy of the datasets generated using our framework on the COMPAS dataset, with gender as the protected attribute. To ensure high disparate impact of the output dataset, we use a prior that assigns weights in a particular way to samples in the dataset. The expectation vector input to the maximum entropy program is either the weighted mean of the dataset or the uniform mean with balanced sensitive attribute. The distribution obtained using the above parameters not only has high disparate impact values (close to 1), and high gender ratio (close to 1). We further show that classifiers trained on such simulated datasets are more fair with respect to disparate impact, and the loss in accuracy as compared to using the raw data is small. In fact, this approach outperforms the state-of-the-art data debiasing methods in [6] and [14].

Other related work. The literature on algorithmic fairness is vast and roughly divided into: preprocessing, in-processing and post-processing methods. Preprocessing methods discussed earlier modifies the data before it is used in any machine learning algorithm. Thus they can be used in a variety of different tasks with existing frameworks which make it the most versatile among the three approaches. The in-processing methods usually change the loss function minimized during training to include fairness constraints. These methods require modification of existing optimization problems and the algorithms used, but they directly enforce fairness in the resulting system. Some examples of these methods include [7, 25, 16]. The post-processing methods modify the outcome of the existing machine learning models by changing the decision boundary. These methods require minimal modification to existing models but their effectiveness can be limited with respect to other methods. Some examples of these methods include [11, 15]. While there is also a vast amount of literature on using maximum entropy models in machine learning (see [10, 22] and the references therein), to the best of our knowledge, our framework is the first usage of maximum entropy distributions for fair machine learning, and the first second order method to compute these distributions; see also [2, 9].

2 Preliminaries

Domain. We are given a dataset from a domain $\Omega := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ where \mathcal{X} denotes the domain of non-protected feature vectors, \mathcal{Y} denotes the domain of outcomes or class labels (considered for fairness metric evaluation), and \mathcal{Z} denotes the domain of protected attributes, such as gender or race. We focus on the case when each attribute is either binary or categorical with domain $\Omega_1, \dots, \Omega_d$. If binary, Ω_i is encoded as $\{0, 1\}$ and if categorical with k possible values, it can be either encoded as Δ_k for some $k \in \mathbb{Z}^+$, where Δ_k denotes the probability simplex used for one-hot encoding of categorical variables, or as the set of k values of the feature, $\{a_1, a_2, \dots, a_k\}$. For example, if Ω_i is a categorical attribute with three values, then the one-hot encoding is $\Omega_i = \{\{1, 0, 0\}, \{0, 1, 0\}, \{0, 0, 1\}\}$. Thus,

$$\Omega := \Omega_1 \times \dots \times \Omega_d. \quad (\text{Partition of } \Omega \text{ into individual attributes})$$

For a domain $\Omega = \Omega_1 \times \dots \times \Omega_d$, the convex hull of Ω , denoted by $\text{conv}(\Omega)$, is defined as $\text{conv}(\Omega_1) \times \dots \times \text{conv}(\Omega_d)$, where for each i , $\text{conv}(\Omega_i)$ is a probability simplex of dimension $|\Omega_i|$. Thus, $\text{conv}(\Omega) \subseteq \mathbb{R}^n$ where $n := \sum_{i=1}^d |\Omega_i|$. We also require partition of $[d]$ to three index sets I_x , I_y , and I_z where I_z denotes the indices of protected attributes, I_y denotes the indices of non-protected attributes with respect to which we compute fairness metrics such as recidivism result, and I_x denotes the

remaining non-protected attributes. We denote the corresponding dimensions with $\mathcal{X} := \times_{i \in I_x} \Omega_i$, $\mathcal{Y} := \times_{i \in I_y} \Omega_i$, and $\mathcal{Z} := \times_{i \in I_z} \Omega_i$, i.e.,

$$\Omega = \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}. \quad (\text{Partition of } \Omega \text{ into non-protected, output, and protected attributes})$$

Notions of fairness. To quantify how “fair” a distribution is, we use the following two notions, the first does not rely on a class label, but the second one does.

Definition 2.1 (Representation rate). For $\tau \in (0, 1]$, a distribution $p : \Omega \rightarrow [0, 1]$ is said to have representation rate τ w.r.t. a protected attribute $\ell \in I_z$ if $\forall z_i, z_j \in \Omega_\ell$, $\frac{p[Z=z_i]}{p[Z=z_j]} \geq \tau$, where Z is distributed according to the marginal of p restricted to Ω_ℓ .

Definition 2.2 (Statistical rate and disparate impact). For $\tau \in (0, 1]$, a distribution $p : \Omega \rightarrow [0, 1]$ is said to have statistical rate τ w.r.t. a protected attribute $\ell \in I_z$ and for a class label $y \in \mathcal{Y}$ if $\forall z_i, z_j \in \Omega_\ell$, $\frac{p[Y=y|Z=z_i]}{p[Y=y|Z=z_j]} \geq \tau$, where Y is the random variable when p is restricted to \mathcal{Y} and Z corresponds to the random variable when p is restricted to Ω_ℓ . If p has statistical rate 1 w.r.t. Ω_ℓ and y , then it is said to have no disparate impact w.r.t. Ω_ℓ and y .

Remark 2.3. In certain cases it might be desirable to have an upper and lower bound on the probabilities in the definitions above rather than requiring all of them be close to each other. For a distribution $p : \Omega \rightarrow [0, 1]$, a protected attribute $r \in I_z$, and numbers $l = \{l_z\}_{z \in \Omega_r}$ and $u = \{u_z\}_{z \in \Omega_r}$, say that:

1. p is (l, u) -representative for Ω_r , if for all $z \in \Omega_r$, $l_z \leq p[Z = z] \leq u_z$, and
2. p is (l, u) -statistical for Ω_r , if for any fixed class label $y \in \mathcal{Y}$ and for all $z \in \Omega_r$, $l_z \leq p[Y = y | Z = z] \leq u_z$.

Our framework extends to these two definitions, we omit the details from this version of the paper.

Maximum entropy distributions. Given $\Omega \subseteq \mathbb{R}^n$, a prior distribution $q : \Omega \rightarrow [0, 1]$ and a expectation or marginal vector $\theta \in \text{conv}(\Omega)$, the maximum entropy distribution $p^* : \Omega \rightarrow [0, 1]$ is the maximizer of the following concave program,

$$\sup_{p \in \mathbb{R}_{\geq 0}^{|\Omega|}} \sum_{\alpha \in \Omega} p(\alpha) \log \frac{q(\alpha)}{p(\alpha)}, \quad \text{s.t. } \sum_{\alpha \in \Omega} \alpha p(\alpha) = \theta, \text{ and } \sum_{\alpha \in \Omega} p(\alpha) = 1. \quad (\text{primal-MaxEnt})$$

The objective can also be viewed as minimizing the KL-divergence w.r.t. the prior q . The domain Ω is exponentially large in n , making (primal-MaxEnt) computationally infeasible to solve due to an equal number of variables. One observation is that the dual of (primal-MaxEnt) is convex, has only n variables, and it sufficient to solve it provided θ lies in the relative interior of $\text{conv}(\Omega)$ [22].

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

where the function $h_{\theta, q} : \mathbb{R}^n \rightarrow \mathbb{R}$ denotes the dual max entropy program. If λ^* is a global minimizer of $h_{\theta, q}$, then p^* can be computed as $p^*(\alpha) = \frac{q(\alpha) e^{\langle \lambda^*, \alpha \rangle}}{\sum_{\beta \in \Omega} q(\beta) e^{\langle \lambda^*, \beta \rangle}}$, see [10, 22].

3 Our framework

Entropy maximizing distributions. In our framework, the basic input consists of Ω_i for $i = 1, \dots, d$ and N samples $\mathcal{S} := \{\alpha_i\}_{i=1}^N$ from $\Omega = \Omega_1 \times \dots \times \Omega_d \in \mathbb{R}^n$. We use the following starting probabilistic model for the dataset: the entropy maximizing distribution p^* on Ω corresponding to a uniform prior ($q(\cdot) = 1$) and $\theta = \hat{\theta} := \frac{1}{N} \sum_{i=1}^N \alpha_i$. Such distributions are used widely as probabilistic models of data [10] and are inspired by the maximum-entropy principle [12, 13].

However, since we expect N to be much smaller than $|\Omega|$, it is desirable to have a way to move from a distribution that is supported only on \mathcal{S} to that on all of Ω in a “smooth” manner. Towards this, we introduce a parameter $C \in [0, 1]$ that results in a prior $q_C(\alpha) = \frac{C}{|\Omega|} + (1 - C) \sum_{i=1}^N \mathbb{1}(\alpha_i = \alpha)$, where $\mathbb{1}(\cdot)$ is the indicator function. Note that the smoothing parameter C controls the importance given to unseen points over the samples in \mathcal{S} ; if $C = 0$, the points in $\Omega \setminus \mathcal{S}$ are assigned zero probability mass and the corresponding distribution q_C assigns non-zero probability only to points in \mathcal{S} , while if $C = 1$, then all the points in Ω are assigned equal probability.

Representationally fair distributions. To ensure that the learned distribution has representation rate τ with respect to a protected attribute Ω_ℓ and a fixed $y \in \mathcal{Y}$, for a given τ, ℓ and y , we use a property of the entropy maximizing distribution corresponding to θ that $p^*(Z = z) = \theta(Z = z)$; see [22]. Thus, if it is true that $\hat{\theta}(Z = z_i)/\hat{\theta}(Z = z_j) \geq \tau$ for all $z_i, z_j \in \Omega_\ell$ (said to be τ -balanced), it follows that if we compute p^* using $\hat{\theta}$, $p^*(Z = z_i)/p^*(Z = z_j) \geq \tau$; see first bullet of Theorem 4.5. However, if $\hat{\theta}$ does not satisfy this property, then our framework allows one to input a θ that satisfies this property. Alternatively, we can set θ to be the closest point to $\hat{\theta}$ in $\text{conv}(\Omega)$ that is τ -balanced.

Statistically fair distributions. Ensuring that the learned distribution has statistical rate τ is more complicated. One of the issues is that the prior distribution q may be very far from having statistical rate τ . We first give a method to rescale the weights of the samples (w_i for α_i) in the prior distribution to ensure that the resulting distribution satisfies statistical parity with rate 1 (see Lemma 4.4):

$$q_C^w(\alpha) = \frac{C}{|\Omega|} + (1 - C) \sum_{i=1}^N \mathbb{1}(\alpha_i = \alpha) \cdot w_i. \quad (1)$$

Our main structural result is to show that the entropy maximizing distribution corresponding to this prior is approximately τ -statistically fair; see second bullet of Theorem 4.5

Computability of maximum entropy distributions. All of our approaches reduce the problem of learning “fair” distribution from (potentially biased) samples to that of solving primal-MaxEnt for an appropriate choice of q and θ . Since the prior distribution $q(\cdot)$ is not uniform in general, the optimal distribution p^* is not a product distribution. Thus, even in the simplest case when $\Omega_i = \{0, 1\}$ for each $i \in [d]$, the number of variables in primal-MaxEnt is $|\Omega| = 2^d$, exponential in d .

Thus, like previous works on the computability of maximum entropy distributions, we focus on the computability of dual-MaxEnt. Recent prior works have deployed the ellipsoid algorithm based framework whose running time depends 1) polynomially on the $\|\lambda^*\|_2$, and 2) an efficient algorithm (polynomial in d) to evaluate $h_{\theta, q}$. Both of these are non-trivial problems – (1) was resolved very recently in [23] and we can use their bound directly. Towards (2), we show that our prior distribution q_C^w has this nice property that not only one can evaluate $h_{\theta, q}$, but also its gradient and hessian (Lemma 4.3). Our main technical contribution is a second-order method to solve dual-MaxEnt that comes with provable guarantees (Theorem 4.1) and scales well in practice. The following summarizes our framework.

Input: - Domain $\Omega_1, \dots, \Omega_d$ along with indices of class labels I_y and protected attributes I_z ;
- Dataset $\mathcal{S} = \{\alpha_i\}_{i \in [N]} \subseteq \Omega$;
- Target expected vector $\theta \in \text{conv}(\Omega)$ that is τ -balanced;
- Smoothing parameter $0 \leq C \leq 1$.
Output: A solution λ^* to dual-MaxEnt corresponding to θ and $q = q_C$ when representational fairness is desired, and corresponding to θ and $q = q_C^w$ when statistical parity is desired. λ^* is enough to describe the maximum entropy distribution p^* as $p^*(\alpha) \propto q(\alpha)e^{\langle \lambda^*, \alpha \rangle}$ for any $\alpha \in \Omega$.

4 Theoretical results

Theorem 4.1 (Main Result 1: Fast second-order algorithm for dual-MaxEnt). *There exists an algorithm such that, given sets $\Omega_1, \dots, \Omega_d$, N samples from $\Omega \subseteq \mathbb{R}^n$, smoothing parameter $C > 0$ and error margin $\varepsilon > 0$, it can compute $\hat{\lambda} \in \mathbb{R}^n$ that is ε -approximate solution to dual-MaxEnt with $\tilde{O}\left(\left(N + cd + n^{\omega - \frac{3}{2}} \log\left(\frac{n}{\varepsilon}\right)\right) n^4 d \log\left(\frac{1}{C\varepsilon}\right)^2\right)$ arithmetic operations where $c := \max_{i \in [d]} |\Omega_i|$ and $\omega \sim 2.3 \dots$ is the matrix multiplication coefficient.*

The algorithm is based on a second-order framework suggested by [2, 9] for the special case of matrix scaling. It requires a convex function, f , with a unique global minimizer, to be “second order robust”¹, and an R_ε such that the ℓ_1 -ball of radius R_ε contains an ε -approximate minimizer of f . The second-order robustness condition allows one to give lower and upper bound for the Taylor approximation of f in a small ball using the gradient and the Hessian of f that leads to provable guarantee on the decrease of $f(x)$ at each iteration; see Algorithm 1. The correctness and run-time of Algorithm 1 follows with minimal modification to original proof of [2, 1] and is presented in Section 7. A key observation, missing from prior work, is that the dual maximum entropy program is second-order robust function in order to apply the box-constrained Newton method.

Lemma 4.2 (Second-order robustness of the dual-MaxEnt function). *Given $\Omega \subseteq \mathbb{R}^n$ with $D_\Omega := \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_1$, a $q : \Omega \rightarrow [0, 1]$ and the target expected vector $\theta \in \text{conv}(\Omega)$, the dual maximum entropy function $h_{\theta, q}(\lambda) := \log\left(\sum_{\alpha \in \Omega} q(\alpha)e^{\langle \lambda, \alpha - \theta \rangle}\right)$ is $4D_\Omega$ -second order robust.*

The proof of this lemma appears in Section 8. The unique global minimizer condition can be satisfied by adding an ℓ_2 -regularization term with sufficiently small coefficient; we apply algorithm 1 to the function $f_\varepsilon(\lambda) := h_{\theta, q}(\lambda) + \frac{\varepsilon}{nR_\varepsilon^2} \|\lambda\|_2^2$, where $R_\varepsilon := O(nd \log(n \max_{i \in [d]} |\Omega_i| / C\varepsilon))$. Any ε -approximate minimizer of $h_{\theta, q}$ in the ℓ_1 -ball of radius R_ε is 2ε -minimizer of f_ε . The proof follows from Lemma 9.5, which provides a bound on the size of approximate solutions of dual-MaxEnt.

Crucially, we show that one can efficiently compute the value of dual-MaxEnt, its gradient, and its Hessian from given samples. If $S_{q, \lambda}[f]$ denotes $\sum_{\alpha \in \Omega} f(\alpha) \cdot q(\alpha)e^{\langle \lambda, \alpha \rangle}$, then one needs to compute the exponential sums $S_{q_C^w, \lambda}[1]$, $S_{q_C^w, \lambda}[x]$, and $S_{q_C^w, \lambda}[xx^\top]$ for given x .

Lemma 4.3 (Oracles for the dual objective function). *Given $\Omega = \Omega_1 \times \dots \times \Omega_d \subseteq \mathbb{R}^n$, $C \in [0, 1]$, samples $\mathcal{S} := \{\alpha_i\}_{i \in [N]} \subseteq \Omega$ with normalizing weights $w \in \Delta_{N-1}$, and $\lambda \in \mathbb{R}^n$, one can compute $S_{q_C^w, \lambda}[1]$, $S_{q_C^w, \lambda}[x]$, and $S_{q_C^w, \lambda}[xx^\top]$ with $O(n^2(N + d \max_{i \in [d]} |\Omega_i|))$ arithmetic operations where q_C^w is the prior distribution estimated from \mathcal{S} , w and C as in Eq. (1).*

¹A convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be α -second order robust, if for all $x, y \in \mathbb{R}^n$ with $\|y\|_\infty \leq 1$ satisfies $|D^3 f(x)[y, y, y]| \leq \alpha D^2 f(x)[y, y]$ where $D^k f(x)[y, \dots, y] := \frac{d^k}{dt^k} f(x + ty) \Big|_{t=0}$.

Algorithm 1 Box constrained Newton's method, [2]

- 1: **Input:** first and second order oracle access to α -second-order robust function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, starting point $x \in \mathbb{R}^n$, promise of ℓ_1 ball of radius R_ε that contains ε -approximate minimizer of f , $\varepsilon > 0$, and number of iterations T
- 2: **for** $i = 1$ **to** T **do**
- 3: compute a $\frac{\varepsilon}{2\alpha R_\varepsilon}$ -approximate minimizer y_ε of the following convex quadratic program,

$$\inf_{y \in \mathbb{R}^n} \langle \nabla f_x, y \rangle + \frac{1}{2e} y^\top \nabla^2 f_x y \quad \text{s.t.} \quad \|y\|_\infty \leq 1/\alpha \quad \text{and} \quad \|x + y\|_\infty \leq R_\varepsilon \quad (2)$$

where $f_x(y) := f(x + y)$.

- 4: $x \leftarrow x + y_\varepsilon/e^2$
 - 5: **end for**
-

Fairness guarantees. Let Ω_ℓ be a protected attribute and $y \in \mathcal{Y}$. Let Y denote the random variable when p^* is restricted to \mathcal{Y} and let Z when p^* is restricted to Ω_ℓ . Then, $\frac{p^*(Y=y|Z=z_1)}{p^*(Y=y|Z=z_2)} = \frac{p^*(Y=y, Z=z_1)}{p^*(Y=y, Z=z_2)} \cdot \frac{p^*(Z=z_2)}{p^*(Z=z_1)}$. The ratio $p^*(Z=z_2)/p^*(Z=z_1)$ can be controlled using the expected value θ taken as input to the program, since $p^*(Z=z) = \theta(Z=z)$. Therefore, to ensure a particular statistical rate, we need to ensure that $p^*(Y=y, Z=z)$ is roughly equal for all values $z \in \Omega_\ell$.

To ensure that $p^*(Y=y, Z=z)$ is equal (or approximately equal) for all values $z \in \Omega_\ell$, we choose a prior distribution which satisfies this property. The first and obvious choice for the weights $\{w_i\}_{i=1}^N$ in Eq. (1) of prior distribution is the uniform weights to all samples. We can show that for this choice of weights, the probability $p(Y=y, Z=z)$ is not equal for all $z \in \Omega_\ell$ and depends on the samples chosen; see Section 12. We propose a prior with different weights to ensure that the above property is satisfied. To make the probability $q_C(Z=z, Y=y)$ independent of the count $c(y, z)$, we can assign weights inversely proportional to $c(y, z)$. Algorithm 2 describes our approach for estimating desired weights from samples. The following lemma establishes the correctness of Algorithm 2.

Lemma 4.4 (Joint distribution is fair in case of weights obtained from Algorithm 2; see Section 11). *Let q_C^w be the prior distribution computed according to Eq. (1) and using the weights obtained from Algorithm 2. Let Y denote the random variable when p^* is restricted to \mathcal{Y} and let Z denote the random variable when p^* is restricted to Ω_ℓ . Then for a fixed $y \in \mathcal{Y}$ and $\forall z_1, \dots, z_k \in \Omega_\ell$,*

$$q_C^w(Z=z_1, Y=y) = q_C^w(Z=z_2, Y=y) = \dots = q_C^w(Z=z_k, Y=y).$$

As mentioned earlier, given that the joint probability of y and z does not vary for different values of z , we can control the statistical rate of the max-entropy distribution using the expected vector θ . Using the prior distribution q_C^w , with weights from Algorithm 2, and θ , we can prove the following bound on the fairness of max-entropy distributions.

Theorem 4.5 (Main Result 2: Fairness guarantees on the output distribution). *Given samples $\mathcal{S} \subseteq \Omega \subseteq \mathbb{R}^n$ and parameters $\tau, C \in [0, 1]$, let q_C^w be the prior distribution computed using the weights obtained from Algorithm 2. Let D_Ω denote the ℓ_2 -diameter of Ω , i.e., $D_\Omega = \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_2$. Let θ denote the given expected vector and θ_C denote the expected vector corresponding to q_C^w . Let p^* denote the max-entropy distribution obtained corresponding to the prior distribution q_C^w and expected value θ . Let Y denote the random variable when p^* is restricted to \mathcal{Y} and let Z denote the random variable when p^* is restricted to the domain of the protected attribute Ω_ℓ . Then, the max-entropy distribution p^* satisfies the following bounds:*

Algorithm 2 Reweighting samples to construct the prior distribution

```

1: Input: samples  $\mathcal{S} := \{(X_i, Y_i, Z_i)\}_{i \in N} \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ 
2: for  $y \in \mathcal{Y}, z \in \mathcal{Z}$  do
3:    $c(y) \leftarrow \sum_{i \in [N]} 1_{Y_i=y}$ 
4:    $c(y, z) \leftarrow \sum_{i \in [N]} 1_{Y_i=y, Z_i=z}$ 
5: end for
6: for  $i = 1$  to  $N$  do
7:    $w_i \leftarrow c(y_i)/c(y_i, z_i)$ 
8: end for
9:  $W \leftarrow \sum_{i \in [N]} w_i$ 
10: return  $\{w_i/W\}_{i \in [N]}$ 

```

- (Representation rate) If θ satisfies the property that $\theta(Z=z_i)/\theta(Z=z_j) \geq \tau$, then $\frac{p^*(Z=z_i)}{p^*(Z=z_j)} \geq \tau$.
- (Statistical rate) If θ satisfies the property that $\theta(Z=z_i)/\theta(Z=z_j) \geq \tau$, for all values $z_i, z_j \in \Omega_\ell$ and $\|\theta - \theta_C\|_1 \leq \varepsilon_\theta$, then for a fixed class label $y \in \mathcal{Y}$

$$\frac{p^*[Y=y|Z=z_i]}{p^*[Y=y|Z=z_j]} \geq \tau \left(1 - \frac{2\sqrt{R \cdot \varepsilon_\theta} \cdot |\Omega_\ell| \cdot |\mathcal{Y}|}{C - \sqrt{R \cdot \varepsilon_\theta} \cdot |\Omega_\ell| \cdot |\mathcal{Y}|} \right).$$

In all the above bounds, $R = O(n^{3/2} \cdot (\log \frac{|\Omega|}{C} + \log \frac{nD_\Omega}{\varepsilon_\theta}))$.

Thus, one can pick C to be small enough to ensure that ε_θ is small and that the statistical rate of p^* is close to τ . The proof of the above theorem is presented in the Section 11. Theorem 4.5 shows the one can control the statistical rate or ensure bounded fairness of the output distribution using the expected value, upto a certain error margin.

Price of fairness. Using the recent results of [23], we are also able to bound the distance of the max-entropy distribution from the prior distribution as a function of the input expected vector; this can be interpreted as *price of fairness*. We provide details in Section 10.

5 Empirical results

Datasets: We evaluate our approach on the **COMPAS** dataset [3, 18], which contains information on criminal defendants at the time of trial (including crime history, age, sex, and race), and then tracked defendants for a period post-trial to record instances of recidivism (in this case marked as re-arrest). In our simulations we use two versions of the dataset: The *small* dataset contains sex, race, age, priors count, and charge degree as features, and uses a binary marker of recidivism within two years as the label. The age attribute is categorized into three categories, younger than 25, between 25 and 45, and older than 45, and the priors count attribute is categorized in to three categories (no prior crime, between 1 and 3, and more than 3). We only consider data for convicted criminals labeled as being either White or Black. Consequently, the domain Ω_S for this version is $\{0, 1\}^4 \times \Delta_2^2$ and contains 144 possible unique points. The *large* dataset consists of 19 attributes and samples from 6 different racial categories and included additional features such as the type of prior and juvenile prior counts. Overall, the domain contains approximately 1.4×10^{11} possible unique points, and the results are presented in Section 14. We use gender as the protected attribute, and take “no recidivism” to be the favorable outcome.

Algorithms and Baselines: Given training data $\mathcal{S} \subseteq \Omega$, we estimate different maximum entropy distributions with given parameters using \mathcal{S} . We use two kinds of prior distributions: (1)

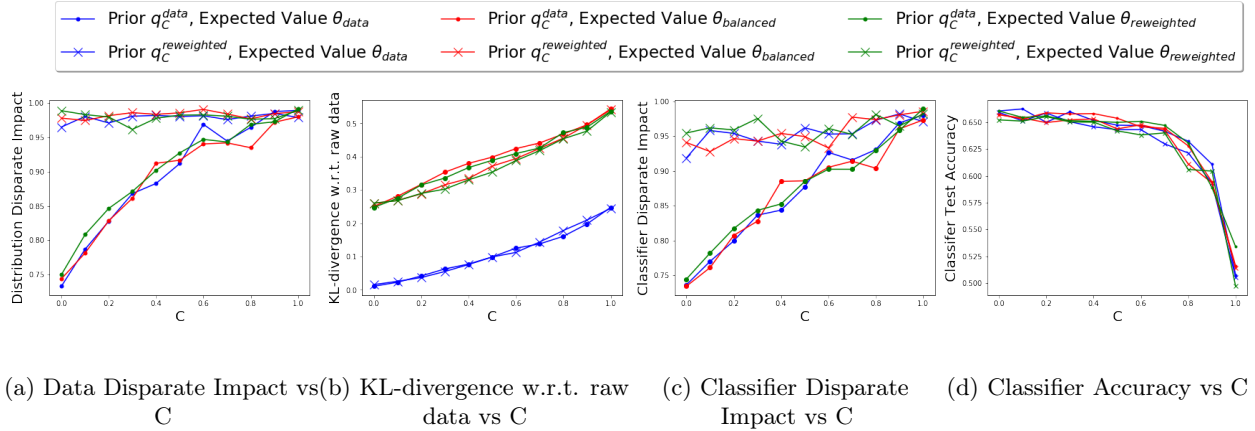


Figure 1: Comparison of max-entropy distributions with different priors and expectation vectors. Note that a value of $C = 1$ effectively would result in sampling uniformly at random from the entire domain. Hence, as expected, we see fairness increase and accuracy decrease as C increases. (a) Data disparate impact. We observe that using q_C^{rewt} is more fair with respect to disparate impact than using q_C^{data} . Surprisingly, the value of C does not significantly affect the results for q_C^{rewt} . (b) KL-divergence between the empirical distributions as compared with the raw data. We observe that this value is smaller when using the expected vector θ_{data} . (c) Classifier disparate impact. Similar to before, we observe that using the q_C^{rewt} prior results in a fairer outcome. Here there is a slight increase in fairness as C is increased even for q_C^{rewt} . (d) Classifier accuracy. We observe that there is no significant difference in accuracy across different metrics and priors. This is surprising, especially in light of the significant differences with respect to how well they capture the raw data as depicted in Figure 1(b).

q_C^{data} assigns uniform weights to the samples, i.e., $w = \{1/N\}_{i=1}^N$. (2) q_C^{rewt} assigns weights returned by the Algorithm 2. We use three kinds of expectation vectors: (a) The expected value of the dataset S , $\theta_{\text{data}} := \left(\sum_{i \in [N]} \frac{1}{N} X_i, \sum_{i \in [N]} \frac{1}{N} Y_i, \sum_{i \in [N]} \frac{1}{N} Z_i \right)$. The resulting max-entropy distribution is our best guess for the underlying distribution without any modification for fairness. (b) The same as above except that the marginal of the sensitive attribute is changed to ensure balanced representation, i.e., $\theta_{\text{bal}} := \left(\sum_{i \in [N]} \frac{1}{N} X_i, \sum_{i \in [N]} \frac{1}{N} Y_i, \sum_{z \in \mathcal{Z}} z/|\mathcal{Z}| \right)$. (c) The weighted mean of the samples, i.e., $\theta_{\text{rewt}} := \left(\sum_{i \in [N]} w_i X_i, \sum_{i \in [N]} w_i Y_i, \sum_{i \in [N]} w_i Z_i \right)$, as obtained from Algorithm 2.

This results in six distributions; we generate a synthetic dataset from each distribution to use in our evaluation. We also compare against the raw data, another reweighting method [14], and an optimized preprocessing method [6]. For classification-based metrics, we used a decision tree classifier (with gini information criterion as splitting rule). We also evaluated a Gaussian naive Bayes classifier; the results were similar and are presented in Section 13.

Metrics: We evaluate both the simulated datasets and the classifiers trained on these datasets with respect to metrics of fairness and accuracy. We consider disparate impact, the ratio between the probability of observing a favourable outcome given unprivileged group membership and the probability of observing a favorable outcome given privileged group membership, as a metric of data fairness. More precisely, $\min \{ \mathbb{P}[Y=1|Z=0]/\mathbb{P}[Y=1|Z=1], \mathbb{P}[Y=1|Z=1]/\mathbb{P}[Y=1|Z=0] \}$. We also consider the overall gender ratio of the simulated data, i.e., $\min \{ \mathbb{P}[Z=0]/\mathbb{P}[Z=1], \mathbb{P}[Z=1]/\mathbb{P}[Z=0] \}$. For both metrics, the a value closer to 1 is considered “more fair”. To evaluate the fit of the simulated dataset with the original data using the KL-divergence of the empirical distributions of the datasets. The smaller the

Table 1: Comparison of max-entropy distribution with the state-of-the-art. The max-entropy distribution uses parameter $C = 0.5$, prior q_C^{rewt} and expected value θ_{rewt} or θ_{bal} . "DI" denotes Disparate Impact. The first column shows that the max-entropy distribution has high DI. The gender ratio of the max-entropy distribution dataset is also much higher than other output dataset, due to the balanced marginals used. The DI of the decision tree classifier is high for the max-entropy dataset as well. Despite the fact that the KL-divergence of the max-entropy distributions from the raw data is high, it does not affect the classifier accuracy significantly.

Dataset	Data DI	KL-div	Gender Ratio	Class. DI	Class. Accuracy
Raw Data	0.73	0	0.24	0.74	0.67
Max-Entropy using $q_C^{\text{rewt}}, \theta_{\text{rewt}}$	0.95	0.35	0.98	0.95	0.66
Max-Entropy using $q_C^{\text{rewt}}, \theta_{\text{bal}}$	0.97	0.35	0.99	0.97	0.65
Optimized pre-processing [6]	0.87	0.02	0.24	0.93	0.66
Re-weighting samples [14]	0.99	0.14	0.24	0.97	0.67

value of KL-divergence, the closer the simulated distribution resembles the original data.

For the classifier trained on the generated data, we again compute the disparate impact using predictions to evaluate the effects of different training data on the fairness of the classifier. Let \hat{Y} be the random variable corresponding to the classifier prediction, then the disparate impact with respect to the classifier is $\min \left\{ \frac{\mathbb{P}[\hat{Y}=1|Z=0]}{\mathbb{P}[\hat{Y}=1|Z=1]}, \frac{\mathbb{P}[\hat{Y}=1|Z=1]}{\mathbb{P}[\hat{Y}=1|Z=0]} \right\}$. In addition, we report the classifier accuracy when trained on each simulated dataset.

Evaluation: We perform five fold cross-validation. For each dataset, we fit six different model of maximum entropy distribution as discussed earlier with different C values, the parameter that controls the amount of uniformly distributed mass. We used C values in the range $[0, 1]$ with increments 0.1.

5.1 Results

Comparison across priors and expected value vectors: We first evaluate the dataset generated using max-entropy distributions with different combinations of prior weights and expected value mentioned earlier. The results for this evaluation are present in Figure 1.

Figure 1 shows that the max-entropy distributions obtained using prior q_C^{rewt} achieve higher disparate impact values than the distributions obtained using q_C^{data} . However, the KL-divergence of the max-entropy distributions obtained using expected value θ_{rewt} or θ_{bal} are higher as well, as seen in Figure 1b. As the samples in the raw dataset are unbalanced with respect to gender, the distributions using balanced marginal distributions are expected to have a larger divergence from the empirical distribution of raw data than the distributions using the expected value of data. The gender ratio of the distributions using θ_{rewt} or θ_{bal} are also high, as seen in Figure 2 in the Section 13.

The classifier results are presented in Figure 1c and Figure 1d. Once again the the max-entropy distributions obtained using prior distribution q_C^{rewt} achieve higher classifier disparate impact values than the distributions obtained using q_C^{data} . The accuracy of the classifiers trained on datasets obtained using prior distribution q_C^{rewt} is slightly lower than the accuracy of the classifiers trained on distributions obtained using sample uniform weights. However, it is interesting to note that the significant difference in "accuracy" of the data all but disappears when passed through the classifier.

Importantly, the accuracy drops sharply as the value of C increases as $C = 1$ assigns equal probability mass to all points in the domain and ignores the original samples. This suggests a C value in the low-to-mid range would likely optimize accuracy and disparate impact simultaneously.

Comparison with the state-of-the-art: Table 1 presents the comparison of the dataset generated from the max-entropy distribution using prior distribution q_C^{rewt} , expected value θ_{rewt} or θ_{bal} and $C = 0.5$ with the state-of-the-art. The max-entropy distributions have better disparate impact values (close to 1) than the raw data or data from the optimized pre-processing algorithm [6]. Even though it has worse KL-divergence from the empirical distribution from raw data than other algorithms, it does not affect the accuracy of the classifier trained on the generated data, which is only slightly smaller than the accuracy of the classifier trained on raw data. The gender ratio of the dataset generated from the max-entropy distribution (close to 1) is also much higher than other algorithms. In conjunction, this suggests that using our approach to generate simulated datasets for use in training can lead to fairer classifiers without a significant loss to accuracy.

6 Limitations and future work

This paper provides an approach to control the fairness properties of an output distribution by changing the inputs (i.e., the expected vector and prior distribution) to the max-entropy program. The max-entropy framework can, however, be utilized in many different ways to satisfy various other properties. Firstly, one could also consider the usage of other prior distributions and re-weighting algorithms to satisfy various other kinds of group fairness constraints. More generally, our current fairness definitions support conditioning on a single sensitive attribute. While the presentation considered sensitive attributes with a binary value, it trivially generalizes to the multiple-value case (for a single attribute).

Given multiple sensitive attributes, one could pool them together to form a larger categorical sensitive attribute that captures intersectional types. This would allow us to control for fairness across intersectional types in the same manner by treating each one as a different value of the same sensitive attribute, and hence the method can be used to address fairness across intersectional types. However, if we want to consider fairness across multiple sensitive attributes separately, then this approach would likely not suffice, and we would need to provide constraints on second-order (or larger) moments in addition to the existing constraints. This results in an optimization problem that is significantly harder and would be an interesting direction for future work.

7 Second order box-constrained Newton’s method

Given a second-order robust function, we expand upon the details of finding the approximate minimizer using Algorithm 1. We first look at the number of iterations required by Algorithm 1 to return an approximate minimizer of an α -second order robust function and then bound the time taken by each iteration.

7.1 Bounding the number of iterations of Algorithm 1

The algorithm is based on a second-order optimization framework recently suggested by [2, 9] for the special case of matrix scaling. In particular, we can prove the following theorem regarding the number of iterations of the algorithm. The correctness and run-time of this algorithm follows with minimal modification to original proof of [2].

Theorem 7.1 (Run time of the Box constrained Newton's method, [2]). *Given access to the first and second order oracles for α -second order robust function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with unique global minimum, promise of ℓ_∞ ball of radius R_ε contains ε -approximate minimizer of f , starting point $x \in \mathbb{R}^n$ with $\|x\|_\infty \leq R_\varepsilon$, and $\varepsilon > 0$, algorithm 1 runs for $O\left(\alpha R_\varepsilon \log\left(\frac{\text{var}_{R_\varepsilon}(f)}{\varepsilon}\right)\right)$ iterations and outputs 3ε -approximate minimizer of f where $\text{var}_{R_\varepsilon}(f) := \max_{x,y:\|x\|_1,\|y\|_1 \leq R_\varepsilon} f(x) - f(y)$.*

In order to prove Theorem 7.1, we need following simple lemma by [1] about the lower and upper bounds on the second order Taylor approximation of a second order robust univariate function.

Lemma 7.2 (Bounds for second order Taylor approximation[1]). *If $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is an α -second order robust convex function, then for any $y \in \mathbb{R}^m$ and $x \in \mathbb{R}^m$ with $\|x\|_\infty \leq s$, the univariate function $f_{x,y}(t) := f(x + ty)$ satisfies*

$$\forall |t| \leq \frac{1}{\alpha}, f_{x,y}(0) + f'_{x,y}(0)t + \frac{1}{2e} f''_{x,y}(0)t^2 \leq f_{x,y}(t) \quad (3)$$

and

$$\forall |t| \leq \frac{1}{\alpha}, f_{x,y}(t) \leq f_{x,y}(0) + f'_{x,y}(0)t + \frac{e}{2} f''_{x,y}(0)t^2 \quad (4)$$

where $f'_{x,y}$ and $f''_{x,y}$ are first and second derivatives of $f_{x,y}$, respectively.

Next, we present a lemma to quantify progress made by algorithm 1 at each iteration. We remark that the lemma and its proof is the modification of Lemma C.4 presented by [1].

Lemma 7.3. *Let $x_\varepsilon \in \mathbb{R}^n$ be ε -approximate minimizer of α -second order robust function f for given $\varepsilon > 0$ with $\|x_\varepsilon\|_\infty \leq R_\varepsilon$. If one iteration of algorithm 1 applied to point $x_0 \in \mathbb{R}^n$ with $\|x_0\|_\infty \leq R_\varepsilon$, then resulting point x_1 satisfies*

$$f(x_1) - f(x_\varepsilon) \leq \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right) (f(x_0) - f(x_\varepsilon)) + \frac{\varepsilon}{2e^2\alpha R_\varepsilon}.$$

Proof. Let us denote the update toward x_ε with $y_\varepsilon := x_\varepsilon - x_0$. By our assumption $\|x_\varepsilon\|_\infty \leq R_\varepsilon$ and therefor, $\|x_0 + y_\varepsilon\|_\infty \leq R_\varepsilon$. In particular, for any constant $\beta \in [0, 1]$, $\|x_0 + \beta y_\varepsilon\|_\infty \leq R_\varepsilon$ since $\|x_0\|_\infty \leq R_\varepsilon$. Thus, there is a feasible vector for the inner optimization problem (2) in the direction of y^* . Since, $\|x_0\|_\infty \leq R_\varepsilon$, we also have $\|y_\varepsilon\|_\infty = \|x_\varepsilon - x_0\|_\infty \leq 2R_\varepsilon$. Thus, a desired solution to the inner optimization problem $y'_\varepsilon := \frac{y_\varepsilon}{2\alpha R_\varepsilon}$ is feasible. Similarly for any $x \in \mathbb{R}^m$ with $\|x\|_\infty \leq R_\varepsilon$, if $y = x - x_0$, then $\|y\|_\infty \leq 2R_\varepsilon$. If $\beta := \min\left(1, \frac{1}{\alpha\|y\|_\infty}\right)$, then $x' := \beta x$ is a feasible direction. Let us define the univariate function $f_{x_0,y}(t) := f_{x_0}(ty) = f(x_0 + ty)$. $f_{x_0,y}$ is a convex function since it is a restriction of a convex function f to a line. Hence

$$\begin{aligned} f_{x_0}(0) - f_{x_0}\left(\frac{y}{2\alpha R_\varepsilon}\right) &= f_{x_0,y}(0) - f_{x_0,y}\left(\frac{1}{2\alpha R_\varepsilon}\right) \\ &\geq \frac{1}{2\alpha R_\varepsilon} (f_{x_0,y}(0) - f_{x_0,y}(1)) \\ &= \frac{1}{2\alpha R_\varepsilon} (f_{x_0}(0) - f_{x_0}(y)) \end{aligned} \quad (5)$$

by the convexity of $f_{x_0, y}$. Let us denote the gradient of f at x_0 with g and the Hessian of f at x_0 with H . If we apply Lemma 7.2 to $f_{x_0, y}$, then we get

$$\begin{aligned}\langle g, y' \rangle + \frac{1}{2e} y'^\top H y' &= \langle g, y \rangle \beta + \frac{1}{2e} y^\top H y \beta^2 \\ &\leq f_{x_0}(\beta y) - f_{x_0}(0) \\ &= f_{x_0}(y') - f_{x_0}(0),\end{aligned}\tag{6}$$

$$\begin{aligned}\langle g, y' \rangle + \frac{e}{2} x y^\top H y' &= \langle g, y \rangle \beta + \frac{e}{2} y^\top H y \beta^2 \\ &\geq f_{x_0}(\beta y) - f_{x_0}(0) \\ &= f_{x_0}(y') - f_{x_0}(0).\end{aligned}\tag{7}$$

Let us denote $\varepsilon/(2\alpha R_\varepsilon)$ -approximate minimizer of the inner optimization problem (2) by y^\star . Since y'_ε is a feasible vector, we have

$$\langle g, y^\star \rangle + \frac{1}{2e} (y^\star)^\top H y^\star \leq \frac{\varepsilon}{2\alpha R_\varepsilon} + \langle g, y'_\varepsilon \rangle + \frac{1}{2e} y'_\varepsilon{}^\top H y'_\varepsilon\tag{8}$$

Combining (8) with (6) for a desired solution y'_ε , we get

$$\begin{aligned}\langle g, y^\star \rangle + \frac{1}{2e} (y^\star)^\top H y^\star &\leq - (f_{x_0}(0) - f_{x_0}(y'_\varepsilon)) + \frac{\varepsilon}{2\alpha R_\varepsilon} \\ &\stackrel{(5)}{\leq} - \frac{1}{2\alpha R_\varepsilon} (f_{x_0}(0) - f_{x_0}(y_\varepsilon) - \varepsilon) \\ &= - \frac{1}{2\alpha R_\varepsilon} (f(x_0) - f(x_\varepsilon) - \varepsilon).\end{aligned}\tag{9}$$

Combining (8) with (7) for the approximate solution y^\star , we get

$$\begin{aligned}\langle g, y^\star \rangle + \frac{1}{2e} (y^\star)^\top H y^\star &= e^2 \left(\langle g, \frac{y^\star}{e^2} \rangle + \frac{e}{2} \frac{(y^\star)^\top}{e^2} H \frac{y^\star}{e^2} \right) \\ &\stackrel{(7)}{\geq} - e^2 \left(f_{x_0}(0) - f_{x_0} \left(\frac{y^\star}{e^2} \right) \right) \\ &= - e^2 (f(x_0) - f(x_1)).\end{aligned}\tag{10}$$

Combining (9) and (10), we get

$$- \frac{1}{2\alpha R_\varepsilon} (f(x_0) - f(x_\varepsilon)) + \frac{\varepsilon}{2\alpha R_\varepsilon} \geq - e^2 (f(x_0) - f(x_1)).$$

Equivalently,

$$f(x_1) - f(x_\varepsilon) \leq \left(1 - \frac{1}{2e^2 \alpha R_\varepsilon} \right) (f(x_0) - f(x_\varepsilon)) + \frac{\varepsilon}{2e^2 \alpha R_\varepsilon}.$$

□

Proof of Theorem 7.1. Let x_0, x_1, \dots be the sequence of points generated by the algorithm 1. We note that by construction $\|x_i\|_\infty \leq R_\varepsilon$ for any $i \in \mathbb{Z}_{\geq 0}$. Thus, we can apply Lemma 7.3. If x_ε is a ε

approximate minimizer of f with $\|x_\varepsilon\|_\infty \leq R_\varepsilon$, then

$$\begin{aligned}
f(x_{t+1}) - f(x_\varepsilon) &\leq \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right)^t (f(x_0) - f(x_\varepsilon)) + \left(\frac{\varepsilon}{2e^2\alpha R_\varepsilon}\right) \sum_{i=0}^t \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right)^i \\
&\leq \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right)^t (f(x_0) - f(x_\varepsilon)) + \left(\frac{\varepsilon}{2e^2\alpha R_\varepsilon}\right) \sum_{i=0}^\infty \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right)^i \\
&= \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right)^t (f(x_0) - f(x_\varepsilon)) + 2e^2\alpha R_\varepsilon \left(\frac{\varepsilon}{2e^2\alpha R_\varepsilon}\right) \\
&= \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right)^t (f(x_0) - f(x_\varepsilon)) + \varepsilon \\
&\leq \left(1 - \frac{1}{2e^2\alpha R_\varepsilon}\right)^t \text{var}_{R_\varepsilon}(f) + \varepsilon
\end{aligned}$$

Thus, if $T := O\left(\alpha R_\varepsilon \log\left(\frac{\text{var}_{R_\varepsilon}(f)}{\varepsilon}\right)\right)$, then x_T satisfies

$$f(x_T) - f(x_\varepsilon) \leq 2\varepsilon.$$

Consequently, x_T is 3ε -approximate minimizer of f . □

7.2 Solving the inner optimization problem

We next bound the time taken by each iteration. Note that the inner optimization program, i.e., the program 2, is a convex quadratic optimization problem since f is a convex function, thus it can be solved with standard convex optimization techniques, in particular using primal path following algorithm. We have the following standard run-time bound for the primal path-following algorithm.

Theorem 7.4 (Convex quadratic optimization). *There is a primal path following algorithm, given positive definite matrix $A \in \mathbb{R}^{n \times n}$, a vector b , bounds $L, U \in \mathbb{R}^n$ and $\varepsilon > 0$ outputs a ε -approximate minimizer of*

$$\begin{aligned}
&\inf_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x + b^\top x \\
&s.t. \quad L \preceq x \preceq U
\end{aligned}$$

with $O\left(n^{1/2+\omega} \log\left(\frac{\|A\|_2 \|b\|_2 r}{\varepsilon}\right)\right)$ arithmetic operations where $r := \max_{j \in [n]} U_j - L_j$, ω is the fast matrix multiplication coefficient, and $\|A\|_2$ denotes the largest eigenvalue of A .

Using Theorem 7.1 and Theorem 7.4, we get a bound on the runtime of the box-constrained Newton's method, given access to first and second order oracles.

8 Proof of second-order robustness: Proof of Lemma 4.2

We use $S_{q,\lambda}[f]$ to denote $\sum_{\alpha \in \Omega} q(\alpha) f(\alpha) e^{\langle \lambda, \alpha \rangle}$ for given prior distribution $q : \Omega \rightarrow [0, 1]$ and vector $\lambda \in \mathbb{R}^n$ for brevity in the remainder of this section. We recall Lemma 4.2.

Lemma 8.1 (Second-order robustness of the dual-MaxEnt function). *Given $\Omega \subseteq \mathbb{R}^n$ with $D_\Omega := \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_1$, a $q : \Omega \rightarrow [0, 1]$ and the target expected vector $\theta \in \text{conv}(\Omega)$, the dual maximum entropy function $h_{\theta,q}(\lambda) := \log\left(\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha - \theta \rangle}\right)$ is $4D_\Omega$ -second order robust.*

Before proving the lemma, we state and prove the following general claim in the proof.

Claim 8.2. *Let X be a real valued random variable over the discrete set Ω with $|X| \leq r$ for some constant $r \in \mathbb{R}_+$. Then,*

$$|\mathbb{E}[X^3] - \mathbb{E}[X^2] \mathbb{E}[X]| \leq 2r(\mathbb{E}[X^2] - \mathbb{E}[X]^2).$$

Proof. Let us denote the probability mass function of X with p . Then,

$$\begin{aligned} & \mathbb{E}[X^3] - \mathbb{E}[X^2] \mathbb{E}[X] \\ &= \sum_{\alpha \in \Omega} X(\alpha)^3 p(\alpha) - \sum_{\alpha, \beta \in \Omega} X(\alpha)^2 X(\beta) p(\alpha) p(\beta) \\ &= \frac{1}{2} \sum_{\alpha, \beta \in \Omega} (X(\alpha)^3 - X(\alpha)^2 X(\beta)) p(\alpha) p(\beta) \\ &\quad + \frac{1}{2} \sum_{\alpha, \beta \in \Omega} (X(\beta)^3 - X(\alpha) X(\beta)^2) p(\alpha) p(\beta) \\ &= \frac{1}{2} \sum_{\alpha, \beta \in \Omega} (X(\alpha) - X(\beta))^2 (X(\alpha) + X(\beta)) p(\alpha) p(\beta). \end{aligned}$$

We also note that, $|X(\alpha) + X(\beta)| \leq 2r$ for any $\alpha, \beta \in \Omega$ as $|X| \leq r$. Therefore,

$$\begin{aligned} & |\mathbb{E}[X^3] - \mathbb{E}[X^2] \mathbb{E}[X]| \\ &= \frac{1}{2} \left| \sum_{\alpha, \beta \in \Omega} (X(\alpha) - X(\beta))^2 (X(\alpha) + X(\beta)) p(\alpha) p(\beta) \right| \\ &\leq r \sum_{\alpha, \beta \in \Omega} (X(\alpha) - X(\beta))^2 p(\alpha) p(\beta) \\ &= 2r(\mathbb{E}[X^2] - \mathbb{E}[X]^2). \end{aligned}$$

□

Proof of Lemma 4.2. Let us fix a point $\lambda_0 \in \mathbb{R}^n$ and a direction $\lambda_1 \in \mathbb{R}^n$ with $\|\lambda_1\|_\infty \leq 1$. We need to verify that

$$|D^3 h_{\theta, q}(\lambda_0)[\lambda_1, \lambda_1, \lambda_1]| \leq 4D_\Omega D^2 h_{\theta, q}(\lambda_0)[\lambda_1, \lambda_1] \quad (11)$$

to show that $h_{\theta, q}$ is $4D_\Omega$ -second order robust. A simple calculation shows that

$$\begin{aligned} D^2 h_{\theta, q}(\lambda_0)[\lambda_1, \lambda_1] &= \frac{S_{q, \lambda_0}[\langle x, \lambda_1 \rangle^2]}{S_{q, \lambda_0}[1]} - \frac{S_{q, \lambda_0}[\langle x, \lambda_1 \rangle]^2}{S_{q, \lambda_0}[1]^2}, \\ D^3 h_{\theta, q}(\lambda_0)[\lambda_1, \lambda_1, \lambda_1] &= \frac{S_{q, \lambda_0}[\langle x, \lambda_1 \rangle^3]}{S_{q, \lambda_0}[1]} + \frac{2S_{q, \lambda_0}[\langle x, \lambda_1 \rangle]^3}{S_{q, \lambda_0}[1]^3} - \frac{3S_{q, \lambda_0}[\langle x, \lambda_1 \rangle^2]S_{q, \lambda_0}[\langle x, \lambda_1 \rangle]}{S_{q, \lambda_0}[1]^2}. \end{aligned}$$

We note that shifting Ω by a vector does not affect the function $h_{\theta, q}$ as θ would be also shifted by the same vector. Hence without loss of generality we assume that $0 \in \Omega$, which implies that $\|\alpha\|_1 \leq D_\Omega$ for $\alpha \in \Omega$. Consequently, we have

$$\begin{aligned} \left| \frac{S_{q, \lambda_0}[\langle x, \lambda_1 \rangle^3]}{S_{q, \lambda_0}[1]^3} - \frac{S_{q, \lambda_0}[\langle x, \lambda_1 \rangle^2]S_{q, \lambda_0}[\langle x, \lambda_1 \rangle]}{S_{q, \lambda_0}[1]^2} \right| &= \left| \frac{S_{q, \lambda_0}[\langle x, \lambda_1 \rangle]}{S_{q, \lambda_0}[1]} \right| D^2 h_{\theta, q}(\lambda_0)[\lambda_1, \lambda_1] \\ &\leq D_\Omega D^2 h_{\theta, q}(\lambda_0)[\lambda_1, \lambda_1] \end{aligned} \quad (12)$$

as $|\langle \alpha, \lambda_1 \rangle| \leq \|\alpha\|_1 \|\lambda\|_\infty \leq D_\Omega$ by Cauchy-Schwartz inequality for any $\alpha \in \Omega$. On the other hand, if we define the probability distribution $p_\lambda(\alpha) := q(\alpha)e^{\langle \lambda, \alpha \rangle} / S_{q,\lambda}[1]$, then $S_{q,\lambda}[f] / S_{q,\lambda}[1] = \mathbb{E}[p_\lambda][f]$.

Using Claim 8.2, we get

$$|\mathbb{E}[p_{\lambda_0}][\langle \lambda_1, x \rangle^3] - \mathbb{E}[p_{\lambda_0}][\langle \lambda_1, x \rangle^2] \mathbb{E}[p_{\lambda_0}][\langle \lambda_1, x \rangle]| \leq 2D_\Omega(\mathbb{E}[p_{\lambda_0}][\langle \lambda_1, x \rangle^2] - \mathbb{E}[p_{\lambda_0}][\langle \lambda_1, x \rangle]^2)$$

or equivalently

$$\left| \frac{S_{q,\lambda_0}[\langle x, \lambda_1 \rangle^3]}{S_{q,\lambda_0}[1]} - \frac{S_{q,\lambda_0}[\langle x, \lambda_1 \rangle^2] S_{q,\lambda_0}[\langle x, \lambda_1 \rangle]}{S_{q,\lambda_0}[1]^2} \right| \leq 2D_\Omega D^2 h_{\theta,q}(\lambda_0)[\lambda_1, \lambda_1] \quad (13)$$

Combining (12) and (13) with the triangle inequality, we get (11). Therefore, $h_{\theta,q}$ is a $4D_\Omega$ second-order robust function. \square

9 Fast computation of max-entropy distribution: Proof of Theorem 4.1

In this Section we present a proof of Theorem 4.1 using the second-order robustness of the max-entropy distribution and the run-time bounds of the box-constrained Newton's method. However, we will need fast first and second-order oracles to the dual program. Hence, we first prove Lemma 4.3.

9.1 Computability of the gradient and hessian of dual program

The main reason for employing Laplace smoothing is one can efficiently compute the value of dual maximum entropy program, its gradient, and its Hessian from given samples. For simplicity of notation, we drop the subscript w from the q . If $S_{q,\lambda}[f]$ denotes $\sum_{\alpha \in \Omega} f(\alpha) \cdot q(\alpha)e^{\langle \lambda, \alpha \rangle}$, then one needs $S_{q_C,\lambda}[1]$, $S_{q_C,\lambda}[x]$, and $S_{q_C,\lambda}[xx^\top]$, i.e., generalized counting oracles, to compute the gradient and the Hessian.

Lemma 9.1 (Oracles for the dual objective function). *Given $\Omega = \Omega_1 \times \dots \times \Omega_d \subseteq \mathbb{R}^n$, $C \in [0, 1]$, samples $\mathcal{S} := \{\alpha_i\}_{i \in [N]} \subseteq \Omega$ with normalizing weights $w \in \Delta_{N-1}$, and $\lambda \in \mathbb{R}^n$, one can compute $S_{q_C,\lambda}[1]$, $S_{q_C,\lambda}[x]$, and $S_{q_C,\lambda}[xx^\top]$ with $O(n^2(N + d \max_{i \in [d]} |\Omega_i|))$ arithmetic operations where q_C is the prior distribution estimated from, \mathcal{S} , w and C as in Eq. (1)*

Proof. Let u be the uniform distribution over Ω and let us treat w as a distribution over Ω by combining weights of duplicate samples in \mathcal{S} . We note that, $S_{q_C,\lambda}[f] = (1 - C)S_{w,\lambda}[f] + CS_{u,\lambda}[f]$ for any function f . Since w is non-zero for at most N points, one can directly compute $S_{w,\lambda}[f]$ by evaluating f at N points. In particular, one can compute $S_{w,\lambda}[1]$, $S_{w,\lambda}[x]$ and $S_{w,\lambda}[xx^\top]$ with $O(n^2N)$ arithmetic operations.

We note that uniform part splits over $\Omega_1, \dots, \Omega_d$. Let us write λ as $(\lambda_1, \dots, \lambda_d)$, where λ_i corresponds to i th attribute and let us define variables

$$\begin{aligned} \bar{\alpha}_i &:= (0, \dots, 0, \alpha_i, 0, \dots, 0) \in \mathbb{R}^n, \\ s_i^0 &:= \sum_{\alpha_i \in \Omega_i} e^{\langle \lambda_i, \alpha_i \rangle}, \\ s_i^1 &:= \sum_{\alpha_i \in \Omega_i} \bar{\alpha}_i e^{\langle \lambda_i, \alpha_i \rangle}, \\ s_i^2 &:= \sum_{\alpha_i \in \Omega_i} \bar{\alpha}_i \bar{\alpha}_i^\top e^{\langle \lambda_i, \alpha_i \rangle}, \end{aligned}$$

for all $i \in [d]$ and $\alpha_i \in \Omega_i$. Then, we have

$$\begin{aligned} S_{u,\lambda}[1] &= \frac{1}{|\Omega|} \prod_{i \in [d]} s_i^0, \\ S_{u,\lambda}[x] &= \frac{1}{|\Omega|} \sum_{i \in [d]} \left(s_i^1 \prod_{j \neq i} s_j^0 \right), \\ S_{u,\lambda}[xx^\top] &= \frac{1}{|\Omega|} \sum_{i \in [d]} \left[s_i^2 \prod_{j \neq i} s_j^0 + \sum_{j \neq i} s_i^1 (s_j^1)^\top \prod_{k \neq i,j} s_k^0 \right]. \end{aligned}$$

Clearly, we can compute variable s_i^0, s_i^1, s_i^2 with $O(n^2 \max_{i \in [d]} |\Omega_i|)$ arithmetic operations. Therefore, one can compute $S_{q_C,\lambda}[1]$, $S_{q_C,\lambda}[x]$, and $S_{q_C,\lambda}[xx^\top]$ with $O(n^2(N + d \max_{i \in [d]} |\Omega_i|))$ arithmetic operations. \square

9.2 Runtime of primal path following algorithm

In order to conclude the run time of primal path following algorithm when it used to minimize (2) with $f := f_\varepsilon$, we need following simple bounds on the gradient and Hessian of $h_{\theta,\varepsilon}$.

Lemma 9.2 (Size of gradient and Hessian of dual maximum entropy program). *Given a finite domain $\Omega \subseteq \mathbb{R}^n$ with $D_\Omega := \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_1$, expected value $\theta \in \text{conv}(\Omega)$, prior distribution $q : \Omega \rightarrow [0, 1]$, and a point $\lambda \in \mathbb{R}^n$, $h_{\theta,q}$ satisfies*

1. $\|\nabla h_{\theta,q}(\lambda)\|_2 \leq D_\Omega \sqrt{n}$,
2. $\|\nabla^2 h_{\theta,q}(\lambda)\|_2 \leq D_\Omega^2$, and
3. $|h_{\theta,q}(\lambda)| \leq D_\Omega \|\lambda\|_\infty$.

Proof. Without loss of generality, we assume that $\max_{\alpha \in \Omega} \|\alpha\|_1 \leq D_\Omega$. This assumption does not effect the generality of the statement, as the value of $h_{\theta,q}$, as well as its first and second derivatives are invariant under translation of the points in Ω with a common vector. Let us assume that $D_\Omega = \|\alpha_1 - \alpha_0\|_1$ for some $\alpha_0, \alpha_1 \in \Omega$. Then, we consider $\tilde{\Omega} := \{\tilde{\alpha} := \alpha - \alpha_0 \mid \alpha \in \Omega\}$ instead of Ω . We note that

$$\begin{aligned} h_{\tilde{\theta},q}(\lambda) &= \log \left(\sum_{\tilde{\alpha} \in \tilde{\Omega}} q(\alpha) e^{\langle \lambda, \tilde{\alpha} - \tilde{\theta} \rangle} \right) \\ &= \log \left(\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha - \theta \rangle} \right) = h_{\theta,q}(\lambda). \end{aligned}$$

Thus, k th order derivative of $h_{\tilde{\theta}}$ is equal to the k th order derivative of h_θ . Furthermore,

$$\max_{\tilde{\alpha} \in \tilde{\Omega}} \|\tilde{\alpha}\|_1 = \max_{\alpha \in \Omega} \|\alpha - \alpha_0\|_1 \leq \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_1 = D_\Omega.$$

Therefore, we can assume without loss of generality that $\max_{\alpha \in \Omega} \|\alpha\|_1 \leq D_\Omega$.

1. A simple calculation shows that

$$\nabla h_{\theta,q}(\lambda) = \frac{\sum_{\alpha \in \Omega} (\alpha - \theta) q(\alpha) e^{\langle \lambda, \alpha \rangle}}{\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle}} - \theta.$$

We remark that

$$\left| \sum_{\alpha \in \Omega} (\alpha - \theta)_j q(\alpha) e^{\langle \lambda, \alpha \rangle} \right| \leq D_\Omega \sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle},$$

for each $j \in [n]$ as $|\alpha_j - \theta_j| \leq \|\alpha - \theta\|_1 \leq D_\Omega$ for any $\alpha \in \Omega$. Therefore, we have $\|\nabla h_{\theta,q}(\lambda)\|_\infty \leq D_\Omega$ and $\|\nabla h_{\theta,q}(\lambda)\|_2 \leq D_\Omega \sqrt{n}$.

2. A simple calculation shows that

$$\begin{aligned} \nabla^2 h_{\theta,q}(\lambda) &= \frac{\sum_{\alpha \in \Omega} \alpha \alpha^\top q(\alpha) e^{\langle \lambda, \alpha \rangle}}{\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle}} \\ &\quad - \left(\frac{\sum_{\alpha \in \Omega} \alpha q(\alpha) e^{\langle \lambda, \alpha \rangle}}{\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle}} \right) \left(\frac{\sum_{\alpha \in \Omega} \alpha q(\alpha) e^{\langle \lambda, \alpha \rangle}}{\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle}} \right)^\top \end{aligned}$$

which is bounded above by

$$\frac{\sum_{\alpha \in \Omega} \alpha \alpha^\top q(\alpha) e^{\langle \lambda, \alpha \rangle}}{\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle}}.$$

This matrix is a positive semidefinite matrix, hence its largest eigenvalue is smaller than its trace. Therefore, we have

$$\begin{aligned} \|\nabla^2 h_{\theta,q}(\lambda)\|_2 &\leq \text{Tr} \left[\frac{\sum_{\alpha \in \Omega} \alpha \alpha^\top q(\alpha) e^{\langle \lambda, \alpha \rangle}}{\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle}} \right] \\ &= \sum_{\beta \in \Omega} \text{Tr} \left[\frac{\beta \beta^\top p(\beta) e^{-\langle \beta, y_t \rangle}}{\sum_{\alpha \in \Omega} p(\alpha) e^{-\langle \alpha, y_t \rangle}} \right] \\ &= \sum_{\beta \in \Omega} \frac{\|\beta\|_2^2 p(\beta) e^{-\langle \beta, y_t \rangle}}{\sum_{\alpha \in \Omega} p(\alpha) e^{-\langle \alpha, y_t \rangle}} \\ &\leq D_\Omega^2 \end{aligned}$$

as $\|\beta\|_2^2 \leq \|\beta\|_1^2 \leq D_\Omega^2$ for $\beta \in \Omega$.

3. We note that $|\langle \lambda, \alpha - \theta \rangle| \leq \|\lambda\|_\infty \|\alpha - \theta\|_1 \leq D_\Omega \|\lambda\|_\infty$ by Cauchy-Schwartz inequality for any $\alpha \in \Omega$. Thus, we have

$$\begin{aligned} -D_\Omega \|\lambda\|_\infty &= \log \left(\sum_{\alpha \in \Omega} q(\alpha) e^{-D_\Omega \|\lambda\|_\infty} \right) \\ &\leq \log \left(\sum_{\alpha \in \Omega} q(\alpha) e^{\langle \lambda, \alpha \rangle} \right) \\ &\leq \log \left(\sum_{\alpha \in \Omega} q(\alpha) e^{D_\Omega \|\lambda\|_\infty} \right) = D_\Omega \|\lambda\|_\infty. \end{aligned}$$

Therefore, $|h_{\theta,q}(\lambda)| \leq D_\Omega \|\lambda\|_\infty$.

□

9.3 Proof of runtime of max-entropy distribution computation

Combining second order robustness of $h_{\theta,q}$, size of its gradient and Hessian with primal path following algorithm and box-constrained Newton methods, we get a fast algorithm to minimize $h_{\theta,q}$. We start by stating a slightly more general result than Theorem 4.1.

Theorem 9.3 (Fast second-order algorithm for dual-MaxEnt – general). *There exists an algorithm given a finite domain $\Omega \subseteq \mathbb{R}^n$ with $D_\Omega := \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_1$, prior distribution $q : \Omega \rightarrow [0, 1]$, $\theta \in \text{conv}(\Omega)$, $\varepsilon > 0$, and the promise of ℓ_1 ball of radius R_ε which contains ε -approximate minimizer of $h_{\theta,q}$, outputs a 6ε -approximate minimizer of $h_{\theta,q}$ with at most*

$$\tilde{O} \left(\left(\mathcal{T}_H + n^{1/2+\omega} \log(nD_\Omega/\varepsilon) \right) D_\Omega R_\varepsilon \log(D_\Omega \cdot R_\varepsilon / \varepsilon) \right)$$

many arithmetic operations where \mathcal{T}_H is number of arithmetic operations required to compute the gradient and the Hessian of $h_{\theta,q}$.

Proof. The proof of this theorem follows directly from Theorem 7.4 and Theorem 7.1. The time taken by each primal path-following computation iteration is $n^{1/2+\omega} \log(nD_\Omega/\varepsilon)$, using the ε -value in Algorithm 1 and gradient and Hessian bounds from Lemma 9.2. Also in each iteration, we need to compute the gradient and Hessian, which takes \mathcal{T}_H operations.

For the max-entropy dual function, $|h_{\theta,q}(\lambda)| \leq D_\Omega \|\lambda\|_\infty$ from Lemma 9.2. Therefore, $\text{var}_{R_\varepsilon}(f)$ in this case is $O(D_\Omega \cdot R_\varepsilon)$. Combining these bounds, we get the above theorem. \square

We now provide the bit complexity of approximate minimizers of (dual-MaxEnt) when Laplace smoothed prior distribution is used. To provide bounds on the size of approximate solutions of dual program, we will use the following result from [23], who showed that in a large class of polytopes a ε -approximate solution can be found in a small ℓ_1 ball around the origin.

Theorem 9.4 (Size of approximate minimizers of dual maximum entropy program, [23], Theorem 5.1). *If $\Omega \subseteq \mathbb{R}^n$ is a finite set whose convex hull can be represented as $\{x \in \mathbb{R}^n \mid \forall i \in I, \langle a_i, x \rangle \leq b_i\}$ for some integers $a_i, b_i \in \mathbb{Z}$ with $|a_i| \leq M$ for each $i \in I$, then given prior distribution $q : \Omega \rightarrow [0, 1]$, expected value θ and error parameter $\varepsilon > 0$, there is an ε -approximate minimizer of $h_{\theta,q}$ in ℓ_1 ball of radius $R_\varepsilon := O \left(nM \log \left(\frac{|\Omega| \kappa n}{\varepsilon} \right) \right)$ where $\kappa := \max_{\alpha \in \Omega} q(\alpha) / \min_{\alpha \in \Omega} q(\alpha)$.*

Lemma 9.5 (Size of approximate solutions of dual maximum entropy program). *Given a finite domain $\Omega := \Omega_1 \times \dots \times \Omega_d \subseteq [0, 1]^n$ that corresponds to n -dimensional encodings of numerical, binary, and categorical d attributes, N samples from Ω , uniformly distributed mass, $C > 0$, target expected value θ , and $\varepsilon > 0$, if q_C denotes the corresponding Laplace smoothed prior distribution, then there exists an ε approximate minimizer of h_{θ,q_C} in ℓ_1 ball of radius $R_\varepsilon := O \left(nd \log \left(n^{\max_{i \in [d]} |\Omega_i| / C\varepsilon} \right) \right)$.*

Proof. We note that domain Ω is the Cartesian product of intervals and standard simplices by construction. Each interval can be represented by two inequality constraints of form $0 \leq \langle e_i, x \rangle \leq 1$ and simplices can be represented with the constraints of the form $\langle \sum_{i \in I} e_i, x \rangle = 1$ where e_i is the standard basis element for \mathbb{R}^n and $I \subseteq [n]$ is an index set. Thus, the parameter M in Theorem 9.4 is 1. We also note that, $|\Omega| = \prod_{i \in [d]} |\Omega_i| \leq (\max_{i \in [d]} |\Omega_i|)^d$. On the other hand, if the estimated prior distribution is q , then $q(\alpha) \leq 1$ for any $\alpha \in \Omega$ and $q(\alpha) \geq C/|\Omega|$ by construction. Therefore,

$$R_\varepsilon = O \left(nM \log \left(\frac{|\Omega| \kappa n}{\varepsilon} \right) \right) = O \left(nd \log \left(n^{\max_{i \in [d]} |\Omega_i| / C\varepsilon} \right) \right)$$

where $\kappa = \max_{\alpha \in \Omega} q(\alpha) / \min_{\alpha \in \Omega} q(\alpha)$. \square

We are now ready to prove Theorem 4.1. We recall it for completeness.

Theorem 9.6 (Fast second-order algorithm for dual-MaxEnt). *There exists an algorithm such that, given sets $\Omega_1, \dots, \Omega_d$, N samples from $\Omega \subseteq \mathbb{R}^n$, smoothing parameter $C > 0$ and error margin $\varepsilon > 0$, it can compute $\hat{\lambda} \in \mathbb{R}^n$ that is ε -approximate solution to dual-MaxEnt with $\tilde{O}\left(\left(N + cd + n^{\omega - \frac{3}{2}} \log\left(\frac{n}{\varepsilon}\right)\right) n^4 d \log\left(\frac{1}{C\varepsilon}\right)^2\right)$ arithmetic operations where $c := \max_{i \in [d]} |\Omega_i|$ and $\omega \sim 2.3 \dots$ is the matrix multiplication coefficient.*

Proof. Let q_C denotes the Laplace smoothed prior distribution. By Lemma 4.3, one can compute $S_{q_C, \lambda}[1]$, $S_{q_C, \lambda}[x]$, and $S_{q_C, \lambda}[xx^\top]$ with $O(n^2(N + d \max_{i \in [d]} |\Omega_i|))$ many arithmetic operations where $S_{q_C, \lambda}[f] := \sum_{\alpha \in \Omega} f(\alpha) \cdot q_C(\alpha) e^{\langle \lambda, \alpha \rangle}$. We note that given $\theta \in \text{conv}(\Omega)$, the gradient of h_{θ, q_C} is

$$\frac{S_{q_C, \lambda}[x]}{S_{q_C, \lambda}[1]} - \theta,$$

and the Hessian of h_{θ, q_C} is

$$\frac{S_{q_C, \lambda}[xx^\top]}{S_{q_C, \lambda}[1]} - \left(\frac{S_{q_C, \lambda}[x]}{S_{q_C, \lambda}[1]}\right) \left(\frac{S_{q_C, \lambda}[x]}{S_{q_C, \lambda}[1]}\right)^\top.$$

Thus, the computation of the gradient and Hessian of h_{θ, q_C} requires $O(n^2(N + d \max_{i \in [d]} |\Omega_i|))$ many arithmetic operations. We substitute the values of R_ε , $D_\Omega = O(n)$ and \mathcal{T}_H in Theorem 9.3, to get the following bound,

$$\tilde{O}\left(\left(N + cd + n^{\omega - \frac{3}{2}} \log\left(\frac{n}{\varepsilon}\right)\right) n^4 d \log\left(\frac{1}{C\varepsilon}\right)^2\right)$$

many arithmetic operations by Theorem 9.3. □

10 Price of fairness

The ideal choice of θ to be used in the max-entropy program is the expected value of the prior distribution q_C . However, we allow the user to vary the value of θ to ensure different statistical rates of the max-entropy distribution. In this section, we quantify and bound the change in the max-entropy distribution due to the change in the expected value θ . This quantity is referred to as the “price of fairness” in our framework. We use the recent work of [23] on the stability of max-entropy distributions to achieve a bound on the price of fairness. We often drop the superscript w from q_C^w for brevity of notation.

For the prior distribution q_C , let θ_C denote the expected value of this distribution. We have the following general bound on the price of fairness.

Theorem 10.1 (Distance between output and prior distribution in terms of input expected value). *Given samples $\mathcal{S} \subseteq \Omega \subseteq \mathbb{R}^n$ and $0 \leq C \leq 1$, let q_C be the prior distribution defined in Eq. (1). Let D_Ω denote the diameter of Ω , i.e. $D_\Omega = \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_2$. Let θ denote the desired expected value (taken as input to the framework) and let θ_C denote the expected value of distribution q_C .*

For every $\varepsilon \geq 0$, if $\|\theta - \theta_C\|_1 \leq \varepsilon$ then

$$\|p^\star - q_C\|_1 \leq \sqrt{R \cdot \varepsilon},$$

where p^\star is the max-entropy distribution obtained using prior distribution q_C and expected value θ and $R = O(n^{3/2} \cdot (\log \frac{|\Omega|}{C} + \log \frac{n D_\Omega}{\varepsilon}))$.

We will use the results of [23] to obtain the above bound. They prove the following general theorem regarding the stability of max-entropy distributions.

Theorem 10.2 (Stability of max-entropy distributions [23]). *Given a domain $\Omega \subseteq \mathbb{Z}^n$ and let R_Ω denote the ℓ_2 -diameter of Ω . Let $fc(\Omega)$ denote the unary facet complexity of $\text{conv}(\Omega)$. Then for every function $q : \Omega \rightarrow \mathbb{R}_0$ and for every $\varepsilon > 0$, there exists a number $R = O(n^{3/2} \cdot fc(\Omega) \cdot (L_q + \log |\Omega| + \log \frac{nR_\Omega}{\varepsilon}))$ such that for every $\theta_1, \theta_2 \in \text{conv}(\Omega)$, if $\|\theta_1 - \theta_2\|_1 \leq \varepsilon$ then*

$$\|p^{\theta_1} - p^{\theta_2}\|_1 \leq \sqrt{R \cdot \varepsilon},$$

where p^θ is the max-entropy distribution corresponding to prior q and expected value θ and $L_q := \max_{\alpha \in \Omega} |\log q(\alpha)|$.

Informally, the unary facet complexity of a polytope is the smallest number $M \in \mathbb{N}$ such that the polytope ($\text{conv}(\Omega)$ in this case) can be described by linear inequalities with coefficients in $\{-M, \dots, 0, \dots, M\}$. We refer the reader to [23] for a detailed discussion on unary facet complexity.

We will now use the above theorem to prove the price of fairness bound.

Proof of Theorem 10.1. The domain and the prior used in Theorem 10.1 satisfy the conditions of 10.2. Since the domain Ω has only binary or categorical attributes, the polytope can be described using linear inequalities with $O(1)$ coefficients. Therefore $fc(\Omega) = O(1)$.

We next compute $L_{q_C} := \max_{\alpha \in \Omega} |\log q_C(\alpha)|$. Since q_C is a probability distribution, its value is upper bounded by 1. To find the maximum of absolute value of log of q_C , we need a lower bound on the probability value. From the definition of q_C , we know that

$$q_C(\alpha) \geq \frac{C}{|\Omega|}.$$

Therefore,

$$L_{q_C} := \max_{\alpha \in \Omega} |\log q_C(\alpha)| \leq \log \frac{|\Omega|}{C}.$$

Since $C < 1$, $\log \frac{|\Omega|}{C} + \log |\Omega|$ is $O(\log \frac{|\Omega|}{C})$. Furthermore note that the max-entropy distribution with prior q_C and expected value θ_C is the distribution q_C itself. Substituting these values in Theorem 10.2, we get the bound mentioned in Theorem 10.1. \square

We are specifically interested in the case when the input specifies the desired statistical rate. Suppose that the input expected value θ has the same value as θ_C (expected value of q_C) for all indices other than I_Z , i.e., the sensitive attributes, and for the sensitive attribute, θ satisfies the condition $\theta(Z=z_1)/\theta(Z=z_2) \geq \tau$, for all values z_1, z_2 of any sensitive attribute Z .

Consider the case of a single binary sensitive attribute Z for which the input θ_1 satisfies the above condition for a given τ . It can be shown that in this case $\tau/2 \leq \theta(Z=1), \theta(Z=0) \leq 1/2\tau$. Therefore in this case $\|\theta - \theta_C\|_1 \leq \max\{|\theta_C(Z=1) - 1/2\tau|, |\theta_C(Z=1) - \tau/2|\} + \max\{|\theta_C(Z=0) - 1/2\tau|, |\theta_C(Z=0) - \tau/2|\}$. Using this bound, we can obtain a τ -dependent bound on the price of fairness in this case.

We can similarly obtain a τ -dependent bound on the price of fairness in the general case as a corollary of the above theorem.

Corollary 10.3 (τ -dependent price of fairness bound). *Let $\Omega \subset \mathbb{R}^n$ be the domain and let Ω_ℓ denote the domain of a sensitive attribute.*

Given samples $S \in \Omega \subset \mathbb{R}^n$ and $0 \leq C, \tau \leq 1$, let q_C be the prior distribution defined in Eq. (1). Let D_Ω denote the ℓ_2 -diameter of Ω . Let θ denote the desired expected value (taken as input to the framework) and let θ_C denote the expected value of distribution q_C . Suppose that θ and θ_C have the same value for all indices other than I_ℓ , i.e., the sensitive attributes and θ satisfies the condition $\theta(Z=z_1)/\theta(Z=z_2) \geq \tau$, for all values $z_1, z_2 \in \Omega_\ell$. Then

$$\|p^\star - q_C\|_1 \leq \sqrt{R \cdot M},$$

where p^\star is the max-entropy distribution obtained using prior distribution q_C and expected value θ , $R = O(n^{3/2} \cdot (\log \frac{|\Omega|}{C} + \log \frac{nD_\Omega}{\varepsilon}))$ and $M = \sum_{z \in \Omega_\ell} \max\{|\theta_C(Z=z) - 1/(\Omega_\ell \tau)|, |\theta_C(Z=z) - \tau/\Omega_\ell|\}$.

Proof. From the input $\frac{\theta(Z=z_1)}{\theta(Z=z_2)} \geq \tau$, for all values $z_1, z_2 \in \Omega_\ell$. Since $\sum_{z \in \Omega_\ell} \theta(Z=z) = 1$, we get that for all $z \in \Omega_\ell$,

$$\frac{\tau}{|\Omega_\ell|} \leq \theta(Z=z) \leq \frac{1}{\tau|\Omega_\ell|}.$$

Therefore, the difference between θ and θ_C is

$$\|\theta - \theta_C\|_1 \leq \sum_{z \in \Omega_\ell} \max\{|\theta_C(Z=z) - 1/(\Omega_\ell \tau)|, |\theta_C(Z=z) - \tau/\Omega_\ell|\}.$$

Substituting the above term in place of ε in Theorem 10.1 gives us the bound in the corollary. \square

11 Fairness guarantees - Proof of Theorem 4.5

We first recall our main theorem on fairness guarantees. Again, we drop the superscript w from q_C^w for brevity.

Theorem 11.1 (Fairness guarantees on the output distribution). *Given samples $\mathcal{S} \subseteq \Omega \subseteq \mathbb{R}^n$ and parameters $\tau, C \in [0, 1]$, let q_C be the prior distribution computed using the weights obtained from Algorithm 2. Let D_Ω denote the ℓ_2 -diameter of Ω , i.e., $D_\Omega = \max_{\alpha, \beta \in \Omega} \|\alpha - \beta\|_2$. Let θ denote the given expected vector and θ_C denote the expected vector corresponding to q_C . Let p^\star denote the max-entropy distribution obtained corresponding to the prior distribution q_C and expected value θ . Let Y denote the random variable when p^\star is restricted to \mathcal{Y} and let Z denote the random variable when p^\star is restricted to the domain of the protected attribute Ω_ℓ . Then, the max-entropy distribution p^\star satisfies the following bounds:*

- (Representation rate) If θ satisfies the property that $\theta(Z=z_i)/\theta(Z=z_j) \geq \tau$, then $\frac{p^\star(Z=z_i)}{p^\star(Z=z_j)} \geq \tau$.
- (Statistical rate) If θ satisfies the property that $\theta(Z=z_i)/\theta(Z=z_j) \geq \tau$, for all values $z_i, z_j \in \Omega_\ell$ and $\|\theta - \theta_C\|_1 \leq \varepsilon_\theta$, then for a fixed class label $y \in \mathcal{Y}$

$$\frac{p^\star[Y=y|Z=z_i]}{p^\star[Y=y|Z=z_j]} \geq \tau \left(1 - \frac{2\sqrt{R \cdot \varepsilon_\theta} \cdot |\Omega_\ell| \cdot |\mathcal{Y}|}{C - \sqrt{R \cdot \varepsilon_\theta} \cdot |\Omega_\ell| \cdot |\mathcal{Y}|} \right).$$

In all the above bounds, $R = O(n^{3/2} \cdot (\log \frac{|\Omega|}{C} + \log \frac{nD_\Omega}{\varepsilon_\theta}))$.

We will need the following lemma and Theorem 10.1 for the proof.

Lemma 11.2 (Joint distribution is fair in case of weights obtained from Algorithm 2). *Let q_C be the prior distribution computed according to Eq. (1) and using the weights obtained from Algorithm 2. Let Y, Z denote the random variables for class label and sensitive attribute respectively. Then for a given $y \in \mathcal{Y}$ and for all values $z_1, z_2, \dots, z_k \in \Omega_\ell$,*

$$q_C(Z = z_1, Y = y) = q_C(Z = z_2, Y = y) = \dots = q_C(Z = z_k, Y = y).$$

Proof of Lemma 11.2. For any value $z_i \in \Omega_\ell$,

$$\begin{aligned}
q_C(Z = z_i, Y = y) &= q_C(Z = z_i, Y = y) \\
&= \sum_{\alpha \in \Omega \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) \\
&= \left(\sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) + \sum_{\alpha \in \Omega \setminus S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) \right).
\end{aligned}$$

We analyze each of the two parts individually. From Algorithm 2,

$$\begin{aligned}
\sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) &= \sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} \frac{C}{|\Omega|} + \frac{1-C}{W} \frac{c(y)}{c(y, z_i)} \\
&= \frac{C}{|\Omega|} c(y, z_i) + \frac{1-C}{W} c(y, z_i) \frac{c(y)}{c(y, z_i)} \\
&= \frac{C}{|\Omega|} c(y, z_i) + \frac{(1-C) \cdot c(y)}{W}. \\
\sum_{\alpha \notin S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) &= \sum_{\alpha \notin S \mid y(\alpha)=y, z(\alpha)=z_i} \frac{C}{|\Omega|} \\
&= \frac{C}{|\Omega|} c'(y, z_i).
\end{aligned}$$

where $c'(y, z_i)$ is the number of elements not in S with class label y and sensitive attribute z_i . Therefore,

$$\begin{aligned}
\sum_{\alpha \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) &= \sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) + \sum_{\alpha \notin S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) \\
&= \frac{C}{|\Omega|} c(y, z_i) + \frac{(1-C) \cdot c(y)}{W} + \frac{C}{|\Omega|} c'(y, z_i) \\
&= \frac{C}{|\Omega|} (c(y, z_i) + c'(y, z_i)) + \frac{(1-C) \cdot c(y)}{W} \\
&= \frac{C}{|\mathcal{Y}| \cdot |\Omega_\ell|} + \frac{(1-C) \cdot c(y)}{W}
\end{aligned}$$

Note that first part of the above term is independent of y and z_i as it counts all possible elements with class label y and sensitive attribute z_i . Hence $q_C(Z = z_i, Y = y)$ is independent of the sensitive attribute z_i . □

We are now ready to prove Theorem 11.1.

Proof of Theorem 11.1. (1) The first bound follows from the definition of max-entropy distribution and the constraints in the program.

(2) For the second bound, suppose we are given that $\|p^* - q_C\|_1 \leq D$. The total variation distance between two distributions is defined as

$$\delta_{TV}(p, q) = \max_{\Omega' \subseteq \Omega} |p(\Omega') - q(\Omega')|.$$

We also know that

$$\delta_{TV}(p^*, q_C) = \frac{1}{2} \|p^* - q_C\|_1 \leq \frac{D}{2}.$$

Let Ω' be the domain of all point $\alpha \in \Omega$ with class label y and sensitive attribute z . Then

$$q_C(Y = y, Z = z) - D \leq p^*(Y = y, Z = z) \leq q_C(Y = y, Z = z) + D.$$

From Lemma 11.2, we know that $q_C(Z = z, Y = y)$ is equal for a fixed y and all z . Therefore, for any two values z_1, z_2

$$|p^*(Z = z_1, Y = y) - p^*(Z = z_2, Y = y)| \leq 2D.$$

We can bound the ratio of these quantities as

$$\frac{p^*(Z = z_1, Y = y)}{p^*(Z = z_2, Y = y)} \geq 1 - \frac{2D}{\max_z p^*(Y = y, Z = z)}.$$

We next attempt to lower bound $\max_z p^*(Y = y, Z = z)$. Once again

$$\begin{aligned} p^*(Z = z, Y = y) &\geq q_C(Y = y, Z = z) - D \\ &\geq \sum_{\alpha | y(\alpha)=y, z(\alpha)=z} q_C(\alpha) - D \\ &\geq \sum_{\alpha | y(\alpha)=y, z(\alpha)=z} \frac{C}{|\Omega|} - D \\ &= \frac{C}{|\Omega|} \cdot \frac{|\Omega|}{|\Omega_\ell| \cdot |\mathcal{Y}|} - D = \frac{C}{|\Omega_\ell| \cdot |\mathcal{Y}|} - D. \end{aligned}$$

Therefore,

$$\frac{p^*(Z = z_1, Y = y)}{p^*(Z = z_2, Y = y)} \geq 1 - \frac{2D \cdot |\Omega_\ell| \cdot |\mathcal{Y}|}{C - D \cdot |\Omega_\ell| \cdot |\mathcal{Y}|},$$

and if $\theta(Z=z_i)/\theta(Z=z_j) \geq \tau$, for all values z_i, z_j of Z , then we get

$$\frac{p^*[Y = y | Z = z_i]}{p^*[Y = y | Z = z_j]} \geq \tau \left(1 - \frac{2D \cdot |\Omega_\ell| \cdot |\mathcal{Y}|}{C - D \cdot |\Omega_\ell| \cdot |\mathcal{Y}|} \right).$$

Using $\|\theta - \theta_C\|_1 \leq \varepsilon_\theta$, by plugging in the upper bound on D from Theorem 10.1 we get the statistical rate bound in the theorem. □

12 Using uniform weights in the prior distribution

As mentioned earlier, using uniform weights $w_i = 1/N$, for all $i \in [N]$, does not lead to a fair prior distribution. We formalize this argument with the lemma stated below.

Lemma 12.1 (Joint distribution is not fair in case of uniform weights). *Let q_C be the prior distribution computed using the uniform weights, i.e., $w_i = 1/N$, for all $i \in [N]$. Let Y, Z denote the random variables for class label and sensitive attribute respectively. Then for a given $y \in \mathcal{Y}$ and $z \in \Omega_\ell$,*

$$q_C(Z = z, Y = y) = \frac{C}{|\Omega_\ell| |\mathcal{Y}|} + \frac{(1 - C) \cdot c(y, z)}{N},$$

where $c(y, z)$ is the count of the number of samples in S with class label y and sensitive attribute z .

Proof. For any value z_i of Z ,

$$\begin{aligned}
q_C(Z = z_i, Y = y) &= q_C(Z = z_i, Y = y) \\
&= \sum_{\alpha \in \Omega \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) \\
&= \left(\sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) + \sum_{\alpha \in \Omega \setminus S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) \right).
\end{aligned}$$

We analyze each of the two parts individually. From Algorithm 2,

$$\begin{aligned}
\sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) &= \sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} \frac{C}{|\Omega|} + \frac{1-C}{N} \\
&= \frac{C}{|\Omega|} c(y, z_i) + \frac{1-C}{N} c(y, z_i).
\end{aligned}$$

$$\begin{aligned}
\sum_{\alpha \notin S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) &= \sum_{\alpha \notin S \mid y(\alpha)=y, z(\alpha)=z_i} \frac{C}{|\Omega|} \\
&= \frac{C}{|\Omega|} c'(y, z_i).
\end{aligned}$$

where $c'(y, z_i)$ is the number of elements not in S with class label y and sensitive attribute z_i . Therefore,

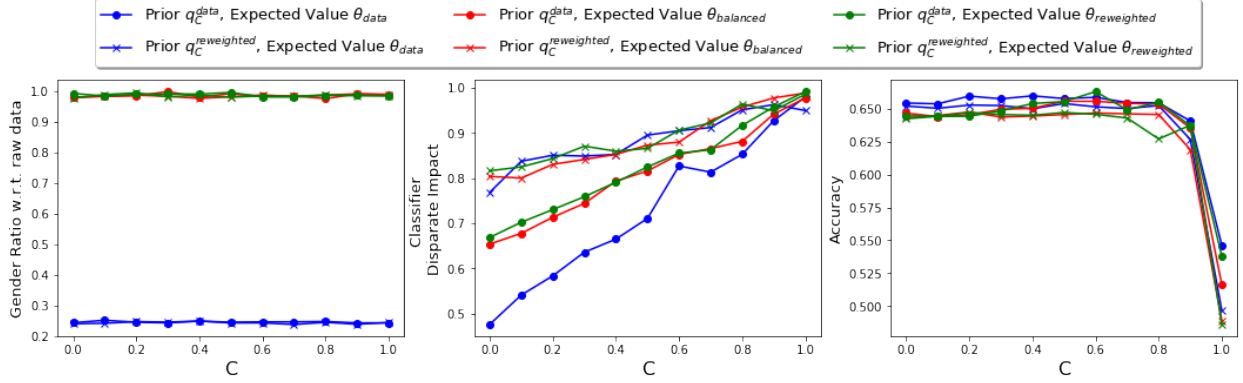
$$\begin{aligned}
\sum_{\alpha \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) &= \sum_{\alpha \in S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) + \sum_{\alpha \notin S \mid y(\alpha)=y, z(\alpha)=z_i} q_C(\alpha) \\
&= \frac{C}{|\Omega|} c(y, z_i) + \frac{1-C}{N} c(y, z_i) + \frac{C}{|\Omega|} c'(y, z_i) \\
&= \frac{C}{|\Omega|} (c(y, z_i) + c'(y, z_i)) + \frac{1-C}{N} c(y, z_i) \\
&= \frac{C}{|\mathcal{Y}| \cdot |\Omega_\ell|} + \frac{1-C}{N} c(y, z_i).
\end{aligned}$$

□

The above lemma shows that the joint probability for a given y and z depends on the number of samples with those values in S when the weights are uniform and hence is not fair in the general case.

13 Other experiments on smaller COMPAS dataset

Figure 2 shows the variation of gender ratio and Gaussian Naive Bayes classifier disparate impact and accuracy, when trained using different max-entropy distributions.



(a) Gender Ratio vs C. (b) Classifier Disparate Impact vs C. (c) Classifier Accuracy vs C.

Figure 2: The figures show the comparison of max-entropy distributions with different prior distributions and expected values. The first figure show the gender ratio of different max-entropy distribution. The second and third figure show the disparate impact and accuracy of Gaussian Naive Bayes classifier trained on the output distribution.

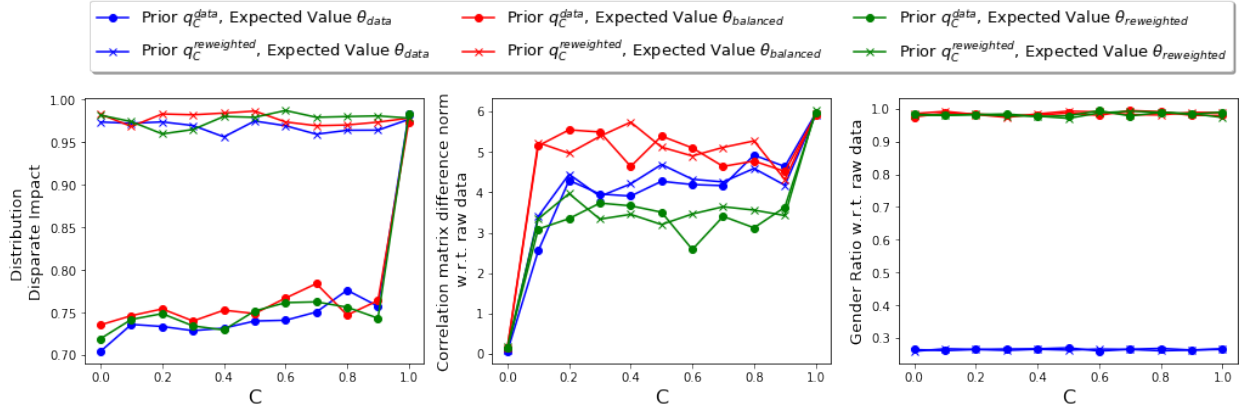


Figure 3: Comparison of disparate impact, gender ratio and KL-divergence from empirical distribution of raw data for max-entropy distributions with different priors and expected values.

14 Experiments on larger COMPAS dataset

In this section, we present the experiments on the larger version of the COMPAS dataset. This dataset consists of attributes sex, race, age, juvenile felony count, juvenile misdemeanor count, juvenile other count, months in jail, priors count, decile score, charge degree, violent crime, violent recidivism, drug related crime, firearm involved, minor involved, road safety hazard, sex offense, fraud and petty crime, with recidivism as the label. We did not exclude any samples and we did not categorize any attributes. The original data contains samples from 6 different races whose age ranged from 18 to 96 with at most 40 prior counts, juvenile felony count, juvenile misdemeanor count, and juvenile other count.

Thus, we model the domain Ω for this version as $\{0, 1\}^8 \times \{0, 1, 2\}^3 \times \{0, 1, \dots, 5\} \times \Delta_6 \times \{0, 1, \dots, 7\}^2 \times \{0, 1, \dots, 10\}^2 \times \{0, 1, \dots, 11\} \times \{0, 1, \dots, 13\}$. Overall the domain contains approximately 1.4×10^{11} different points.

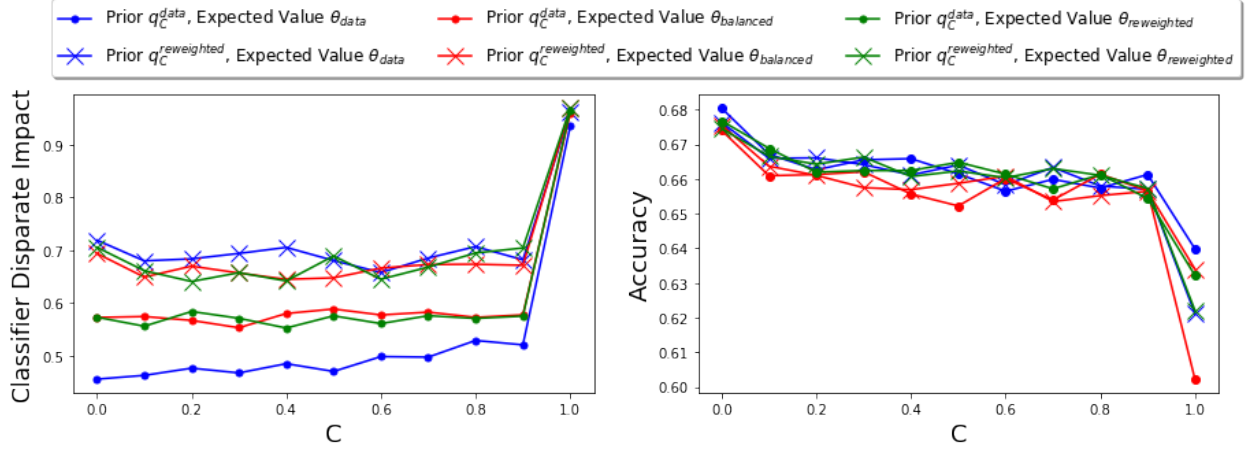


Figure 4: Comparison of Gaussian Naive Bayes classifier trained on data from max-entropy distributions with different prior distributions and expected values.

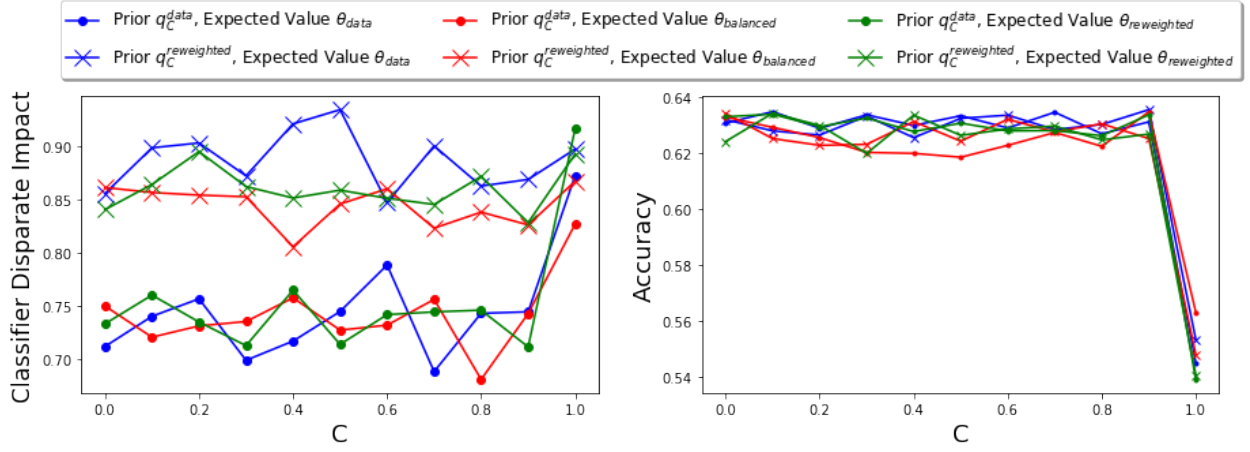


Figure 5: Comparison of Decision Tree classifier trained on data from different max-entropy distributions with different prior distributions and expected values.

14.1 Evaluating the disparate impact and accuracy of generated dataset

We evaluate the dataset generated using different max-entropy algorithms. We run the algorithm with different combinations of prior weights and expected value mentioned earlier. We vary the C value for our framework and measure the disparate impact of the output distribution. We also consider a metric to check how well the max-entropy distribution preserves the pairwise correlation between features. To calculate this, we first calculate the covariance matrix of the output dataset, say $\text{Cov}_{\text{output}}$ and the original raw dataset Cov_{data} , and then report the Frobenius norm of the difference of these matrices, i.e., $\|\text{Cov}_{\text{output}} - \text{Cov}_{\text{data}}\|_F^2$. The lower the value of the norm, the better the output distribution preserves the pairwise correlation. The results for this evaluation are present in Figure 3. Here again the first part of the figure shows that the max-entropy distributions obtained using prior q_C^{rewt} and expected value θ_{rewt} or θ_{bal} achieve higher disparate impact values than the distributions obtained from max-entropy distribution obtained using uniform weights on samples. Similarly the gender ratio of max-entropy distributions using prior distribution q_C^{rewt} and expected value θ_{rewt} or θ_{bal} are close to 1.0.

Table 2: Comparison of max-entropy distribution on large COMPAS dataset with the state-of-the-art. The max-entropy distribution uses parameter $C = 0.5$, prior q_C^{rewt} and expected value θ_{rewt} or θ_{bal} . "DI" denotes Disparate Impact. The first column shows that the max-entropy distribution has high DI value. The gender ratio of the max-entropy distribution dataset is also much higher than other output dataset, due to the balanced marginals used. The DI of the decision tree classifier is high for the max-entropy dataset as well.

Data source	Data DI	Correlation matrix difference norm	Gender Ratio	Classifier DI	Classifier Accuracy
Raw Data	0.71	0	0.24	0.71	0.66
Max-Entropy Distribution using $q_C^{\text{rewt}}, \theta_{\text{rewt}}$	0.98	3.2	0.97	0.85	0.63
Max-Entropy Distribution using $q_C^{\text{rewt}}, \theta_{\text{bal}}$	0.99	5.12	0.99	0.86	0.62
Re-weighting samples [14]	1.0	0.14	0.26	0.79	0.62

14.2 Evaluating the disparate impact and accuracy of classifier trained on generated dataset

As mentioned earlier, we use the generated datasets to train a Gaussian Naive Bayes and the Decision Tree Classifier and evaluate the fairness and the accuracy of the resulting classifier.

Firstly, we again vary the C value for our framework and measure the disparate impact of the output of the classifier as well as the accuracy. The results for this evaluation using Gaussian Naive Bayes are present in Figure 4 and using Decision Tree Classifier are present in Figure 5. As expected, once again the the max-entropy distributions obtained using prior distribution q_C^{rewt} achieve higher disparate impact values than the distributions obtained from max-entropy distribution obtained using uniform weights on samples. The accuracy also drops as the value of C tends to 1. This is again because the prior distribution in case of $C = 1$ assigns equal probability mass to all points in the domain.

14.3 Comparison with other algorithms

We compare the performance of the max-entropy distributions with the raw data and re-weighting algorithm [14]. The results are presented in Table 2. From the table, one can see that the correlation matrix norm difference is smallest for the max-entropy distribution. At the same time the data and classifier disparate impact are high.

References

- [1] Zeyuan Allen Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing, Los Angeles, CA, USA, June 25-29, 2018*.
- [2] Zeyuan Allen Zhu, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Much faster algorithms for matrix scaling. In *FOCS'17: Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science, 2017*.
- [3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. COMPAS recidivism risk score data and analysis, 2016.

- [4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.
- [5] Toon Calders and Indrė Žliobaitė. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, pages 43–57. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
- [6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.
- [7] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328. ACM, 2019.
- [8] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 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*, pages 286–305, 2017.
- [9] Michael B. Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. 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.
- [10] Miroslav Dudik. Maximum entropy density estimation and modeling geographic distributions of species, 2007.
- [11] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323, 2016.
- [12] Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, May 1957.
- [13] Edwin T. Jaynes. Information theory and statistical mechanics. II. *Physical Review*, 108:171–190, October 1957.
- [14] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication, 2009. IC4 2009.*, pages 1–6. IEEE, 2009.
- [15] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, editors, *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pages 924–929. IEEE Computer Society, 2012.
- [16] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*, volume 7524 of *Lecture Notes in Computer Science*, pages 35–50. Springer, 2012.

- [17] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 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, pages 3819–3828. ACM, 2015.
- [18] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica* (5 2016), 9, 2016.
- [19] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.
- [20] Cathy O’Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown/Archetype, 2016.
- [21] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness GAN. *arXiv preprint arXiv:1805.09910*, 2018.
- [22] Mohit Singh and Nisheeth K Vishnoi. Entropy, optimization and counting. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 50–59. ACM, 2014.
- [23] Damian Straszak and Nisheeth K Vishnoi. Maximum entropy distributions: Bit complexity and stability. *COLT*, 2019.
- [24] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.
- [25] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.