Bayesian Fairness

Christos Dimitrakakis,[1,2] **Yang Liu,**[3] **David C. Parkes,**[4] **Goran Radanovic**[4]
[1]University of Oslo; [2]Chalmers; [3]University of California, Santa Cruz; [4]Harvard
christos.dimitrakakis@gmail.com, yangliu@ucsc.edu, jparkes@eecs.harvard.edu, gradanovic@g.harvard.edu

Abstract

We consider the problem of how decision making can be fair when the underlying probabilistic model of the world is not known with certainty. We argue that recent notions of fairness in machine learning need to explicitly incorporate parameter uncertainty, hence we introduce the notion of *Bayesian fairness* as a suitable candidate for fair decision rules. Using *balance*, a definition of fairness introduced in (Kleinberg, Mullainathan, and Raghavan 2016), we show how a Bayesian perspective can lead to well-performing and fair decision rules even under high uncertainty.

Introduction

Fairness is an important property of algorithmic systems in settings where decisions are made that affect individuals in a population, for example in the context of loan decisions, college admissions, hiring decision, or bail decisions.

Recognizing this, there has been considerable emphasis in recent work on developing definitions of fairness in the context of machine learning algorithms. In this paper, we take a closer look at informational aspects of fairness. In particular, by adopting a Bayesian viewpoint, we explicitly take into account model uncertainty, something that turns out to be crucial for fairness.

Uncertainty about the underlying probabilistic model of the world has two main effects. Firstly, many notions of fairness have been defined with respect to latent variables, including model parameters. This means that we need to take into account uncertainty about these latent variables and parameters. Secondly, in many problems our decisions determine the data that we will collect in the future. Ignoring uncertainty may magnify subtle biases in our model.

By viewing fairness through a Bayesian perspective, we avoid these problems. In particular, we demonstrate that Bayesian policies can allow for suitable trade offs to be made between utility and fairness, taking into account uncertainty about model parameters.

We consider a setting where a *decision maker* (DM) makes a sequence of decisions through some chosen *policy* π to maximize her *expected utility* u. However, the DM must trade off utility with some *fairness criterion* f. We assume the existence of some underlying probability law P, so that the decision problem, when P is known, can be written as:

$$\max_{\pi} (1-\lambda)\, \mathbb{E}_P^{\pi}\, u - \lambda\, \mathbb{E}_P^{\pi}\, f, \tag{1}$$

where λ is the DM's trade-off between fairness and utility.[1] We adopt a Bayesian viewpoint and assume the DM has *belief* β over some family of distributions $\mathscr{P} \triangleq \{ P_\theta \mid \theta \in \Theta \}$, which may contain the actual law, i.e. $P_{\theta^*} = P$ for some θ^*.

The DM's policy π defines the actions $a_t \in \mathcal{A}$ the DM takes at different (discrete) times t depending on the available information. More precisely, at time t the DM observes some *data* $x_t \in \mathcal{X}$, and depending on her belief β_t makes a *decision* $a_t \in \mathcal{A}$, so that $\pi(a_t \mid \beta_t, x_t)$ defines a probability over actions for every possible belief and observation. The DM has a utility function, modeled here with structure $u : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y}$ is a set of *outcomes* (in a loan setting, was the loan repaid on time?). The fairness concept we focus on is a Bayesian version of *balance* (Kleinberg, Mullainathan, and Raghavan 2016), which is also a generalization of the equality of opportunity (Hardt, Price, and Srebro 2016).

The amount of uncertainty about the model parameters directly influences the interpretation of the balance condition. Informally, the more uncertain we are, the more stochastic the decision rule will need to be.

Our contributions. In this paper, we develop a Bayesian framework for fairness that recognizes that there can be a high degree of uncertainty about model parameters and latent variables, and especially when not a lot of data has been collected, or in sequential settings. In particular, we propose that the DM should take into account how unfair she would be under all possible models, weighted by their probability. Fairness is a property of the decision rule with respect to the true model, and it is this that is used to *measure* fairness. On the other hand, the appropriate way to achieve fairness depends on the DM's information, and it is this that is used to derive *algorithms*. In order to work without model approximations, we illustrate the approach in a simple setting. We show that the policies that are obtained are qualitatively and

[1]We do not consider the alternative constrained problem, i.e. $\max \{ \mathbb{E}_P^{\pi}\, u \mid \mathbb{E}_P^{\pi}\, f \leq \epsilon \}$, in the present paper.

quantitatively different when we consider uncertainty and adopt a Bayesian viewpoint in comparison to when we do not.

Given that the Bayesian approach to fairness takes into account uncertainty and makes explicit consideration of the DM's information, we can also use the approach to select policies that influence the data we collect, and thus our knowledge about the model. This is an important informational feedback effect, and one that a Bayesian methodology can provide in a principled way. We provide experimental results on the COMPAS dataset (Larson et al. 2016) as well as artificial data, showing the robustness of the Bayesian approach, and comparing against methods that define fairness measures according to a single, marginalized model (e.g. (Hardt, Price, and Srebro 2016)). While we mainly treat the non-sequential setting, where the data is fixed, we can also accommodate sequential, bandits-style settings, as explained in later sections. The results there provide a vivid illustration of what can go wrong with a certainty-equivalent approach to achieving fairness.

All missing proofs and details can be found in our supplementary materials.

Related work. Algorithmic fairness has been studied quite extensively in recent work. But we are not aware of work that adopts a Bayesian perspective. For instance, (Dwork et al. 2012; Chouldechova 2016; Corbett-Davies et al. 2017; Kleinberg, Mullainathan, and Raghavan 2016; Kilbertus et al. 2017) studied fairness under a setting where the model is known. (Corbett-Davies et al. 2017) have considered how to satisfy fairness considerations while also maximizing expected utility. In this paper, we focus on notions of fairness related to notions of conditional independence, the specifics of which are discussed in the next section.

(Dwork et al. 2012) consider an individual-fairness approach, and look for decision rules that are smooth in a sense that similar individuals are treated similarly.

The recent work of (Russell et al. 2017) considers the problem of uncertainty from the point of view of causal modeling, with the three main differences to the present work being: (a) they consider a PAC-like setting, rather than the Bayesian framework; (b) we show that the effect of uncertainty remains important even without varying the counterfactual assumptions; and (c) the Bayesian framework easily admits a sequential setting. (Jabbari et al. 2016) and (Joseph et al. 2016) study fairness in sequential decision making settings, but not from a Bayesian viewpoint.

There is also research on questions of fairness in other machine learning contexts, such as clustering (Chierichetti et al. 2017), natural language processing (Blodgett and O'Connor 2017) and recommendation systems (Celis and Vishnoi 2017).

Preliminaries

(Chouldechova 2016) considers the problem of fair prediction with disparate impact. She defines an action (a "statistic" in her paper) a as *test-fair* with respect to the *outcome* y and *sensitive variable* z if y is independent of z under the action and parameter θ, i.e. if $y \perp\!\!\!\perp z \mid a, \theta$. While the author does not explicitly discuss the distribution P_θ, it is implicitly assumed to be that of the true model. We slightly generalize the definition of disparate impact as follows:

Definition 1 (Calibrated decision rule)**.** A decision rule $\pi(a \mid x)$ is *calibrated* with respect to some distribution P_θ if y, z are independent for all actions a taken, i.e. if

$$P_\theta^\pi(y, z \mid a) = P_\theta^\pi(y \mid a) P_\theta^\pi(z \mid a), \tag{2}$$

where P_θ^π is the distribution induced by P_θ and the decision rule π.

(Kleinberg, Mullainathan, and Raghavan 2016) also consider two balance conditions (one for each label class), which we re-interpret as follows. Here, we simplify the notation of the decision rule so that $\pi(a \mid x)$ corresponds to the probability of taking action a given observation x.

Definition 2 (Balanced decision rule)**.** A decision rule $\pi(a \mid x)$ is *balanced* with respect to some distribution P_θ if a, z are independent for all y, i.e. if

$$P_\theta^\pi(a, z \mid y) = P_\theta^\pi(a \mid y) P_\theta^\pi(z \mid y), \tag{3}$$

where P_θ^π is the distribution induced by P_θ and the decision rule π.

As with (Chouldechova 2016), (Kleinberg, Mullainathan, and Raghavan 2016) also work with the true model. We will slightly generalize the definition, stating balance with respect to any model parameter.

It is known that calibration and balance cannot be achieved simultaneously for non-trivial environments (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2016). This is also true for our more general definitions, as we show in Theorem S1 in the Supplementary material.

From a practitioner's perspective, we must choose either calibration or balance. We work with a generalized version of the balance condition, because balance gracefully extends to settings with uncertainty. In particular, balance involves equality in the expectation of a score function (when writing the probabilities as the expectations of a 0-1 indicator function; also depending on an observation x) under different values of a sensitive variable z, conditioned on the true (but latent) outcome y. Consequently, balance can always be satisfied—by using a randomized decision rule that is independent of x. This is not the case for the calibration condition under model uncertainty, because calibration criteria depends highly on the details of a model.

Bayesian Formulation

We first introduce a concrete, statistical decision problem. The true (latent) outcome y is generated independently of the DM's decision, with a probability distribution that depends on the available information x. There also exists a sensitive attribute variable z, which may be dependent on x.[2]

[2]Depending on the application scenario, z may actually be a subset of x and thus directly observable, while in other scenarios it may be latent. Here we focus on the case where z is not directly observed.

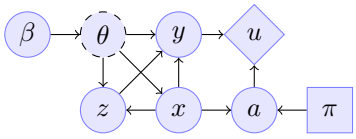


Figure 1: A Bayesian decision problem with observations x, outcome y, action a, sensitive variable z, utility u, unknown parameter θ, belief β and policy π. The joint distribution of x, y, z is fully determined by the unknown parameter θ, while the conditional distribution of actions a given observations x is given by the selected policy π. The DM's utility function is u, while the fairness of the policy depends on the problem parameters.

Definition 3 (Statistical decision problem)**.** See Figure 1 for the decision diagram. The DM observes $x \in \mathcal{X}$, then takes a decision $a \in \mathcal{A}$ and obtains utility $u(y, a)$ depending on a true (latent) outcome $y \in \mathcal{Y}$ generated from some distribution $P_\theta(y \mid x)$. The DM has a belief $\beta \in \mathcal{B}$ in the form of a probability distribution on parameters $\theta \in \Theta$ on a family $\mathscr{P} \triangleq \{ P_\theta(y \mid x) \mid \theta \in \Theta \}$ of distributions. In the Bayesian case, the belief β is a posterior formed through a prior and available data. The DM has a utility function $u : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, with utility depending on the DM's action and the outcome.

For simplicity, we will assume that $\mathcal{X}$, $\mathcal{A}$, and $\mathcal{Y}$, are finite sets, whereas Θ is a subset of $\mathbb{R}^n$. We focus on Bayesian decision rules, i.e. rules whose decisions depend upon a posterior belief β. The Bayes-optimal decision rule, ignoring fairness, is defined below.

Definition 4 (Bayes-optimal decision rule)**.** The *Bayes-optimal decision rule* $\pi^* : \mathcal{B} \times \mathcal{X} \to \mathcal{A}$ is a deterministic policy that maximizes the utility in expectation, i.e. takes action $\pi^*(\beta, x) \in \arg\max_{a \in \mathcal{A}} u_\beta(a \mid x)$, with $u_\beta(a \mid x) \triangleq \sum_y u(y, a)\, \mathbb{P}_\beta(y \mid x)$, where $\mathbb{P}_\beta(y \mid x) \triangleq \int_\Theta P_\theta(y \mid x)\, \mathrm{d}\beta(\theta)$ is the marginal distribution over outcomes conditional on the observations according to the DM's belief β.

The Bayes-optimal decision rule does not directly depend on the sensitive variable z. We are interested in settings with multiple time periods. At time t, the DM observes x_t and makes a decision a_t using policy π_t and obtains some instantaneous payoff $U_t = u(y_t, a_t)$ and fairness violation F_t. The DM's utility is the sum of instantaneous payoffs over time, $U \triangleq \sum_{t=1}^{T} u(y_t, a_t)$ and she is interested in finding a policy maximising U in expectation.

Although the Bayes-optimal decision rule brings the highest expected reward to the DM, it may be unfair. In the sequel, we will define analogs of the *balance* notion of fairness in terms of decision rules π, and investigate appropriate decision rules, that possibly result in randomized policies. In particular, we shall consider a utility function that combines the DM's utility with the societal benefit that comes from fairness, and search for Bayes-optimal decision rules with respect to this new, combined utility.

In particular, we define a Bayesian analogue of the maximization problem (1) as:

$$\begin{aligned} &\max_\pi (1-\lambda)\, \mathbb{E}^\pi_\beta\, u - \lambda\, \mathbb{E}^\pi_\beta\, f \\ &= \max_\pi \int_\Theta \left[(1-\lambda)\, \mathbb{E}^\pi_\theta\, u - \lambda\, \mathbb{E}^\pi_\theta\, f \right] \mathrm{d}\beta(\theta). \end{aligned} \tag{4}$$

To make this concrete, in the sequel we shall define the appropriate Bayesian version of the balance condition.

Bayesian Balance

In the Bayesian setting, we would like our decisions to take into account their impact on all possible models. That is, fairness is measured with respect to the true model.

It turns out that sometimes only a trivial decision rule can satisfy a strong form of balance in a setting with model uncertainty. In particular, what if we insist that balance must hold exactly, for all possible model parameters?

Theorem 1. *A trivial decision rule of the form $\pi(a \mid x) = p_a$ can always satisfy balance for a Bayesian decision problem. However, it may be the only balanced decision rule, even when a non-trivial balanced policy can be found for every possible $\theta \in \Theta$.*

The proof, as well as an example illustrating this result, are in the supplementary materials.

For this reason, we consider the the p-norm of the deviation from fairness with respect to our belief β:

Definition 5 (Bayesian Balance)**.** We say that a decision rule π is (α, p)-*Bayes-balanced with respect to belief* β if:

$$\begin{aligned} f(\pi) \triangleq \int_\Theta \sum_{a,y,z} \Bigg| \sum_x \pi(a|x) [P_\theta(x, z|y) \\ - P_\theta(x|y) P_\theta(z|y)] \Bigg|^p \mathrm{d}\beta(\theta) \leq \alpha^p. \end{aligned} \tag{5}$$

This definition captures the expected deviation from balance of policy π, for a Bayesian DM under their belief β. It measures the deviation of policy π from perfect balance with respect to each possible parameter θ, and weighs this deviation according to the probability of that model. This provides a graceful trade-off between achieving near-balance in the most likely models, while avoiding extreme unfairness in less likely ones.

Why not use a single point estimate for the model, instead of the full Bayesian approach? This would entail simply measuring balance (and utility) with respect to the marginal model, $\mathbb{P}_\beta \triangleq \int_\Theta P_\theta\, \mathrm{d}\beta(\theta)$.

Definition 6 (Marginal balance)**.** A decision rule $\pi(\cdot)$ is (α, p)-*marginal-Balanced with respect to belief* β if $\forall a, y, z$:

$$\sum_{a,y,z} \left| \sum_x \pi(a|x) \left[\mathbb{P}_\beta(x, z|y) - \mathbb{P}_\beta(x|y)\, \mathbb{P}_\beta(z|y) \right] \right|^p \leq \alpha. \tag{6}$$

One problem with this definition, which we will see in our experimental results, is that the decision policy may be very unfair towards other, high-probability models that are different from the marginal model.

Still, both balance conditions can provide a bound on balance with respect to the true model. For this, denote the true underlying model as θ^*, and define the (ϵ, δ)-accurate belief.

Definition 7. We call $\beta(\theta)$ an (ϵ, δ)-*accurate belief with respect to the true model* $\theta^* \in \Theta$, if with β-probability at least $1 - \delta$, $\forall x, y, z$:

$$|P_\theta(x|y,z) - P_{\theta^*}(x|y,z)| \leq \epsilon,\ |P_\theta(x|y) - P_{\theta^*}(x|y)| \leq \epsilon,$$

i.e. the set Θ_ϵ for which the above conditions hold has measure $\beta(\Theta_\epsilon) \geq 1 - \delta$.

Under some conditions, the balance achieved through either definition provides an approximation to balance under the true model, as shown by the following theorem.

Theorem 2. *If a decision rule satisfies either* $(\alpha, 1)$-*marginal-balance or* $(\alpha, 1)$-*Bayes-balance for* β *or both, and* β *is* (ϵ, δ)-*accurate, then the resulting decision rule is a*

$$(\alpha + 2|\mathcal{A}| \cdot |\mathcal{Z}| \cdot |\mathcal{Y}| \cdot (\epsilon + \delta), 1)\text{-}balanced$$

decision rule w.r.t. the true model θ^*.

This theorem says that if our belief β is concentrated around the true model P_{θ^*}, and our decision rule is fair with respect to either definition, then it is also fair with respect to the true model.

The Sequential setting

We can also extend the approach to a sequential setting, where the information learned by the DM about the environment depends on the action.

For example, if we approve a loan, we will only later discover if the loan is paid off on time. This information will in turn affect our future decisions. Analogous to other sequential decision making problems such as Markov decision processes (Puterman 1994), we need to solve the following optimization problem over a time horizon T:

$$\max_{\pi} \mathbb{E}_{\beta_1} \left[\sum_{t=1}^{T} (1-\lambda) U_t - \lambda F_t \right], \tag{7}$$

where π now must explicitly map future beliefs β_t to probabilities over actions. If the data that the DM obtains depends on her decisions a_t, then she must consider adaptive policies, as the next belief depends on the data obtained by the policy.

We can reformulate the maximization problem so as to explicitly include the future changes in belief:

$$\begin{aligned} V^*(\beta_t) \triangleq \sup_{\pi_t} & \mathbb{E}^{\pi_t}_{\beta_t} \left[(1-\lambda) U_t - \lambda F_t \right] \\ & + \sum_{\beta_{t+1}} V^*(\beta_{t+1})\, \mathbb{P}^{\pi_t}_{\beta_t}(\beta_{t+1}), \end{aligned} \tag{8}$$

under the mild assumption that the set of reachable next beliefs is finite (easily satisfied when the set of outcomes is finite). This now features the tradeoff between explore (obtaining new knowledge) and exploit (maximizing utility).

However, just as in the bandits case (c.f. Duff 2002), the above computation is intractable, as the policy space is exponential in T. For this reason, in this paper we only consider *myopic policies* that select a policy (and decision) that is optimal for the current step t, trading utility and fairness as well as the value of the information at any particular single step. A specific instance of this type of sequential version of the problem is a later section.

Algorithms

We compare the Bayesian framework with the simpler, marginal-model approach. In particular, for the Bayesian framework, we directly optimize (4). Using the marginal simplification, we maximize (1) with respect to the marginal model $\mathbb{P}_\beta$.

Balance gradient descent

We have a family of models $\{ P_\theta \}$ with a corresponding subjective distribution $\beta(\theta)$. In order to derive algorithms, we shall focus on the quantity:

$$C(\pi, \theta) \triangleq \sum_{y,z} \Big\| \sum_{x} \pi(a \mid x) \Delta_\theta(x, y, z) \Big\|_p, \tag{9}$$

This isthe deviation from balance for decision rule π under parameter θ, where

$$\Delta_\theta(x, y, z) \triangleq P_\theta(x, z \mid y) - P_\theta(x \mid y) P_\theta(z \mid y). \tag{10}$$

Given this, the Bayesian balance of the policy is $f(\pi) = \int_\Theta C(\pi, \theta)\, \mathrm{d}\beta(\theta)$.

In order to find a rule that trades-off utility for balance, we maximize a convex combination of the expected utility and deviation specified in (4). In particular, we look for a parametrized rule π_w solving the following unconstrained maximization problem:

$$\begin{aligned} \max_{\pi_w} & \int_\Theta V_\theta(\pi_w)\, \mathrm{d}\beta(\theta), \\ & V_\theta(\pi_w) \triangleq (1-\lambda)\, \mathbb{E}^{\pi_w}_\theta\, u - \lambda C(\pi_w, \theta) \end{aligned} \tag{11}$$

To perform this maximization, we use parametrized policies and stochastic gradient descent. In particular, for a finite set $\mathcal{X}$ and $\mathcal{Y}$, the policies can be defined in terms of parameters $w_{xa} = \pi(a \mid x)$. Then we can perform stochastic gradient descent as detailed in the Supplementary materials, by sampling $\theta \sim \beta$, and calculating the gradient for each sampled θ. For the marginal decision rule, we employ the same approach, but instead of sampling the parameters from the posterior, we use the parameters of the marginal model.

Experiments

We study the utility-fairness trade-off on artificial and real data sets. We compare our approach, which uses a decision rule based on the full Bayesian problem, to classical approaches such as (Hardt, Price, and Srebro 2016) which optimize the DM's policy with respect to a single model. We show that the Bayesian approach gracefully handles fairness, even with high model uncertainty, while a marginal approach can be blatantly unfair. For an unbiased comparison, we assume the same prior parameter distribution. We consider a model where posteriors can be calculated in closed-form, in order to focus on the choice of policy. However, our

algorithm is generally applicable, and could be combined for example with MCMC inference.

Performance is evaluated with respect to the actual balance and utility achieved: for the synthetic data this is measured according to the actual data-generating distribution, while for the COMPAS data this is the empirical distribution on a holdout set.

The algorithm for optimizing policies uses stochastic gradient descent. In particular, the Bayesian policy minimizes (5) by sampling θ from the posterior distribution β and then taking a step in the gradient direction. The marginal policy simply performs steepest gradient descent for the marginal model.

The results shown in Figures 2–5 display the performance of the corresponding Bayesian or marginal decision rule for different value of λ as more data is acquired. In the first two experiments, we assume that no matter what the decision of the DM is, z_t, y_t are always observed after the DM's decision and so the model is fully updated. In that setting, it is not necessary for the DM to take into account the information generated by actions. However, in the third experiment, described below, the values of z_t and y_t are only observed when the DM makes the decision $a_t = 1$, and the DM faces a generalized exploration problem.

The model we employ throughout is a discrete Bayesian network model, with finite $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{A}$. The models are thus described through multinomial distributions that capture the dependency between different random variables. The available data is used to calculate a *posterior* distribution $\beta(\theta)$. From this, we calculate both a marginal balanced rule as well as a Bayesian balanced rule. The former uses the marginal model directly, while the latter uses $k = 16$ samples from the posterior distribution.[3] We tested these approaches both on synthetic data and on the COMPAS dataset. The conjugate prior distribution to this model is a Dirichlet-product. The graphical model is fully connected, and the model uses the factorization $P_\theta(x, y, z) = P_\theta(y \mid x, z)P_\theta(x \mid z)P_\theta(z)$. We used this simple modeling choice throughout the paper, apart from the small experiment on synthetic data in the following section (Experiments on synthetic data). In all cases where a Dirichlet prior was used, the Dirichlet prior parameters were set equal to $1/2$.

Experiments on synthetic data

Here we consider a discrete decision problem, with $|\mathcal{X}| = 8$, $|\mathcal{Y}| = |\mathcal{Z}| = |\mathcal{A}| = 2$, and $u(y, a) = \mathbb{I}\{y = a\}$. We generate 100 observations from this model. We perform the experiment 10 times, each time generating data from a fully connected, discrete Bayesian network with uniformly randomly selected parameters. Unlike the rest of the paper, in this example, the prior distribution has finite support on only 8 models. This means that the posterior will have effectively converged to the true model after 100 observations.

As can be seen in Figure 2, the relative performance of the Bayesian approach w.r.t. the marginal approach increases as we put more emphasis on fairness (Figure 2 (a) cares nothing about fairness.). In some cases (e.g. Figure 2 (c)), value for the marginal approach decreases at the beginning and eventually reaches the same value as the Bayesian approach after enough data has been observed. This conforms with our hypothesis that one should take into account model uncertainty. The fact that both approaches converge toward the maximum value is in accordance with our formal results (Theorem 2).

Finally, Figure 3 and its extended version (Figure S1 in supplementary materials) more clearly shows how well the two different solutions perform with respect to the utility fairness trade-off. As we vary λ and the amount of data, both methods achieve the same utility. However the Bayesian approach consistently achieves lower fairness violations for similar U.

Experiments on COMPAS data

For the COMPAS dataset, we consider a discretization where fields such as the number of offenses are converted to binary features.[4] We used the first 6000 observations for training and the remaining 1214 observations for validation. Two attributes are sensitive (sex, race), while six attributes (relating to prior convictions and age) are used for the policy. With discretization, there are a total of 12 distinct values for the sensitive attributes and 141 for the features that are used for the underlying model. The task is to predict recidivism over the next two years, with DM utility function $u(a, y) = \mathbb{I}\{a = y\}$.

Figure 4 and its extended version (Figure S2 in the supplementary materials) show the results of applying our analysis to the COMPAS dataset used by ProPublica. Since in this case the true model is unknown, the results are calculated with respect to the marginal model estimated on the holdout set. In this scenario we can see that when we only focus on classification performance, the marginal and Bayesian decision rules perform equally well. However, when we place more emphasis on fairness, we observe that the Bayesian approach dominates. [5]

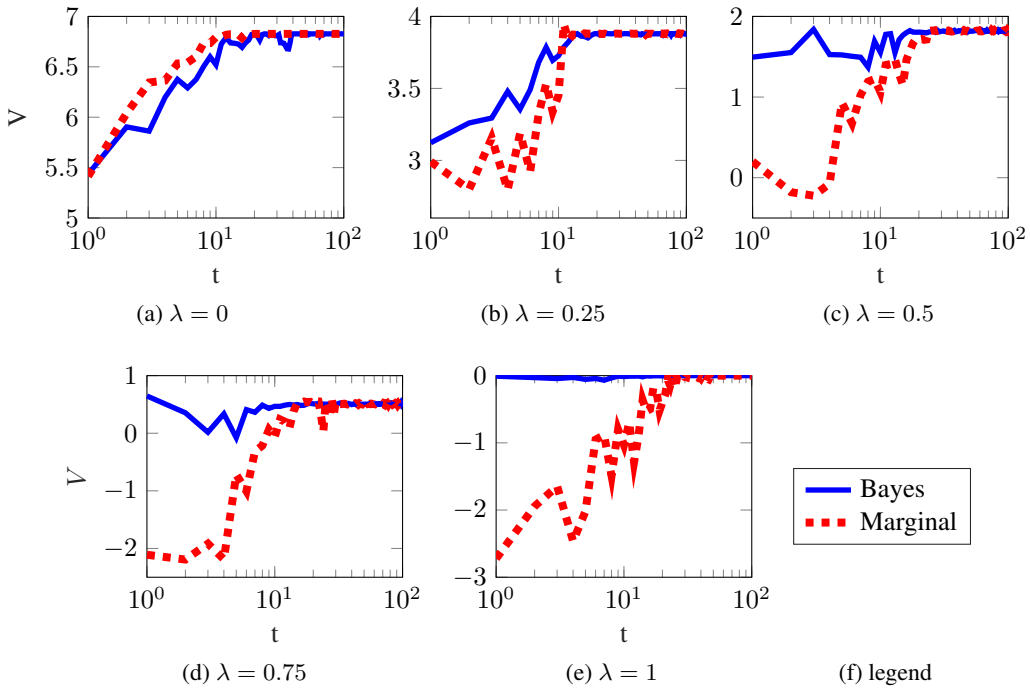
Sequential allocation

Suppose now that the DM, at each time t, observes x_t and has a choice of actions $a_t \in \{0, 1\}$. Both actions are to predict whether $y_t \in \{0, 1\}$ and have the following side-effect: the DM only observes y_t, z_t upon decision $a_t = 1$, and otherwise only observes x_t. The utility is not directly observed by the DM, and is measured against the empirical model in the holdout set, as before. We use the same COMPAS dataset, and the results are broadly similar, apart from

[3] We found empirically that 16 was a sufficient number for stable behaviour and efficient computation. For $k = 1$ the algorithm reduces to an approximation of Thompson sampling.

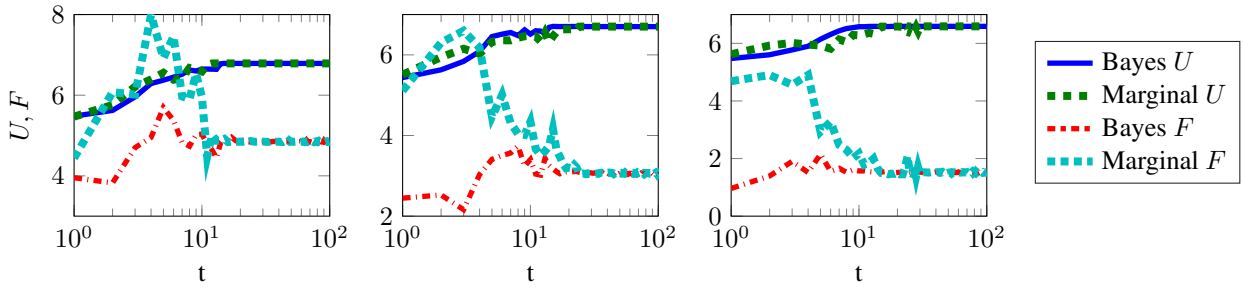
[4] We arrived at the specific discretization through cross validating the performance of a discrete Bayesian classifier over possible discretizations.

[5] The measured performance performance may not monotonically increase with respect to the (rather small) holdout set. Even if we had converged to the true model, measuring with respect to an empirical estimate is problematic, as it will be ϵ-close to the true model. This is particularly important for fairness considerations.



(a) $\lambda = 0$ (b) $\lambda = 0.25$ (c) $\lambda = 0.5$ (d) $\lambda = 0.75$ (e) $\lambda = 1$ (f) legend

Figure 2: **Synthetic data.** Test of the effect of the amount of data on the decisions of the Bayesian balance versus marginal balance approach, for different values of the λ parameter, with evaluation with respect to the true model. As more weight is placed on guaranteeing fairness, we see that the Bayesian approach is better able to guarantee fairness for the true model. The plots show the average performance over 10 runs, with an initially uniform prior over a set of 8 models, one of which is the correct one. In this setting $|\mathcal{A}| = |\mathcal{Y}| = |\mathcal{Z}| = 2$ and $|\mathcal{X}| = 8$.



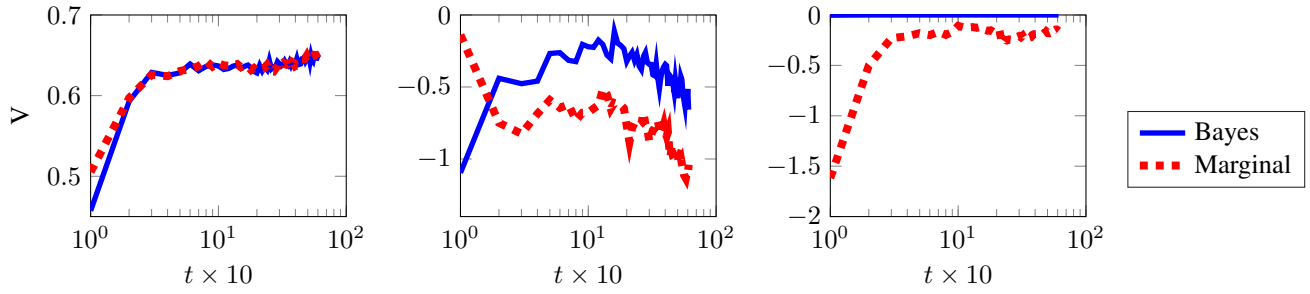
(a) $\lambda = 0.25$ (b) $\lambda = 0.5$ (c) $\lambda = 0.75$ (d) legend

Figure 3: **Synthetic data, utility-fairness trade-off.** This plot is generated from the same data as Figure 2. However, now we are plotting the utility and fairness of each individual policy separately. In all cases, it can be seen that the Bayesian policy achieves the same utility as the non-Bayesian policy, while achieving a lower fairness violation.

the fact that the Bayesian decision rule appears to remain consistent and robust (blue and solid lines in Figure 5) in this setting, while the marginal one's performance degrades. This is because the Bayesian decision rule explicitly takes uncertainty into account, while the marginal decision rule does not. The results are shown in Figure 5 and its extended version (Figure S3 in supplementary materials). The larger discrepancy between the Bayesian case in Figure 5(a) implies that explicitly modelling uncertainty is also crucial for utility in this case.

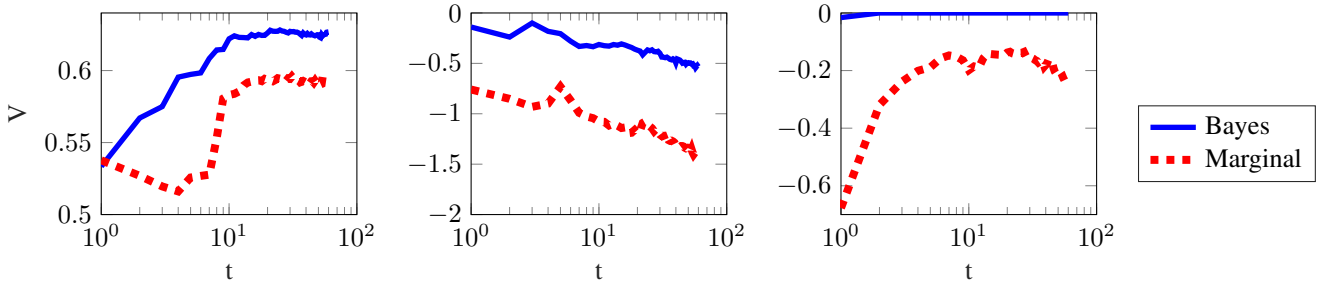
Conclusion

Existing fairness criteria can be hard to satisfy or verify in a learning setting because they are defined for the true model. Recognizing this, we develop a Bayesian framework for fairness, which allows a decision maker to explicitly reason about uncertainty about the true model and thus the extent to which a decision will, or will not be, fair. Beyond this, the Bayesian approach is helpful because it points to the importance of the informational aspects of fairness, and in par-



(a) $\lambda = 0$ (b) $\lambda = 0.5$ (c) $\lambda = 1$ (d) legend

Figure 4: **COMPAS dataset.** Demonstration of balance on the COMPAS dataset. The plots show the value measured on the holdout set for the **Bayes** and **Marginal** balance. Figures (a-c) show the utility achieved under different choices of λ as we we observe each of the 6,000 training data points. Utility and fairness are measured on the empirical distribution of the remaining data and it can be seen that the Bayesian approach dominates as soon as fairness becomes important, i.e. $\lambda > 0$.



(a) $\lambda = 0$ (b) $\lambda = 0.5$ (c) $\lambda = 1$ (d) legend

Figure 5: **Sequential allocation** Performance measured with respect to the empirical model of the holdout COMPAS data, when the DM's actions affect which data will be seen. This means that whenever a prisoner was not released, then the dependent variable y will remain unseen. For that reason, the performance of the Bayesian approach dominates the classical approach even when fairness is not an issue, i.e. $\lambda = 0$.

ticular for sequential decisions and the role that they play in both their current actions but their ability to censor or enable additional information acquisition.

Acknowledgments The project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 608743, the Swedish Research Council grant 2015-05410 and an SNSF Early Postdoc Mobility fellowship.

References

Blodgett, S. L., and O'Connor, B. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *CoRR* abs/1707.00061.

Celis, L. E., and Vishnoi, N. K. 2017. Fair personalization. *CoRR* abs/1707.02260.

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2017. Fair learning in markovian environments. *FATML*.

Chouldechova, A. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Technical Report 1610.07524, arXiv.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. Technical Report 1701.08230, arXiv.

Duff, M. O. 2002. *Optimal Learning Computational Procedures for Bayes-adaptive Markov Decision Processes*. Ph.D. Dissertation, University of Massachusetts at Amherst.

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

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *NIPS*, 3315–3323.

Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2016. Fair learning in markovian environments. *arXiv preprint arXiv:1611.03071*.

Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2016. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. Technical Report 1706.02744, arXiv.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. Technical Report 1609.05807, arXiv.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. Propublica COMPAS git-hub repository. https://github.com/propublica/compas-analysis/.

Puterman, M. L. 1994. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. New Jersey, US: John Wiley & Sons.

Russell, C.; Kusner, M. J.; Loftus, J.; and Silva, R. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, 6414–6423.