Bayesian Fairness

Christos Dimitrakakis, Yang Liu, David Parkes, Goran Radanovic

Bayesian Fairness

Christos Dimitrakakis
University of Oslo, Chalmers University
christos.dimitrakakis@gmail.com

Yang Liu
UC Santa Cruz
yangliu@ucsc.edu

David C. Parkes
Harvard University
parkes@eecs.harvard.edu

Goran Radanovic
Harvard University
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 a desirable property of policies applied to a population of individuals. For example, college admissions should be decided on variables that inform about merit, but fairness may also require taking into account the fact that certain communities are inherently disadvantaged. At the same time, a person should not feel that another in a similar situation obtained an unfair advantage. All this must be taken into account while still optimizing a decision maker's utility function.

Much of the recent work on fairness in machine learning has focused on analysing sometimes conflicting definitions. In this paper we do not focus on proposing new definitions or algorithms. We instead take a closer look at informational aspects of fairness. In particular, by adopting a Bayesian viewpoint, we can explicitly take into account model uncertainty, something that turns out to be crucial for fairness.

Uncertainty about the underlying reality has two main effects. Firstly, most notions of fairness are defined with respect to some latent variables, including model parameters. This means that we need to take into account uncertainty in order to be fair. Secondly, in many problems our decisions determine what data we will collect in the future. Ignoring uncertainty may magnify subtle biases in our model.

By viewing fairness through a Bayesian decision theoretic perspective, we avoid these problems. In particular, we demonstrate that Bayesian policies can optimally trade off utility and fairness by explicitly 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 maximise her expected utility u. However, the DM must trade off utility with some fairness constraint 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] In this paper we adopt a Bayesian viewpoint and assume the DM has some 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 what 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's objective is to maximize her expected utility. We model this as a function with structure $u : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y}$ is a set of *outcomes*. The fairness concept we focus on in this paper is a Bayesian version of *balance* [Kleinberg, Mullainathan, and Raghavan, 2016], which depends on the policy at time t. In the Bayesian setting, information is central. The amount of uncertainty about the model parameters directly influences how fairness can be achieved. Informally, the more uncertain we are, the more stochastic the decision rule is.

Our contributions. In this paper, we develop a framework for fairness that is defined as being appropriate to the available information for the DM. The motivation for the Bayesian framework is that there can be a high degree of uncertainty, particularly when not a lot of data has been collected, or in sequential settings. This informational notion of fairness is central to our discussion. It entails that the DM should take into account how unfair she would be under all possible models, weighted by their probability. While the fairness concepts we use are

[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.

grounded in conditional independence [Chouldechova, 2016; Kleinberg, Mullainathan, and Raghavan, 2016; Hardt, Price, and Srebro, 2016] type of notions of fairness, we employ a Bayesian decision theoretic methodology. In particular, we cleanly separate model parameters from the DM's information, and the decision rule used by the DM. Fairness can thus be seen as a property of the decision rule with respect to the true model (which is used to *measure* fairness), while achieving it depends on the DM's information (which is used to derive *algorithms*). The Bayesian approach we adopt for fair decision making is generally applicable. In this paper, however, we focus on a simple setting so that we can work without model approximations, and proceed directly to the effect of uncertainty on fairness. The policies we obtain are qualitatively and quantitatively different when we consider uncertainty (by being Bayesian) compared to when we do not.

The Bayesian algorithms we develop, based on gradient descent, take into account uncertainty by considering fairness with respect to the DM's information. This inherent modeling of uncertainty allows us to select better policies when those policies influence the data we collect, and thus our knowledge about the model. This is an important informational feedback effect, 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 Sections *The Sequential setting* and *Sequential allocation*. The results 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. Recently algorithmic fairness has been studied quite extensively in the context of statistical decision making. 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 the one-shot statistical decision making framework. [Jabbari et al., 2016; Joseph et al., 2016] kicked off the study of fairness in sequential decision making settings. Besides, there is also a trending line of research on fairness in other machine learning topics, such as clustering [Chierichetti et al., 2017], natural language processing [Blodgett and O'Connor, 2017] and recommendation systems [Celis and Vishnoi, 2017]. While the aforementioned works focused on fairness in a specific context, such as classification, [Corbett-Davies et al., 2017] have considered how to satisfy some of the above fairness constraints while maximizing expected utility. For a given model, they find a decision rule that maximizes expected utility while satisfying fairness constraints. [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. Finally, we'd like to mention the recent work of [Russell et al., 2017], which considers the problem of uncertainty from the point of view of causal modeling, with the three main differences 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, which is the main focus of that paper; (c) the Bayesian framework easily admits a sequential setting.

In this paper, we focus on notions of fairness related to notions of conditional independence, discussed next.

Preliminaries

[Chouldechova, 2016] considers the problem of fair prediction with disparate impact. She defines an action[2] a as *test-fair* with respect to the outcome y and the 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 it 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, which we re-interpret as follows:

Definition 2 (Balanced decision rule)**.** A decision rule [3] $\pi(a \mid x)$ 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.

These authors also work with the true model, while we will slightly generalize the definition, stating balance with respect to any model parameter.

Unfortunately, the calibration and balanced conditions cannot be achieved simultaneously for non-trivial environments [Kleinberg, Mullainathan, and Raghavan, 2016]. This is also true for our more general definitions, as we show in Theorem 3 in the Supplementary material. From a practitioner's perspective, we must choose either the calibration condition or the balanced conditions in order to find a fair decision rule. We work with the balanced condition, because it gracefully degrades to settings with uncertainty. In particular, balance involves equality in the expectation of a score

[2] Called a "statistic" in their paper.

[3] Here we simplified the notation of the decision rule so that $\pi(a \mid x)$ corresponds to the probability of taking action a given observation x.

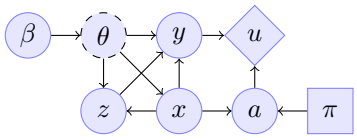


Figure 1: The basic Bayesian decision problem with observations x, outcome y, action a, sensitive variable z, utility u, unknown parameter θ, belief β and policy π. The joint distribution of x, y, z is fully determined by the unknown parameter θ, while the conditional distribution of actions a given 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.

function (when writing the probabilities as the expectations of 0-1 score functions; 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 trivial, for example randomized decision rule, being independent of x. The same, however, does not hold for the calibration condition under model uncertainty. Note that there also exist other fairness notions that go beyond disparate treatment [Zafar et al., 2017]. This merits future studies, and is out of the scope of the current draft.

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.[4]

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 and discrete, whereas Θ will be 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 for a given posterior and utility, but 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.

[4]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.

The Bayes-optimal decision rule does not directly depend on the sensitive variable z. We are interested in operating over 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. As always, 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. Note the decision problem and its variables stay unchanged over time.

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 due to fairness, and search for the Bayes-optimal decision rules with respect to this new, combined utility.

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

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

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

Bayesian Balance

In the Bayesian setting, we would like our decisions to take into account the impact on all possible models. While perfect balance is generally achievable, it turns out that sometimes only a trivial decision rule can satisfy it in a setting with model uncertainty (where 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 $\pi(\cdot)$ is (α, p)-Bayes-balanced with respect to β if:

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

This definition captures the expected deviation from balance of policy π, for a Bayesian DM under their belief β. It measures the deviation of the specific policy π from perfect balance with respect to each possible parameter θ, and weighs it 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 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, which we will see in our experimental results, is that the DM would assume that the marginal model is the correct one, and may be very unfair towards other high-probability models.

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. that 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 extend the approach to a sequential setting, where the information learned depends on the action. For example, if we grant a loan application, we will only later discover if the loan is going to be paid off. This will 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 what the data obtained by the policy was.

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

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

under the mild assumption that the set of reachable next beliefs is finite (easily satisfied when the set of outcomes is finite). This formulation is not different from standard MDP formulation (e.g., the reinforcement learning settings) that features the trade-offs between *exploration* (obtaining new knowledges) and *exploitation* (maximizing utilities). We know in these settings a myopic policy will lead to sub-optimal solutions.

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 'single-step' information. A specific instance of this type of sequential version of the problem is experimentally studied in Section *Sequential allocation*.

Algorithms

The algorithms we employ in this paper are based on gradient descent. We compare the full Bayesian framework with the simpler approach of assuming that the marginal model is the true one. 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

Again, as in the Bayesian setting, 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:

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

to be the deviation from balance for decision rule π under parameter θ, where

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

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

In order to find a rule trading off utility for balance, we can maximize a convex combination of the expected utility

and deviation specified in (4). In particular, we can look for a parametrized rule π_w solving the following unconstrained maximization problem.

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

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 Section *Gradient calculations for optimal balance decision* of 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. The approach is otherwise identical.

Experiments

In this section 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 simply optimizes the DM's policy with respect to a single model. Rather than introducing a new fairness metric, we use a generalized version of the balance metric in Kleinberg, Mullainathan, and Raghavan [2016], which is also a generalization of the equality of opportunity in Hardt, Price, and Srebro [2016]. We see that the Bayesian approach very gracefully handles fairness, even with high model uncertainty, while a marginal approach can be blatantly unfair. For a fair comparison, in both cases we assume the same prior distribution for the parameters. We focus on a simple model where posterior distributions can be calculated in closed-form, in order to focus on the choice of policy, rather than the case with approximate inference. However, our algorithm is generally applicable and could be combined with e.g. MCMC inference.

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

The algorithm for optimising 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 expected future information for her actions. However, in the third experiment, described in Section *Sequential allocation*, 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 an approximate 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.[5] We tested these approaches both on synthetic data and on the COMPAS dataset. The conjugate prior distribution to this model is a simple Dirichlet-product, as the network is discrete. The graphical model is fully connected, so 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. In all cases where a Dirichlet prior was used, the Dirichlet prior parameters were all set equal to $1/2$.

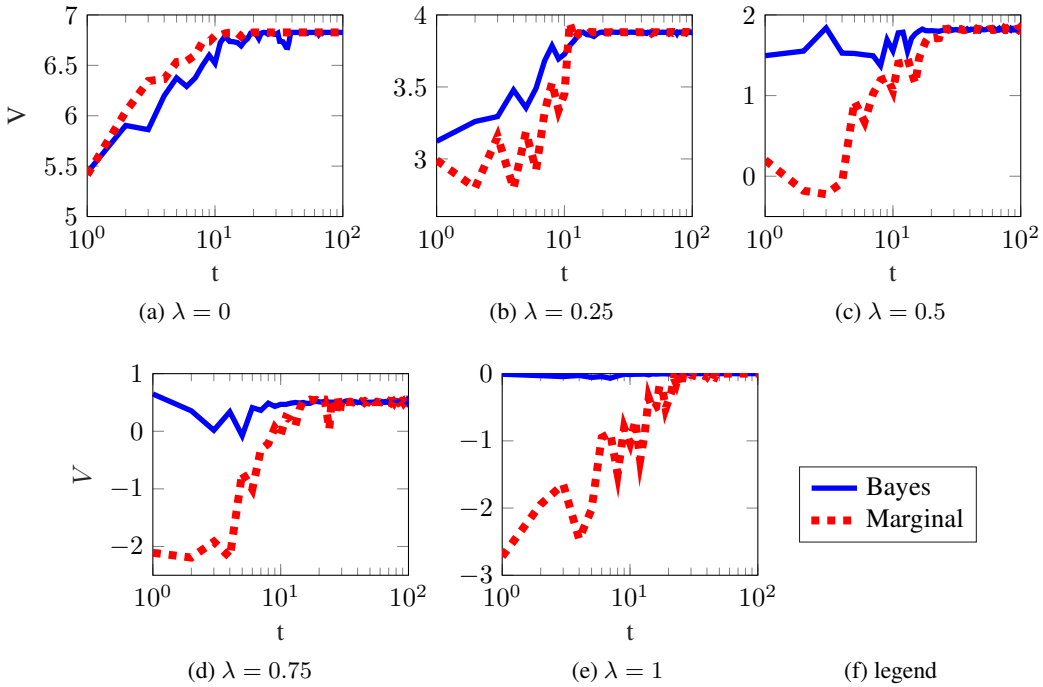
Experiments on synthetic data.

Here we consider a discrete decision problem, with $|\mathcal{X}| = 8$, $|\mathcal{Y}| = |\mathcal{Z}| = |\mathcal{A}| = 2$, and $u(y, a) = \mathbb{I}\{y = a\}$. In our first experiment, we generate 100 observations from this model. We performed 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 sufficient amount of data is received. 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 6 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.

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



(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 effect of amount of data for Bayesian versus marginal decision rules, for different values of the λ parameter, 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 $|\mathcal{X}| = 8$.

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.[6] We used the first 6000 data points for training and the remaining 1214 points for validation. Two attributes are sensitive (sex, race), while 6 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 observables used for the underlying model. The prediction is whether or not there is recidivism in the next two years, with utility function $u(a, y) = \mathbb{I}\{a = y\}$.

Figure 4 and its extended version (Figure 7 in supplementary materials) show the results of applying our analysis to the COMPAS dataset used by ProPublica. Since in this case the true model was 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, as we place more emphasis on fairness, we observe that the Bayesian approach dominates. [7]

Sequential allocation.

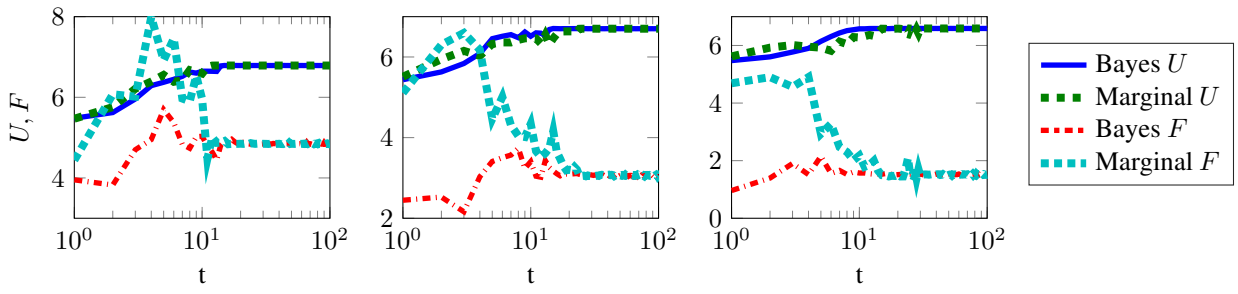
Here the DM, at each time t observes x_t and has a choice of actions $a_t \in \{0, 1\}$. The action both predicts $y_t \in \{0, 1\}$ and has the following side-effect: the DM only observes y_t, z_t after he makes the choice $a_t = 1$, otherwise only x_t. The utility is not directly observed by the DM, 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 the fact that the Bayesian decision rule appears to remain robust in this setting, while the marginal one's performance degrades. We presume that this is because that the Bayesian decision rule explicitly taking into account uncertainty leads to more robust performance relative to the marginal decision rule, which does not. The results are shown in Figure 5 and its extended version (Figure 8 in supplementary materials). The larger discrepancy between for the Bayesian case in Figure 5(a) implies that explicitly modelling uncertainty is also crucial for utility in this case.

Conclusion and future directions

Existing fairness criteria may be hard to satisfy or verify in a learning setting because they are defined for the true model.

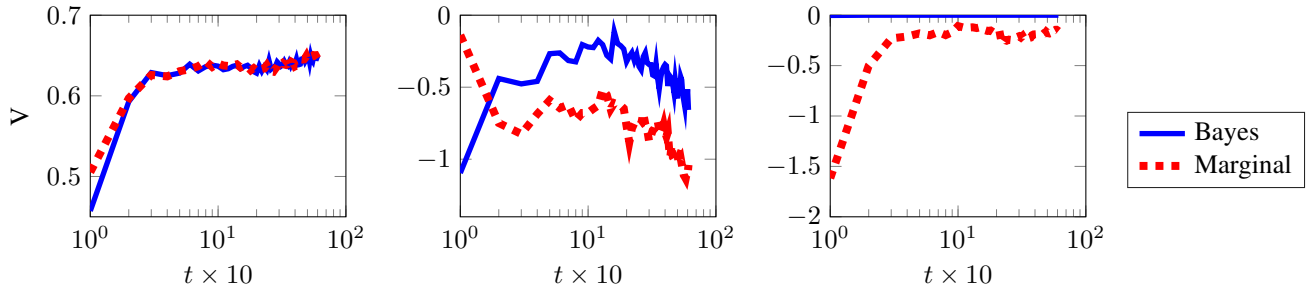
[6] We arrived at the specific discretization through cross validating the performance of a discrete Bayesian classifier over possible discretizations.

[7] We note here that 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-far away from the true model. This is particularly important for fairness considerations.



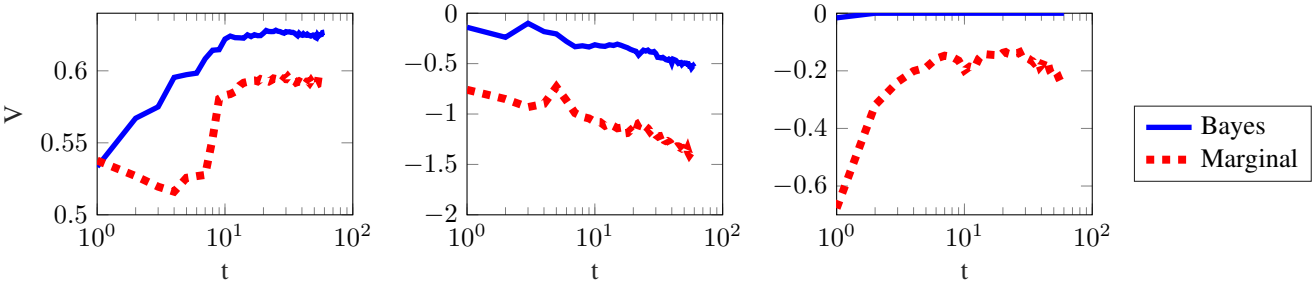
(a) $\lambda = 0.25$ (b) $\lambda = 0.5$ (c) $\lambda = 0.75$ (d) legend

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.



(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. $\lambda = 0$.

For that reason, we develop a Bayesian fairness framework, which deals explicitly with the information available to the decision maker. Our framework allows us to more adequately incorporate uncertainty into fairness considerations. We believe that a further exploration of the informational aspects of fairness, and in particular for sequential decision problems in the Bayesian setting, will be extremely fruitful.

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.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. International World Wide Web Conferences Steering Committee.

Supplementary materials for "Bayesian Fairness"

Impossibility result

Theorem 3. *Calibration and balance conditions cannot hold simultaneously, except: (i) if there exists perfect decision rules that there exists a, y s.t. $P^\pi_\theta(y \mid a) = 0$ or $P^\pi_\theta(a \mid y) = 0$, or (ii) z is independent of y that for each z, $P^\pi_\theta(z|y) \equiv$ const., $\forall y$.*

Proof. We prove by contradiction. Using Bayes rule we first have

$$P^\pi_\theta(a, z \mid y) = P^\pi_\theta(y, z \mid a) \cdot \frac{P^\pi_\theta(a \mid y)}{P^\pi_\theta(y \mid a)} . \tag{12}$$

Suppose calibration condition holds, that is

$$P^\pi_\theta(y, z \mid a) = P^\pi_\theta(y \mid a) P^\pi_\theta(z \mid a)$$

Plug above into Eqn. (12) we have

$$\begin{aligned} P^\pi_\theta(a, z \mid y) &= P^\pi_\theta(y \mid a) P^\pi_\theta(z \mid a) \cdot \frac{P^\pi_\theta(a \mid y)}{P^\pi_\theta(y \mid a)} \\ &= P^\pi_\theta(z \mid a) \cdot P^\pi_\theta(a \mid y). \end{aligned}$$

On the other hand, if balanced condition holds too, we have

$$P^\pi_\theta(a, z \mid y) = P^\pi_\theta(a \mid y) \cdot P^\pi_\theta(z \mid y)$$

Together we have that

$$P^\pi_\theta(z \mid a) \cdot P^\pi_\theta(a \mid y) = P^\pi_\theta(a \mid y) \cdot P^\pi_\theta(z \mid y) \to P^\pi_\theta(z \mid a) = P^\pi_\theta(z \mid y),$$

which does not hold when condition (ii) does not hold, completing the proof. □

Trivial decision rules for balance

Theorem 4. *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 $\theta \in \Theta$.*

Proof. For the first part, notice that Eqn. (3) can be always satisfied trivially if $\pi(a \mid x) = p_a$, i.e. we ignore the observations when taking our actions. For the second part, we can rewrite Eqn. (3) as

$$\begin{aligned} \sum_x \pi(a|x) \left[P_\theta(x, z|y) - P_\theta(x|y) P_\theta(z|y)\right] = 0 \\ \sum_x \pi(a|x) \Delta_\theta(x, y, z) = 0, \end{aligned}$$

where the Δ term is only dependent on the parameters. This condition can be satisfied in two ways: the first is if the model θ makes x, z conditionally independent on y. The second is if the vector $\pi(a \mid \cdot)$ is orthogonal to $\Delta_\theta(\cdot, y, z)$. If $|\mathcal{X}| > |\mathcal{Y} \times \mathcal{Z}|$, then, for any θ, we can always find a policy vector $\pi(a \mid \cdot)$ that is orthogonal to all vectors $\Delta_\theta(\cdot, y, z)$. However, if these vectors across θ have exactly degree of freedom being 1 (since they add up to 0, thus the rank of them can be at most the full rank - 1), then no single policy can be orthogonal to all, as otherwise the degree of freedom for this set of vectors will be at least 2. □

EXAMPLE 1. In this balance example, there are two models. In the first model, for some value y, we have:

$$P_\theta(x = 0|y) = 1/4, \qquad P_\theta(x = 0|y, z = 1) = 1/4 - \epsilon, \tag{13}$$
$$P_\theta(x = 1|y) = 1/4, \qquad P_\theta(x = 1|y, z = 1) = 1/4 + \epsilon, \tag{14}$$
$$P_\theta(x = 2|y) = 1/4, \qquad P_\theta(x = 2|y, z = 1) = 1/4 + \epsilon, \tag{15}$$
$$\tag{16}$$

so that

$$P_\theta(x=0|y) - P_\theta(x=0|y,z=1) = \epsilon, \tag{17}$$
$$P_\theta(x=1|y) - P_\theta(x=1|y,z=1) = -\epsilon \tag{18}$$
$$P_\theta(x=2|y) - P_\theta(x=2|y,z=1) = -\epsilon \tag{19}$$

Similarly, we can construct models θ' and θ'' so that the corresponding differences are $(-\epsilon, \epsilon, -\epsilon)$ and $(\epsilon, \epsilon, -\epsilon)$. For any policy $\pi(a \mid x)$ consider the vector $\pi_a = (\pi(a = 1 \mid x))_{x=1}^3$. Note that we can make the policy orthogonal to the first model simply by setting $\pi_a = (1/2, 1/2, 1)$.

Proof of Theorem 2

Proof. We show the proof for Bayes-balance condition, while the proof for Marginal-balance resembles similarities. Denote the $(1-\delta)$-event that θ drawn from $\beta(\theta)$ that is ϵ close to the true model θ^* in all the conditional probabilities $P_\theta(x|y,z), P_\theta(x|y)$ as $\mathcal{E}$, then we have:

$$\begin{aligned}
&\Big|\sum_x \pi(a|x)\left[P_{\theta^*}(x,z|y) - P_{\theta^*}(x|y)P_{\theta^*}(z|y)\right]\Big| \\
&= \left|\int_{\theta\in\mathcal{E}} \sum_x \pi(a|x)\left[P_{\theta^*}(x,z|y) - P_{\theta^*}(x|y)P_{\theta^*}(z|y)\right]\right. \\
&\left. + \int_{\theta\notin\mathcal{E}} \sum_x \pi(a|x)\left[P_{\theta^*}(x,z|y) - P_{\theta^*}(x|y)P_{\theta^*}(z|y)\right]\right| \\
&\le \left|\int_{\theta\in\mathcal{E}} \sum_x \pi(a|x)\left[P_{\theta}(x,z|y) - P_{\theta}(x|y)P_{\theta}(z|y)\right] + 2\epsilon\right. \\
&\left. + \int_{\theta\notin\mathcal{E}} \sum_x \pi(a|x)\left[P_{\theta^*}(x,z|y) - P_{\theta}(x|y)P_{\theta}(z|y)\right] + 2\delta\right| \\
&\le \Big|\sum_x \pi(a|x) \int_\Theta [P_\theta(x,z|y) - P_\theta(x|y)P_\theta(z|y)]\Big| + 2(\epsilon+\delta).
\end{aligned}$$

Summing over all a, y, z gives us the results. □

Gradient calculations for optimal balance decision

For simplicity, let us define the vector in $\mathcal{P}^{\mathcal{A}}$:

$$c_w(y,z) = \sum_x \pi_w(\cdot \mid x)\Delta(x,y,z),$$

so that

$$f_\lambda(w) = u(\beta, \pi_w) - \lambda \sum_{y,z} c_w(y,z)^\top c_w(y,z).$$

Now

$$\begin{aligned}
&\nabla_w \left(c_w(y,z)^\top c_w(y,z)\right) \\
&= \nabla_w \sum_a c_w(y,z)_a^2 \\
&= \sum_a 2c_w(y,z)_a \nabla_w c_w(y,z)_a \\
\nabla_w c_w(y,z)_a &= \sum_x \nabla_w \pi_w(a \mid x)\Delta(x,y,z),
\end{aligned}$$

while

$$\nabla u(\beta, \pi_w) = \int_{\mathcal{X}} \mathrm{d}\,\mathbb{P}_\beta(x) \nabla_w \pi_w(a \mid x)\, \mathbb{E}_\beta(u \mid x, a) \tag{20}$$

Combining the two terms, we have

$$\begin{aligned}\nabla_w f_\lambda(w) = \int_{\mathcal{X}} \nabla_w \pi_w(a \mid x) \big[\mathrm{d}\,\mathbb{P}_\beta(x)\, \mathbb{E}_\beta(U \mid x, a) \\ - 2\lambda \sum_{y,z} c_w(y,z)_a \Delta(x,y,z)\, \mathrm{d}\Lambda(x), \big].\end{aligned}$$

where Λ is the Lebesgue measure. We now derive the gradient for the $\nabla_w \pi_w$ term. We consider two parameterizations.

Independent policy parameters. When $\pi(a \mid x) = w_{ax}$, we obtain

$$\partial \pi(a' \mid x') / \partial ax = \mathbb{I}\{ax = a'x'\}$$

. This unfortunately requires projecting the policy parameters back to the simplex. For this reason, it might be better to use a parameterization that allows unconstrained optimization.

Softmax policy parameters. When

$$\pi(a \mid x) = e^{w_{ax}} / \sum_{a'} e^{w_{a'x}},$$

we have the following gradients:

$$\begin{aligned}\partial \pi(a \mid x) / \partial ax &= e^{w_{ax}} \sum_{a' \neq a} e^{w_{a'x}} \left(\sum_{a'} e^{w_{a'x}} \right)^{-2} \\ \partial \pi(a \mid x) / \partial a'x &= e^{w_{ax} + w_{a'x}} \left(\sum_{a''} e^{w_{a''x}} \right)^{-2}, \; a \neq a' \\ \partial \pi(a \mid x) / \partial a'x' &= 0, \; ax \neq a'x'.\end{aligned}$$

Empirical formulation.

For infinite $\mathcal{X}$, it may be more efficient to rewrite (**??**) as

$$\begin{aligned}0 &= \int_{\mathcal{X}} \pi(a \mid x)\, \mathrm{d}\,[P(x, z \mid y) - P(x \mid y) P(z \mid y)] && (21) \\ &= \int_{\mathcal{X}} \pi(a \mid x)\, [P(z \mid y, x) - P(z \mid y)]\; \mathrm{d}P(x \mid y) && (22) \\ &= \int_{\mathcal{X}} \pi(a \mid x)\, [P(z \mid y, x) - P(z \mid y)]\, \frac{P(y \mid x)}{P(y)}\, \mathrm{d}P(x) && (23) \\ &\approx \sum_{x \sim P_\theta(x)} \pi(a \mid x)\, [P(z \mid y, x) - P(z \mid y)]\, \frac{P(y \mid x)}{P(y)} && (24)\end{aligned}$$

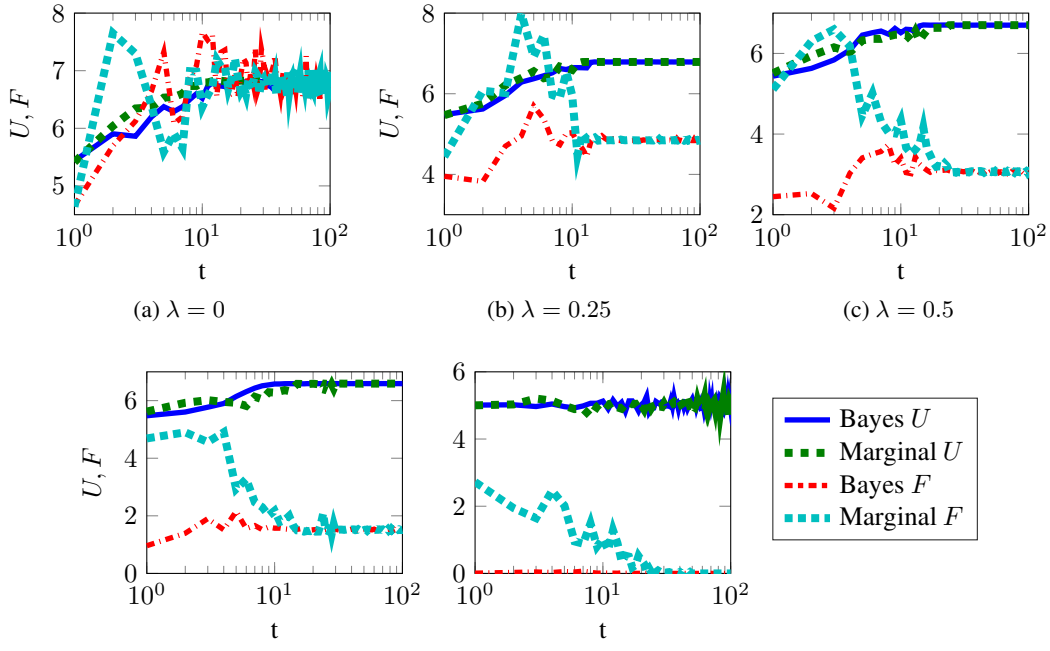
simplifying by dropping the $P(y)$ term:

$$0 \approx \sum_{x \sim P_\theta(x)} \pi(a \mid x)\, [P(z \mid y, x) - P(z \mid y)]\, P(y \mid x), \tag{25}$$

This allows us to approximate the integral by sampling x, and can be useful for e.g. regression problems.

Complete figures

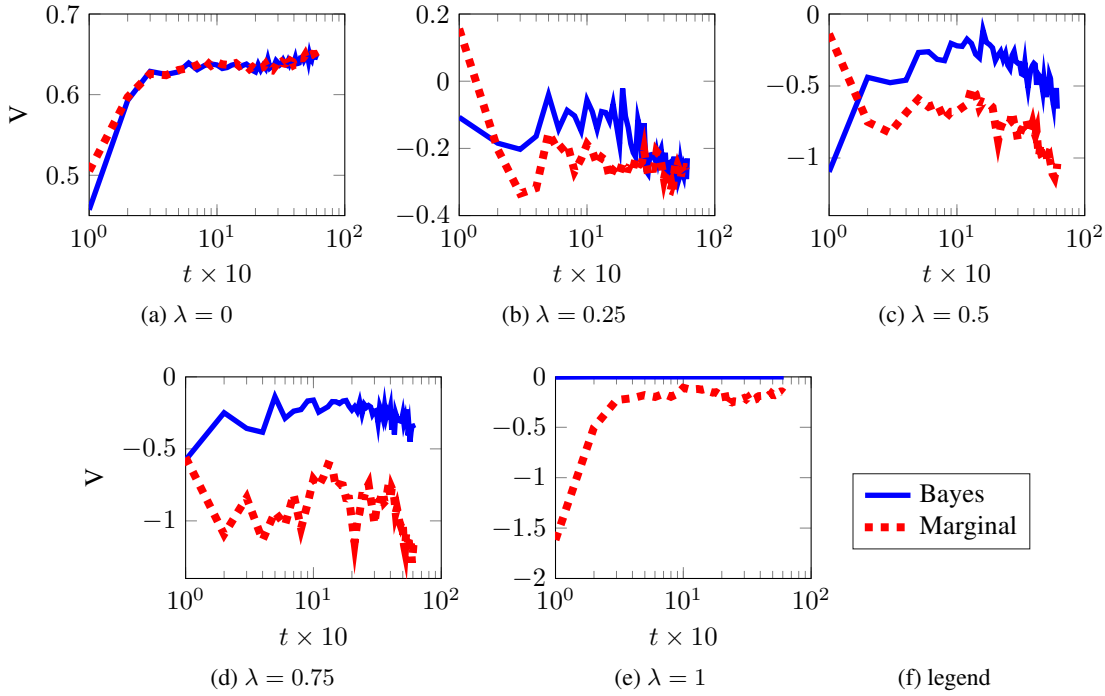
This section has complete versions of the figures which could not fit in the main text.



(a) $\lambda = 0$ (b) $\lambda = 0.25$ (c) $\lambda = 0.5$

(d) $\lambda = 0.75$ (e) $\lambda = 1$ (f) legend

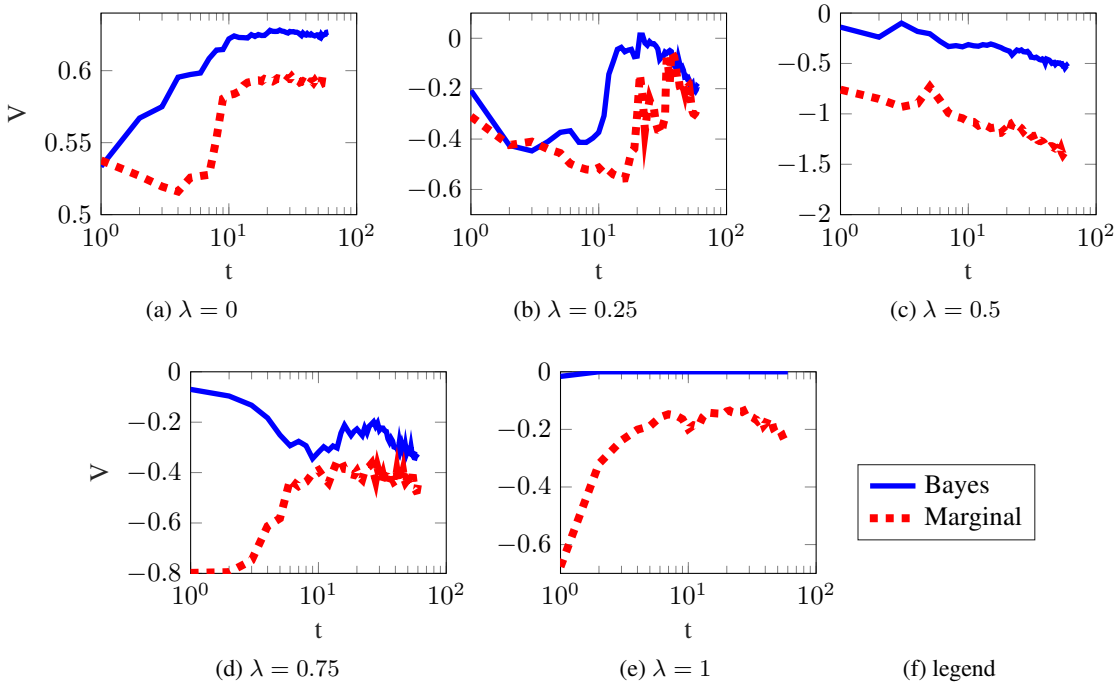
Figure 6: **Synthetic data, utility-fairness trade-off.** This plot is generated from the same data as Figure 2. However, now we are plotting the utility and fairness of each individual policy separately. In all cases, it can be seen that the Bayesian policy achieves the same utility as the non-Bayesian policy, while achieving a lower fairness violation.



(a) $\lambda = 0$ (b) $\lambda = 0.25$ (c) $\lambda = 0.5$

(d) $\lambda = 0.75$ (e) $\lambda = 1$ (f) legend

Figure 7: **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-e) 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) $\lambda = 0.25$ (c) $\lambda = 0.5$

(d) $\lambda = 0.75$ (e) $\lambda = 1$ (f) legend

Figure 8: **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 wif 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. $\lambda = 0$.