Subjective Fairness

Fairness is in the eye of the beholder

Christos Dimitrakakis Yang Liu David Parkes
Goran Radanovic

June 2, 2017

1 Introduction

Fairness is a desirable property of decision rules applied to a population of individuals. For example, college admissions should be decided on variables describing merit, but may also need to take into account the fact that certain communities are inherently disadvantaged. At the same time, individuals should not feel that another individual in a similar situation obtained an unfair advantage. All this must be taken into account while still caring about optimizing for a decision maker's utility function.

In particular, for a given distribution over a population, we wish to derive a decision rule that takes into account a merit variable, but also ensures fairness for members of disadvantaged groups. The problem becomes even more challenging when we take into account potential uncertainties in decision making models, which can even make strict notions of fairness impossible to satisfy.

As an example, consider the problem of fair prediction with disparate impact as defined by Chouldechova [2016]. Informally, their formulation defines a statistic a such that true category y (also called outcome or true label) is conditionally independent of a sensitive variable z given the statistic and the model parameters θ, i.e. $y \perp\!\!\!\perp z \mid a, \theta$. When we face uncertainties in our modeling assumptions, the natural thing is to impose that the conditional independence holds if we marginalize the parameters out, i.e. $y \perp\!\!\!\perp z \mid a$. As we argue later in the paper, such a condition is impossible to satisfy, even if it holds for every possible parameter value, and we must incorporate subjectivity when model parameters are uncertain.

We instead develop a natural, and widely applicable framework for fairness that relies on the *available information*. We develop algorithms for achieving a few different notions of fairness within the subjective framework, and in particualr recently proposed concepts of fairness that are grounded in concepts of similarity and conditional independence. We argue that a suitable notion of similarity in the Bayesian setting is distributional similarity conditioned on the observations. For the latter, as independence is difficult to achieve uniformly in the Bayesian setting, we suggest a relaxation, for which we provide a small experimental demonstration.

2 Constraints on distributions and impossibility results

Chouldechova [2016] considers the problem of fair prediction with disparate impact. In this context, a statistic $a : \Omega \to \mathcal{A}$ is test-fair in some underlying probability space $(\Omega, \Sigma, P_\theta)$, with respect to the *outcome* $y : \Omega \to \mathcal{Y}$ and *sensitive* variable $z : \Omega \to \mathcal{Z}$, if y is independent of z under the statistic and parameter θ, i.e. if $y \perp\!\!\!\perp z \mid a, \theta$. While the authors do not explicitly discuss the distribution P_θ, it must be known in order for us to be able to find a statistic a satisfying this property.

Kleinberg et al. [2016] consider feature vectors $x : \Omega \to \mathcal{X}$ and groups $z : \Omega \to \mathcal{Z}$ that correspond to the sensitive variables of Chouldechova [2016]. Their decision rule consists of a conditional distribution $\pi(a \mid x)$ that assigns individuals to a finite $\mathcal{A}$ called "bins". There is also a score function $U : \mathcal{A} \to [0, 1]$ corresponding to the utility of that bin. Finally, there is a categorical variable $y : \Omega \to \{0, 1\}$. They consider three fairness conditions, which we re-interpret below, subsuming two of them into a general "balance" condition:

$$U(a) = P^\pi_\theta(y = 1 \mid a, z) \qquad \text{(calibration)}$$
$$\mathbb{E}^\pi_\theta(U \mid y, z) = \mathbb{E}^\pi_\theta(U \mid y, z'). \qquad \text{(balance)}$$

The authors show that these cannot can be simultaneously achieved under the distribution P^π_θ induced by the underlying latent parameter θ and the decision rule π. However, a sufficient condition for *balance* is the independence: $U \perp\!\!\!\perp z \mid y, \theta, \pi$, i.e. when $P^\pi_\theta(U \mid y, z) = P^\pi_\theta(U \mid y)$. In contrast to the condition in Chouldechova [2016], it is the decision variable U that is made independent given the ground-truth variable y, instead of the other way around. A stronger sufficient condition is the independence of the statistic a itself, rather than the utility variable.

Decision problems. While the aforementioned works focused on classification, Dwork et al. [2012] consider general decision rules $\pi : \mathcal{X} \to \mathcal{A}$ maximising expected utility, under a fairness constraint:

$$\sup \{ \mathbb{E}^\pi U \mid \Delta(\pi(\cdot \mid x), \pi(\cdot \mid y)) \leq \rho(x, y) \, \forall x, y \in \mathcal{X} \}, \tag{1}$$

where $\Delta(P, Q)$ is a divergence between distributions P, Q, and $\rho(x, y)$ is a metric in observation space. This formalizes a type of fairness where similar people are treated similarly. Even though there is no inherent notion of a sensitive variable, it's possible to capture it in the observation metric to a limited extent. In a similar decision-theoretic context, Corbett-Davies et al. [2017] consider tradeoffs between utility maximization and satisfaction of constraints on distributions, such as balance.

On **sequential decision problems**, such as multi-armed bandits and reinforcement learning, Joseph et al. [2016] define an algorithm as fair if it plays arms with highest means most of the time. Jabbari et al. [2016] study a similar notion for Markovian environments, whereby the algorithm's is fair the algorithm is more likely to play actions that have a higher utility under the optimal policy. However, this notion of fairness relies on oracle knowledge, while we focus on subjectivity.

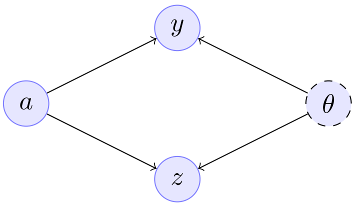


Figure 1: Graphical model of the fairness as equity problem. Here, y, z are clearly independent given a, θ. However, when we marginalize over θ they become dependent.

2.1 Fairness, subjectivity and learning.

Unfortunately it may may impossible to find a non-trivial decision rule under uncertainty, even for the case where only a single notion of fairness is appropriate. While such a notion may be satisfied for the correct underlying distribution P_θ, it may not hold for an arbitrary estimator: In the Bayesian case, even if all P_θ in a family satisfy a fairness condition, the marginal distribution may not.

Going back to our example from the introduction, we can easily see the impossibility of satisfying the disparate impact condition of Chouldechova [2016] from Figure 1. In this case, even if $y \perp\!\!\!\perp z \mid a, \theta$, the same independence does not hold if θ is latent, no matter how a is distributed. This is clear from the graphical model, as y, z are connected through θ. When θ is observed, it is blocking and independence can be achieved.

$$\begin{aligned} P_\beta(y, z \mid a) &= \sum_\theta P_\theta(y, z \mid a)\beta(\theta) && (2) \\ &= \sum_\theta P_\theta(y \mid a) P_\theta(z \mid a)\beta(\theta) && (3) \\ &\neq P_\beta(y \mid a) P_\beta(z \mid a). && (4) \end{aligned}$$

In the Bayesian setting, even if all the members of our family satisfy a fairness condition, uncertainty about the correct distribution can create unfairness. In the following, we begin by a definition of the Bayes optimal rule, and then consider two achievable fairness constraints: smoothness and balance.

3 Fair Bayes decision rules

We begin with a simple statistical decision problem, where the decision maker observes $x \in \mathcal{X}$, then takes a decision $a \in \mathcal{A}$ and obtains utility $U(y, a)$ depending on an outcome $y \in \mathcal{Y}$ generated from some unknown distribution $P_\theta(x \mid y)$.

We have some prior information $\beta \in \mathcal{B}$ in the form of a probability distribution on parameters $\theta \in \Theta$, an observation space $\mathcal{X}$, an outcome space $\mathcal{Y}$, and a family $\mathcal{F} \triangleq \{ P_\theta(y \mid x) \mid \theta \in \Theta \}$ of distributions.

We consider a simple factorization, where $\beta(\theta \mid x, a) = \beta(\theta)$, and a utility function $U : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ that only depends on our action and the outcome. The decision diagram is given in Figure 2.

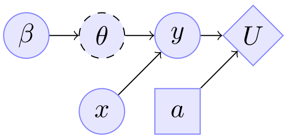


Figure 2: The basic Bayesian decision problem with observations x, latent outcome y, action a, utility U, unknown parameter θ, prior β.

Definition 1 (Bayes decision rule)**.** The Bayes 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

$$\pi^*(\beta, x) \in \operatorname*{arg\,max}_{a \in \mathcal{A}} U_{\beta|x}(a) \tag{5}$$

$$U_{\beta|x}(a) \triangleq \mathbb{E}_\beta(U \mid a) = \sum_y U(y, a) P_{\beta|x}(y), \tag{6}$$

where $P_{\beta|x}(y) = \int_\Theta P_\theta(y \mid x) \, \mathrm{d}\beta(\theta)$ is the marginal distribution over outcomes conditional on the observations.

We can define analogues of the fairness concepts listed previously in terms of decision rules π, rather than any specific random variable. We will focus on Bayesian decision rules, i.e. rules whose decisions solely depend upon a posterior distribution.

3.1 Fairness as smoothness

The fairness notion in Dwork et al. [2012] requires action distributions to be similar for similar individuals, but the question of what "similar" is left open to the context. In the Bayesian setting, there is a natural candidate: the relevant information about an individual encoded in the distribution of outcomes $P_{\beta|x}(y)$:

Definition 2 (Smoothly fair Bayesian decision rule)**.** A stochastic decision rule $\pi : \mathcal{B} \times \mathcal{X} \to \mathbb{\Delta}(\mathcal{A})$ on $\langle U, \Theta, \mathcal{X}, \mathcal{Y}, \mathcal{A} \rangle$ is $(k, \beta \mid x)$-fair with respect to the observation x for some prior β if

$$\Delta\left(\pi(\beta, x), \pi(\beta, x')\right) \leq k D\left(P_{\beta|x} \middle\| P_{\beta|x'}\right), \tag{7}$$

where $D\left(P \| Q\right)$ is some divergence between measures P and Q, and $\Delta\left(P, Q\right)$ is a (possibly different) divergence. It is simply (k, β)-prior-fair with respect to the prior β, if

$$\Delta\left(\pi(\beta, x), \pi(\beta', x)\right) \leq k D\left(\beta \| \beta'\right), \qquad \forall x \in \mathcal{X}. \tag{8}$$

The first definition ensures that, people whose merit distribution is similar under the same distribution β on parameters, will have a similar treatment. The second definition looks instead at how the treatment of one particular person who differ under a changing belief β on parameters.

We analyze Bayesian procedures that are commonly used in reinforcement learning problems. The first algorithm samples the underlying parameter θ from the posterior distribution and selects the decision that is optimal for the sample (i.e. marginalising over outcomes y). This is commonly known as Thompson samplingThompson [1933] in the reinforcement learning literatureWang et al. [2005], Agrawal and Goyal [2012].

Lemma 1. *The Thompson sampling algorithm, i.e. making a decision a with probability equal to*

$$\pi^{Th}(\beta \mid x)(a) = \beta\left(\{\, \theta \in \Theta \mid \mathbb{E}_\theta(U \mid a) > \mathbb{E}_\theta(U \mid a') \,\} \mid x\right), \tag{9}$$

is prior-fair when Δ is the KL divergence.

Proof. Thompson sampling can be seen as first drawing some $\hat{\theta} \sim \beta(\theta \mid x)$ and then selecting the decision maximising $\mathbb{E}_{\hat{\theta}}(U \mid a) = \int_{\mathcal{Y}} U(y, a) \, \mathrm{d}P_{\hat{\theta}}(y \mid x)$, which depends only on $\hat{\theta}$. Through the post-processing inequality, we obtain the required result. □

The second algorithm instead samples the latent outcome y, marginalising over parameters θ and corresponds to sampling actions according to their stochastic dominance.

Lemma 2. *Stochastic dominance sampling, i.e. making a decision a with probability equal to*

$$\pi^{SD}(\beta \mid x)(a) = P_{\beta|x}\left(\{\, y \in \mathcal{Y} \mid (U(y, a) > U(y, a') \,\}\right), \tag{10}$$

is observation-fair when Δ is the KL divergence.

Proof. This is in line with the proof of the previous lemma, but now we are drawing outcomes y instead of parameters θ. □

For single-shot decisions, sampling has no informational advantage, as there is no exploration-exploitation trade-off. This makes the following result natural, but one should keep in mind that it also applies to fact that the (intractable) Bayes-optimal sequential decision rule is also deterministic.

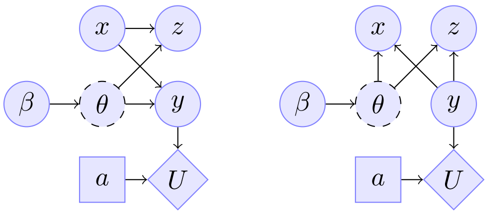
Lemma 3. *The Bayes-expected utility of stochastic dominance sampling is lower than that of Thompson sampling, which in turn is lower than that of the Bayes decision rule.*

Proof. Let $a^*(x) \triangleq \max_a \mathbb{E}(U \mid a, x)$ for some arbitrary distribution. Then for any randomized policy π: $\mathbb{E}^\pi(U \mid x) = \int_{\mathcal{A}} \mathbb{E}(U \mid a, x) \, \mathrm{d}\pi(a \mid x) \leq \int_{\mathcal{A}} \mathbb{E}(U \mid a^*, x) \, \mathrm{d}\pi(a \mid x) = \mathbb{E}(U \mid a^*, x) \, \mathrm{d}\pi(a \mid x)$. We first apply this to the distribution $\beta(\theta)$ to obtain the result for Thompson sampling. For stochastic dominance, note that $\mathbb{P}_\beta(X > Y) = \int_\Theta \mathbb{P}_\theta(X > Y) \, \mathrm{d}\beta(\theta)$. As stochastic dominance can be implemented by first sampling a parameter and then sampling a dominant variable under this parameter, we can reapply this fact and obtain the final result. □

We now consider a simple extension of Thompson sampling, which takes k parameter samples. As $k \to \infty$, the algorithm becomes closer to the Bayes-optimal.

Lemma 4. *The k-Thompson sampling algorithm π_k^{Th}, sampling $\Theta_k \sim \beta^k$ and taking action*

$$a^*(x) \in \operatorname*{arg\,max}_{a \in \mathcal{A}} \frac{1}{k} \sum_{\theta \in \Theta_k} \int_{\mathcal{Y}} U(y, \theta) \, \mathrm{d}P_\theta(y \mid x), \tag{11}$$



(a) Discriminative (b) Generative

Figure 3: The Bayesian decision problem with a sensitive variable z.

is (k, β)*-fair, i.e.*

$$\varDelta\left(\pi_k^{Th}(\beta, x), \pi_k^{Th}(\beta, x')\right) \leq kD\left(P_{\beta|x} \middle\| P_{\beta|x'}\right), \tag{12}$$

and $(k^{-1/3}, \beta)$*-optimal for* $U \in [0, 1]$.

Proof. The first claim follows from the fact that we obtain k independent draws and from the additive property of the KL divergence. Optimality follows from standard concentration inequalities. For example, from Hoeffding's inequality, and bounded rewards, we obtain that with probability $1-\delta$ our k-sample reward is $\sqrt{\frac{\ln 1/\delta}{2k}}$-close to optimal. The result follows by choosing $\delta = k^{-1/3}$, and using the boundedness of the utility. □

Although these algorithms are defined for simple decision problems, they can be easily applied to sequential problems. This can be achieved if e.g. our decision space is the set of all behavioural policies. Now we look at another notion of fairness, which is instead places constraints on conditional independence of random variables.

3.2 Fairness as balance

The balance requirement in Kleinberg et al. [2016] involves equality in the expectation of a score function (depending on an observation x) under different values of a sensitive variable z, conditioned on the true (but latent) outcome y.

We can distinguish two cases for the dependency structure between x, y, z. The first is a discriminative model satisfying $y \perp\!\!\!\perp z \mid x, \theta$, our decision depends on x, but also on z, as x generates z. For a generative model satisfying $x \perp\!\!\!\perp z \mid y, \theta$, we only have that x is not independent of z if θ is latent. However, it might still be possible to obtain a decision rule satisfying the following balance condition, analogously to Kleinberg et al. [2016]:

Definition 3 (Balanced decision rule)**.** A decision rule[1] $\pi(a \mid x)$ is balanced with respect to P if a, z are independent under β for all y, i.e. if

$$P^{\pi}(a, z \mid y) = P^{\pi}(a \mid y) P^{\pi}(z \mid y), \tag{13}$$

where P^{π} is the distribution induced by P and the decision rule π.

[1] 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.

Expanding the left hand side, we obtain

$$
\begin{aligned}
P^{\pi}(a, z \mid y) &= P^{\pi}(a \mid z, y) P(z \mid y) && (14)\\
&= \sum_x \pi(a \mid x) P(x \mid y, z) P(z \mid y) && (15)
\end{aligned}
$$

while for the right hand side, we have

$$
\begin{aligned}
P^{\pi}(a \mid y) &= \sum_x P^{\pi}(a \mid x, y) P(x \mid y) && (16)\\
&= \sum_x \pi(a \mid x) P(x \mid y). && (17)
\end{aligned}
$$

Equating the two terms, we obtain that

$$
\sum_x \pi(a \mid x) \left[P(x, z \mid y) - P(x \mid y) P(z \mid y) \right] = 0, \qquad \forall a, y, z, \qquad (18)
$$

This implies two sufficient conditions: The first is $x \perp\!\!\!\perp z \mid y$, but this is something that depends entirely on P, rather than the decision rule itself. Under the generative model of Figure 3b, this is achieved when θ is known, but not necessarily under β. The second is that the policy probabilities for every possible action a are orthogonal to the difference between the distribution of x for every possible choice of z, y.

Empirical formulation. For infinite $\mathcal{X}$, it may be more efficient to rewrite (18) as

$$
\begin{aligned}
0 &= \int_{\mathcal{X}} \pi(a \mid x) \, \mathrm{d} \left[P(x, z \mid y) - P(x \mid y) P(z \mid y) \right] && (19)\\
&= \int_{\mathcal{X}} \pi(a \mid x) \left[P(z \mid y, x) - P(z \mid y) \right] \mathrm{d}P(x \mid y) && (20)\\
&= \int_{\mathcal{X}} \pi(a \mid x) \left[P(z \mid y, x) - P(z \mid y) \right] \frac{P(y \mid x)}{P(y)} \mathrm{d}P(x) && (21)\\
&\approx \sum_{x \sim P_\theta(x)} \pi(a \mid x) \left[P(z \mid y, x) - P(z \mid y) \right] \frac{P(y \mid x)}{P(y)}. && (22)
\end{aligned}
$$

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

3.2.1 The optimally balanced decision rule.

Such a rule would maximize expected utility under the balance constraint. In our setting, however, we have no fixed P, but rather a belief β on a family of distributions P_θ. This inherently makes our decision rule subjectively fair.

We consider two variants of subjective fairness in this context: *credible fairness* and its natural relaxation, *marginal fairness.*

Credible fairness. In the Bayesian setting, we have a family of models $\{ P_\theta \}$ with a corresponding subjective distribution $\beta(\theta)$. As no single model is correct we need to consider fairness for every possible model. Rather than looking for a single useful decision rule π such that:

$$\sum_x \pi(a \mid x) \left[P_\theta(x, z \mid y) - P_\theta(x \mid y) P_\theta(z \mid y)\right] = 0, \qquad \forall x, y, z, \theta$$

i.e. so that the constraint holds for all members $\{ P_\theta \mid \theta \in \Theta \}$ of the family, we instead choose to minimize some penalty function averaged over all models. For that reason, let us first define

$$C(\pi, \theta) \triangleq \Big\| \sum_x \pi(a \mid x) D_\theta(x, y, z) \Big\|_p^q \tag{23}$$

to be the deviation from balance for decision rule π under parameter θ. We now use this in the following definition:

Definition 4. A decision rule is ϵ-credibly fair under β, and $p, q \geq 0$, when $\epsilon \geq \int_\Theta C(\pi, \theta) \, \mathrm{d}\beta(\theta)$.

In order to find a rule trading off utility for balance, we can maximize a convex combination of the expected utility and deviation. In particular, we can look for a rule π solving the following unconstrained maximization problem.

$$\max_\pi \int_\Theta f_\theta(\pi) \, \mathrm{d}\beta(\theta) \tag{24}$$

$$f_\theta(w) \triangleq (1 - \lambda) \, \mathbb{E}_\theta^\pi \, U - \lambda C(\pi, \theta) \tag{25}$$

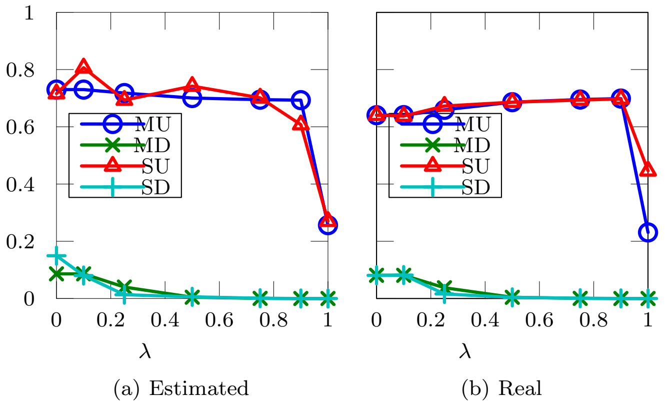
The rule maximising $\mathbb{E}_\beta^\pi f$ is a Bayes decision rule (Def. 1) for f under the distribution β. To find this, we can restrict ourselves to parametric decision rules, and use stochastic gradient descent. As in Section 3.1, if we only take $k < \infty$ samples to perform the gradient ascent over, the policy becomes an instance of k-Thompson sampling. When $p = q = 2$, the gradient procedure is quite simple, and is given in App. A,

Marginal fairness. As an approximation, we can maximize with respect to the marginal model, i.e.

$$f_\beta(w) \triangleq (1 - \lambda) \, \mathbb{E}_\beta^\pi \, U - \lambda \sum_{z,y} \left| \sum_x \pi_w(a \mid x) D_\beta(x, y, z) \right|^p,$$

where $\mathbb{E}_\beta, D_\beta$ are defined in terms of the marginal distribution $\mathbb{P}_\beta \triangleq \int_\Theta P_\theta \, \mathrm{d}\beta(\theta)$. This could be a useful formulation if calculating the marginal is easier than sampling.

A demonstration. Here we consider a discrete decision problem, for which we generate 100 observations. In particular $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{Z}| = |\mathcal{A}| = 4$, and $U(y, a) = \mathbb{I}\{y = a\}$. The true model has a dependence of approximately 0.26. These observations are used to calculate a posterior distribution $\beta(\theta)$. The prior distribution is a simple Dirichlet-product, as the network is discrete, and we assume we know the structure, which is $y \to x \to z$. We then find two decision rules. The first uses the $k = 10$ samples from the full posterior distribution, and the second uses the marginal model. Figure 4 shows the estimated[2] and achieved[3] utility (SU, MU) and dependence (SD, MD) for the sampling-based and marginal decision rule respectively. Both rules are successful in trading off utility and dependency, with the rules based on the complete distribution being slightly better. The results match the estimated performance rather well for this toy problem.



(a) Estimated

(b) Real

Figure 4: Demonstration of balanced Bayes. The plots show a summary of the estimated and real utility (U) and dependence (D) for a marginal (M) or sampling-based (S) decision rule trained with the independent parametrization. Figure 4a shows the utility and dependence as measured during the optimization, while Figure 4b shows the actual utility and dependence for the marginal and credible decision rules for the true underlying distribution, as we vary λ.

4 Conclusion

Existing fairness criteria may be hard to satisfy or verify in a learning setting because they are defined for the true model. For that reason, we develop a subjective fairness framework, which can encompass existing criteria from a Bayesian viewpoint. This allows us to make some new connections, the stochastic gradient ascent algorithm, for example, can simultaneously satisfy similarity and independence definitions.

In addition, although the smoothness fairness criterion implies a stochastic decision rule, this rule can become deterministic when the information about individuals results in a singular distribution of the merit variable y. This implies a link between the independence and similarity conditions for certain information sets. In that case, it's also easy to see that the independence criteria will also be met. We believe that a further exploration of the informational aspects of fairness, and in particular in the Bayesian setting, will be extremely fruitful.

[2] The function components, U, C targeted by the gradient algorithm.

[3] Measured using the true distribution P_θ.

References

Shipra Agrawal and Navi Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT 2012*, 2012.

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

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. Technical Report 1701.08230, arXiv, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaaron Roth. Fair learning in Markovian environments. Technical Report 1611.03107, arXiv, 2016.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaaron Roth. Fairness in learning: Classic and contextual bandits. Technical Report 1605.07139, arXiv, 2016.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. Technical Report 1609.05807, arXiv, 2016.

W.R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of two Samples. *Biometrika*, 25(3-4):285–294, 1933.

Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *ICML '05*, pages 956–963, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi: http://doi.acm.org/10.1145/1102351.1102472.

A Gradient calculations.

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

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

so that

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

Now

$$\nabla_w \left(c_w(y,z)^\top c_w(y,z)\right) = \nabla_w \sum_a c_w(y,z)_a^2 \tag{26}$$

$$= \sum_a 2c_w(y,z)_a \nabla_w c_w(y,z)_a \tag{27}$$

$$\nabla_w c_w(y,z)_a = \sum_x \nabla_w \pi_w(a \mid x) D(x,y,z), \tag{28}$$

while

$$\nabla U(\beta, \pi_w) = \int_{\mathcal{X}} \mathrm{d}\,\mathbb{P}_\beta(x) \nabla_w \pi_w(a \mid x)\, \mathbb{E}_\beta(U \mid x, a) \tag{29}$$

Combining the two terms, we have

$$\nabla_w f_\lambda(w) = \int_{\mathcal{X}} \nabla_w \pi_w(a \mid x) \big[\mathrm{d}\,\mathbb{P}_\beta(x)\, \mathbb{E}_\beta(U \mid x, a) \tag{30}$$

$$- 2\lambda \sum_{y,z} c_w(y,z)_a D(x,y,z)\, \mathrm{d}\mathit{\Lambda}(x), \big]. \tag{31}$$

where $\mathit{\Lambda}$ 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 $\partial \pi(a' \mid x')/\partial ax = \mathbb{I}\left\{ax = a'x'\right\}$. 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 $\pi(a \mid x) = e^{w_{ax}} / \sum_{a'} e^{w_{a'x}}$.

$$\partial \pi(a \mid x)/\partial ax = e^{w_{ax}} \sum_{a' \neq a} e^{w_{a'x}} \left(\sum_{a'} e^{w_{a'x}}\right)^{-2} \tag{32}$$

$$\partial \pi(a \mid x)/\partial a'x = e^{w_{ax} + w_{a'x}} \left(\sum_{a''} e^{w_{a''x}}\right)^{-2}, \qquad a \neq a' \tag{33}$$

$$\partial \pi(a \mid x)/\partial a'x' = 0, \qquad ax \neq a'x'. \tag{34}$$