Approximate Inference for the Bayesian Fairness Framework

Andreas Athanasopoulos[1], Amanda Belfrage[2], David Berg Marklund[2] and Christos Dimitrakakis[1,2,3]

[1]University of Neuchatel, Neuchatel, Switzerland
[2]Chalmers University, Gothenburg, Sweden
[3]University of Oslo, Oslo, Norway

Abstract

As the impact of Artificial Intelligence systems and applications on everyday life increases, algorithmic fairness undoubtedly constitutes one of the major problems in our modern society. In the current paper, we extend the work of Dimitrakakis et al. on Bayesian fairness [1] that incorporates models uncertainty to achieve fairness, proposing a practical algorithm with the aim to scale the framework for a broader range of applications. We begin by applying the bootstrap technique as a scalable alternative to approximate the posterior distribution of parameters of the fully Bayesian viewpoint. To make the Bayesian fairness framework applicable to more general data settings, we define an empirical formulation suitable for the continuous case. We experimentally demonstrate the potential of the framework from an extensive evaluation study on a real dataset and different decision settings.

Keywords

Bayesian Fairness, Algorithmic Fairness, Machine Learning, Decision Making

1. Introduction and Background

Algorithmic fairness is increasingly a major concern, in particular for AI systems. Fairness either relates to individuals (e.g. meritocracy) or groups (e.g. discrimination) [2]. We study the problem of fair decision-making under uncertainty. We adopt a Bayesian viewpoint, where fairness is a property of the decision rule π and the underlying problem parameter θ. However, while π is chosen by the decision maker (DM), the parameter is known only up to probability distribution. Hence, the DM must choose π so as to balance the utility-fairness trade-off by marginalising over the Bayesian posterior. However, as this is not always practical, in this paper we propose some approximations. We begin by explaining the basic framework in the remainder of this section, before offering algorithmic solutions and experimental results.

Fair Decision Rules: As in[1], we consider decision problems where the DM observes an individual's features $x \in X$, as well as a sensitive variable $z \in Z$ (such as gender), takes an action $a \in A$, resulting in an outcome $y \in Y$. The observed variables are generated from some distribution $P_{\theta(y|x,a,z)}$, $P_{\theta^*}(x, y)$. The decision-maker's actions a are produced according to a policy $\pi(a|x)$ that defines a probability over actions for every possible observation.

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
andreas.athanasopoulos@unine.ch (A. Athanasopoulos); amanda.belfrage@gmail.com (A. Belfrage); david@bergmarklund.se (D. B. Marklund); christos.dimitrakakis@gmail.com (C. Dimitrakakis)

Since the DM does not know the true parameter θ^*, they have a belief $\beta \in B$ over a family of distribution $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ that it may contain the actual law, i.e. P_θ for some θ. The belief β expresses the uncertainty of the decision maker about the word. In the Bayesian case, the belief β is a posterior formed through his prior distribution $P(\theta)$ and the available data.

The DM wishes to find a policy π maximizing expected utility $E^\pi_\beta[u]$, where $u(a, y)$ is a utility function dependent on the DM's action and the unknown outcome. At the same time, though they must take into account fairness constraints F. In particular, the optimal policy for a given belief β will maximize the following utility-fairness tradeoff:

$$T(\pi, \beta) = \mathbb{E}^\pi_\beta[U] - \lambda\, \mathbb{E}^\pi_\beta[F] = \int_\Theta \mathbb{E}^\pi_\theta[U] d\beta(\theta) - \int_\Theta \lambda\, \mathbb{E}^\pi_\theta[F] d\beta(\theta) \tag{1}$$

This idea was used in [1], which focused on fairness constraints related to balance [3], that states that the actions a should be independent of the sensitive variable z if conditioned to the true outcome y i.e $a \perp z \mid y$. The resulting deviation from fairness can be formalized as follows:

$$\mathbb{E}^\pi_\theta[F] = \sum_{a,z,y} \| \sum_x \pi(a|x)[P_\theta(x, z|y) - P_\theta(x|y)P_\theta(z|y)] \|^p_q \tag{2}$$

Combining the above fairness deviation (2) with (1) results in the **Bayesian Balance Rule**, while on the other hand, replacing the unknown P_θ with the marginal model $\mathbb{P}_\beta = \int_\Theta P_\theta d\beta(\theta)$, results in the **Marginal Balance Rule**, which in practice can be very unfair for high-probability models. This approximates the expected fairness violation as:

$$\mathbb{E}^\pi_{\mathbb{P}_\beta}[F] = \sum_{a,z,y} \| \sum_x \pi(a|x)[\mathbb{P}_\beta(x|z, y) - \mathbb{P}_\beta(x|y)]\, \mathbb{P}_\beta(z|y) \|^p_q \tag{3}$$

Related Work. The academic community has recently expressed growing concerns about the relationship between uncertainty and algorithmic fairness. The concept of decision-making under uncertainty, that we adopt, was first formally introduced by Dimitrakakis et al., while the significance of incorporating model uncertainty was also highlighted in [4] from a point of view of causal modeling. In a more recent study, Ali et al. also argues that we have to consider epistemic uncertainty, ignoring aleatoric uncertainty, in a fair classification setting by incorporating predictive multiplicity. Another interesting research direction is the work of Singh et al. that considers fairness in ranking under uncertainty. Additionally, a broader discussion about uncertainty as a form of transparency is presented by Bhatt et al..

Contribution. The previous work in the Bayesian fairness setting [1] focused on simple discrete models, where posterior distributions of the parameters can be calculated in closed form. We instead focus on the case where Bayesian inference is not closed-form, in particular for continuous data. We consider both replacing the posterior calculations with bootstrapping and the analytical version of the balance metric with an empirical approximation. Finally, we experimentally evaluate our extensions by comparing the policies obtained from Bayesian and the Marginal balanced rule on the COMPAS [8] dataset.

2. Algorithm for Bayesian Fairness

In the context of the Bayesian fairness framework, the decision maker has a belief β that is expressed through a distribution $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ over different θ parameters. Instead of using the posterior distribution, as in Bayesian inference, one can use the sampling distribution of an estimator $\hat{\theta}$ obtained by sub-sampling data D_i from the original dataset D. The idea of the bootstrapping method is to get different datasets D_i by sampling N different datasets of size $|D|$ with replacement, called bootstrap datasets. We then obtain a set of n samples $\theta_1, \ldots, \theta_n$, where $\theta_i = \hat{\theta}(D_i)$, and our optimisation problem becomes

$$\pi^*_{\text{Bootstrap}} = \arg\max_{\pi} \sum_{i=1}^{n} \mathbb{E}^{\pi}_{\theta_i}[U - \lambda F]. \tag{4}$$

In addition to the bootstrapping method, we need an empirical estimate of the fairness violation 2 when our observations x are continuous variables. In particular, for infinite X finite Z, Y and A the equation 2 becomes:

$$\mathbb{E}^{\pi}_{\theta}[F] \approx \sum_{a,z,y} \| \frac{1}{N} \sum_{i}^{N} \pi(a|x_i)[P_\theta(z|x_i, y) - P_\theta(z|y)] \frac{P_\theta(y|x_i)}{P_\theta(y)} \|^p_q \tag{5}$$

For the above equation we need to estimate $P_\theta(y)$, $P_\theta(z|y)$, $P_\theta(y|x)$, $P_\theta(z|x, y)$. In the case where Y, Z is finite it's easy to estimate $P_\theta(y)$, $P_\theta(z|y)$ with discrete models, whereas for $P_\theta(y|x)$, $P_\theta(z|x, y)$ we can use continuous predictive models to estimate them. To find the optimal policy of 1 we can make use of a gradient descent algorithm.

3. Experiments

We empirically evaluate our approach by comparing the policies obtained from the different balance rules on the COMPAS dataset. We performed two different experiments to study the effects of our algorithmic solution using discrete and continuous versions of the dataset. More specifically, in the discrete case, we can compare the bootstrap and the marginal methods to the closed-form Bayesian approach by employing a fully discrete Bayesian network model as described in [1]. In the second experiment, we only compare the policies obtained from Bootstrap and the marginal rules using the empirical formulation 5. To evaluate our policies, we calculate the fairness F and utility U with respect to the empirical model from a hold-out dataset, performing the experiments 50 times, each time shuffling our dataset and using 6000 (70%) data points for training and the rest 1214 (30%) for testing, reporting the average and the variance of the aforementioned metrics. Our code https://github.com/a-athanasopoulos/Bayesian-fairness reproduces all presented experiments.

Experimental Setup. To study how DM uncertainty affects fairness we consider the following experimental setting which is operating over multiple steps t, each time increasing the amount of training data. Intuitively, in the early steps of the process, the DM is more uncertain about the model parameters as he has assessed fess data compared to the later steps,

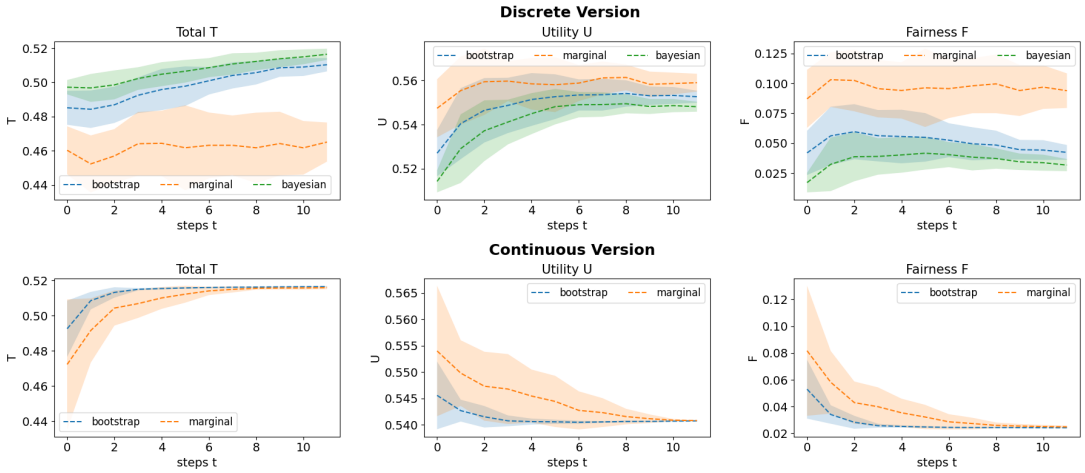

Figure 1: In the figure, we illustrate the total utility T, the utility U, and the Fairness F of the Discrete and Continuous experiments on the test-set and for $= 1$. At every step t, the policies were trained using $n = t * 500$ data points. We report the average performance together with the stand deviation on the test set for the different optimization approaches according to the labels over 50 runs.

so we expect the Bayesian methods to perform better regarding fairness F as we account for uncertainty. We use 12 steps each time considering $d_t = t * 500$ data points.

Models and Policies. The graphical model is fully connected, so the model uses the factorization $P_{(x,y,z)=P_{(x|z,y)}P_{(z|y)}P_{(y)}}$ from which we can calculate all the relevant models to optimize our policy. To form the posterior distribution for the discrete case we use the Dirichlet-Multinomial model with a non-informative initial Dirichlet parameter of $1/2$ from which we can sample different model parameters, while we can also calculate the marginal model in closed form. We use parameterized policies of the form $\pi(a|x) = w_{ax}$. For the continuous case, we calculate the marginal models of $P_\theta(y)$, $P_\theta(z|y)$ as in the discrete case, while for the models $P_\theta(y|x)$, $P_\theta(z|x, y)$ and the policy $\pi(a|x)$ that contains the continuous observations we use logistic regression models. The sampling distribution of the bootstrap method uses the marginal model using different bootstrap datasets for both continuous and discrete cases as described in section 2.

The Algorithm. We use gradient descent to optimize the policies by minimizing the negative total expected utility T as in 1. In particular, for the Bayesian approaches we sample a model from the $\mathcal{P}$ distribution and then take a step in the gradient direction in each iteration. More specifically for both bootstrapping and the fully Bayesian approach, we use 16 models to form the $\mathcal{P}$, obtained either by using a different bootstrap dataset or directly sampling from the true posterior accordingly. The marginal policies simply perform the steepest gradient descent for the marginal model. We make 1500 gradient decent iterations in each policy update, with a learning rate of 0.01.

Results. In Figure 1 we illustrate the different trade-offs between utility U and fairness F for both discrete and continuous experiment versions in each step of the process using $\lambda = 1$. Both experiments indicate that the policies obtained from the Bayesian approaches outperform

the marginal one in both fairness F and total utility T. In particular, for the continuous version, the effect is more evident in the early steps of the algorithm where we have a limited amount of data and thus greater model uncertainty. For the discrete data, we observe a constant advantage of the Bayesian approach over the marginal one. In addition, we can say that the bootstrap method is a good approximation of the Bayesian posterior. Finally, the policies obtained from the Bayesian approaches result in less variance in both continuous and discrete experiments.

4. Conclusion and Future Directions

In this work, we offer an algorithm solution to scale the concept of Bayesian fairness [1]. We propose the use of bootstrapping as an alternative to the posterior distribution of the parameters, along with an empirical formulation of the balance fairness metric suitable for continuous data scenarios. The results of our empirical study emphasize the significant role of accounting for uncertainty in the context of algorithmic fairness. Interesting research directions include the consideration of alternative fairness metrics, the incorporation of Bayesian fairness in reinforcement learning, and the application of the framework in modern deep learning models to investigate the advantages of Bayesian fairness in huge and complex datasets with higher degrees of uncertainty.

References

[1] C. Dimitrakakis, Y. Liu, D. C. Parkes, G. Radanovic, Bayesian fairness, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019). doi:10.1609/aaai.v33i01.3301509.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[3] J. M. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, CoRR abs/1609.05807 (2016). URL: http://arxiv.org/abs/1609.05807. arXiv:1609.05807.

[4] C. Russell, M. J. Kusner, J. Loftus, R. Silva, When worlds collide: integrating different counterfactual assumptions in fairness, Advances in neural information processing systems 30 (2017).

[5] J. Ali, P. Lahoti, K. P. Gummadi, Accounting for model uncertainty in algorithmic discrimination, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 336–345. URL: https://doi.org/10.1145/3461702.3462630. doi:10.1145/3461702.3462630.

[6] A. Singh, D. Kempe, T. Joachims, Fairness in ranking under uncertainty, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 11896–11908. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf.

[7] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, M. Srikumar, A. Weller, A. Xiang, Uncertainty

as a form of transparency: Measuring, communicating, and using uncertainty, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, Association for Computing Machinery, 2021, p. 401–413. doi:10.1145/3461702.3462571.

[8] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of data and analytics, Auerbach Publications, 2016, pp. 254–264.