Generalized Bayesian Inference with Fairness Constraints

Tin Lok James Ng

Trinity College Dublin, Ireland

June 15, 2023

Machine Learning Algorithms

Machine learning algorithms, which leverage data to make predictions, are playing an increasingly significant role in our everyday lives.

- Credit Scoring
- Customer Churn Prediction
- Healthcare Diagnosis
- Personalized Recommendations

Fair Machine Learning

There is growing awareness of the potential biases that can be present in machine learning algorithms due to biased or unrepresentative training data.

These biases can result in discriminatory behavior towards certain populations.

- Racial bias in predictive Policing
- Bias in loan approval
- Algorithmic hiring bias

Machine Learning Fairness in Insurance

- Potential biases may arise in underwriting, pricing, claims processing, etc.
- Methodologies for discrimination free insurance pricing (Xin and Huang, 2022; Lindholm et al., 2023).

Fairness Metrics

The scientific literature on fairness in machine learning has indeed emphasized two key aspects (Mehrabi et al., 2021):

1. measuring and assessing fairness,
2. mitigating unfairness when necessary.

A growing number of fairness definitions:

- observational vs. causality-based criteria,
- group vs. individual criteria.
- fairness is a multi-faceted concept,
- conflicts between different fairness definitions.

Unfairness Mitigation

Unfairness mitigation methods in machine learning can be categorized into three main approaches:

- Pre-processing approaches: modifying the training data before it is fed into the learning algorithm.
- In-processing approaches: modify the learning algorithm to incorporate fairness constraints.
- Post-processing approaches: modifying the output of a trained model.

Bayesian Approaches to Fair Machine Learning

Bayesian approaches to fair machine learning have been relatively understudied and under-utilized. However, they offer several benefits:

- incorporation of parameter uncertainty,
- quantification of uncertainty in fairness metrics,
- transparent decision-making.

Many fairness metrics or their relaxations can be represented as convex constraints.

Bayesian inference with parameter constraints arises in various contexts.

Bayesian Inference with Fairness Constraints (Extension of Sen et al. (2018))

We incorporate fairness metrics as fairness constraints into the Bayesian inference.

- Let Θ be the parameter space with norm $||\cdot||$.
- Let $\tilde{\Theta}_n \subset \Theta$ be a closed and non-empty constraint set which may depend on the data $U^{(n)} := (U_1, \ldots, U_n)$.
- Define the projection operator $T_{\tilde{\Theta}_n} : \Theta \to \mathcal{P}(\tilde{\Theta}_n)$:

$$T_{\tilde{\Theta}_n}(\theta) = \{\tilde{\theta} \in \tilde{\Theta}_n : ||\theta - \tilde{\theta}|| = \text{dist}(\theta, \tilde{\Theta}_n)\}, \tag{1}$$

where

$$\text{dist}(\theta, \tilde{\Theta}_n) = \inf\{||\theta - \tilde{\theta}|| : \tilde{\theta} \in \tilde{\Theta}_n\}.$$

- $T_{\tilde{\Theta}_n}(\theta)$ is the set of best approximation points $\tilde{\theta} \in \tilde{\Theta}_n$ for θ.

Bayesian Inference with Fairness Constraints

- Suppose further the set $\tilde{\Theta}_n$ is also convex. Then $T_{\tilde{\Theta}_n}(\theta)$ exists and is unique for all $\theta \in \Theta$, and $T_{\tilde{\Theta}_n}$ becomes a measurable map from Θ to $\tilde{\Theta}_n$.
- Let Π_Θ be a prior distribution on $(\Theta, \mathcal{B}_\Theta)$ that places positive mass on $\tilde{\Theta}_n$, and $\Pi_\Theta^{(n)}$ be the corresponding posterior distribution.
- $T_{\tilde{\Theta}_n}$ induces the measure $\tilde{\Pi}_{\tilde{\Theta}_n}^{(n)}$ on $(\tilde{\Theta}_n, \mathcal{B}_{\tilde{\Theta}_n})$ such that for any $\tilde{B} \in \mathcal{B}_{\tilde{\Theta}_n}$

$$\tilde{\Pi}_{\tilde{\Theta}_n}^{(n)}(\tilde{B}) = \Pi_\Theta^{(n)}(T_{\tilde{\Theta}_n}^{-1}\tilde{B}), \tag{2}$$

  where $T_{\tilde{\Theta}_n}^{-1}\tilde{B} = \{\theta \in \Theta : T_{\tilde{\Theta}_n}\theta \in \tilde{B}\}$.
- $\Pi_\Theta^{(n)}$ and $\tilde{\Pi}_{\tilde{\Theta}_n}^{(n)}$ are referred to as the unconstrained posterior and constrained posterior, respectively.

Asymptotic concentration of constrained posterior distributions

- Let Θ be a separate Hilbert space.
- let $\tilde{\Theta}_n$ be a sequence of subsets of Θ that are non-empty, closed, and convex under ($\mathbb{P}_{\theta_0}$ - a.s.)
- Suppose d_Θ is bi-Lipschitz with respect to $(\Theta, ||\cdot||)$, i.e., there exists a constant $c \geq 1$ such that $c^{-1}||\theta - \theta'|| \leq d_\Theta(\theta, \theta') \leq c||\theta - \theta'||$ for all $\theta, \theta' \in \Theta$.
- Suppose the unconstrained posterior concentrates at $\theta_0 \in \Theta$ with rate ϵ_n.

Then the constrained posteriors concentrate at the sequence $T_{\tilde{\Theta}_n}\theta_0$, $n = 1, 2, \ldots$ with rate at least ϵ_n:

$$\tilde{\Pi}^{(n)}_{\tilde{\Theta}_n}\{\theta : d_\Theta(\theta, T_{\tilde{\Theta}_n}\theta_0) > c^2 M_n \epsilon_n\} \to 0,$$

in $\mathbb{P}^{(n)}_{\theta_0}$-probability for every $M_n \to \infty$.

Sampling from $\tilde{\Pi}^{(n)}_{\tilde{\Theta}_n}$

Sampling from $\tilde{\Pi}^{(n)}_{\tilde{\Theta}_n}$ is straightforward.

1. We ignore the constraint set $\tilde{\Theta}_n$ and sample from the unconstrained posterior, $\theta_1, \ldots, \theta_m \sim \Pi_\Theta$.
2. Apply the map $T_{\tilde{\Theta}_n}$ to the sample $\theta_1, \ldots, \theta_m$ to obtain $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$.

It follows that $\tilde{\theta}_1, \ldots, \tilde{\theta}_m \sim \tilde{\Pi}^{(n)}_{\tilde{\Theta}_n}$.

Gibbs Posterior (Bissiri et al., 2016; Syring and Martin, 2023)

Bayesian framework requires specifying a statistical model/likelihood which has a number of potentially negative consequences:

- Risk of model mis-specification,
- Complicated models with many nuisance parameters,
- Quantity of interest (e.g. a quantile) may be independent of a statistical model.

Gibbs posterior replaces the log-likelihood function with the *empirical risk*

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(u_i).$$

For example, in a regression setting, $u = (x, y)$ and θ is a function, $\ell_\theta(u) = |y - \theta(x)|$.

Gibbs posterior (with learning rate $\omega > 0$) is defined as

$$\Pi_\Theta^{(n)} \propto e^{-\omega n R_n(\theta)} \Pi(\theta).$$

Constrained Gibbs Posterior

Given closed, non-empty, and convex sets $\tilde{\Theta}_n$, and a projection map $T_{\tilde{\Theta}_n}$, *constrained Gibbs posteriors* $\tilde{\Pi}_{\tilde{\Theta}_n}$ can be defined analogously.

Sampling from the constrained Gibbs posterior $\tilde{\Pi}_{\tilde{\Theta}_n}$ is also analogous to sampling from the constrained posterior.

The constrained Gibbs posteriors can be shown to concentrate around $T_{\tilde{\Theta}_n}\theta^*$ as $n \to \infty$ where θ^* is the *risk minimizer*.

Application: Adult (Census Income) Data Set

Classification task: predict whether the annual income of a person exceeds 50,000 US dollars.

- 48,842 instances
- 15 features (6 numerical and 9 categorical)
- We focus on 4 features (age, capital gain, education number of years, hours per week) for illustration purposes.

Constraint: bound the difference in false positive rates for males and females.

Application: Adult (Census Income) Data Set

We consider the loss function

$$1\{Y \neq \phi_\theta(X)\},$$

where

$$\phi_\theta(X) = 1\{\theta^T X > 0\}.$$

The corresponding empirical risk is

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} 1\{y_i \neq \phi_\theta(x_i)\}.$$

The constraint set $\tilde{\Theta}_n$ is defined by bounding the squared difference between false positive rates for males and females above by a constant c.

Results ($c = 0.01$)

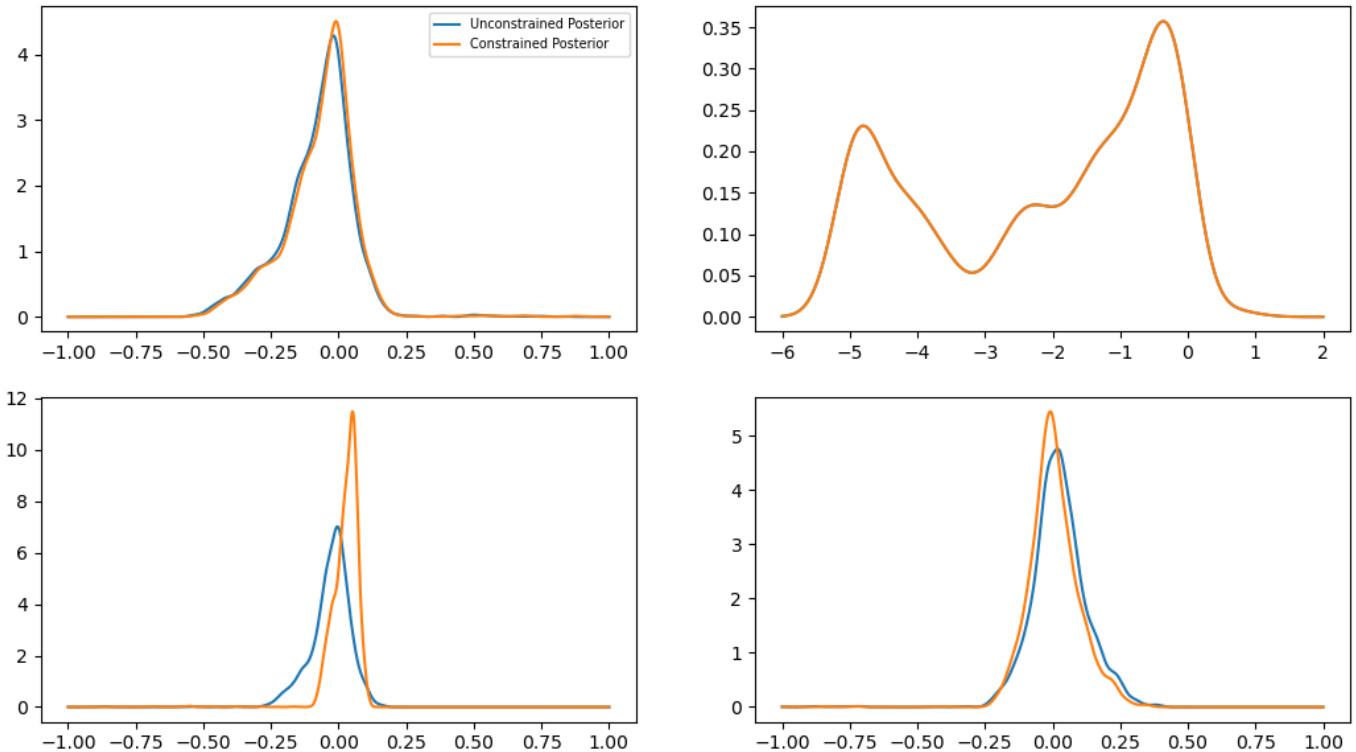

Figure: Top left: age, top right: capital gain, bottom left: education number of years, bottom right: hours per week

Results ($c = 0.1$)

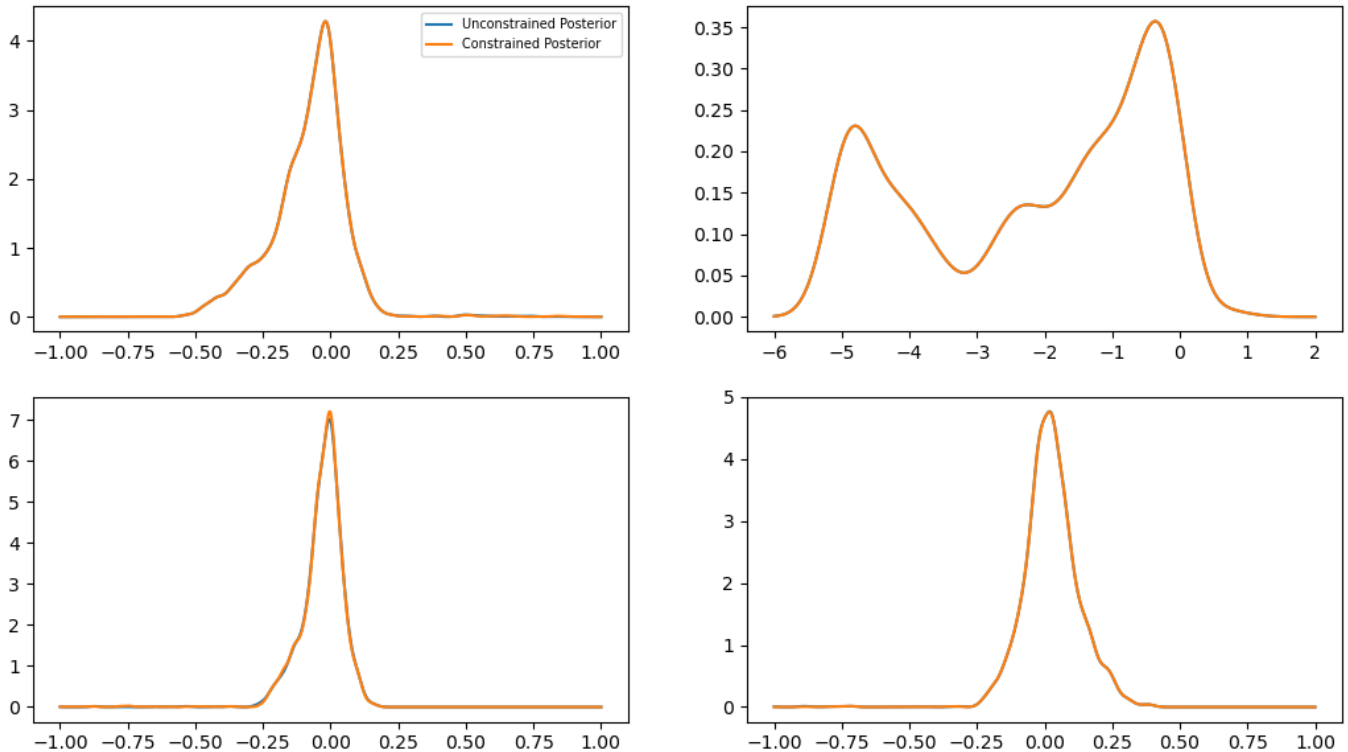

Figure: Top left: age, top right: capital gain, bottom left: education number of years, bottom right: hours per week

Selected References

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016), "A general framework for updating belief distributions," *Journal of the Royal Statistical Society Series B*, 78, 1103–1130.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2023), "What is fair? Proxy discrimination vs. demographic disparities in insurance pricing," .

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021), "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, 54.

Sen, D., Patra, S., and Dunson, D. (2018), "Constrained inference through posterior projections," .

Syring, N. and Martin, R. (2023), "Gibbs posterior concentration rates under sub-exponential type losses," *Bernoulli*, 29, 1080 – 1108.

Xin, X. and Huang, F. (2022), "Anti-discrimination insurance pricing: regulations, fairness criteria, and models," *Fairness Criteria, and Models*.