JMLR: Workshop and Conference Proceedings 1–9

Individual Fairness in Bayesian Neural Networks

Alice Doherty *Trinity College Dublin* ADOHERT@TCD.IE
Matthew Wicker *The Alan Turing Institute* M.WICKER@TURING.AC.UK
Luca Laurenti *Delft University of Technology* L.LAURENTI@TUDELFT.NL
Andrea Patane *Trinity College Dublin* APATANE@TCD.IE

Abstract

We study Individual Fairness (IF) for Bayesian neural networks (BNNs). Specifically, we consider the ϵ-δ-individual fairness notion, which requires that, for any pair of input points that are ϵ-similar according to a given similarity metrics, the output of the BNN is within a given tolerance $\delta > 0$. We leverage bounds on statistical sampling over the input space and the relationship between adversarial robustness and individual fairness to derive a framework for the systematic estimation of ϵ-δ-IF, designing *Fair-FGSM* and *Fair-PGD* as global, fairness-aware extensions to gradient-based attacks for BNNs. We empirically study IF of a variety of approximately inferred BNNs with different architectures on fairness benchmarks, and compare against deterministic models learnt using frequentist techniques. Interestingly, we find that BNNs trained by means of approximate Bayesian inference consistently tend to be markedly more individually fair than their deterministic counterparts.

1. Introduction

Deep learning models have achieved state-of-the-art performance in a wide variety of tasks (Goodfellow et al., 2016). Several calls for caution have, however, recently been raised about their deployment in tasks where fairness is of concern (Barocas and Selbst, 2016). In fact, Neural Networks (NNs) have been found to reinforce negative biases from sensitive datasets (Bolukbasi et al., 2016), discriminating against individuals on the basis of attributes such as gender or race. To address this, research efforts have been directed at both measuring the fairness of NNs, their de-biased training, as well as defining precise notions of fairness.

Given a BNN f and a similarity metric between individuals d_{fair}, which encodes a task-dependent notion of similarity (Ilvento, 2020), *Individual Fairness* (IF) enforces that all pairs of *similar* individuals in the input space get treated similarly by f (Dwork et al., 2012). As opposed to the statistical nature of *group fairness* (Mehrabi et al., 2021), IF aims at computing worst-case bias measures over a model input space. Consequently, albeit being defined on the full input space and over a fairness similarity metric, because of its worst-case nature, IF has been linked to adversarial robustness (Yeom and Fredrikson, 2021). As it has recently been show that Bayesian Neural Networks (BNNs) have a tendency to be less fragile to adversarial attacks than their frequentist counter-parts (Carbone et al., 2020), it is natural to wonder whether approximate Bayesian inference may also have a positive impact over the IF of a neural network. However, to the best of our knowledge, no work has been conducted along these lines of inquire.

In this paper, we investigate the IF of BNNs and empirically evaluate it on various benchmarks. While exact computations of IF in BNNs is infeasible due to their non-convexity,

© A. Doherty, M. Wicker, L. Laurenti & A. Patane.

we exploit the relationship between IF and adversarial robustness (Yeom and Fredrikson, 2021; Benussi et al., 2022) to develop a framework for the adaptation of adversarial attack methods for IF. In particular, we explicitly instantiate *Fair-FGSM* and *Fair-PGD* as extensions of their corresponding adversarial attacks (Goodfellow et al., 2015; Athalye et al., 2018) by employing gradient steps modifications and projections specific to d_{fair} metrics commonly used in the fairness literature. Furthermore, while attack methods estimate IF locally around a given individual, we use concentration inequalities (Boucheron et al., 2004) to statistically bound the worst-case expected IF over all pairs of individuals.

We perform an empirical evaluation of IF in BNNs on a selection of benchmarks, including the Adult dataset (Dua and Graff, 2017) and Folktables (Ding et al., 2021), and on a variety of different architectural, approximate Bayesian inference, and similarity metric parameters. We compare the results obtained by BNNs with those obtained by deterministic NNs and deep ensembles. We find that BNNs consistently outperform their deterministic counterparts in terms of individual fairness. That is, albeit still learning certain biases, all things being equal, BNNs show a tendency to be *fairer* than their deterministic counterparts. We finish the paper with an empirical analysis inquiring this interesting property and a discussion on why approximate Bayesian inference may lead to fairer prediction models.

2. Individual Fairness for BNNs

We consider a NN $f^w : X \to \mathbb{R}^m$ trained on a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ composed by N points, where w is the aggregated vector of weights and biases, $X \subset \mathbb{R}^n$ is the input space, and n and m are respectively the input and output dimensions. For simplicity, we focus on the case of binary classification, however, our results can be trivially extended to the regression and multi-class case. In a Bayesian setting, w is sampled from a random variable $\mathbf{w}$ whose posterior distribution is obtained by application the Bayes rule: $p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w)$, where $p(w)$ is the prior and $p(\mathcal{D}|w)$ is the likelihood. The posterior predictive distribution π over an input x is then obtained by averaging the posterior distribution over a sigmoid function σ, i.e., $\pi(x) = \int \sigma(f^w(x))p(w|\mathcal{D})dw$. To define individual fairness for BNNs, we adopt the ϵ-δ-IF (John et al., 2020), which depends on a similarity metric $d_{fair} : X \times X \to \mathbb{R}_{\geq 0}$ encapsulating task-dependent information about sensitive attributes.

Definition 1 *Given $\epsilon \geq 0$ and $\delta \geq 0$, we say that a BNN $f^{\mathbf{w}}$ is ϵ-δ-individually fair iff $\forall x', x'' \in X s.t.\ d_{fair}(x', x'') \leq \epsilon \implies |\pi(x') - \pi(x'')| \leq \delta$, that is, if the predictive distribution of ϵ-similar individuals are δ-close to each other.*

Notice that while related to notion of adversarial robustness for BNNs (Cardelli et al., 2019; Wicker et al., 2020), Definition 1 has some key differences. First, rather than being a local definition specific to a test point and its neighbourhood, individual fairness looks at the worst-case behaviour over the whole input space, i.e., simultaneously looking at neighbourhoods around all the input points in X. Furthermore, while adversarial attacks are generally developed for an ℓ_p metric, individual fairness is built around a task-specific similarity metric d_{fair}. Intuitively, d_{fair} needs to encode the desired notion of similarity between individuals (Dwork et al., 2012). While a number of metrics have been discussed

in the literature (Ilvento, 2020), we focus on the following ones, which are widely used and can be automatically learnt from data (Yurochkin et al., 2020):[1]

- **Weighted ℓ_p.** In this case $d_{\text{fair}}(x', x'')$ is defined as a weighted version of an ℓ_p metric, i.e. $d_{\text{fair}}(x', x'') = \sqrt[p]{\sum_{i=1}^{n} \theta_i |x'_i - x''_i|^p}$, where the weights θ_i can be set accordingly to their correlation with the sensitive attribute.
- **Mahalanobis distance.** In this case we have $d_{\text{fair}}(x', x'') = \sqrt{(x' - x'')^T S^{-1} (x' - x'')}$, for a given positive semi-definite (SPD) matrix S. Intuitively, S accounts for the intra-correlation of features to capture latent dependencies w.r.t. the sensitive features.

3. Fairness attacks

Definition 1 could be reformulated according to the following optimization problem

$$\delta^* = \max_{x' \in X} \max_{\substack{x'' \in X \\ d_{fair}(x', x'') \leq \epsilon}} |\pi(x') - \pi(x'')|. \tag{1}$$

Checking IF is then equivalent to check whether $\delta^* \leq \delta$ or not. In Eqn (1), the inner maximization problem finds the point in a specific d_{fair}-neighbourhood that maximizes the variation in the predictive posterior, while the outer one considers the global part of the definition. We proceed by solving the inner-optimisation problem with gradient-based techniques, and rely on statistical methods to solve with high confidence the outer problem.

Inner Problem Because of the non-convexity of NN architectures, exact computation of the inner maximization problem in Eqn (1) is generally infeasible. We proceed by adapting gradient-based optimization techniques commonly used to detect adversarial attacks to our setting. In particular, in the case of the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015), given the log-loss L associated to the likelihood function (e.g., the binary entropy for the sigmoid), and a point x with associated label y, we obtain that:

$$x_{x',\max} \approx x' + \eta \cdot \text{sgn}(\mathbb{E}_{w \sim p(w|\mathcal{D})}[\nabla_x L(f^w(x), y)]).$$

$\eta \in \mathbb{R}^n$ scales the attack strength according to ϵ and d_{fair}. It is the first ingredient that needs to be adapted for ϵ-δ-IF. In the case of the weighted ℓ_p metric, η can be simply defined as the vector with entries $\eta_i = \epsilon/\sqrt{\theta_i}$. For the Mahalanobis distance, given x' and ϵ, $d_{fair}(x', x'') \leq \epsilon$ describes a hyper-ellipsoid centred in x'. We can hence set η to be the vector whose generic component is scaled according to the ellipsoid axis, i.e., $\eta_i = \epsilon\sqrt{S_{ii}}$. Furthermore, while one is guaranteed to remain inside the d_{fair}-neighbourhood in the ℓ_p case, for the Mahalanobis metric one needs to check whether the attack is inside the ϵ-d_{fair} ball or not. In the latter case, we set up a local line search problem to project $x_{x',\max}$ back to a point inside the ball. Other gradient-based attacks such as PGD (Madry et al., 2018) or CW (Carlini and Wagner, 2017) can be similarly adapted by taking consecutive gradients. Similarly, other more general attacks could be employed (Yuan et al., 2020; Däubener et al., 2020) as well as training techniques (Liu et al., 2019; Wicker et al., 2021). In Section 4, we will consider both FGSM and PGD, which we denote with *Fair-FGSM* and *Fair-PGD* to distinguish them from the adversarial attack counter-parts.

1. While not directly investigated in this paper, our fairness attacks can be generalised to similarity metrics built over embeddings (Ruoss et al., 2020), by attacking the embeddings as well.

Outer Problem The final step needed for the computation ϵ-δ-IF is to check whether:

$$\max_{x' \in X} \left(\pi(x_{x',\max}) - \pi(x')\right) \leq \delta. \tag{2}$$

Again, exact computation of Eqn (2) is infeasible as it would require solving a non-convex optimisation problem over a possibly large input space. We relax the problem and instead compute the probability that a point sapled from X is fair. In order to do that we iid sample n_{samples} points from X and build the following empirical estimator: $\hat{p} = \sum_{i=1}^{n_{\text{samples}}} \frac{\phi(x_i)}{n_{\text{samples}}}$, where $\phi(x_i) = \begin{cases} 1 & \text{if } \pi(x_{x_i,\max}) - \pi(x_i) \leq \delta \\ 0 & \text{otherwise} \end{cases}$. Note that $\phi(x_i)$ is a binary variable. Consequently, $\hat{p}$ approximate the probability that a point sampled from X satisfies the inner condition in Eqn (1). Lemma 2 below is an application of the Chernoff bound (Hellman and Raviv, 1970) and guarantees that if n_{samples} is large, then the error introduced by the empirical estimator can be made arbitrarily small with high confidence.

Lemma 2 *Consider $n_{samples}$ points sampled iid from X. Call $p = \mathbb{E}_{x' \sim P}[\phi(x')]$. Then for $\gamma, \theta > 0$ such that $n_{samples} > \frac{1}{2\theta^2} \log\left(\frac{2}{\gamma}\right)$, it holds that $P(|\hat{p} - p| > \theta) \leq \gamma$.*

Therefore, in order to approximate the solution of the optimisation problem in Equation (2) with statistical confidences θ and γ we compute the smallest δ such that $\hat{p} = 1$.

4. Results

We employ *Fair-FGSM* and *Fair-PGD* to evaluate and compare the IF of deterministic Neural Networks (dNNs) and Bayesian Neural Networks (BNNs). We approximate the BNNs' posterior either with VI (in particular using variational online Gauss-Newton (Khan et al., 2018)) or with Hamiltonian Monte Carlo (HMC) (Neal, 2012). The NNs are trained to solve the taks associated to the Adult (Dua and Graff, 2017) and the Folktables (Ding et al., 2021) datasets, where we take gender as the sensitive attribute.[2]

Parametric Analysis We evaluate how IF is affected by the architecture. Namely, across the two similarity metrics, we begin by fixing $\epsilon = 0.1$ and analysing how δ is affected by the numbers of hidden layers and of neurons per layer, in ranges typically used for these datasets (Yurochkin et al., 2020; John et al., 2020; Yeom and Fredrikson, 2021; Benussi et al., 2022). The results for these analyses are plotted in the heatmaps in Figure 1. The three columns of heatmaps respectively show the IF estimations for models learned with dNN, VI, and HMC. The first two rows give the results for the Adult dataset, respectively for *Fair-FGSM* with the weighted ℓ_p metric and *Fair-PGD* with the Mahalanobis similarity metric. The last two rows show the analogous results on the Folks dataset.

We immediately observe how, in the overwhelming majority of the cases, the models learnt by HMC are markedly fairer (i.e., they have lower values of δ) than their deterministic counter-parts. The same generally applies also to VI models, except for the IF obtained on the Adult dataset by means of *Fair-PGD* for the Mahalanobis metric (second row, second column). This combined observations almost perfectly mirrors the behaviour of approximate

2. Code to run the experiments can be found at https://github.com/alicedoherty/bayesian-fairness.

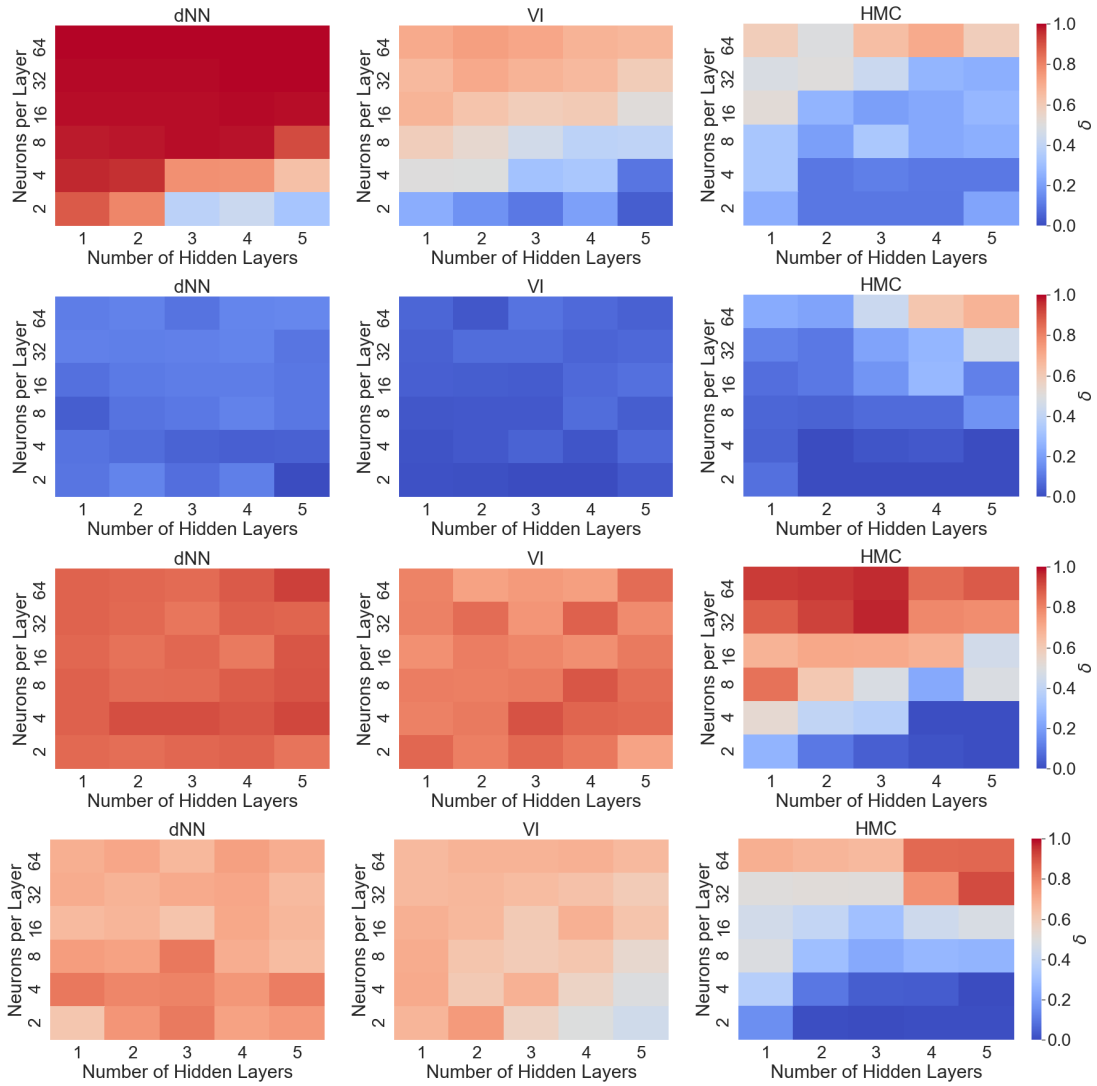

Figure 1: Heatmaps for the estimations of individual fairness (δ) across various architectures, inference methods (dNN, VI or HMC), datasets, and similarity metrics. The first two rows give results for the Adult dataset (respectively *Fair-FGSM* with ℓ_p, and *Fair-PGD* with Mahalanobis). The last two for the Folktables dataset.

Bayesian inference in adversarial robustness settings (Carbone et al., 2020). Interestingly, both adversarial robustness and overfitting generally worsen as the model becomes deeper. Instead, we observe that more often than not this pattern is reversed in IF. We posit that while the NNs are indeed becoming more fragile to attacks as they get deeper, it is also less reliant on each specific input feature, rather it builds on a non-trivial representation of them – therefore having a tendency to perform better on IF tasks. On the other hand, as expected, when the number of neurons per layer increases, we get less fair models. In the

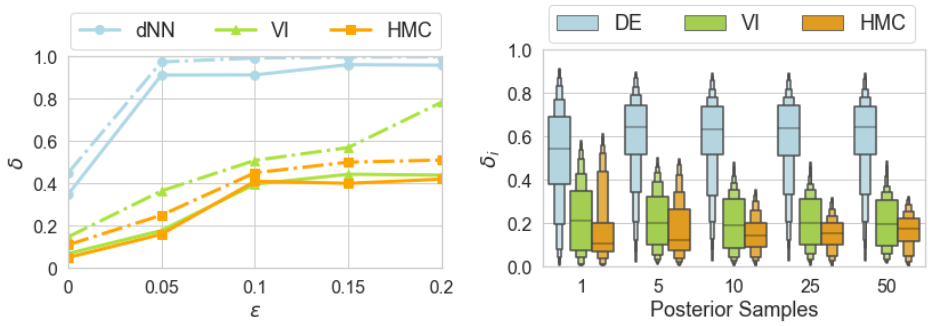

Figure 2: **Left:** Fairness for different maximum distances ϵ for deterministic and Bayesian NNs with 3 layers and either 16 (thick lines) or 32 (dashed lines) neurons per layer. **Right:** Changes in empirical distribution of fairness across deterministic and Bayesian predictors computed over an increasing number of posterior samples.

left plot of Figure 2 we analyse how individual fairness is affected by ϵ. We give a selection of the results for an architecture with 3 layers and 16 or 32 neurons trained on the Adult dataset. Of course, we observe that as ϵ increases δ increases as well. Interestingly, we find that BNNs' tendency to be more fair is consistent across the different values of epsilon.

Analysis of Posterior Predictive We experimentally inquire on the reason behind BNNs' improved individual fairness. To this end, for a given BNN posterior predictive distribution, we solve the inner problem of Section 3 over 100 randomly sampled test points. Rather than doing this over the full posterior, we randomly sample 15 times a number k of weights realisation from the posterior, and compute the predictive posterior over them. We investigate values of k ranging from 1 (i.e., a single deterministic-like network extracted from the BNN) to 50. We also compare BNNs with an averaging of deterministic NNs in the form of a Deep Ensemble (DE) (Lakshminarayanan et al., 2017). The results for this analysis are given in the right plot of Figure 2. Surprisingly, even when using only one posterior sample from the BNN distribution, the models learnt with Bayesian approximation method are already significantly more fair than with the frequentist DE technique. Furthermore, especially in the case of HMC, the more samples are drawn from the posterior distribution at prediction time, the fairer the model becomes – with a 50% bias reduction when going from 1 to 50. On the other hand, notice that averaging does not help with DE.

5. Conclusion

In this paper we extended adversarial attack techniques to study individual fairness (IF) for BNNs. On a set of experiments we empirically showed that BNNs tend to be intrinsically fairer than their deterministic counter-parts. Furthermore, we showed how the particular approximate inference method has an impact on IF, with more approximate methods being less fair compared to approaches such as HMC. Finally, we have empirically shown how increased fairness is likely due to a combination of Bayesian training and Bayesian averaging, which may have a beneficial effect in reducing the magnitude of the gradients.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Elias Benussi, Andrea Patane, Matthew Wicker, Luca Laurenti, and Marta Kwiatkowska. Individual fairness guarantees for neural networks. *International Joint Conference on Artificial Intelligence*, 2022.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pages 208–240, 2004.

Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, and Guido Sanguinetti. Robustness of bayesian neural networks to gradient-based attacks. *Advances in Neural Information Processing Systems*, 33:15602–15613, 2020.

Luca Cardelli, Marta Kwiatkowska, Luca Laurenti, Nicola Paoletti, Andrea Patane, and Matthew Wicker. Statistical guarantees for the robustness of bayesian neural networks. *International Joint Conference on Artificial Intelligence*, 2019.

Nicholas Carlini and David Wagner. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017.

Sina Däubener, Joel Frank, Thorsten Holz, and Asja Fischer. Efficient calculation of adversarial examples for bayesian neural networks. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34:6478–6490, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

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, 2012.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Martin Hellman and Josef Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pages 749–758. PMLR, 2020.

Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *International conference on machine learning*, pages 2611–2620. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh. Adv-bnn: Improved adversarial defense through robust bayesian neural network. *International Conference on Learning Representations*, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. *Advances in Neural Information Processing Systems*, 33: 7584–7596, 2020.

Matthew Wicker, Luca Laurenti, Andrea Patane, and Marta Kwiatkowska. Probabilistic safety for bayesian neural networks. In *Conference on Uncertainty in Artificial Intelligence*, pages 1198–1207. PMLR, 2020.

Matthew Wicker, Luca Laurenti, Andrea Patane, Zhuotong Chen, Zheng Zhang, and Marta Kwiatkowska. Bayesian inference with certifiable adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 2431–2439. PMLR, 2021.

Samuel Yeom and Matt Fredrikson. Individual fairness revisited: transferring techniques from adversarial robustness. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 437–443, 2021.

Matthew Yuan, Matthew Wicker, and Luca Laurenti. Gradient-free adversarial attacks for bayesian neural networks. *Third Symposium on Advances in Approximate Bayesian Inference*, 2020.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *International Conference on Learning Representations*, 2020.