Inframarginality Audit of Group-Fairness

Arpita Biswas* Siddharth Barman* Amit Deshpande† Amit Sharma†

Consider algorithmic decisions in societally critical domains such as healthcare [9, 16], education [28], criminal justice [3, 6], policing [26, 15] or finance [14]. Typically, machine learning algorithms train models on the available data so as to maximize accuracy that often leads to differential outcomes or errors across various sensitive groups, e.g., race or gender [3, 24, 4, 10, 7, 8]. Previous work on group-fairness achieves equality or near-equality of certain metrics for such groups, typically, via a trade-off between accuracy and group-fairness. The popular group-fair notions studied in classification are disparate impact [18, 13, 30], statistical parity [19, 31, 12], and equalized odds [17, 20, 29]. All these notions share the common characteristic of equalizing one or more performance metrics (such as predictive prevalence, false positive rate, false negative rate) across groups. Thus, the group-fair algorithms aim to maximize accuracy subject to near-equality of one or multiple performance metrics across different groups. Such group-fairness constraints are often non-convex, and the group-level adjustments made by group-fair algorithms are often unfair to similar individuals. For instance, consider the medical domain where doctors assess the severity of a person's illness (risk probability) and prioritize treatment accordingly. It may be acceptable to prioritize patients based on a given reliable estimate of a patient's risk, but any deviation from this rule to incorporate group-fairness (with good intentions) may deprive some high-risk people from receiving treatment, thus being individually unfair.

In this paper, we study *infra-marginality* [26, 11], a concept which measures the deviation from individual-fairness. To remove infra-marginality, Simoiu et al. ([26]) propose taking decisions using a single threshold on the true outcome probability, whenever possible. Such a single-threshold classifier implements the high-level idea that legislation should apply equally to everyone and not be based on their group identities.

Our Contributions

- Our conceptual contribution is a quantitative measure of the degree of infra-marginality η_ϵ^C, extending past work that introduced infra-marginality as an important consideration for individual fairness [26, 11].
- We show that a classifier of high accuracy can be used to reliably estimate the infra-marginality of group-fair classifiers.
- We propose a method to audit classifiers. We evaluate on real-world datasets, Adult Income [21] and Medical datasets [2], and find that infra-marginality is an important concern: some group-fair classifiers suffer from high infra-marginality, close to a random classifier.
- We discuss that our definition is still useful to capture the trade-off between accuracy and infra-marginality with respect to any given threshold. In particular, high accuracy may not always imply low infra-marginality, especially when the decision threshold is different from 0.5.

We hope that our work motivates the inclusion of infra-marginality as an individual fairness metric in the growing work on auditing bias of algorithmic decisions [22, 27, 23, 1, 5, 25].

Acknowledgements

A part of this work was done when Arpita Biswas was an intern at Microsoft Research India, and she also gratefully acknowledges the support of a

*Indian Institute of Science. {arpitab,barman}@iisc.ac.in
†Microsoft Research. {amitdesh,amshar}@microsoft.com

Google PhD Fellowship Award. Siddharth Barman gratefully acknowledges the support of a Ramanujan Fellowship (SERB - SB/S2/RJN-128/2015) and a Pratiksha Trust Young Investigator Award.

References

[1] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

[2] Agency for Healthcare Research and Quality. Medical Expenditure Panel Survey. `https://meps.ahrq.gov/mepsweb/`, 2016.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: Theres software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 2016. https://www.propublica.org/article/machine-bias-riskassessments- in-criminal-sentencing.

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

[5] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

[7] Arpita Biswas and Suvam Mukherjee. Fairness Through the Lens of Proportional Equality. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS, Montreal, QC, Canada, 2019*, pages 1832–1834, 2019.

[8] Arpita Biswas and Suvam Mukherjee. Ensuring fairness under prior probability shifts. *arXiv preprint arXiv:2005.03474*, 2020.

[9] Irene Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems NIPS*, pages 3543–3554, 2018.

[10] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018.

[11] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[14] Mark J Furletti. An overview and history of credit reporting. 2002. http://dx.doi.org/10.2139/ssrn.927487.

[15] Sharad Goel, Maya Perelman, Ravi Shroff, and David Alan Sklansky. Combatting police discrimination in the age of big data. *New Criminal Law Review: In International and Interdisciplinary Journal*, 20(2):181–232, 2017.

[16] Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 2018.

[17] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[18] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[19] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[20] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science*, 2017.

[21] Ronny Kohavi and Barry Becker. Adult Data Set . `https://archive.ics.uci.edu/ml/datasets/adult`, 1996. [accessed 20-February-2019].

[22] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. Accountable algorithms. *U. Pa. L. Rev.*, 165:633, 2016.

[23] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 626–633, 2017.

[24] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5): 582–638, 2014.

[25] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[26] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[27] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310, 2018.

[28] Robin D Tierney. Fairness in classroom assessment. *SAGE handbook of research on classroom assessment*, pages 125–144, 2013.

[29] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953, 2017.

[30] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[31] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.