Fair Machine Learning Through Post-processing: The Case of Predictive Parity

Joachim Baumann[1,2,*], Anikó Hannák[1] and Christoph Heitz[2]

[1]*University of Zurich, Zurich, Switzerland*
[2]*Zurich University of Applied Sciences, Zurich, Switzerland*

Abstract
Post-processing is a bias mitigation technique proposed by the algorithmic fairness community to ensure the fairness of decision making systems that rely on machine learning (ML). Several works have provided solutions to optimally post-process ML-based systems for taking decisions that are fair w.r.t. specific group fairness criteria such as *statistical parity* (*SP*) or *equality of opportunity* (*EOP*) [1, 2]: here, optimal decision rules always take the form of lower-bound threshold rules. We investigate the case of another important fairness criterion called *predictive parity*. We show that for this notion of fairness, the optimum decision rules are different: In some cases, the optimum decision rule consists in applying an *upper-bound* threshold rule for (at least) one group. This result is counter-intuitive: For a decision maker, it may be optimal to leave out the most promising individuals of a group in order to generate predictive parity in a globally optimal way. This is in contrast to the analogous solutions for SP and EOP. Furthermore, even if between-group fairness is achieved, within-group fairness may be created. We encourage readers to consult the complete manuscript [3], which was published at FAccT 2022.

Keywords
Fairness, predictive parity, post-processing, optimal decision rules, group fairness, sufficiency

Background Prediction-based binary decision systems are not fair by default. In order to measure and eventually correct for discrimination against certain social groups, different mathematical notions of so-called *group fairness criteria* have been proposed [4, 5]. One line of research is concerned with optimal post-processing of ML models, deriving decision rules that satisfy some group fairness constraint while still leading to efficient decisions [1, 2, 6, 7]. Following this approach, we formulate the goal of fairness as a constrained optimization problem for a decision maker, assuming that goal is to maximize a decision maker's utility function while satisfying some fairness constraint [8]. Such optimal decision rules have been derived for the group fairness criteria (conditional) statistical parity, equality of opportunity (also called True Positive Rate (TPR) parity), False Positive Rate (FPR) parity, and Equalized Odds (EO) [1, 2, 6]. It has been shown that lower-bound threshold rules characterize optimal decision rules that satisfy these fairness constraints.[1]

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
*Corresponding author.
baumann@ifi.uzh.ch (J. Baumann); hannak@ifi.uzh.ch (A. Hannák); christoph.heitz@zhaw.ch (C. Heitz)
0000-0003-2019-4829 (J. Baumann); 0000-0002-0612-6320 (A. Hannák); 0000-0002-6683-4150 (C. Heitz)

[1]In the fair ML literature, so-called *thresholding*, applied to the prediction $p = P[Y = 1]$, is arguably the most typical decision rule for probabilistic classifiers, also because of its conceptual similarity to the way humans take decisions [9, 10]. *Lower-bound threshold rules* are decision rules given by $D = 1$, if $p > t$, $D = 0$ else. For the EO

Table 1
Group fairness criteria

Criterion	Equation
Predictive parity	$P(Y=1\|D=1,A=0)=P(Y=1\|D=1,A=1)$
FOR parity	$P(Y=1\|D=0,A=0)=P(Y=1\|D=0,A=1)$
Sufficiency	$P(Y=1\|D=i,A=0)=P(Y=1\|D=i,A=1), i\in\{0,1\}$

In computer science and in philosophy literature, *predictive parity* (also known as parity of positive predictive values (PPV) or precision across groups) is often mentioned as one of the main fairness criteria [4, 5, 10–22]. Related fairness criteria are false omission rate (FOR) parity and sufficiency. Most prominent is probably the case of the 2016 debate surrounding the recidivism risk prediction tool COMPAS [11]. In response to [23] suggesting that the tool systematically disadvantages black defendants, Northpointe (the developers of COMPAS) claimed that their tool is fair because it satisfies predictive parity and FOR parity [24].[2]

Research gap Optimal post-processing solutions are unknown for fairness criteria that condition on the decision, namely, *predictive parity*, *FOR parity*, and *sufficiency* (which combines the former two) – see Table 1 for the definitions w.r.t the decision D, label Y, and binary groups $A=\{0,1\}$. We close this gap by deriving optimal decision rules that satisfy these group fairness criteria through post-processing.

Findings We provide formal proof showing that optimal decision rules satisfying predictive parity or FOR parity take the form of group-specific threshold rules, as has been found for other fairness criteria. However, surprisingly, under some conditions (depending on the populations and the applied utility function), *upper-bound* thresholds are optimal: a decision maker would assign a positive decision ($D=1$) to individuals with a *low* probability of belonging to the positive class ($Y=1$). This is visualized in Figure 1 where the probability (p) density functions are shown for two groups 0 and 1 and the colored parts represent those individuals that receive a positive decision: Without any fairness constraints, a single uniform lower-bound threshold would be optimal (i.e., $D=1$ if $p>t_0$), resulting in different PPVs for the two groups (denoted by PPV_{t0} in Figure 1). To ensure predictive parity, it is optimal to apply a lower-bound threshold to Group 0 (i.e., $D=1$ if $p>t_1$) and an upper-bound threshold to Group 1 (i.e., $D=1$ if $p<t_2$), resulting in a PPV of $PPV_{t1,t2}$ for both groups. In this situation, any rational decision maker is willing to omit the most promising individuals from Group 1 in order to achieve predictive parity – which is highly counter-intuitive.

Furthermore, we provide a solution for the optimal decision rules that satisfy sufficiency. We find that this definition of fairness requires randomization (similar to the EO criterion [2]).

criterion, randomization involving two such thresholds is needed to satisfy EOP and FPR parity simultaneously [2].

[2] In addition to recidivism prediction, predictive parity is also prevalent in predictive policing [25] (where the metric is usually called *hit rate* or *outcome test*) and in personalized online ads (where the notion of *click through rates* [26], which is an equivalent metric, is omnipresent).

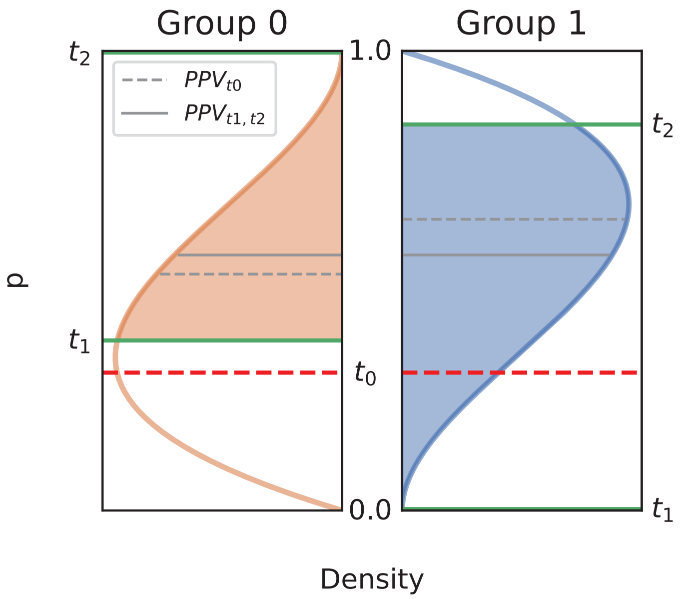

Figure 1: Optimal decisions under predictive parity may require an upper-bound threshold rule. The uniform threshold t_0 denotes the optimal unconstrained decision rule. The positive predictive values (PPVs) that result from this decision rule are denoted by PPV_{t0}. The group-specific thresholds $[t_1, t_2]$ (where t_1 denotes the lower- and t_2 the upper-bound, meaning that any individuals with a probability $p \in [t_1, t_2]$ is assigned the decision $D = 1$ and $D = 0$ otherwise) denote the optimal decision rule under predictive parity. The PPVs that result from this optimal fair decision rule are denoted by $PPV_{t1,t2}$.

Recently, we have conducted additional experiments, showing that the solution provided in this paper is effective in mitigating many different types of bias that can be present in ML-based decision making systems [27]. These experiments show that post-processing techniques [1–3] can cope with historical biases on the features or labels and even with measurement bias on the features. However, measurement bias on the label is particularly difficult to mitigate, and existing (post-processing) solutions are limited since they rely on the biased proxy of the label.

Ethical implications In many cases, individuals with $Y = 1$ have, morally speaking, a higher claim to a positive decision $D = 1$ than individuals with $Y = 0$, and vice versa. For example, in the case of COMPAS, this means that individuals with a *lower* probability of recidivism (i.e., a low $p = P[Y = 1]$) should preferably be released ($D = 0$). However, requiring a rational decision maker to fulfill predictive parity can result in releasing individuals with *higher* recidivism probabilities instead. This represents a case of *within-group unfairness*: achieving between-group fairness at the expense of within-group fairness may be problematic from an ethical perspective.

Society increasingly calls for fairer algorithms. At least for the group fairness criteria predictive parity, FOR parity, and sufficiency, our work shows that imposing such fairness criteria on utility-maximizing decision makers may lead to ethically problematic outcomes.

Acknowledgments

We thank the other members of our project and colleagues (Corinna Hertweck, Eleonora Viganò, Ulrich Leicht-Deobald, Serhiy Kandul, Markus Christen, Nicolò Pagan, Stefania Ionescu, Aleksandra Urman, and Leonore Röseler) for their helpful comments and suggestions. We also thank the anonymous reviewers for their feedback. This work was supported by Innosuisse – grant number 44692.1 IP-SBM – and by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF) – grant number 187473.

References

[1] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic Decision Making and the Cost of Fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 797–806. doi:10.1145/3097983.3098095.

[2] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, pp. 3323–3331. arXiv:1610.02413.

[3] J. Baumann, A. Hannák, C. Heitz, Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2315–2326. doi:https://doi.org/10.1145/3531146.3534645.

[4] A. Narayanan, Translation tutorial: 21 fairness definitions and their politics, in: Proc. Conf. Fairness Accountability Transp., New York, USA, 2018.

[5] S. Verma, J. Rubin, Fairness Definitions Explained, in: Proceedings of the International Workshop on Software Fairness, FairWare '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–7. doi:10.1145/3194770.3194776.

[6] Z. C. Lipton, A. Chouldechova, J. McAuley, Does mitigating ML's impact disparity require treatment disparity?, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates, Inc., 2018, pp. 8136–8146. URL: https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf.

[7] A. K. Menon, R. C. Williamson, The cost of fairness in binary classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, New York, NY, USA, 2018, pp. 107–118. URL: http://proceedings.mlr.press/v81/menon18a.html.

[8] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 8 (2021) 141–163. doi:10.1146/annurev-statistics-042720-125902.

[9] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human Decisions and Machine Predictions*, The Quarterly Journal of Economics 133 (2017) 237–293. doi:10.1093/qje/qjx032.

[10] S. Caton, C. Haas, Fairness in Machine Learning: A Survey, 2020. arXiv:2010.04053.

[11] A. Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, Big data 5 (2017) 153–163. doi:10.1089/big.2016.0047.
[12] M. Kearns, A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design, Oxford University Press, Inc., USA, 2019.
[13] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019.
[14] D. Pessach, E. Shmueli, A Review on Fairness in Machine Learning, ACM Comput. Surv. 55 (2022). doi:10.1145/3494672.
[15] D. Leben, Normative Principles for Evaluating Fairness in Machine Learning, Association for Computing Machinery, New York, NY, USA, 2020, pp. 86–92. URL: https://doi.org/10.1145/3375627.3375808.
[16] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods & Research 50 (2021) 3–44. doi:10.1177/0049124118782533.
[17] K. Makhlouf, S. Zhioua, C. Palamidessi, On the Applicability of Machine Learning Fairness Notions, SIGKDD Explor. Newsl. 23 (2021) 14–23. doi:10.1145/3468507.3468511.
[18] J. Baumann, C. Heitz, Group Fairness in Prediction-Based Decision Making: From Moral Assessment to Implementation, in: 2022 9th Swiss Conference on Data Science (SDS), 2022, pp. 19–25. doi:10.1109/SDS54800.2022.00011.
[19] M. Loi, A. Herlitz, H. Heidari, A Philosophical Theory of Fairness for Prediction-Based Decisions, SSRN Electronic Journal (2019). doi:10.2139/ssrn.3450300.
[20] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, 2016. arXiv:1609.05807v2.
[21] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, On the (im)possibility of fairness, 2016. URL: https://arxiv.org/abs/1609.07236. arXiv:1609.07236.
[22] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On Fairness and Calibration, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., 2017, pp. 5684–5693.
[23] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica, May 23 (2016) 139–159. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[24] W. Dieterich, C. Mendoza, T. Brennan, COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Technical Report, Northpoint Inc, 2016. URL: https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/.
[25] C. Simoiu, S. Corbett-Davies, S. Goel, The problem of infra-marginality in outcome tests for discrimination, The Annals of Applied Statistics 11 (2017) 1193–1216. doi:10.1214/17-AOAS1058.
[26] X. Wang, W. Li, Y. Cui, R. Zhang, J. Mao, Click-through rate estimation for rare events in online advertising, in: Online multimedia advertising: Techniques and technologies, IGI Global, 2011, pp. 1–12.
[27] J. Baumann, A. Castelnovo, R. Crupi, N. Inverardi, D. Regoli, Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3593013.3594058.