The need for transparent demographic group trade-offs in Credit Risk and Income Classification

Ananth Balashankar *[1,2] and Alyssa Lees[1]

[1]Google AI, New York
[2]New York University

Abstract. Prevalent methodology towards constructing fair machine learning (ML) systems, is to enforce a strict equality metric for demographic groups based on protected attributes like race and gender. While definitions of fairness in philosophy are varied, mitigating bias in ML classifiers often relies on demographic parity-based constraints across sub-populations. However, enforcing such constraints blindly can lead to undesirable trade-offs between group-level accuracy if groups possess different underlying sampled population metrics, an occurrence that is surprisingly common in real-world applications like credit risk and income classification. Similarly, attempts to relax hard constraints may lead to unintentional degradation in classification performance, without benefit to any demographic group. In these increasingly likely scenarios, we make the case for transparent human intervention in making the trade-offs between the accuracies of demographic groups. We propose that transparency in trade-offs between demographic groups should be a key tenet of ML design and implementation. Our evaluation demonstrates that a transparent human-in-the-loop trade-off technique based on the Pareto principle increases both overall and group-level accuracy by 9.5% and 9.6% respectively, in two commonly explored UCI datasets for credit risk and income classification.

1 Introduction

In recent discussions of ethical ML algorithms, evaluating fairness has been frequently predicated on defining constraints based on specific protected attributes, such as race or gender [1, 2]. These attributes should **not** demonstrate conditionally discriminative behavior while learning classification targets. If care is not taken in the construction of an ML model, works such as [3] and [4] have shown that inequalities in underlying data distributions can be amplified in the predicted output, leading to runaway feedback loops. Recent works [5] have argued that examining the intersectionality of multiple protected attributes is crucial for establishing coherent standards of fairness. However, real-world data subpopulations often display varying underlying sampling distributions, bias and

* Corresponding author: ananth@nyu.edu

noise. We argue that principles towards *fair* ML should encourage transparency in the trade-offs between demographic group accuracy in a classification task and at a minimum be able to reflect their true underlying population distributions. At a fixed sample size, as the number of protected attributes increases, the intersectional subgroup populations tend to decrease in size. In these scenarios, it is evident that any classifier which does not perform worse on all groups can never be *fair* [6]. To overcome this downside, we look to the rich literature of "individual fairness" which defines fairness with respect to a similarity metric between two individuals and enforces that similar individuals are treated similarly, within an error bound [7, 8, 5]. We find this definition to be useful in allowing us to continue to ensure that minority demographic group populations perform at their best accuracy while ensuring that majority demographic groups do not suffer a large decrease in group-level accuracy.

Using this transparent Pareto-principle of Efficiency [9], popular in social welfare and economics, we argue that trade-offs between demographic group accuracy undertaken by ML algorithms in high-stakes applications like credit risk and income classification [10] should be made transparent in order to be examined against socio-technical norms in that application domain [11–13]. We have been motivated by the insight that many fairness problems in existing classification tasks for specific subpopulations can be remedied by controlled data collection, subject to ethical considerations [14, 15]. As such, we suggest that in the spirit of achieving fair outcomes, when learning on datasets with varying demographic group sample sizes, how we weigh the loss suffered by each demographic group can be a critical choice and should be transparent.

In the domain of credit risk assessment, the trade-off between the accuracy of demographic groups has implications on financial justice across demographic groups. For example, older married male individuals have better accuracy than younger single female individuals for credit risk assessment. This means that even a seemingly group-blind ML algorithm can have significantly different accuracy across demographic groups. Similarly, in the income classification task, Caucasian male individuals have much better baseline accuracy than non-Caucasian female individuals in the United States. Therefore, to build transparent and fair ML systems, we show that the trade-offs between these demographic groups cannot be avoided, but rather should be an integral part of the transparent design of any socio-technical ML system. We illustrate one such transparent trade-off mechanism by arguing for efficiency based on the Pareto principle, where degradation in the accuracy of one group should not occur without improving another group's accuracy. In this paper, we compare our transparent Pareto-principle based trade-off with several other strict equality-based constraints and demonstrate an increase in 9.5% and 9.6% overall and group-level accuracy respectively on both the credit risk and income classification tasks.

2 Motivation: Trade-offs in the real world

COMPAS A ML model (COMPAS tool) was used for determining the risk of recidivism in Broward County, Florida, USA. ProPublica [16] found in an independent investigation involving 18,610 people over 2 years that black males were twice as likely to be misclassified by the model as high risk as compared to white males. This scenario highlights the critical need for auditing existing decision-making systems (including the ones based on human experts) and understanding the trade-offs made in their design. In such a high stakes scenario, ideally, a decision-making system that achieves the highest group level metrics (such as accuracy) is required. By incorporating inductive biases based on racial and social justice, one could hope to achieve the end objective of improving the Pareto front transparently. If we do not attempt to evaluate and discover Pareto efficient classifiers, a domain expert choosing a classifier might end up making trade-offs of accuracy and fairness among inefficient classifiers.

Gender Shades Certain image recognition models were discovered to have lower accuracy for one particular group (darker females) than other groups in the Gender Shades project [17]. The intervention undertaken to resolve this discrepancy involved collecting better data for the poor performing group (females with darker skin tone). The progress from such interventions amounts to discovering better group accuracies on the Pareto frontier, as opposed to restricting the models to strict equality among groups. Here too, the authors of the project, Buolamwini and Gebru, advocate for a complete ban of ML models for facial recognition tasks since these models are not advanced enough to perform with high accuracy on all groups independent of skin tone and gender, without encoding spurious correlations. Hence, a ML model needs to be transparent in the trade-offs that it implicitly makes to gain socio-technical acceptance in the real world.

3 Transparent Trade-offs

The Pareto frontier has been used to characterize the trade-offs between more than one dimension in multiple objective learning [18, 19]. It characterizes solutions such that no point on the Pareto curve dominates another point on all the dimensions across which we measure an objective. Evaluating the Pareto curve for any ML classifier can be critical in making transparent trade-offs between demographic groups [20].

3.1 Pareto front in ML based models

In our analysis of the German Credit and Adult Census Datasets, we take an example of a feedforward neural network model with up to 3 layers with each layer containing 256, 128 and 64 hidden units respectively. We then perform a sweep of the hyperparameters by varying the depth of the neural network,

learning rates and L1 and L2 regularization parameters [21], and the training data made available to train the network (specific demographic groups versus the entire dataset). Each network was trained multiple times with randomized seeds for initializing the parameters of the network. This gives us a wide range of group-level accuracies along each of the demographic groups we slice the accuracy of the model. We then constructed the Pareto front of these group-level accuracies after varying the hyper-parameters, with each group corresponding to a dimension of the Pareto front. Note that visualization of the Pareto front can be tricky, given that in most real-world applications, the demographic groups are more than three. Hence, we need a principled approach using which a domain practitioner can argue about their choice of a specific classifier on the Pareto front. In figure 1, we see that in simulated data with two demographic groups, a domain expert can trade-off one group's performance with another by choosing different points on the Pareto front. Also, we can see that a trade-off is inevitable unless we assume that the Pareto front exactly intersects with the hyperplane where all demographic groups perform equally (x=y in case of two dimensions).

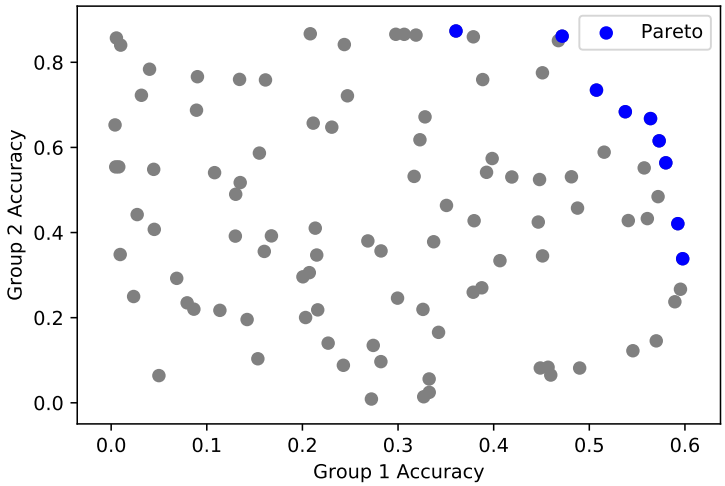

Fig. 1: An illustration of a two group-setting plotting group-level accuracy and its corresponding Pareto front (in blue) shows that demographic group trade-offs are implicit and unavoidable in ML systems

3.2 Pareto Trade-offs

Having established that a trade-off between group-level accuracy should be conducted on the Pareto front, we now provide an example where such a trade-off is transparent and based on a Pareto Efficient and fair principle. In this principle, a domain expert might choose a classifier where each group's performance sacrifices accuracy equally. For example, in the German Credit risk assessment task, older male individuals can achieve their best accuracy of 91% among all the points on the Pareto front, whereas younger female individuals can achieve only 73%. In this case, the Pareto-based trade-off would advocate for a classifier that achieves 89% and 71.4% on the two groups respectively, each of them about 2.2% below (Pareto Loss) their respective optimal choices on the Pareto front. This choice is different than the one a domain expert would choose based on the principle of strict equality or Demographic Parity [22] between the groups (both groups at 73%, i.e. zero Parity Loss). We acknowledge that both of these choices might be valid in different contexts based on the principles the corresponding algorithmic decision-making system prescribes. But, the choice needs to be transparent and cannot be masked behind the objective of minimizing overall classification error. This transparency allows people who apply for credit to contend the trade-offs and the corresponding principles in automated decision-making systems. Hence, with transparency, the people who were previously left out of the decision-making systems' design can be involved and provide them the ability to appeal the trade-offs made by such ML models.

4 Evaluation

4.1 Baselines

We compare our transparent trade-off approach with optimization techniques that use fairness constraints such as Equality Constraint [3], Adversarial [23], and Min-Max fairness [24]. [3] aims to lower the sum of absolute discrepancy of all group accuracy from the overall accuracy (Parity loss), while [23] adversarially attempts to nudge the classifier such that it cannot predict the protected attributes. [24] aims to maximize the accuracy of the least performing demographic group.

4.2 UCI Adult Dataset

The UCI Census Adult dataset focuses on the prediction of income as a binary variable ($> \$50K, <= \$50$K) based on demographic information. Protected attributes selected are gender and race and are denoted as binary categorical variables. We consider the 4 groups at the intersection of the protected attributes, to overcome the limitations of group fairness as outlined in [25]. The dataset has 48,842 instances out of which 20% is held out as test data, while the remaining is used for training and cross-validation. There are 14 attributes out of which 6 are continuous and 8 variables are categorical. Table 1 shows the Pareto

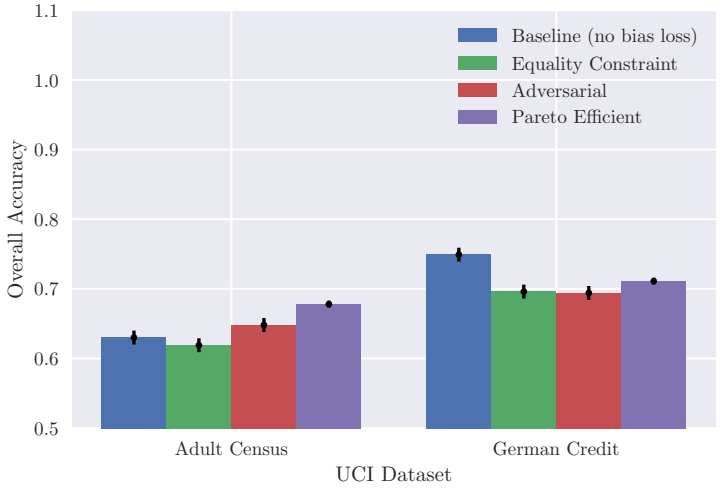

Fig. 2: Comparison for 2 UCI datasets showing that the pareto-based transparent trade-off achieves better overall accuracy than other fairness constrained classifiers.

loss, i.e how much each group deviates from the pseudo-optimal of the respective group for the UCI Census Adult dataset. Based on the Pareto principle, we were able to choose an optimal point on the Pareto front that ensured that each of the demographic groups perform optimally. In our transparent trade-off on the Pareto front, each of the groups has better individual accuracy than the other approaches and thus better overall accuracy as shown in Fig 2. Fig 3 demonstrates that our approach arrives at a better classifier on all demographic groups. Some groups even exceed the baseline accuracy (computed using the average of all unconstrained optimization results) due to an extensive swap of the hyperparameters and transparently choosing the Pareto optimal classifier.

4.3 UCI German Credit Dataset

The UCI German Credit risk assessment dataset involves predicting credit type as a binary label (good or bad) from demographic information where the protected attributes selected are age, gender and personal status. Each of these protected attributes is binarized and the intersection of these 3 attributes is considered as the groups in our study. There are 1000 instances in the dataset with a total of 20 categorical attributes. We hold out a random 20 % as test data over which we present the results. The evaluation of this dataset is determined by a cost matrix where the false positives are considered 5 times more costly than a false negative. The final accuracy reported takes this into account. Sim-

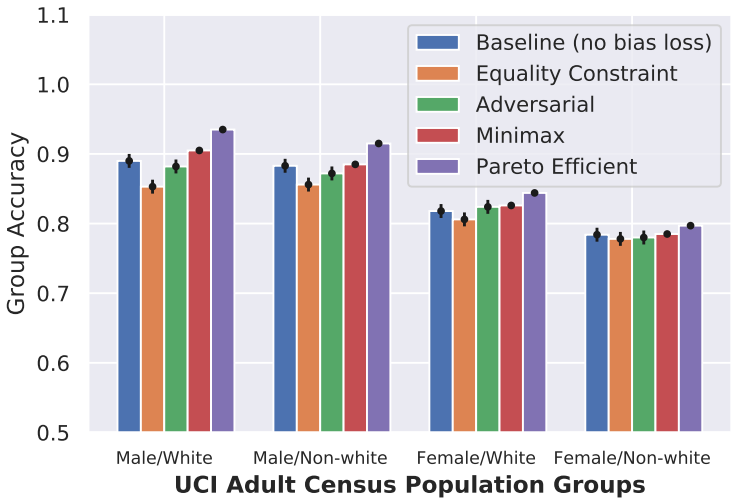

Fig. 3: Group accuracy comparison shows that we achieve Pareto dominating group level accuracy for all groups in UCI Adult dataset.

Model	**FPR**	**FNR**	**Parity Loss**	**Pareto Loss**
Baseline (no bias loss)	0.253	0.747	0.199	0.016
Equality Constraint[3]	0.283	**0.712**	**0.167**	0.133
Adversarial [23]	0.224	0.769	0.226	0.077
Min-max [24]	0.202	0.773	0.218	0.075
Pareto Efficient	**0.165**	0.830	0.250	**0.000**

Table 1: Comparison of test losses in UCI Adult dataset. Our Pareto-based trade-off has no difference as compared to the Pareto optimal group-accuracy, while [3] minimizes Parity loss.

ilar to the UCI Adult Dataset, in Figure 4, we see that choosing a point based on our Pareto principle, we increase the group-level accuracies as compared to the equality constraints [22], adversarial loss [23] and minimax [24] optimization techniques. The 5 groups (out of the total 8) are shown in the UCI German Credit Dataset, as the rest of the groups do not have enough samples (< 100).

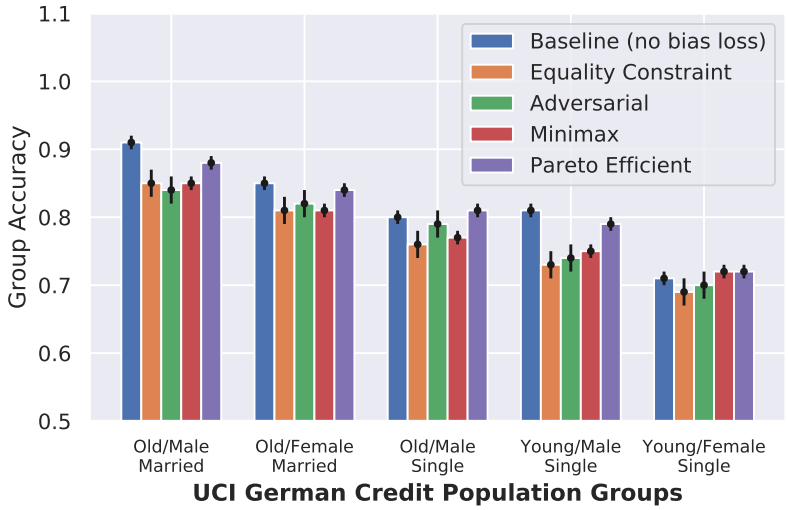

Fig. 4: Group accuracy comparison showing we achieve optimal group level accuracy for all groups in UCI German Credit dataset among constrained classifiers.

4.4 Sample Size Inconsistencies

The use of explicit demographic attributes in real-world scenarios is sometimes a hard constraint. One example is legislation enforcing fairness around disparate impact [26, 27]. In simplified examples, exploring the intersectionality of protected attributes may be appropriate. For example, in this paper, we explore two gender and two race subgroups in the evaluation of the UCI Adult dataset, which translates to four separate groups. It is conceivable that in a real-world application, the intersection of gender and race subgroups could extend into **many** different groups. As the intersectionality of groups grows, a group's sample size will likely be insufficient. In the case of the UCI German credit risk assessment dataset, the attribute - marriage status, with five possible values, is treated as a protected attribute along with gender and age. However, in the dataset, there were **no** samples containing both the attributes of young, female and being married.

Table 2: Comparison of sample complexity ranking for Probably Approximately Metric Fairness with actual subgroup sizes of subgroups

#	Group	Complexity Rank	Sample Size (Rank)
UCI Adult Dataset			
1	Male/White	1	2,129 (1)
2	Male/Non-white	3	8,642 (3)
3	Female/White	4	2,616 (2)
4	Female/Non-white	2	19,174 (4)
UCI German Credit Dataset			
1	Old/Male	3	50 (1)
2	Old/Female	4	310 (3)
3	Young/Male	2	548 (4)
4	Young/Female	1	92 (2)

Despite the impossibility results of achieving fairness in the extreme case of subgroup sized one, there is still a need to highlight cases where simple (linear) models are inadequately applied in datasets with complex underlying subgroup distributions [16, 28]. The ability to transparently argue about the trade-offs made in designing the required model along with the limitations of small sample sizes for certain demographic groups will guide the choices made by practitioners and ML researchers. Through our work, we see that even an ML model that does not explicitly perform a trade-off between demographic groups has already decided the trade-off implicitly.

Using the theory of sample complexity based on Rademacher complexity [29, 30], if we assume all the hypotheses are linear, we can rank the hardness of learning the target for each demographic group, and order them (Table 2 - higher numbered rank has higher complexity values). In the UCI Adult Census dataset, the ordering of the actual subgroup sample sizes ($4 > 2 > 3 > 1$) reveals that new samples are needed to match the desired sample complexity ordering ($3 > 2 > 4 > 1$). Specifically, more samples for subgroup 3 (Female/White) need to be gathered than for subgroup 2 (Male/Non-white) to ensure the ordering of actual sample sizes aligns with that of the sample complexities. Similarly, in the German Credit Dataset, Table 2 shows disparity in the order of the actual sample sizes ($3 > 2 > 4 > 1$) as compared to desired sample complexity ($2 > 1 > 3 > 4$). This implies that in the UCI German Credit dataset, more new samples from group 2 (Old/Female) than from group 3 (Young/Male) should be drawn for us to make a balanced and transparent choice while performing trade-offs. Similarly, more samples from subgroup 1 (Old/Male) need to be collected than from subgroup 4 (Young/Female) to remove any inversion in the ranking of complexities and actual group sample sizes to ensure that the trade-offs are not performed inefficiently due to insufficient sample sizes.

5 Conclusion

We advocate for transparency in the demographic group accuracy trade-offs in high-stakes real-world applications like credit risk and income classification tasks. We demonstrate that transparency in how we balance group-level accuracies can lead to better classifiers being explored on the Pareto front while improving overall accuracy too by 9.5%. Further, we caveat that trade-offs on demographic groups with smaller sample sizes should be taken into account and appropriate data collection exercises should be conducted. We argue that for the development of an ethical AI framework for policy and decision-makers, transparency in the group-level accuracy trade-offs is critical. Future work to extend this analysis to more complex ML models may provide principled standards for transparent trade-offs between groups in other application domains along with mechanisms to contest them.

References

1. Corey D. Johnson Rafael Salamanca Jr. Vincent J. Gentile Robert E. Cornegy Jr. Jumaane D. Williams Ben Kallos Carlos Menchaca James Vacca, Helen K. Rosenthal. A local law in relation to automated decision systems used by agencies, 2018.
2. P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.
3. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.
4. Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. *CoRR*, abs/1706.09847, 2017.
5. Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*, 2018.
6. Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118, New York, NY, USA, 23–24 Feb 2018. PMLR.
7. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *ITCS*, 2012.
8. Cynthia Dwork and Christina Ilvento. Fairness under composition. In *ITCS*, 2019.
9. Parke Godfrey, Ryan Shipley, and Jarek Gryz. Algorithms and analyses for maximal vector computation. *The VLDB Journal*, 16(1):5–28, January 2007.
10. Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.
11. A D. Selbst, D Boyd, S A. Friedler, S Venkatasubramanian, and J Vertesi. Fairness and abstraction in sociotechnical systems. FAT* '19.
12. N Grgic-Hlaca. The case for process fairness in learning : Feature selection for fair decision making. 2016.
13. D Madras, E Creager, T Pitassi, and R Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. FAT* '19.

14. Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, 2018.
15. Irene Chen, Fredrik D. Johansson, and David A Sontag. Why is my classifier discriminatory? *CoRR*, abs/1805.12002, 2018.
16. Surya Mau Julia Angwin, Je Larson and Lauren Kirchner. How we analyzed the compas recidivism algorithm. 2016.
17. Dean Foster and Rakesh Vohra. An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188, 1992.
18. J Ali, M B Zafar, A Singla, and K P. Gummadi. Loss-aversively fair classification. AIES '19, page 211–218, 2019.
19. John Rawls. *A theory of justice*. 1971.
20. Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. *CoRR*, abs/1803.02453, 2018.
21. Andrew Y. Ng. Feature selection, ¡i¿l¡/i¿¡sub¿1¡/sub¿ vs. ¡i¿l¡/i¿¡sub¿2¡/sub¿ regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 78, New York, NY, USA, 2004. Association for Computing Machinery.
22. Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.
23. Alex Beutel, Jilin Chen, Zhe Zhao, and Ed Huai hsin Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.
24. N Martinez, M Bertran, and G Sapiro. Minimax pareto fairness: A multi objective perspective. ICML 2020.
25. Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *CoRR*, abs/1711.05144, 2017.
26. Supreme Court of the United States. Griggs v. duke power co. 1971.
27. 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*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.
28. Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? 06 2017.
29. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012.
30. John Shawe-Taylor, Martin Anthony, and Norman Biggs. Bounding sample size with the vapnik-chervonenkis dimension. 42:65–73, 02 1993.