# Leveling up in Intersectional Fairness

The aim of fair machine learning is to develop models that are devoid of any kind of discriminatory behavior towards any subset of the population. In recent years, this task has attracted a broad range of approaches (1; 2; 3), with most of them focusing on a single sensitive axis (4; 5), such as gender or race. However, recent studies (6) have demonstrated that even when fairness can be ensured at the individual sensitive axis levels, significant unfairness can still exist at the intersection level. For example, (7) showed that commercial face recognition tools have been shown to exhibit significantly higher error rates for groups of darker-skinned females than for lighter-skinned males.

**Contribution 1:** To capture this intersectional fairness, several measures have been proposed (8; 9), with the most commonly used being the Differential Fairness (DF) measures (10). Intuitively, DF is the log-ratio of the best-performing group to the worst-performing group for a given performance measure, such as the False Positive Rate (FPR). In this work, we argue that the DF metric does not capture the complete picture. For instance, the DF metric can be trivially minimized by harming all the groups, i.e., by increasing FPR uniformly for all groups. To illustrate this, assume, on the one hand, a model M1 where the group with the lowest FPR is $0.2$ and the one with the highest FPR is $0.4$. On the other hand, a model M2 with the lowest and the highest FPR corresponds to $0.3$ and $0.45$, respectively. According to the DF measure, M2 is significantly fairer than M1 as it is strictly more egalitarian, even though the way M2 has achieved is by harming both the best and worst group. Thus there is a tension between the relative performance between groups and the absolute performance of the groups. In order to capture this tension, we propose $F_\beta$ Differential Fairness measure. Informally, it is a weighted harmonic mean between the relative and absolute performance of the group. It enables practitioners to explore the whole trade-off between them by changing the weight ($\beta$).

**Contribution 2:** Another major challenge in achieving intersectional fairness is the limited data availability for certain subgroups. For instance, in the Twitter Hate Speech dataset (11), the smallest group consists of only $300$ examples, while the largest group contains over $6,000$ examples. This mismatch in the amount of data also reflects the real world, where it can be inherently challenging to collect data for certain groups compared to others.

In order to circumvent the above problem, we propose a simple Maximum Mean Discrepancy (MMD) (12) based data generation mechanism, which relies on the fact that smaller subgroups could be augmented by modifying and combining the data from other subgroups. More specifically, we exploit the structure inherent in the problem of intersectionality by viewing each subgroup as an intersection of other groups. For instance, data for subgroup Female-AfricanAmerican can be derived by combing examples from group Female and group AfricanAmerican. At the time of training, (i) we combine the generated and the real data into one batch, and (ii) instead of randomly sampling examples from the training data, we propose to use equal sampling, where the classifier sees the same number of examples for each group and label. The first step increases the diversity of examples the classifier is trained on, improving generalization, while the latter ensures that equal importance is given to all subgroups instead of focusing more on larger groups.

**Experiments and Conclusion:** Our experiments on various fairness enforcing methods across multiple datasets show that several of these methods improve DF by harming groups. We also empirically evaluate our proposed data generation mechanism over various datasets and find it leads to a more fair representation while not harming the worst of the group. Moreover, apart from superior results, our approach is easy to integrate with existing training pipeline, and can be combined with various in-processing fairness approaches.

## References


[1] Zafar, M. B., I. Valera, M. G. Rogriguez, et al. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[2] Denis, C., R. Elie, M. Hebiri, et al. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*, 2021.

[3] Maheshwari, G., M. Perrot. Fairgrad: Fairness aware gradient descent. *CoRR*, abs/2206.10923, 2022.

[4] Lohaus, M., M. Perrot, U. Von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR, 2020.

[5] Agarwal, A., A. Beygelzimer, M. Dudík, et al. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[6] Yang, F., M. Cisse, O. Koyejo. Fairness with overlapping groups; a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[7] Buolamwini, J., T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler, C. Wilson, eds., *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, vol. 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018.

[8] Kearns, M. J., S. Neel, A. Roth, et al. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. G. Dy, A. Krause, eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, vol. 80 of *Proceedings of Machine Learning Research*, pages 2569–2577. PMLR, 2018.

[9] Hébert-Johnson, Ú., M. P. Kim, O. Reingold, et al. Multicalibration: Calibration for the (computationally-identifiable) masses. In J. G. Dy, A. Krause, eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, vol. 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

[10] Foulds, J. R., R. Islam, K. N. Keya, et al. An intersectional definition of fairness. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pages 1918–1921. IEEE, 2020.

[11] Huang, X., L. Xing, F. Dernoncourt, et al. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448. European Language Resources Association, Marseille, France, 2020.

[12] Gretton, A., K. M. Borgwardt, M. J. Rasch, et al. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.