Continual Audit of Individual Fairness in Deployed Classifiers via Prediction Sensitivity

Krystal Maughan [1] Ivoline Ngong [1] Joe Near [1]

[1]College of Engineering and Mathematical Sciences

Group Fairness Classifiers Do Not Tell The Whole Story.

Many existing approaches use *metrics* that offer formal processes for "measuring" fairness. *Group fairness* metrics [12, 14] measure disparate treatment of groups in aggregate. These metrics are useful to demonstrate unfairness, but previous work has shown that group-fair classifiers can still make clearly unfair predictions for individuals.

Prediction Sensitivity

- Let x represent an input and $\mathcal{F}(x)$ represent an output prediction. Our gradient, which represents the change in prediction over x, is represented $\nabla\mathcal{F}$. We estimate how changes in x would affect the prediction $\mathcal{F}(x)$ using the gradient $\nabla\mathcal{F}$.

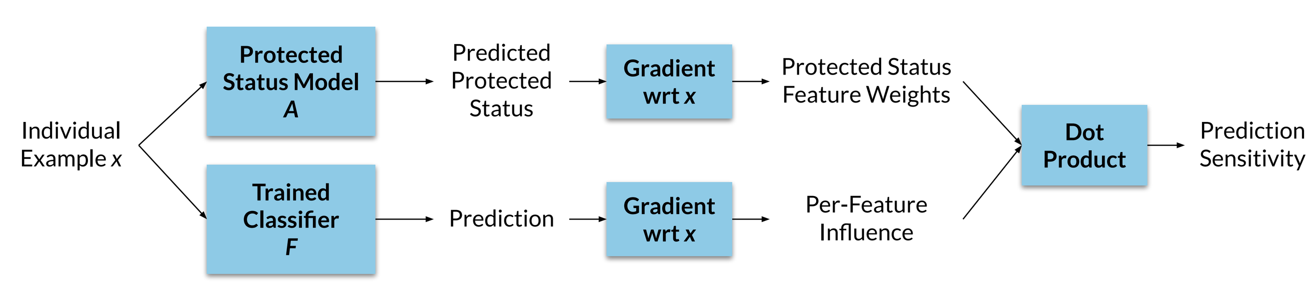

Figure 1. Overview of calculating prediction sensitivity. Prediction sensitivity is based on measurements of each feature's contribution to both protected stastus and the classifier's prediction.

Calculating Prediction Sensitivity

We consider a classifier $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}$ and an individual input . We would like to know if for all other individuals y, $|\mathcal{F}(x) - \mathcal{F}(y)| \leq d(x, y)$ under the similarity metric d, as required by individual fairness.

A mapping $M : V \to \Delta(A)$ satisfies the (D, d)-Lipschitz property if for every pair of individuals $x, y \in V$:

$$D(M(x), M(y)) \leq d(x, y)$$

Let $T_d \in \mathbb{R}^m$ be a *similarity transformation* for the distance metric d if $\|T_d\|_1 = 1$ (weights sum to 1) and for all $x, y \in \mathbb{R}^m$:

$$\|(1 - T)(T_d \circ x - T_d \circ y)\| = d(x, y)$$

where $\circ$ is the Hadamard (elementwise) product.

The ***prediction sensitivity*** $PS(x) \in \mathbb{R}$ for an example x is defined as:

$$PS(x) = (T_d(x)) \cdot abs(\nabla\mathcal{F}(x))$$

where $\nabla\mathcal{F}(x))$ is the gradient of $\mathcal{F}(x)$ (with respect to x) and *abs* denotes element-wise absolute value.

Modeling Individual Fairness with Synthetic Data

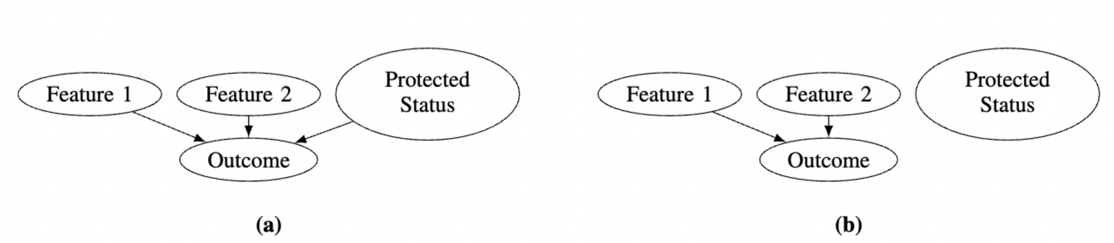

Figure 2. Causal graphs for synthetic data. **(a)** shows a causal graph for "biased" synthetic data, in which a causal relationship exists between protected status and outcome. **(b)** shows a modified causal graph that removes this relationship. Data generated according to model (b) can be used to train classifiers that satisfy individual fairness.

Our Contribution: Prediction Sensitivity

1. We propose *prediction sensitivity*, a gradient-based method for measuring individual fairness.
2. We prove that prediction sensitivity is an *upper bound* on individual fairness.
3. We show how to use prediction sensitivity to detect biased predictions at the individual level in *deployed models*.
4. We present experimental results suggesting that prediction sensitivity is *effective* for detecting biased predictions.

Experiments and Evaluations

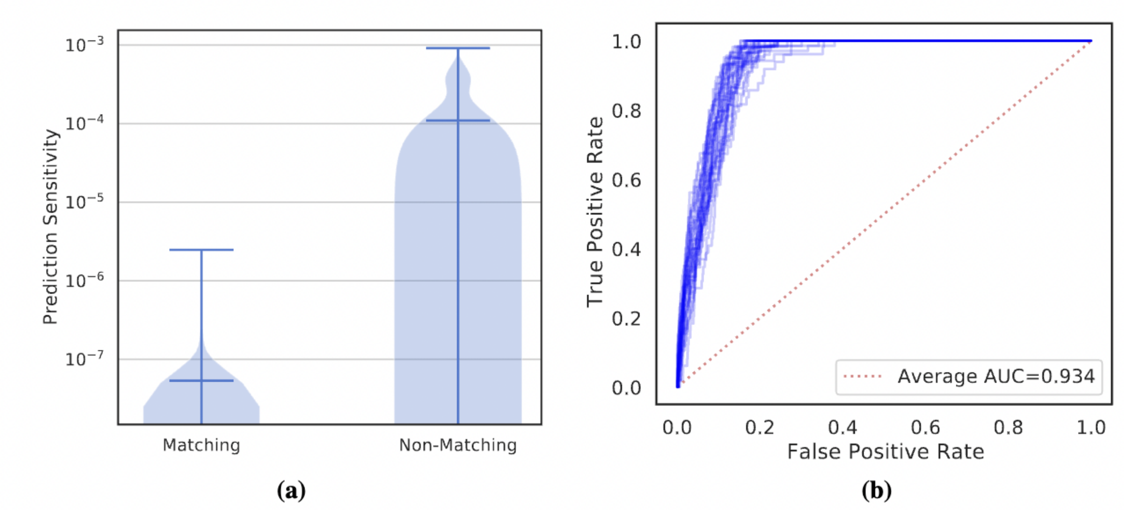

Figure 3. Using prediction sensitivity to audit models trained on synthetic data. **(a)** shows that prediction sensitivity is low for members of the match set, but high for non-members (note the logarithmic scale in the vertical axis). **(b)** shows that a distinguisher based on prediction sensitivity is effective at detecting failures of individual fairness.

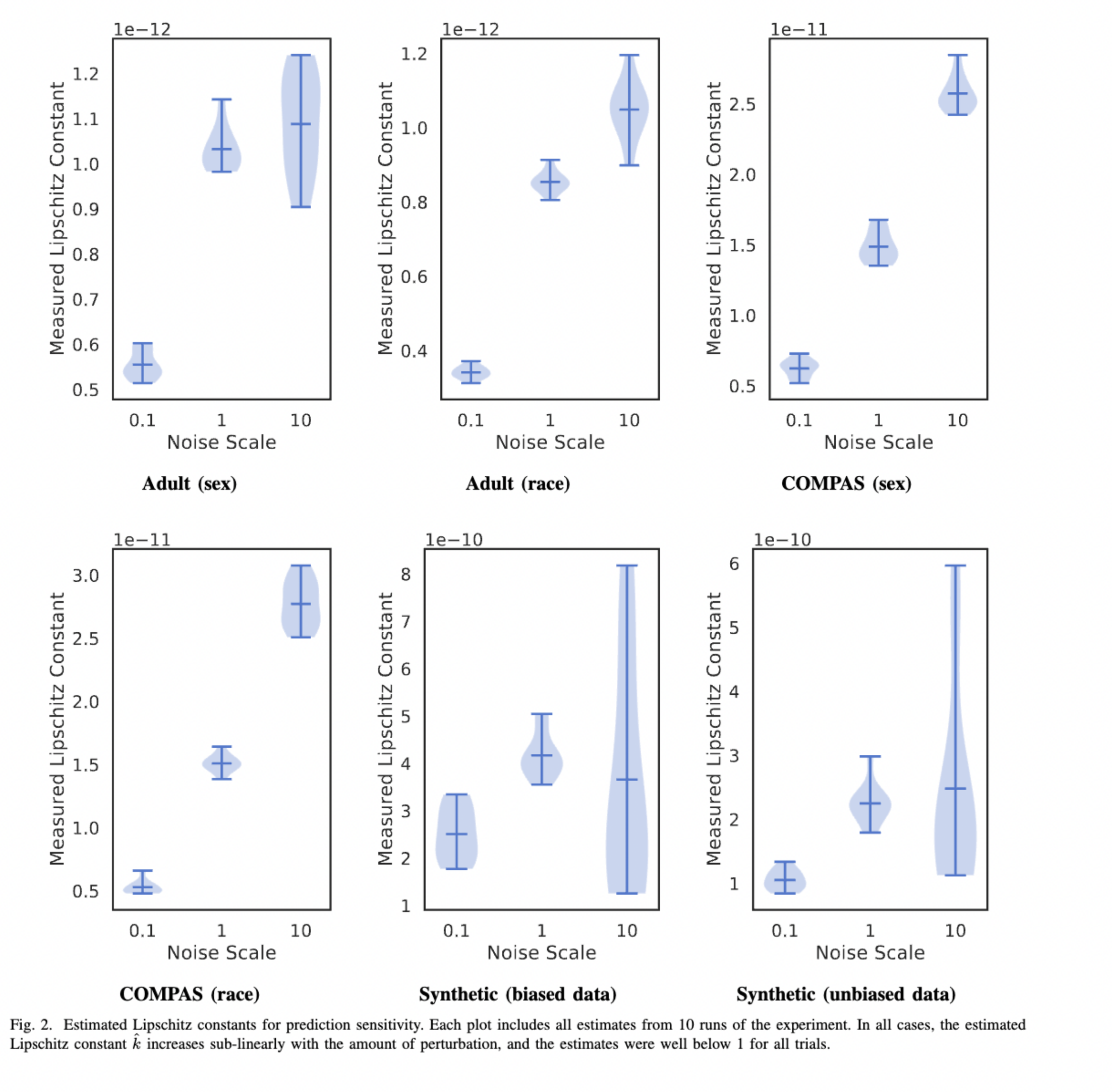

Figure 4. Estimated Lipschitz constants for prediction sensitivity. Each plot includes all estimates from 10 runs of the experiment. In all cases, the estimated Lipschitz constant $\hat{k}$ increases sub-linearly with the amount of perturbation, and the estimates were well below 1 for all trials.

Conclusion

- **Our results** suggest that prediction sensitivity is effective at detecting unfair predictions, but they also reflect the inherent challenge of this task.
- **Individual predictions** with extremely high prediction sensitivity are likely to be blatantly unfair, and should be easily detected using prediction sensitivity; however, borderline cases may be more difficult to detect.

References

[1] Adult dataset (UCI machine learning repository). `https://archive.ics.uci.edu/ml/datasets/adult`.

[2] COMPAS dataset (Propublica). `https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis`.

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[5] Atılım Günes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, 18(1):5595–5637, 2017.

[6] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[7] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

[8] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

[9] 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*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[10] 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.

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[12] 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, 2015.

[13] Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems*, pages 11722–11732, 2019.

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

[15] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9737–9748, 2019.

[16] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.

[18] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In *Advances in neural information processing systems*, pages 981–990, 2017.

[19] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Fairmatch: A graph-based approach for improving aggregate diversity in recommender systems. *arXiv preprint arXiv:2005.01148*, 2020.

[20] Krystal Maughan and Joseph P. Near. Towards a measure of individual fairness for deep learning. *CoRR*, abs/2009.13650, 2020. Presented at the 4th Workshop on Mechanism Design for Social Good (MD4SG '20).

[21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

[22] Niels JS Morch, Ulrik Kjems, Lars Kai Hansen, Claus Svarer, Ian Law, Benny Lautrup, Steve Strother, and Kelly Rehm. Visualization of neural networks using saliency maps. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 2085–2090. IEEE, 1995.

Krystal.Maughan@uvm.edu