# Parametrised Data Sampling for Fairness Optimisation

Vladimiro G. Zelaya, Paolo Missier and Dennis Prangle

KDD XAI Workshop, 5 August 2019
Anchorage, Alaska

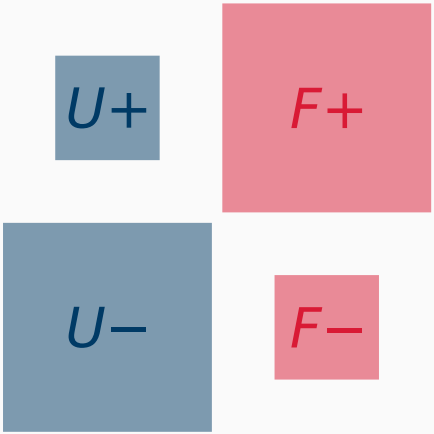
EPSRC
Engineering and Physical Sciences Research Council

Newcastle University
Digital Institute

# What is This Talk About?

- Method for correcting *classifier fairness*
- *Model* and *definition* agnostic
- *Tune* correction level to optimise fairness

# Population Subgroups



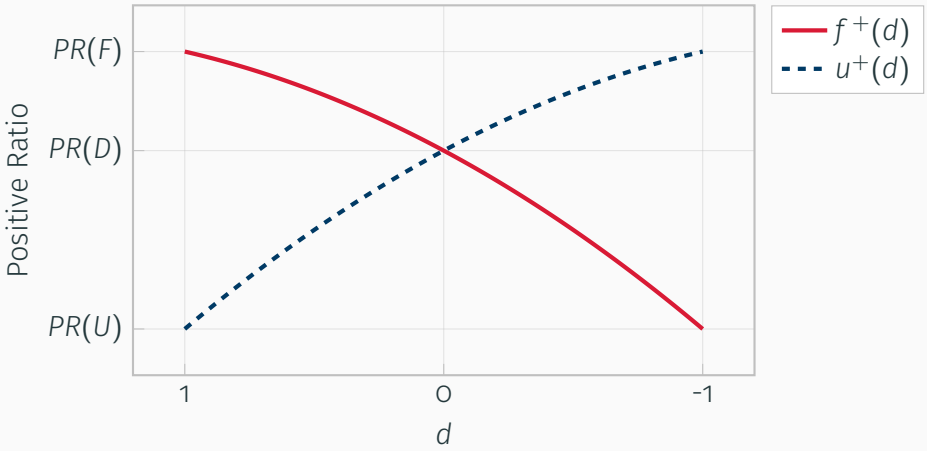

By Protected Attribute (PA):

- Favoured ($F$)
- Unfavoured ($U$)

By Class Label:

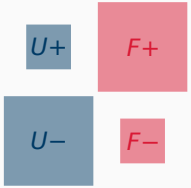
- Positive ($+$)
- Negative ($-$)

# Train Set Correction

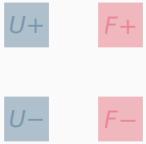


# Sampling Strategies

Original Data

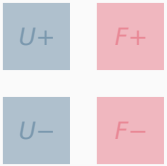



Under




SMOTE [1]

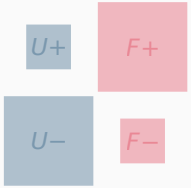

Preferential [2]



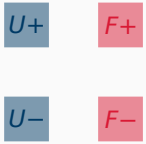

[1] [Chawla et al., 2002]
[2] [Kamiran and Calders, 2010]

# Sampling Strategies

Original Data

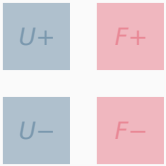



Under




SMOTE [1]

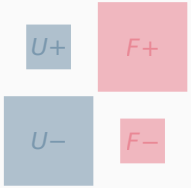

Preferential [2]



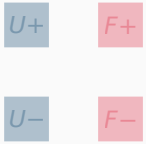

[1] [Chawla et al., 2002]
[2] [Kamiran and Calders, 2010]

# Sampling Strategies

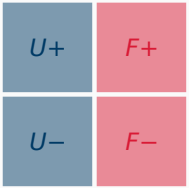
Original Data



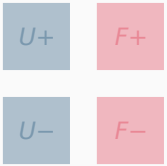

Under




SMOTE [1]




Preferential [2]

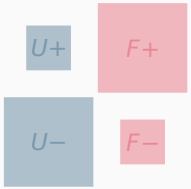



[1] [Chawla et al., 2002]

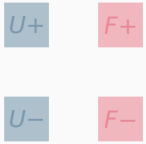
[2] [Kamiran and Calders, 2010]

# Sampling Strategies

Original Data

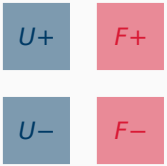



Under




SMOTE [1]


Preferential [2]




[1] [Chawla et al., 2002]

[2] [Kamiran and Calders, 2010]

# Ratio Form of Fairness Definitions

Equality Form

$$P\left(\hat{Y} = 1 \mid PA = U\right) = P\left(\hat{Y} = 1 \mid PA = F\right)$$

Ratio Form

$$\frac{P\left(\hat{Y} = 1 \mid PA = U\right)}{P\left(\hat{Y} = 1 \mid PA = F\right)} = 1$$

# Ratio Form of Fairness Definitions

Equality Form $$P\left(\hat{Y}=1 \mid PA=U\right) = P\left(\hat{Y}=1 \mid PA=F\right)$$

Ratio Form $$\frac{P\left(\hat{Y}=1 \mid PA=U\right)}{P\left(\hat{Y}=1 \mid PA=F\right)} = 1$$

# Some Fairness Ratios

Demographic Parity

$$DPR = \frac{P\left(\hat{Y} = 1 \mid PA = U\right)}{P\left(\hat{Y} = 1 \mid PA = F\right)}$$

Equality of Opportunity

$$EOR = \frac{P\left(\hat{Y} = 1 \mid PA = U, Y = 1\right)}{P\left(\hat{Y} = 1 \mid PA = F, Y = 1\right)}$$

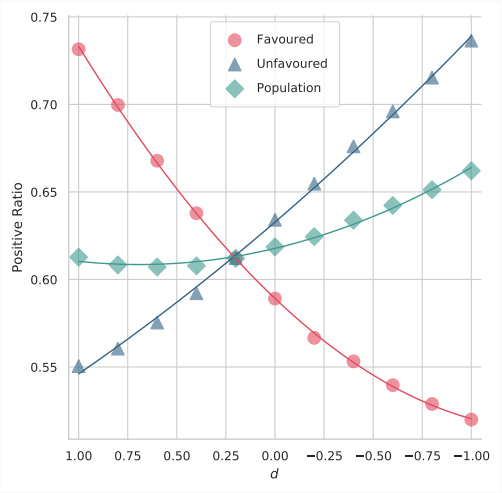
Counterfactual (Proxy)

$$CFR = \frac{PR\left(Test_{PA \leftarrow U}\right)}{PR\left(Test_{PA \leftarrow F}\right)}$$

# Experiments

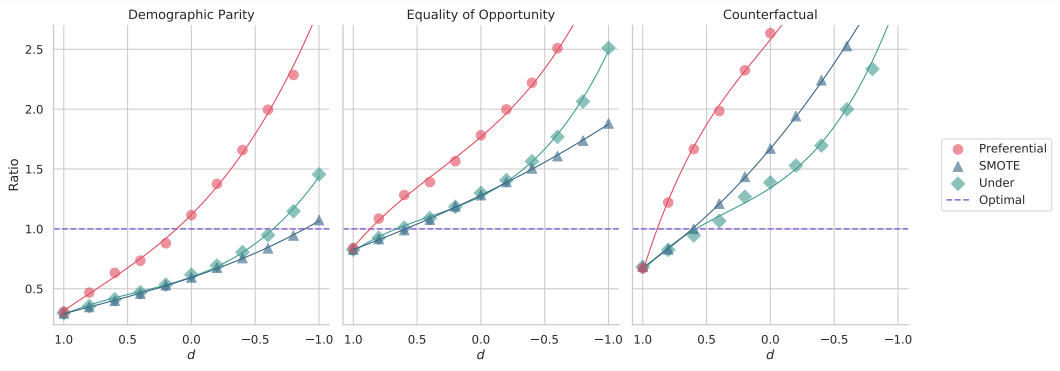
| Dataset | Protected | Favoured | Positive Class | Instances |
|---|---|---|---|---|
| COMPAS | Race | White | Won't reoffend | 6907 |
| Credit | Gender | Male | Will repay loan | 1000 |
| Income | Gender | Male | Income > $50k | 48842 |

# Effects on Test Set (COMPAS, Undersampling)



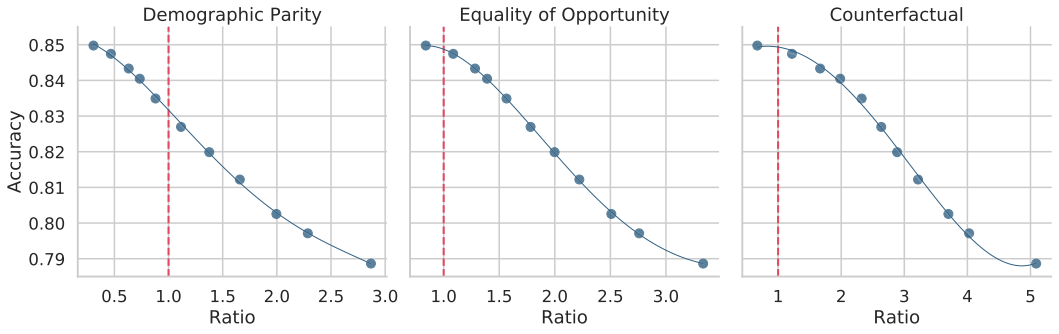

- Effect is correlated with correction
- But it occurs to a different extent
- Intersection is *not* at $d = 0$

# Optimal Correction by Fairness and Sampling



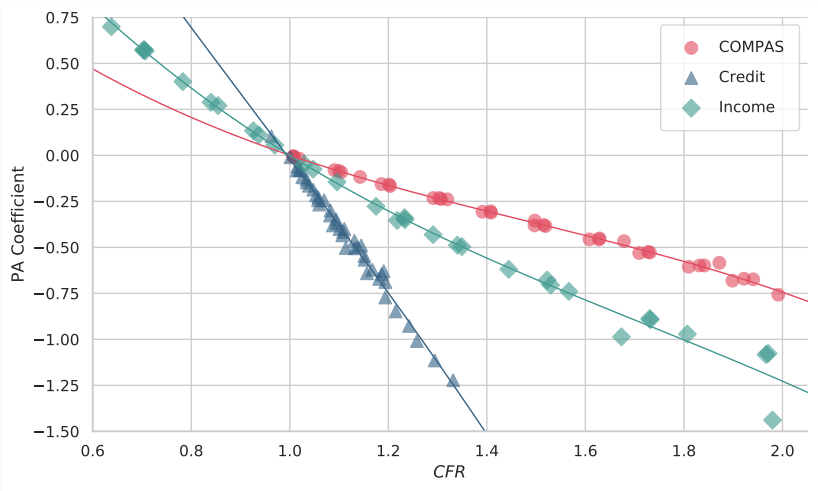

Plots for *Income* dataset

# Accuracy vs Fairness Trade-off




Plots for *Income* dataset corrected by Preferential Sampling

# Conjecture: *A LR model is FTU* $\Leftrightarrow$ *CFR* $= 1$



## Conclusions

- Fairness-agnostic optimisation with a relatively small loss in accuracy
- Ideal correction level is definition dependant
- Different sampling strategies produced similar results

## Future Work

- Generalise to non-binary cases (easy for PA)
- Optimise for more than one fairness definition
- Optimise for fairness and accuracy

# Thank You!

## Any Questions?

c.v.gonzalez-zelaya2@ncl.ac.uk

These slides: `https://git.io/fjHK5`

# For Further Reading


- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002).
  **SMOTE: Synthetic Minority Over-sampling Technique.**
  *Journal of Artificial Intelligence Research*, 16:321–357.
- Kamiran, F. and Calders, T. (2010).
  **Classification with no discrimination by preferential sampling.**
  In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6. Citeseer.