Week 9 Notes

By: Ethan Chriswell, Kyle McManus, Ryan Hildebrandt

TUESDAY OCTOBER 22, 2019

FAIR PREDICTION WITH DISPARATE IMPACT

Introduction

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time

COMPAS

Developed by Northpointe Inc.

Risk Assessment used to determine recidivism likelihood

With such heavy impact we need to know if systems are bias

ProPublica Analysis to see if COMPAS is bias

Notation

- Let S=s(x) denote the risk Score.
- We let R $\epsilon \{b, w]$ denote the group to which an individual belongs
- We denote the outcome indicator by $Y \in \{0,1)$, with Y=1 indicating that the given individual goes on to recidivate.
- We introduce the quantity sHR, which denotes the high-risk score threshold.
- Defendants whose score S exceeds sHR will be referred to as high risk, while the remaining defendants will be referred to as low risk.

Four Different Fairness Criteria

CALIBRATION

PREDICTIVE PARITY

ERROR RATE BALANCE

STATISTICAL (DEMOGRAPHIC) PARITY

Collaboration

A score S=S(x) is said to be well-calibrated if it reflects the same likelihood of recidivism irrespective of the individuals’ group membership

In their response to the ProPublica investigation, Flores et al. [6] verify that COMPAS is well-calibrated using logistic regression modeling

Predictive Parity

A score s = S(x) satisfies predictive parity at a threshold sHR if the likelihood of recidivism among high-risk offenders is the same regardless of group membership.

$$\mathbb{P}(Y=1 \,||\, S>sHR, R=b)$$
$$= \mathbb{P}(Y=1 \,||\, S>sHR, R=w).$$

Predictive parity at a given threshold sHR amounts to requiring that the positive predictive value (PPV) of the classifier Y^ = 1S>sHR be the same across groups

Northpointe’s refutation of the ProPublica analysis shows that COMPAS satisfies predictive parity for threshold choices of interest.

Error Rate Balance

A score S = s(x) satisfies error rate balance at a threshold sHR if the false positive and false negative error rates are equal across groups

$\mathbb{P}\ (S>sHR\ ,||Y=0\ ,\ R=b)=\mathbb{P}(S>sHR\ ,Y=0,\ R=w)$
$\mathbb{P}\ (S<sHR\ ,||Y=1\ ,\ R=b)=\mathbb{P}(S<sHR\ ,Y=1,\ R=w)$

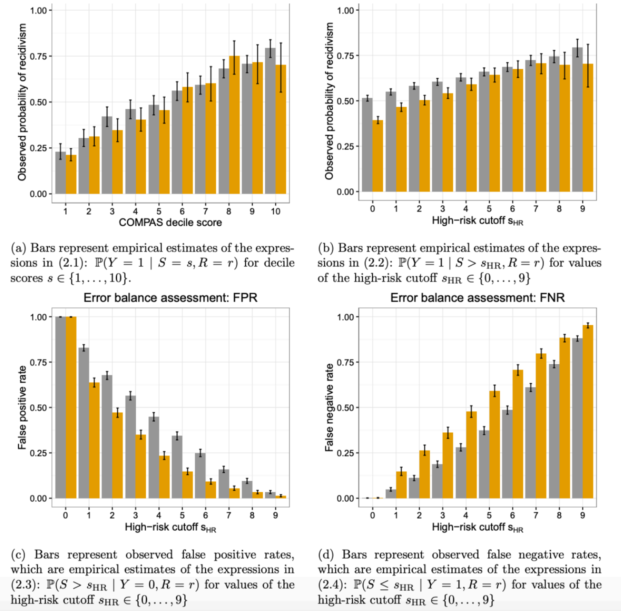
ProPublica analysis considered a threshold or sHR=4, which they showed leads to considerable imbalance in both false positive and false negative rates.

Error rate balance is also closely connected to the notions of equalized odds and equal opportunity

Statistical (Demographic) Parity

- ➢A score S(x) satisfies statistical parity at a threshold sHR if the proportion of individuals classified as high-risk is the same for each group.
- ➢$\mathbb{P}(S > s_{HR\,,} \,|R = b) = \mathbb{P}(S > s_{HR\,,} \,|R = w)$
- ➢Other names: Demographic parity, equal acceptance rates, group fairness.

Does Compas Satisfy



(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \ldots, 10\}$.

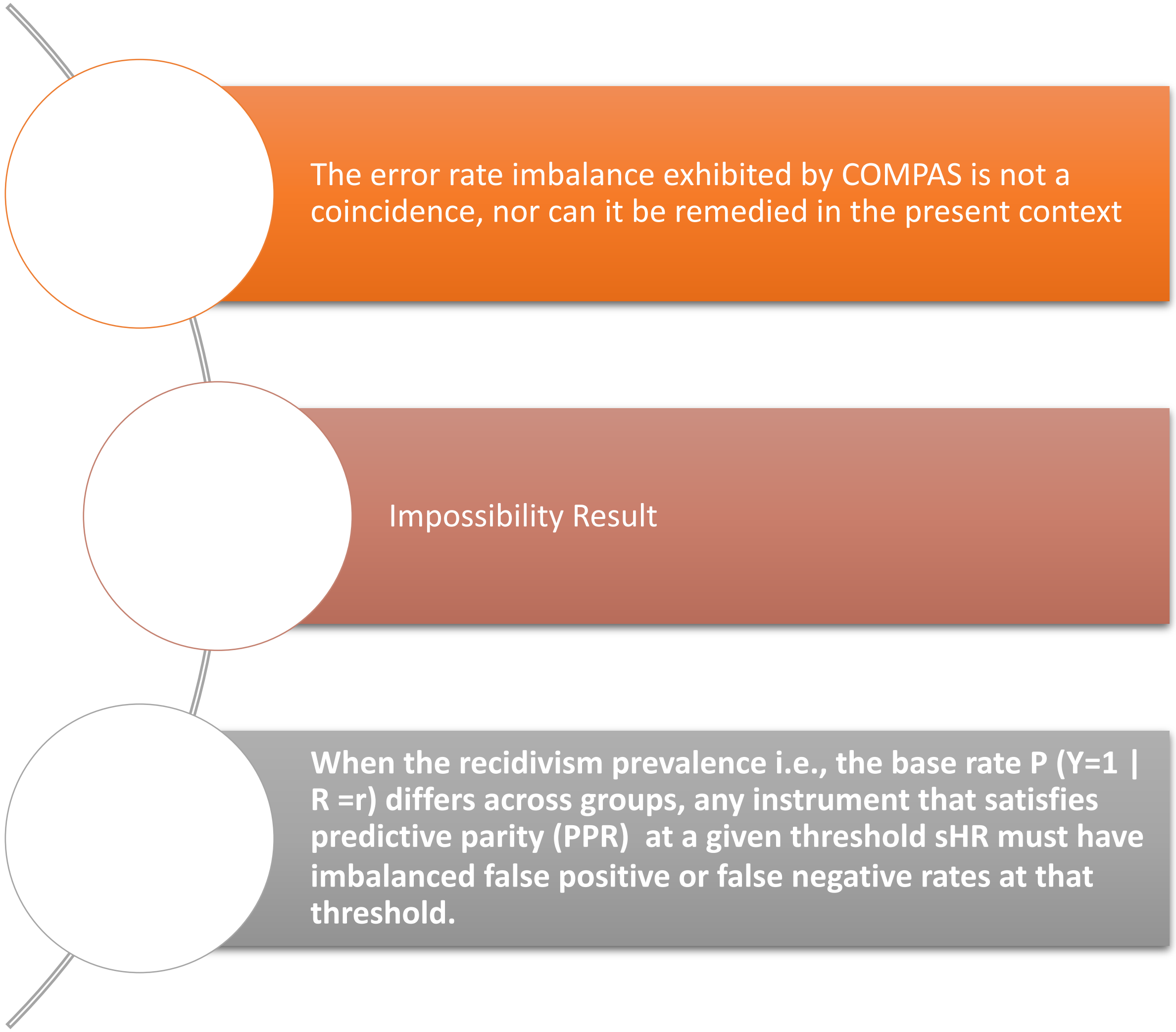
(b) Bars represent empirical estimates of the expressions in (2.2): $\mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$

(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3): $\mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$

(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4): $\mathbb{P}(S \le s_{\mathrm{HR}} \mid Y = 1, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$

- Calibration Assessment
  - Yes
  - 95% Confidence Intervals Intercept for Black and White groups
- Predictive Parity
  - Yes
  - 95% Confidence Intervals Intercept for Black and White groups
- Error Balance
  - No
  - Higher False Positive Rates for Black Groups is Higher
- Statistical Parity
  - No
  - Higher False Negative Rates for White Groups is Higher

Conclusion



The error rate imbalance exhibited by COMPAS is not a coincidence, nor can it be remedied in the present context

Impossibility Result

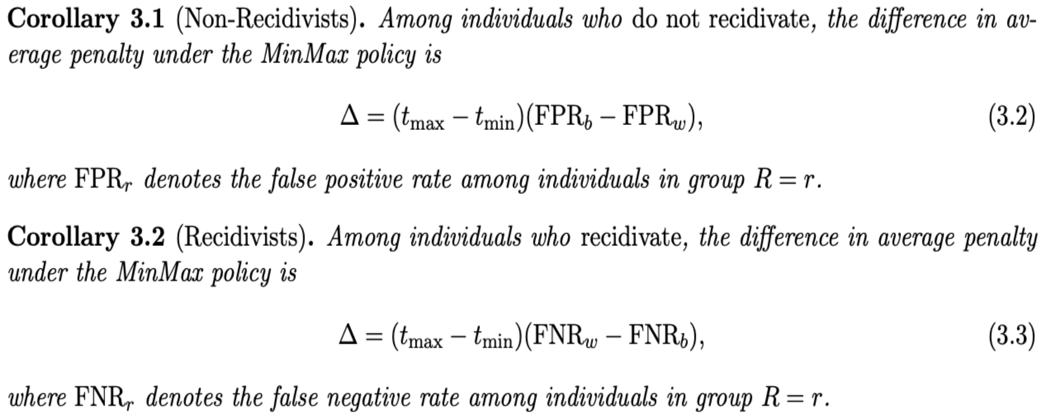
When the recidivism prevalence i.e., the base rate P (Y=1 | R =r) differs across groups, any instrument that satisfies predictive parity (PPR) at a given threshold sHR must have imbalanced false positive or false negative rates at that threshold.

$$\mathrm{FPR} = \frac{p}{1-p}\,\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}).$$

THURSDAY OCTOBER 24, 2019

EQUALITY OF OPPORTUNITY IN SUPERVISED LEARNING

Does this lead to disparate impact?



Corollary 3.1 (Non-Recidivists). *Among individuals who* do not recidivate, *the difference in average penalty under the MinMax policy is*

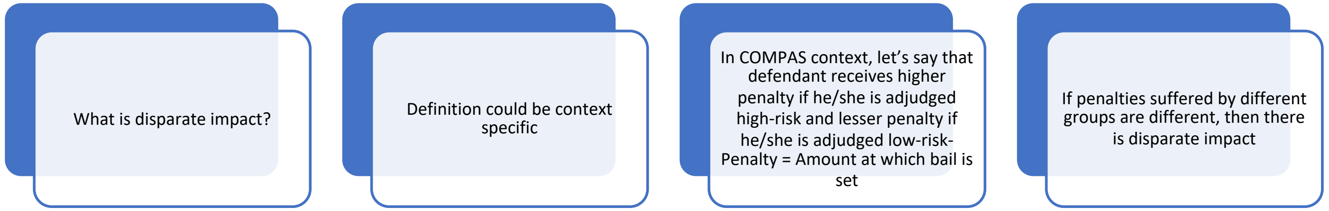
$$\Delta = (t_{\max} - t_{\min})(\mathrm{FPR}_b - \mathrm{FPR}_w), \qquad (3.2)$$

where FPR_r *denotes the false positive rate among individuals in group* $R = r$.

Corollary 3.2 (Recidivists). *Among individuals who* recidivate, *the difference in average penalty under the MinMax policy is*

$$\Delta = (t_{\max} - t_{\min})(\mathrm{FNR}_w - \mathrm{FNR}_b), \qquad (3.3)$$

where FNR_r *denotes the false negative rate among individuals in group* $R = r$.



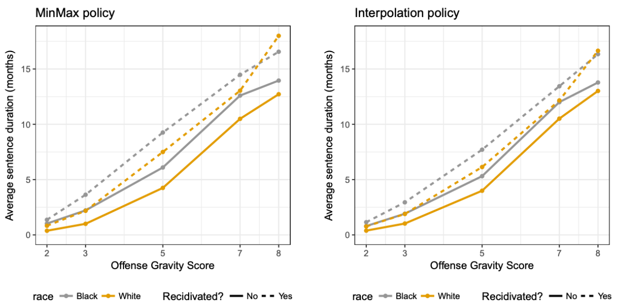
What is disparate impact?

Definition could be context specific

In COMPAS context, let's say that defendant receives higher penalty if he/she is adjudged high-risk and lesser penalty if he/she is adjudged low-risk-Penalty = Amount at which bail is set

If penalties suffered by different groups are different, then there is disparate impact

Important to see the Distinction



- Unfairness in ML relates more to predictions made by ML model
- Disparate impact deals more with what you do with the predictions
- Might be the case that unfair ML does not lead to disparate impact if it is not actually used to make decisions
  - Just using your ML model to understand how your features are correlated with you target variable

Three Ways of Fixing Biases

PREPROCESS DATA

POSTPROCESSING DATA

MODIFY ML ALGORITHM

What is their high-level approach?

Don’t change anything in the ML training pipeline

Train models like you usually do

Do some post-processing on the outputs of the ML model

Make it fairer

Make this process oblivious to the training set

2 Fairness criteria that they try to establish

Equalized odds
 - Across your two categories: FPR should be equal, TPR should be equal

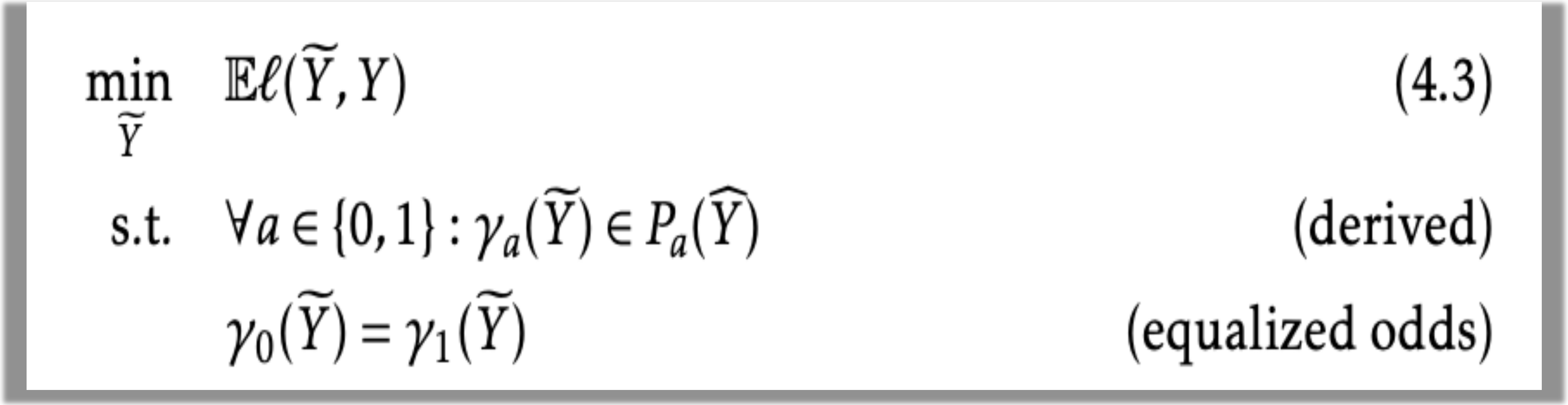
Equal opportunity

We say that a binary predictor Yb satisfies equal opportunity with respect to A and Y if only satisfied for the positive class

Equal opportunity is a weaker, though still interesting, notion of non-discrimination, and thus typically

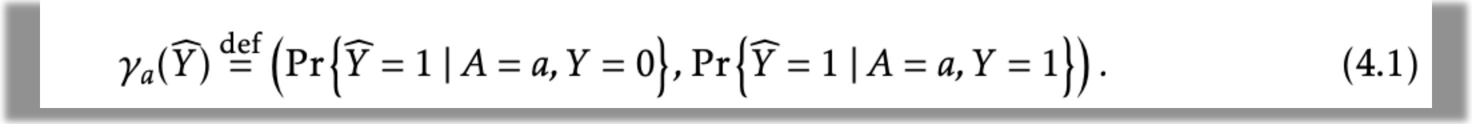
$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\} = \Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in \{0,1\}$$

Deriving from Binary Predictor

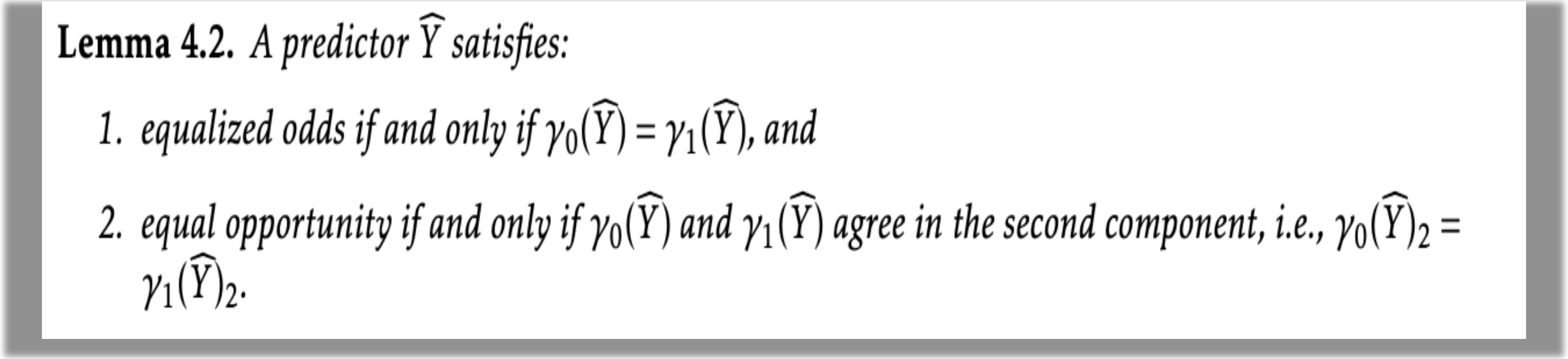


$$
\begin{aligned}
\min_{\widetilde{Y}} \quad & \mathbb{E}\ell(\widetilde{Y}, Y) && \text{(4.3)} \\
\text{s.t.} \quad & \forall a \in \{0,1\} : \gamma_a(\widetilde{Y}) \in P_a(\widehat{Y}) && \text{(derived)} \\
& \gamma_0(\widetilde{Y}) = \gamma_1(\widetilde{Y}) && \text{(equalized odds)}
\end{aligned}
$$

Deriving from Binary Predictor pt.2

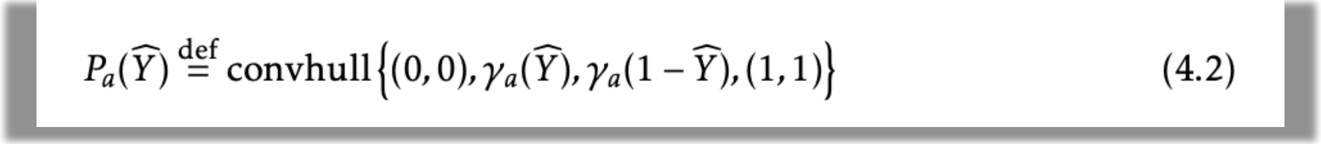


$$\gamma_a(\widehat{Y}) \stackrel{\text{def}}{=} \left(\Pr\left\{\widehat{Y}=1 \mid A=a, Y=0\right\}, \Pr\left\{\widehat{Y}=1 \mid A=a, Y=1\right\}\right). \tag{4.1}$$



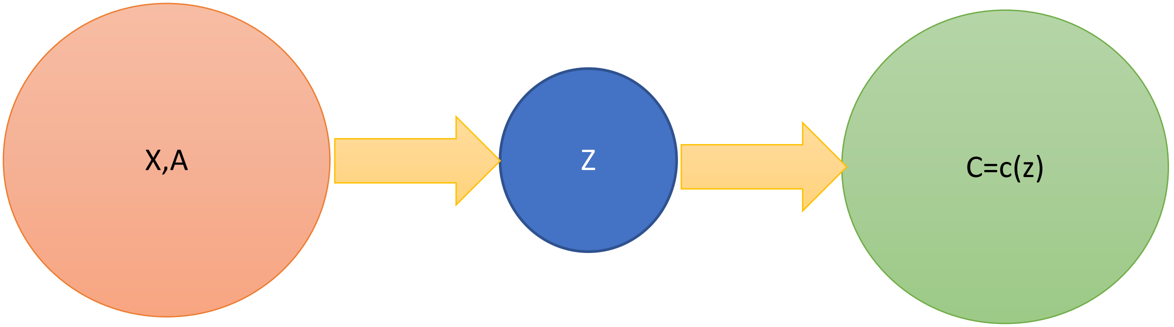
Lemma 4.2. *A predictor $\widehat{Y}$ satisfies:*

1. *equalized odds if and only if $\gamma_0(\widehat{Y}) = \gamma_1(\widehat{Y})$, and*
2. *equal opportunity if and only if $\gamma_0(\widehat{Y})$ and $\gamma_1(\widehat{Y})$ agree in the second component, i.e., $\gamma_0(\widehat{Y})_2 = \gamma_1(\widehat{Y})_2$.*



$$P_a(\widehat{Y}) \stackrel{\text{def}}{=} \operatorname{convhull}\left\{(0,0), \gamma_a(\widehat{Y}), \gamma_a(1-\widehat{Y}), (1,1)\right\} \tag{4.2}$$

Another Approach: Pre-Processing



- Remove the sensitive attribute
- Remove all features with sensitive attribute
- Brute Force Method?

  For all the features that are somewhat correlated with the sensitive attribute
- What is a more sophisticated method?

THAT'S IT!

QUESTIONS?