Equality of Opportunity in Supervised Learning

Paper by Moritz Hardt, Eric Price and Nathan Sebro

Presented by Fabian Bosshard

Importance of Fairness

The Scenario

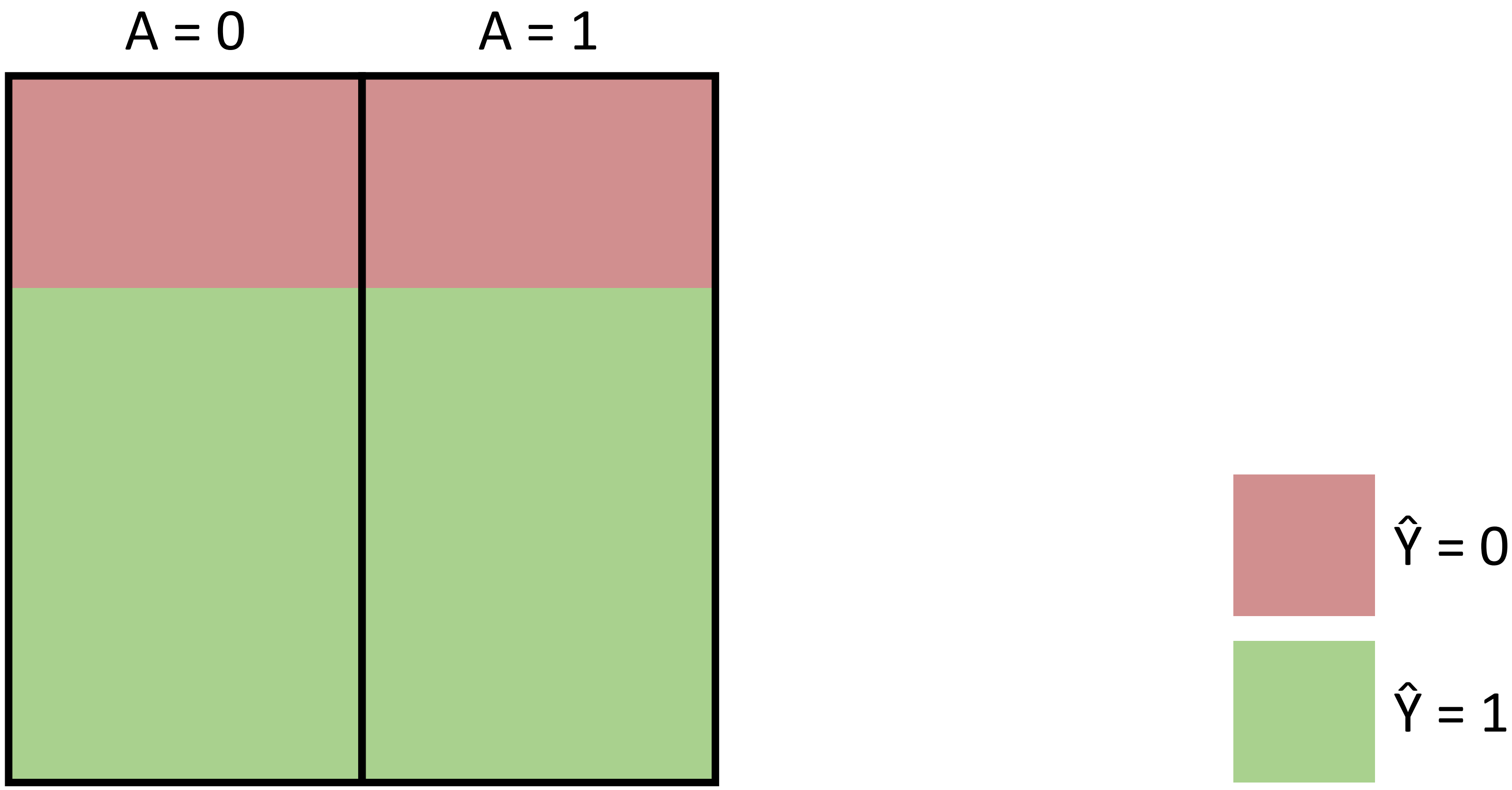
- Y: The variable to predict, binary, truthful
  “Pays back the loan”, “Will not relapse”, …
- A: The protected attribute, binary
  Ethnicity, Gender, Sexual orientation, …
- X: Other attributes
  Profession, Wealth, ZIP code, …
- Ŷ: The prediction of Y
  Produced by any classifier (SVM, Neural Network, Hand crafted, …)

What is Fairness?

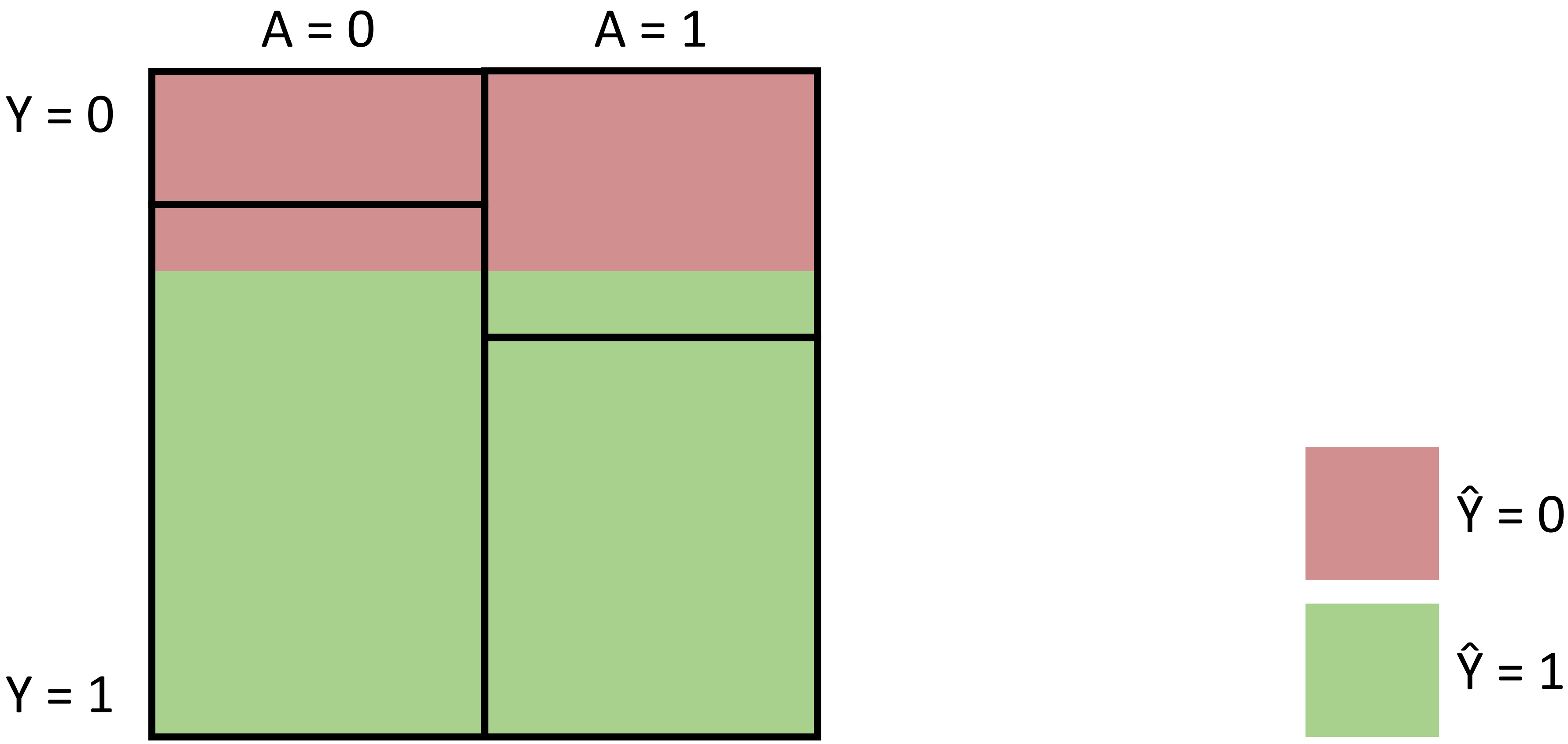
Oblivious Measure

Only depends on the joint distribution of $(Y, A, \hat{Y})$

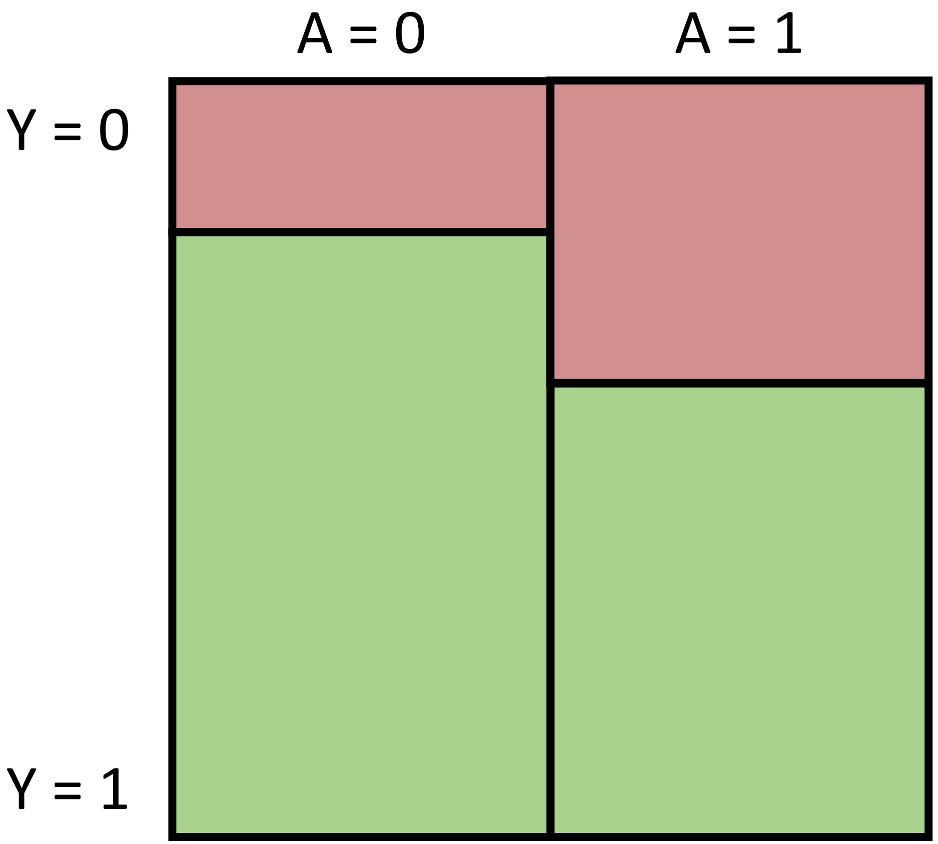
Demographic Parity



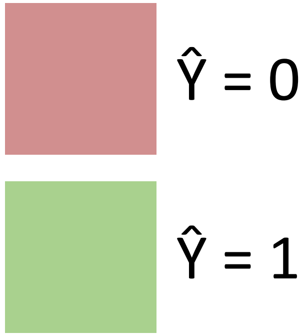
Demographic Parity



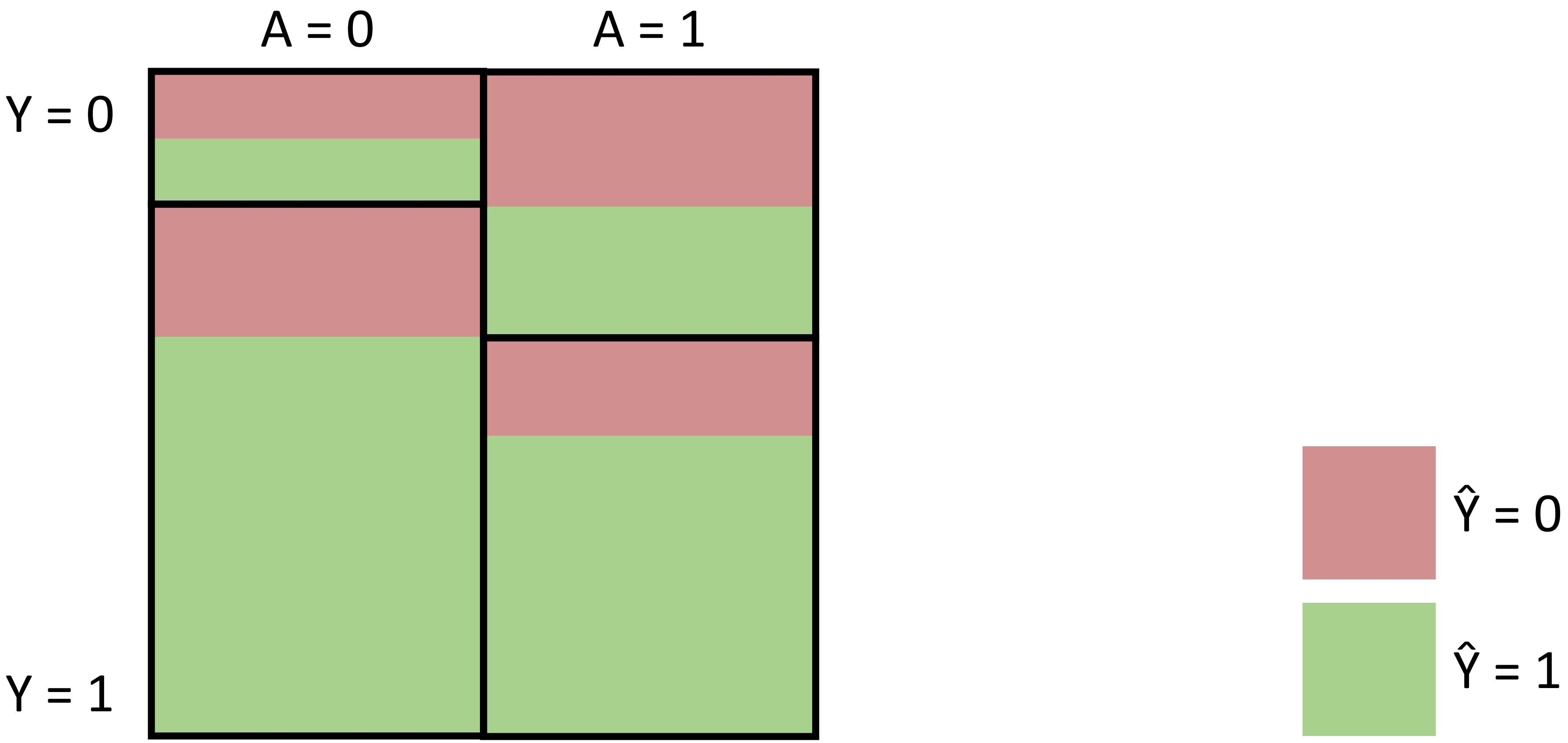
Demographic Parity - Issue



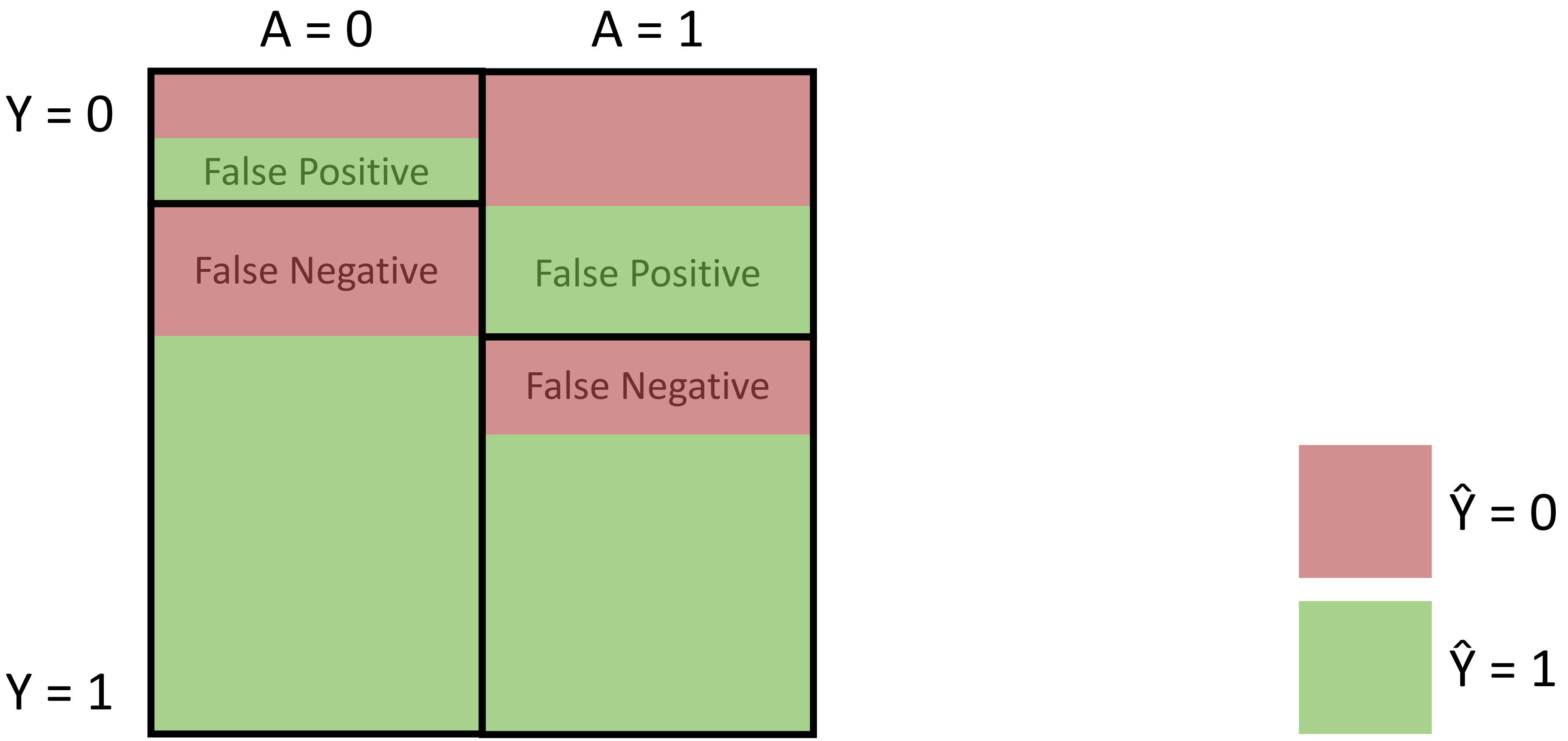
Perfect Classifier…
… but Demographic Parity violated!



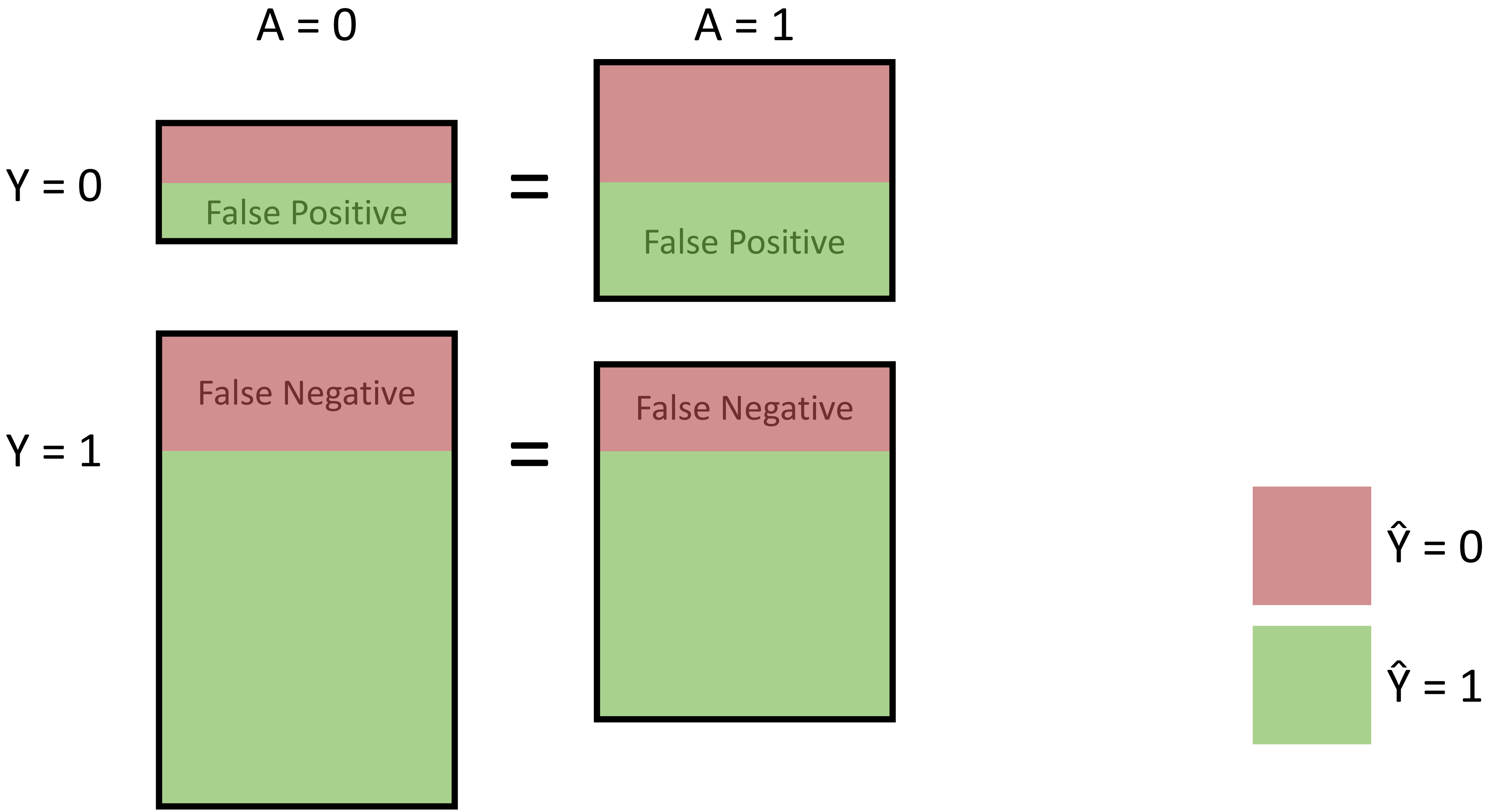
Equalized Odds



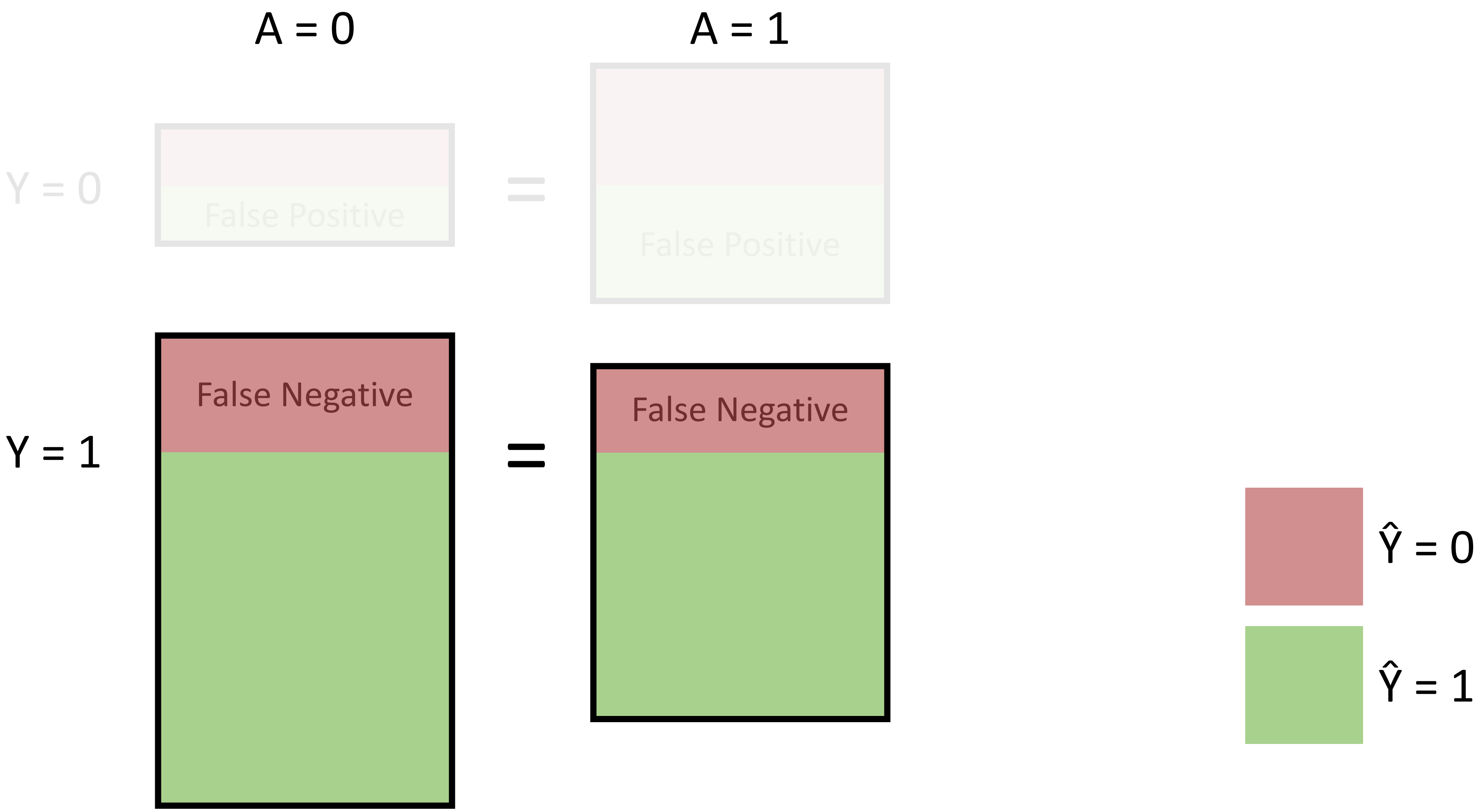
Equalized Odds



Equalized Odds

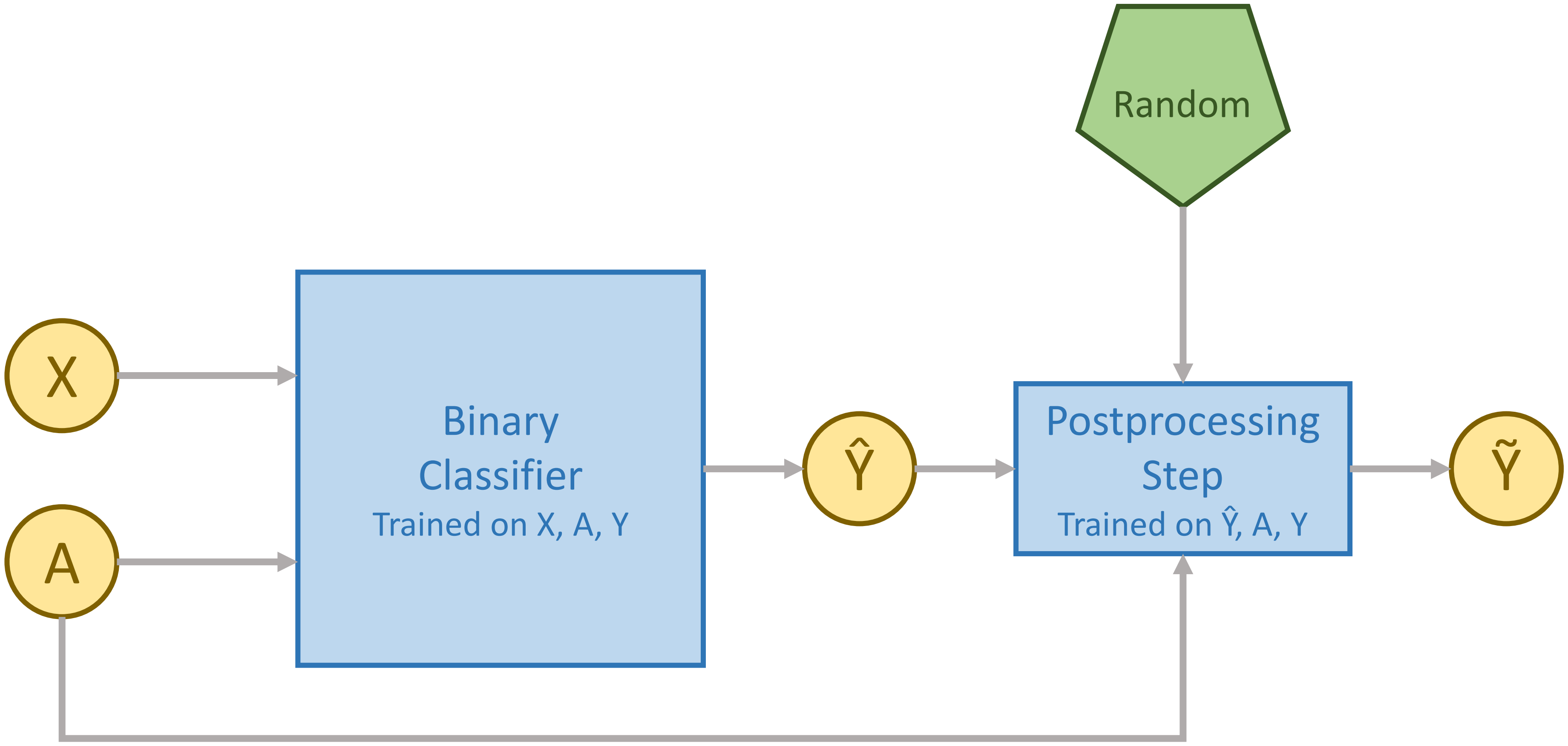


Equalized Opportunity

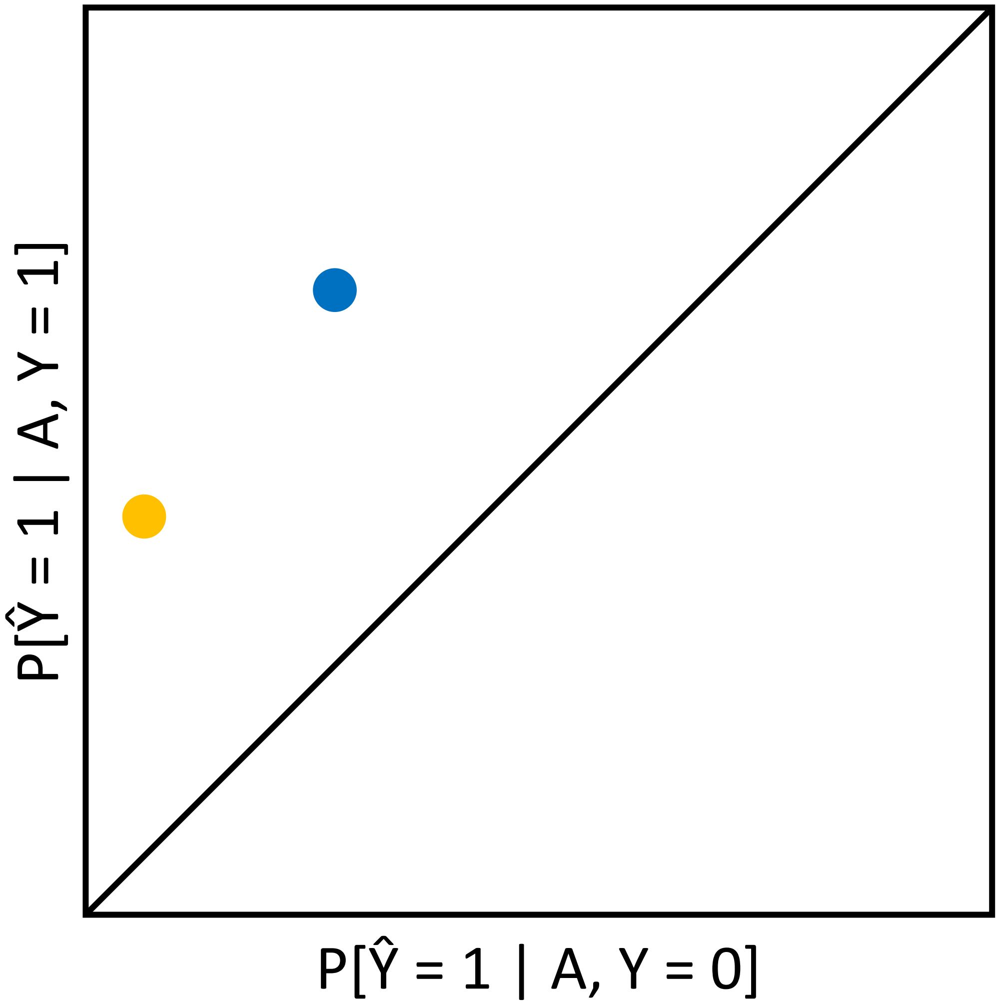


Achieving Equalized Odds

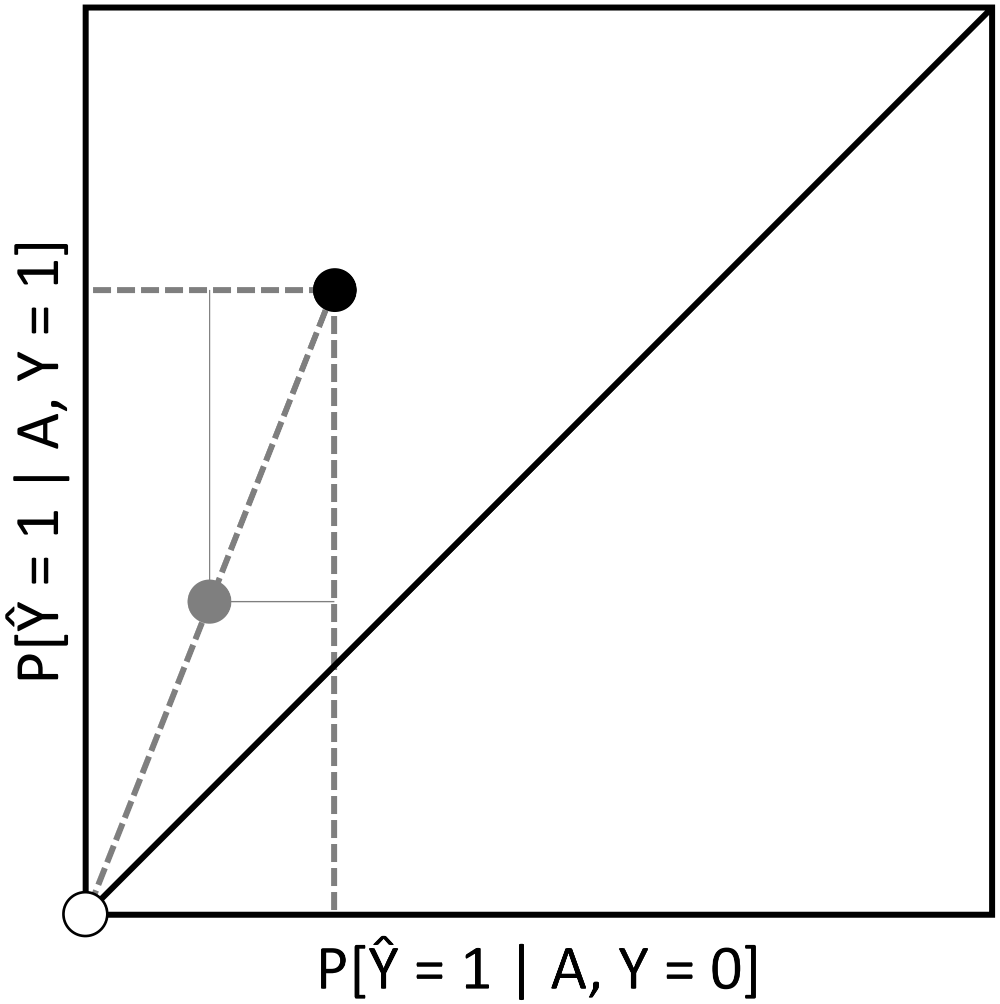
General Procedure



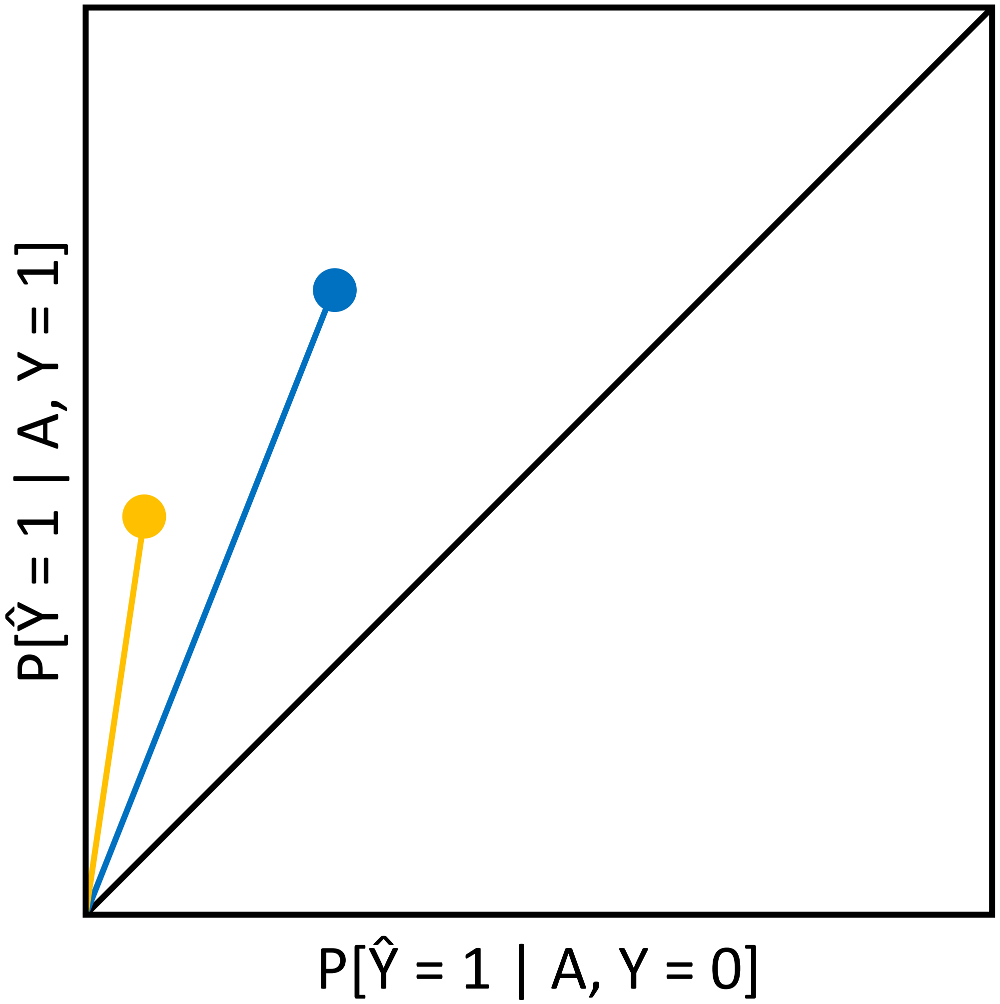
Training of the Postprocessing



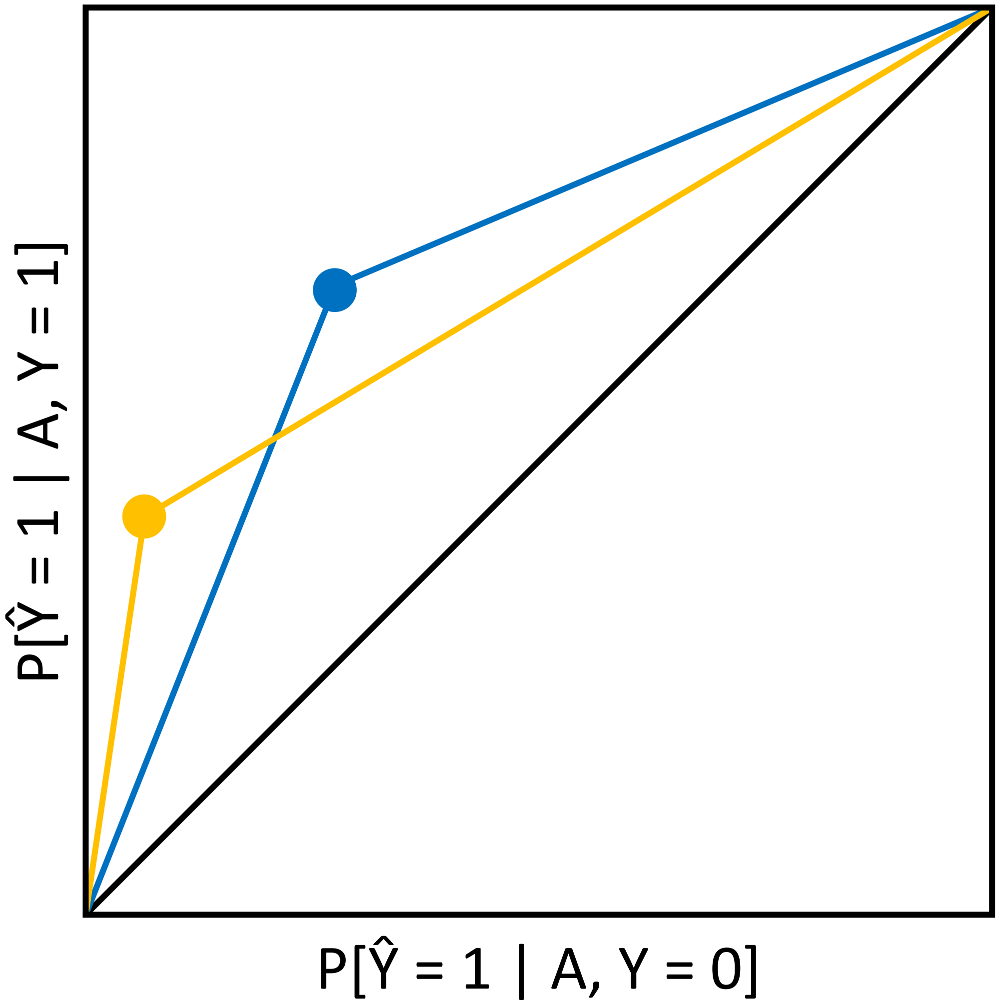
Deriving Classifiers



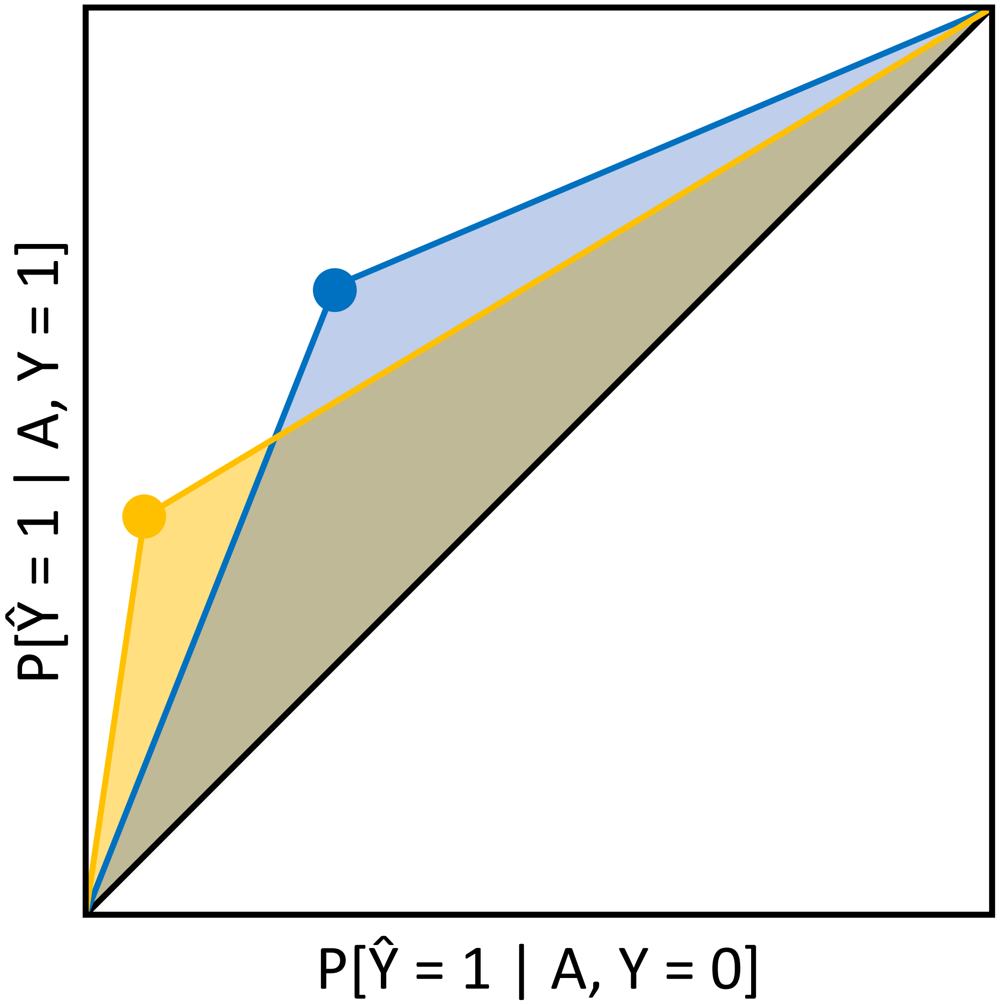
Training of the Postprocessing



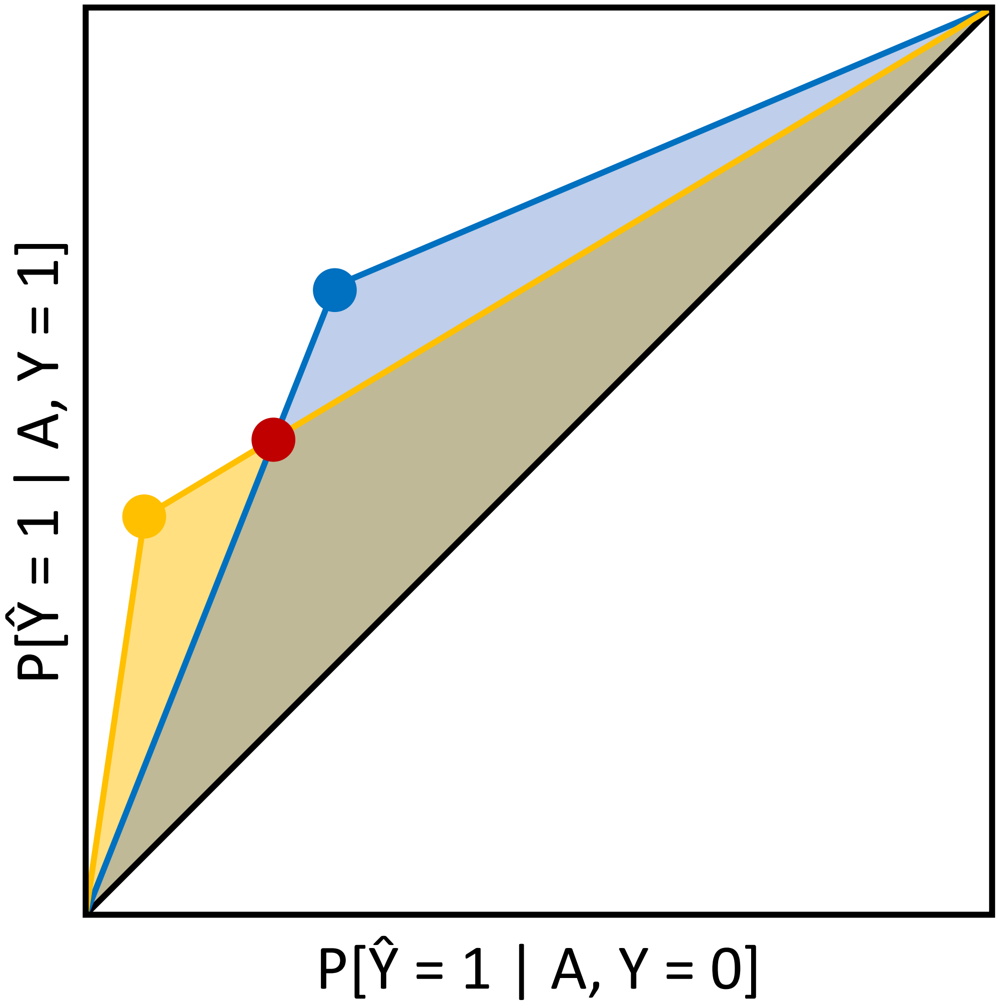
Training of the Postprocessing



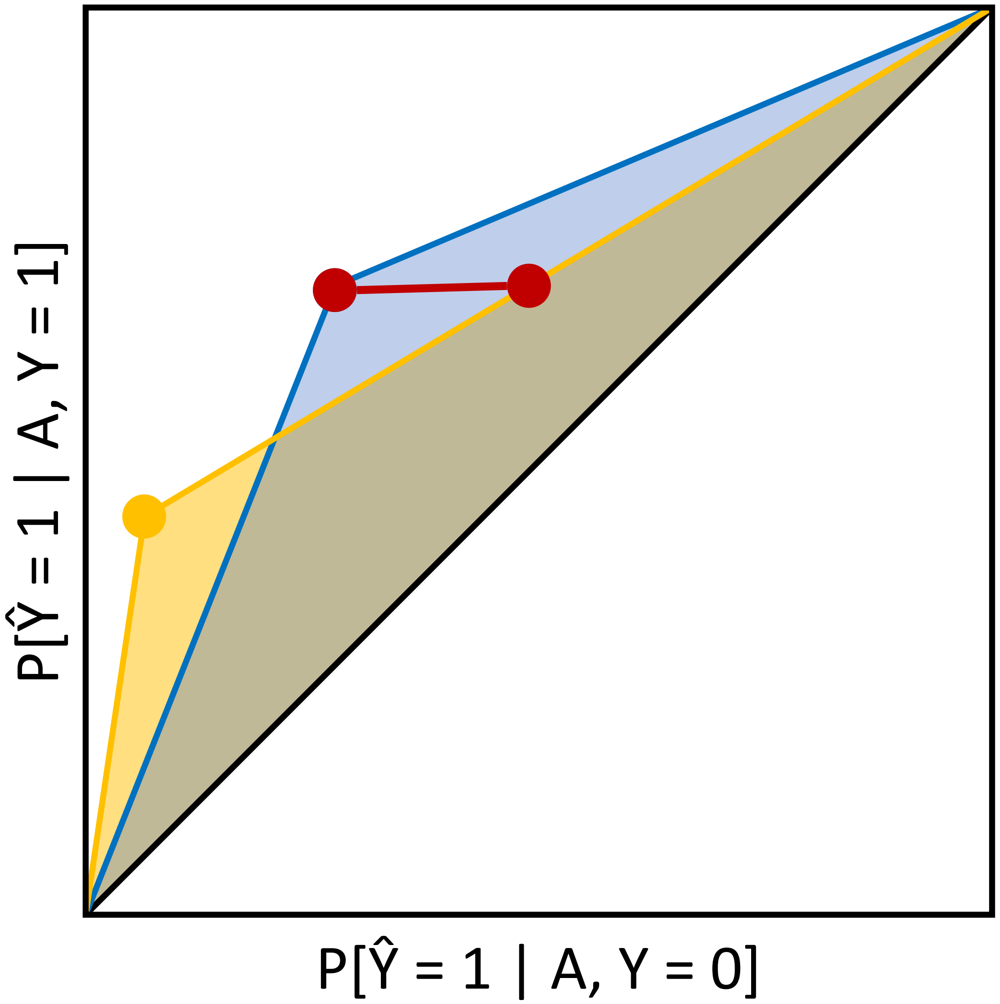
Training of the Postprocessing



Training of the Postprocessing



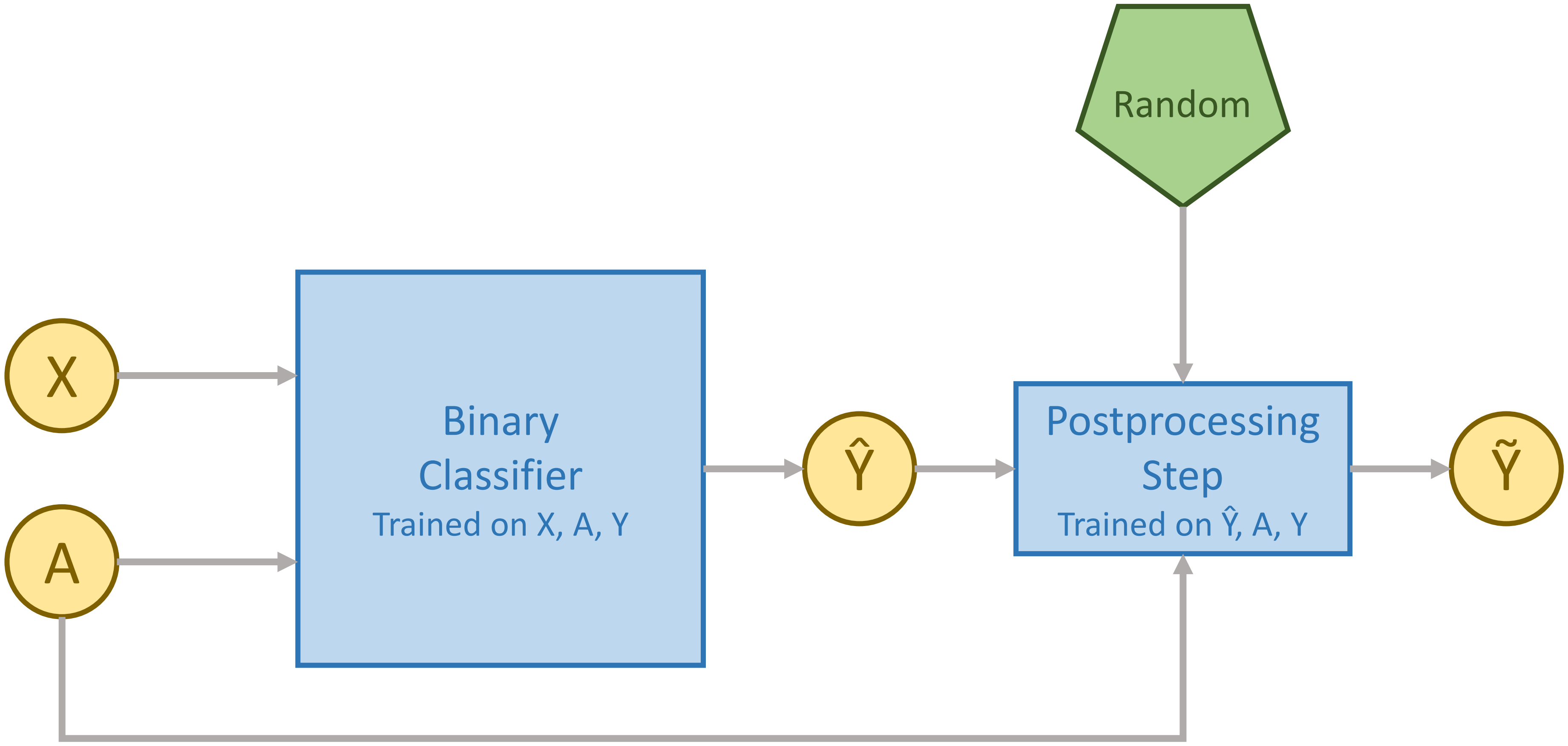
Training of the Postprocessing (Equal Opportunity)



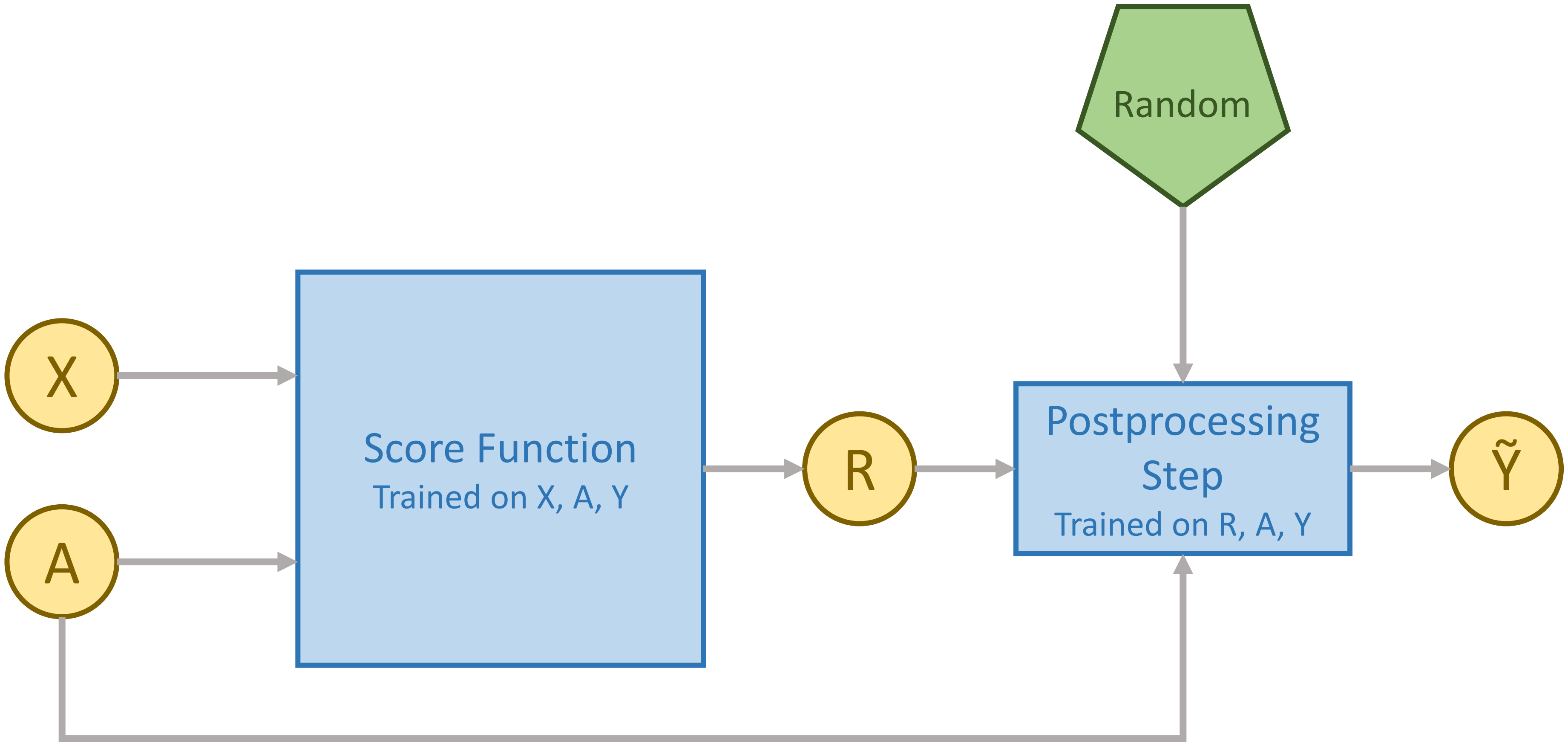
Training using Linear Programming

- Four Parameters: $P[\tilde{Y} = 1 \mid \hat{Y} = \hat{y}, A = a]$, $\hat{y} \in \{0, 1\}$, $a \in \{0, 1\}$
- Objective: minimize $\mathbb{E}[\ell(\tilde{Y}, Y)]$
- Constraints:
  - In the feasible region for each A
  - Same False Positive Rate (X-Axis)
  - Same False Negative Rate (Y-Axis)

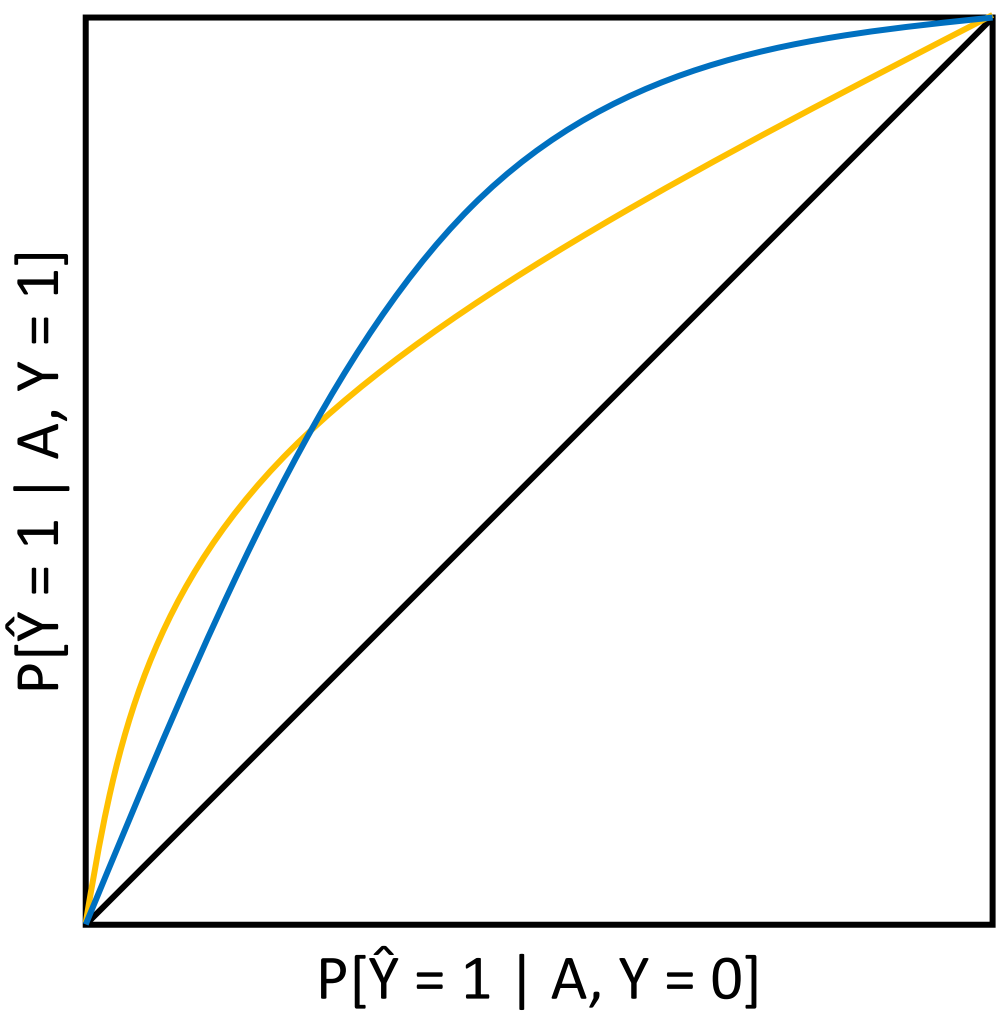
General Procedure



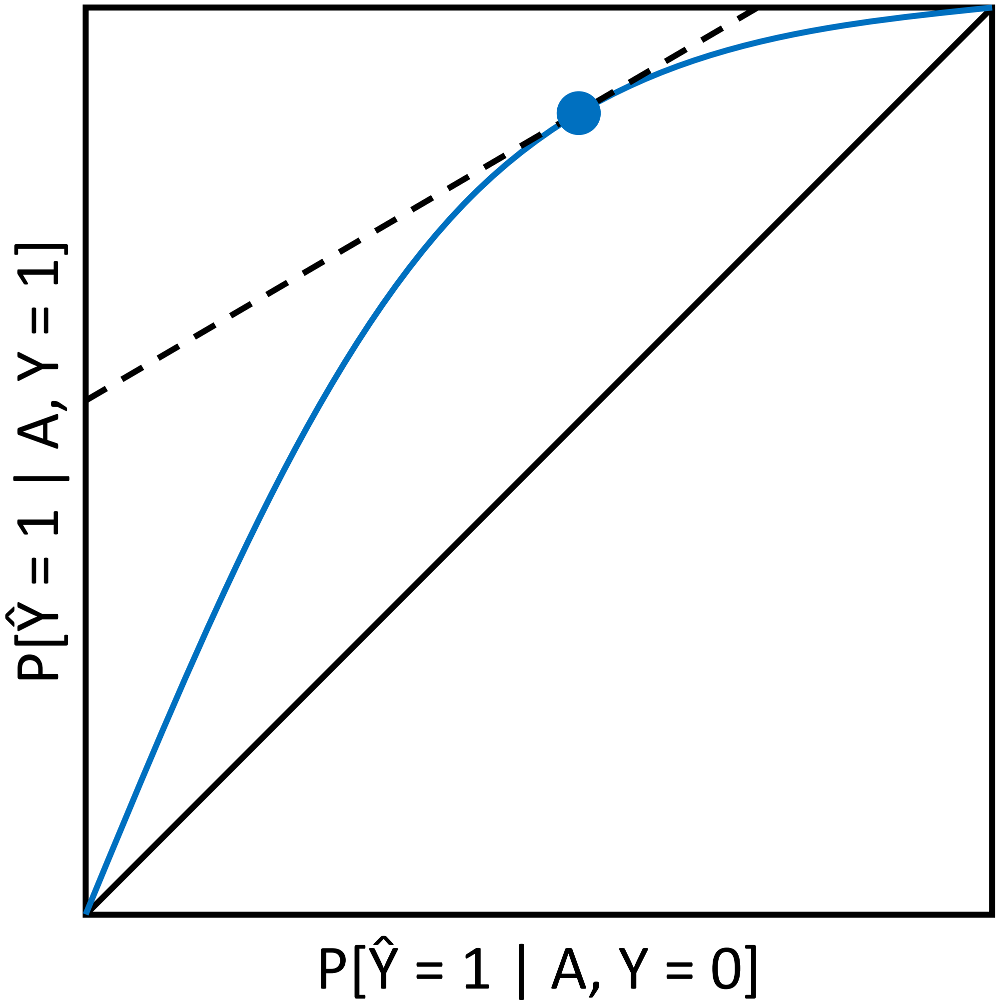
General Procedure



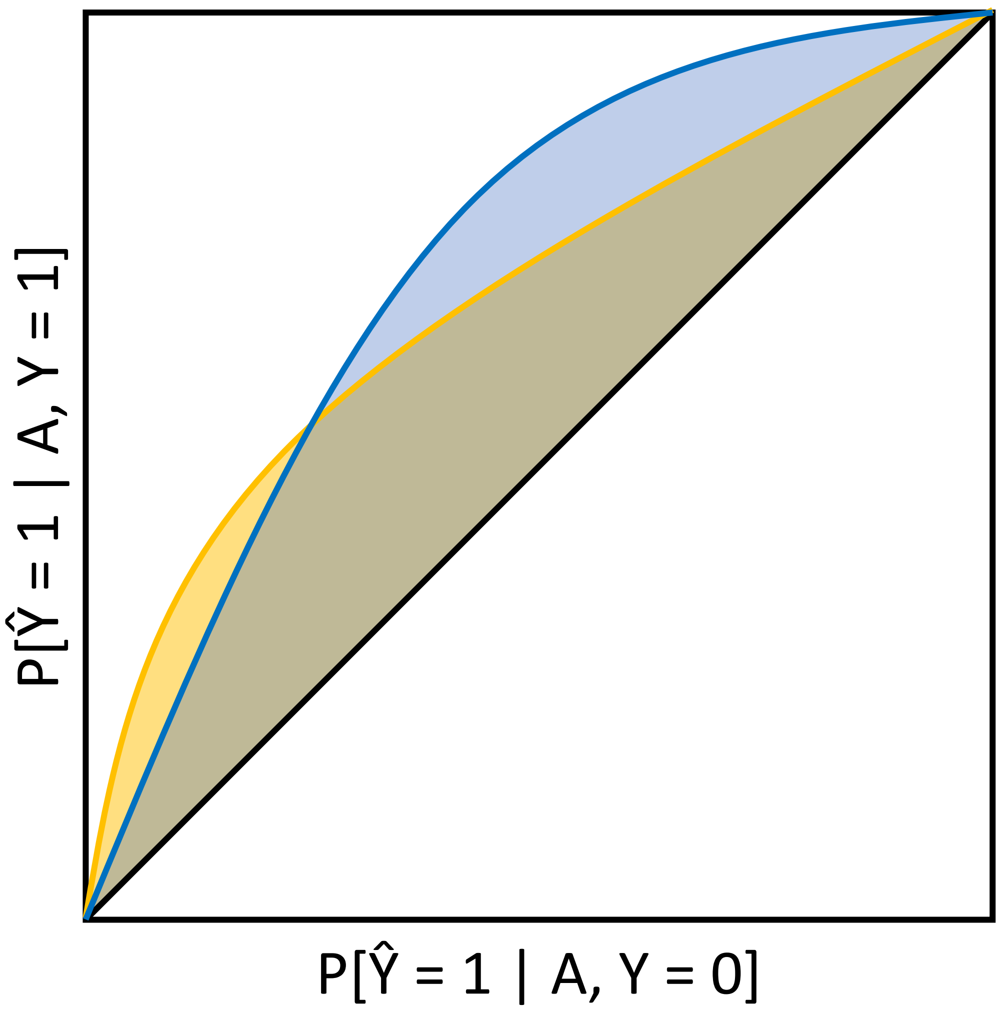
Training of the Postprocessing



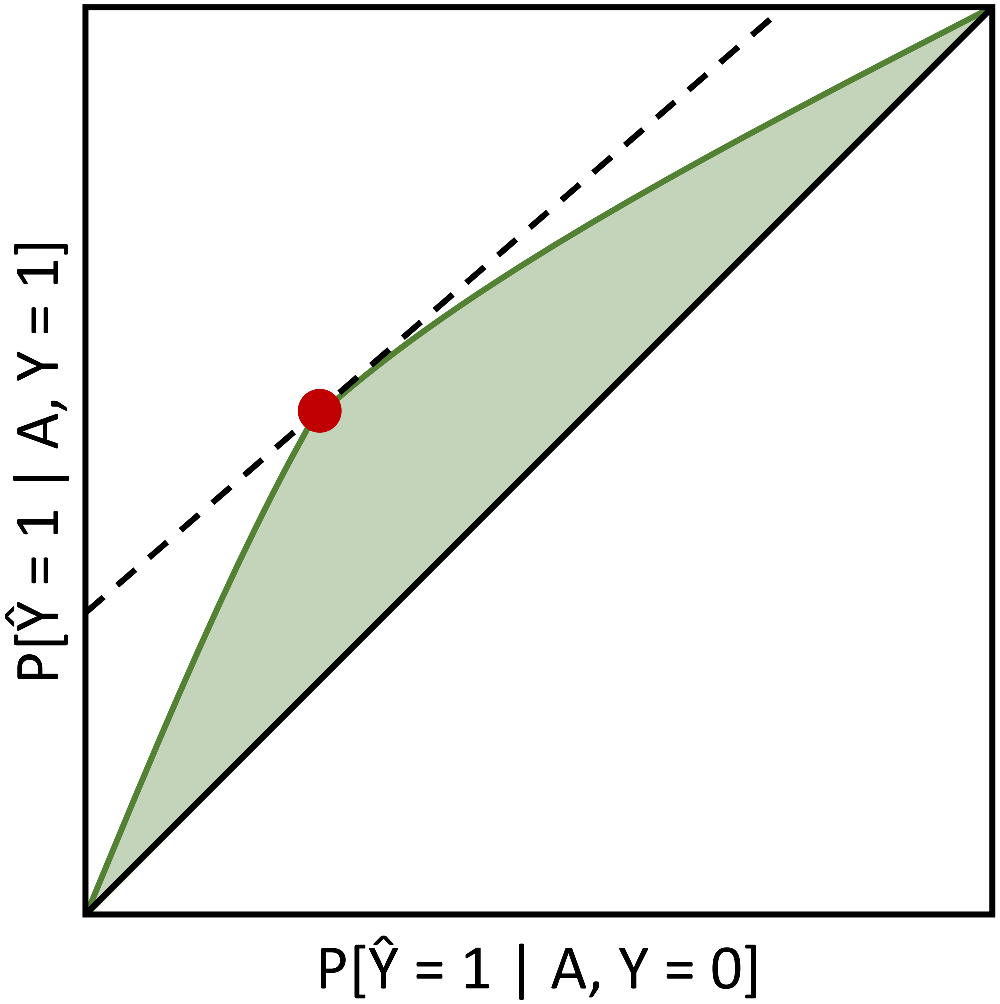
Finding the Optimal Classifier



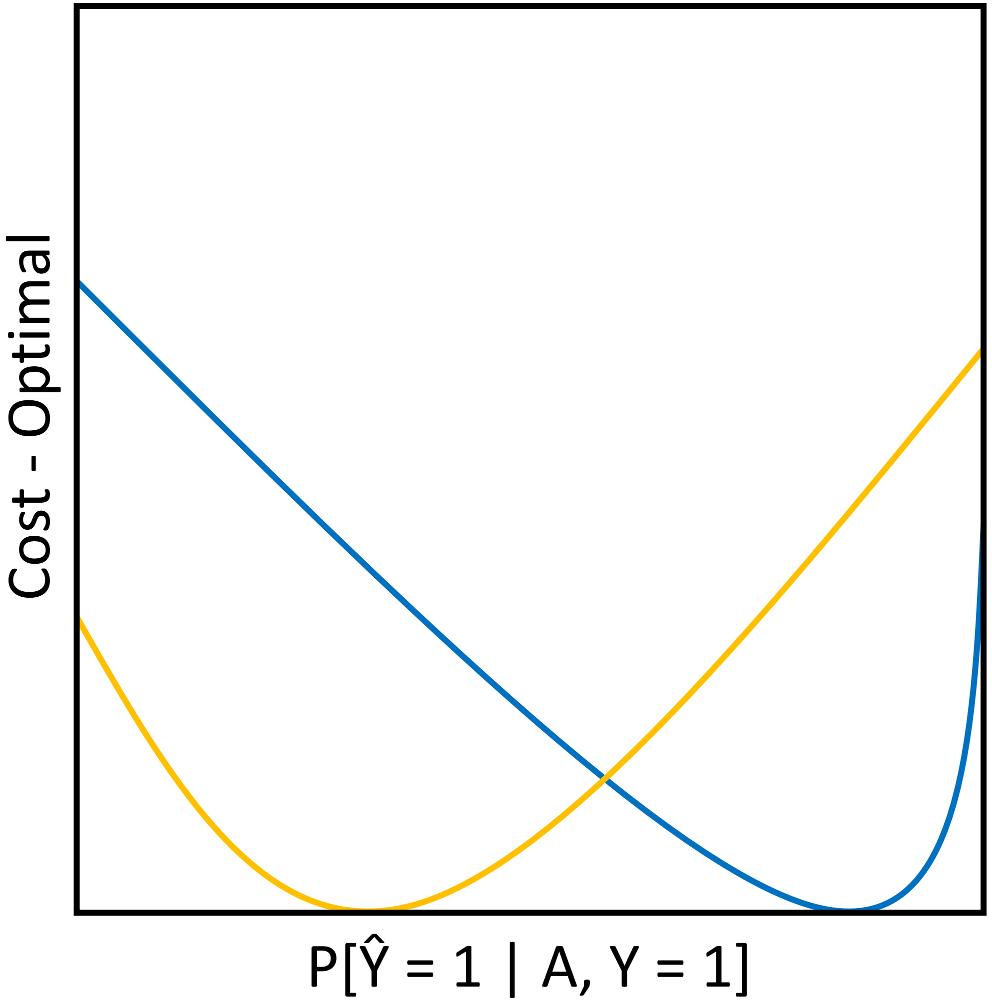
Training of the Postprocessing



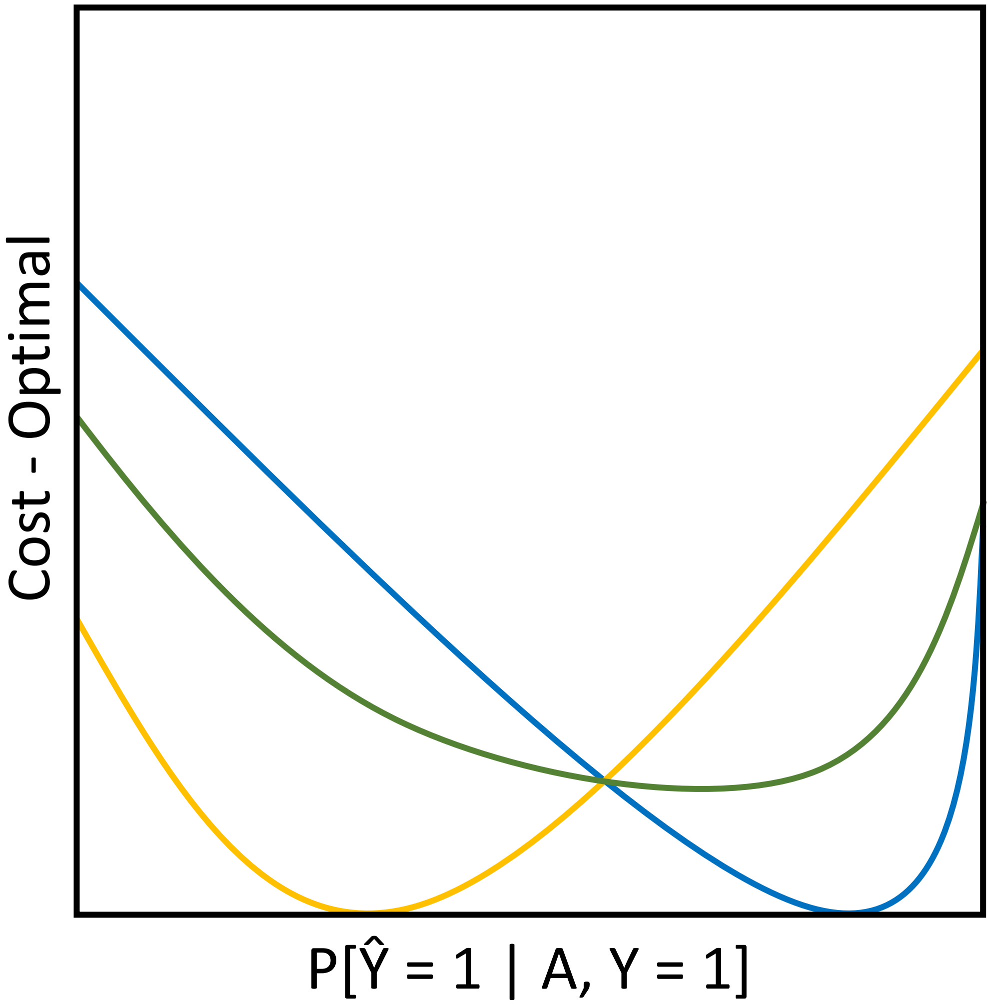
Training of the Postprocessing



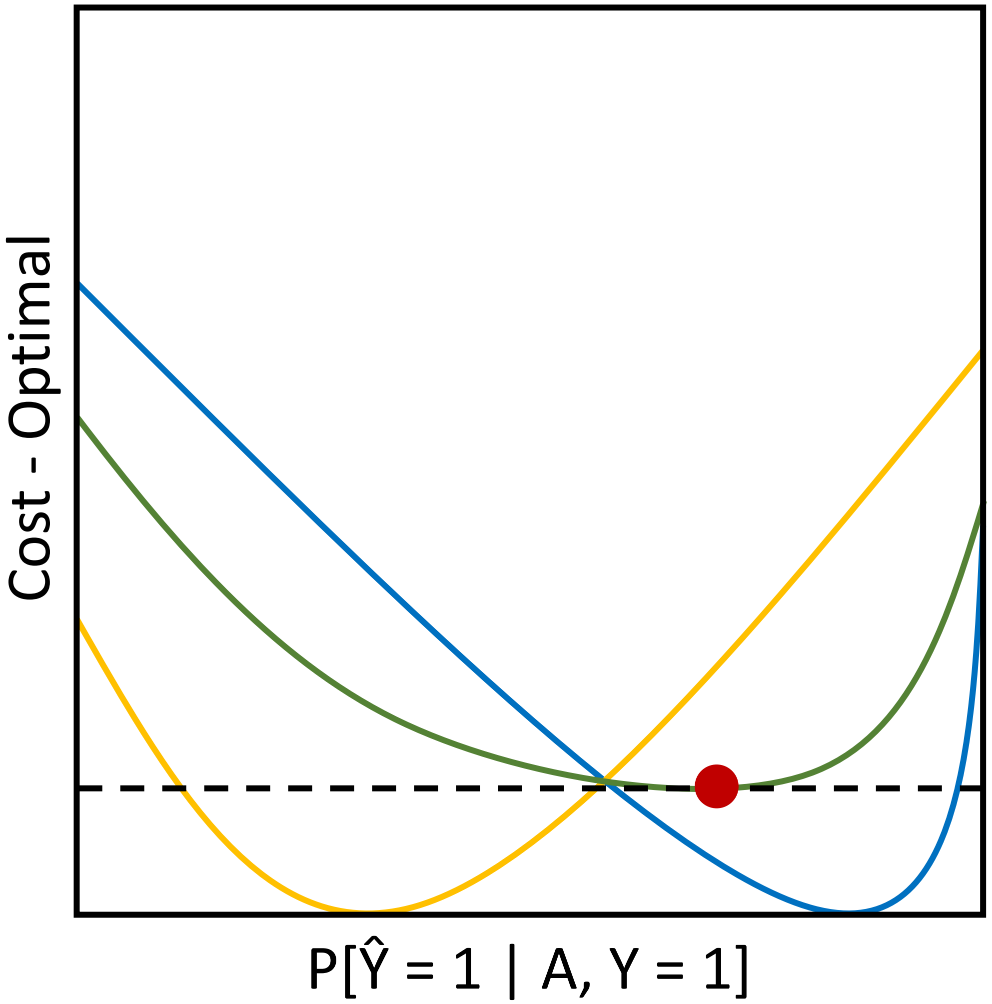
Training of the Postprocessing (Equal Opportunity)



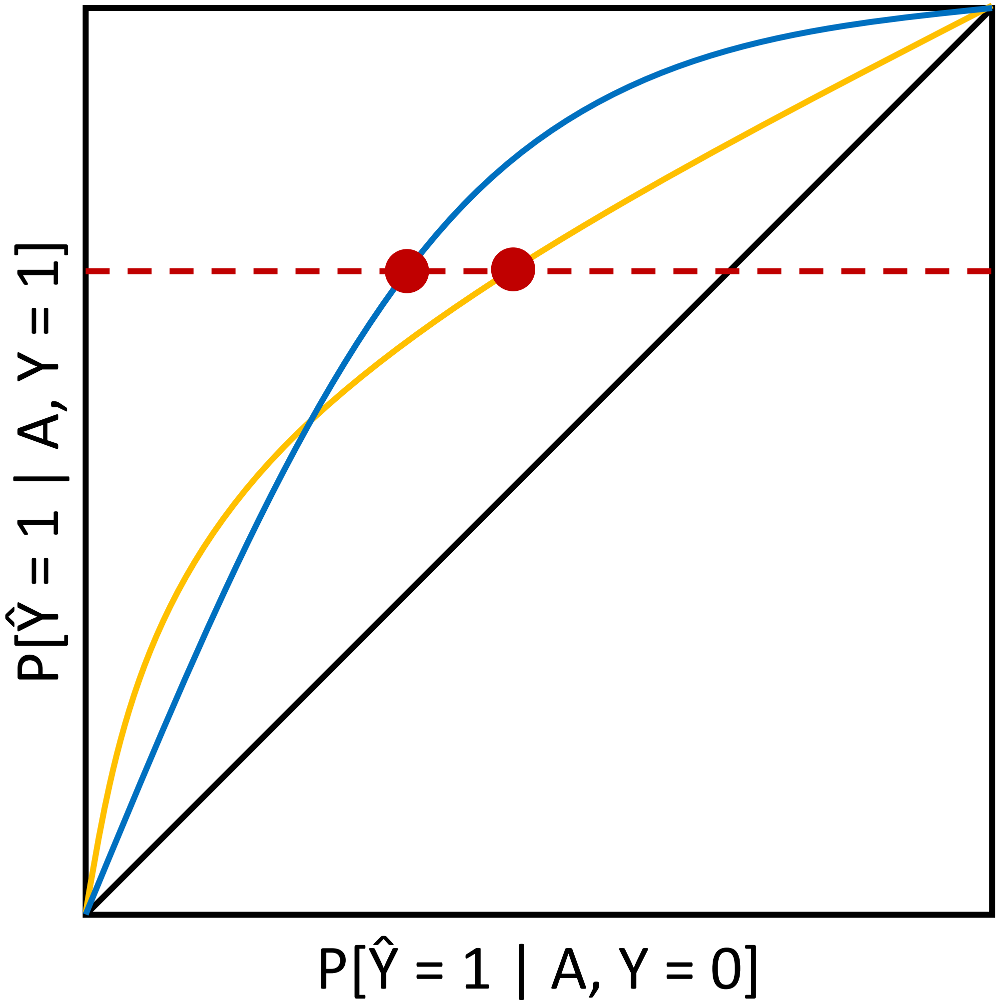
Training of the Postprocessing (Equal Opportunity)



Training of the Postprocessing (Equal Opportunity)



Training of the Postprocessing (Equal Opportunity)



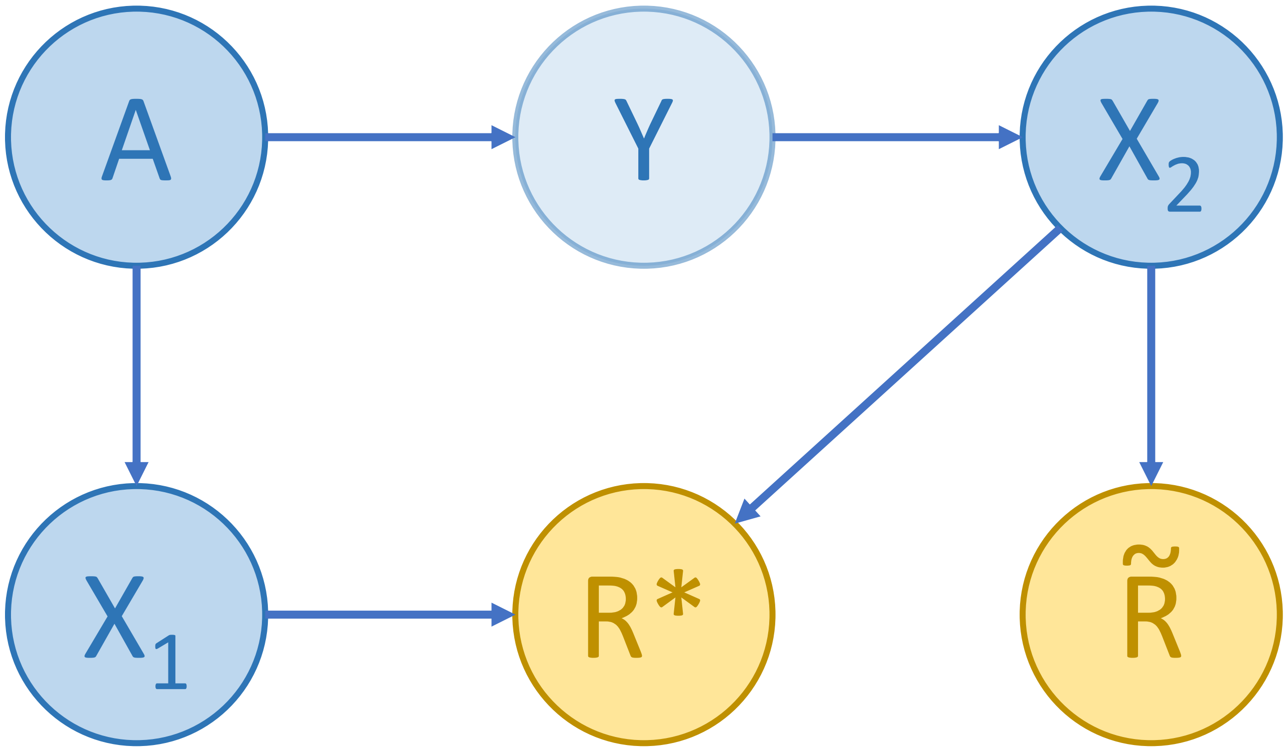
Theoretical Results - Near Optimality

- Bayes optimal regressor R^*: $r^*(x, a) = \mathbb{E}[Y \mid X = x, A = a]$

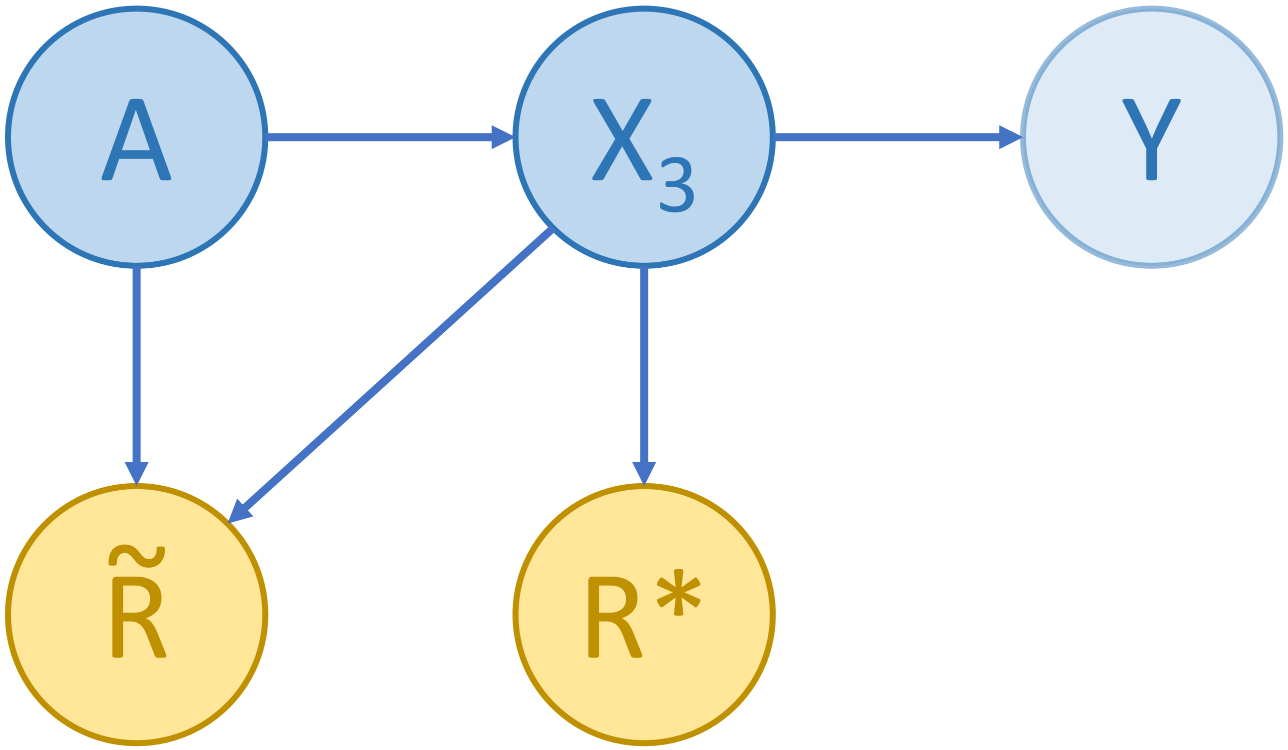
$$\underbrace{\mathbb{E}[\ell(\tilde{Y}, Y)]}_{\text{Loss of derived equalized odds predictor}} - \underbrace{\mathbb{E}[\ell(Y^*, Y)]}_{\text{Best achievable loss for an equalized odds predictor}} \leq \underbrace{d(R, R^*)}_{\text{How far the given regressor is from optimal}}$$

Examples

Scenario I

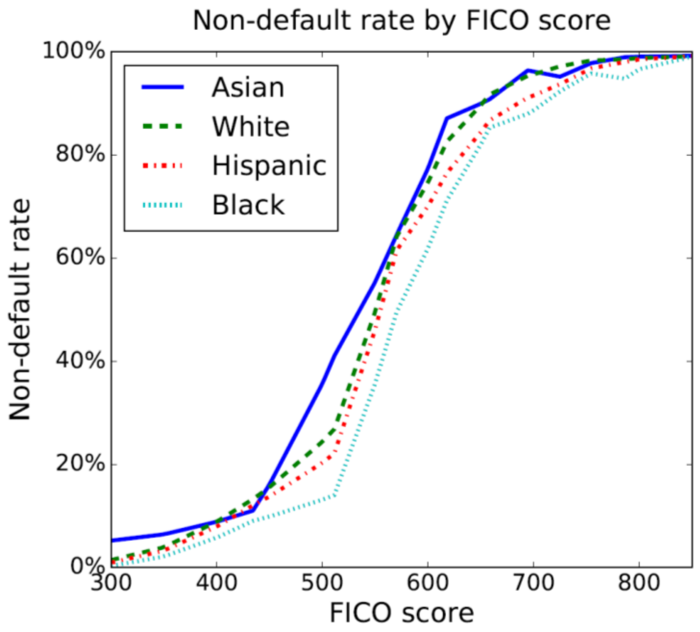


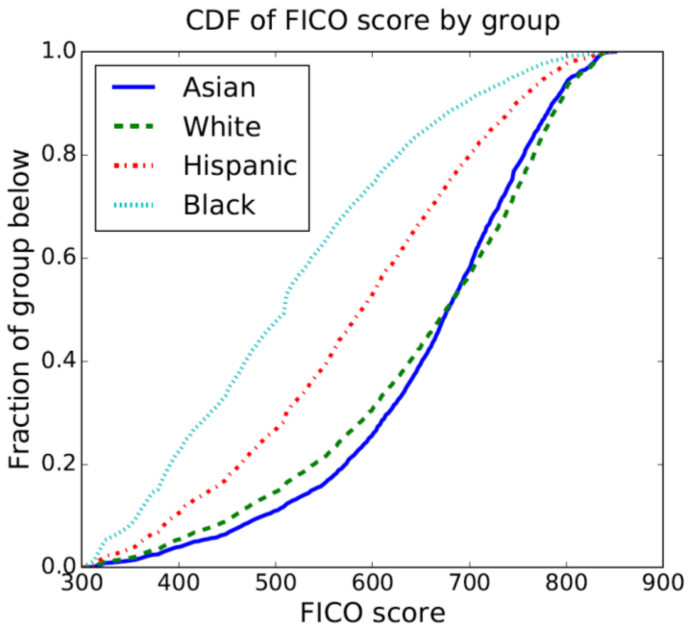
Scenario II



FICO Scores

- Predict credit worthiness
- Y: Failed to pay debt for 90 days on at least one account in 18-24 months
- X: Some features
- A: Race
- Finding a Threshold

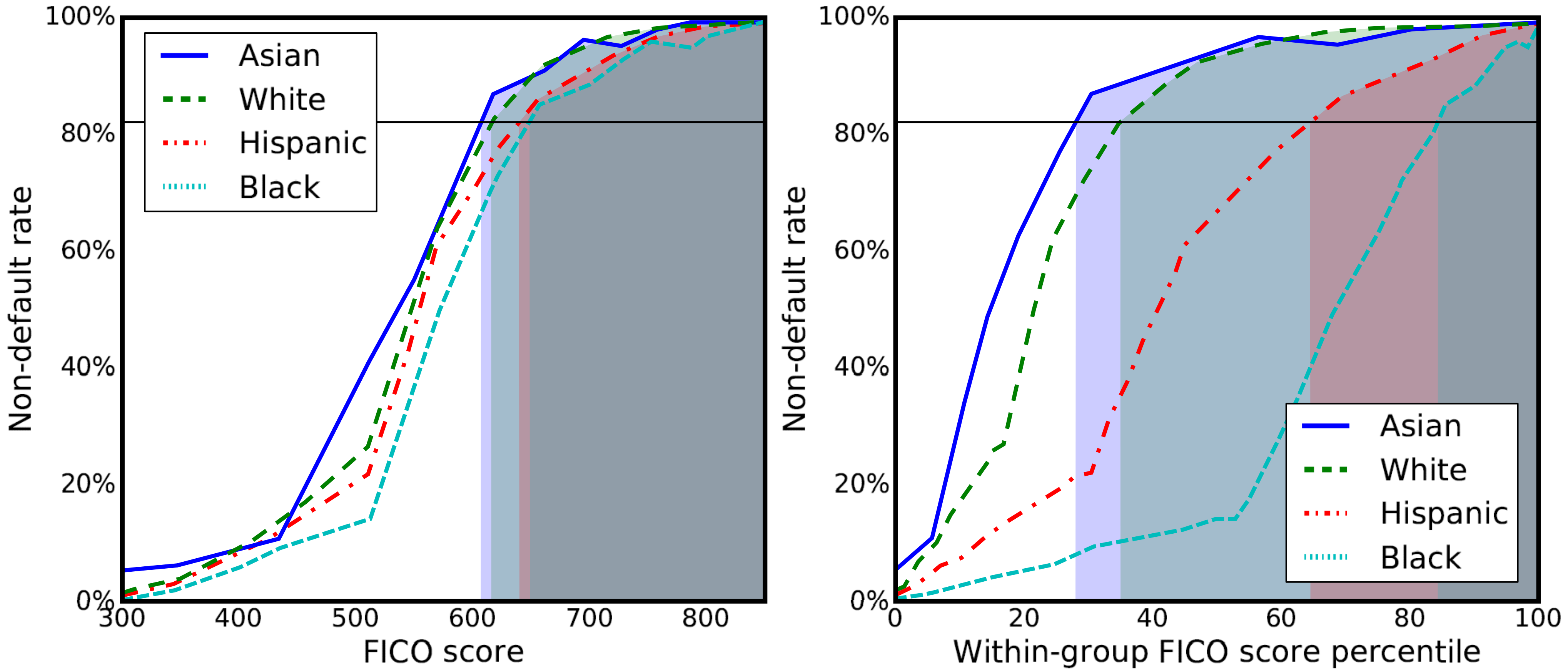




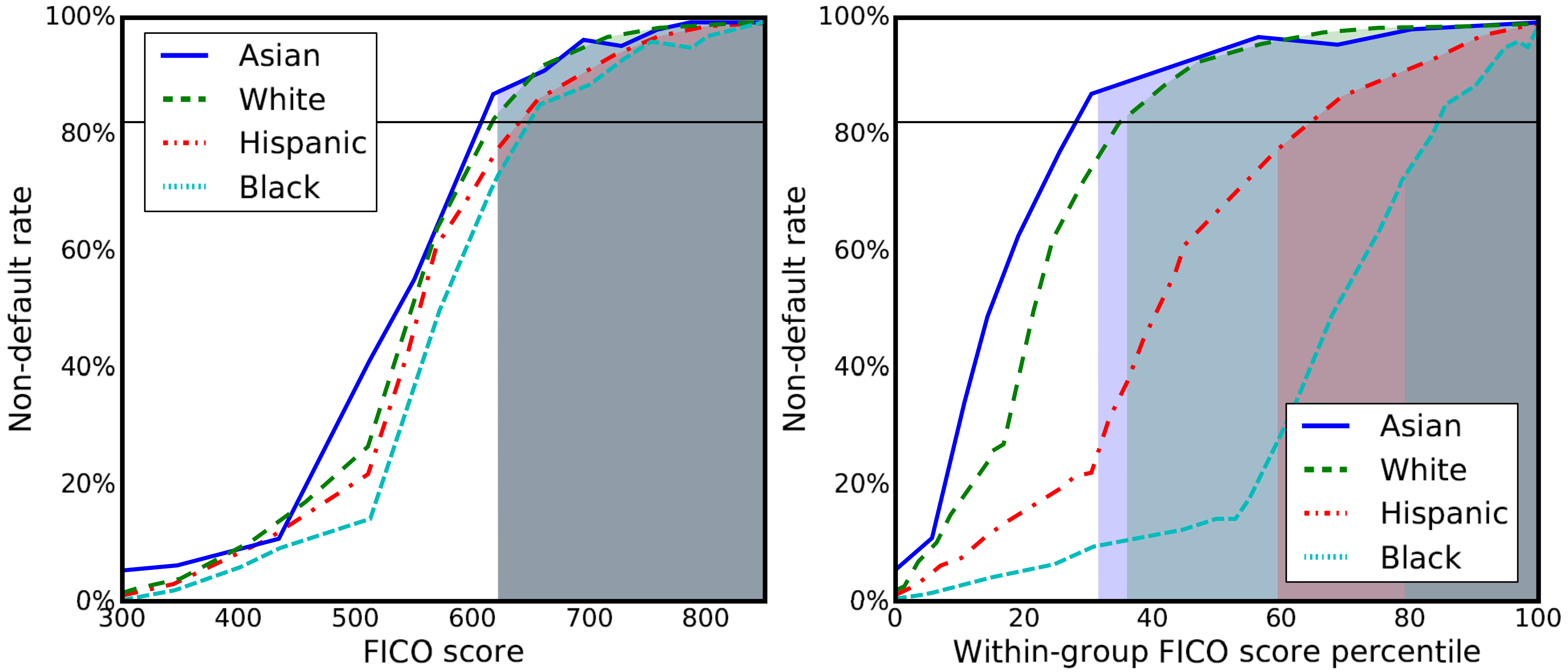
Different Constraints

- Max Profit
  No constraints
- Race Blind
  Same threshold for all groups
- Demographic Parity
  Same fraction of people that qualify for all groups
- Equal Opportunity
- Equalized Odds

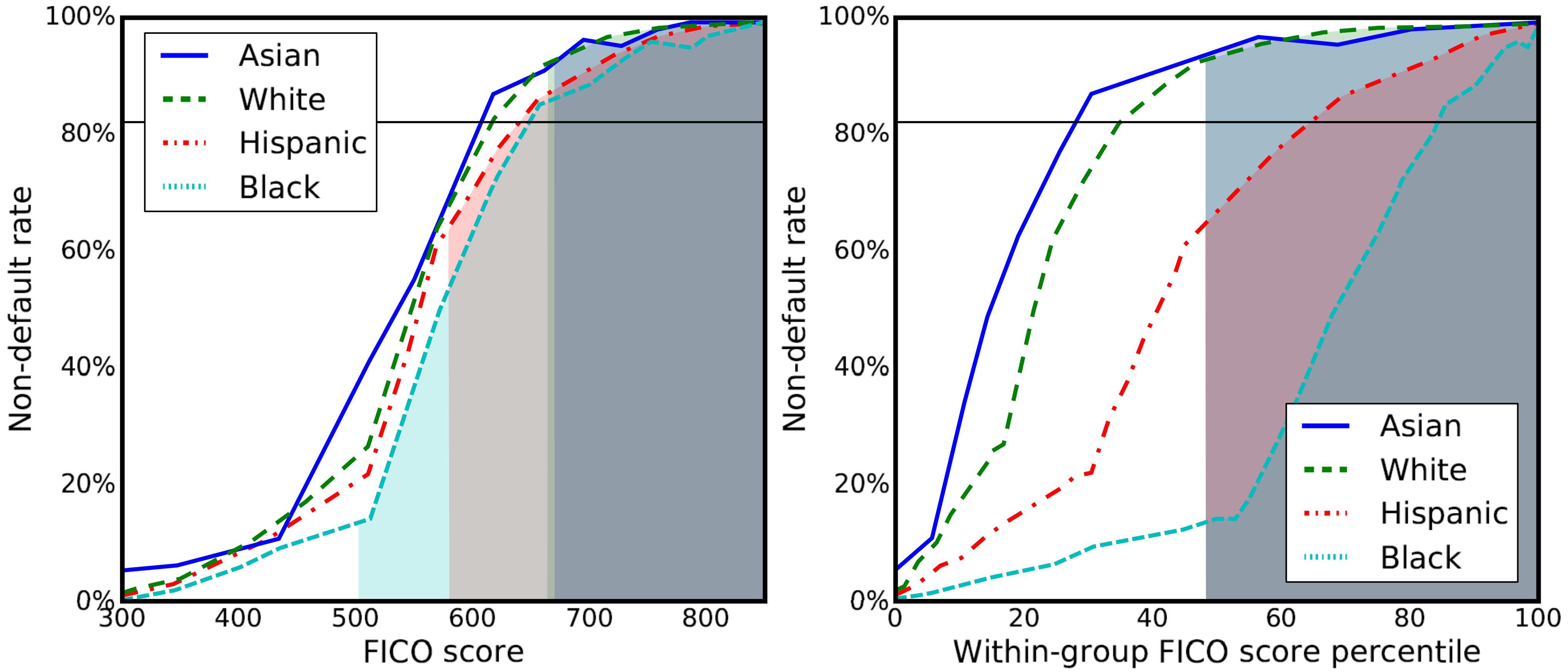
FICO Scores – Max Profit



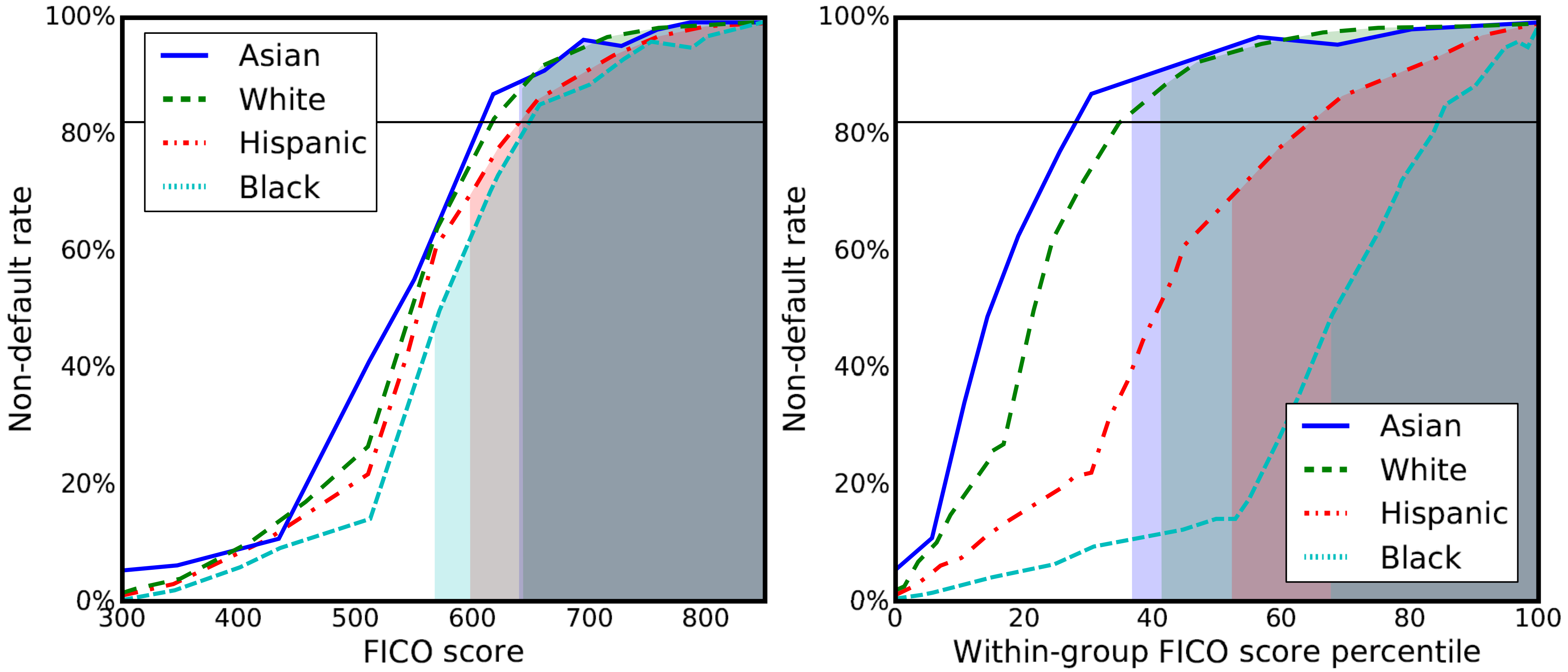
FICO Scores – Single Threshold



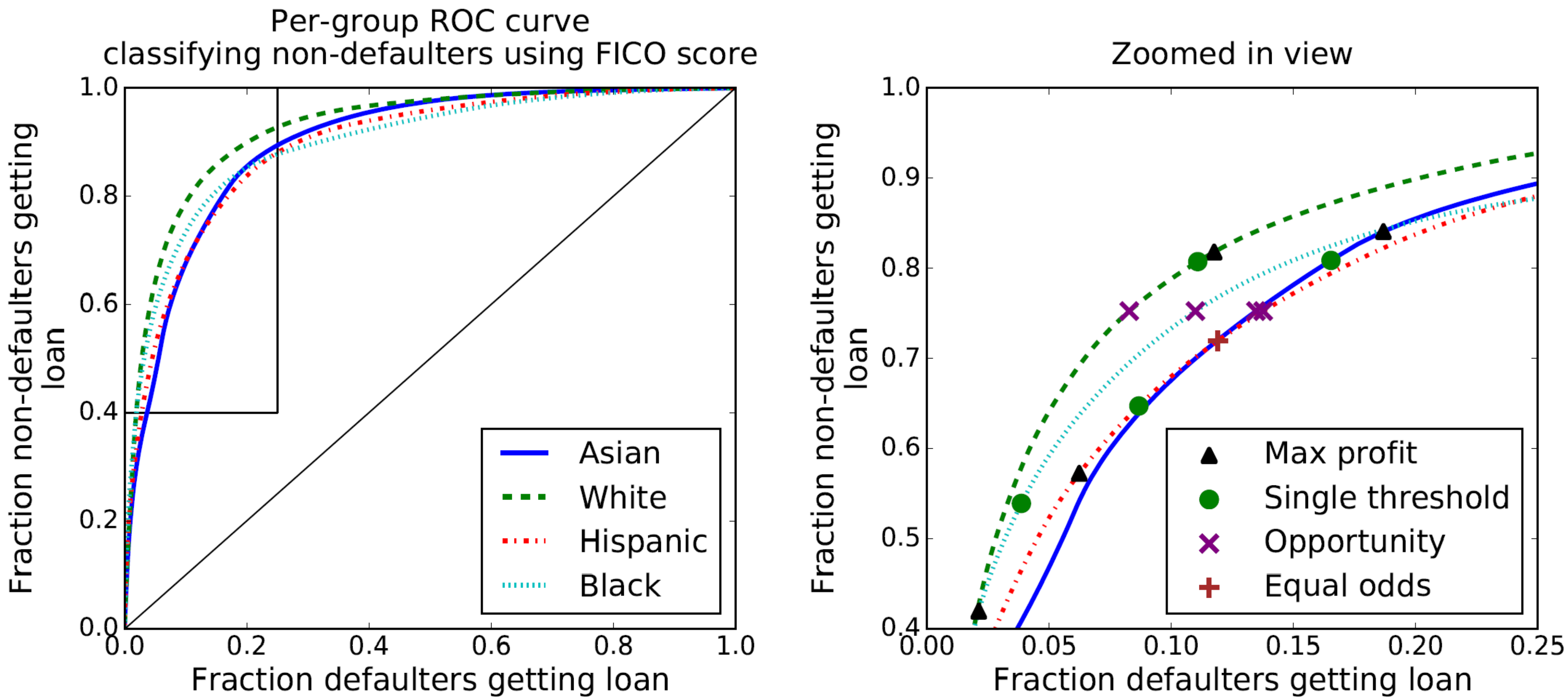
FICO Scores – Demographic Parity



FICO Scores – Equal Opportunity



FICO Scores


Per-group ROC curve
classifying non-defaulters using FICO score
Fraction non-defaulters getting loan
Fraction defaulters getting loan
0.0
0.2
0.4
0.6
0.8
1.0
Asian
White
Hispanic
Black
Zoomed in view
Fraction non-defaulters getting loan
Fraction defaulters getting loan
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.00
0.05
0.10
0.15
0.20
0.25
Max profit
Single threshold
Opportunity
Equal odds

Effects

- Classifier performance is reduced to that on the worst-classified group
- Decision maker cannot simply ignore a group
- Incentivized to gather better data

Conclusion

- Proposed a definition of fairness
- Practical algorithm to derive fair classifiers
- Issues pointed out
- Practical application

My Opinion

+ Very practical
+ Allows “real” predictor
- Shortcomings of obliviousness

• It is a good “last thing you can try” to achieve “fairness”
• Societal issues cannot be fixed by tuning ML classifiers

Questions and Discussion