Binary choice with asymmetric loss in a data-rich environment: theory and an application to algorithmic fairness

Andrii Babii [1] Xi Chen [1] Eric Ghysels [1] Rohit Kumar [2]

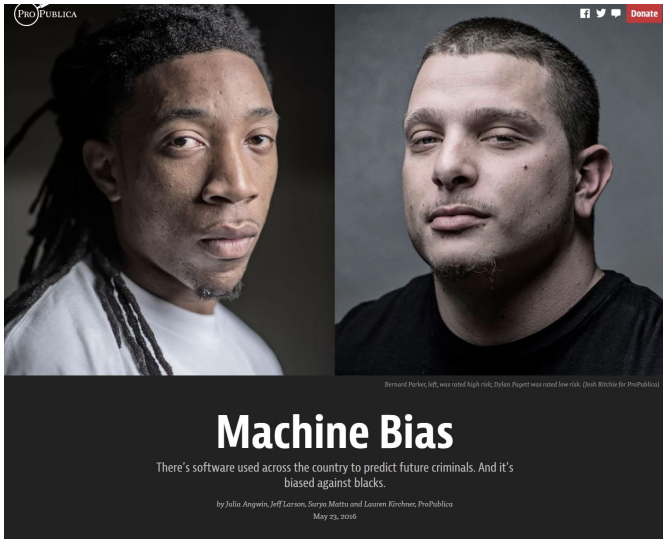
[1]UNC Chapel Hill

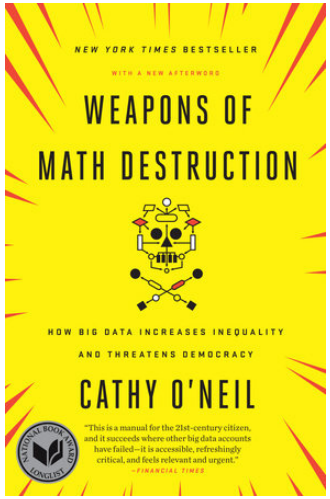
[2]IIT Delhi

Paris 2026

The University of North Carolina at Chapel Hill

Motivation: algorithmic discrimination





... college admission, firing, ad delivery, search engine bias...

Motivation: a concrete decision problem

1. **A judge must decide: detain or release on a bail?**
   - Wrongly jailing an innocent person is costly (false positive).
   - Wrongly releasing a recidivist is costly (false negative).
   - These costs are not the same, and they may differ across groups.
2. Same structure arises in hiring, lending, medical testing, ad delivery...
3. **The gap:**
   - Economists have studied this (Manski, 1975; Elliott, 2013), but existing methods require NP-hard optimization.
   - ML has scalable algorithms, but typically assumes symmetric costs.
   - We bridge both.

Main idea

Recipe

To make optimal binary decisions with asymmetric losses:

1. Specify a loss function $\ell(f, y, x)$ encoding your asymmetries and group preferences.
2. Compute weights ω_i directly from ℓ.
3. Run weighted logistic regression, LASSO, boosting, SVM, or deep learning.

No distributional assumptions on $\Pr(Y = 1|X)$ are needed.

Main contributions

1. **Practical:** asymmetric binary decisions reduce to weighted standard ML — logit, LASSO, boosting, SVM, deep learning.

2. **Theoretical:** excess risk bounds for all methods; asymmetric deep learning is minimax optimal.

3. **Economic:** optimal rule is a covariate-driven threshold $\text{sign}(\eta(x) - c(x))$, where $c(x)$ is the cost-benefit ratio. Connects to Kleinberg et al. (QJE, 2018) and Rambachan et al. (AER P&P, 2020).

4. **Empirical:** application to COMPAS bail data — can equalize error rates across racial groups by adjusting loss weights.

Related literature

1. Asymmetric regression: Koenker and Bassett (ECMA, 1978); Newey and Powell (ECMA, 1987) $\rightarrow$ *well-developed for continuous outcomes, not binary decisions.*

2. Nonparametric binary choice: Manski (JoE, 1975), Elliott and Lielli (JoE, 2013) $\rightarrow$ NP-hard, *no scalable algorithms with guarantees.*

3. Binary classification: Vapnik (1995), Zhang (AoS, 2004), Bartlett, Jordan, McAuliffe (JASA, 2006) $\rightarrow$ *assumes symmetric losses.*

4. Neural network regressions: Chen (Handbook, 2007), Farrell, Liang, Misra (ECMA, 2021) $\rightarrow$ *regression, not classification.*

5. Bail decisions: Kleinberg et al. (QJE, 2018) $\rightarrow$ *motivates our empirical application.*

Roadmap

1. Binary decisions and loss functions
   - Examples of loss functions
   - Optimal decision
   - Empirical risk minimization
2. Convexification
   - Convexified empirical risk minimization (ERM)
   - Optimal decision
   - Examples: asymmetric logit and ML
3. Excess risk bounds for asymmetric ML
   - Convexification result
   - Linear decision rules
   - Shallow and Deep learning
4. Monte Carlo experiments
5. Racial bias and recidivism revised

Binary decisions

1. Notation:
   - $Y \in \{-1, 1\}$, target variable;
   - $X \in \mathcal{X} \subset \mathbb{R}^p$, covariates;
   - $f : \mathcal{X} \to \{-1, 1\}$, binary decision/prediction/choice;
   - $\ell : (f, y, x) \mapsto \ell(f(x), y, x)$, loss function.

2. Objective: risk minimization over all possible binary decisions $f(X)$

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)}[\ell(f(X), Y, X)].$$

3. See Elliott and Lielli (2013, JoE) for an equivalent utility-based framework.

Example 1: preferences towards protected group

Example (Binary decision with a protected group)

The loss function

$$\ell(f(x), y, g) = \psi_g \mathbb{1}\{f(x) \neq y\}$$

has different weights $\psi_g > 0$ for the group $g \in \{0, 1\}$.
$\psi_1 > \psi_0$ means that group $g = 1$ is protected.

More generally, we can have different losses for keeping a non-recidivist in jail and releasing a recidivist

$$\ell(f(x), y, g) = \underbrace{\varphi_g \mathbb{1}\{f(x) = 1, y = -1\}}_{\text{false positive}} + \underbrace{\psi_g \mathbb{1}\{f(x) = -1, y = 1\}}_{\text{false negative}}.$$

Example 2: Rambachan et al., 2020 (AER P&P)

Example (Social planner with a disadvantaged group)

The social planner's loss/welfare function is

$$\ell(f(z,g),y,g) = \psi_g \ell(f(z,g),y,z,g),$$

where $\psi_g > 0$ are the generalized social welfare weights placed on group $g \in \{0,1\}$.

If $\psi_1 > \psi_0$, then the outcomes associated with a disadvantaged group $g = 1$ are valued more.

Choice of the loss function

The decision maker has to specify preferences through the loss function: $\ell(f, y, x) = \ell_{f,y}(x)$, where $f, y \in \{-1, 1\}$ and $x \in \mathcal{X}$

pred. \ true	$Y = 1$	$Y = -1$
$f = 1$	$\ell_{1,1}(x)$	$\ell_{1,-1}(x)$
$f = -1$	$\ell_{-1,1}(x)$	$\ell_{-1,-1}(x)$

Two choices:

- covariates: group, economic costs/benefits;
- functional forms.

Optimal decision

1. The optimal decision rule f^* achieves the smallest risk, denoted

$$\mathcal{R}^* = \inf_{f:\mathcal{X}\to\{-1,1\}} \mathbb{E}[\ell(f(X), Y, X)],$$

where $X \in \mathbb{R}^p$ can have the group membership indicator $G \in \{0, 1\}$.

2. Minimax approach: construct a data-driven binary decision rule $\hat{f}_n : \mathcal{X} \to \{-1, 1\}$ from an i.i.d. sample $(Y_i, X_i)_{i=1}^n$ minimizing the expected excess risk

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\mathcal{R}(\hat{f}_n) - \mathcal{R}^*\right]$$

for a large class of distributions of $(Y, X) \sim P \in \mathcal{P}$.

Empirical risk minimization (ERM)

1. We obtain an equivalent characterization of the optimal decision f^*

$$\inf_{f:\mathcal{X}\to\{-1,1\}} \mathbb{E}[\omega \mathbb{1}\{-Yf(X) \geq 0\}],$$

   where $\omega = \omega(Y, X)$ is computed from the loss function ℓ.

2. Empirical risk minimization problem: non-smooth, non-convex, NP-hard

$$\inf_{f:\mathcal{X}\to\{-1,1\}} \frac{1}{n}\sum_{i=1}^{n} \omega_i \mathbb{1}\{-Y_i f(X_i) \geq 0\}.$$

Comment: standard binary classification when $\omega_i = 1, \forall i \leq n$

$$\inf_{f:\mathcal{X}\to\{-1,1\}} \Pr(Y \neq f(X)).$$

Roadmap

1. Binary decisions and loss functions
   - Examples of loss functions
   - Optimal decision
   - Empirical risk minimization
2. Convexification
   - Convexified empirical risk minimization (ERM)
   - Optimal decision
   - Examples: asymmetric logit and ML
3. Excess risk bounds for asymmetric ML
   - Convexification result
   - Linear decision rules
   - Shallow and Deep learning
4. Monte Carlo experiments
5. Racial bias and recidivism revised

Convexified ERM

1. Instead, we take $\hat{f}_n$ solving

$$\inf_{f:\mathcal{X}\to[-1,1]} \frac{1}{n}\sum_{i=1}^{n} \omega_i \phi(-Y_i f(X_i)),$$

   where $\phi(z) \geq \mathbb{1}\{z \geq 0\}$ convexifies the ERM.
2. Data-driven binary decision: $\operatorname{sign}(\hat{f}_n) \in \{-1, 1\}$.
3. $\hat{f}_n$ is an estimator of f^*_ϕ that minimizes the convexified risk

$$\mathcal{R}_\phi(f) = \mathbb{E}[\omega\phi(-Yf(X))].$$

4. How does f^*_ϕ compare to the binary decision rule f^* that is optimal with respect to the risk $\mathcal{R}$?

Optimal decision

Theorem

Under mild conditions for a generic class of ϕ

$$\operatorname{sign}(f_\phi^*(x)) = \operatorname{sign}(f^*(x)) = \operatorname{sign}(\eta(x) - c(x)),$$

where $\eta(x) = \Pr(Y = 1|X = x)$ *and*

$$c(x) = \frac{\ell_{1,-1}(x) - \ell_{-1,-1}(x)}{\ell_{1,-1}(x) - \ell_{-1,-1}(x) + \ell_{-1,1}(x) - \ell_{1,1}(x)}.$$

Economic interpretation:

- $c(x)$ = fraction of total error cost due to false positives.
- **Predict** $f(x) = 1$ **when** $\Pr(Y = 1|X = x)$ **exceeds this fraction.**
- Symmetric case: $c(x) = 1/2$, standard classification.
- When false positives are costlier (e.g., jailing the innocent): threshold increases, fewer detained.

Examples of convexifying functions

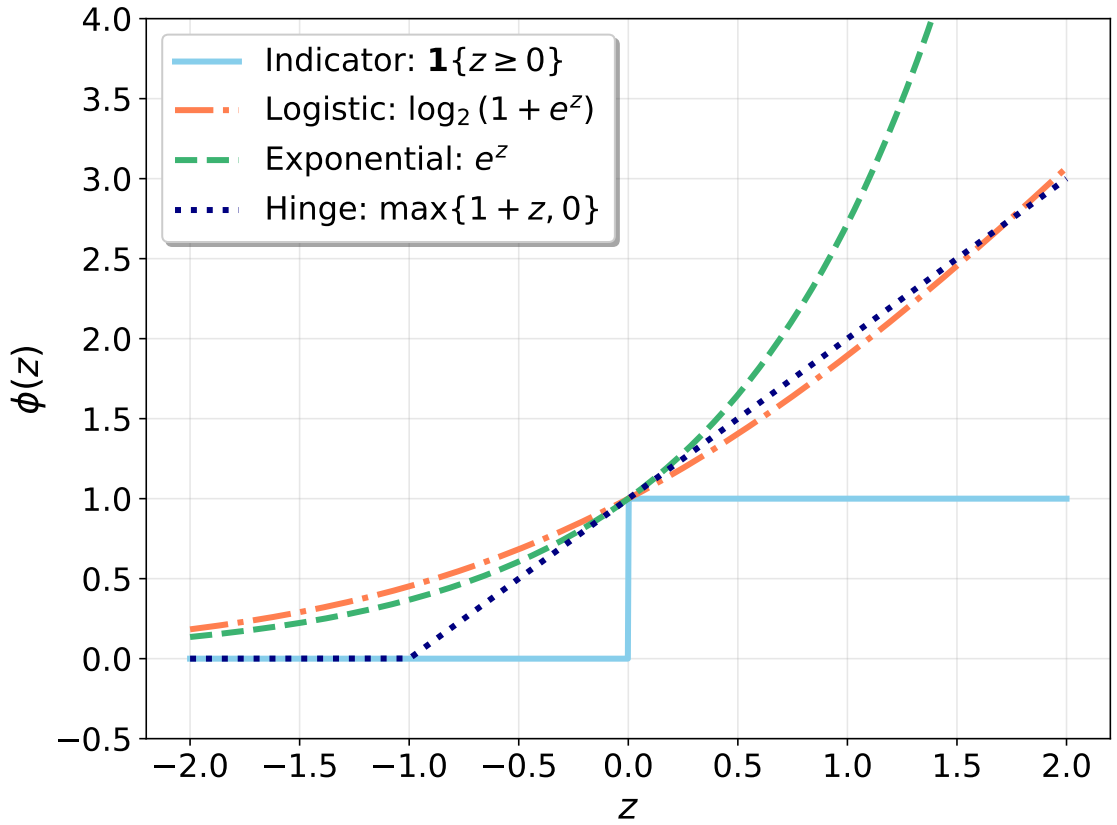


Figure: Convexifications corresponding to Logit/Boosting/LASSO and SVM/Deep learning.

Example 1: Asymmetric Logit/Boosting (XGBoost)

Example (Logistic convexification, $\phi(z) = \log(1 + e^z)$)

1. Log-likelihood of Logit is proportional to

$$f \mapsto \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + e^{-Y_i f(X_i)}\right).$$

2. Our methods: asymmetric Logit/Boosting (XGBoost)

$$f \mapsto \frac{1}{n} \sum_{i=1}^{n} \omega_i \log \left(1 + e^{-Y_i f(X_i)}\right).$$

Comments:

- reweighting for the asymmetry of ℓ through ω_i;
- Extreme Gradient Boosting (XGBoost): minimize via the functional gradient decent with early stopping.

Example 2: Asymmetric LASSO

Example (Logistic convexification, $\phi(z) = \log(1 + e^z)$)

1. LASSO minimizes

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-Y_i X_i^\top \theta}\right) + \lambda_n |\theta|_1$$

2. Our method: asymmetric LASSO

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \omega_i \log\left(1 + e^{-Y_i X_i^\top \theta}\right) + \lambda_n |\theta|_1.$$

Example 3: Asymmetric SVM/Deep learning

Example (Hinge convexification, $\phi(z) = (1+z)_+$)

1. Support vector machines (SVM), Vapnik (1995)

$$f \mapsto \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+$$

2. Our method: asymmetric SVM

$$f \mapsto \frac{1}{n}\sum_{i=1}^{n}\omega_i(1 - Y_i f(X_i))_+.$$

Comment: solved over the reproducing kernel Hilbert space (SVM) or the neural network sieve (deep learning).

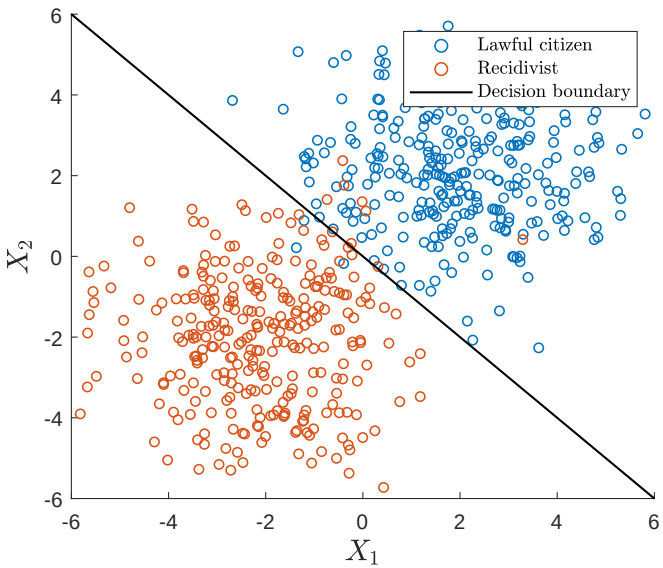
Roadmap

Margin/noise assumption (MA)

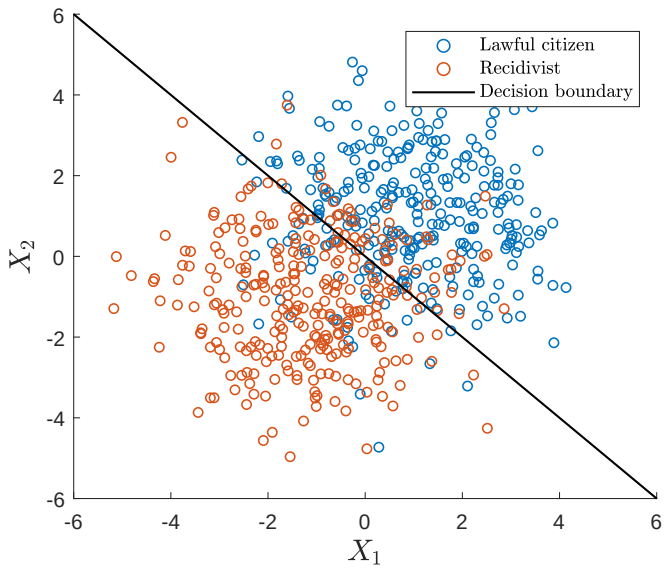
Class of distributions of (Y, X): $\mathcal{P}(\alpha)$ with $X \sim P_X$ satisfying

$$P_X\left(\{x \in \mathcal{X} : \ |\eta(x) - c(x)| \leq u\}\right) \leq C u^{\alpha}, \qquad \forall u > 0.$$

In economics terms: the problem is easier when $\eta(x) = \Pr(Y = 1|X = x)$ is well-separated from the threshold $c(x)$ — i.e., most cases are "clear calls."



(a) Easier (α large)



(b) Harder (α small)

Convexification result

Theorem

Suppose that ϕ is convex, non-decreasing, Lipschitz, and satisfies a mild curvature condition (CFA) with parameter $\gamma \in (0, 1]$. Then under the margin assumption (MA)

$$\mathcal{R}(\operatorname{sign}(f)) - \mathcal{R}^* \lesssim \left[\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}} .$$

Translation: if you solve the convex (weighted ML) problem well, you automatically get a good binary decision.

- The exponent $\frac{\gamma(\alpha+1)}{\gamma\alpha+1}$ is the price of convexification:
  - close to 1 when classes are well-separated (α large);
  - closer to γ when the boundary is noisy (α small).
- All standard convexifications satisfy CFA: logistic/exponential ($\gamma = 1$) and hinge ($\gamma = 1/2$).

Asymmetric linear decision rules

1. Convexified weighted ERM

$$\inf_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \omega_i \phi(-Y_i f(X_i)),$$

   solved over the parametric class $\mathscr{F} = \left\{ f_\theta(x) = x^\top \theta,\ \theta \in \mathbb{R}^p \right\}$.

2. Example: logistic convexifying function,

$$\phi(z) = \log_2(1 + e^z).$$

Asymmetric parametric predictions

Theorem

For the logistic, exponential, and hinge convexification, under stated assumptions

$$\mathbb{E}_P[\mathcal{R}(\operatorname{sign}(\hat{f}_n)) - \mathcal{R}^*] \lesssim \underbrace{\left(\frac{p}{n}\right)^{\frac{1+\alpha}{2+\alpha}}}_{\textit{Variance}} + \underbrace{\left[\inf_{f \in \mathscr{F}} \mathcal{R}(f) - \mathcal{R}^*\right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}}}_{\textit{Bias}}.$$

Comments:

1. for a small number of covariates p, the "variance term" scales at a rate between $O(n^{-1/2})$ and $O(n^{-1})$ depending on the noise α.
2. Asymmetric Logit is a good choice in the parametric approach (where we ignore the approximation error).
3. For LASSO, we show that $O(p/n)$ improves to $O(s \log p/n)$.

Asymmetric shallow and deep learning

$$\inf_{f \in \mathscr{F}_n} \frac{1}{n} \sum_{i=1}^{n} \omega_i (1 - Y_i f(X_i))_+,$$

where $\mathscr{F}_n$ is a neural network class of increasing complexity as $n \uparrow \infty$.

- single layer neural network with W_n neurons (width) and activation function σ_0

$$\theta_n(x) = \sum_{j=1}^{W_n} b_j \sigma_0(a_j^\top x + a_{0,j}) + b_0$$

- shallow learning architecture feeds θ_n in two neurons with rectified linear unit (ReLU) activation function σ

$$\mathscr{F}_n^{\mathrm{SL}} = \{x \mapsto \sigma(\theta_n(x) + c(x)d + 1) - \sigma(\theta_n(x) + c(x)d - 1)\}.$$

Neural network architecture

- Key design: the asymmetry function $c(x)$ enters as a skip connection, bypassing hidden layers and feeding directly into 2 outer ReLU neurons.

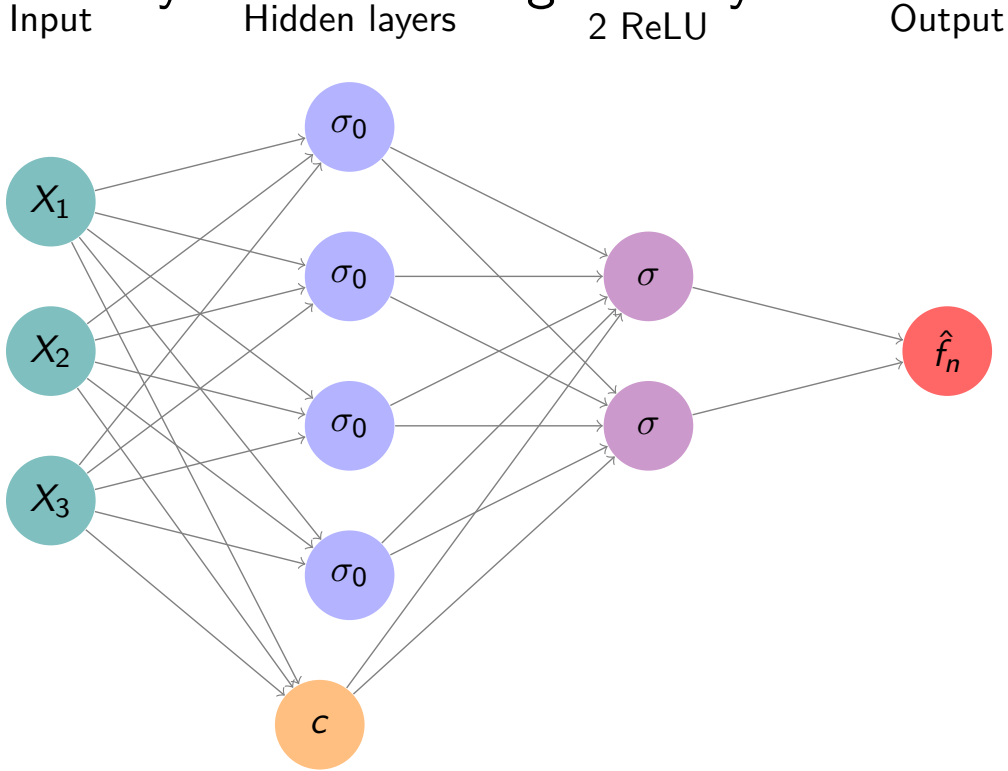


Figure: The orange neuron computes $c(X)$ from the loss function and feeds it as a skip connection into 2 ReLU neurons. Deep learning extends this with multiple hidden layers.

Deep learning architecture

- Deep learning: more flexibility when both width $W_n \to \infty$ and depth $L_n \to \infty$ are tuned.

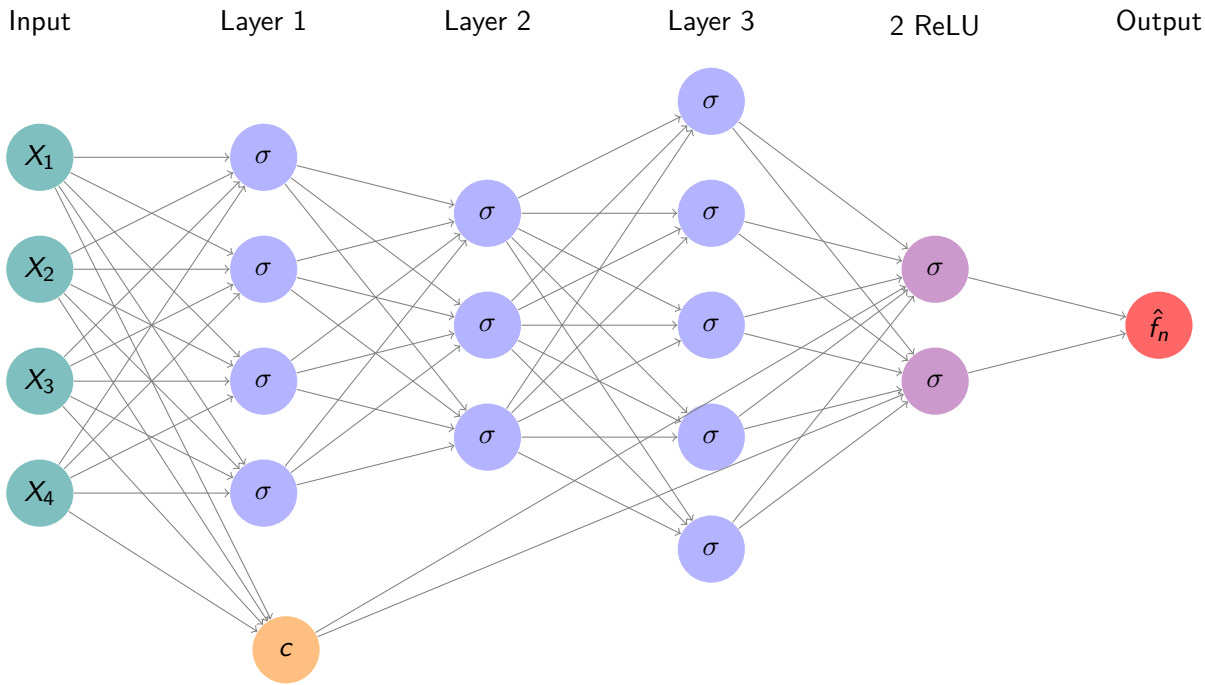


Figure: Deep learning architecture: $d = 4$ covariates, $L = 3$ hidden layers, 2 outer ReLU neurons with $c(X)$ skip connection.

Theory for shallow & deep learning

Theorem

Under appropriate regularity conditions when $W_n \to \infty$ (shallow) or $W_n L_n \to \infty$ (deep) for hinge convexification

$$\sup_{P \in \mathcal{P}(\alpha,\beta)} \mathbb{E}_P \left[\mathcal{R}(\operatorname{sign}(\hat{f}_n)) - \mathcal{R}^* \right] \lesssim \left(\frac{\log^k n}{n} \right)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}},$$

where α is the margin parameter and β is the Hölder smoothness of $\eta(x) = \Pr(Y = 1|X = x)$.

Comments:

1. $k = 2$ for shallow learning and $k = 6$ for deep learning.
2. The rate can approach $O(n^{-1})$ for large α.
3. Matches the minimax lower bound in the symmetric case apart from the $\log^k n$ factor; see Audibert and Tsybakov, 2007 (AoS).

Roadmap

1. Binary decisions and loss functions
   - Examples of loss functions
   - Optimal decision
   - Empirical risk minimization
2. Convexification
   - Convexified empirical risk minimization (ERM)
   - Optimal decision
   - Examples: asymmetric logit and ML
3. Excess risk bounds for asymmetric ML
   - Convexification result
   - Linear decision rules
   - Shallow and Deep learning
4. Monte Carlo experiments
5. Racial bias and recidivism revised

Simulation design

- Stylized example of a social planner with a protected group.
- Probit model

$$Y = \text{sign}\left(2G + Z^\top \gamma + \tau\left(\frac{1}{d}\sum_{j=1}^{d} Z_j^2 + 2Z_1 \sum_{j=2}^{d} Z_j\right) - \varepsilon\right),$$

- $G \sim \text{Bernoulli}(\rho)$ and $\varepsilon, Z_1, \ldots, Z_d \sim_{i.i.d.} N(0,1)$.
- $\rho = 0.2$ fraction in group $G = 1$, $d = 15$ covariates.
- τ controls non-linearities: quadratic and interaction terms.
- Loss function: classification with two groups $G \in \{0, 1\}$

$$\ell(f(X), Y, G) = \varphi_G \mathbb{1}\{f(X) = 1, Y = -1\} + \psi_G \mathbb{1}\{f(X) = -1, Y = 1\}.$$

Benchmark: symmetric case, $\psi_G = \varphi_G = 1$

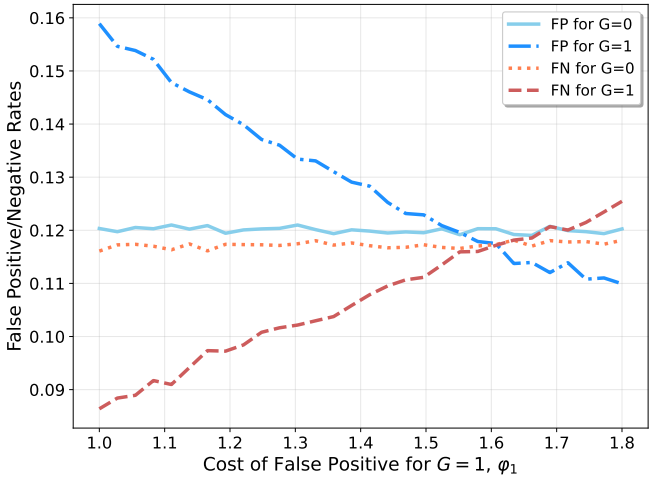
Table: MC simulations: $n = 1{,}000$, $\rho = 0.2$. Symmetric classification.

		Nonlinear DGP ($\tau = 1$)			Linear DGP ($\tau = 0$)		
	G	FP	FN	Error	FP	FN	Error
Logit	0	0.51	0.24	0.35	0.12	0.12	**0.12**
	1	0.69	0.13		0.17	0.08	
Boosting	0	0.30	0.16	**0.22**	0.16	0.13	0.14
	1	0.32	0.15		0.14	0.15	
Shallow Learning	0	0.73	0.10	0.36	0.17	0.12	0.14
	1	0.77	0.07		0.19	0.10	
Deep Learning	0	0.42	0.18	0.27	0.19	0.15	0.17
	1	0.45	0.15		0.20	0.13	

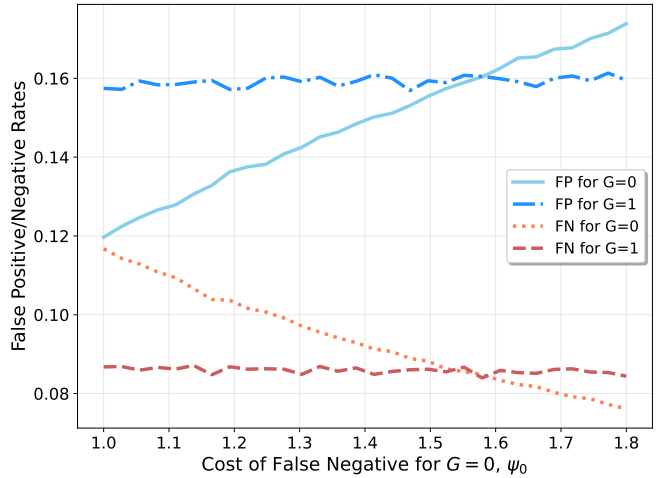
FP/FN = false positive/negative probability, Error = misclassification rate.

Algorithmic Affirmative Action

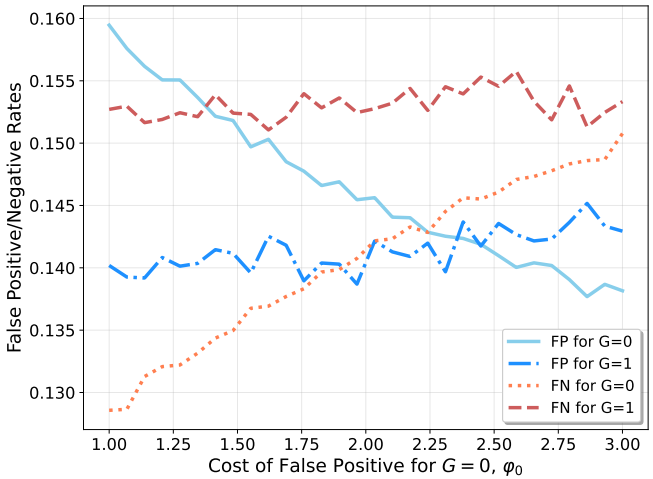
Equalize FP rates across groups by adjusting φ_1 (FP cost for group $G = 1$).



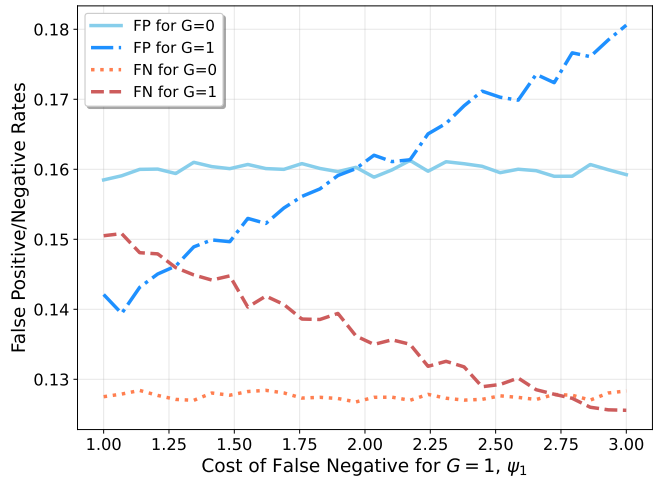
(a) Logit: FP



(b) Logit: FN



(c) Boosting: FP



(d) Boosting: FN

Roadmap

1. Binary decisions and loss functions
   - Examples of loss functions
   - Optimal decision
   - Empirical risk minimization
2. Convexification
   - Convexified empirical risk minimization (ERM)
   - Optimal decision
   - Examples: asymmetric logit and ML
3. Excess risk bounds for asymmetric ML
   - Convexification result
   - Linear decision rules
   - Shallow and Deep learning
4. Monte Carlo experiments
5. Racial bias and recidivism revised

The ProPublica finding

Key fact

African American defendants have **2× higher false positive rates** than other defendants in the COMPAS algorithm — *even without using race as a variable.*

1. Over 10 million people arrested each year in the US (FBI, 2018).
2. Bail decisions made by judges based on the risk of recidivism.
3. Ideal setting for non-causal ML predictions; see Kleinberg et al. (QJE, 2018).
4. Loss function: different losses for keeping a non-recidivist in jail and releasing a recidivist

$$\ell(f(x), y, g) = \underbrace{\varphi_g \mathbb{1}\{f(x) = 1, y = -1\}}_{\text{false positive}} + \underbrace{\psi_g \mathbb{1}\{f(x) = -1, y = 1\}}_{\text{false negative}}.$$

Our fix: adjust loss weights

Approach

Adjust the false-positive weight φ_1 for African Americans. Changes the cost the algorithm assigns to misclassifying this group.

- The social planner explicitly values outcomes of protected groups through the loss function.
- Transparent and interpretable: the weights encode the policy choice.
- Chouldechova (2017): can't equalize all fairness criteria simultaneously when base rates differ. Our framework lets the planner choose which fairness criterion to prioritize via ℓ.

Data

- Dataset compiled by ProPublica to analyse COMPAS algorithm.
- $8,227$ criminal defendants in Broward County, FL (Miami).
- 789 variables corresponding to different types of crime and demographics.
- Gender: 80% male and 20% female.
- Race: 50% African Americans, 34% Caucasian, 16% other.
- Rich crime history covering the degree of crime/misdemeanor and exact chapters from Florida statute.
- Assigned COMPAS scores within 30 days of arrest.
- Actual recidivism rates two years after the release: 33% recidivists and 67% lawful citizens.

LASSO-Logit: standard classification

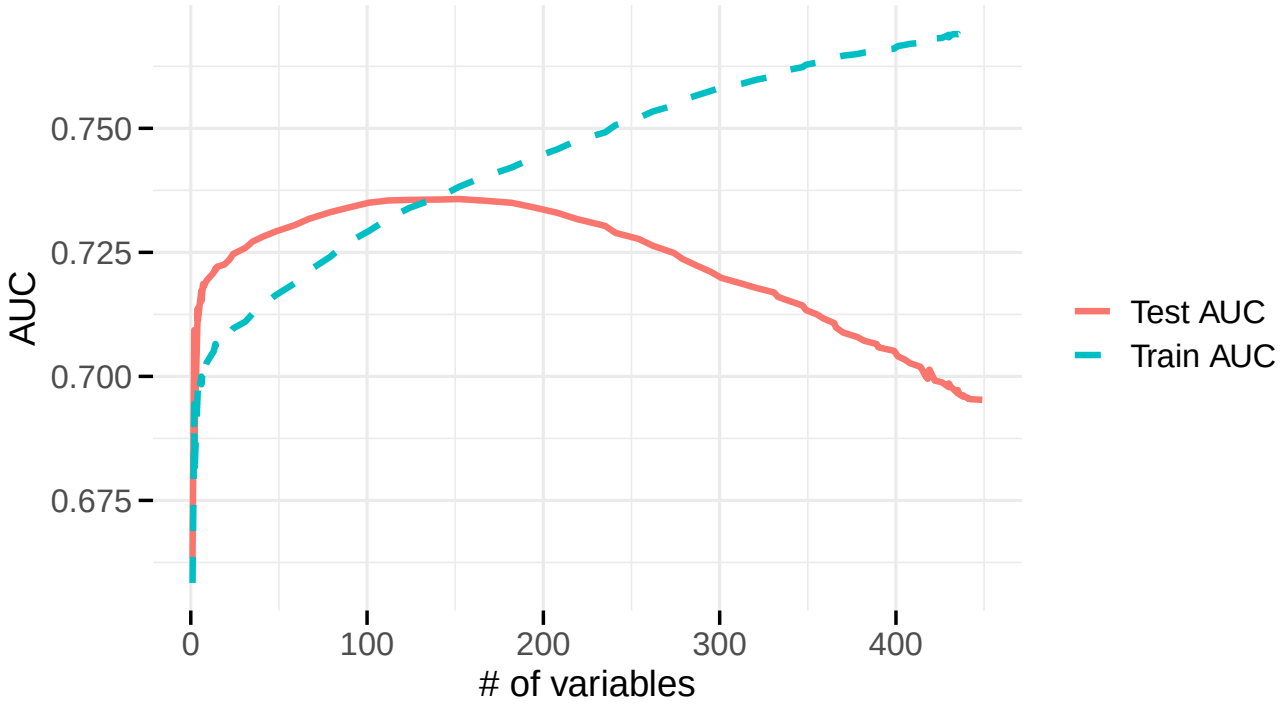


Figure: Training and Test AUC of LASSO-Logit path. The figure shows that the highest test AUC is achieved for the model with 152 covariates.

What predicts recidivism?

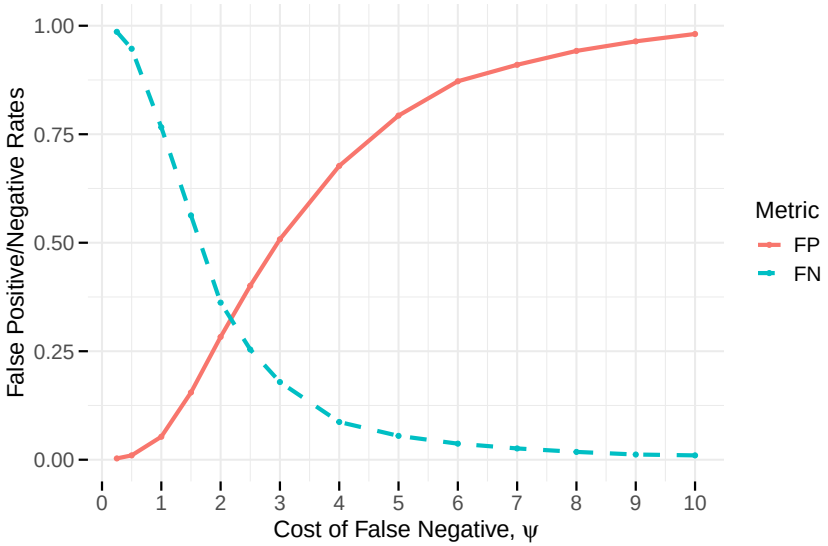
Increase odds of recidivism:

- Obstructing justice (current)
- Theft/robbery (current & past)
- Assault/battery (past)
- African American indicator
- Multiple FL statute charges

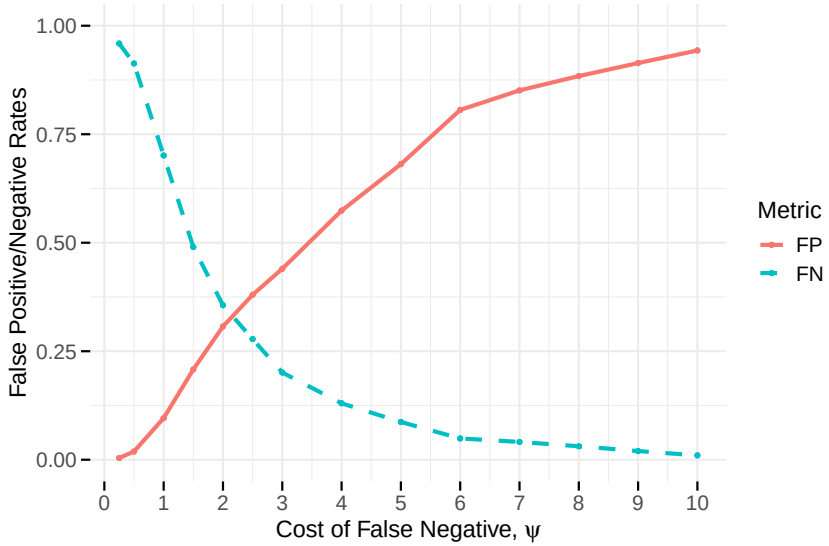
Decrease odds of recidivism:

- Age (-0.56)
- Female gender (-0.38)
- Motor vehicle offenses (current)
- Driving w/o license (past)
- Traffic misdemeanors (past)
- Environmental control charges

Preference-based approach vs. classification ($\psi = 1$)



(a) LASSO Logit: FP and FN rates



(b) Boosting: FP and FN rates

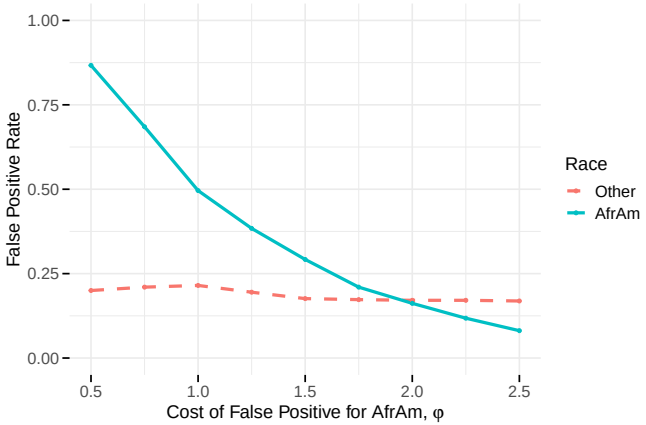
Fairness metrics

1. Balanced Error Rates: achieving similar FP and FN rates across groups
   - False Positive Rate: $\text{FPR}_g = \Pr(\hat{Y} = 1 | Y = -1, G = g)$
   - Fairness: $\text{FPR}_{g=0} \approx \text{FPR}_{g=1}$
   - Among individuals who do **not** reoffend ($Y = 0$), the probability of being flagged high risk is the same across group.
   - Used by ProPublica journalists to criticize COMPAS.

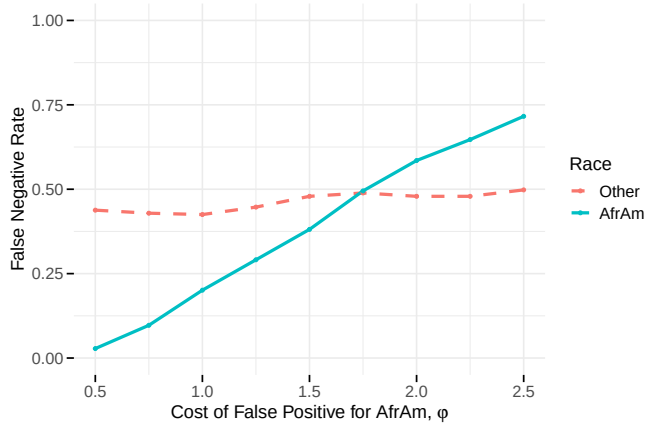
2. Balanced Positive Predicted Values (PPV): achieving similar precision across groups
   - Positive Predicted Value: $\text{PPV}_g = \Pr(Y = 1 | \hat{Y} = 1, G = g)$
   - Fairness: $\text{PPV}_{g=0} \approx \text{PPV}_{g=1}$
   - A high recidivism prediction should have the same meaning across groups.

3. We adjust the loss function weights φ_g and ψ_g to achieve these fairness criteria.

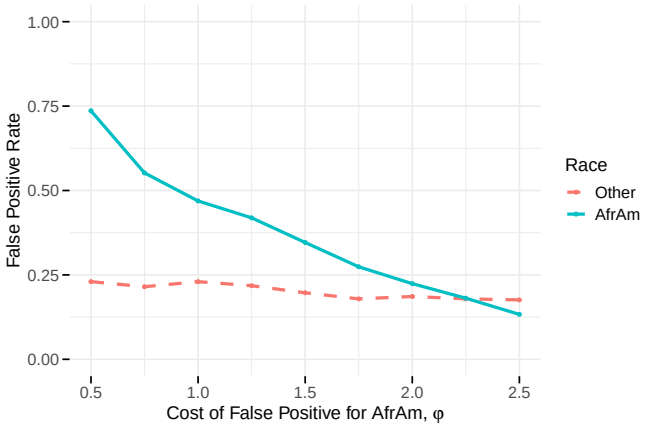
Fairness: Balanced Error Rates



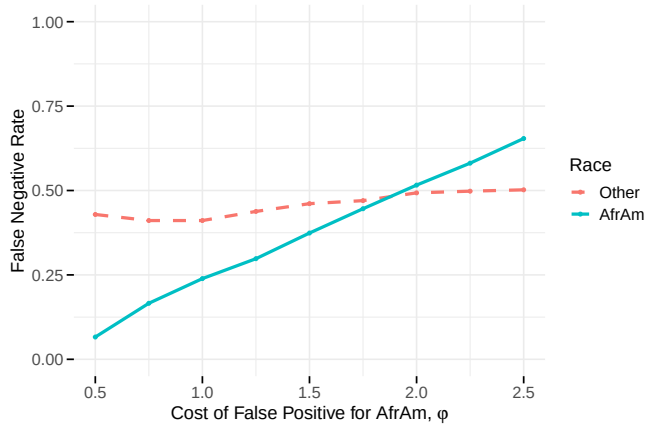
(a) LASSO-Logit: FP rates



(b) LASSO-Logit: FN rates

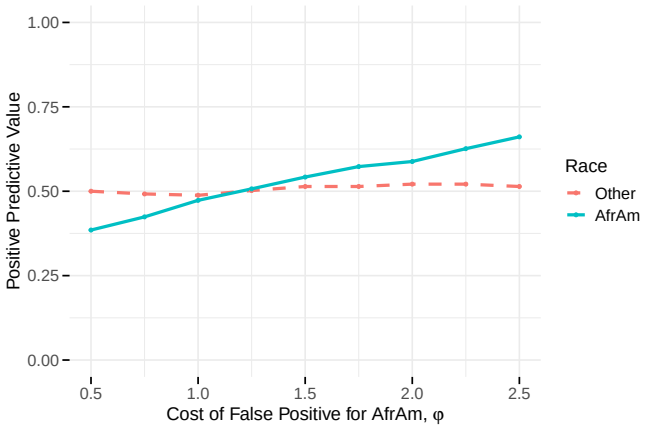


(c) Boosting: FP rates

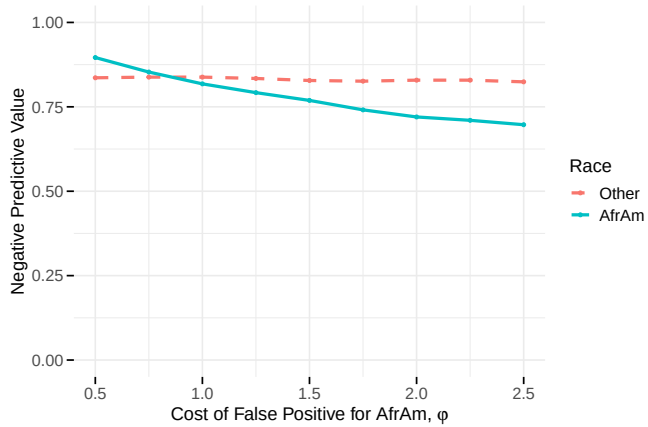


(d) Boosting: FN rates

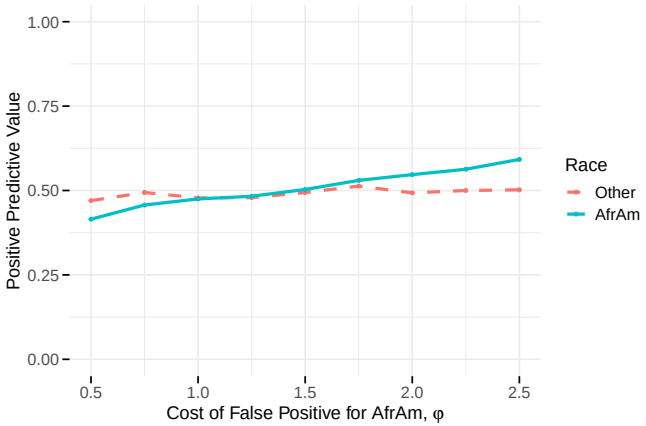
Fairness: Balanced Positive Predicted Values



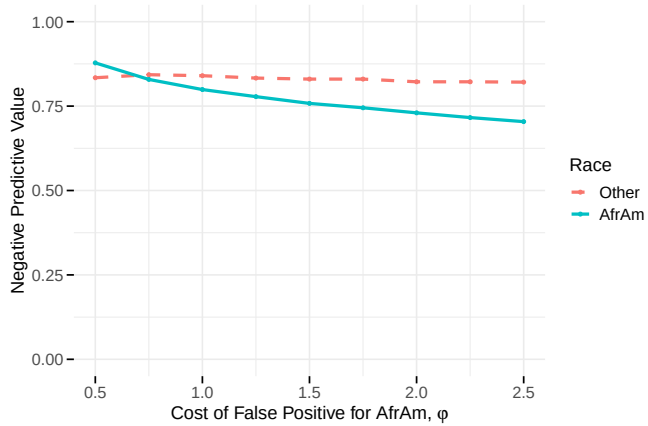
(a) LASSO-Logit: PPV rates



(b) LASSO-Logit: NPV rates



(c) Boosting: PPV rates



(d) Boosting: NPV rates

Concluding remarks: a practical recipe

How to use this in practice

1. **Specify** your loss function $\ell(f, y, x)$ — encode asymmetries and group preferences.
2. **Compute** weights ω_i from ℓ.
3. **Run** weighted logistic regression, LASSO, boosting, or deep learning.
4. The resulting $\text{sign}(\hat{f}_n)$ is a valid, optimal binary decision.

Why this matters:

- Computationally scalable — same cost as standard ML.
- No distributional assumptions on $\Pr(Y = 1|X)$.
- Transparent fairness — policy preferences enter through the loss function, not ad-hoc constraints.
- Biased algorithms could be easier to fix than biased people.

Thank you!

Backup: convexifying function assumption (CFA)

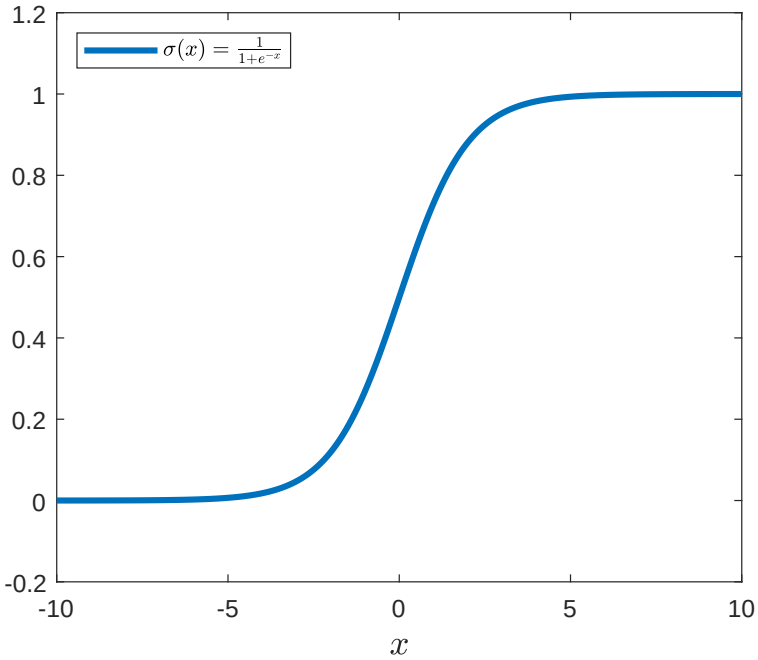
1. $\phi : [0, \infty) \to \mathbf{R}$ is a convex, non-decreasing, and Lipschitz continuous function with $\phi(0) = 1$
2. There exists $C > 0$ and $\gamma \in (0, 1]$ such that

$$|x - c| \leq C \left(x + c - 2xc - \inf_{y \in \mathbf{R}} Q_c(x, y) \right)^{\gamma}, \qquad \forall x, c \in (0, 1),$$

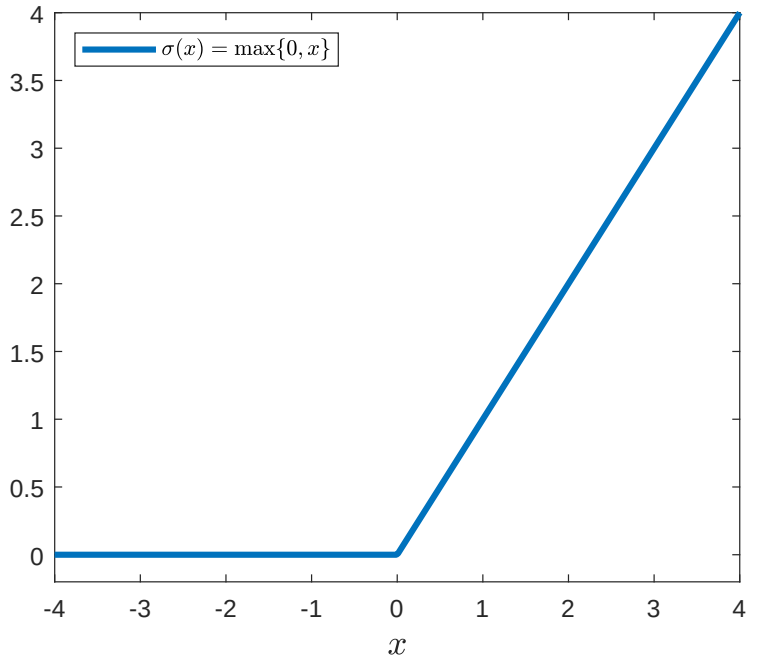
where $Q_c(x, y) := x(1 - c)\phi(-y) + (1 - x)c\phi(y)$.

Satisfied by logistic ($\gamma = 1$), exponential ($\gamma = 1$), and hinge ($\gamma = 1/2$).

Backup: activation functions



(a) Sigmoid, σ_0



(b) ReLU, σ

Figure: Examples of activation functions