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

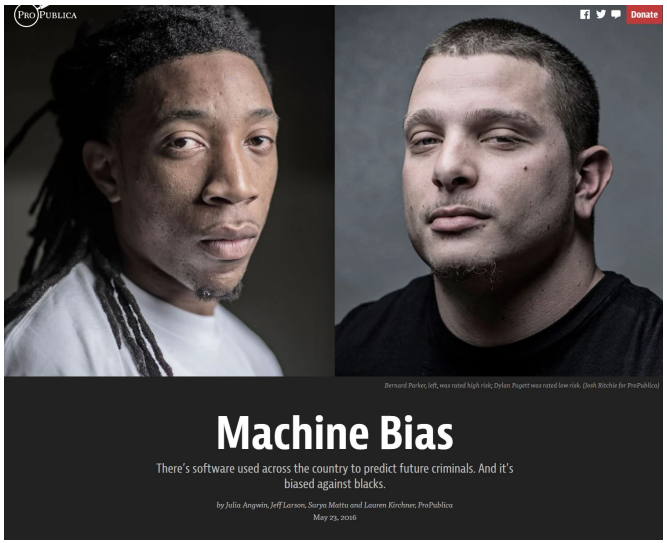
Andrii Babii [1] Xi Chen [1] Eric Ghysels [1] Rohit Kumar [2]

[1]UNC Chapel Hill

[2]ISI, Delhi

2020 IAAE Webinar

Motivation: algorithmic discrimination



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

Motivation

1. The potential for ML algorithms to reproduce and reinforce existing discrimination in the society has been of great concern lately.

2. The upside gains and the downside risk of many economic decisions are not symmetric; see, Granger, 1969, (OR), Granger and Pesaran, 1999 (JoF), and the recent book of Elliot and Timmerman, 2016.

3. How do we adjust ML classification algorithms for asymmetric losses/preferences?
   - controlling false positive/negative mistakes across groups;
   - economic cost/benefit considerations.

4. We focus on the binary choice/decision problems in a data-rich environment with a generic loss functions.

Our contribution: asymmetric Logit and ML

1. Methodological contribution:
   - introduce computationally attractive asymmetric logistic regression and asymmetric ML with a generic loss function.

2. Theoretical contributions:
   - non-asymptotic excess risk bounds for binary decisions produced by asymmetric convexified empirical risk minimization;
   - carefully crafted shallow and deep neural architectures are minimax optimal.

3. Application to bail decisions: racial bias and recidivism revised.

Related economic literature

1. Asymmetric regression: quantile regression, Koenker and Bassett, 1978 (Ecma); asymmetric least-squares, Newey and Powell, 1987 (Ecma), and M-estimators based on a convex loss function.

2. Nonparametric binary choice: Manski, 1975 (JoE), Elliot and Lielli, 2013, (JoE) – NP-hard optimization problems.

3. Neural network regressions:
   - Shallow learning: Chen, 2007 (Handbook of Econometrics, vol. 6B);
   - Deep learning: Farrell, Liang, Misra, 2020 (Ecma, forthcoming).
4. Bail decisions: humans vs. machines, Kleinberg et al., 2018 (QJE).

Roadmap

1. Binary decisions and loss functions
   - Example of loss function
   - Optimal decision
2. Convexification
   - Convexified empirical risk minimization (ERM)
   - Optimal decision
   - Examples: asymmetric logit and ML
3. Excess risk bounds for asymmetric ML
   - Linear decision rules
   - Shallow and Deep learning
4. 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 $f(X)$

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

Example: 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 non-recidivist in jail and releasing a recidivist

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

Choice of the loss function

The decision maker has to specify preferences through the loss function: $\ell(f, y, x) = \ell_{f,y}(x)$

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); Vapnik, 1974

1. Equivalent characterization of the optimal decision f^*

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

   where a and b are computed from the loss function ℓ.

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

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

Roadmap

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

Convexified ERM

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

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

   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

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

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

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

and

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

where $\eta(x) = \Pr(Y = 1|X = x)$ *and* $c(x) = \frac{a(x)+b(x)}{2b(x)}$.

Comments:

1. link between the optimal prediction f^* and the solution to the convexified risk minimization problem f_ϕ^*.
2. in the symmetric binary classification case $a(x) = 0$ and $c(x) = 1/2$.

Examples of convexifying functions

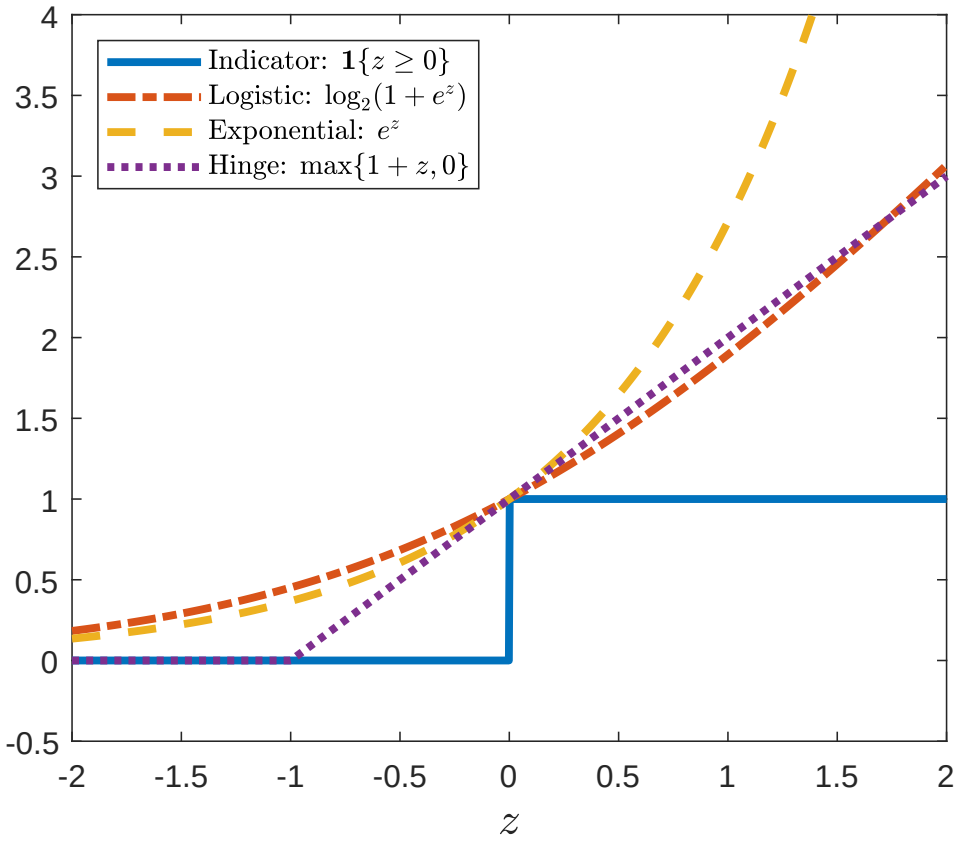


Figure: Convexifications corresponding to Logit/XGBoost and SVM/Deep learning.

Example 1: Asymmetric Logit/XGBoost

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

1. Log-likelihood of Logit

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

2. Our methods: asymmetric Logit/XGBoost

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

Comments:

- logistic regression reweighted for the asymmetry of ℓ through a and b;
- asymmetric extreme gradient boosting (XGBoost).

Example 2: 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

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

Comment: hinge convexification is also a popular choice for symmetric shallow and deep learning problems.

Roadmap

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

Asymmetric linear decision rules

1. Convexified ERM

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

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

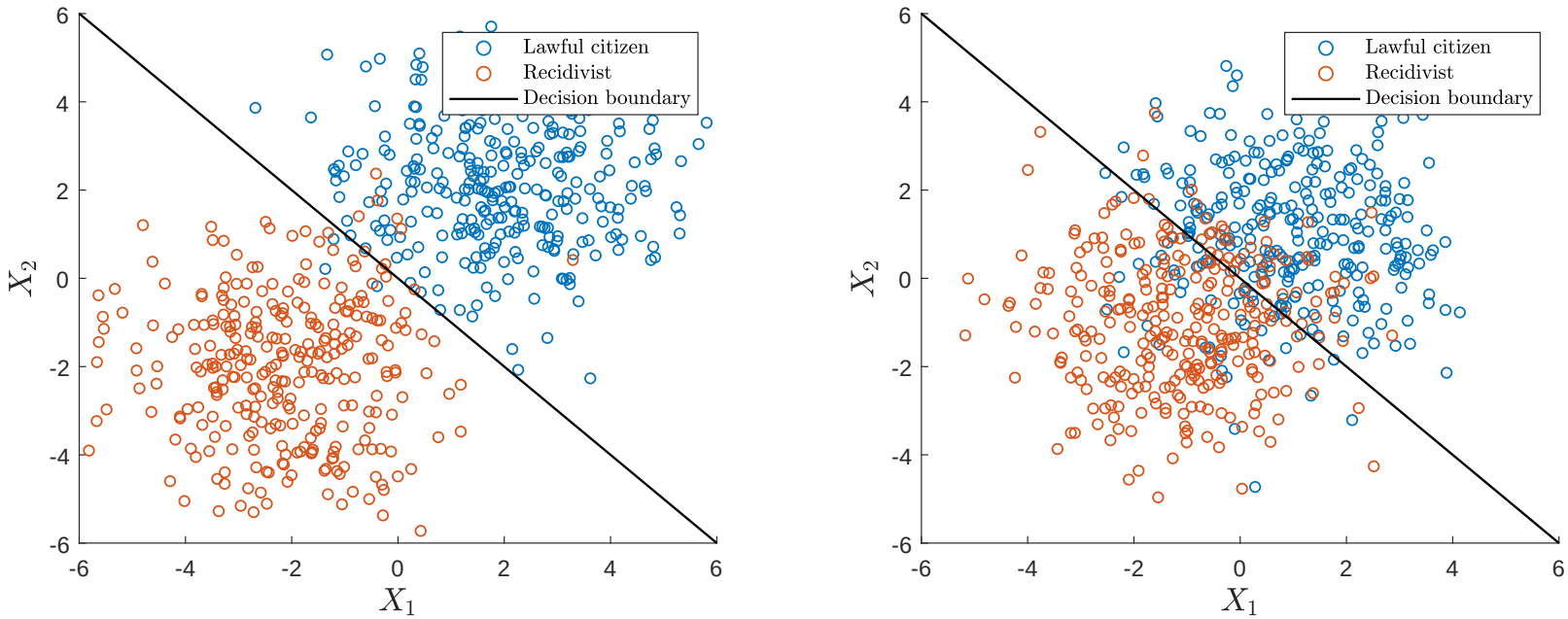
2. Example: logistic convexifying function,

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

Margin/noise assumption

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

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



(a) Easy (b) Difficult

Figure: Restricts how tightly the data are concentrated around the decision boundary, $\{x \in \mathbf{R}^2 : \ \Pr(Y = 1|X = x) = c(x)\}$ (black line)

Asymmetric parametric predictions

Theorem

For logistic and hinge convexification

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

Comments:

1. for a fixed number of covariates p, the first 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. Can add ℓ_1 or ℓ_2 regularization to handle $p \uparrow \infty$.

Asymmetric shallow and deep learning

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

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

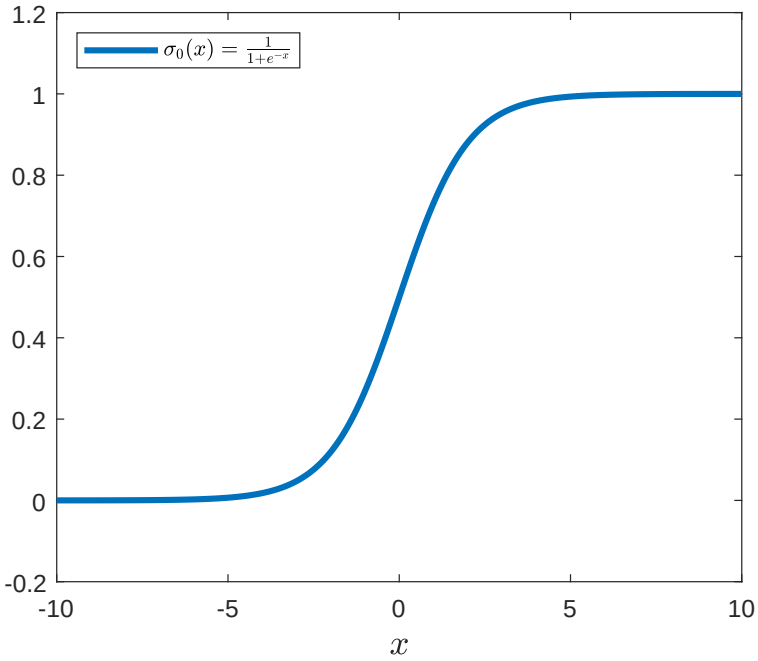
- single layer neural network with W_n neurons (width) and activation function σ_0

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

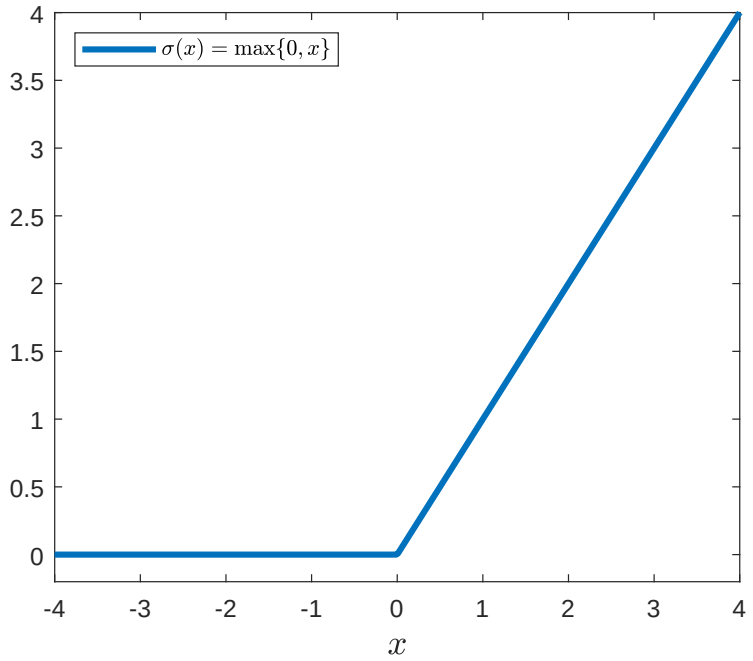
- shallow learning architecture feeds θ_n in two neurons with rectified linear unit (ReLU) activation function σ

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

Activation functions



(a) Sigmoid, σ_0



(b) ReLU, σ

Figure: Examples of activation functions

Shallow learning

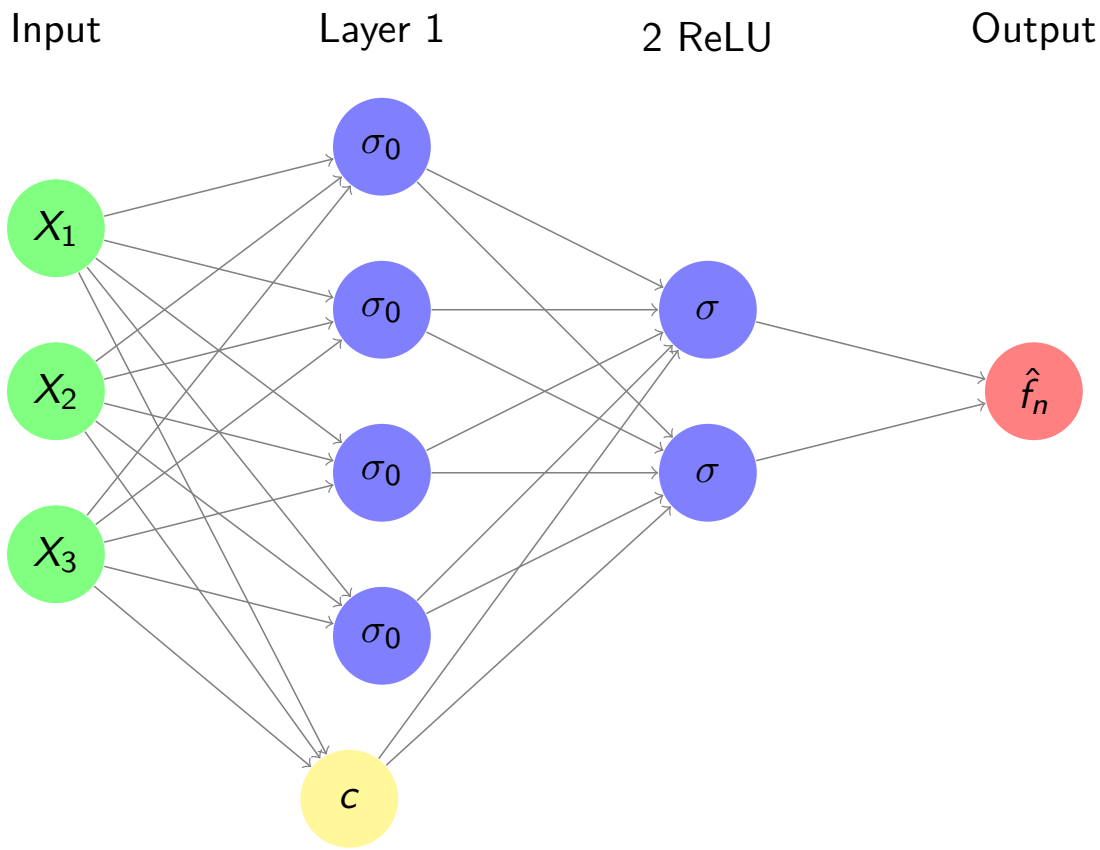


Figure: shallow learning architecture with $d = 3$ covariates, single hidden layer of width $W_n = 4$ sigmoid neurons, and 2 outer ReLU neurons. The yellow neuron takes covariates $X \in \mathbf{R}^d$ as an input and produces $c(X) \in \mathbf{R}$, which is fed directly in 2 ReLU neurons.

Deep learning

1. A single-layer neural networks can approximate any continuous function when the width increases, $W_n \to \infty$; see Gallant and White, 1988 (IEEE) and Hornik, Stinchombe, and White, 1989 (Neural Networks).

2. Alternatively, we can fix the width and increase the depth, $L_n \to \infty$; see Yarotsky, 2018 (PMLR) and Lu, Shen, Yang, Zhang, 2020 (arXiv).

3. Increasing both the width $W_n \to \infty$ and the depth $L_n \to \infty$ gives more flexibility.

4. Recent results suggest that deep learning may adapt efficiently to intrinsic dimensionality and to anisotropic smoothness offsetting the curse of dimensionality; see Suzuki and Nitanda, 2019 (arXiv).

Deep learning

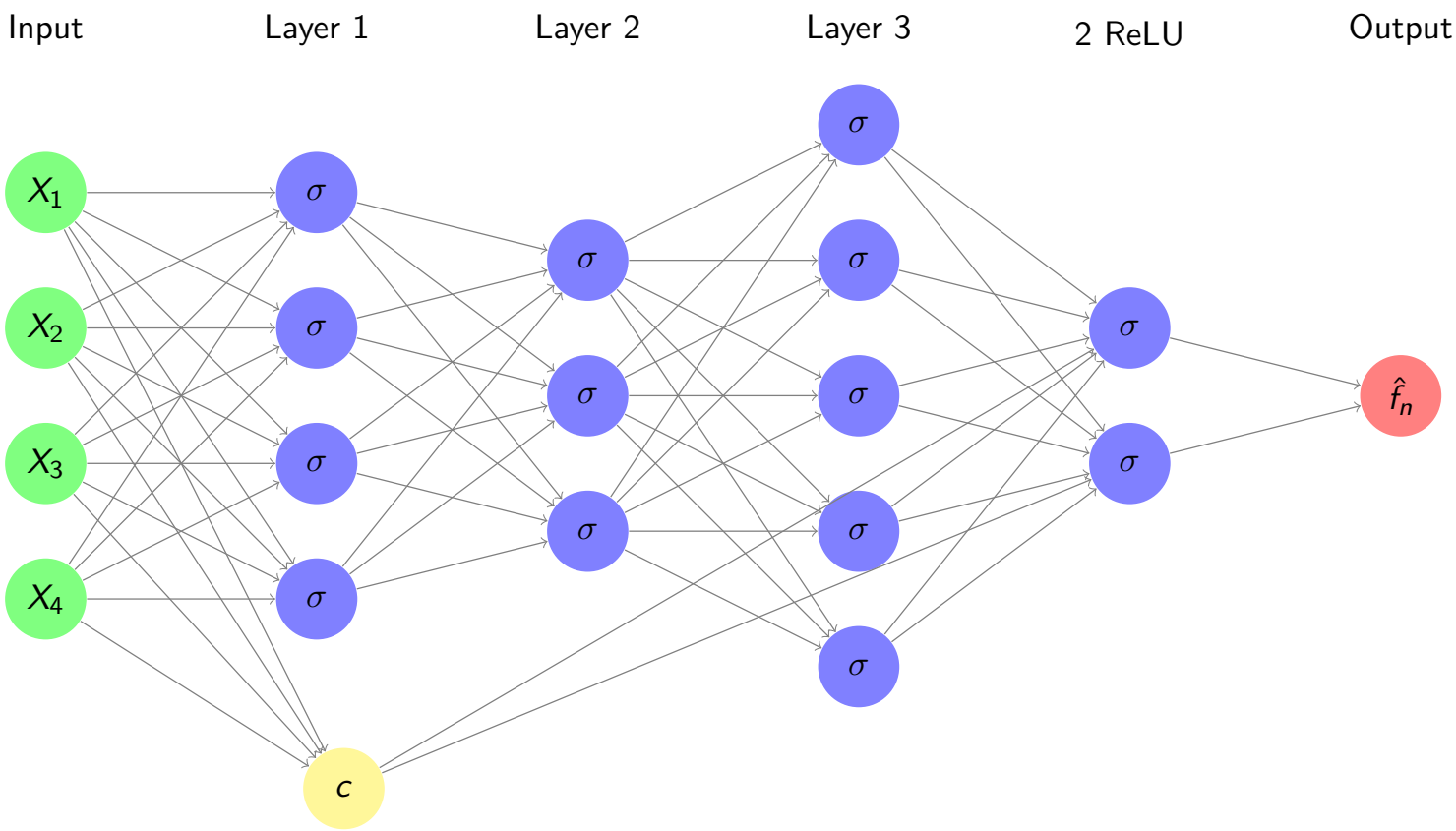


Figure: Deep learning architechture with $d = 4$ covariates, $L = 3$ hidden layers with $(4, 3, 5)$ ReLU neurons, and 2 outer ReLU neurons. The yellow neuron takes covariates $X \in \mathbf{R}^d$ as an input and produces $c(X) \in \mathbf{R}$, which is fed directly in 2 ReLU neurons.

Theory for shallow & deep learning

Theorem

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

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

where α is the margin parameter and β is the Sobolev 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 for the $\log^k n$ factor; see Audibert and Tsybakov, 2007 (AoS).

Roadmap

1. Binary decisions and loss functions
   - Example of loss function
   - Optimal decision
2. Convexification
   - Convexified empirical risk minimization (ERM)
   - Optimal decision
   - Examples: asymmetric logit and ML
3. Excess risk bounds for asymmetric ML
   - Linear decision rules
   - Shallow and Deep learning
4. 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 $G \in \{0, 1\}$

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

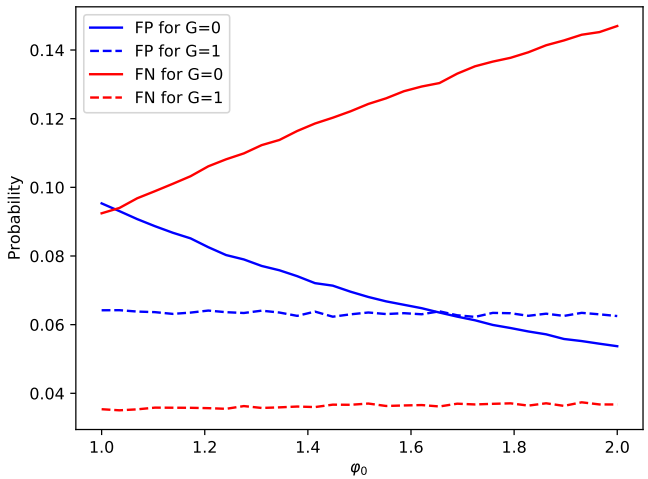
Benchmark: symmetric case, $\psi_G = \varphi_G = 1$

Table: MC simulations: $n = 1,000$ observations

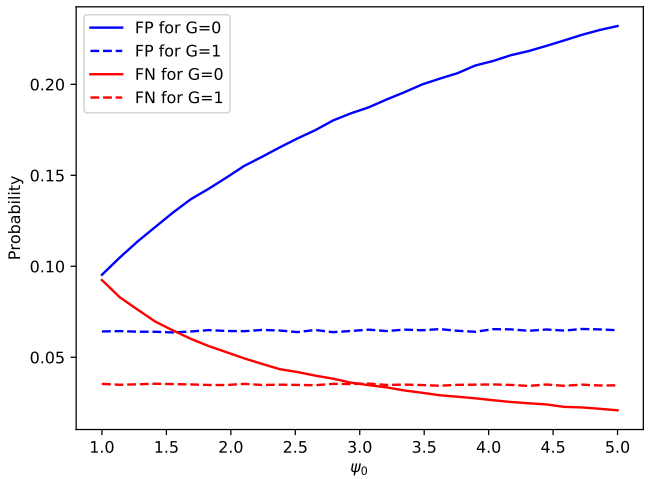
		Nonlinear DGP			Linear DGP		
	G	FP	FN	Error	FP	FN	Error
Logit	0	0.28	0.12	0.37	0.10	0.09	0.17
	1	0.25	0.02		0.06	0.04	
XGBoost	0	0.15	0.09	0.22	0.12	0.10	0.20
	1	0.12	0.06		0.08	0.05	
SVM	0	0.13	0.08	0.20	0.14	0.08	0.20
	1	0.10	0.05		0.04	0.11	
Shallow	0	0.07	0.03	0.09	0.10	0.08	0.17
	1	0.04	0.02		0.07	0.03	
Deep	0	0.06	0.05	0.10	0.10	0.09	0.18
	1	0.04	0.03		0.06	0.05	

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

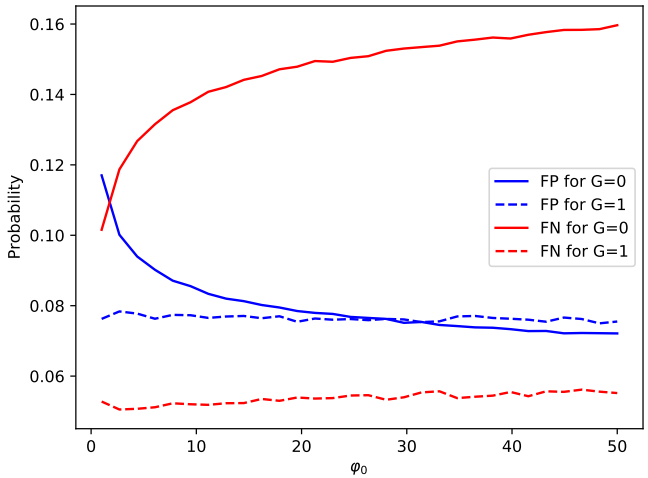
Our method: asymmetric Logit and XGBoost



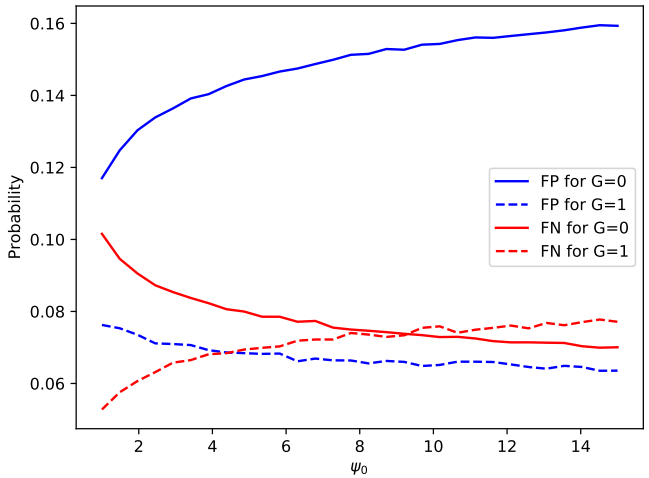
(a) Logit: false positives



(b) Logit: false negatives



(c) XGBoost: false positives



(d) XGBoost: false negatives

Empirical illustration: racial bias and recidivism revised

1. CA spent 3% on prisons and 10% on higher education 30 years ago
   - 11% on prisons and 7.5% on higher education now; see Baughman, 2017
2. Over 10 million people arrested each year in the US (FBI, 2018).
3. Bail decisions made by judges based on the risk of recidivism.
4. Ideal setting for non-causal ML predictions; see Kleinberg et al., 2017 (QJE).
5. Asymmetries: black vs. white, type of misconduct,...
6. Economic costs of crimes and benefits of detention.

Data

- Dataset compiled by ProPublica to analyse COMPAS algorithm.
- $11,181$ criminal defendants in Broward County, FL.
- Gender: 80% male and 20% female.
- Race: 50% African Americans, 34% Caucasian, 16% other.
- Crime history by types: arson, aggravated assault, larceny theft, burglary, vehicle theft, fraud, robbery, rape, murder.
- Assigned COMPAS scores withing 30 days of arrest.
- Actual residivism rates two years after the release: 33% recidivists and 67% lawful citizens.

Economic costs/benefits, Baughman, 2017

	Costs ($)	**Benefits ($)**	
Type of Offense		Min	Max
Murder	10,754,332	4,602,326	18,780,120
Rape/Sexual Assault	266,332	136,191	488,243
Aggravated Assault	126,585	14,715	158,250
Robbery	48,589	12,523	364,898
	Individual Costs		
Loss of Freedom		$(\$1036/90)d_i$	
Loss of Income		$(\$31,028/365)d_i$	
Loss of Housing		$\$1565 m_i$	
Childcare Costs		$(\$1938/365)d_i$	
Violent or Sexual Assault		$(\$136,191(.032)/365)d_i$	
	Public Costs		
Prison Operation Costs		$(\$31,406/365)d_i$	
Loss of Tax		$(\$6391/365)d_i$	
Welfare for Detainee's Family		$(\$8293/365)d_i$	

Preference-based approach vs. binary classification

Our preference-based approach

pred \ true	$Y = 1$	$Y = -1$
$f = 1$	$EBD(c) + ECD(d)$	$\psi_G ECD(d)$
$f = -1$	$\psi_G C(z, c)$	0

1. EBD = economic benefit of detention, ECD = economic cost of detention, C = cost of recidivism.
2. d = duration, c = type of crime, z = other characteristics, and ψ_G preferences towards two groups ($\psi_1 = 2$ and $\psi_0 = 1$).
3. Recall that in the binary classification the loss is

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

Empirical results

	Deep learning		XGBoost	
	Symmetric	Asymmetric	Symmetric	Asymmetric
True Positive Cost	-1677	-2266	-1674	-2332
False Negative Cost	5288	4488	5141	4085
True Negative Cost	0	0	0	0
False Positive Cost	2645	3220	3036	3680
Overall cost	6256	5442	6503	5433
TP Rate	0.27	0.31	0.30	0.30
FP Rate	0.07	0.09	0.09	0.10
AUC	0.65	0.64	0.7	0.68

Comments:

- 13% cost reduction with our preference-based asymmetric deep learning compared to the standard deep learning classification;
- 10% cost reduction with our preference-based asymmetric deep learning compared to the preference-based asymmetric Logit.

Concluding remarks

1. ML classification algorithms are increasingly used for life-changing decisions: bail decisions, hiring, health care...
2. Biased algorithms might be easier to fix and regulate than biased people.
3. We focus on the preference-based ML approach with explicitly stated loss function:
   - economic costs/benefits analysis;
   - preferences towards protected groups.
4. Extremely simple and computationally attractive approach with minimal distributional assumptions.
5. Theory:
   - excess risk bounds for asymmetric Logit and asymmetric ML;
   - asymmetric shallow and deep learning with carefully crafted architectures are minimax optimal.

Thank you!

email: babii.andrii@gmail.com