UNIVERSITY OF AMSTERDAM
Korteweg de Vries Institute for Mathematics

The Risks of Recourse in Binary Classification

2023-10-27

The Risk of Recourse in Binary Classification

- Preprint on ArXiv (2306.00497)
- All work presented was created in collaboration with:

Dr. Tim van Erven

Dr. Damien Garreau

Programme of today

- Introduction to the problem
- Modelling the setting
- Optimal classifiers
- Near Optimal classifiers
- Strategising
- Conclusion

Introduction to the problem

With a peculiar example

Leading example

2 parties:

- Credit Loan Applicant (A)

- Credit Loan Provider (P)

Loan application process:

- (A) provides (P) with a set of features:

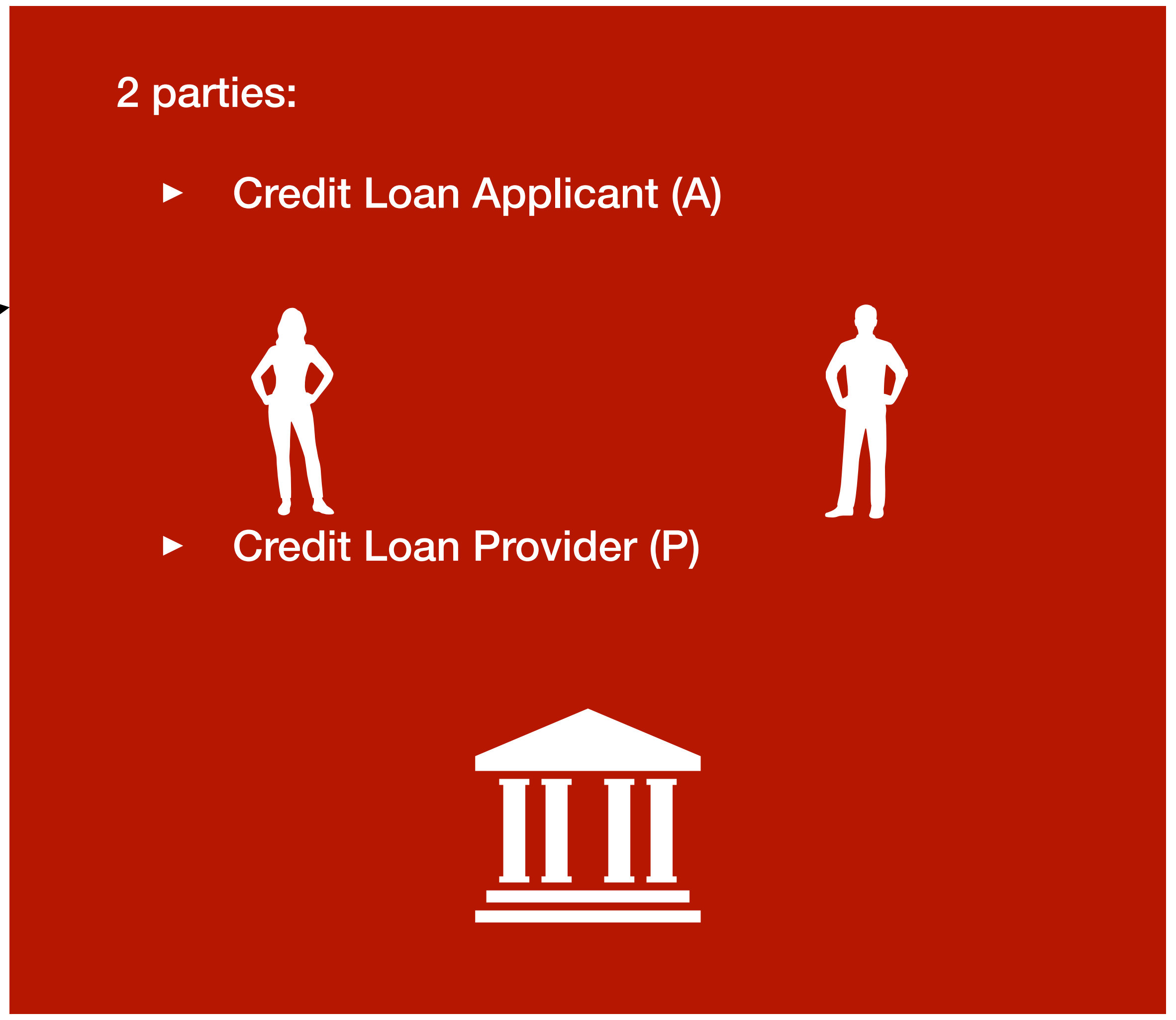
$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

- (P) has an automated decision system f

$$f(x) = 1 \quad \text{if accepted}$$
$$f(x) = -1 \quad \text{if not}$$

- (A) can ask for a counterfactual explanation

Leading example



Loan application process:

- (A) provides (P) with a set of features:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

- (P) has an automated decision system f

$$f(x) = 1 \quad \text{if accepted}$$
$$f(x) = -1 \quad \text{if not}$$

- (A) can ask for a counterfactual explanation

This example us seen as a:

Counterfactual literature

Positive example

Strategic classification

Negative example

Modelling the situation

Model

Learning theoretic setting for classification

$$f: \mathcal{X} \subseteq \mathbb{R}^d \to \{-1,1\}$$

We assume that

$$(X_0, Y) \sim P$$

Loss is measured by counting wrong classifications

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

Goal is to minimize ***expected loss (Risk/Accuracy)***

$$R = \mathbb{E}_P[\ell(f(X_0), Y)] = P(f(X_0) \neq Y).$$

The optimal classifier is the ***Bayes Classifier***

$$f_P^* = \arg\min \mathbb{E}_P[\ell(f(X_0), Y)]$$

Model

Adding recourse

By adding recourse in the mix,

$$X_0 \to X,$$

where X is either X_0 or X^{CF}, we induce a new distribution

$$(X_0, X, Y) \sim Q.$$

Counterfactual point is defined as

$$\varphi(X_0) = X^{\mathrm{CF}} \in \arg\min_{z:\, f(z)=1} c(X_0, z)$$

For simplicity, we assume that every negative X_0 accepts recourse

Risk with **Recourse** is defined as

$$R_Q(f) = \mathbb{E}_Q[\ell(f(X), Y)] = Q(f(X) \neq Y)$$

Note that Q depends on f in general

Model

When is Recourse accepted

The function $r(X_0)$ models how likely X_0 is to accept recourse

In general X_0 does not need to change to X^{CF},

This is modelled by setting

$$X = BX^{\text{CF}} + (1 - B)X_0, \quad B \sim \text{Ber}(r(X_0)).$$

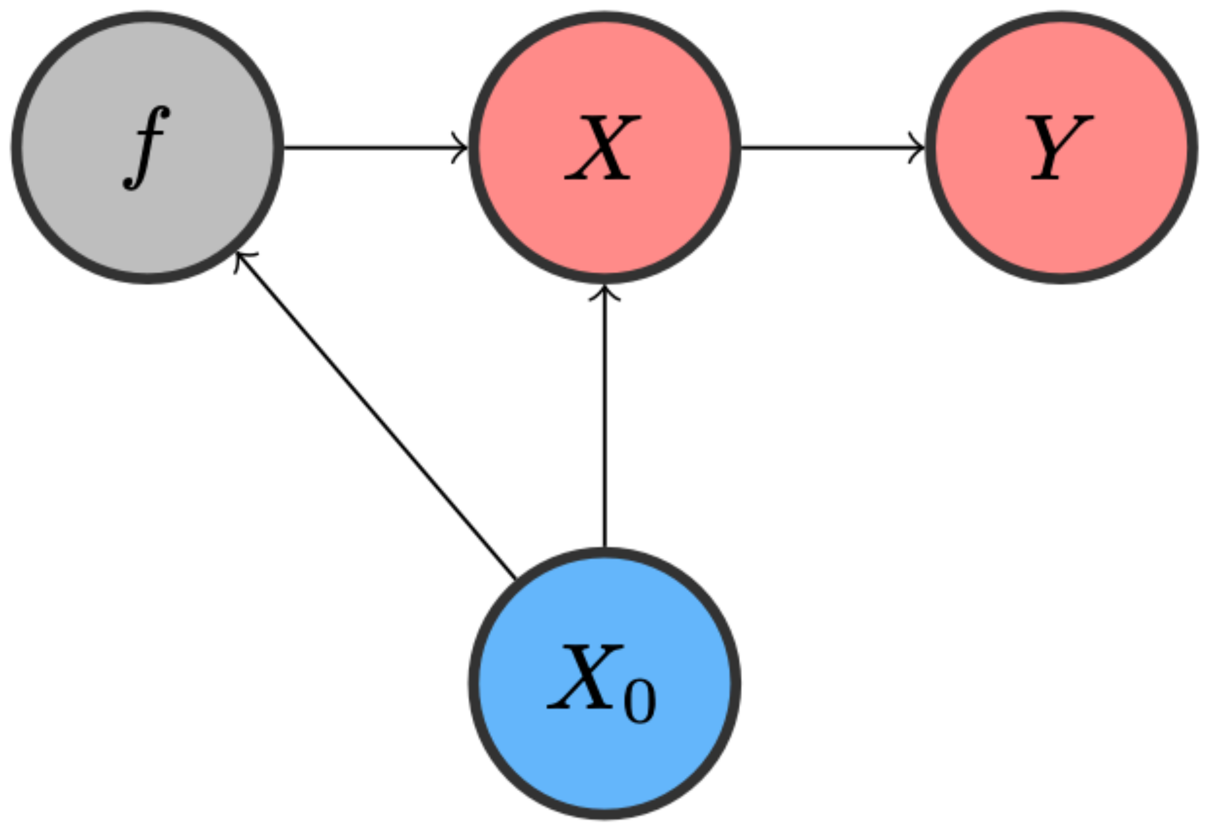
For the rest of the talk we will assume $r(X_0) = 1$

Modelling Q

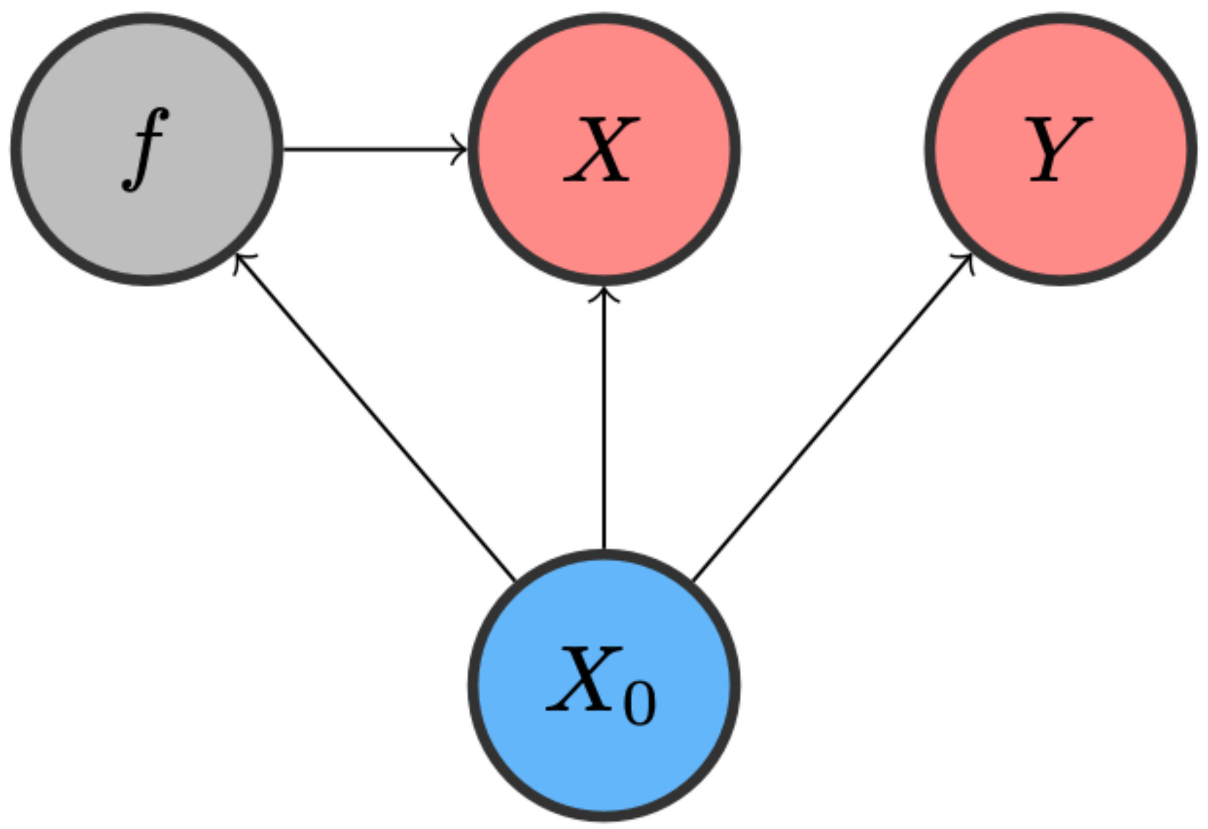
Choice in dependency structure

We define 2 extreme cases:

- **Compliant:** $Q(Y|X_0, X) = P(Y|X)$
- **Defiant:** $Q(Y|X_0, X) = P(Y|X_0)$



Compliant case

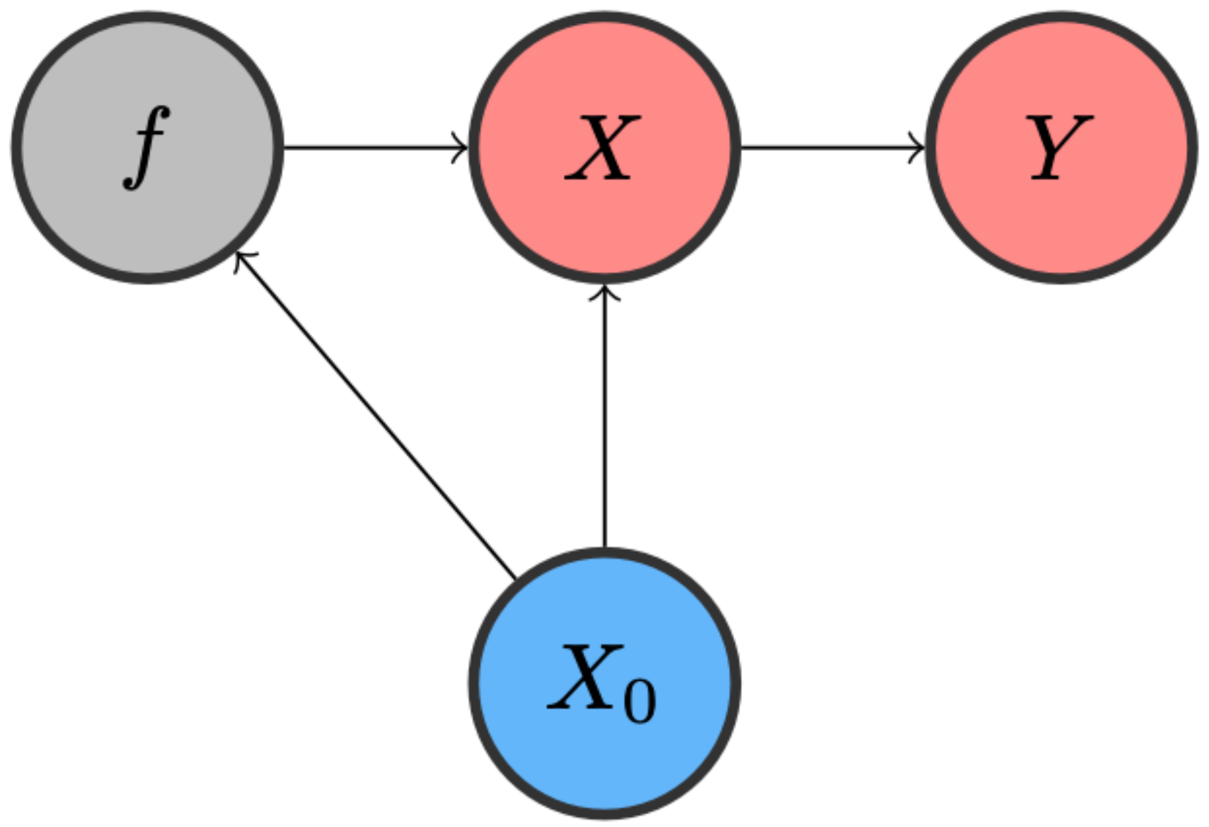


Defiant case

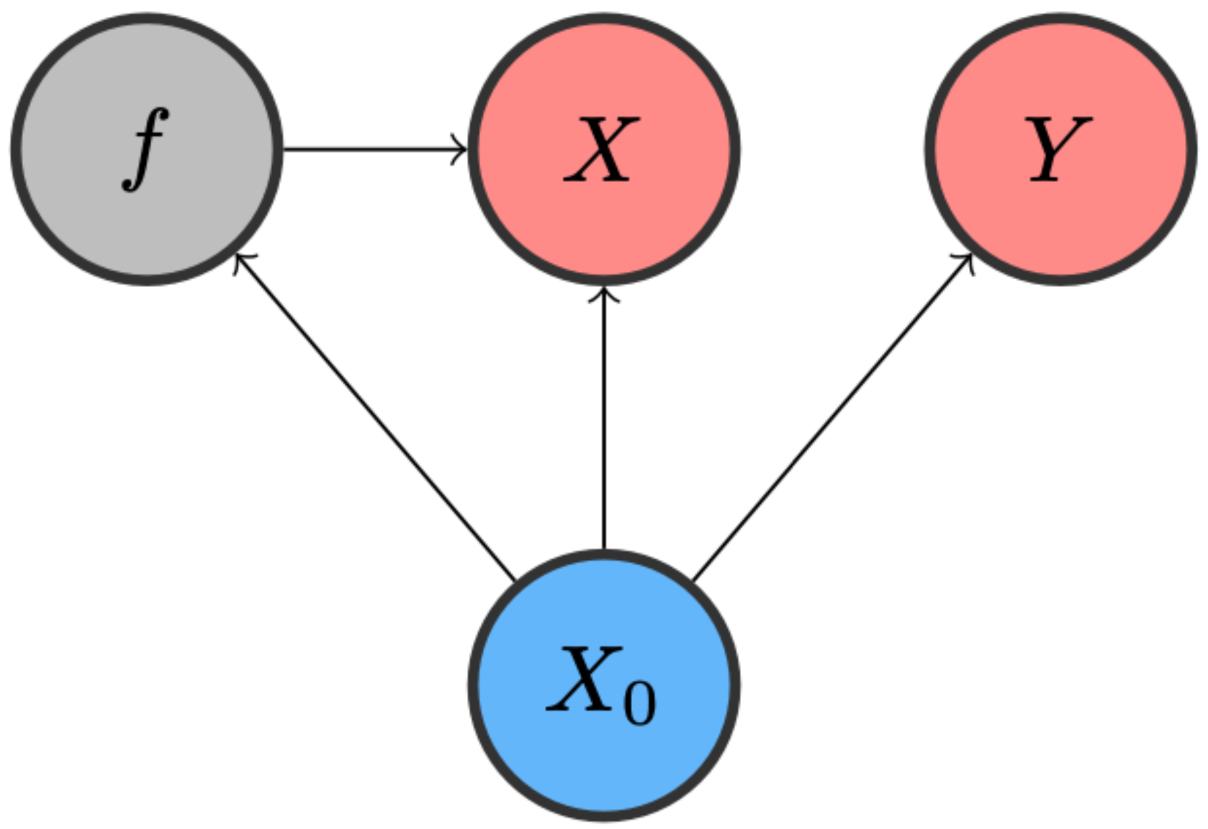
Modelling Q
Examples

Some examples:

- Credit loan application:
  - Compliant: Applicant improves risky behaviour
  - Defiant: Applicant tries to “game the system”
- Medical Diagnosis:
  - Compliant: Patient improves their health
  - Defiant: Patient takes medicine to reduce symptoms
- Job applications:
  - Compliant: Applicant improves their skills
  - Defiant: Applicant improves their CV



Compliant case



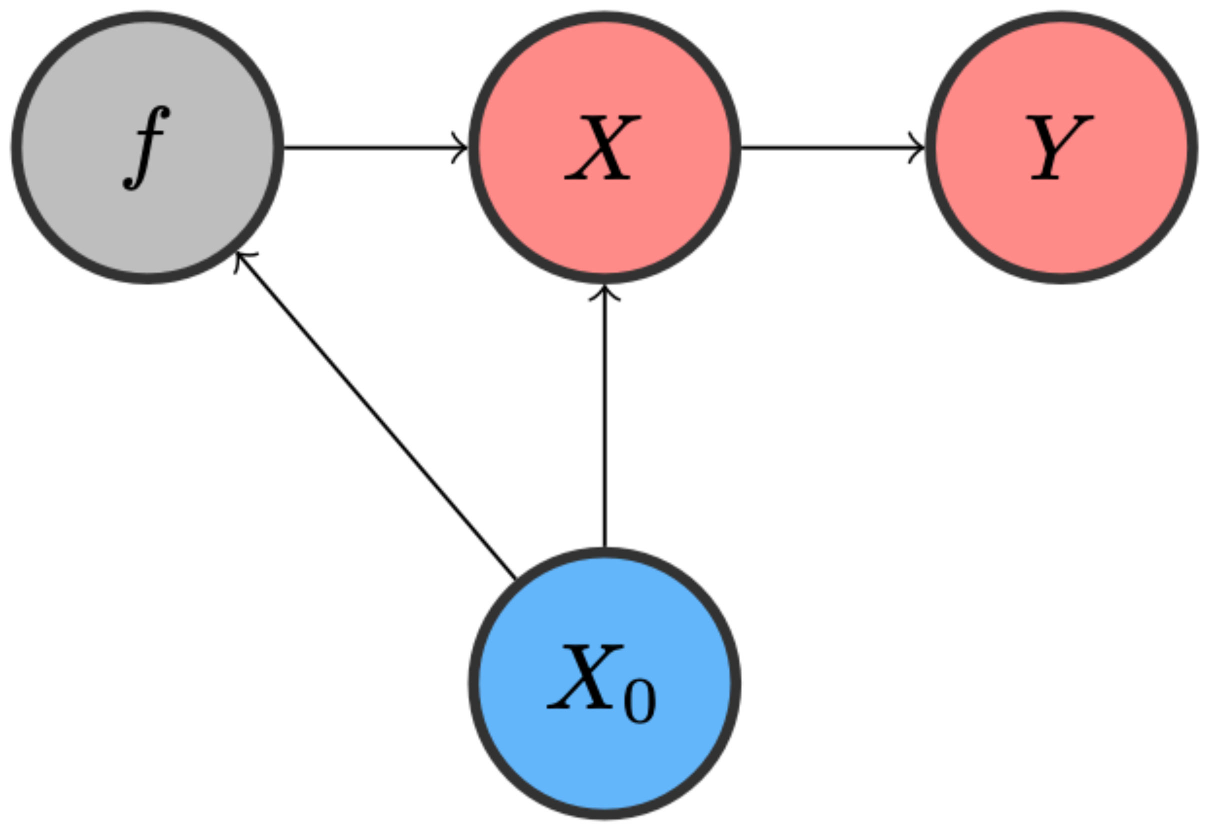
Defiant case

Modelling Q
Causality

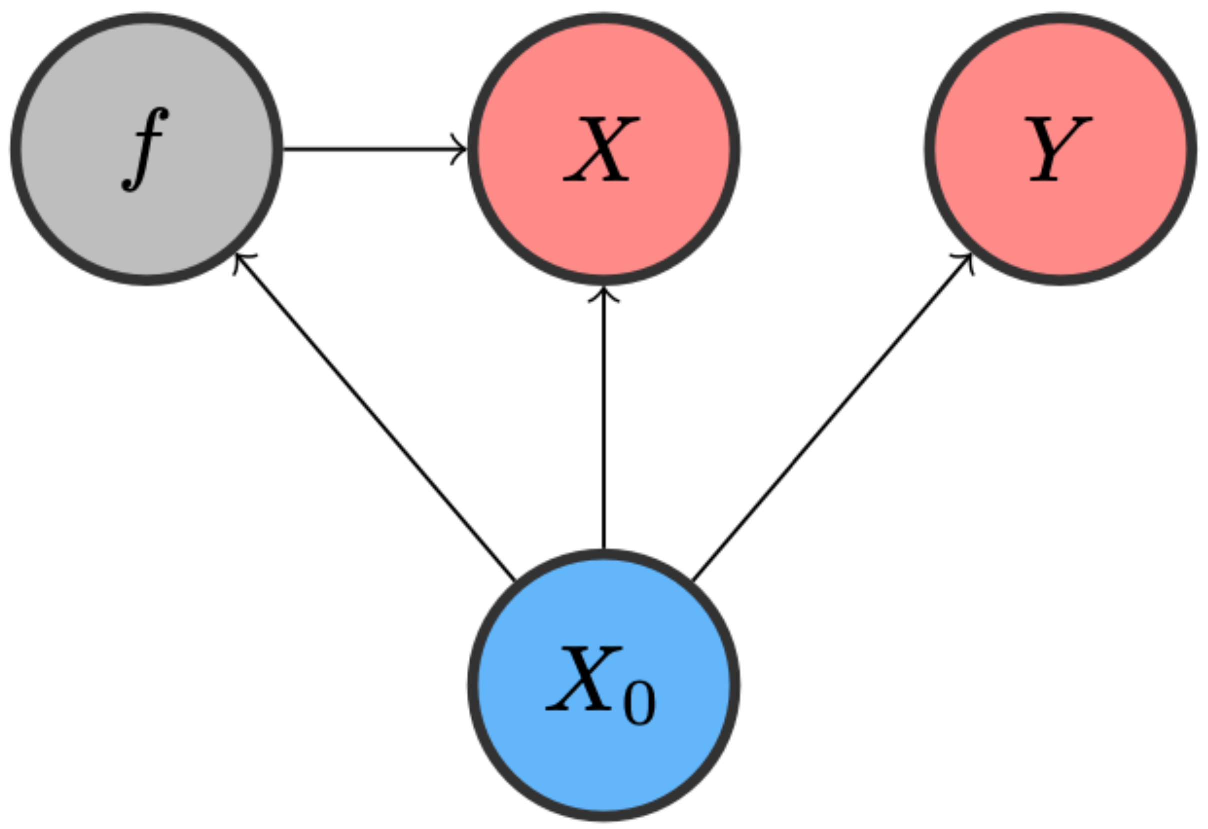
There is a very Causal interpretation to Q.

See *Improvement-Focused Causal Recourse (ICR)* [König et al.] for a more extensive treatment of this view

We view everything on a distributional level



Compliant case



Defiant case

Optimal classifier

Optimal Classifier

Example (Compliant)

We assume that

$$X \mid Y = +1 \sim N(\mu, \Sigma)$$

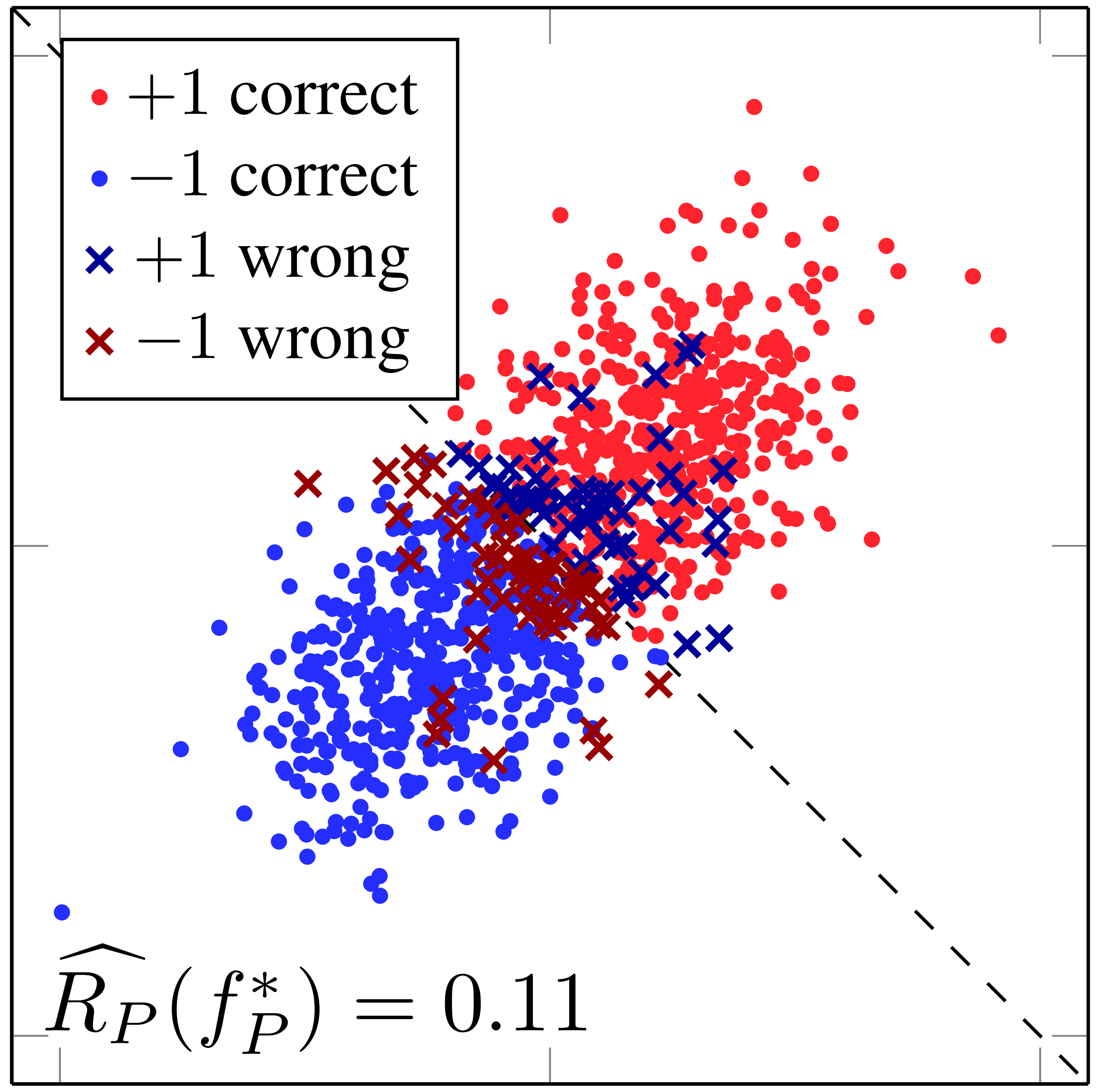
$$X \mid Y = -1 \sim N(\nu, \Sigma)$$

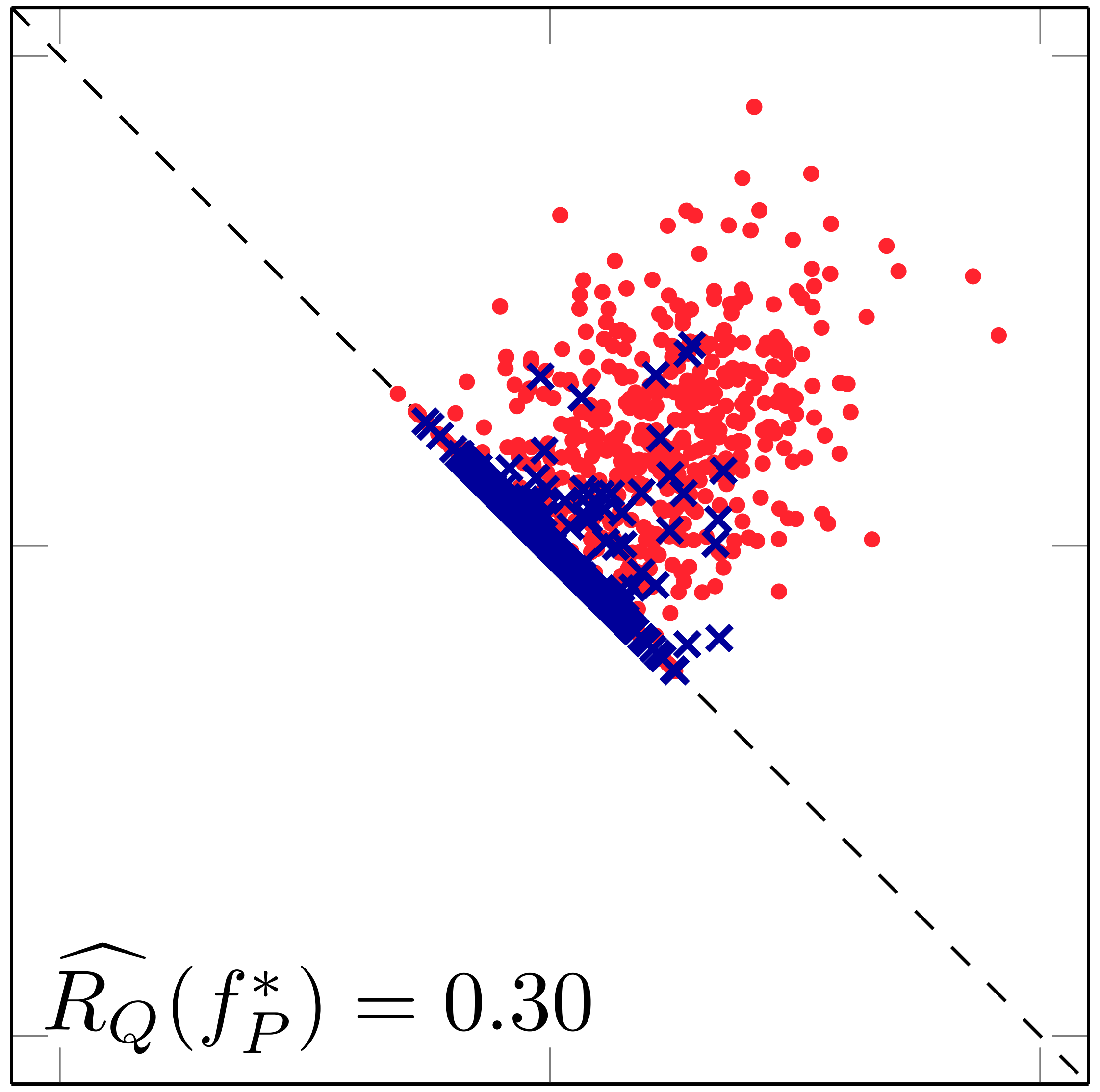
$$P(Y = +1) = P(Y = -1) = \frac{1}{2}$$

Optimal classifier is (assuming $\|\mu\|_{\Sigma^{-1}} = \|\nu\|_{\Sigma^{-1}}$)

$$f_P^*(x) = \text{sign}(\theta^\top x),$$

$$\theta = \Sigma^{-1}(\mu - \nu)$$





- $R_P(f_P^*) = \Phi(\|\mu - \nu\|_{\Sigma^{-1}})$
- $R_Q(f_P^*) = \frac{1}{4} + \frac{1}{2}\Phi(\|\mu - \nu\|_{\Sigma^{-1}})$

$R_Q(f_P^*) > R_P(f_P^*)$ if $R_P(f_P^*) < \frac{1}{2}$

Optimal Classifier

Formal result

Theorem

Let ℓ be the $0/1$ loss and suppose that $P(Y=1 \mid X_0 = x) = \frac{1}{2}$ for all x on the decision boundary of f_P^*, then:

A. For the Compiant case,

$$R_Q(f_P^*) = \frac{1}{2}P(f_P(X_0) = -1) + P(f_P(X_0) = 1, Y = -1) > R_P(f_P^*)$$

B. For the Defiant case,

$$R_Q(f_P^*) = P(Y = -1) > R_P(f_P^*)$$

Optimal Classifier

Proof sketch (Compliant)

$$R_Q(f_P^*) = \frac{1}{2}P(f_P(X_0) = -1) + P(f_P(X_0) = 1, Y = -1) > R_P(f_P^*)$$

- ➡ Every point is now classified as $+1$
- ➡ The mistakes you make are
  - ➡ Original $f_P^*(X_0) = +1$ but $Y = -1$,
  - ➡ Half of the original $f_P^*(X_0) = -1$,
    - ➡ because $P(Y = +1 \mid X) = P(Y = -1 \mid X) = \frac{1}{2}$ on the decision boundary

Optimal Classifier

Proof sketch (Defiant)

$$R_Q(f_P^*) = P(Y = -1) > R_P(f_P^*)$$

- ➡ Every point is now classified as $+1$
- ➡ The mistakes you make are
  - ➡ Original $f_P^*(X_0) = +1$ but $Y = -1$,
  - ➡ Original $f_P^*(X_0) = -1$, but $Y = -1$, because the label does not change in this case

Non-Optimal classifier

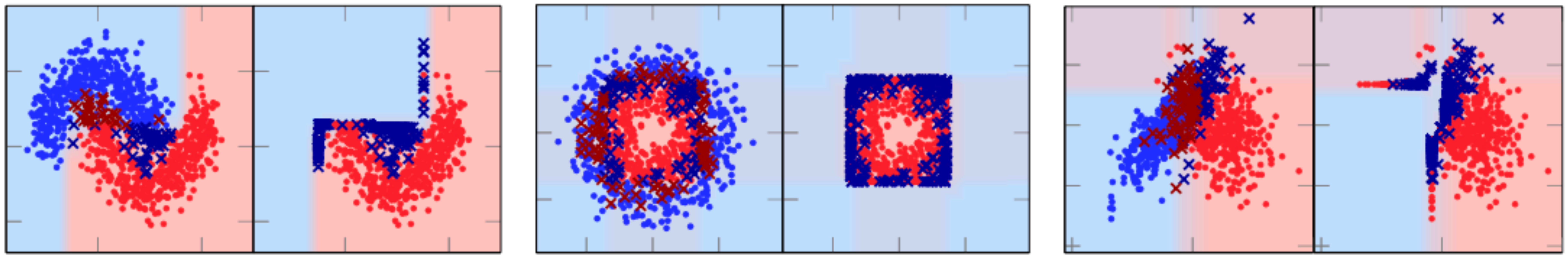
Non-Optimal Classifier

More general

What if we consider non-optimal classifiers?

Need to make some extra assumptions:

- Assume $f(x) = \text{sign}(g(x) - \frac{1}{2})$ for some probabilistic classifier $g(x): \mathcal{X} \to [0,1]$
- The function g is "ϵ-close" to $P(Y = 1 \mid X = x)$ along the decision boundary



$\widehat{R_P}(f) = 0.09$ $\widehat{R_Q}(f) = 0.30$ $\widehat{R_P}(f) = 0.19$ $\widehat{R_Q}(f) = 0.26$ $\widehat{R_P}(f) = 0.13$ $\widehat{R_Q}(f) = 0.33$

ϵ-close assumption

More general

What is the assumption?

Informally:

- The function g is "ϵ-close" to $P(Y=1 \mid X=x)$ along the decision boundary

Formally:

$$\int_{\{x_0:\, g(x_0)<\frac{1}{2}\}} \left|\frac{1}{2} - P(Y=1 \mid X=\varphi(x_0))\right| P(\mathrm{d}x_0) < \epsilon$$

Implied by uniform bound,
$\left|\frac{1}{2} - P(Y=1 \mid X_0=x)\right| < \epsilon$ for all x such that $g(x)=\frac{1}{2}$

Non-Optimal Classifier

Formal result

Theorem

Let ℓ be the $0/1$ loss, $g : \mathcal{X} \rightarrow [0,1]$ a continuous probabilistic classifier and assume the ϵ -condition:

A. For the Compliant case, $R_Q(f)$ is lower and upper bounded by

$$\left(\frac{1}{2} \pm \epsilon\right) P(f(X_0) = -1) + P(f(X_0) = +1, Y = -1)$$

B. For the Defiant case,

$$R_Q(f) = P(Y = -1)$$

Non-Optimal Classifier

Implications

Theorem

Let ℓ be the $0/1$ loss, $g : \mathcal{X} \to [0,1]$ a continuous probabilistic classifier and assume the ϵ -condition:

A. For the Compliant case, $R_Q(f) \geq R_P(f)$ if

$$P(Y = -1 \,|\, f(X_0) = -1) \geq \frac{1}{2} + \epsilon$$

B. For the Defiant case, $R_Q(f) \geq R_P(f)$ if and only if

$$P(Y = -1 \,|\, f(X_0) = -1) \geq \frac{1}{2}$$

Non-Optimal Classifier

Interpretation

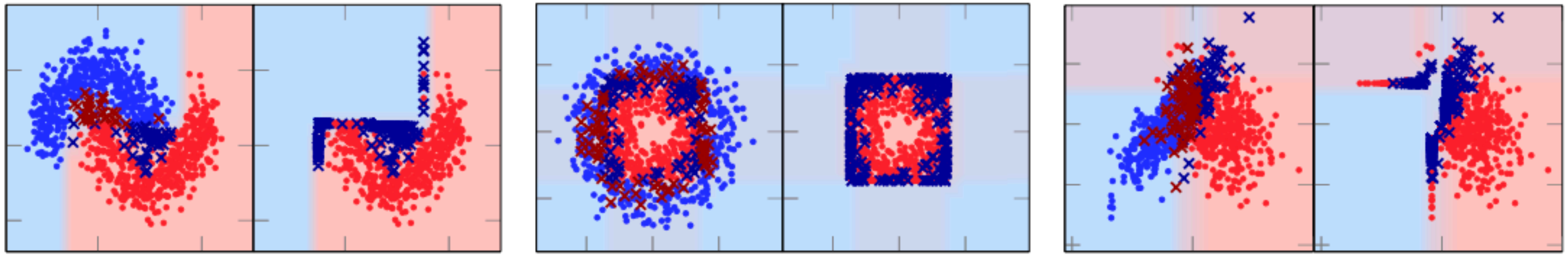
Recourse will harm the risk if

A. For the Compliant case, if f approximates the true conditional distribution and f performs ϵ better on the negative class

B. For the Defiant case, if f performs better than random on the negative class

Non-Optimal Classifier

Experimental results



$\widehat{R_P}(f) = 0.09$ $\widehat{R_Q}(f) = 0.30$ $\widehat{R_P}(f) = 0.19$ $\widehat{R_Q}(f) = 0.26$ $\widehat{R_P}(f) = 0.13$ $\widehat{R_Q}(f) = 0.33$

Table 1: Estimated risks on synthetic data sets. Lower risk is bold.

	Moons data		Circles data		Gaussians data	
	R_P	R_Q	R_P	R_Q	R_P	R_Q
Logistic Regression (LR)	**0.13**	0.33	0.51	**0.34**	**0.14**	0.32
GradientBoostedTrees (GBT)	**0.08**	0.30	**0.19**	0.26	**0.13**	0.33
Decision Tree (DT)	**0.08**	0.29	**0.19**	0.23	**0.13**	0.34
Naive Bayes (NB)	**0.13**	0.33	**0.17**	**0.16**	**0.15**	0.28
QuadraticDiscriminantAnalysis (QDA)	**0.13**	0.33	**0.17**	**0.16**	**0.12**	0.33
Neural Network(4)	**0.12**	0.32	**0.23**	0.30	**0.13**	0.36
Neural Network(4, 4)	**0.04**	0.26	**0.17**	0.22	**0.12**	0.40
Neural Network(8)	**0.04**	0.23	**0.16**	0.20	**0.12**	0.36
Neural Network(8, 16)	**0.04**	0.26	**0.16**	0.18	**0.11**	0.35
Neural Netowrk(8, 16, 8)	**0.04**	0.26	**0.16**	0.18	**0.11**	0.35

Table 2: Estimated risks on real data sets. Lower risk is bold.

	Credit data						Census data						HELOC data					
	Wachter		GS		CoGS		Wachter		GS		CoGS		Wachter		GS		CoGS	
	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q	R_P	R_Q
LR	0.17	**0.05**	0.17	**0.05**	0.17	**0.04**	**0.21**	0.29	**0.21**	0.33	**0.21**	0.32	**0.29**	0.41	**0.29**	0.41	**0.29**	0.44
GBT	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	0.15	**0.04**	**0.15**	0.23	**0.15**	0.33	**0.20**	**0.21**	**0.20**	0.25	**0.20**	0.37
DT	0.29	**0.12**	0.29	**0.05**	0.29	**0.05**	**0.23**	**0.21**	**0.23**	0.43	**0.23**	0.45	**0.19**	0.25	**0.19**	0.21	**0.19**	0.31
NB	0.11	**0.06**	0.11	**0.06**	0.11	**0.07**	**0.19**	0.78	**0.19**	0.76	**0.19**	0.81	**0.29**	0.44	**0.29**	0.43	**0.29**	0.48
QDA	0.12	**0.06**	0.12	**0.06**	0.12	**0.07**	**0.20**	0.78	**0.20**	0.75	**0.20**	0.82	**0.32**	0.46	**0.32**	0.47	**0.32**	0.52
NN(4)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.06**	**0.16**	0.26	**0.16**	0.25	**0.16**	0.26	**0.29**	0.47	**0.29**	0.46	**0.29**	0.50
NN(4, 4)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	**0.15**	0.30	**0.15**	0.27	**0.15**	0.30	**0.29**	0.47	**0.29**	0.47	**0.29**	0.51
NN(8)	**0.06**	**0.06**	**0.06**	**0.06**	**0.06**	**0.07**	**0.16**	0.34	**0.16**	0.33	**0.16**	0.33	**0.28**	0.44	**0.28**	0.46	**0.28**	0.51
NN(8, 16)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	**0.15**	0.36	**0.15**	0.34	**0.15**	0.36	**0.27**	0.42	**0.27**	0.45	**0.27**	0.46
NN(8, 16, 8)	**0.06**	**0.06**	**0.06**	**0.07**	**0.06**	**0.07**	**0.15**	0.36	**0.15**	0.34	**0.15**	0.36	**0.27**	0.42	**0.27**	0.45	**0.27**	0.46

Non-Optimal Classifier

Proof sketch (Compliant)

Upper/lower: $(\frac{1}{2} \pm \epsilon)P(f(X_0) = -1) + P(f(X_0) = +1, Y = -1)$

➡ Every point is now classified as $+1$

➡ The mistakes you make are

- ➡ Original $f(X_0) = +1$ but $Y = -1$,
- ➡ Within ϵ-distance of half the original $f(X_0) = -1$,

➡ Simplify $R_P(f) \leq (\frac{1}{2} - \epsilon)P(f(X_0) = -1)$

Non-Optimal Classifier

Proof sketch (Defiant)

$$R_Q(f) = P(Y = -1) > R_P(f)$$

- ➡ Every point is now classified as $+1$
- ➡ The mistakes you make are
  - ➡ Original $f(X_0) = +1$ but $Y = -1$,
  - ➡ Original $f(X_0) = -1$, but $Y = -1$, because the label does not change in this case

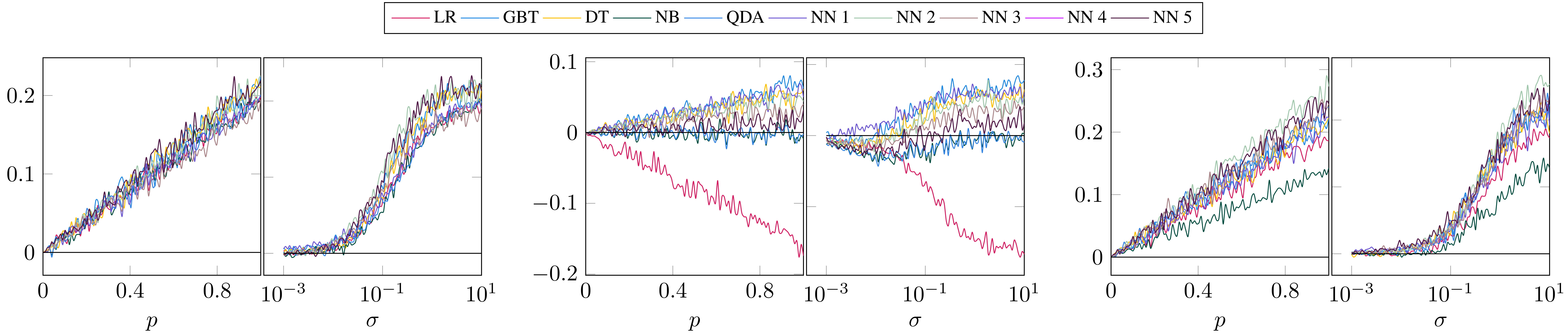
Non-Optimal Classifier

Some more examples

What if $r(x)$ is not constant:

- $r(x) = p \in [0,1]$
- $r(x) = e^{-\frac{1}{2\sigma}\|x-\varphi(x)\|}$

- On y-axis: $R_Q - R_P$
- From L to R: Moons, Circles, Gaussians



Strategising

Strategising

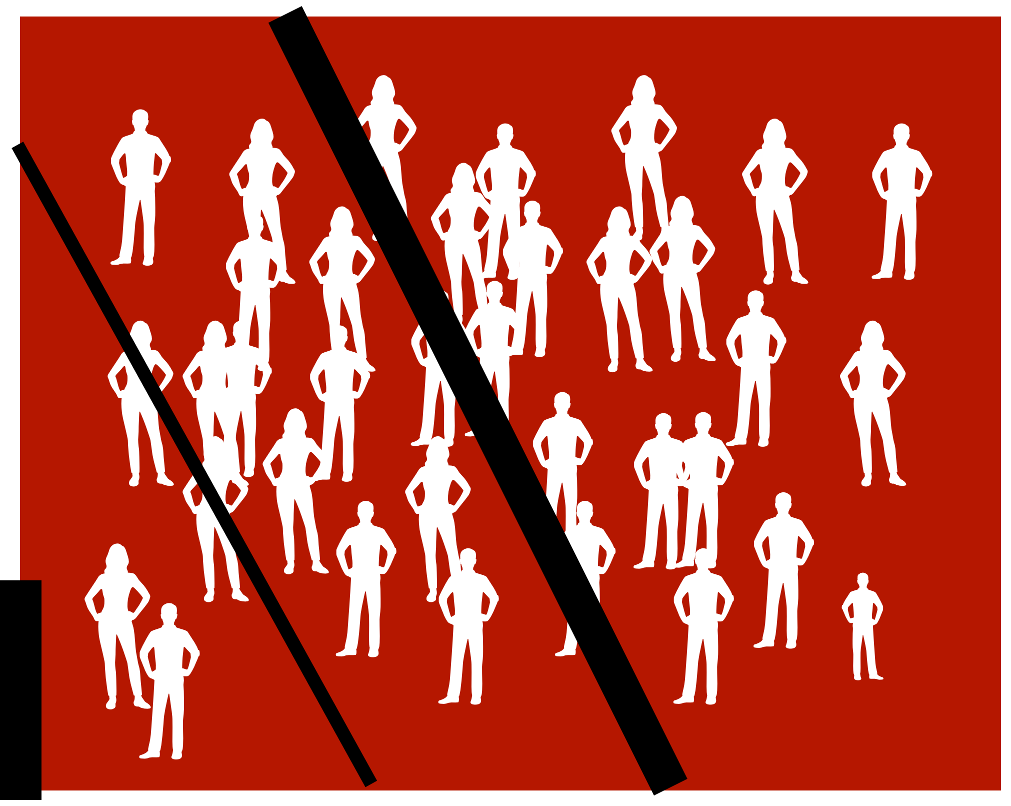
Example

Can (P) strategise against this accuracy drop?

- Need to assume that not everyone gets an explanation, i.e.

$r(x_0) = 1\{\|\varphi(x_0) - x_0\| < D\}$ for some $D > 0$

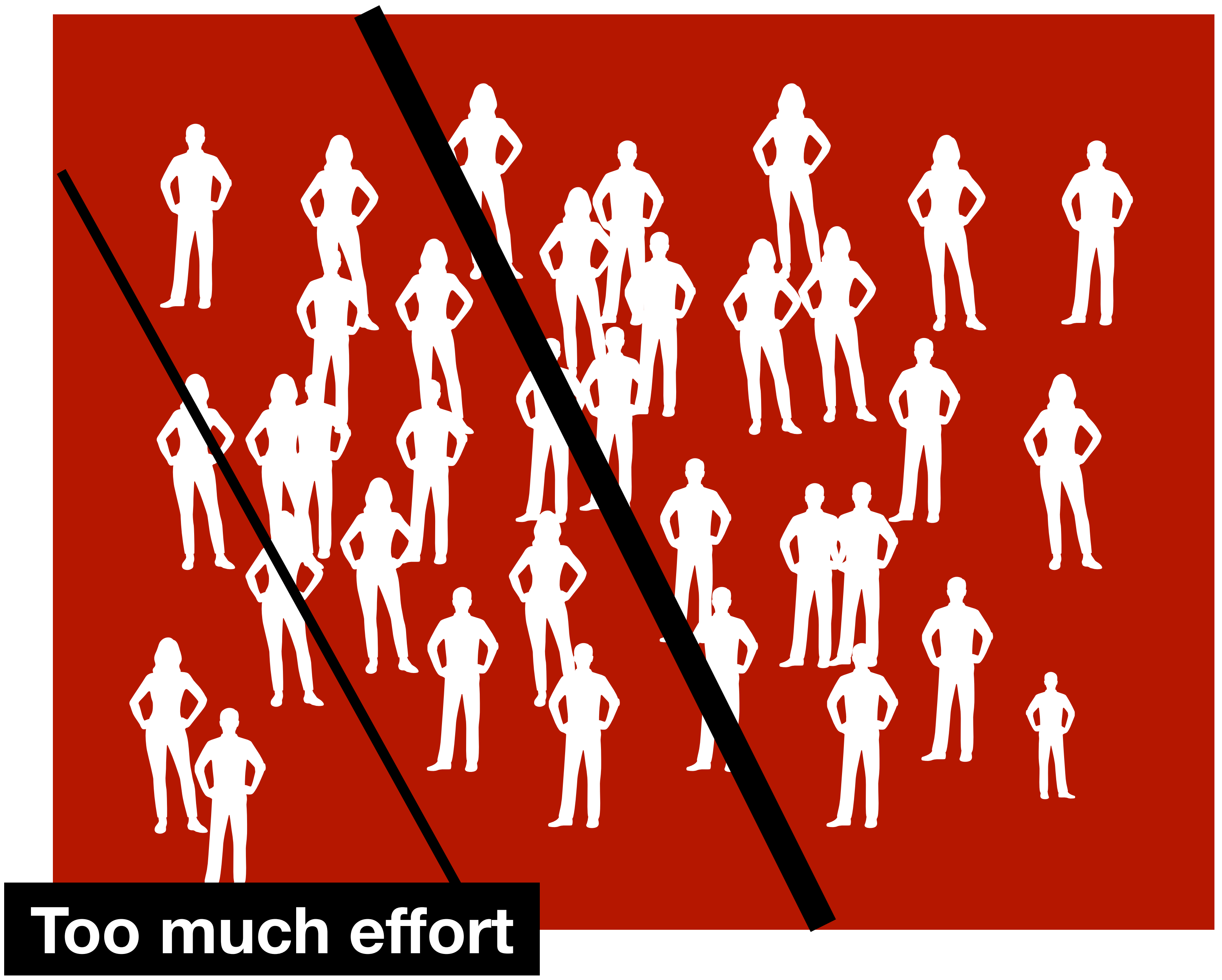
Yes and No

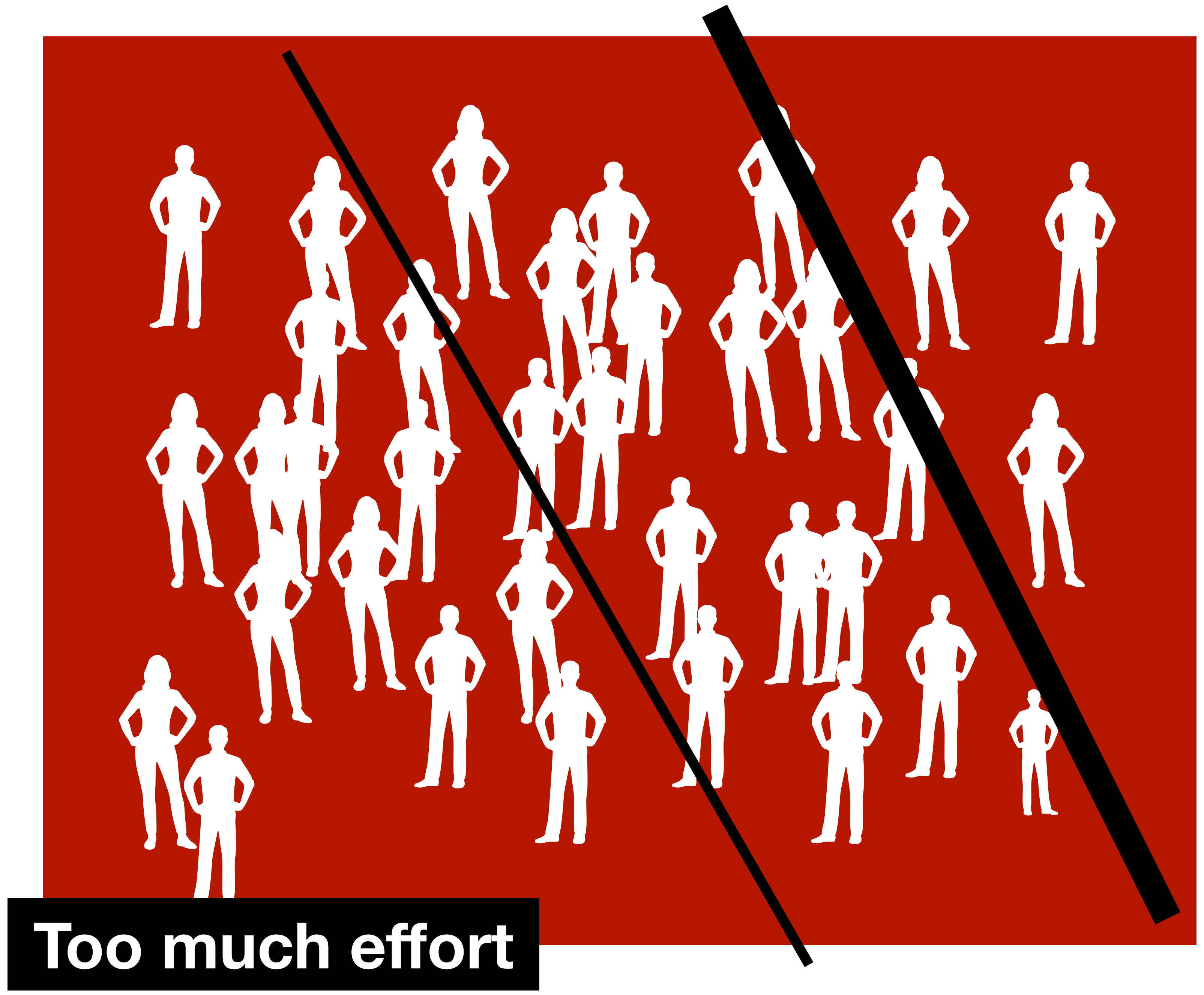


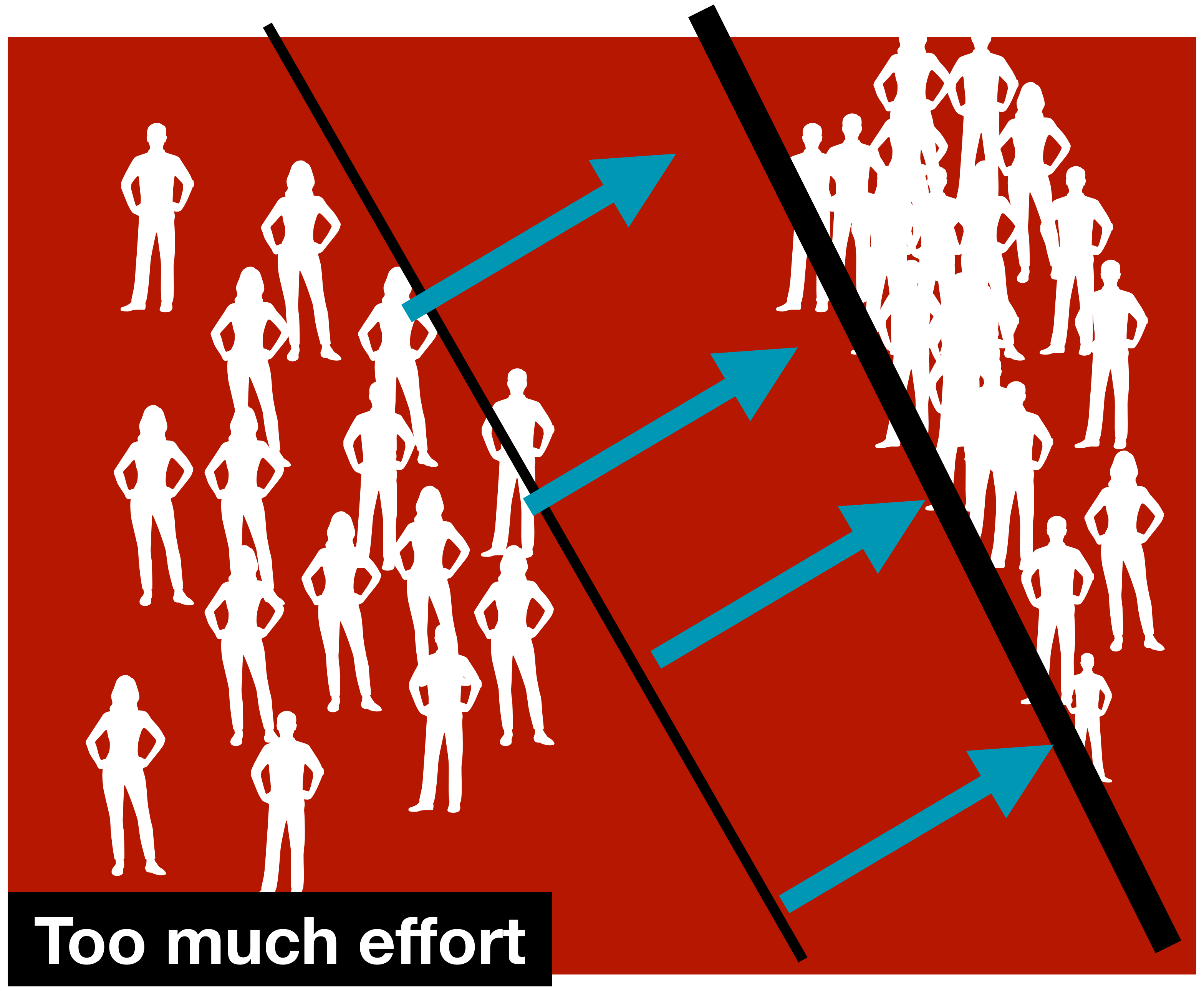
Too much effort

Strategising

Example







Exactly the same people get a loan

More effort

Strategising

- Can generalise this result
  - Defiant case: Can only be as good as before
  - Compliant case: In principle, could have arbitrary improvement, but would increase the cost for every applicant substantially

Strategising

Formal result, a new definitions

Definition

Let $\mathscr{F}$ be some model class, $r(x_0) \in \{0,1\}$and $\varphi^r(x_0) = r(x_0)\varphi(x_0) + (1 - r(x_0))\varphi(x_0)$, define

$$\mathscr{F}_\varphi^r := \{f' : x_0 \mapsto f(\varphi^r(x_0)) \mid f \in \mathscr{F}\}.$$

The set of functions induced by the recourse map.

If $\mathscr{F}_\varphi^r = \mathscr{F}$, we call $\mathscr{F}$ ***invariant under recourse***

For example,

$\mathscr{F} = \{f(x) = \text{sign}(a^\top x + b) \mid a, b \in \mathbb{R}^d\}$ and $r(x_0) = 1\{\|\varphi(x_0) - x_0\| < D\}$.

Then $\mathscr{F} = \mathscr{F}_\varphi$

Strategising

Formal result (Defiant)

Theorem

If $\mathcal{F}$ is a recourse invariant model class, then

$$\min_{f \in \mathcal{F}} R_P(f) = \min_{f \in \mathcal{F}} R_Q(f).$$

You can only do as well as before

Strategising

Formal result (Compliant)

Theorem

If $\mathscr{F}$ is a recourse invariant model class, then

$$\min_{f \in \mathscr{F}} R_Q(f) \leq \min_{f \in \mathscr{F}} R_P(f) - \gamma.$$

Where $\gamma \in \mathbb{R}$ depends on P and f.

Improvement is possible when $\gamma > 0$!

For example, in the Gaussian example

However, there will be a cost for every individual

Some perspectives

Some perspectives

When can Recourse still be beneficial?

Some possible answers:

- In situations where Accuracy is not the most important metric
- When counterfactuals have other explanatory properties
- Should Recourse be replaced by Contestability?

Thank you for your attention!