UNIVERSITY OF AMSTERDAM
Korteweg de Vries Institute for Mathematics

The Risks of Recourse in Binary Classification

2023-11-09

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

- Explainable AI
- Introduction to the problem
- Result for Optimal Classifiers
- Conclusion

Explainable AI

Call for XAI

Some Reasons

- Fairness: Biases can be detected earlier
- Trustworthiness
- Increases reliability
- Regulation

Explanations

Explosion

Methods
CAM with global average pooling [42], [91]
+ Grad-CAM [43] generalizes CAM, utilizing gradient
+ Guided Grad-CAM and Feature Occlusion [68]
+ Respond CAM [44]
+ Multi-layer CAM [92]
LRP (Layer-wise Relevance Propagation) [13], [53]
+ Image classifications. PASCAL VOC 2009 etc [45]
+ Audio classification. AudioMNIST [47]
+ LRP on DeepLight. fMRI data from Human Connectome Project [48]
+ LRP on CNN and on BoW(bag of words)/SVM [49]
+ LRP on compressed domain action recognition algorithm [50]
+ LRP on video deep learning, *selective relevance method* [52]
+ BiLRP [51]
DeepLIFT [57]
Prediction Difference Analysis [58]
Slot Activation Vectors [41]
PRM (Peak Response Mapping) [59]
LIME (Local Interpretable Model-agnostic Explanations) [14]
+ MUSE with LIME [85]
+ Guidelinebased Additive eXplanation optimizes complexity, similar to LIME [93]
# Also listed elsewhere: [56], [69], [71], [94]
Others. Also listed elsewhere: [95]
+ Direct output labels. Training NN via multiple instance learning [65]
+ Image corruption and testing Region of Interest statistically [66]
+ Attention map with autofocus convolutional layer [67]
DeconvNet [72]
Inverting representation with natural image prior [73]
Inversion using CNN [74]
Guided backpropagation [75], [91]
Activation maximization/optimization [38]
+ Activation maximization on DBN (Deep Belief Network) [76]
+ Activation maximization, multifaceted feature visualization [77]
Visualization via regularized optimization [78]
Semantic dictionary [39]
Network dissection [36], [37]
Decision trees
Propositional logic, rule-based [82]
Sparse decision list [83]
Decision sets, rule sets [84], [85]
Encoder-generator framework [86]
Filter Attribute Probability Density Function [87]
MUSE (Model Understanding through Subspace Explanations) [85]

(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

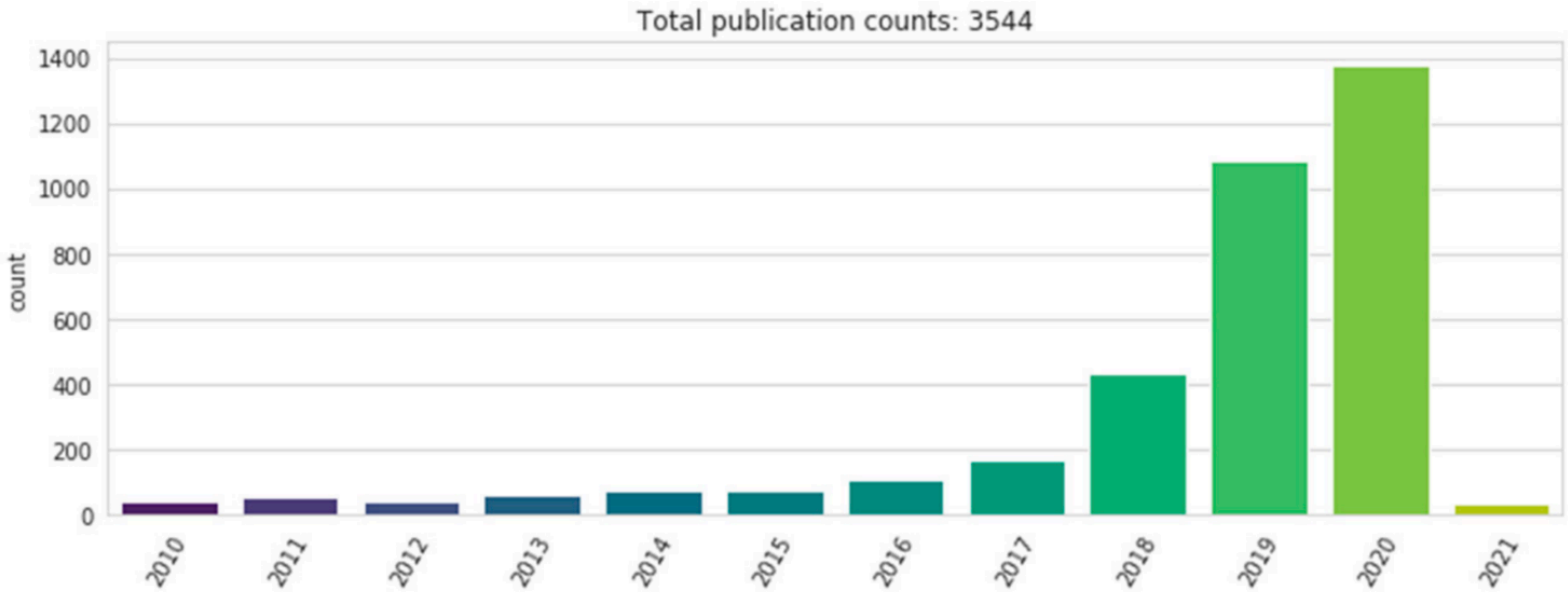
(2014.03)	SEDC [129]
(2015.08)	OAE [51]
(2016.05)	HCLS [110, 112]
(2017.06)	Feature Tweaking [186]
(2017.11)	CF Expl. [196]
(2017.12)	Growing Spheres [114]
(2018.02)	CEM [55]
(2018.02)	POLARIS [209]
(2018.05)	LORE [80]
(2018.06)	Local Foil Trees [190]
(2018.09)	Actionable Recourse [189]
(2018.11)	Weighted CFs [77]
(2019.01)	Efficient Search [175]
(2019.04)	CF Visual Expl. [76]
(2019.05)	MACE [99]
(2019.05)	DiCE [145]
(2019.05)	CERTIFAI [179]
(2019.06)	MACEM [56]
(2019.06)	Expl. using SHAP [165]
(2019.07)	Nearest Observable [201]
(2019.07)	Guided Prototypes [191]
(2019.07)	REVISE [95]
(2019.08)	CLEAR [202]
(2019.08)	MC-BRP [123]
(2019.09)	FACE [162]
(2019.09)	Equalizing Recourse [83]
(2019.10)	Action Sequences [163]
(2019.10)	C-CHVAE [156]
(2019.11)	FOCUS [124]
(2019.12)	Model-based CFs [127]
(2019.12)	LIME-C/SHAP-C [164]
(2019.12)	EMAP [41]
(2019.12)	PRINCE [71]
(2019.12)	LowProFool [18]
(2020.01)	ABELE [79]
(2020.01)	SHAP-based CFs [66]
(2020.02)	CEML [11–13]
(2020.02)	MINT [100]
(2020.03)	ViCE [74]
(2020.03)	Plausible CFs [22]
(2020.04)	SEDC-T [193]
(2020.04)	MOC [52]
(2020.04)	SCOUT [199]
(2020.04)	ASP-based CFs [28]
(2020.05)	CBR-based CFs [103]
(2020.06)	Survival Model CFs [106]
(2020.06)	Probabilistic Recourse [101]
(2020.06)	C-CHVAE [155]
(2020.07)	FRACE [210]
(2020.07)	DACE [96]
(2020.07)	CRUDS [60]
(2020.07)	Gradient Boosted CFs [5]
(2020.08)	Gradual Construction [97]
(2020.08)	DECE [44]
(2020.08)	Time Series CFs [16]
(2020.08)	PermuteAttack [87]
(2020.10)	Fair Causal Recourse [195]
(2020.10)	Recourse Summaries [167]
(2020.10)	Strategic Recourse [43]
(2020.11)	PARE [172]

(2020) A survey of algorithmic recourse: definitions, formulations, solutions, and prospects

Methods	H
Linear probe [101]	.
Regression based on CNN [106]	.
Backwards model for interpretability of linear models [107]	.
GDM (Generative Discriminative Models): ridge regression + least square [100]	.
GAM, GA^{2}M (Generative Additive Model) [82], [102], [103]	.
ProtoAttend [105]	.
Other content-subject-specific models:	N
+ Kinetic model for CBF (cerebral blood flow) [131]	N
+ CNN for PK (Pharmacokinetic) modelling [132]	N
+ CNN for brain midline shift detection [133]	N
+ Group-driven RL (reinforcement learning) on personalized healthcare [134]	N
+ Also see [108]–[112]	N
PCA (Principal Components Analysis), SVD (Singular Value Decomposition)	N
CCA (Canonical Correlation Analysis) [113]	.
SVCCA (Singular Vector Canonical Correlation Analysis) [97] = CCA+SVD	.
F-SVD (Frame Singular Value Decomposition) [114] on electromyography data	.
DWT (Discrete Wavelet Transform) + Neural Network [135]	.
MODWPT (Maximal Overlap Discrete Wavelet Package Transform) [136]	.
GAN-based Multi-stage PCA [118]	.
Estimating probability density with deep feature embedding [119]	.
t-SNE (t-Distributed Stochastic Neighbour Embedding) [77]	.
+ t-SNE on CNN [120]	.
+ t-SNE, activation atlas on GoogleNet [121]	.
+ t-SNE on latent space in meta-material design [122]	.
+ t-SNE on genetic data [137]	.
+ mm-t-SNE on phenotype grouping [138]	.
Laplacian Eigenmaps visualization for Deep Generative Model [124]	.
KNN (k-nearest neighbour) on multi-center low-rank rep. learning (MCLRR) [125]	.
KNN with triplet loss and *query-result activation map pair* [139]	.
Group-based Interpretable NN with RW-based Graph Convolutional Layer [123]	.
TCAV (Testing with Concept Activation Vectors) [96]	.
+ RCV (Regression Concept Vectors) uses TCAV with Br score [140]	.
+ Concept Vectors with UBS [141]	.
+ ACE (Automatic Concept-based Explanations) [56] uses TCAV	.
Influence function [129] helps understand adversarial training points	.
Representer theorem [130]	.
SocRat (Structured-output Causal Rationalizer) [127]	.
Meta-predictors [126]	.
Explanation vector [128]	.
# Also listed elsewhere: [14], [43], [85], [94]	N
# Also listed elsewhere: [14], [60], [85] etc	N
CNN with separable model [142]	.
Information theoretic: Information Bottleneck [98], [99]	.
Database of methods v.s. interpretability [10]	N
Case-Based Reasoning [143]	.

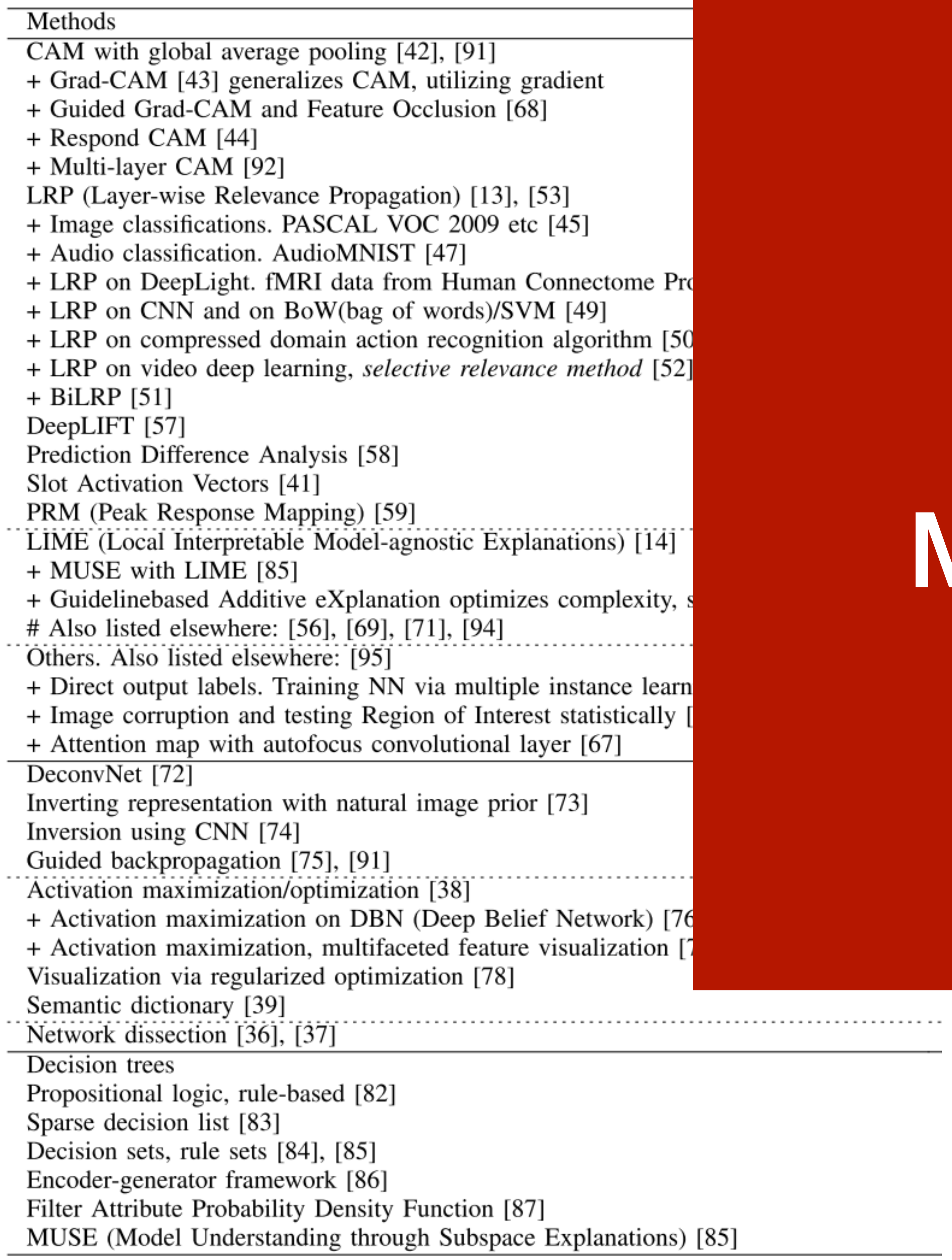
(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

(2021) Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics



Explanations

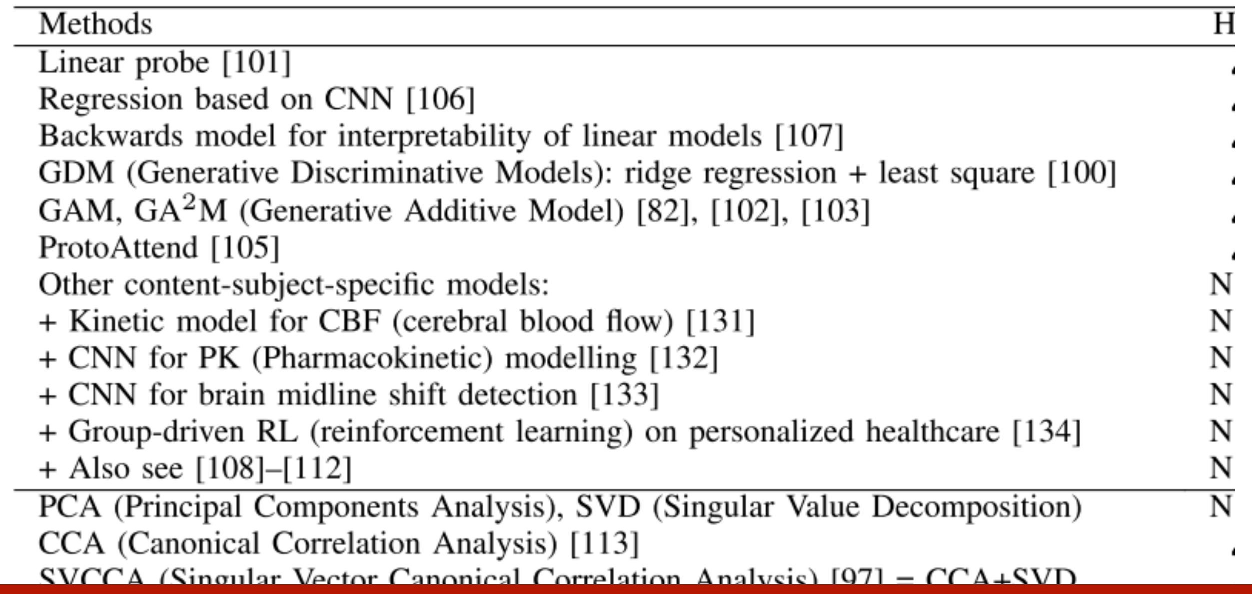
Explosion


Methods
CAM with global average pooling [42], [91]
+ Grad-CAM [43] generalizes CAM, utilizing gradient
+ Guided Grad-CAM and Feature Occlusion [68]
+ Respond CAM [44]
+ Multi-layer CAM [92]
LRP (Layer-wise Relevance Propagation) [13], [53]
+ Image classifications. PASCAL VOC 2009 etc [45]
+ Audio classification. AudioMNIST [47]
+ LRP on DeepLight. fMRI data from Human Connectome Pr
+ LRP on CNN and on BoW(bag of words)/SVM [49]
+ LRP on compressed domain action recognition algorithm [50
+ LRP on video deep learning, *selective relevance method* [52]
+ BiLRP [51]
DeepLIFT [57]
Prediction Difference Analysis [58]
Slot Activation Vectors [41]
PRM (Peak Response Mapping) [59]
LIME (Local Interpretable Model-agnostic Explanations) [14]
+ MUSE with LIME [85]
+ Guidelinebased Additive eXplanation optimizes complexity, s
# Also listed elsewhere: [56], [69], [71], [94]
Others. Also listed elsewhere: [95]
+ Direct output labels. Training NN via multiple instance learn
+ Image corruption and testing Region of Interest statistically [
+ Attention map with autofocus convolutional layer [67]
DeconvNet [72]
Inverting representation with natural image prior [73]
Inversion using CNN [74]
Guided backpropagation [75], [91]
Activation maximization/optimization [38]
+ Activation maximization on DBN (Deep Belief Network) [76
+ Activation maximization, multifaceted feature visualization [7
Visualization via regularized optimization [78]
Semantic dictionary [39]
Network dissection [36], [37]
Decision trees
Propositional logic, rule-based [82]
Sparse decision list [83]
Decision sets, rule sets [84], [85]
Encoder-generator framework [86]
Filter Attribute Probability Density Function [87]
MUSE (Model Understanding through Subspace Explanations) [85]

(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI


(2014.03) SEDC [129]
(2015.08) OAE [51]
(2016.05) HCLS [110, 112]
(2017.06) Feature Tweaking [186]
(2017.11) CF Expl. [196]
(2017.12) Growing Spheres [114]
(2018.02) CEM [55]
(2018.02) POLARIS [209]
(2018.05) LORE [80]
(2018.06) Local Foil Trees [190]
(2018.09) Actionable Recourse [189]
(2018.11) Weighted CFs [77]
(2019.01) Efficient Search [175]
(2019.04) CF Visual Expl. [76]
(2019.05) MACE [99]
(2019.05) DiCE [145]
(2020.08) PermuteAttack [87]
(2020.10) Fair Causal Recourse [195]
(2020.10) Recourse Summaries [167]
(2020.10) Strategic Recourse [43]
(2020.11) PARE [172]

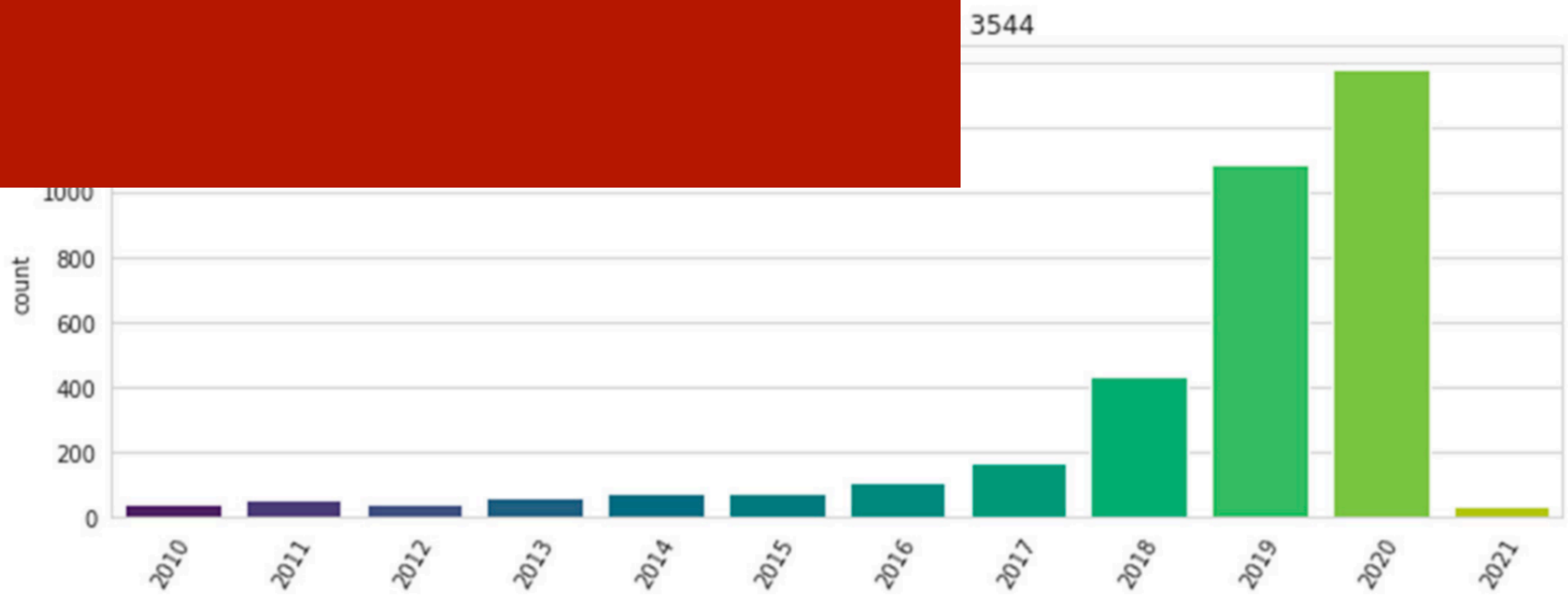
(2020) A survey of algorithmic recourse: definitions, formulations, solutions, and prospects


Methods H
Linear probe [101]
Regression based on CNN [106]
Backwards model for interpretability of linear models [107]
GDM (Generative Discriminative Models): ridge regression + least square [100]
GAM, GA^2M (Generative Additive Model) [82], [102], [103]
ProtoAttend [105]
Other content-subject-specific models: N
+ Kinetic model for CBF (cerebral blood flow) [131] N
+ CNN for PK (Pharmacokinetic) modelling [132] N
+ CNN for brain midline shift detection [133] N
+ Group-driven RL (reinforcement learning) on personalized healthcare [134] N
+ Also see [108]–[112] N
PCA (Principal Components Analysis), SVD (Singular Value Decomposition) N
CCA (Canonical Correlation Analysis) [113]
SVCCA (Singular Vector Canonical Correlation Analysis) [97] = CCA+SVD

(2019) A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

But not a lot of Theoretical/ Mathematical understanding!

(2021) Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics

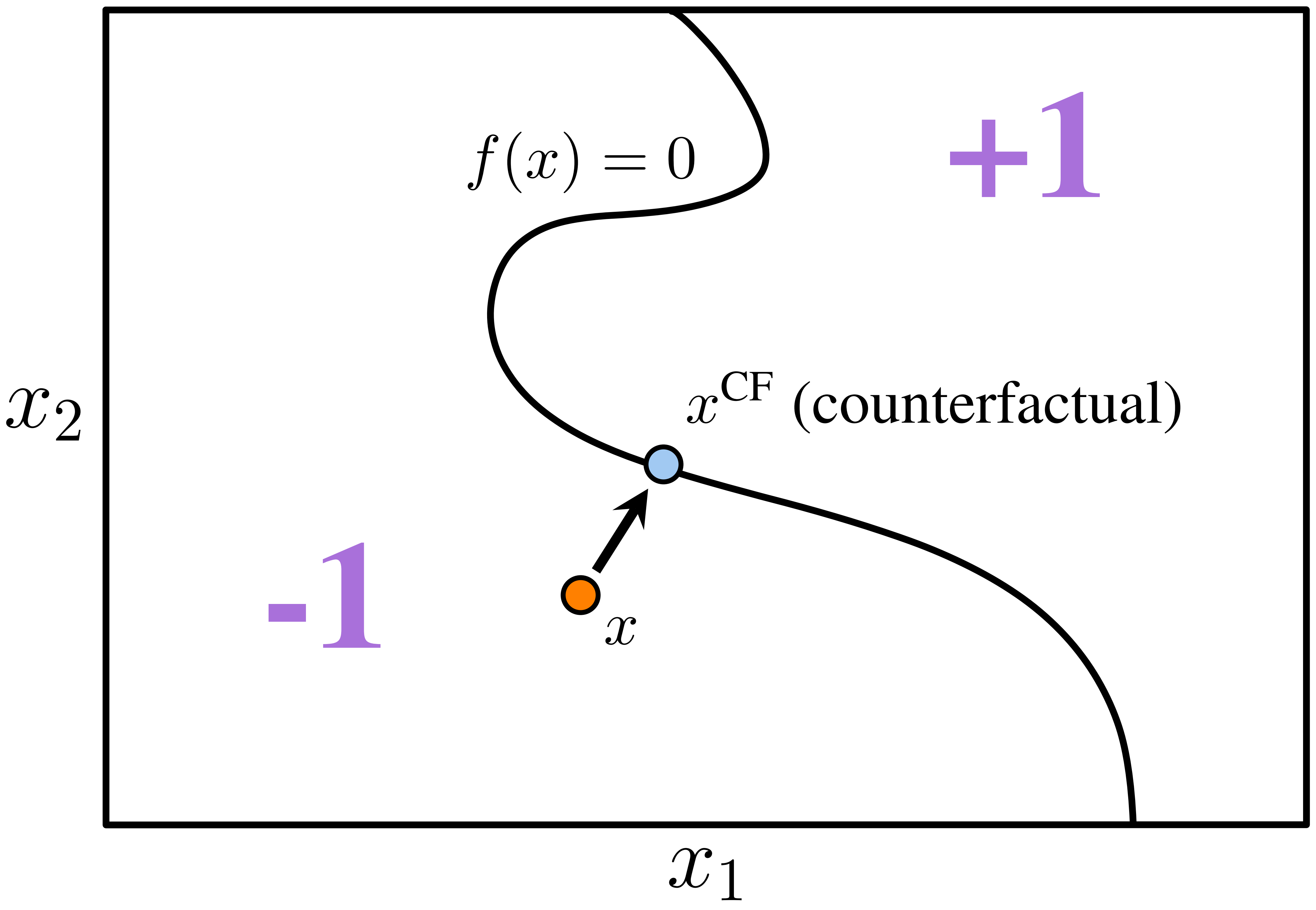


Introduction to the problem

With a peculiar example

Explanations

Counterfactual explanation



"If you would have had an income of € 40 000 instead of €35 000, your loan request would have been approved."

- Tell (A) how to change the decision from -1 to +1
- Minimal cost for (A)
- Provide ***Recourse***

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

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

$$\begin{aligned} f(x) &= +1 \quad \text{if accepted} \\ f(x) &= -1 \quad \text{if not} \end{aligned}$$

- (A) can ask for a counterfactual explanation

This example us seen as a:

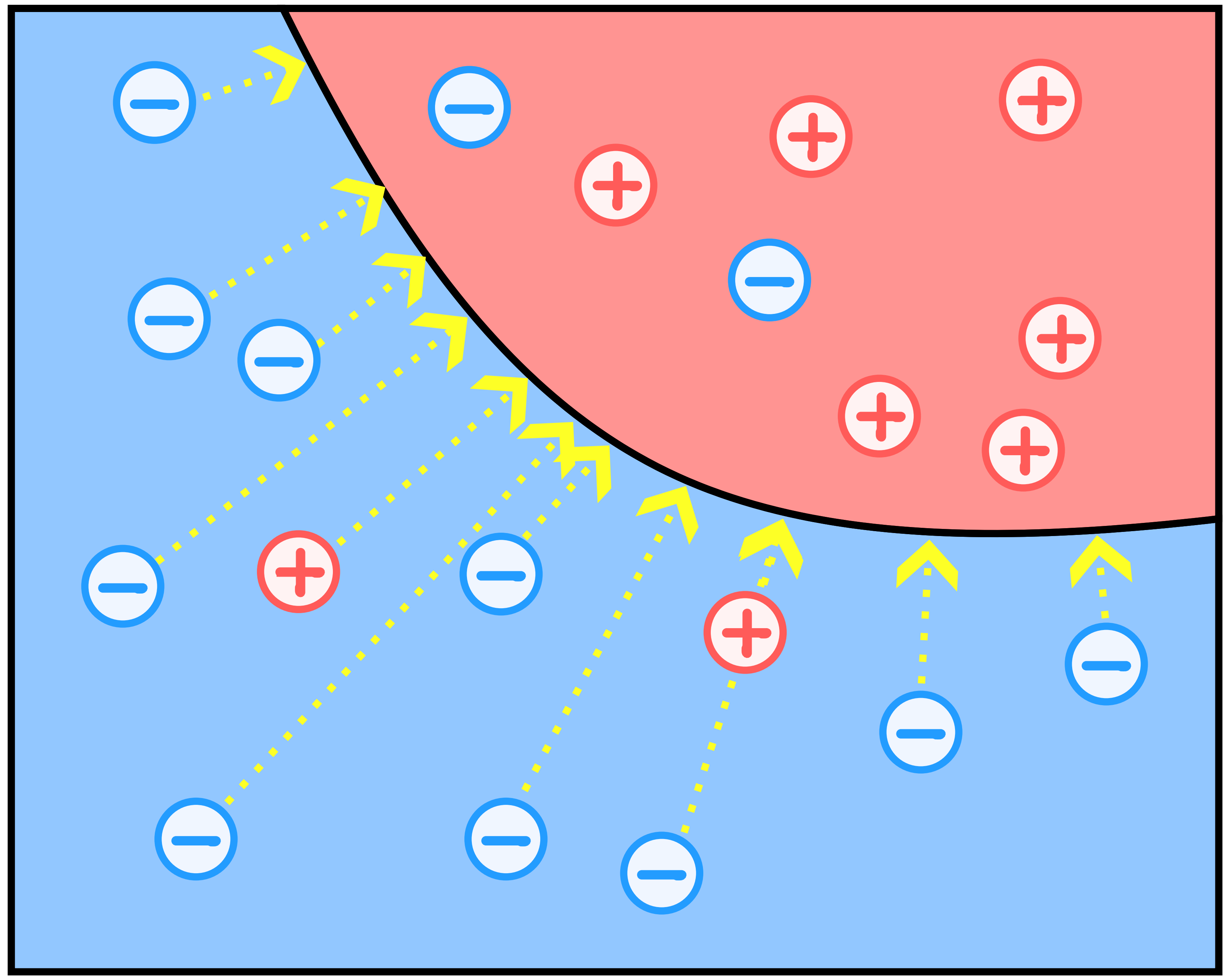
Counterfactual literature

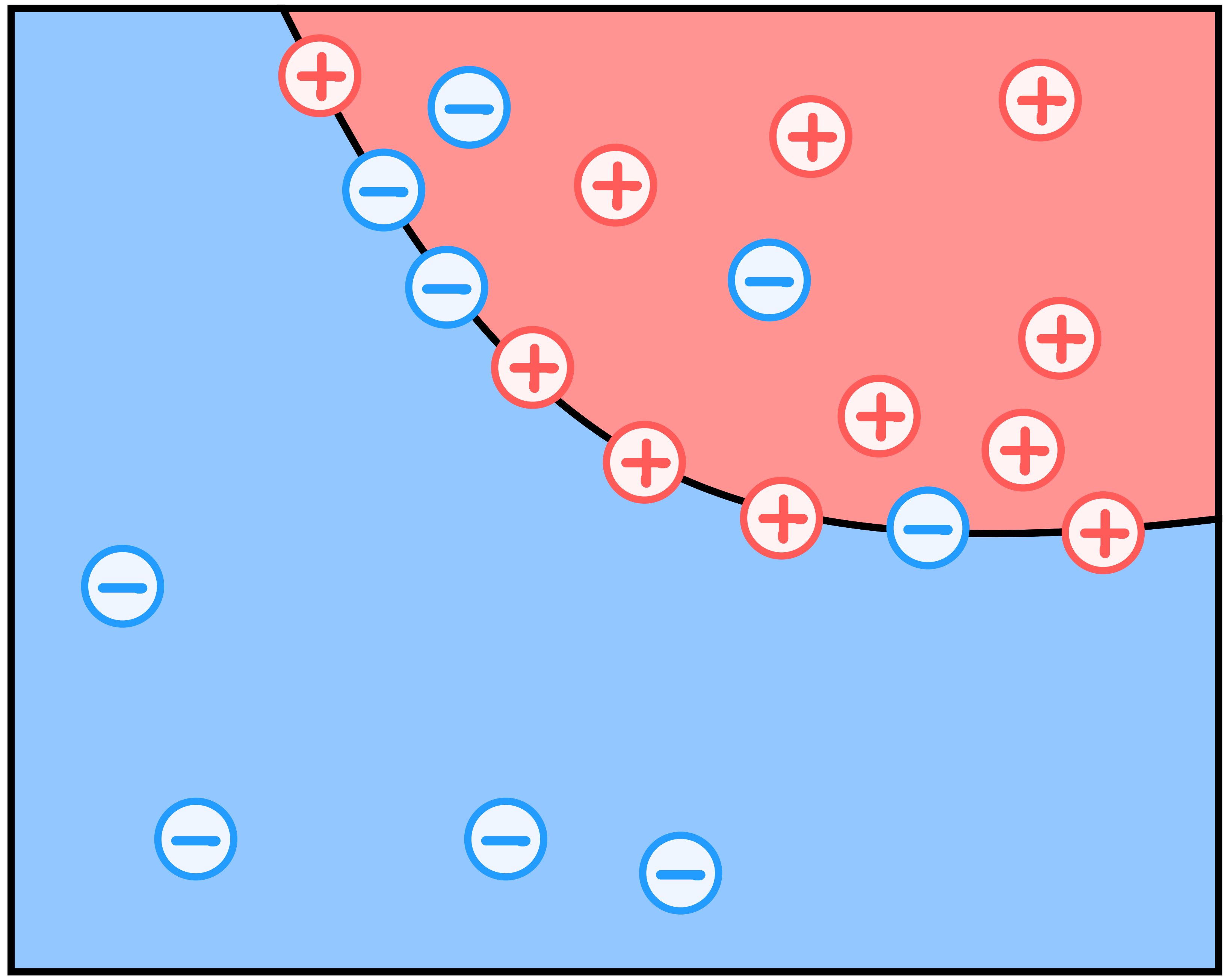
Positive example

Strategic classification

Negative example

Effect of Recourse on Accuracy





What happens to the accuracy?

Model

Learning theoretic setting for classification

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

We assume that

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

We care about ***accuracy***:

$$R_P(f) = P(f(X_0) \neq Y).$$

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

$$f_P^* = \operatorname{sign}\left(P(Y=1 \,|\, X=x) - \frac{1}{2}\right).$$

By adding recourse in the mix,

$$X_0 \to X,$$

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

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

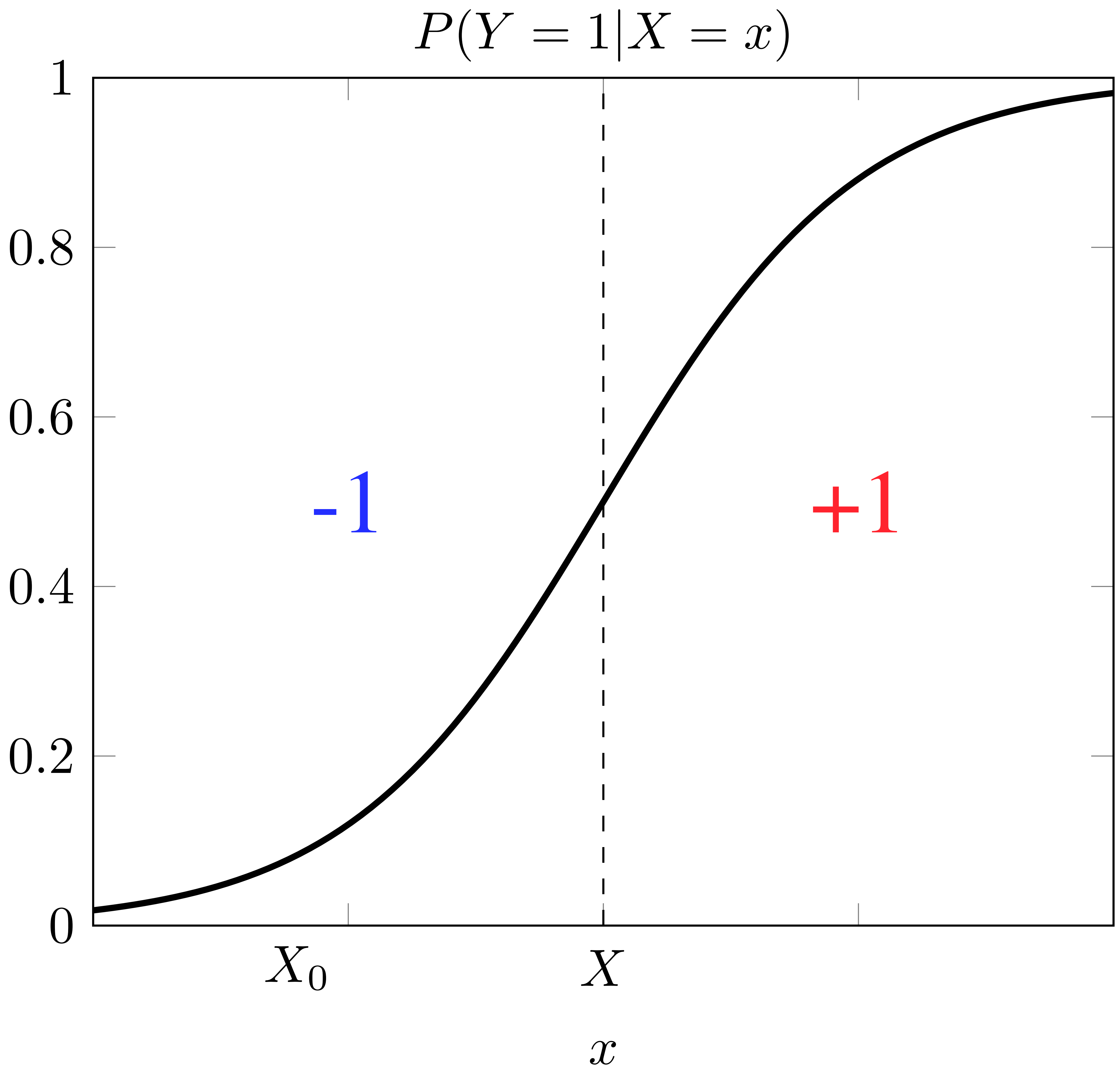
Accuracy with **Recourse** is defined as

$$R_Q(f) = Q(f(X) \neq Y).$$

Note that Q depends on f in general.

Distribution of $Y \,|\, X$ may change.

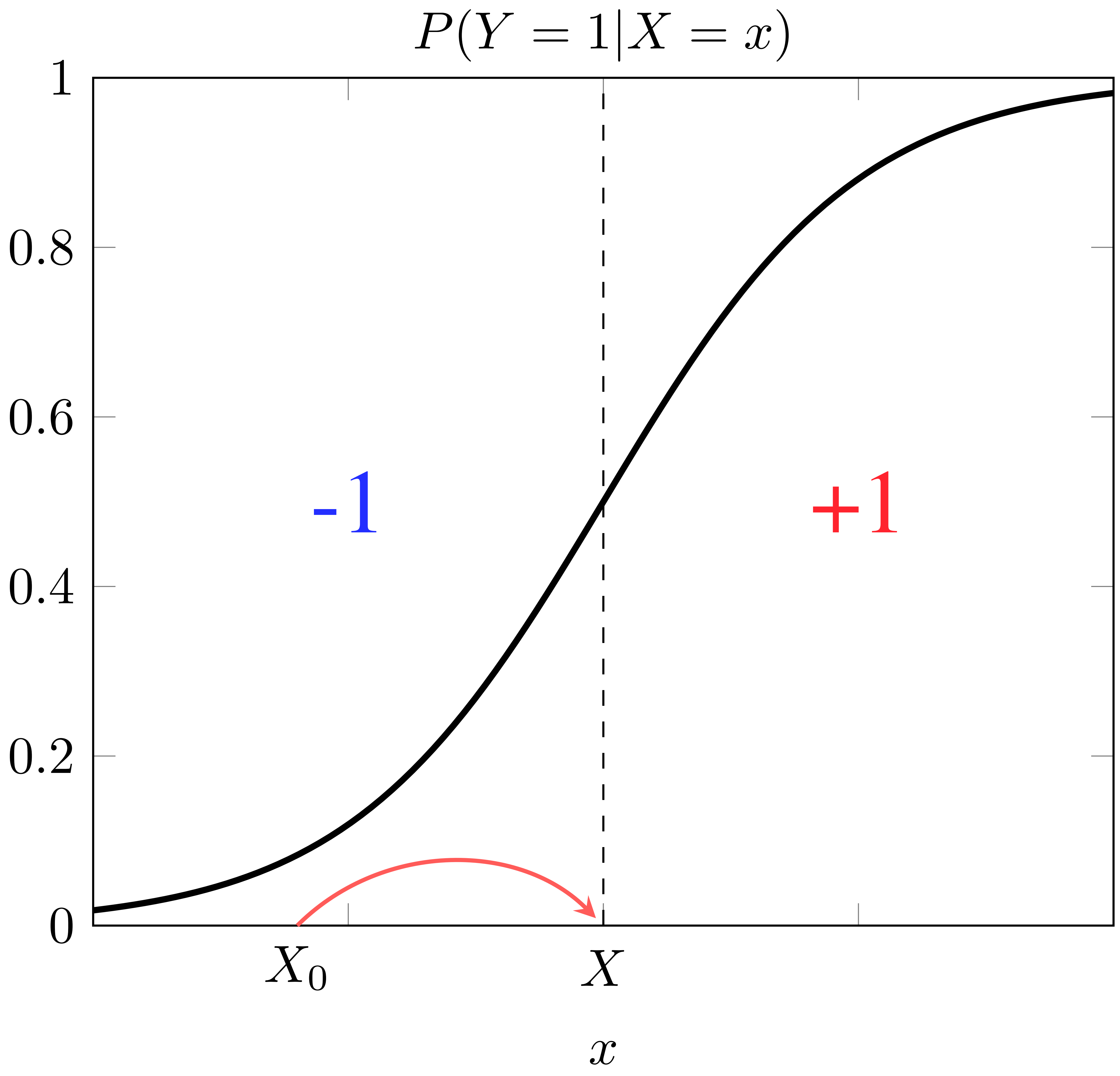
Modelling User Bahaviour



$X_0 \rightarrow X$

Distribution of $Y|X$ may change.

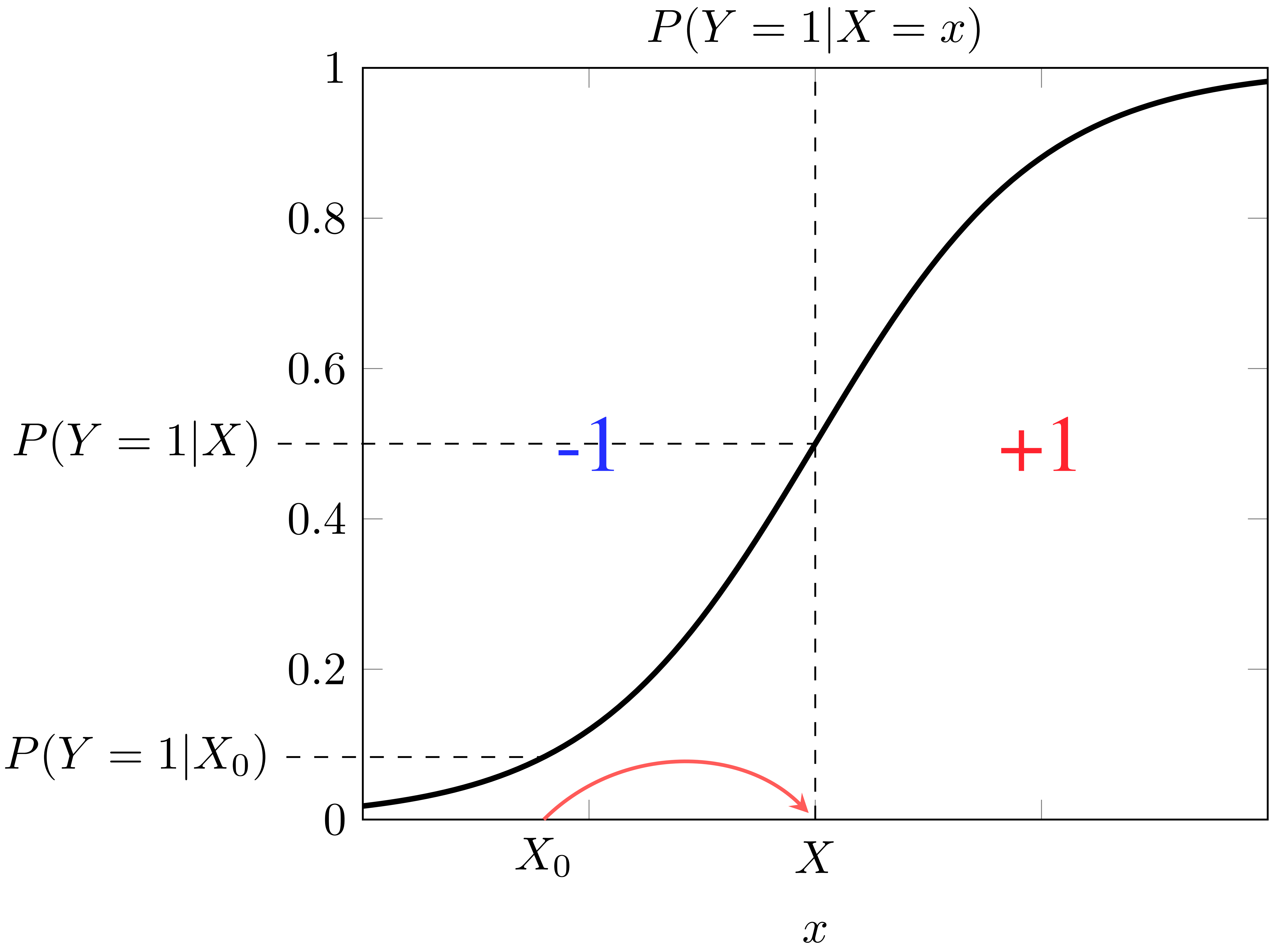
Modelling User Bahaviour



$X_0 \to X$

Distribution of $Y|X$ may change.

Modelling User Bahaviour



$X_0 \rightarrow X$

Distribution of $Y|X$ may change.

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

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

Optimal classifier

Optimal Classifier

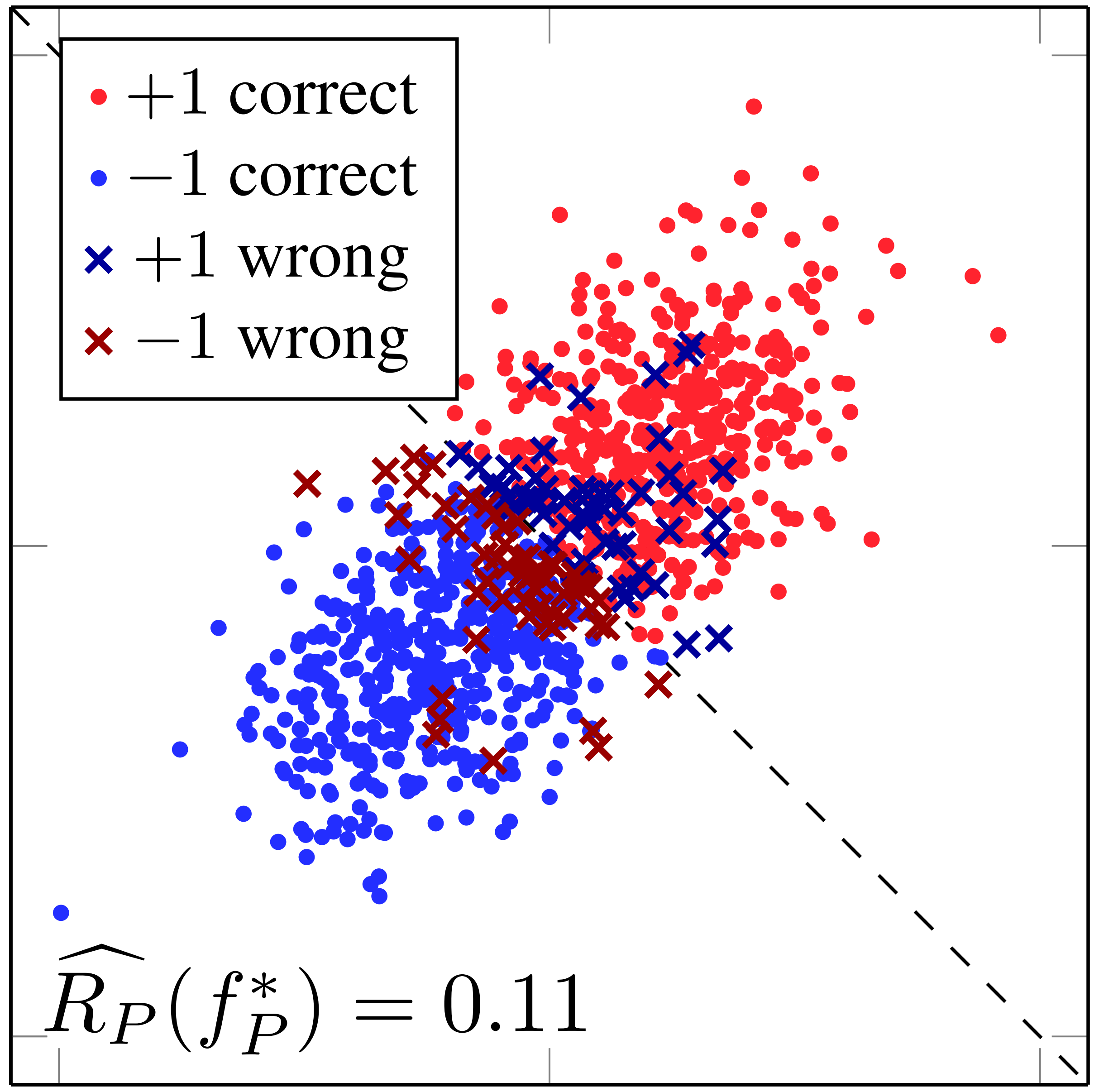
Example (Compliant)

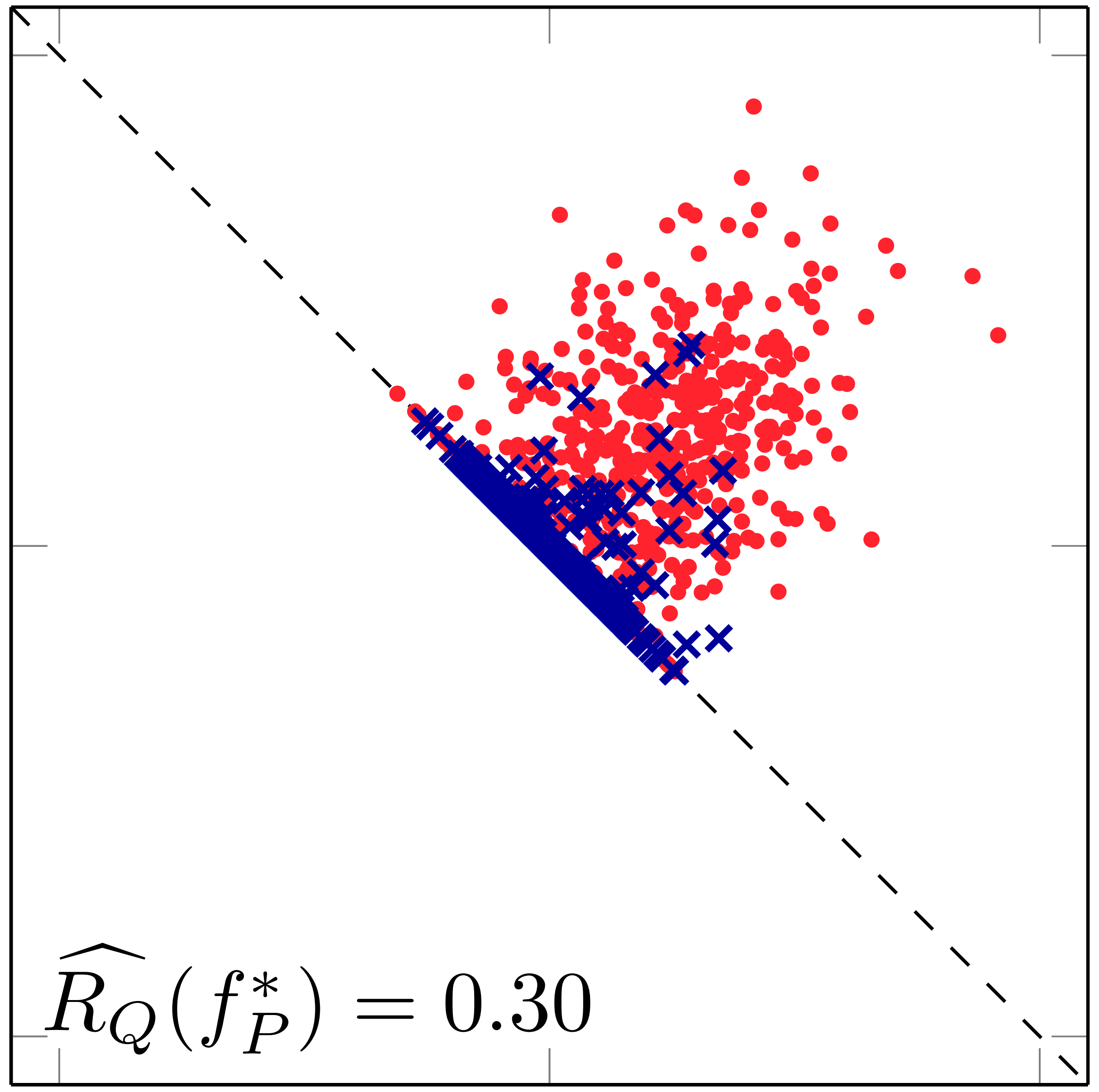
We assume that

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

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

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





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

$$R_Q(f_P^*) > R_P(f_P^*)$$

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 Compliant case,

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

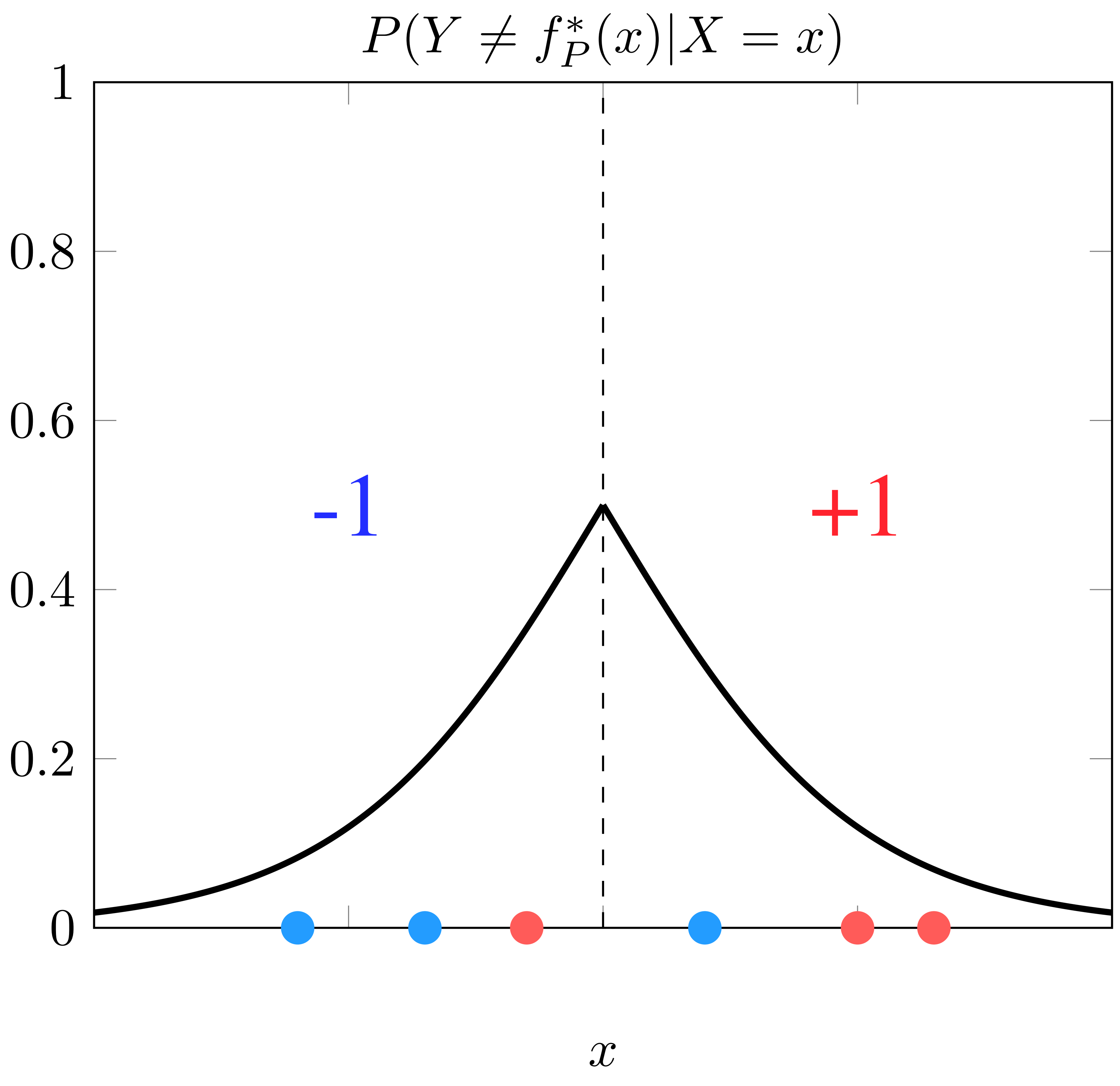
B. For the Defiant case,

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

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 $+1$

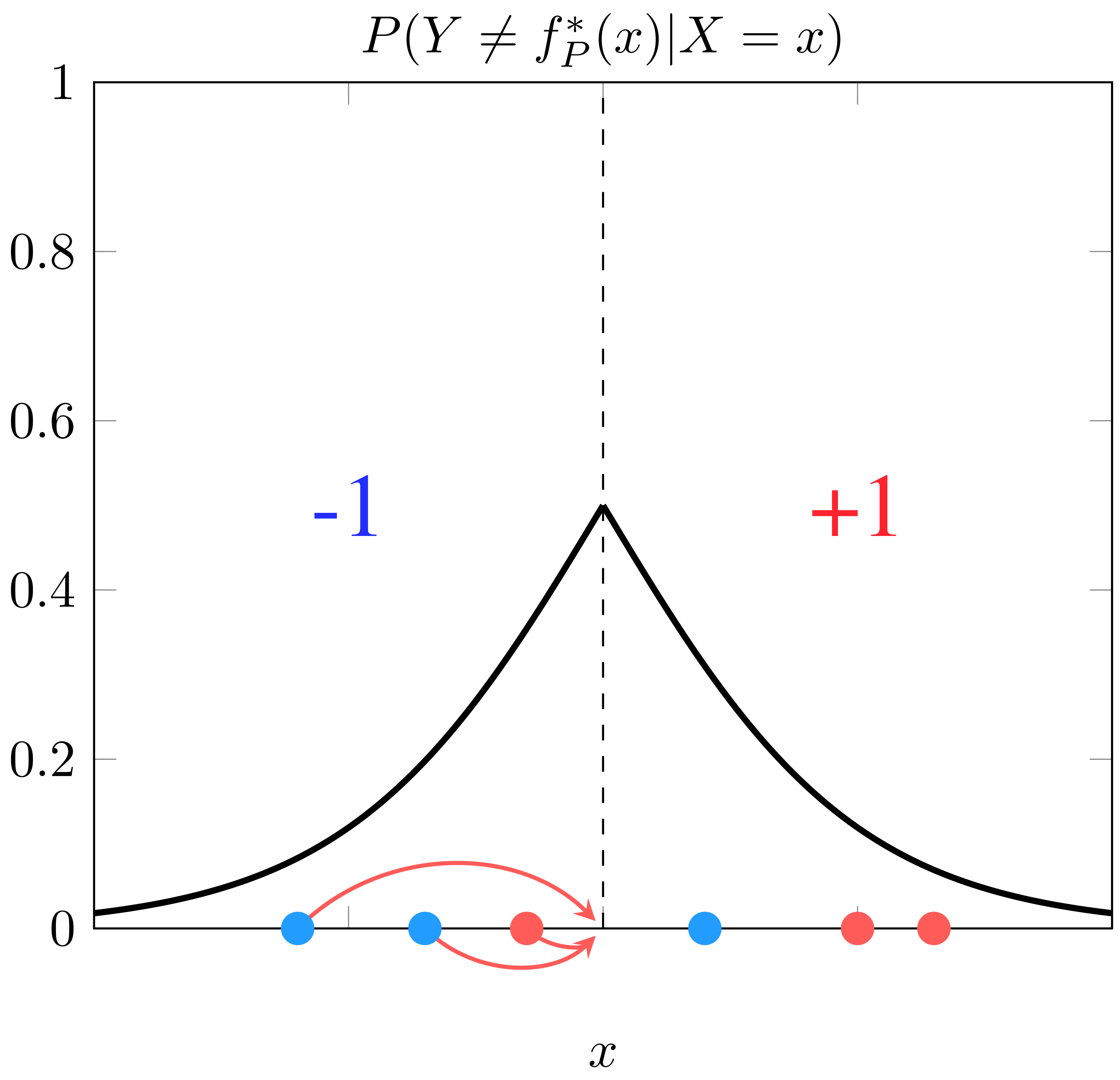
➡ The mistakes you make are

  ➡ Original $f_P^*(X_0) = +1$ but $Y = -1$,

Optimal Classifier

Proof sketch (Compliant)

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



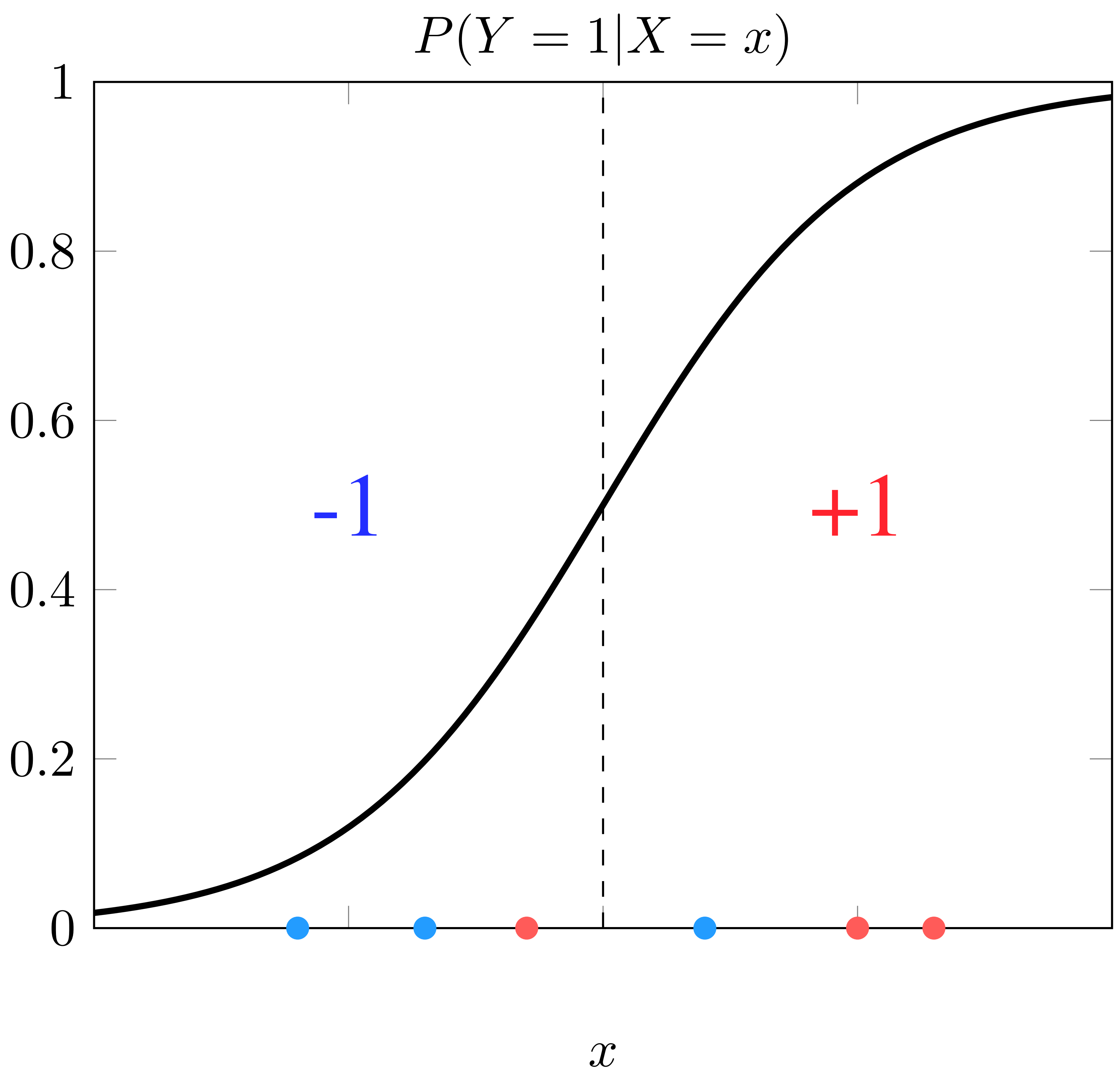
- Every point is now classified as $+1$
- The mistakes you make are
  - Original $f_P^*(X_0) = +1$ but $Y = -1$,
  - Half of the original $f_P^*(X_0) = -1$,
    - $P(Y = +1 \,|\, X) = P(Y = -1 \,|\, X) = \frac{1}{2}$

      on the decision boundary

Optimal Classifier

Proof sketch (Defiant)

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

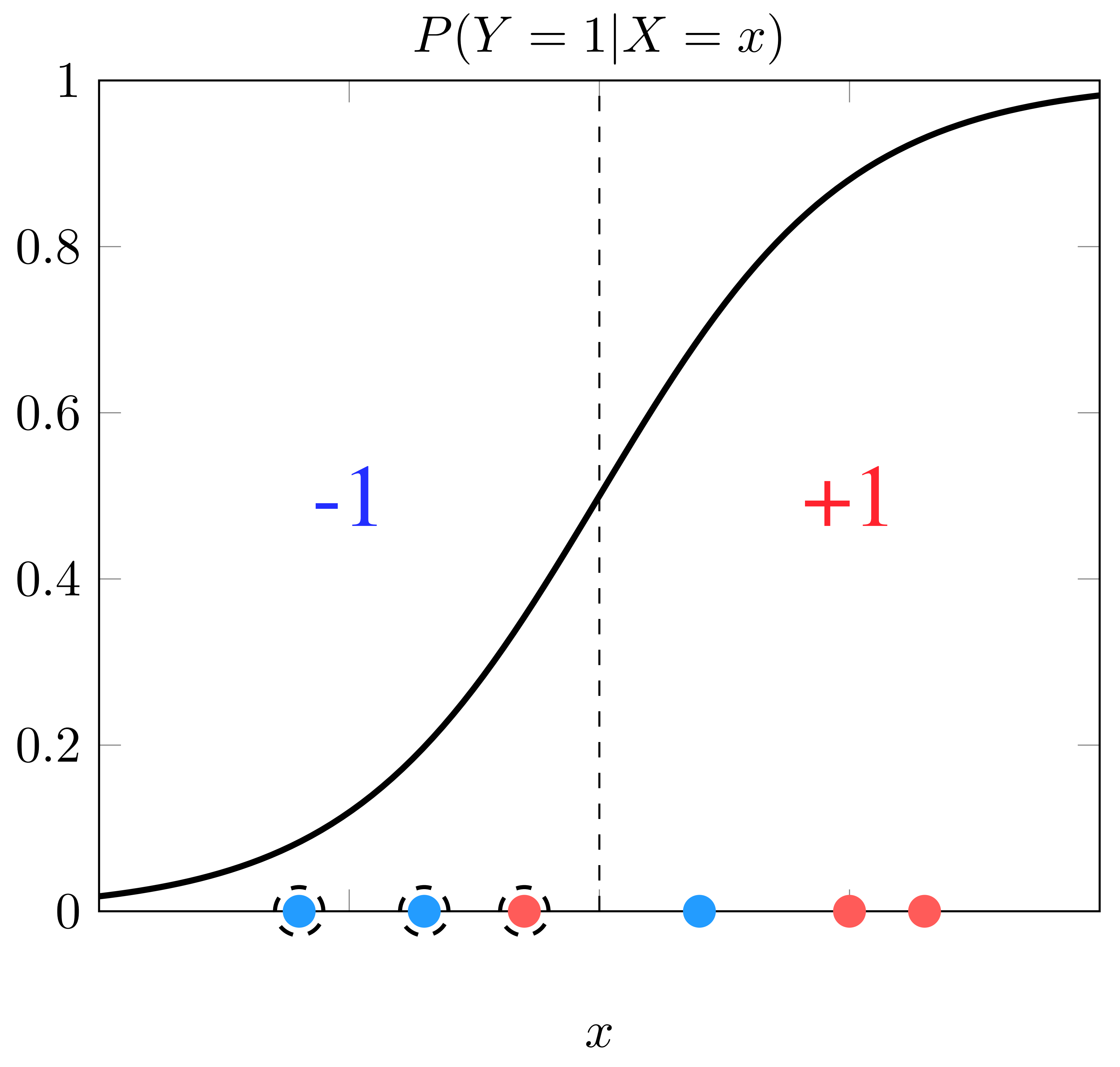


- ➡ Every point is now classified as $+1$
- ➡ The mistakes you make are
  - ➡ Original $f_P^*(X_0) = +1$ but $Y = -1$,

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

Rest of the paper

What else do we show

- Similar results/bounds for Non-Optimal classifiers
- What ML providers can do to strategise against this phenomenon
- More examples and empirical results

Thank you for your attention!