DIFFERENTIABLE LEARNING UNDER TRIAGE

Nastaran Okati, Abir De and Manuel Gomez-Rodriguez

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

IIT Bombay

WHAT IS ALGORITHMIC TRIAGE?

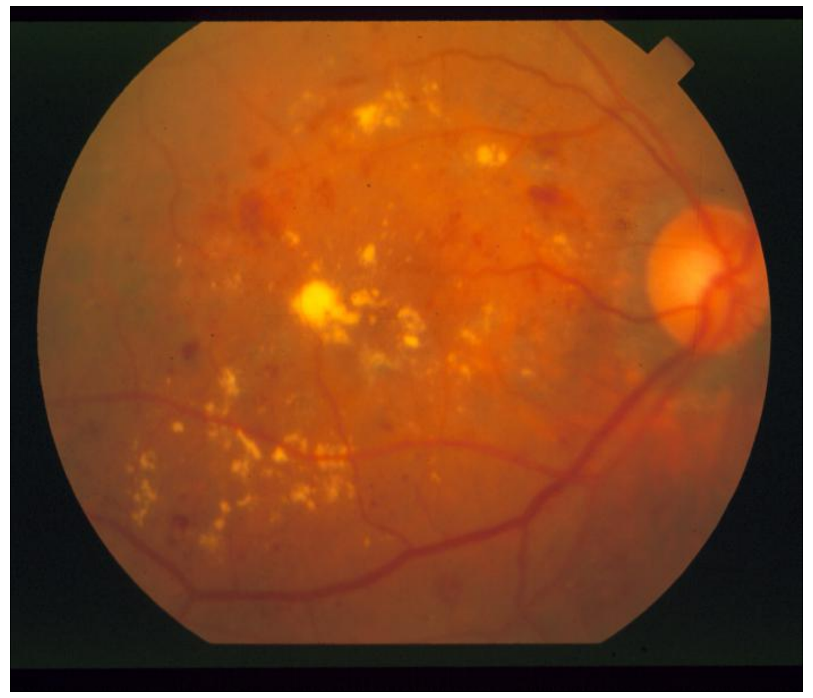
- Machine learning models have surpassed human performance on many tasks.
- There are still some cases on which human has better performance.
- **Algorithmic triage**: balance human and algorithmic predictions.

AI models from Microsoft and Google already surpass human performance on the SuperGLUE language benchmark

[https://venturebeat.com/]

Delving Deep into Rectifiers:
Surpassing Human-Level Performance on ImageNet Classification

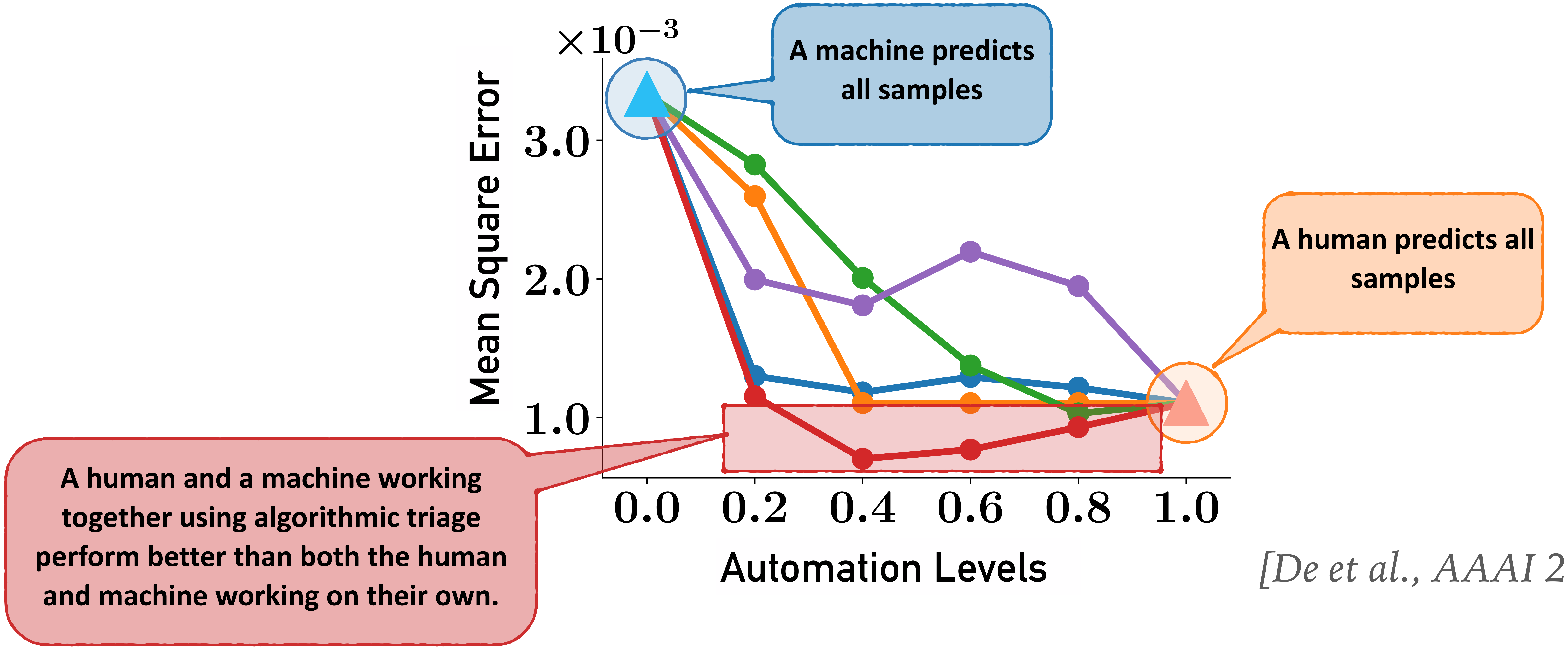
[He et al., CVPR 2015]



A difficult sample for machine due to the existence of yellow spots which suggest the presence of Drusen disease. In this case however, the yellow spots are due to diabetic retinopathy.

WHY ALGORITHMIC TRIAGE?

➤ By working together, **humans** and **machines** are likely to achieve a better performance than each of them on their own.



[De et al., AAAI 2020]

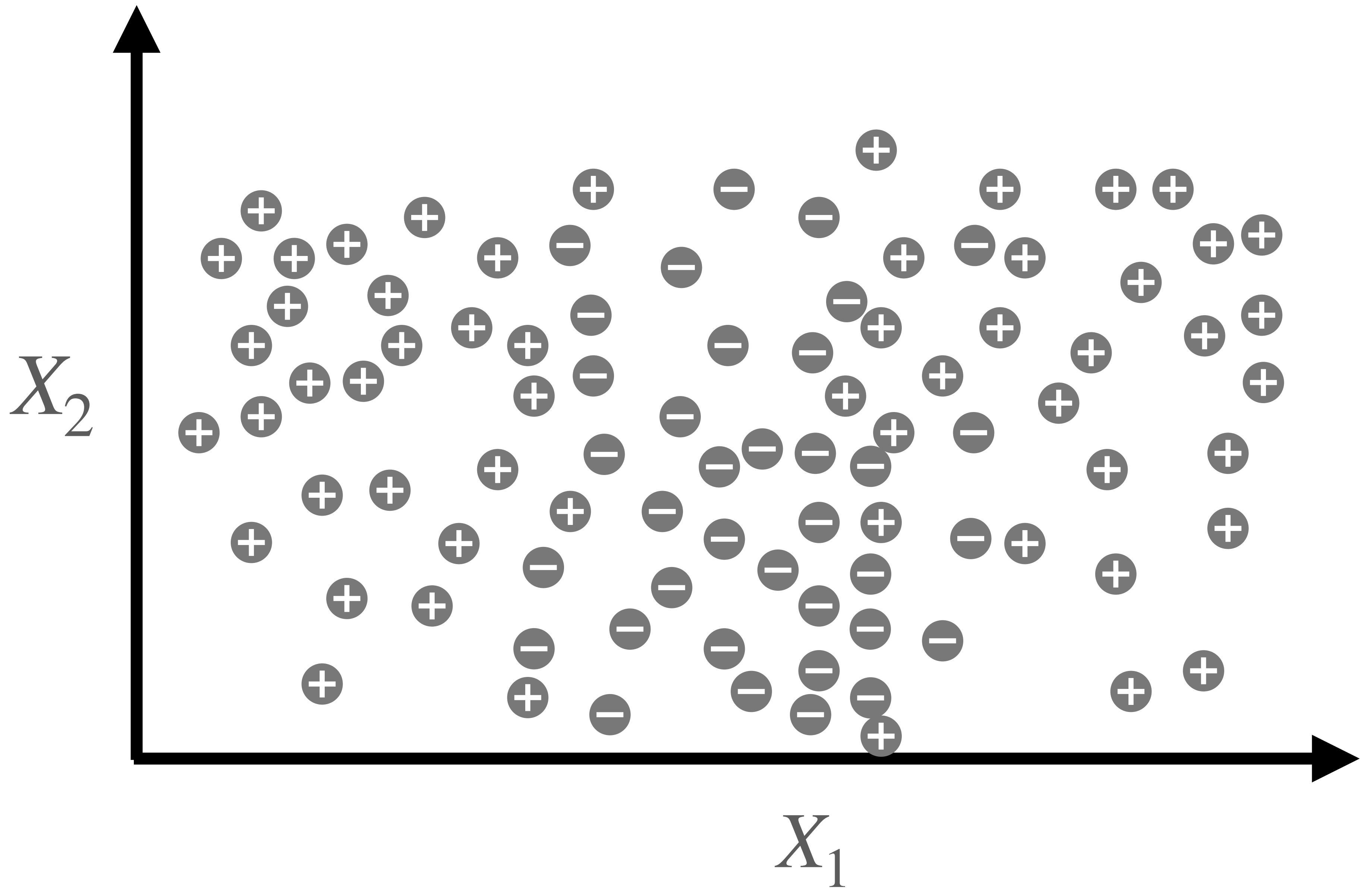
LEARNING UNDER ALGORITHMIC TRIAGE

- Optimize machine model to operate under **different automation levels**.
- A **triage policy** decides which samples should be assigned to the human and which ones should be assigned to the machine.

LEARNING UNDER ALGORITHMIC TRIAGE

- Optimize machine model to operate under **different automation levels**.
- A **triage policy** decides which samples should be assigned to the human and which ones should be assigned to the machine.

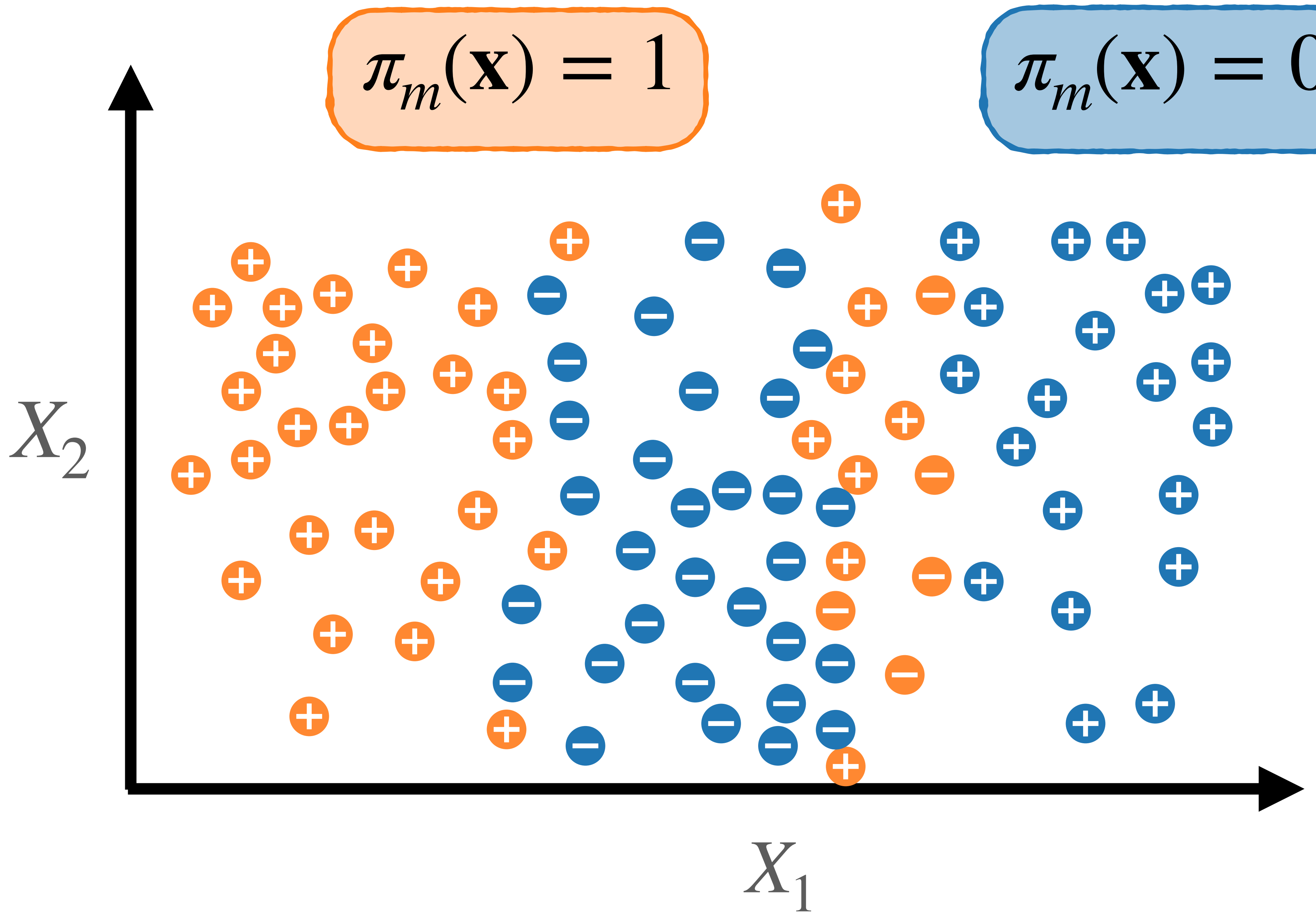
Example Task: Binary Classification. For simplicity, assume that human performs uniformly good across the feature space.



LEARNING UNDER ALGORITHMIC TRIAGE

- Optimize machine model to operate under **different automation levels**.
- A **triage policy** decides which samples should be assigned to the human and which ones should be assigned to the machine.

Triage policy determines which points should be assigned to the human and which ones should be assigned to the machine.

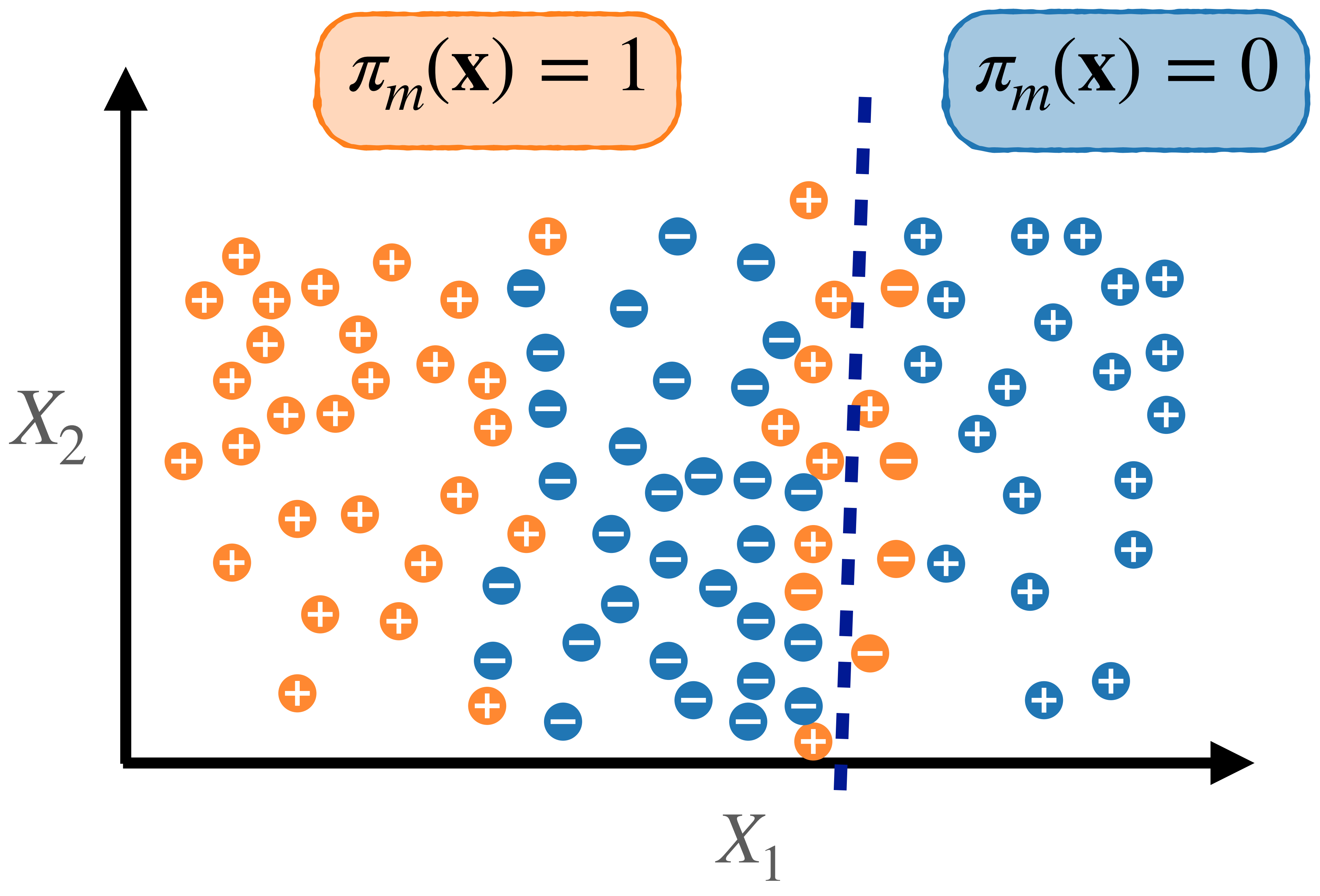


LEARNING UNDER ALGORITHMIC TRIAGE

- Optimize machine model to operate under **different automation levels**.
- A **triage policy** decides which samples should be assigned to the human and which ones should be assigned to the machine.

Triage policy determines which points should be assigned to the human and which ones should be assigned to the machine.

The machine is only trained on the samples for which $\pi_m(\mathbf{x}) = 0$.



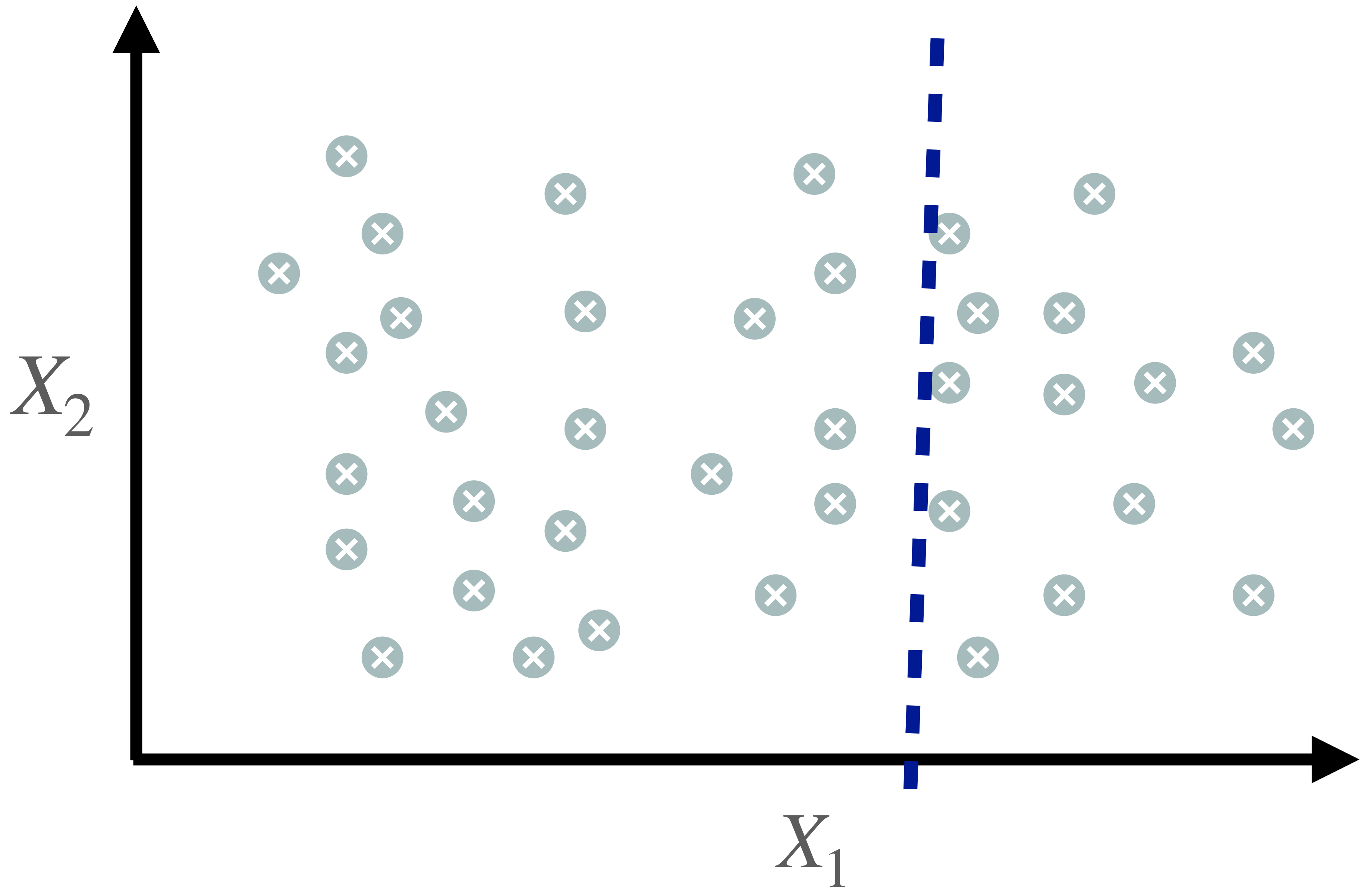
LEARNING UNDER ALGORITHMIC TRIAGE

- Optimize machine model to operate under **different automation levels**.
- A **triage policy** decides which samples should be assigned to the human and which ones should be assigned to the machine.

Triage policy determines which points should be assigned to the human and which ones should be assigned to the machine.

The machine is only trained on the samples for which $\pi_m(\mathbf{x}) = 0$.

An unseen sample is predicted by either the human or the machine.



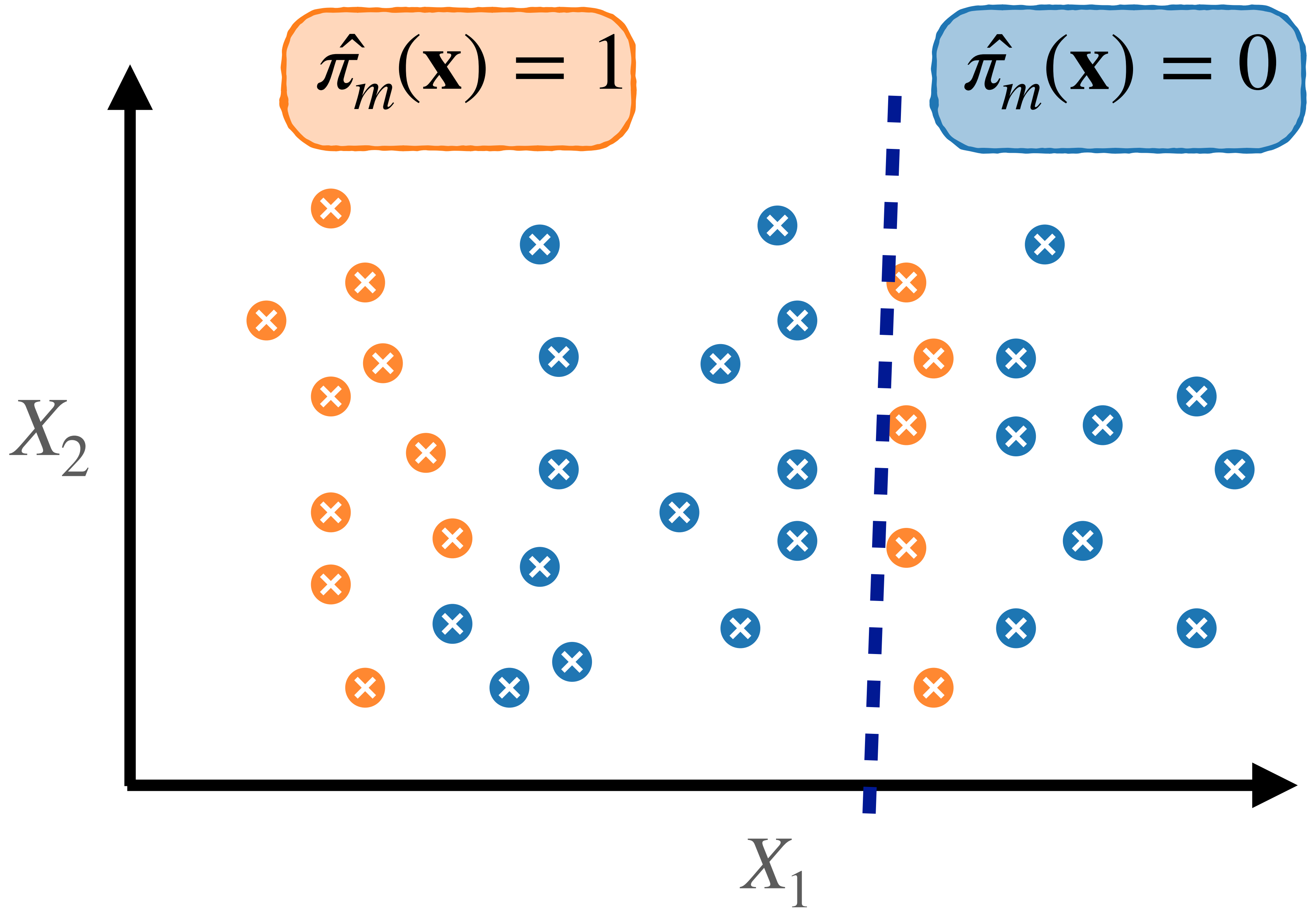
LEARNING UNDER ALGORITHMIC TRIAGE

- Optimize machine model to operate under **different automation levels**.
- A **triage policy** decides which samples should be assigned to the human and which ones should be assigned to the machine.

Triage policy determines which points should be assigned to the human and which ones should be assigned to the machine.

The machine is only trained on the samples for which $\pi_m(\mathbf{x}) = 0$.

An unseen sample is predicted by either the human or the machine.



OUR GOAL IS TO ANSWER THE FOLLOWING:

When is algorithmic triage beneficial?

OUR GOAL IS TO ANSWER THE FOLLOWING:

When is algorithmic triage beneficial?

How does the accuracy of the predictive model and the human affect the triage policy?

OUR GOAL IS TO ANSWER THE FOLLOWING:

When is algorithmic triage beneficial?

How does the accuracy of the predictive model and the human affect the triage policy?

What is the optimal triage policy?

OUR GOAL IS TO ANSWER THE FOLLOWING:

When is algorithmic triage beneficial?

How does the accuracy of the predictive model and the human affect the triage policy?

What is the optimal triage policy?

When is a machine trained under full automation suboptimal given a desired level of triage?

OUR GOAL IS TO ANSWER THE FOLLOWING:

When is algorithmic triage beneficial?

How does the accuracy of the predictive model and the human affect the triage policy?

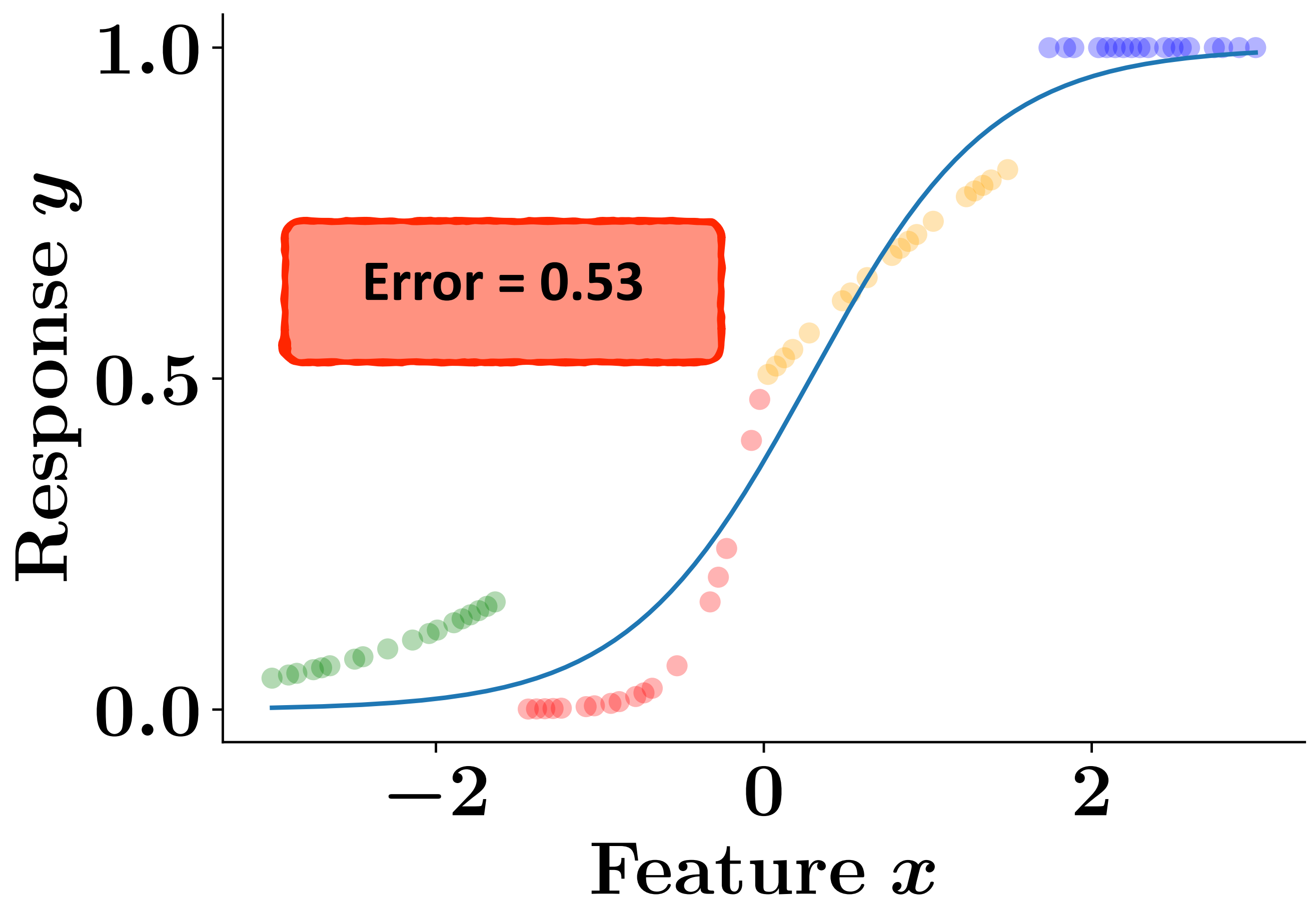
What is the optimal triage policy?

When is a machine trained under full automation suboptimal given a desired level of triage?

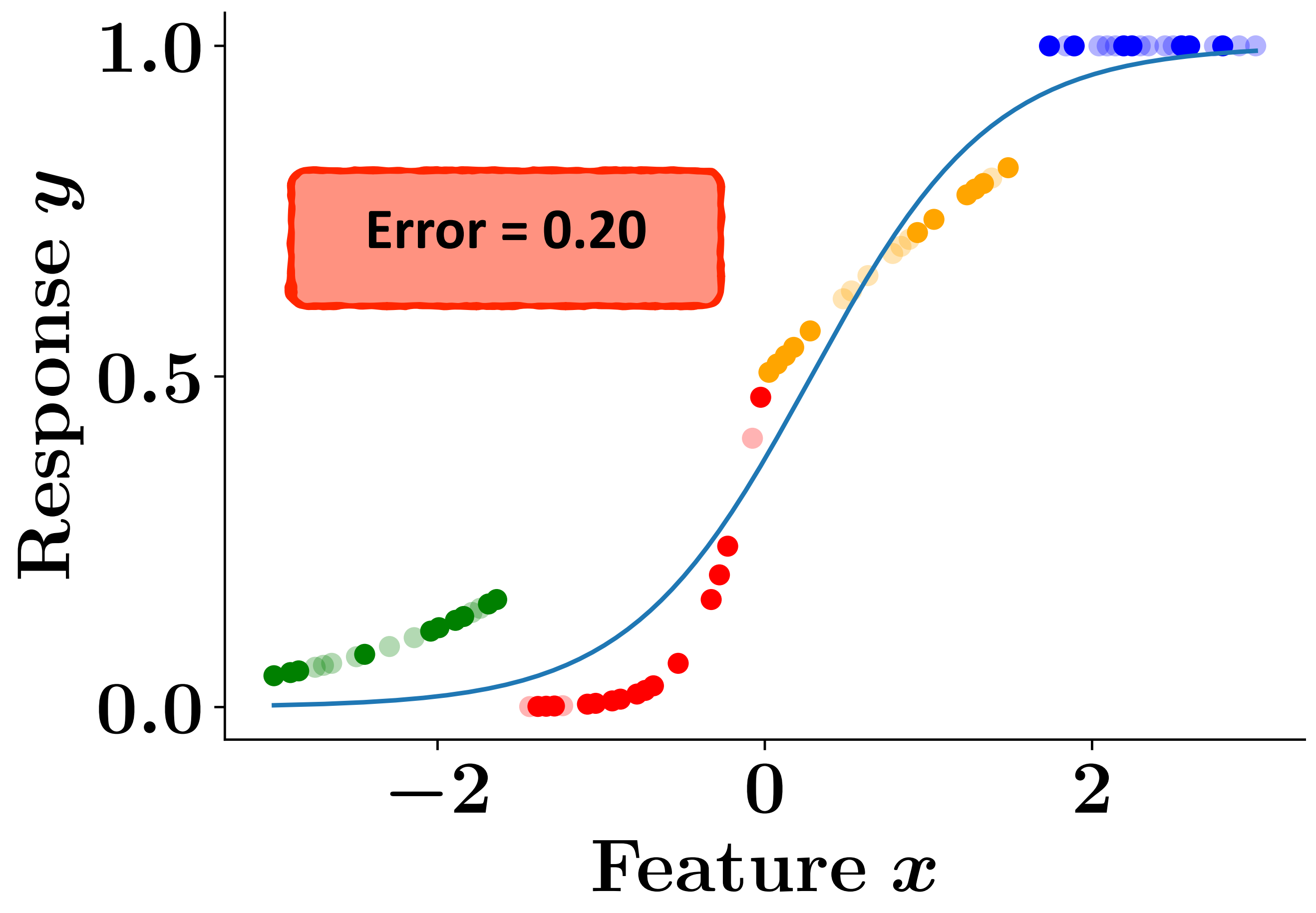
Is there a scalable algorithm for optimizing the machine under algorithmic triage?

ALGORITHMIC TRIAGE IS BENEFICIAL!

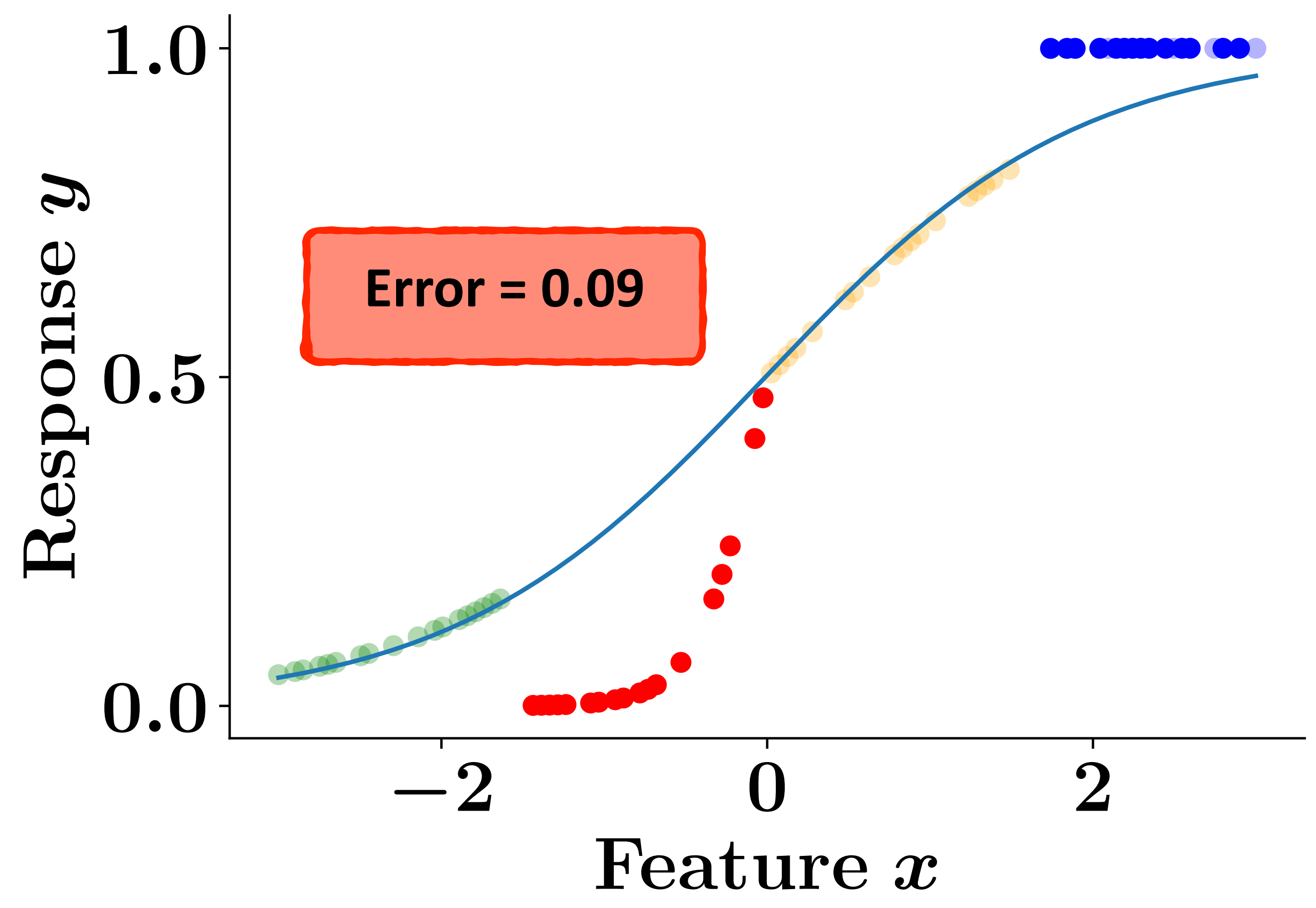
The machine is trained under full automation. No triage at test time.



The machine is trained under full automation. Triage is used at test time.

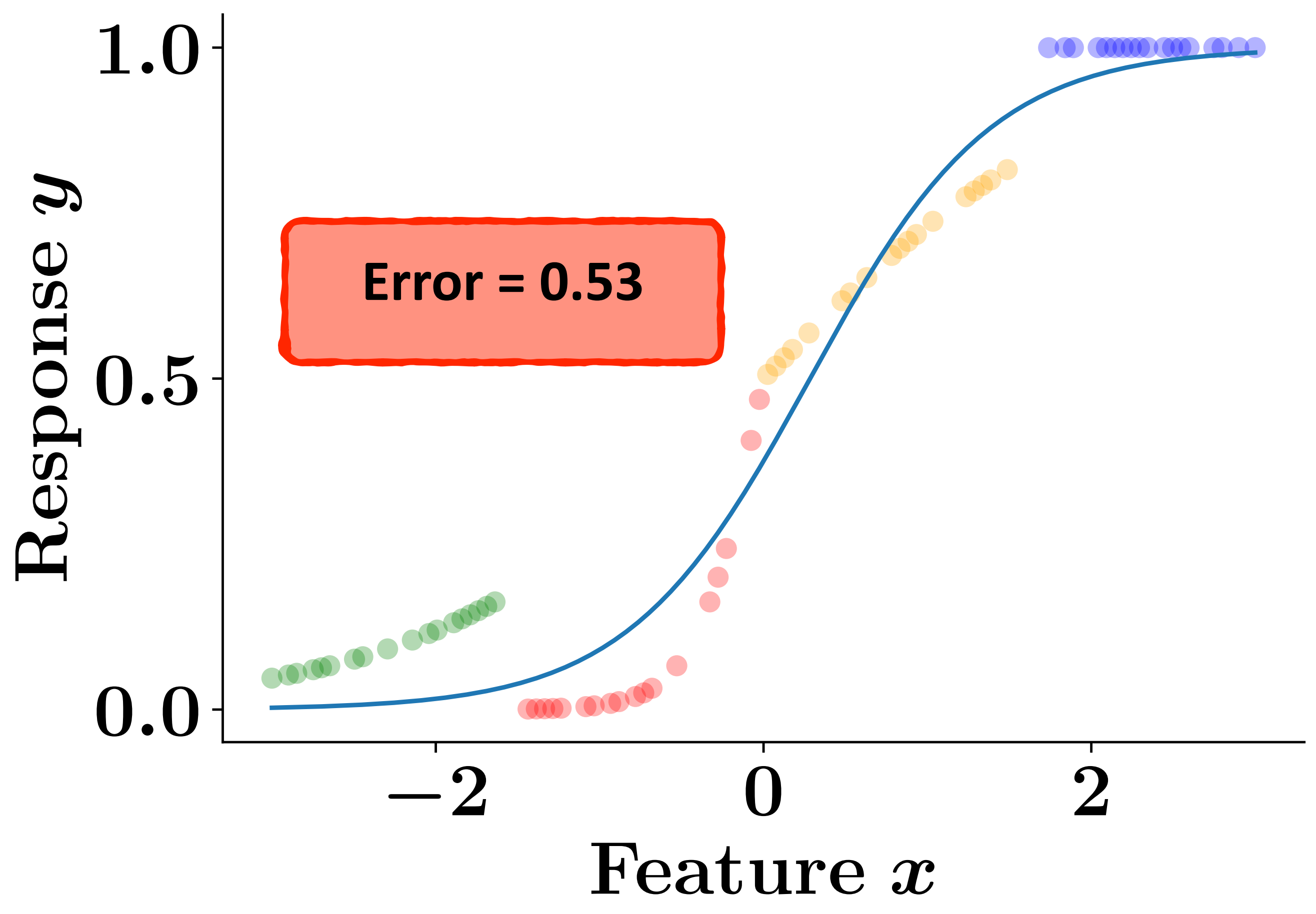


The machine is trained on the points determined by the optimal triage policy. Triage is used at test time.



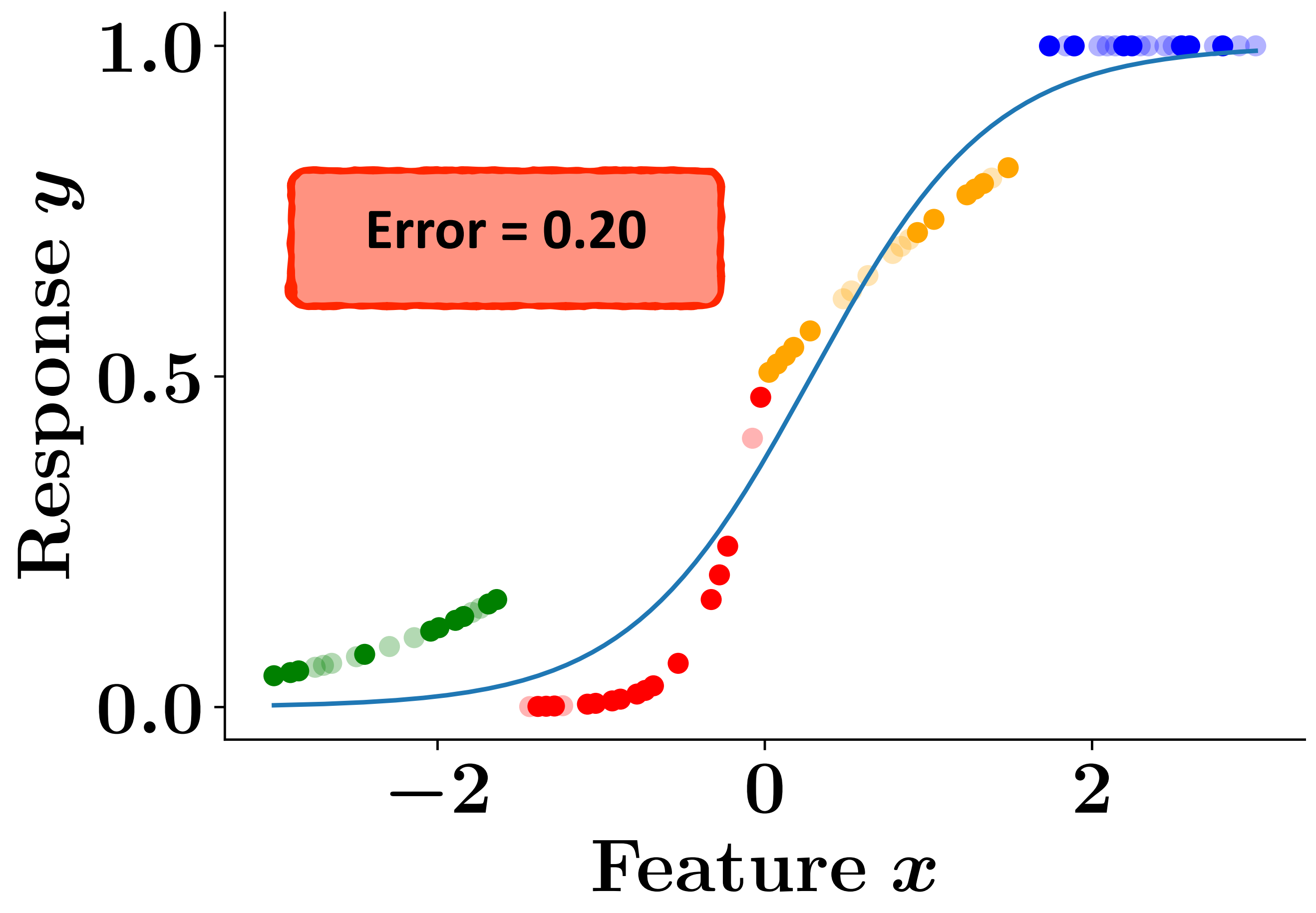
ALGORITHMIC TRIAGE IS BENEFICIAL!

The machine is trained under full automation. No triage at test time.



Error reduced by 62%!

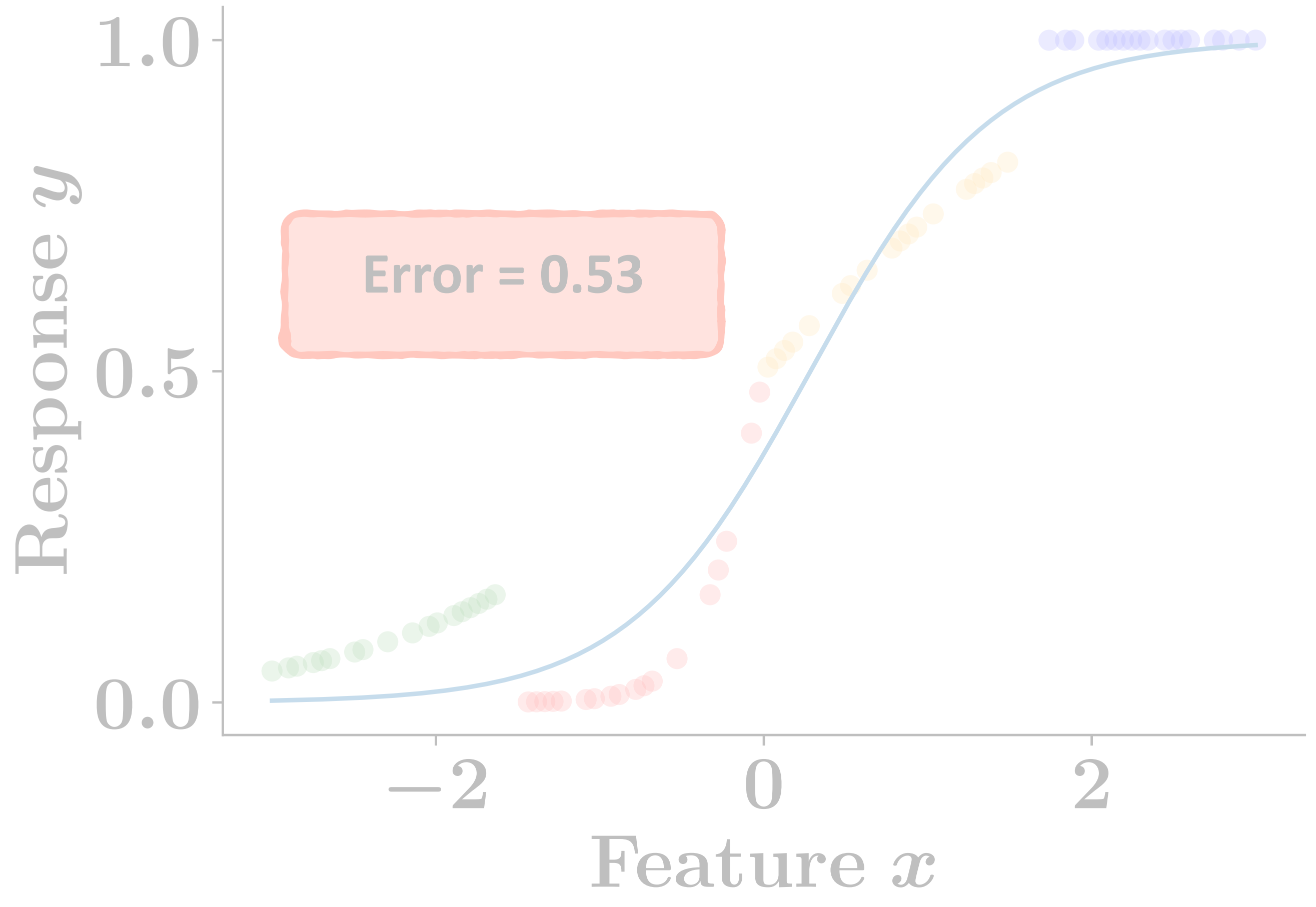
The machine is trained under full automation. Triage is used at test time.



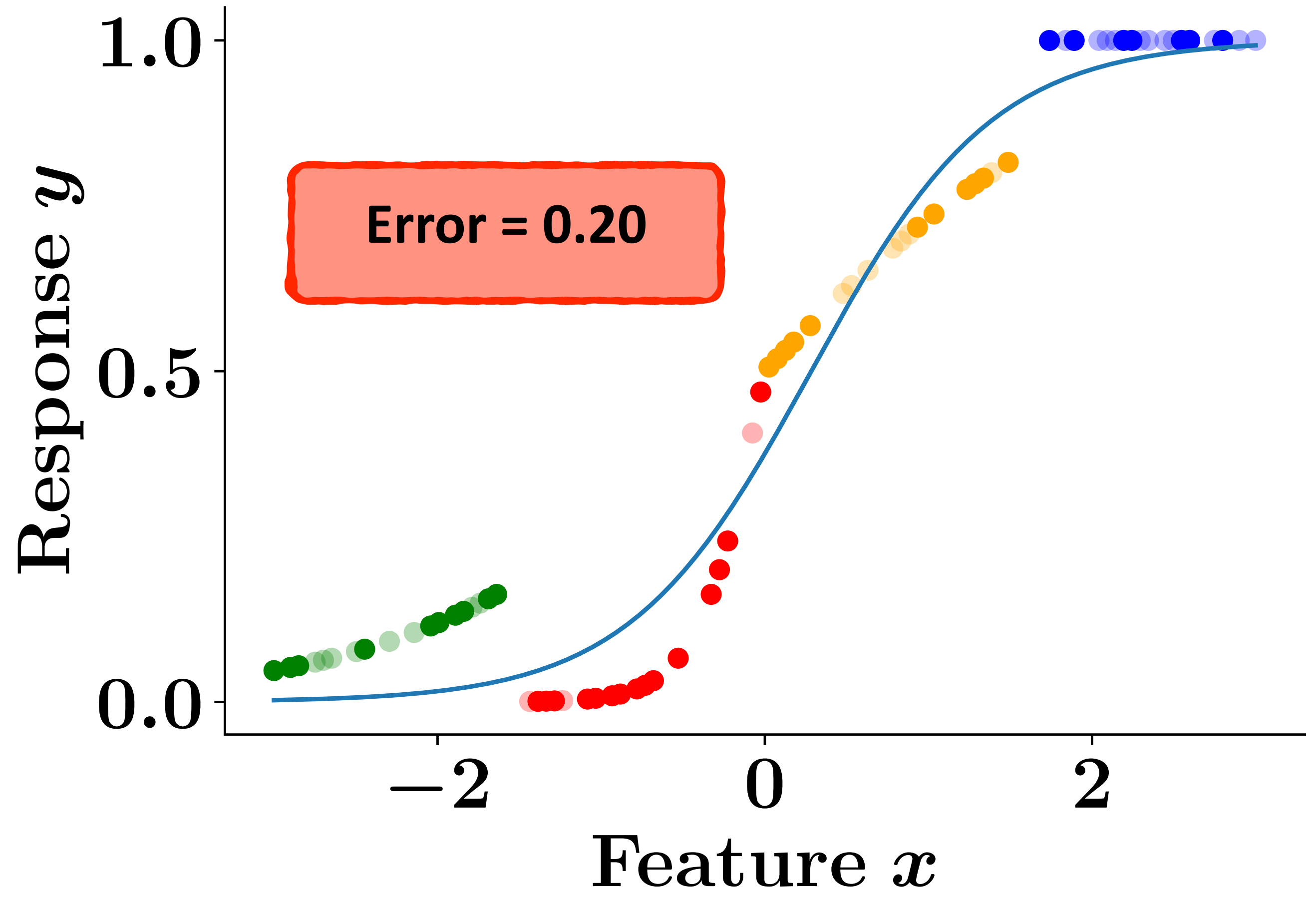
The machine is trained on the points determined by the optimal triage policy. Triage is used at test time.

ALGORITHMIC TRIAGE IS BENEFICIAL!

The machine is trained under full automation. No triage at test time.

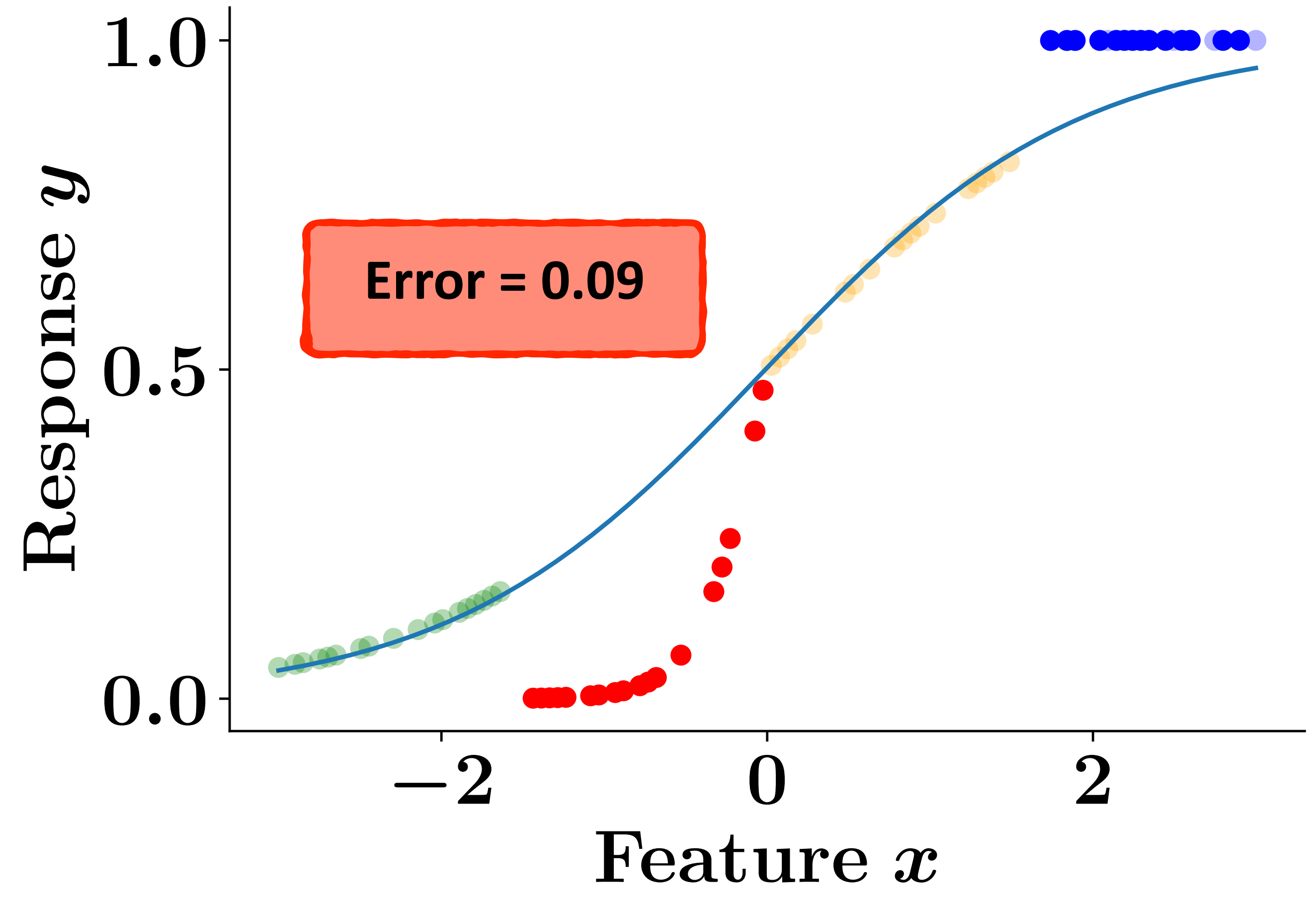


The machine is trained under full automation. Triage is used at test time.



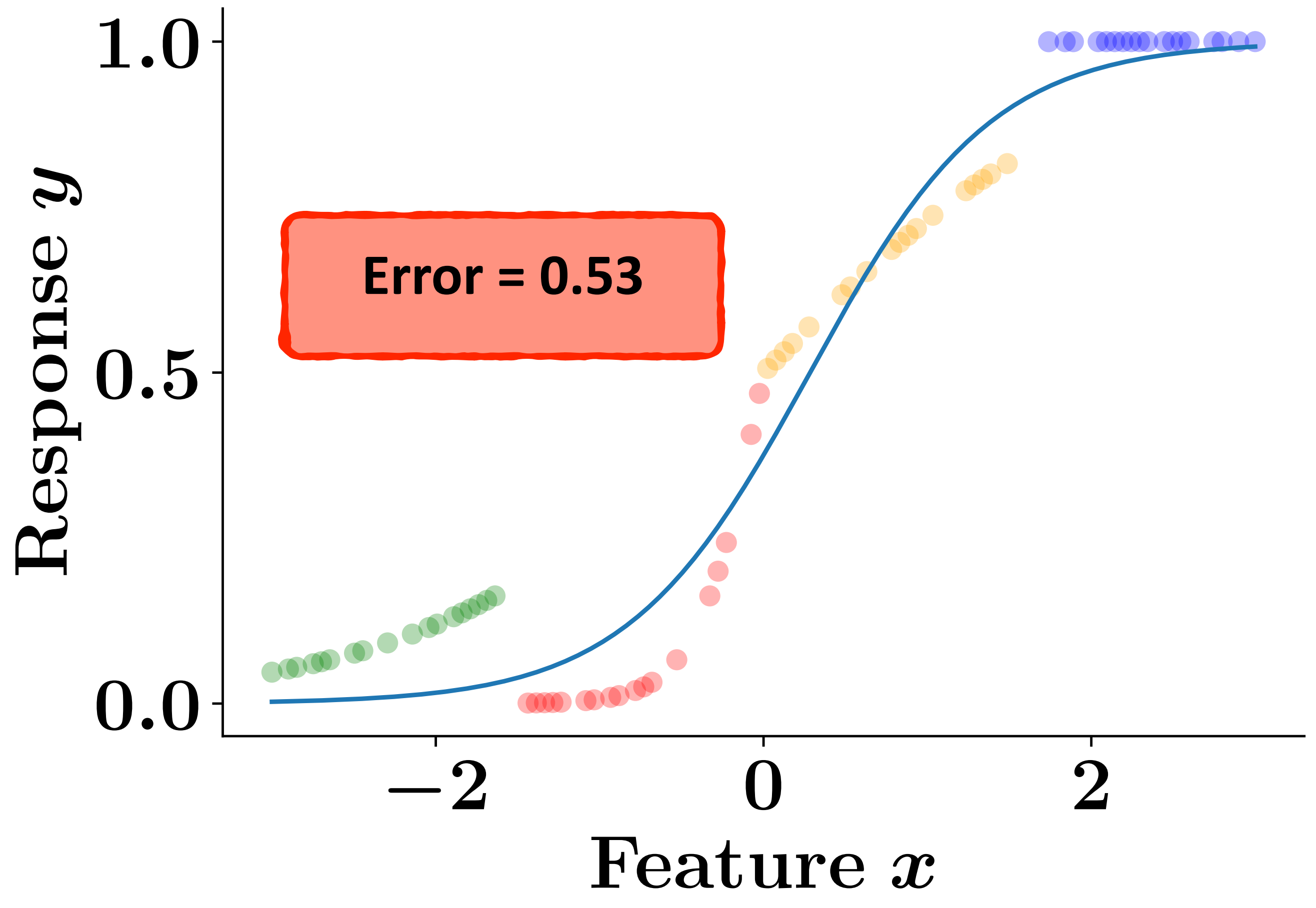
Error reduced by 55%!

The machine is trained on the points determined by the optimal triage policy. Triage is used at test time.



ALGORITHMIC TRIAGE IS BENEFICIAL!

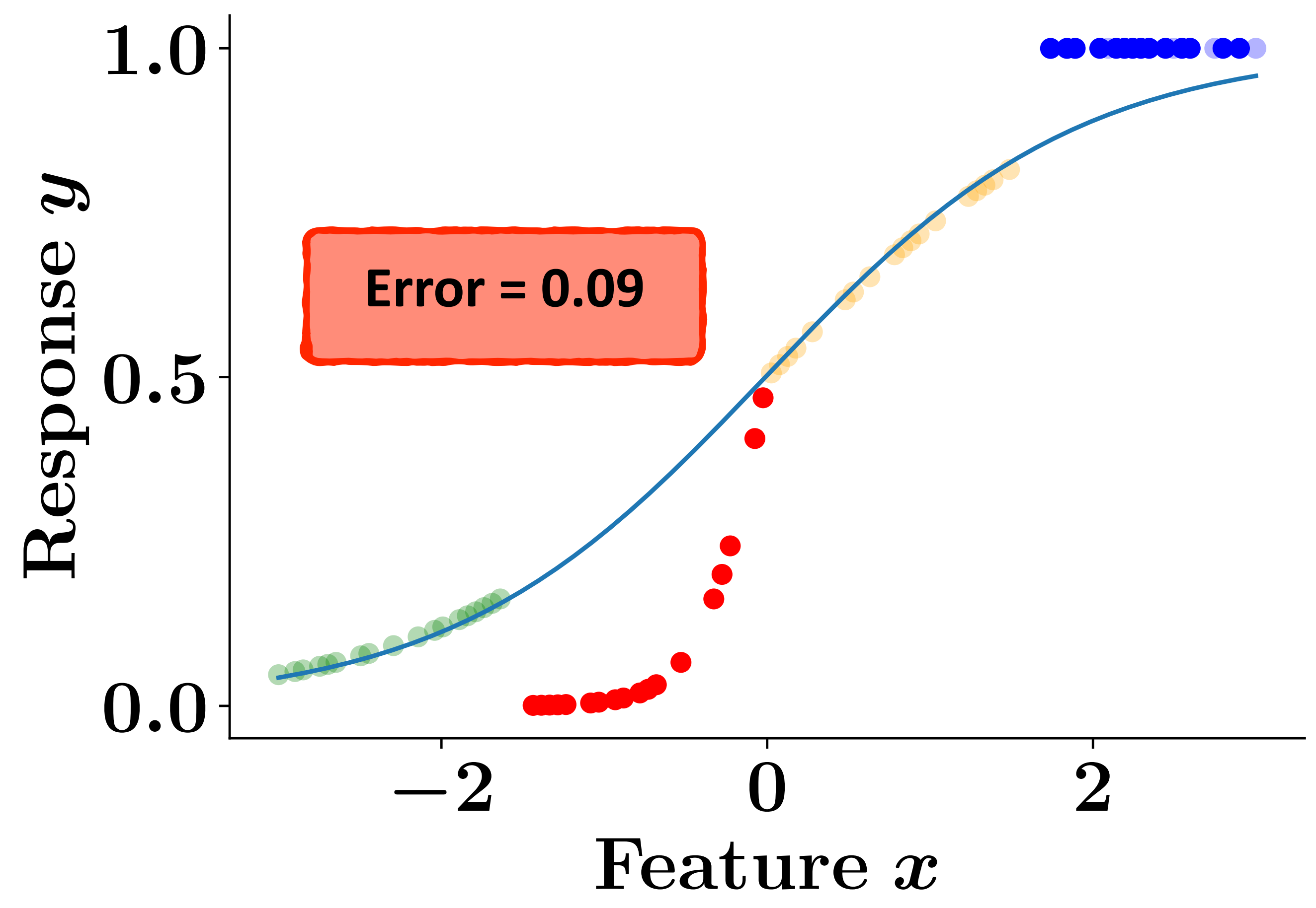
The machine is trained under full automation. No triage at test time.



Error reduced by 83%!

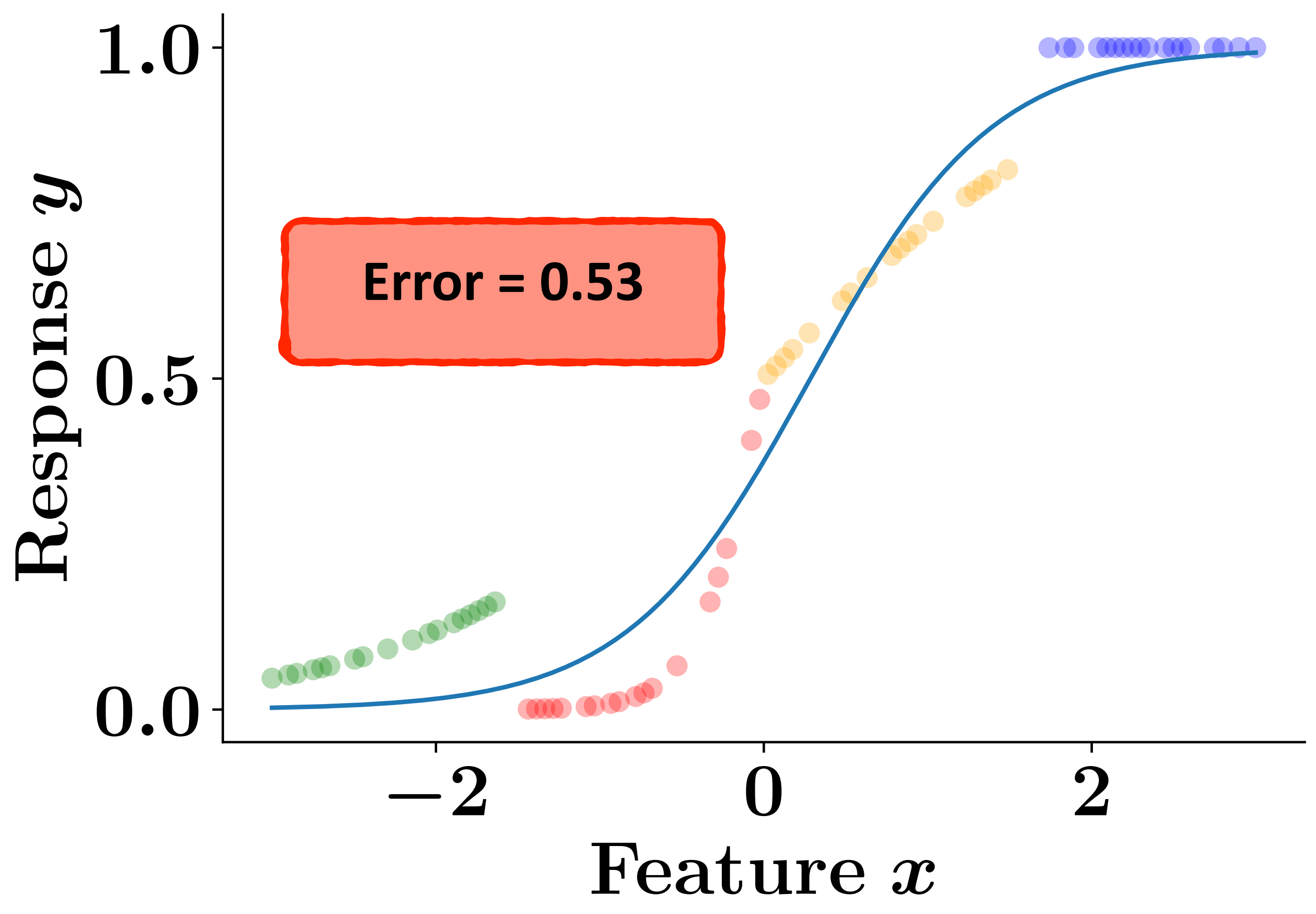
The machine is trained under full automation. Triage is used at test time.

The machine is trained on the points determined by the optimal triage policy. Triage is used at test time.



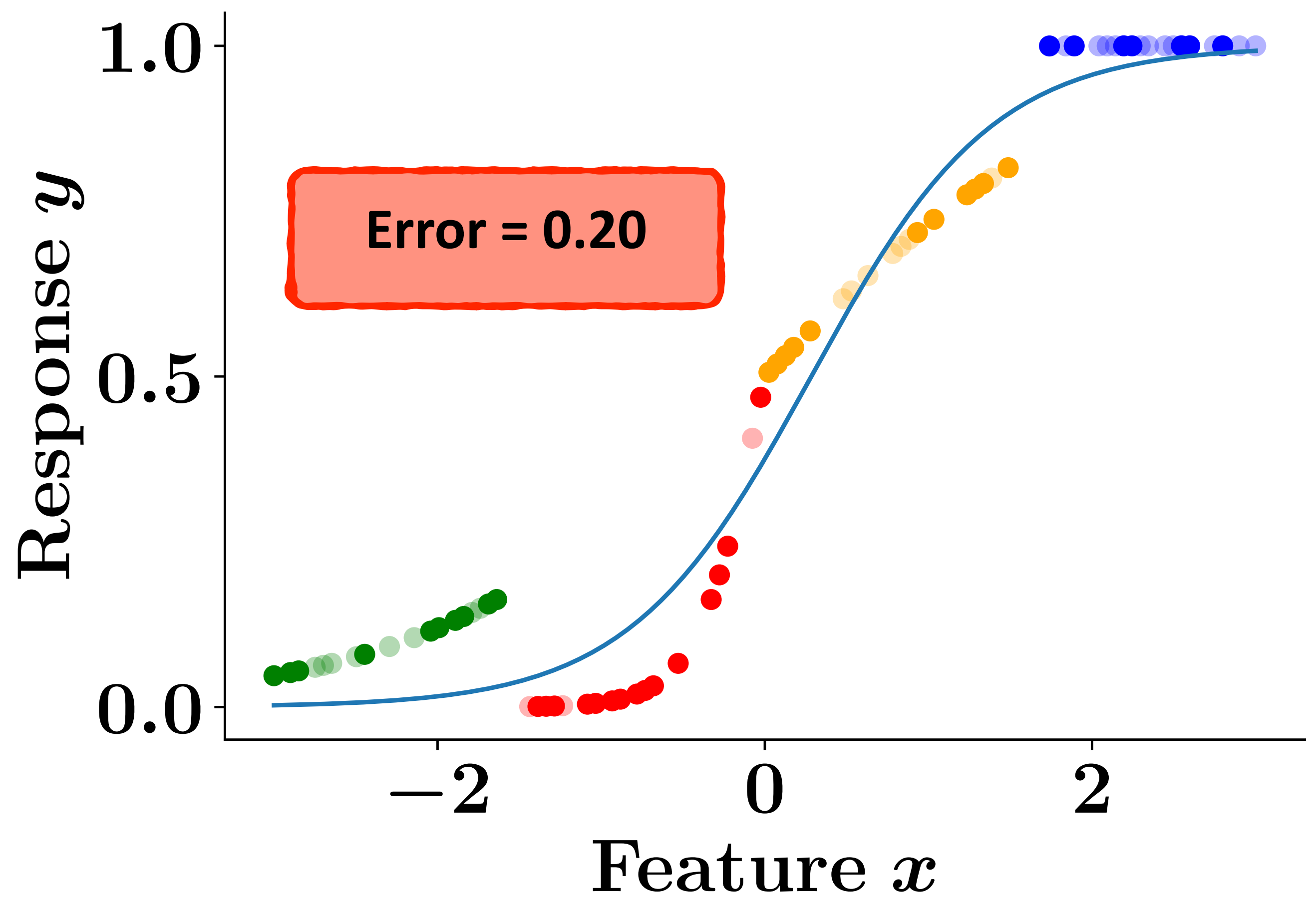
ALGORITHMIC TRIAGE IS BENEFICIAL!

The machine is trained under full automation. No triage at test time.



As long as there exists a subset of samples on which human has better performance than machine, triage is beneficial! (Proposition 2)

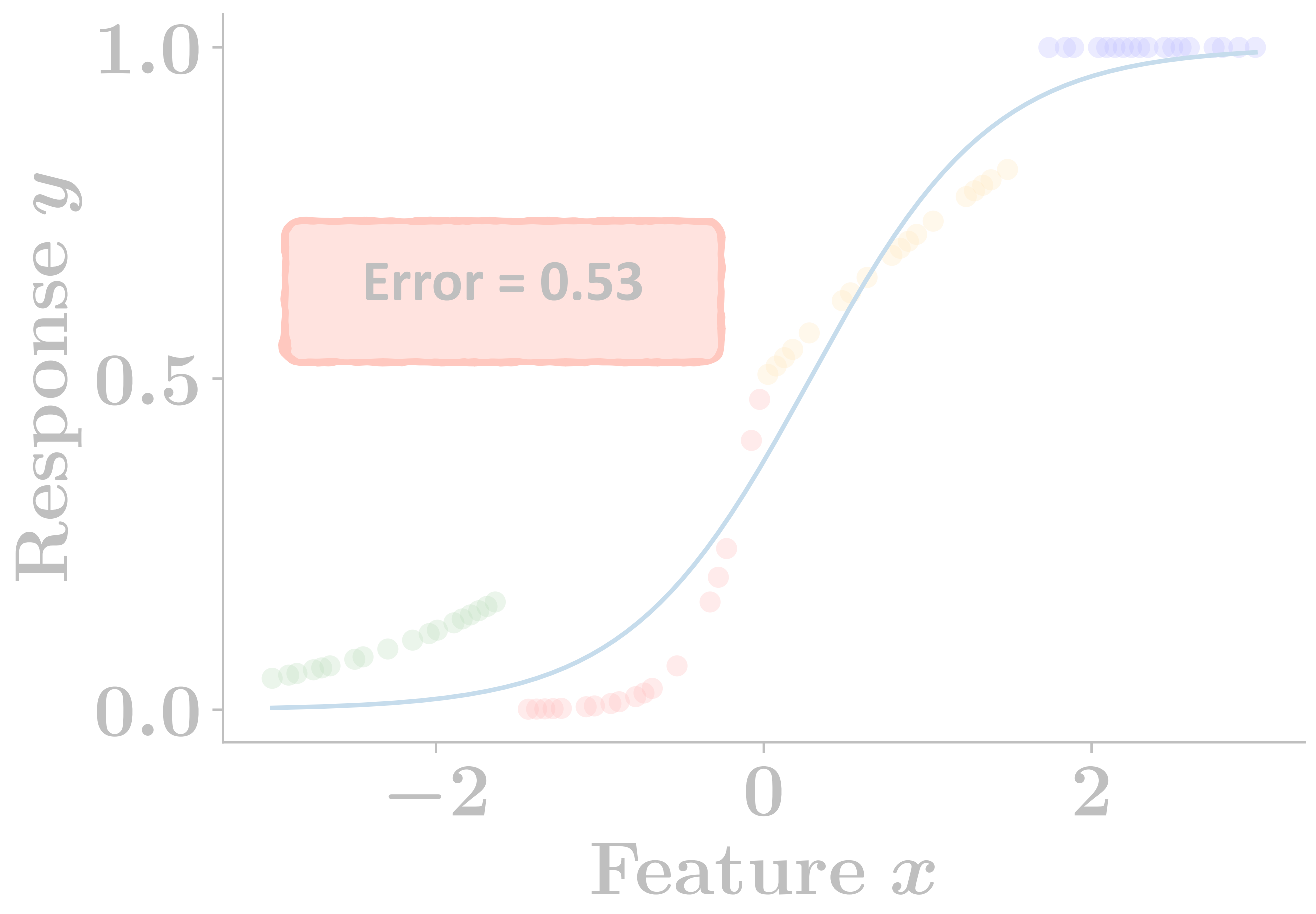
The machine is trained under full automation. Triage is used at test time.



The machine is trained on the points determined by the optimal triage policy. Triage is used at test time.

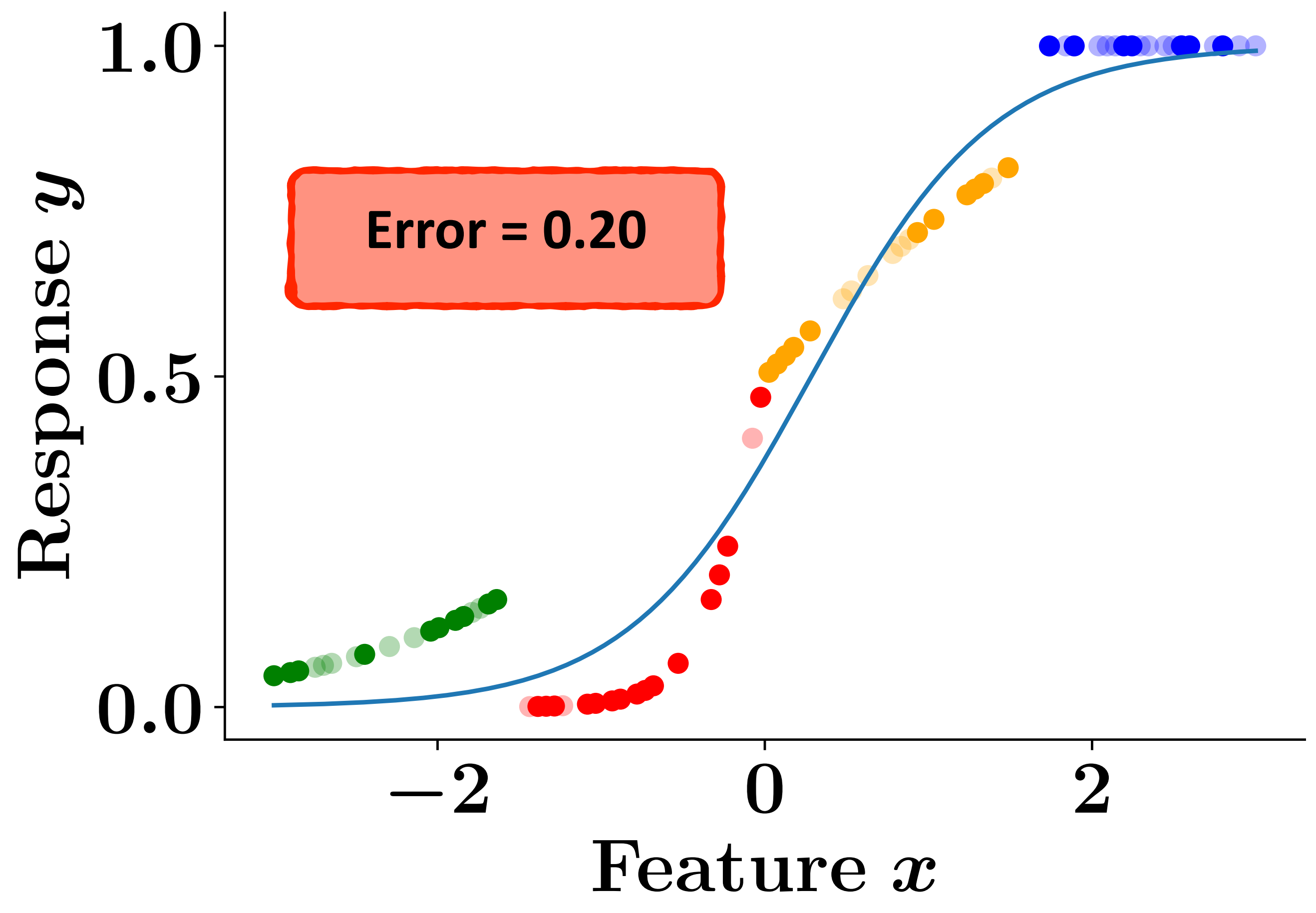
ALGORITHMIC TRIAGE IS BENEFICIAL!

The machine is trained under full automation. No triage at test time.



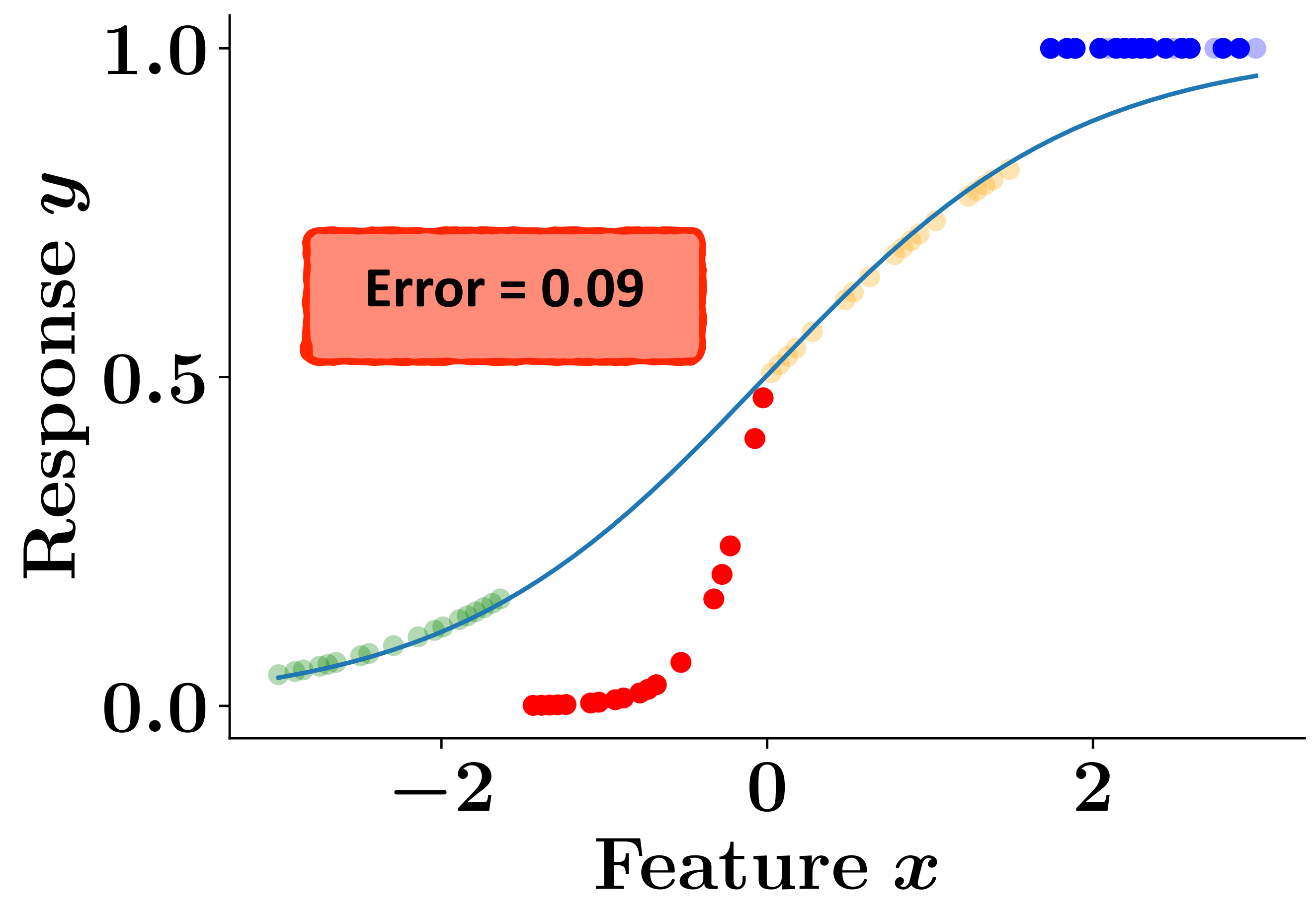
As long as there exists a subset of samples on which human has better performance than machine, triage is beneficial! (Proposition 2)

The machine is trained under full automation. Triage is used at test time.



As long as the machine trained under full automation is not able to perfectly predict the points which are assigned to human, we benefit from training and testing the machine model under triage (Proposition 4).

The machine is trained on the points determined by the optimal triage policy. Triage is used at test time.



SUPERVISED LEARNING UNDER TRIAGE

If $\pi(x) = 0$, assigned to machine

If $\pi(x) = 1$, assigned to human

$$\underset{\pi \in \Pi, m \in \mathcal{M}}{\text{minimize}} \quad L(\pi, m) = \mathbb{E}_{\mathbf{x},y,h}\left[(1 - \pi(\mathbf{x}))\, \ell(m(\mathbf{x}), y) + \pi(\mathbf{x})\, \ell(h, y)\right] \quad \text{subject to } \mathbb{E}_x[\pi(x)] \leq b$$

Set of all triage policies

Set of all predictive models

Machine loss per sample

Human loss per sample

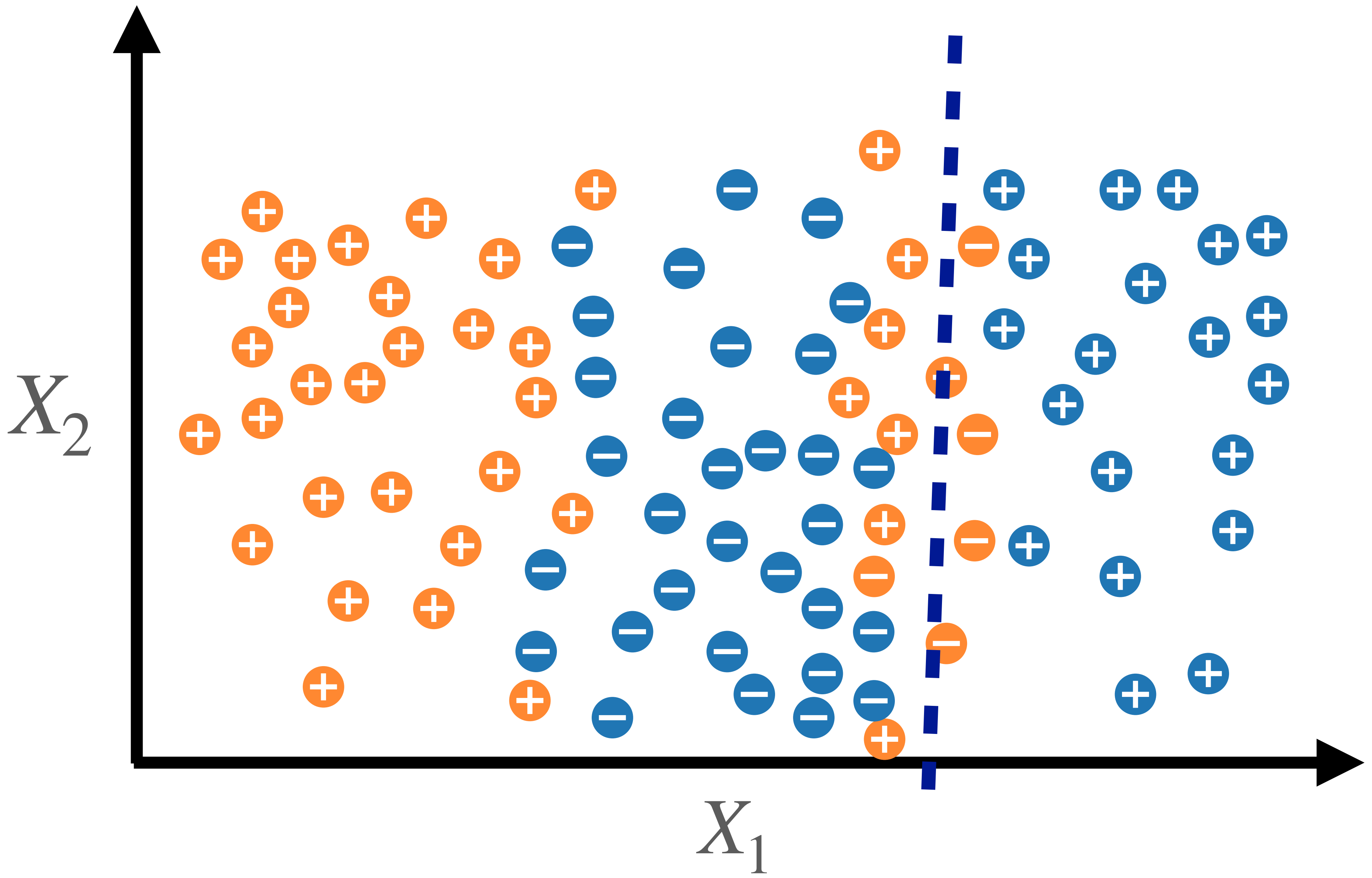
Maximum automation level

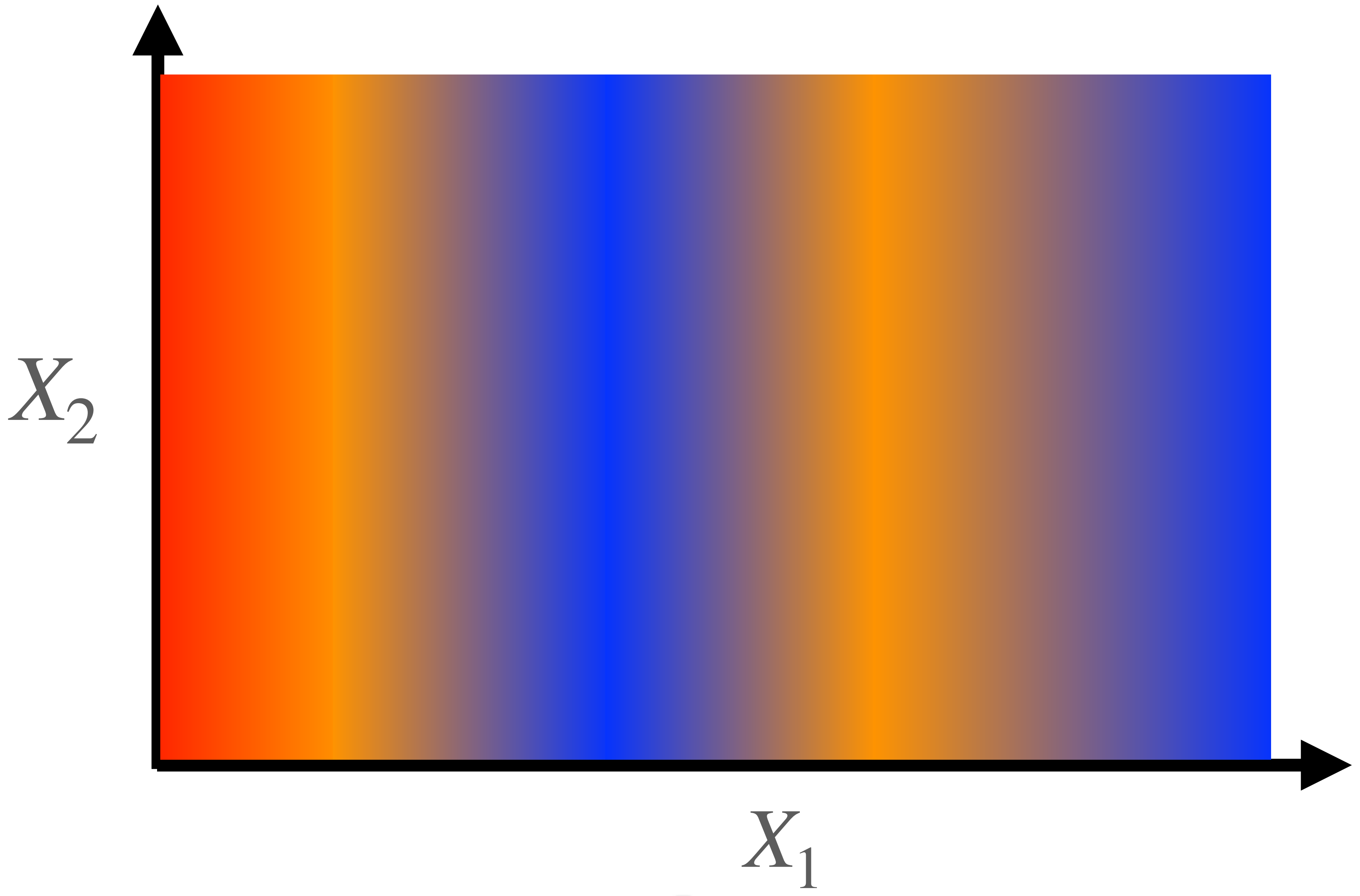
- $\pi(\mathbf{x}) : \mathscr{X} \to \{0,1\}$: the triage policy
- $m(\mathbf{x}) : \mathscr{X} \to \mathscr{Y}$: the predictive model

WHAT IS THE OPTIMAL TRIAGE POLICY?

- A **deterministic threshold rule** on the difference between the model and human loss on a per instance level:

$$\pi^*_{m,b}(\mathbf{x}) = \begin{cases} 1 & \textit{if } \mathbb{E}_{y|\mathbf{x}}\left[\ell(m(\mathbf{x}), y) - \mathbb{E}_{h|\mathbf{x}}\left[\ell(h, y)\right]\right] > t_{P,b,m} \\ 0 & \textit{otherwise,} \end{cases}$$





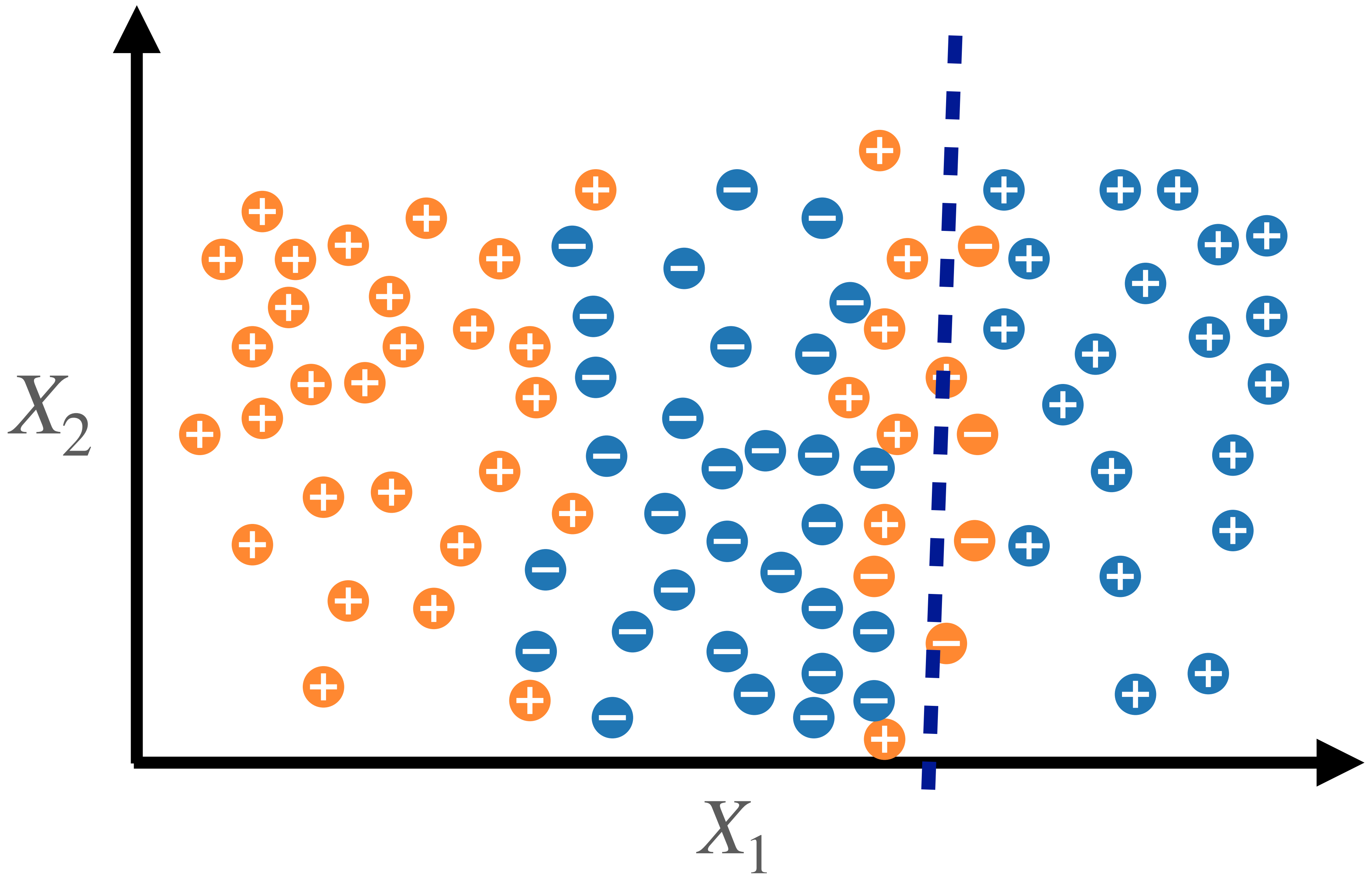
Heatmap showing $\mathbb{E}_{y|\mathbf{x}}\left[\ell(m(\mathbf{x}), y) - \mathbb{E}_{h|\mathbf{x}}\left[\ell(h, y)\right]\right]$

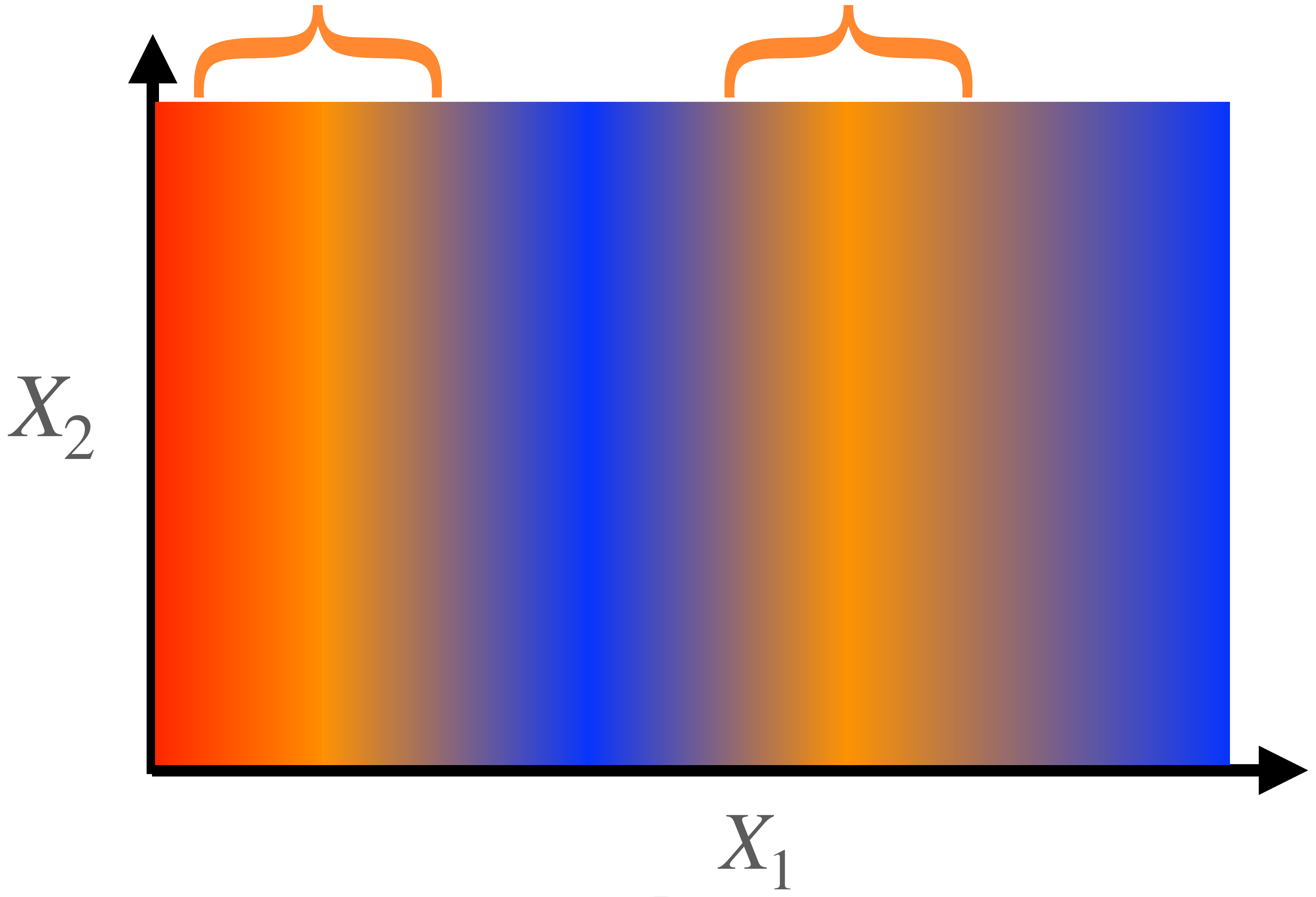
WHAT IS THE OPTIMAL TRIAGE POLICY?

- A **deterministic threshold rule** on the difference between the model and human loss on a per instance level:

$$\pi^*_{m,b}(\mathbf{x}) = \begin{cases} 1 & \textit{if } \mathbb{E}_{y|\mathbf{x}}\left[\ell(m(\mathbf{x}), y) - \mathbb{E}_{h|\mathbf{x}}\left[\ell(h, y)\right]\right] > t_{P,b,m} \\ 0 & \textit{otherwise,} \end{cases}$$

$\pi_m(\mathbf{x}) = 1$



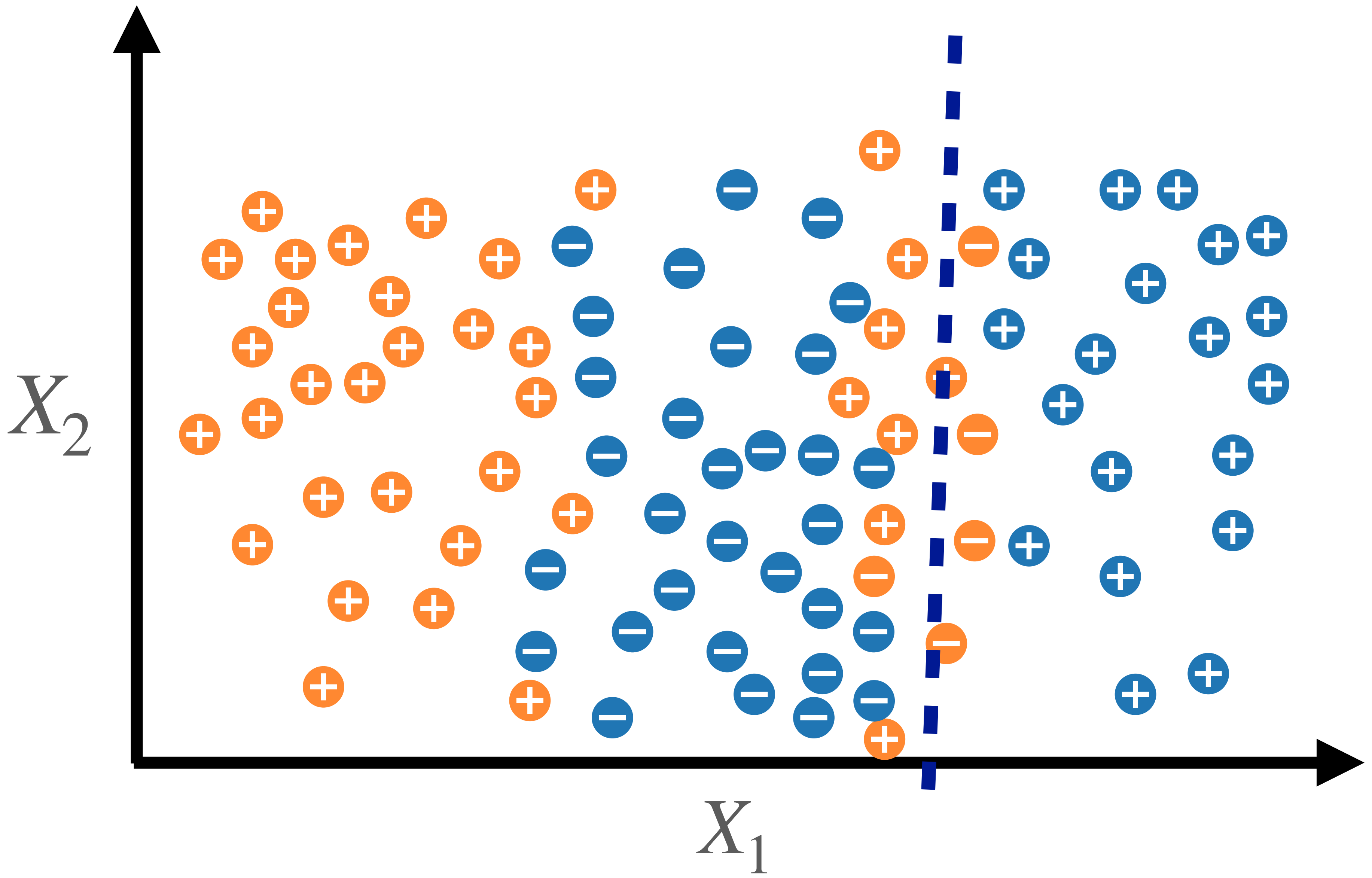


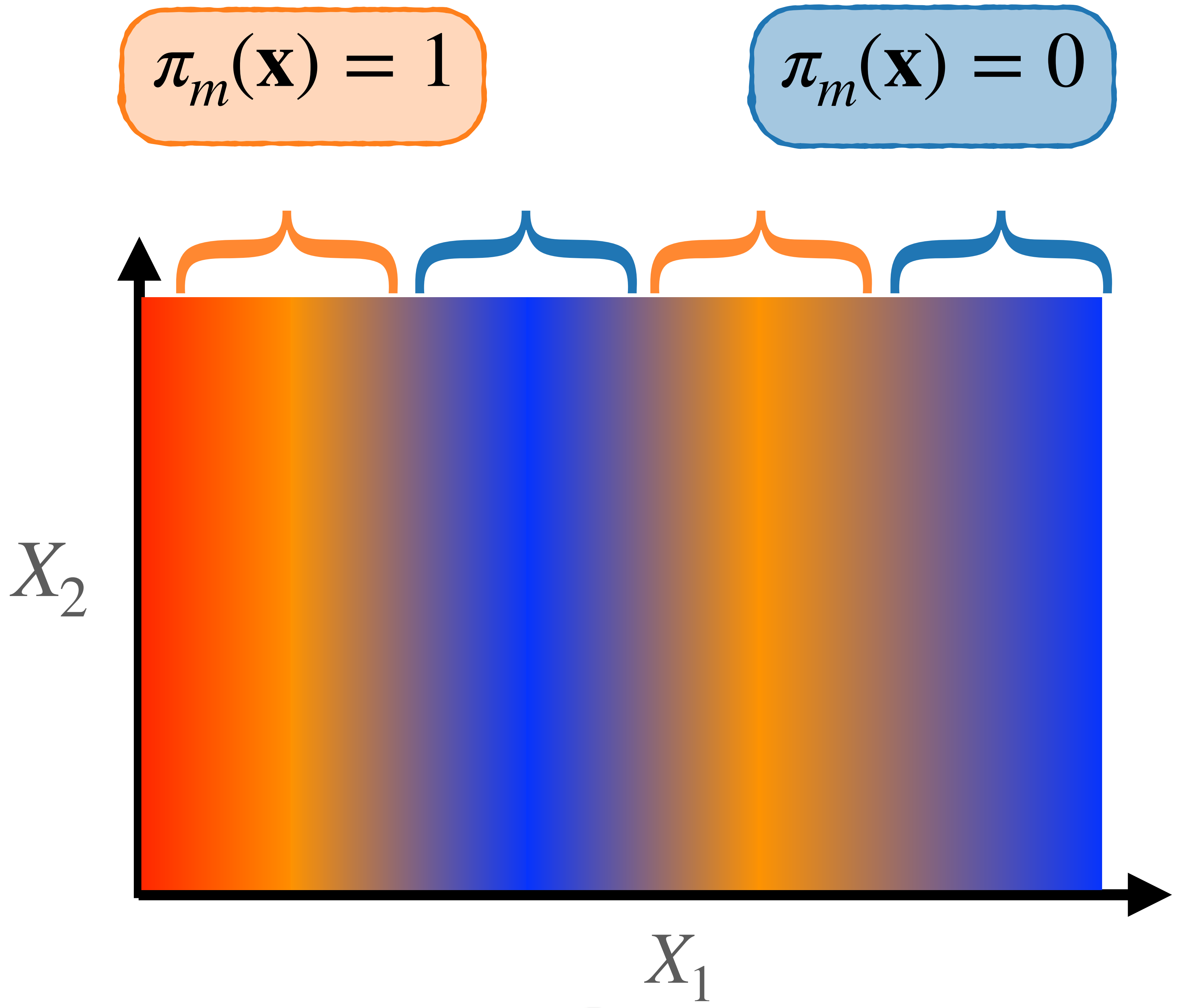
Heatmap showing $\mathbb{E}_{y|\mathbf{x}}\left[\ell(m(\mathbf{x}), y) - \mathbb{E}_{h|\mathbf{x}}\left[\ell(h, y)\right]\right]$

WHAT IS THE OPTIMAL TRIAGE POLICY?

- A **deterministic threshold rule** on the difference between the model and human loss on a per instance level:

$$\pi^*_{m,b}(\mathbf{x}) = \begin{cases} 1 & \textit{if } \mathbb{E}_{y|\mathbf{x}}\left[\ell(m(\mathbf{x}), y) - \mathbb{E}_{h|\mathbf{x}}\left[\ell(h, y)\right]\right] > t_{P,b,m} \\ 0 & \textit{otherwise,} \end{cases}$$





Heatmap showing $\mathbb{E}_{y|\mathbf{x}}\left[\ell(m(\mathbf{x}), y) - \mathbb{E}_{h|\mathbf{x}}\left[\ell(h, y)\right]\right]$

SIMPLE, SCALABLE, EASY TO IMPLEMENT ALGORITHM

function $\textsc{TrainModel}(\theta', \mathcal{D}, M, B, b, \alpha)$

$\quad \theta^{(0)} \leftarrow \theta'$

$\quad$ **for** $i = 0, \ldots, M-1$ **do**

For each mini-batch:

$\quad\quad \mathcal{D}^{(i)} \leftarrow$ the i'th mini batch of $\mathcal{D}$

$\quad\quad \mathcal{D}^{(i)} \leftarrow \textsc{Triage}(\mathcal{D}^{(i)}, b, \theta^{(i)})$

Use the optimal triage policy to find those samples inside the current batch for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0.$

$\quad\quad \nabla \leftarrow 0$

$\quad\quad$ **for** $(\mathbf{x}, y, h) \in \mathcal{D}^{(i)}$ **do**

$\quad\quad\quad \nabla \leftarrow \nabla + \nabla_\theta \, \ell(m_\theta(\mathbf{x}), y)|_{\theta=\theta^{(i)}}$

Calculate the gradient of loss with respect to the parameters of the machine model only on the samples for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0.$

$\quad\quad \theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha \, \frac{\nabla}{B}$

The machine model using the gradients of the points for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0.$

$\quad$ **return** $\theta^{(M)}$

Repeat for T time steps

SIMPLE, SCALABLE, EASY TO IMPLEMENT ALGORITHM

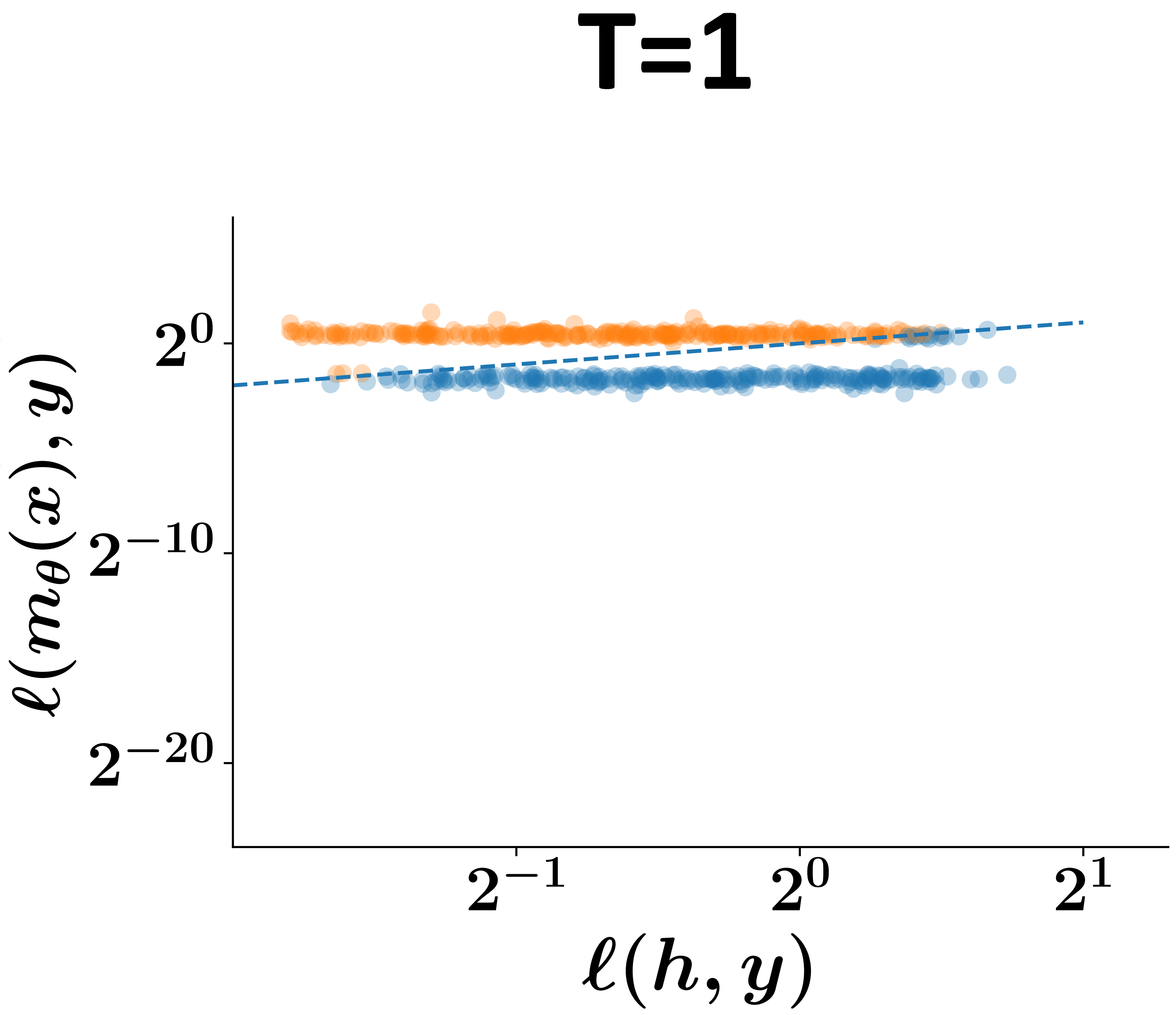
function TRAINMODEL(θ', $\mathcal{D}$, M, B, b, α)
  $\theta^{(0)} \leftarrow \theta'$
  for $i = 0, \ldots, M-1$ **do**
    $\mathcal{D}^{(i)} \leftarrow$ the i'th mini batch of $\mathcal{D}$
    $\mathcal{D}^{(i)} \leftarrow$ TRIAGE($\mathcal{D}^{(i)}, b, \theta^{(i)}$)
    $\nabla \leftarrow 0$
    for $(\mathbf{x}, y, h) \in \mathcal{D}^{(i)}$ **do**
      $\nabla \leftarrow \nabla + \nabla_\theta \, \ell(m_\theta(\mathbf{x}), y)|_{\theta=\theta^{(i)}}$
    $\theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha \frac{\nabla}{B}$
  return $\theta^{(M)}$

For each mini-batch:

Use the optimal triage policy to find those samples inside the current batch for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

Calculate the gradient of loss with respect to the parameters of the machine model only on the samples for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

The machine model using the gradients of the points for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.



SIMPLE, SCALABLE, EASY TO IMPLEMENT ALGORITHM

function TRAINMODEL(θ', $\mathcal{D}$, M, B, b, α)
  $\theta^{(0)} \leftarrow \theta'$
  for $i = 0, \ldots, M-1$ **do**
    $\mathcal{D}^{(i)} \leftarrow$ the i'th mini batch of $\mathcal{D}$
    $\mathcal{D}^{(i)} \leftarrow \text{TRIAGE}(\mathcal{D}^{(i)}, b, \theta^{(i)})$
    $\nabla \leftarrow 0$
    for $(\mathbf{x}, y, h) \in \mathcal{D}^{(i)}$ **do**
      $\nabla \leftarrow \nabla + \nabla_\theta \, \ell(m_\theta(\mathbf{x}), y)|_{\theta=\theta^{(i)}}$
    $\theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha \frac{\nabla}{B}$
  return $\theta^{(M)}$

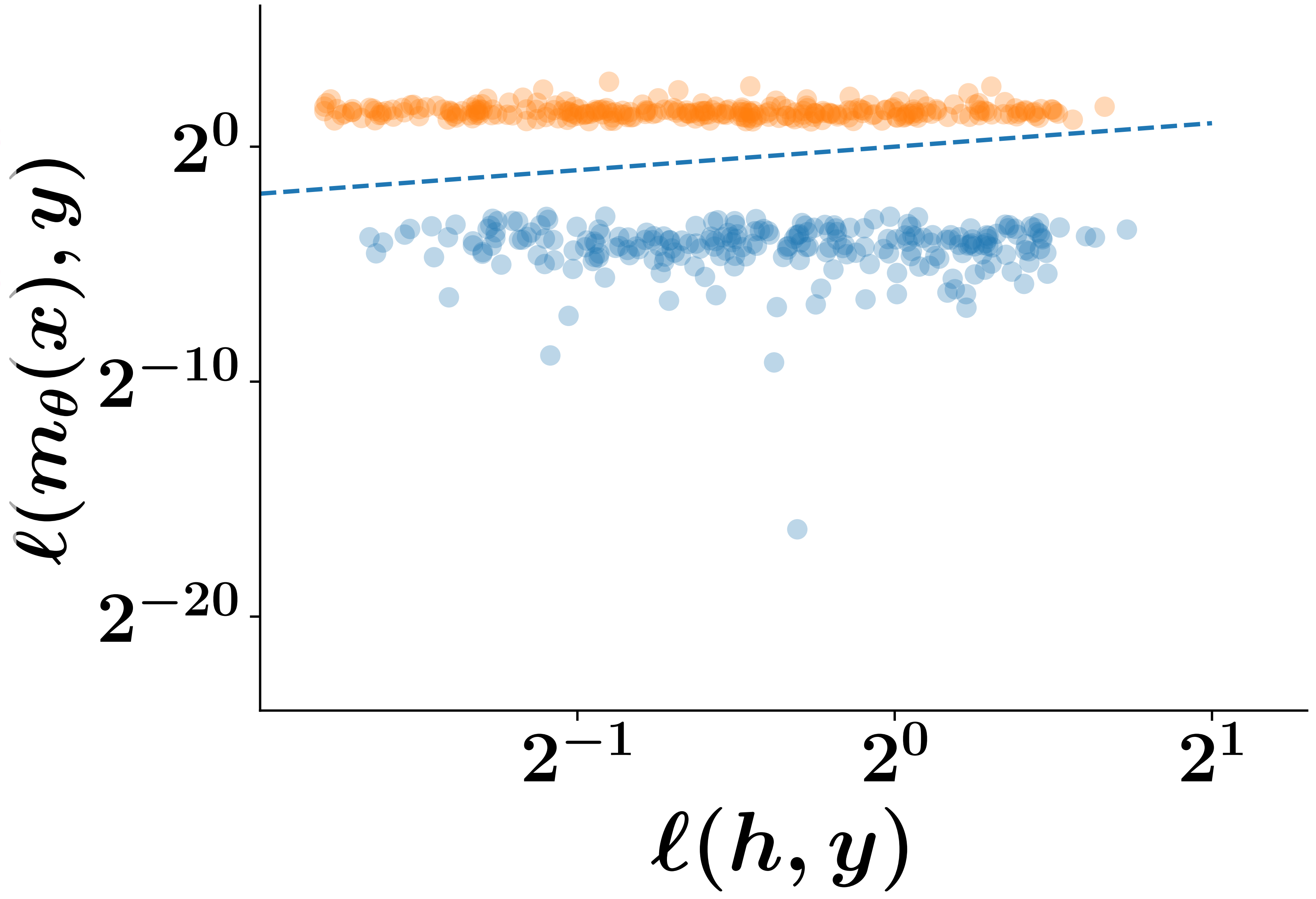
For each mini-batch:

Use the optimal triage policy to find those samples inside the current batch for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

Calculate the gradient of loss with respect to the parameters of the machine model only on the samples for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

The machine model using the gradients of the points for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

T=2



SIMPLE, SCALABLE, EASY TO IMPLEMENT ALGORITHM

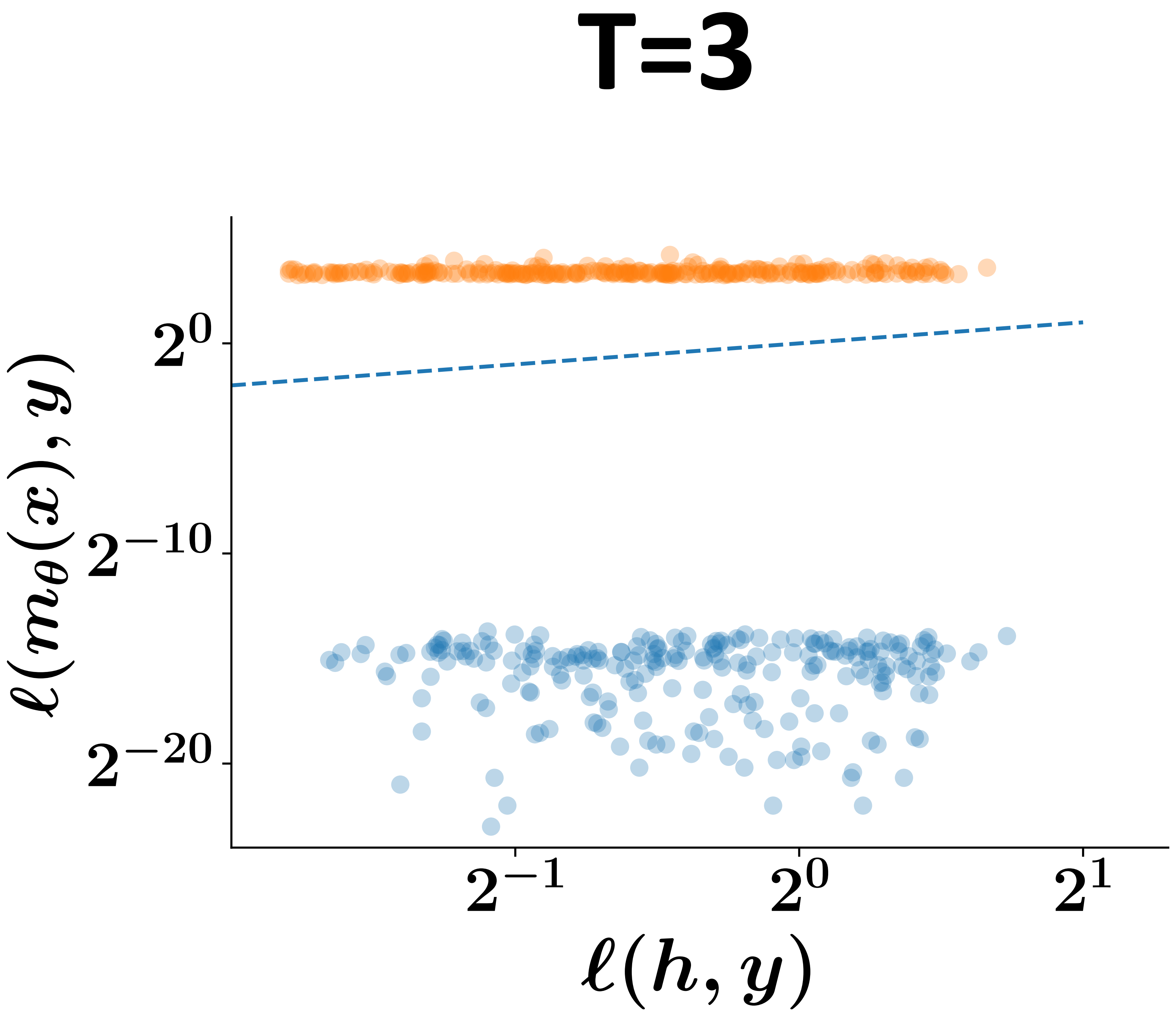
function TRAINMODEL(θ', $\mathcal{D}$, M, B, b, α)
  $\theta^{(0)} \leftarrow \theta'$
  for $i = 0, \ldots, M-1$ **do**
    $\mathcal{D}^{(i)} \leftarrow$ the i'th mini batch of $\mathcal{D}$
    $\mathcal{D}^{(i)} \leftarrow$ TRIAGE($\mathcal{D}^{(i)}, b, \theta^{(i)}$)
    $\nabla \leftarrow 0$
    for $(\mathbf{x}, y, h) \in \mathcal{D}^{(i)}$ **do**
      $\nabla \leftarrow \nabla + \nabla_\theta\, \ell(m_\theta(\mathbf{x}), y)|_{\theta=\theta^{(i)}}$
    $\theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha \frac{\nabla}{B}$
  return $\theta^{(M)}$

For each mini-batch:

Use the optimal triage policy to find those samples inside the current batch for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

Calculate the gradient of loss with respect to the parameters of the machine model only on the samples for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

The machine model using the gradients of the points for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.



SIMPLE, SCALABLE, EASY TO IMPLEMENT ALGORITHM

function TRAINMODEL$(\theta', \mathcal{D}, M, B, b, \alpha)$
$\theta^{(0)} \leftarrow \theta'$
for $i = 0, \ldots, M-1$ **do**
$\mathcal{D}^{(i)} \leftarrow$ the i'th mini batch of $\mathcal{D}$
$\mathcal{D}^{(i)} \leftarrow$ TRIAGE$(\mathcal{D}^{(i)}, b, \theta^{(i)})$
$\nabla \leftarrow 0$
for $(\mathbf{x}, y, h) \in \mathcal{D}^{(i)}$ **do**
$\nabla \leftarrow \nabla + \nabla_\theta\, \ell(m_\theta(\mathbf{x}), y)|_{\theta=\theta^{(i)}}$
$\theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha \frac{\nabla}{B}$
return $\theta^{(M)}$

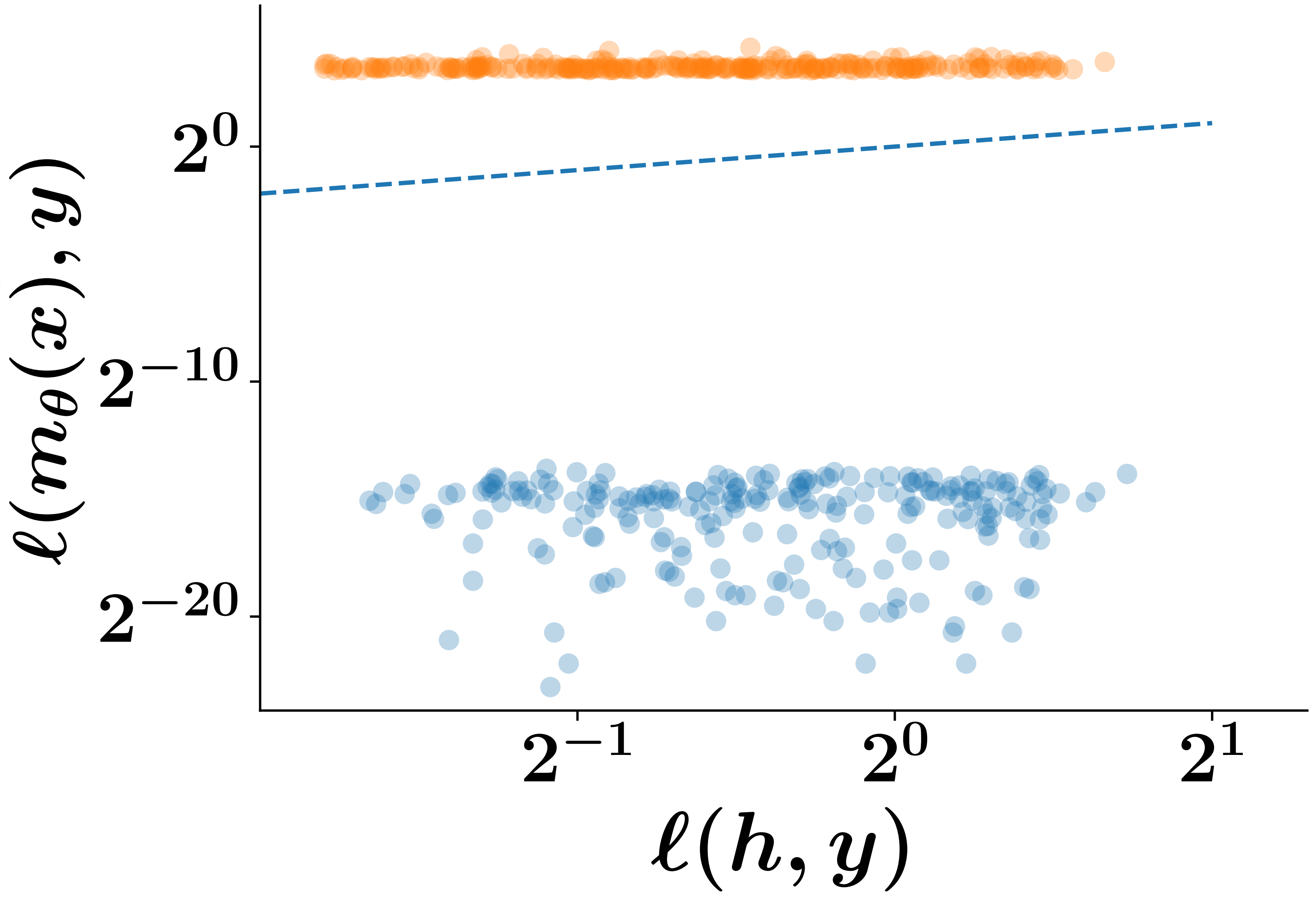
For each mini-batch:

Use the optimal triage policy to find those samples inside the current batch for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

Calculate the gradient of loss with respect to the parameters of the machine model only on the samples for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

The machine model using the gradients of the points for which $\pi_{m_{\theta(i)}}(\mathbf{x}) = 0$.

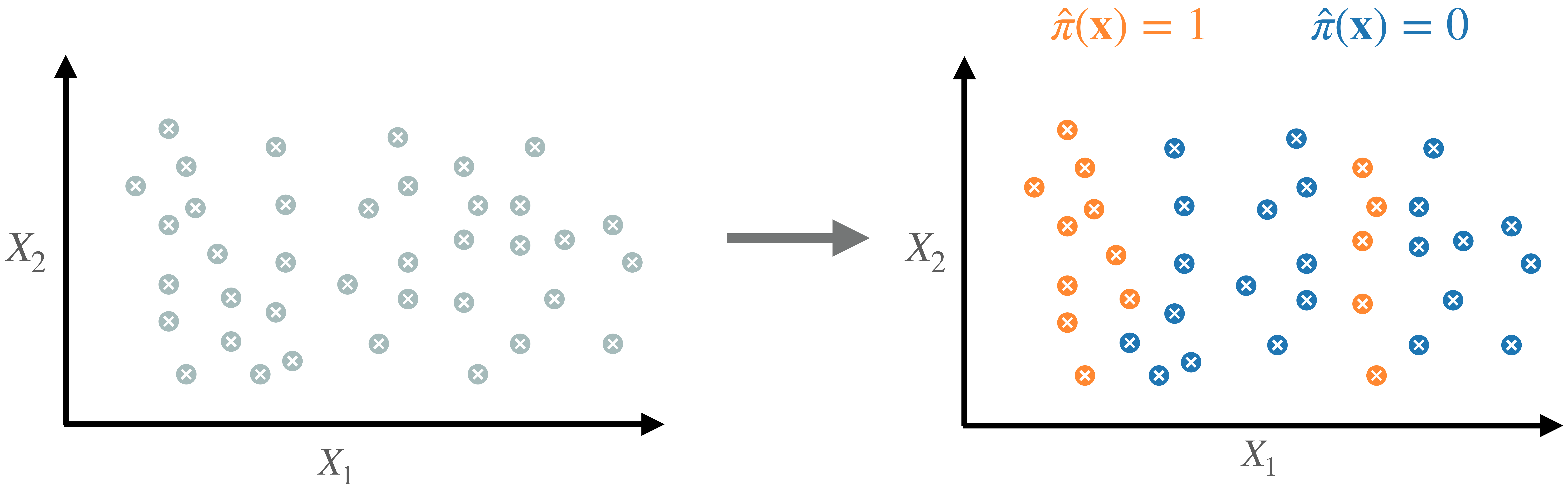
T=3



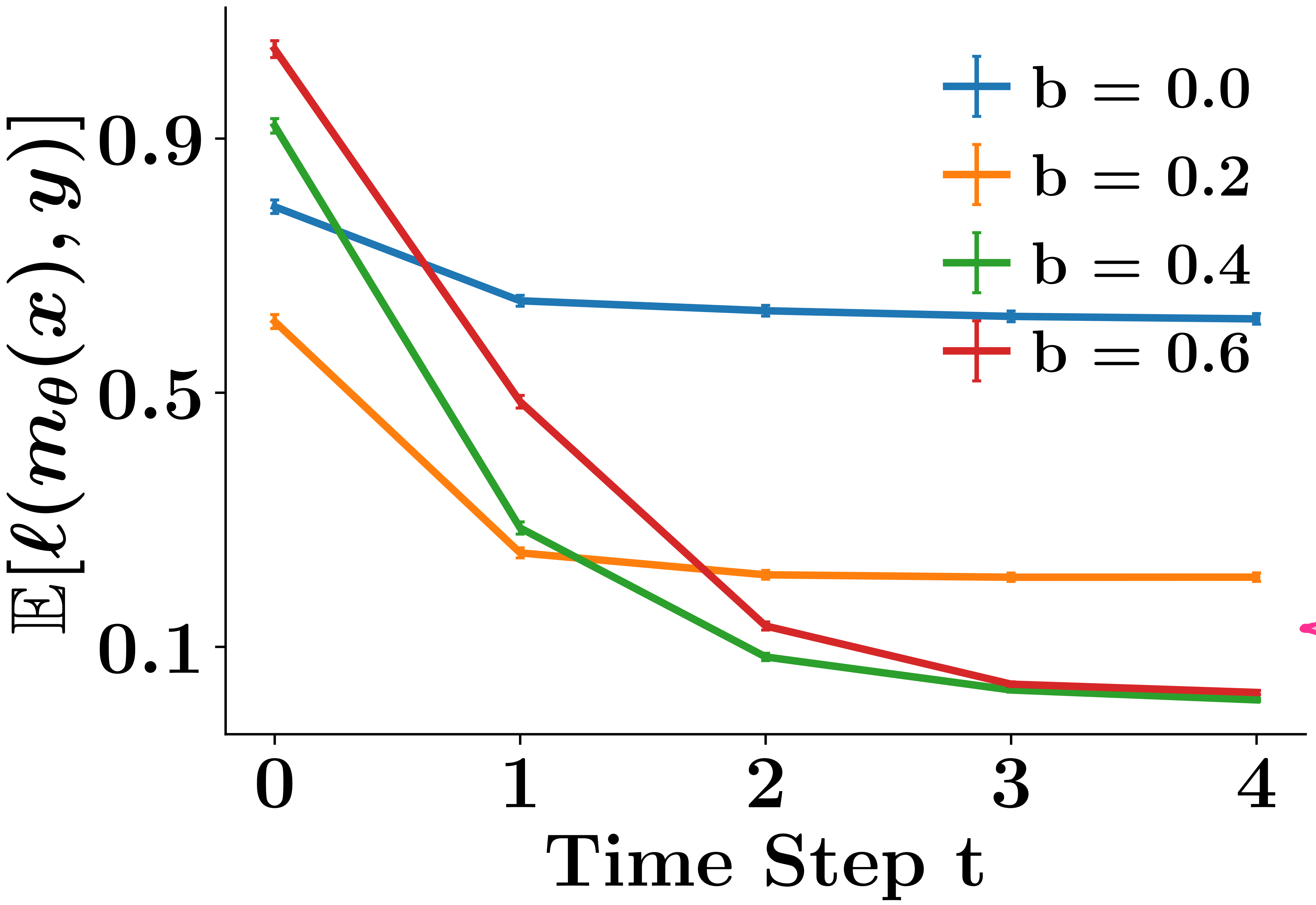
Later into the training process, the predictive model focuses on predicting more accurately the samples that the triage policy hands in to the model.

HOW TO ASSIGN SAMPLES DURING TEST TIME?

- We do not observe the true label and the human prediction at test time.
- We cannot compute the optimal triage policy at test time since it depends on the true label.
- **Solution**: train a parametrised triage policy to **approximate the optimal triage policy**.



HOW DOES TRAINING PERFORMANCE CHANGE UNDER DIFFERENT AUTOMATION LEVELS?



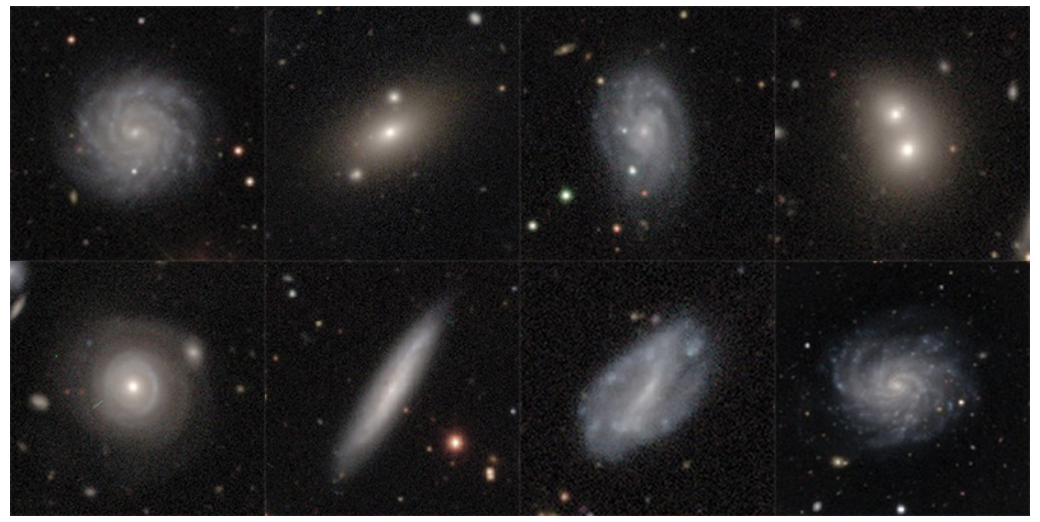
The model is generalising across a large portion of the feature space incuring high error.

The models trained under higher levels of triage needs to generalise across a smaller region of the feature space and achieve much lower error.

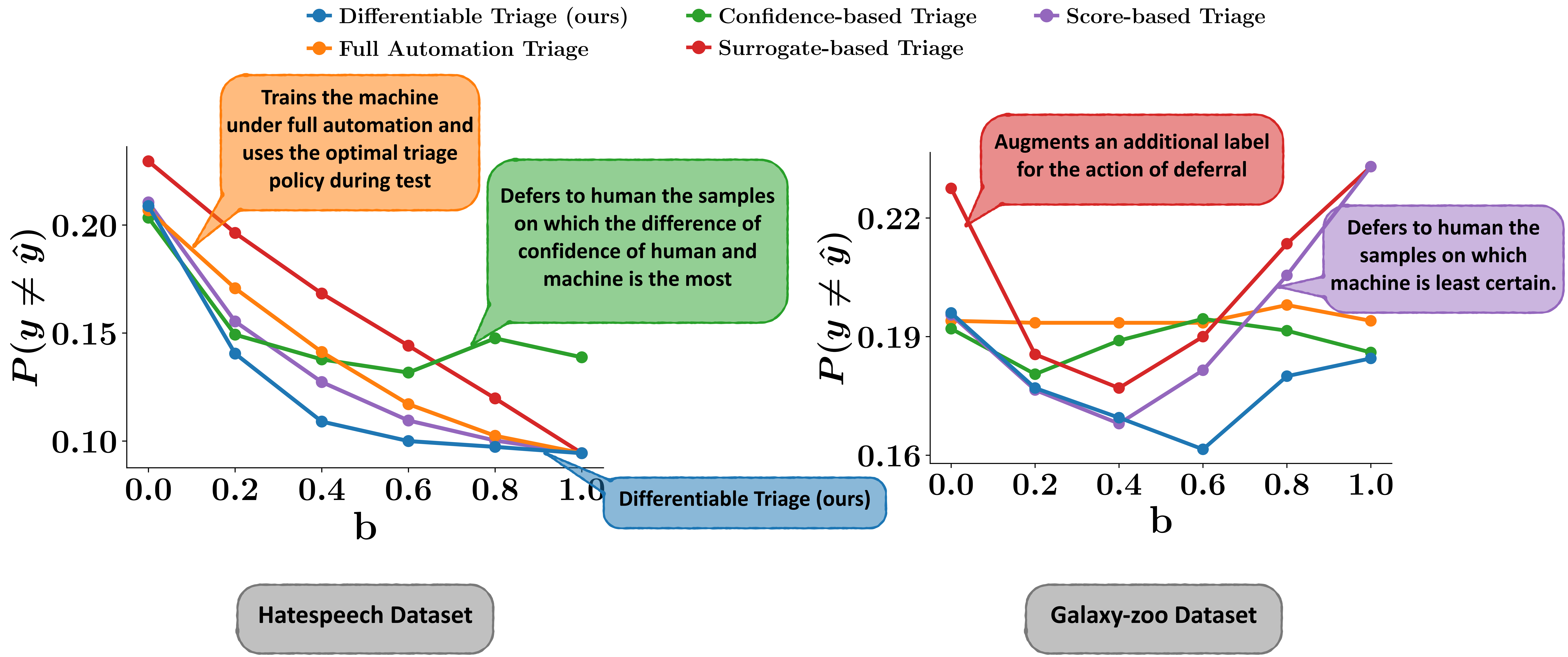
EXPERIMENTS

- Few public dataset with **several human predictions per instance**, necessary to estimate the human loss per instance.
- Hatespeech:
  - 25k tweets
  - Each tweet labeled by 3-5 annotator
  - Labels = {hatespeech, offensive, neither}

- Galaxy-zoo:
  - 10k images
  - Each image labelled by 30+ humans
  - Labels = {early-type, spiral}



AVERAGE TEST MISCLASSIFICATION ERROR



CONCLUSION

- We have contributed towards a better understanding of **supervised learning under algorithmic triage**.

- We have designed a **gradient-based algorithm** for the task of supervised learning under algorithmic triage that is:
  - Easy to implement
  - Is applicable to any differentiable machine learning model
  - Does not increase the complexity of the vanilla SGD
  - Is guaranteed to converge to a local minima