Generalized Linear Models

Prepared by: Joseph Bakarji

What if y is a label?

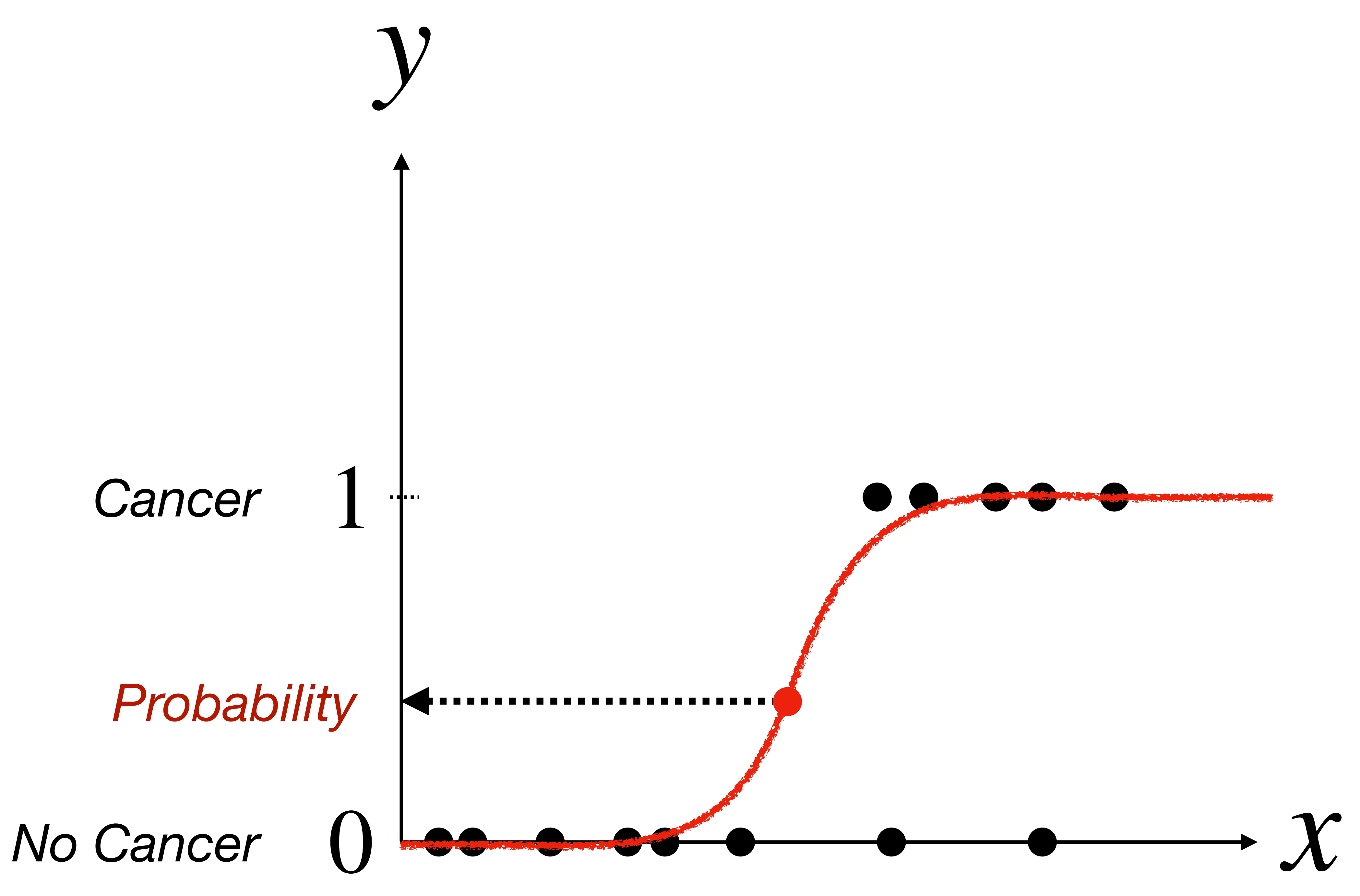
$$x \xrightarrow{h} y$$

Given the data,
find a **function** h,
that predicts y, given x

$$\mathbf{y} = h(\mathbf{x})$$

$$y \in [0,1]$$

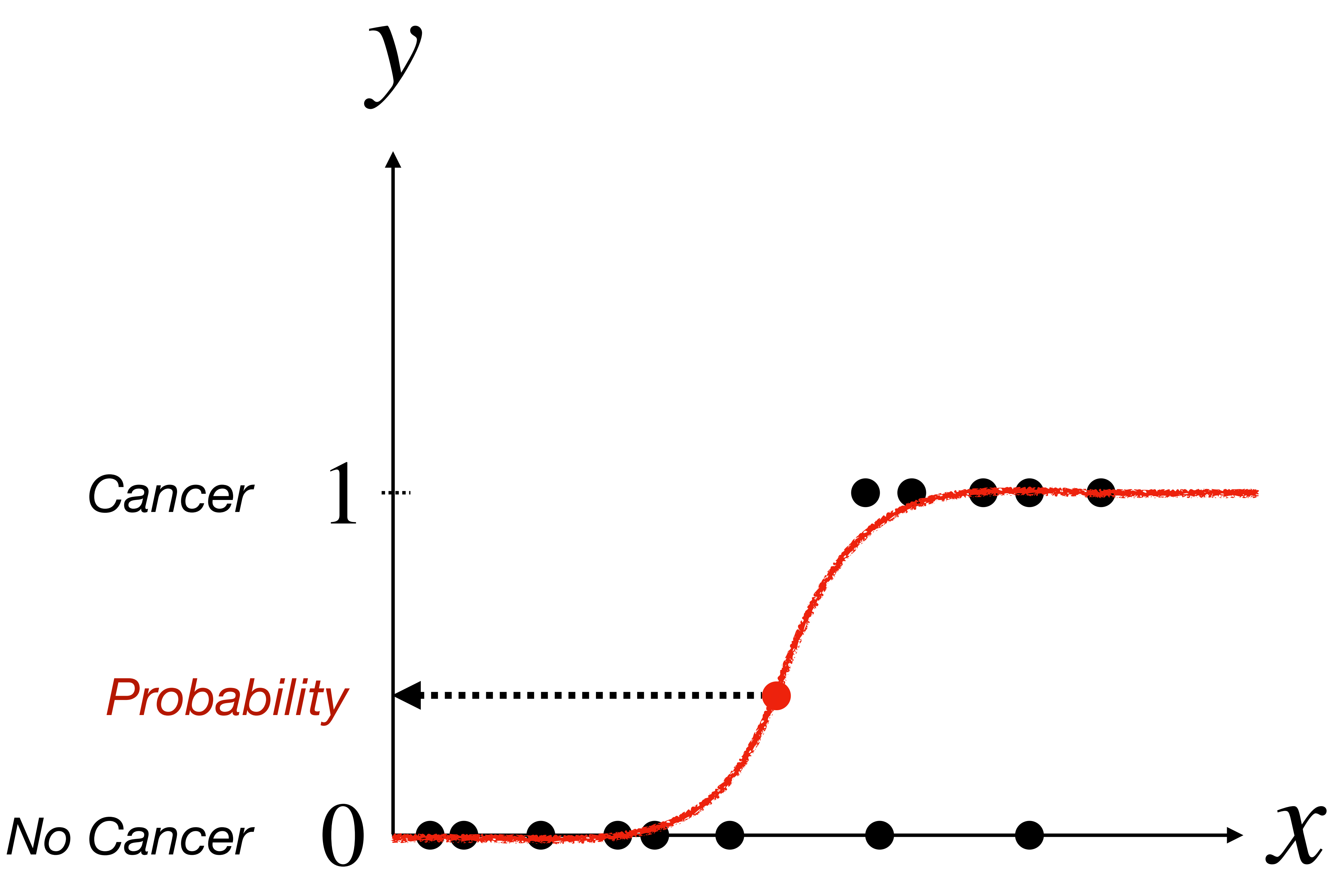
A smooth function that returns probability of occurrence



What if y is a label?

$$y = h_\theta(x) \quad \& \quad y \in [0,1]$$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta^\top x)}}$$

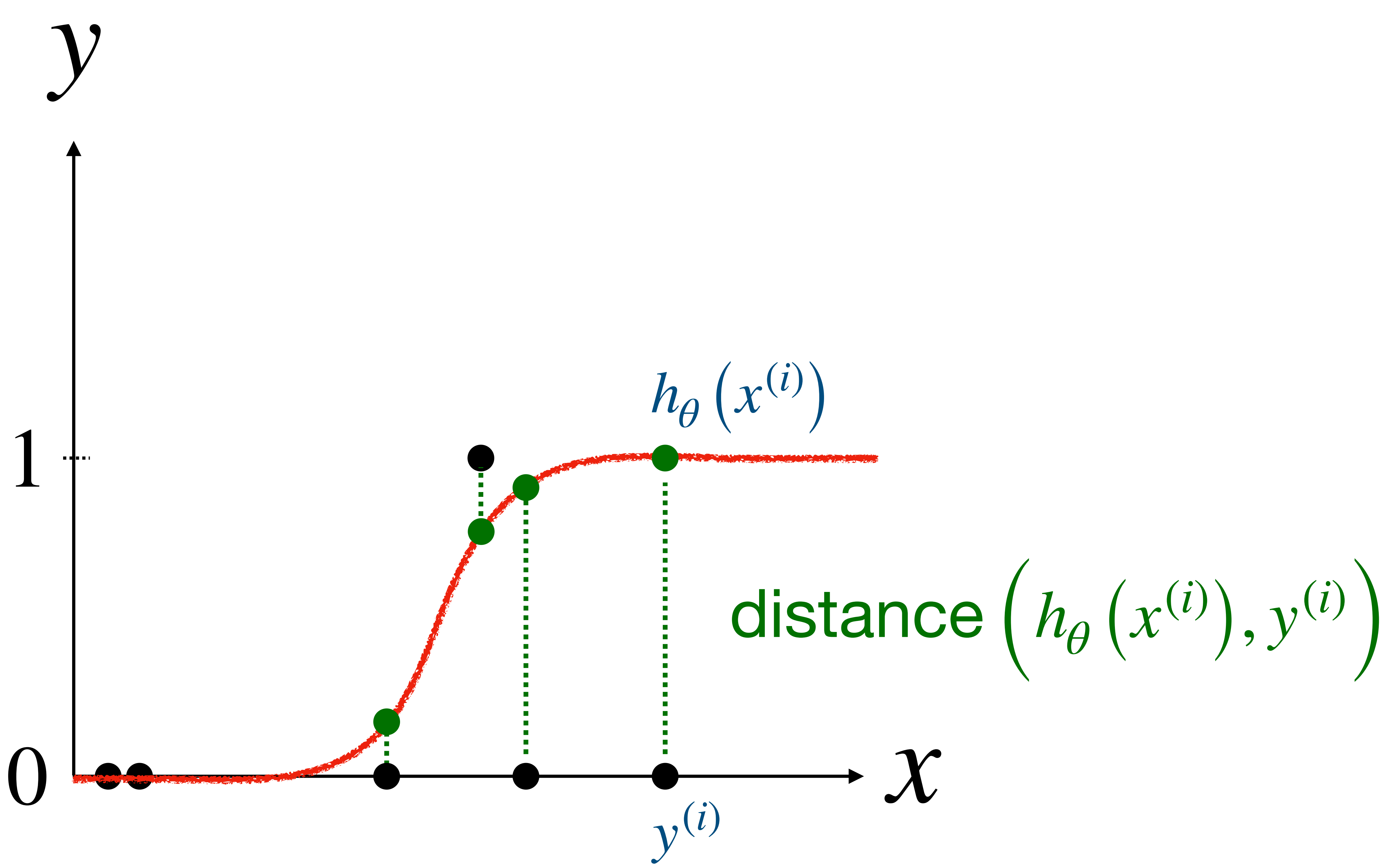


1. **Define a predictor:** the logistic function ✅
2. **Define a loss:** distance between function and data ❓
3. **Optimize loss**
4. **Test model**

Logistic Regression

$$y = h_\theta(x)$$

$$h_\theta(x) = \frac{1}{1 + e^{-\left(\theta^\top x\right)}} = g\left(\theta^\top x\right)$$



Linear predictor
negative log-likelihood or OLS

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{d}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Logistic predictor
Binary-cross entropy loss

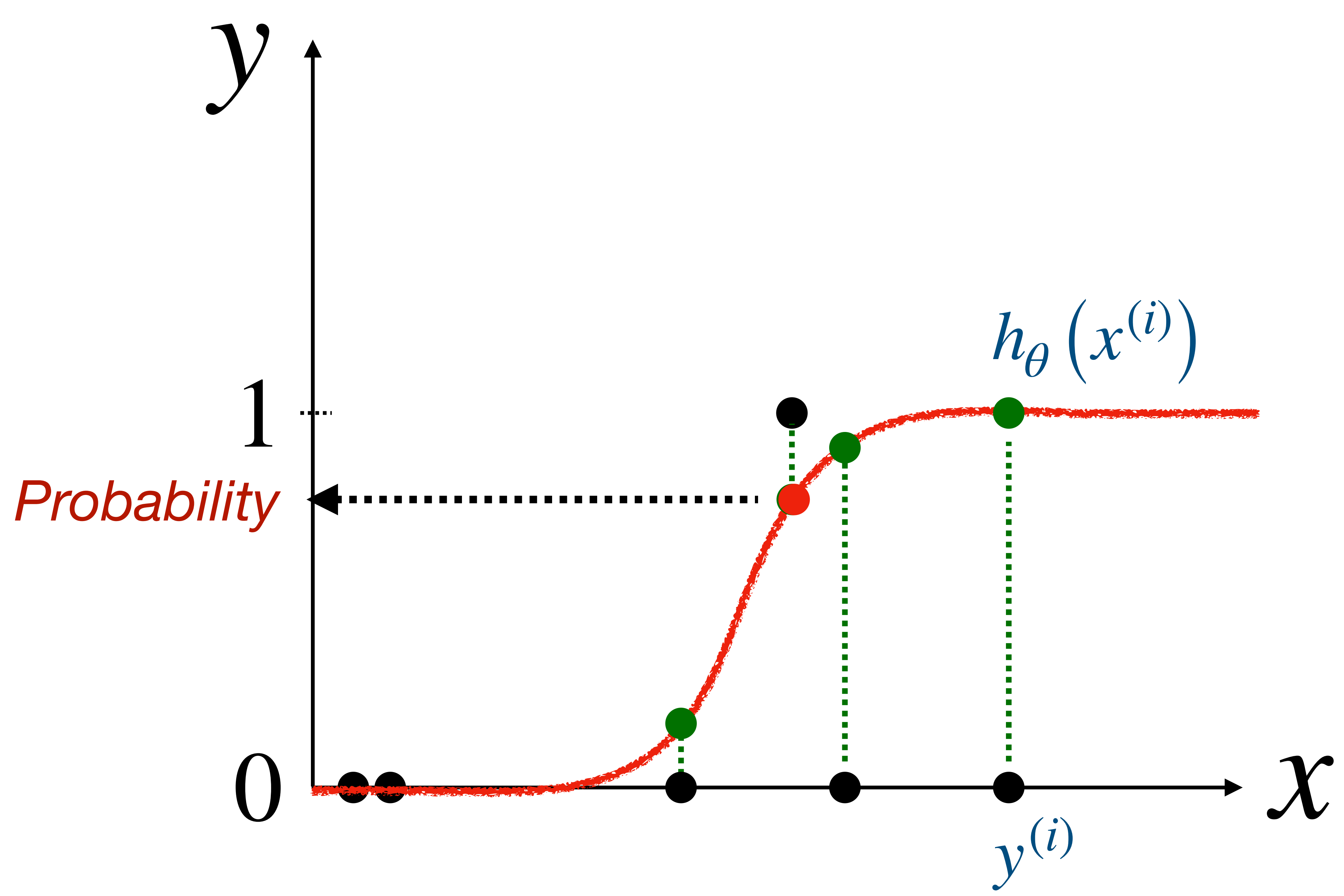
$$\mathcal{L}(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

Gradient descent → Done!

Why not Least Squares?

$$y = h_\theta(x)$$

$$h_\theta(x) = \frac{1}{1 + e^{-\left(\theta^\top x\right)}} = \sigma\left(\theta^\top x\right)$$



Probability of output given input

$$P\left(y = 1 \,\middle|\, x; \theta\right) = h_\theta(x)$$

$$P\left(y = 0 \,\middle|\, x; \theta\right) = 1 - h_\theta(x)$$

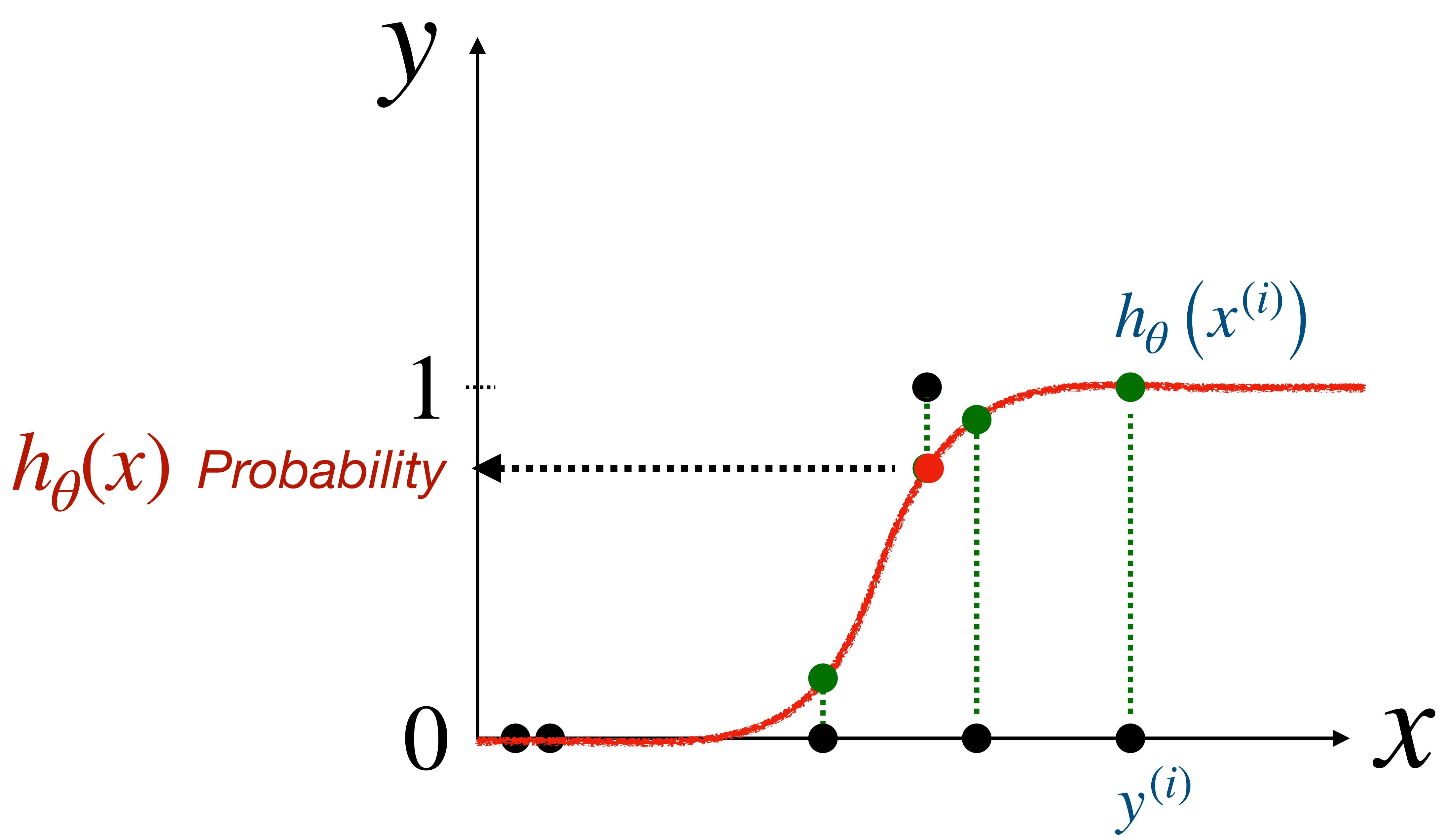
$$p(y \,|\, x; \theta) = \left(h_\theta(x)\right)^{\overset{\text{True label}}{y}} \left(1 - h_\theta(x)\right)^{1-y}$$

Likelihood!

Why not Least Squares?

$$y = h_\theta(x)$$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta^\top x)}} = \sigma\left(\theta^\top x\right)$$



Probability of output given input

$$P\left(y = 1 \,\middle|\, x; \theta\right) = \sigma(\theta^\top x)$$

$$P\left(y = 0 \,\middle|\, x; \theta\right) = 1 - \sigma(\theta^\top x)$$

$\longrightarrow$

True label

$$p(y \,|\, x; \theta) = \left(\sigma(\theta^\top x)\right)^{y} \left(1 - \sigma(\theta^\top x)\right)^{1-y}$$

Likelihood!

Maximize Log-likelihood

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

Update rule

while not converged:

$$\theta := \theta + \alpha \nabla_\theta \mathscr{L}(\theta)$$

Derive

Gradient Descent

for t = 1...T:

$$\theta := \theta - \alpha \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

🤯 **Same as linear regression** 🤔

Generalized Linear Models

Gaussian Distribution $\longrightarrow$ Linear Regression

Bernoulli Distribution $\longrightarrow$ Logistic Regression

Update rule

$$\theta := \theta - \alpha \sum_{i=1}^{n} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x^{(i)}$$

Exponential Family

Family of distributions for which we can derive the same update rule

Assumption: $p(y \mid x; \theta)$ is an exponential family

Data

$$p(y; \eta) = b(y) \exp\left\{\eta^\top y - a(\eta)\right\}$$

Parameters

- $b(y)$ is called the base measure (not depend on η)
- $a(\eta)$ is called the log partition function (not depend on y)
- $a(\eta)$, y and $b(y)$ are scalar. η and y have the same dimensions.

Example 1: Bernoulli Distribution -> Logistic Regression

Data

$$p(y; \eta) = b(y) \exp\left\{\eta^\top y - a(\eta)\right\}$$

Natural Parameters

Bernoulli Distribution

$$p(y; \phi) = \phi^y (1-\phi)^{1-y} = \exp\left\{ y \underbrace{\log \frac{\phi}{1-\phi}}_{\eta} + \underbrace{\log(1-\phi)}_{a(\eta)} \right\}$$

Show that term is only a function of η

Example 2: Gaussian Distribution -> Linear Regression

Data

$$p(y; \eta) = b(y)\ \exp\left\{\eta^\top y - a(\eta)\right\}$$

Natural Parameters

Gaussian Distribution

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\} = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}_{b(y)} \left\{ \underbrace{\mu}_{\eta} y - \underbrace{\frac{1}{2}\mu^2}_{a(\eta)} \right\}$$

Why do we care?

$\theta^\top x$

Data

$$p(y;\eta) = b(y)\,\exp\left\{\eta^\top y - a(\eta)\right\}$$

Natural Parameters

Inference is Easy: $$E[y;\eta] = \frac{da(\eta)}{d\eta} \qquad Var[y;\eta] = \frac{d^2a(\eta)}{d\eta^2}$$

Learning is Easy: **Maximum Likelihood Estimation** leads to **convex** problem in η

Generalized Linear Models

Assumption: $p(y \mid x; \theta)$ is an exponential family

Data Type → Probability Distribution

Binary → Bernoulli ⟶ **Logistic Regression**

Real → Gaussian ⟶ **Linear Regression**

Counts → Poisson

Positive Real → Gamma, Exponential

Distributions → Dirichlet

Generalized Linear Models

Assumption: $p(y \mid x; \theta)$ is an exponential family

The natural parameter is linear in the inputs

$$\boxed{\eta = \theta^\top x}$$

Predictor is a natural consequence

$$\boxed{h_\theta(x) = E[y \mid x; \theta]}$$

Generalized Linear Models

Assumption: $p(y \mid x; \theta)$ is an exponential family

Data Input		Linear Model		Exponential Model
x	$\longrightarrow$	$\theta^\top x$	$\xrightarrow{\eta}$	η, $a(\eta)$, $b(y)$, $\vdots$

Learning

$$\max_{\theta} \log(p(y \mid x; \theta))$$

Inference

$$E[y \mid x; \theta]$$

Update Rule:

$$\theta := \theta - \alpha \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x^{(i)}$$

Terminology

Model parameter — Natural parameter — Canonical parameter

$$\theta \xrightarrow{\theta^\top x} \eta \xrightarrow{g} \phi$$

(Bernoulli)

Canonical response: g

$h_\theta(x)$

Logistic Regression:

$$h_\theta(x) = E[y \mid x; \theta]$$

$$\phi = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-\theta^\top x}}$$

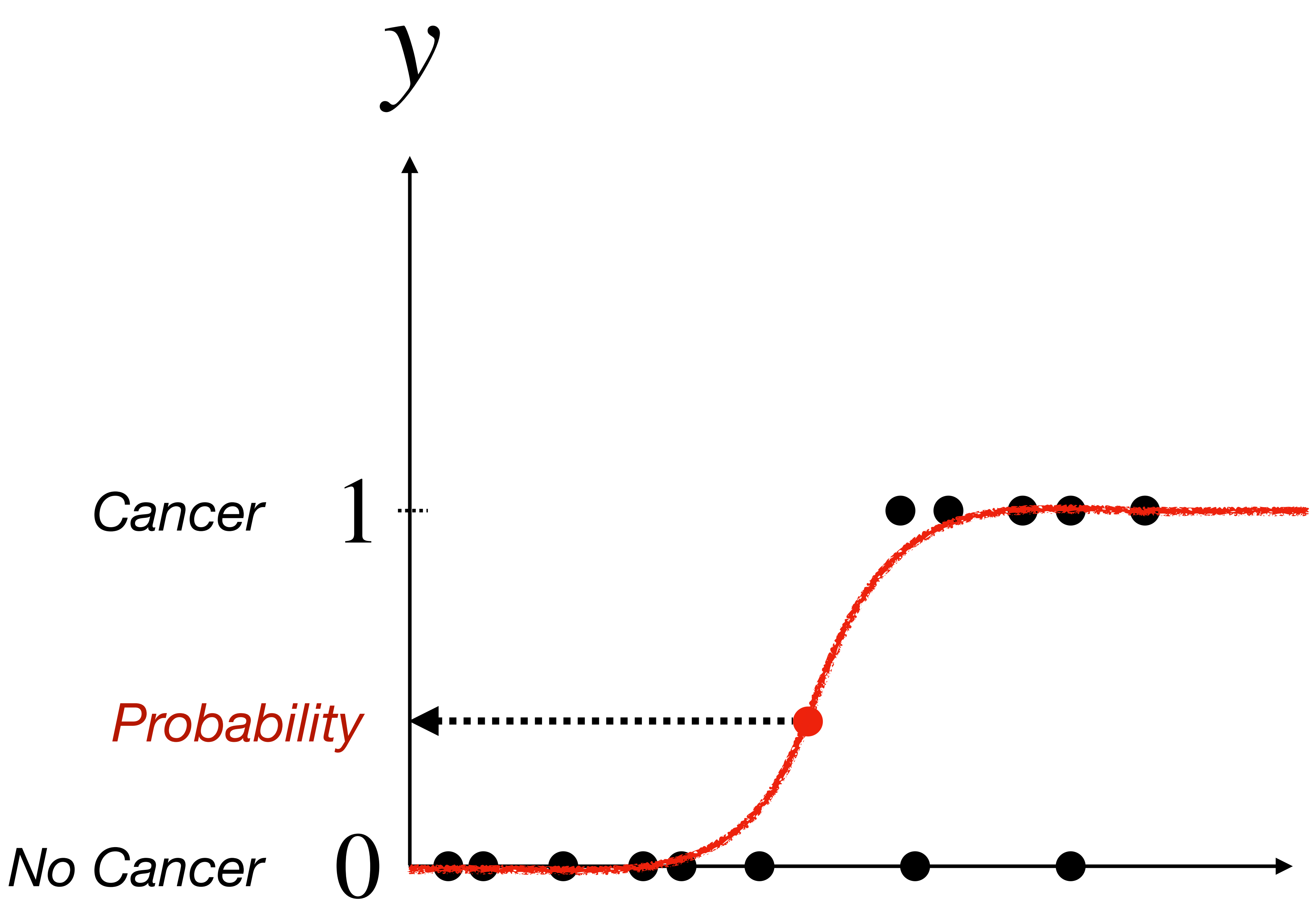
Back to classification

$$x \xrightarrow{h} y$$

$$y = h_\theta(x)$$

$$y \in [0,1]$$

What if we have more outputs?



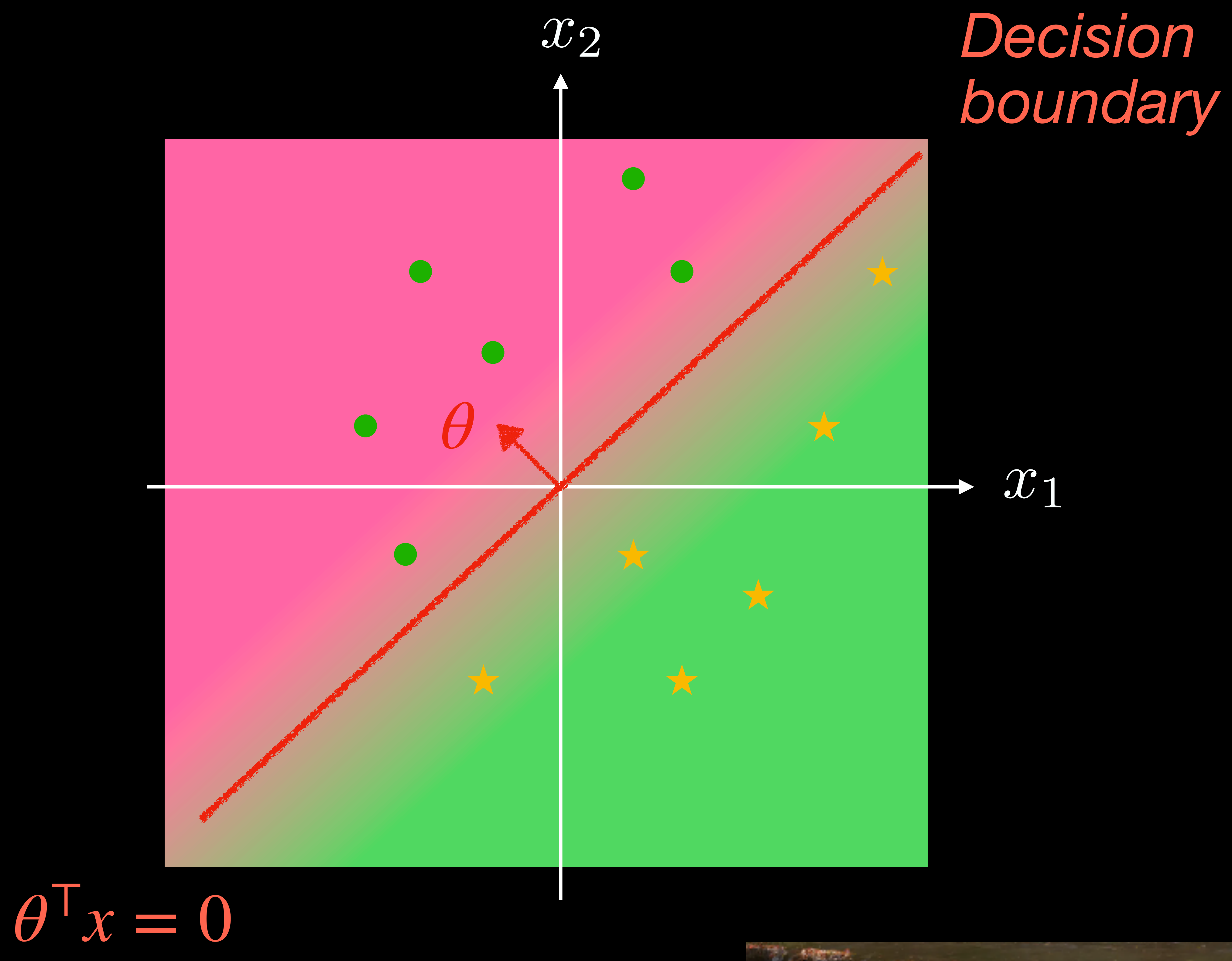
Classification

$x = [x_1, x_2]$

x_1	x_2	y
-2	-1	●
3	1	★
2	3	●
1	-1	★
	⋮	

★ 1

● 0



Logistic Regression

$h_\theta(x) = \sigma(\theta^\top x)$

how confident?

score

$\theta^\top x$

how correct?

margin

$(\theta^\top x) y$

For $y \in [1, -1]$

Multiclass classification - Softmax

$$\mathbf{y} = h_\theta(\mathbf{x})$$

$$\boxed{\mathbf{x}} \longrightarrow \boxed{\mathbf{y}}$$

k discrete values for representing output

One-hot encoding

$$\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\1\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$

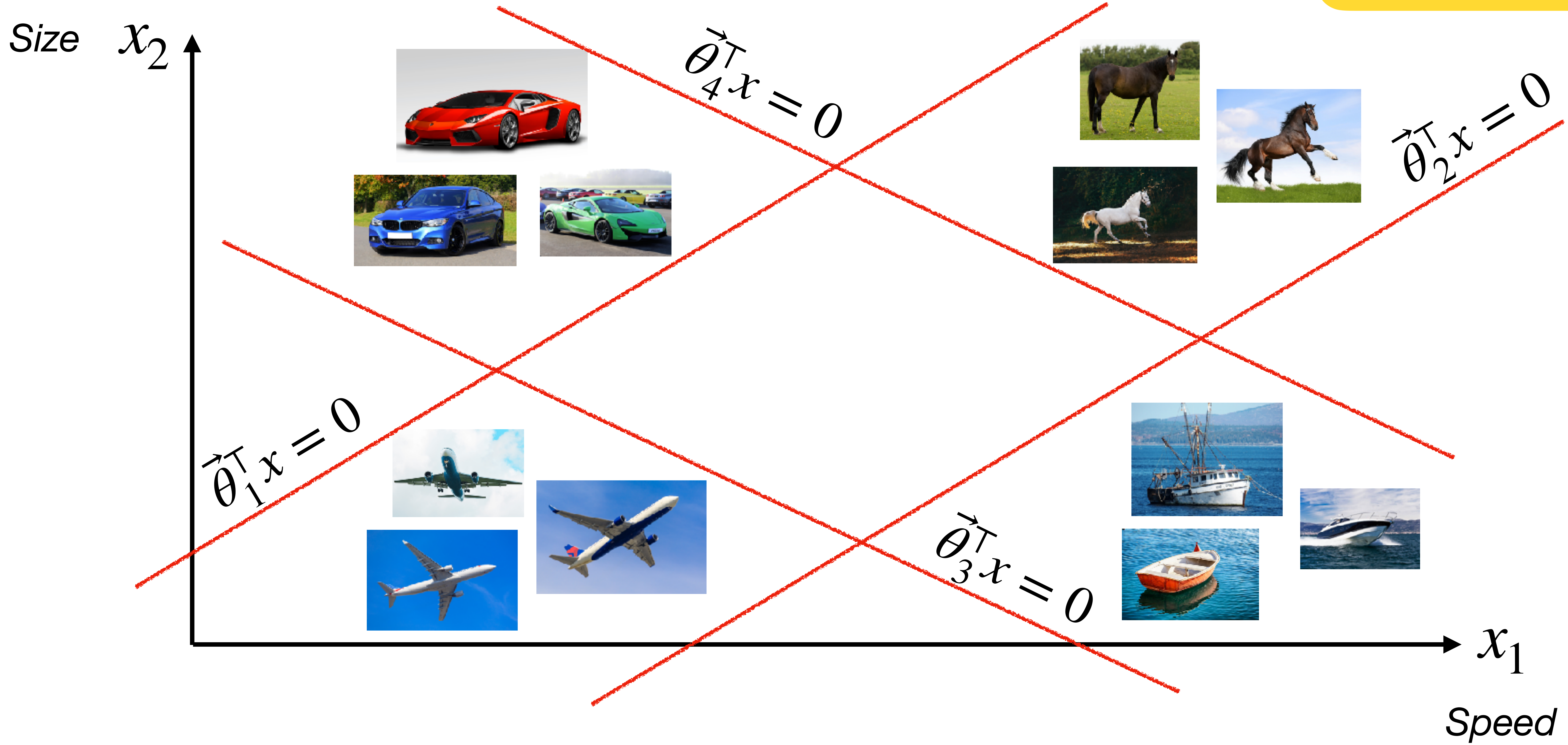
car plane boat horse

Multi-class classification - Softmax

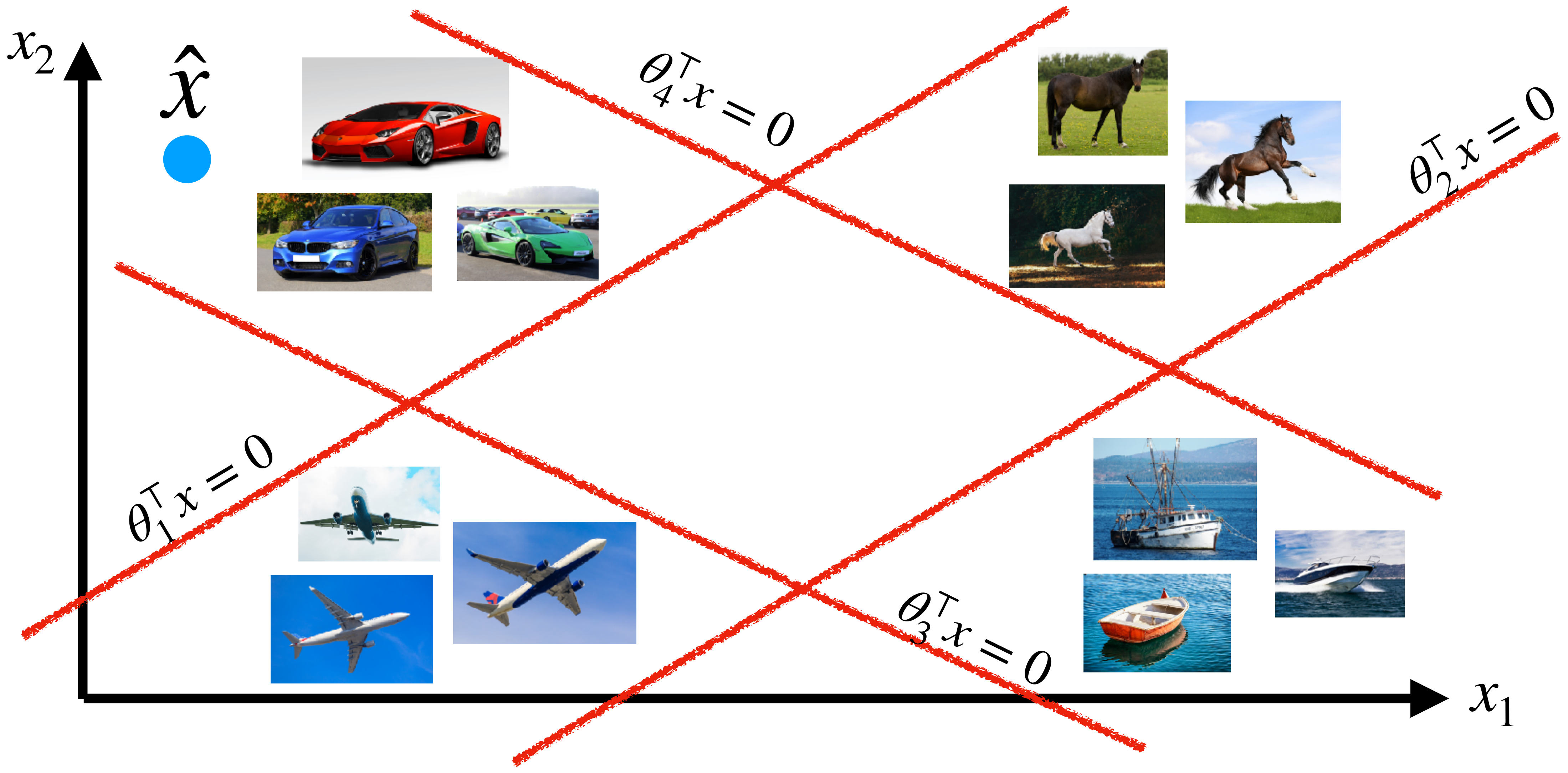
WARNING!!!

Notation Alert

$\vec{\theta}_i$ is a vector



How to turn scores into probabilities?



WARNING!!!

Notation Alert

$\vec{\theta}_i$ **is a vector**

Score

$\vec{\theta}_1^\top \hat{x} = 3$

$\vec{\theta}_2^\top \hat{x} = -0.3$

$\vec{\theta}_3^\top \hat{x} = -0.8$

$\vec{\theta}_4^\top \hat{x} = -22$

$\xrightarrow{\exp}$

Positive Measure

$\exp(3) = 20.1$

$\exp(-0.3) = 0.75$

$\exp(-0.8) = 0.2$

$\exp(-22) = 0.00..1$

Normalize $\longrightarrow$

Softmax

$$\hat{p}(y = i \mid x; \theta) = \frac{\exp\left(\vec{\theta}_i^\top x\right)}{\sum_{j=1}^{k} \exp\left(\vec{\theta}_j^\top x\right)}$$

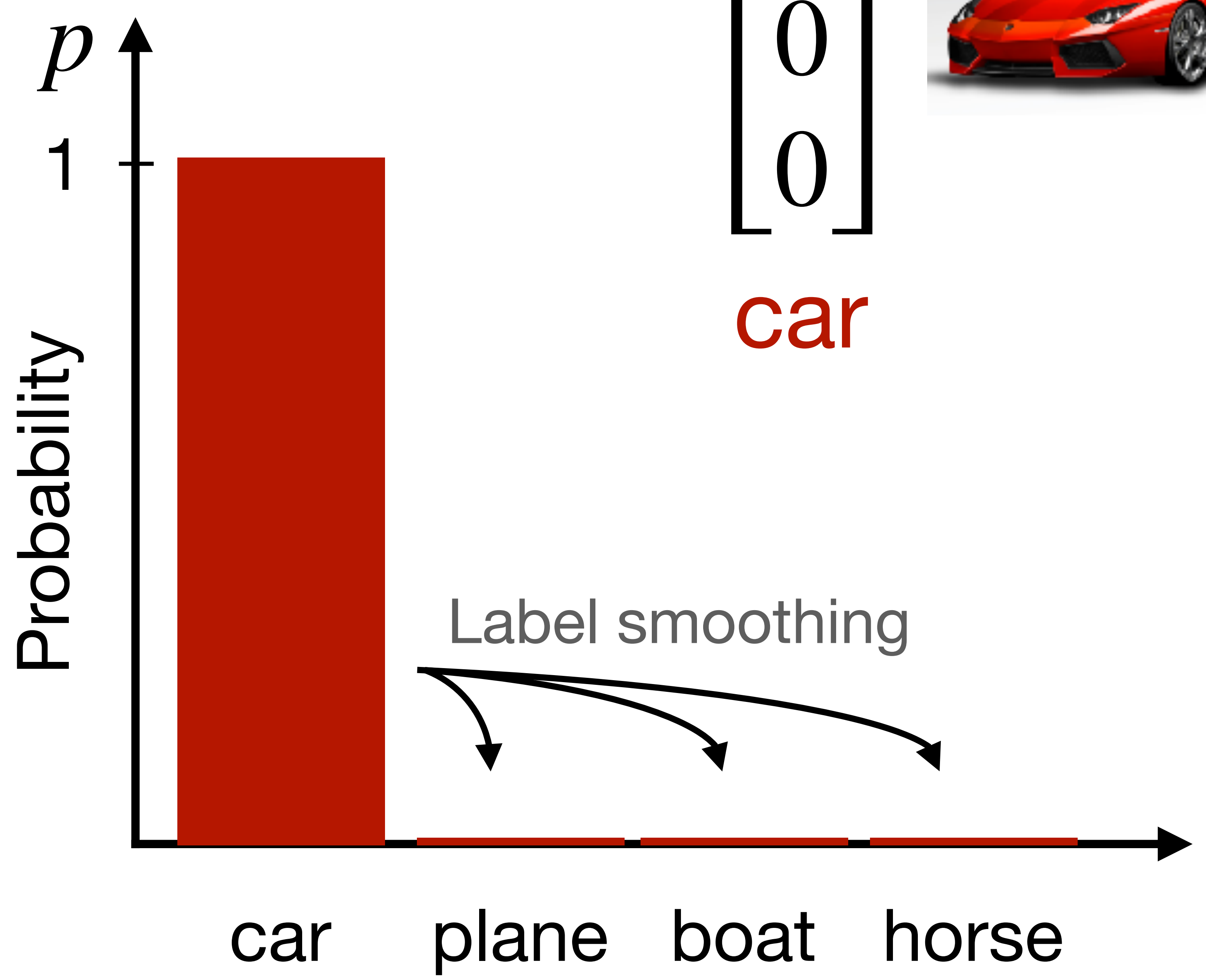
How do you train?

Given Label

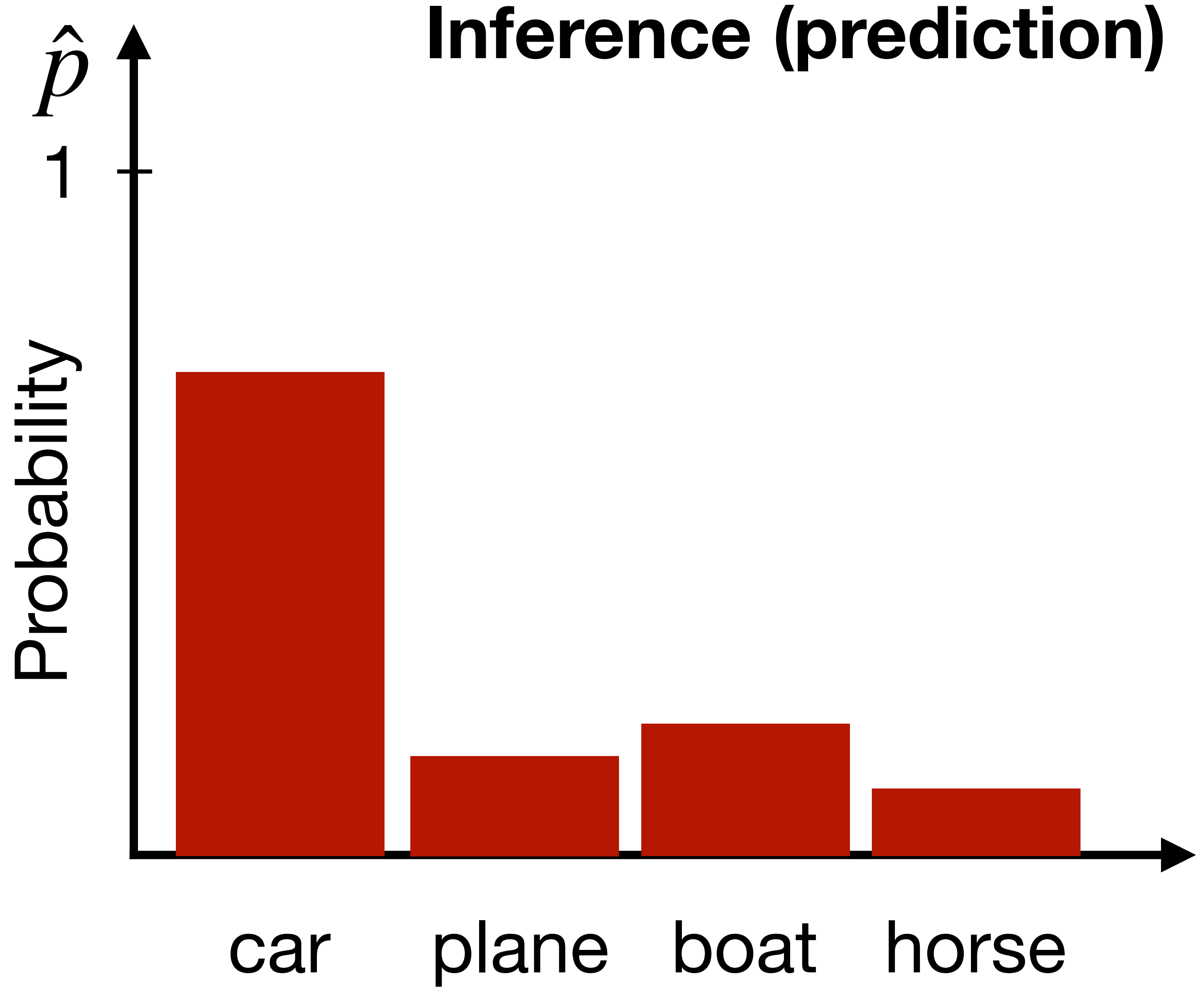
$$\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$$

car

x



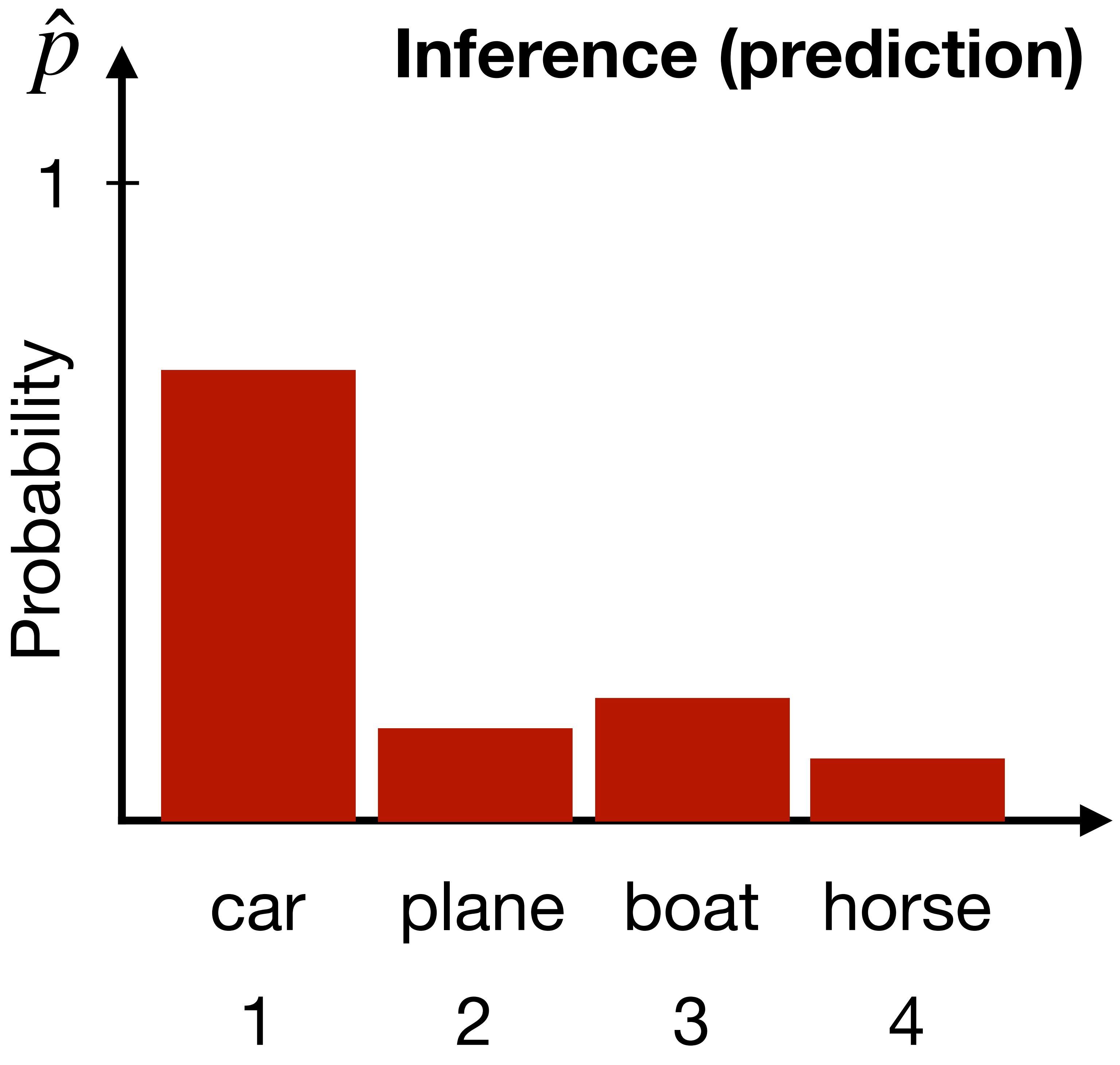
Inference (prediction)



$$\textbf{min} \quad \text{CrossEntropy}(p, \hat{p}) = -\sum_{i=1}^{k} p(y=i)\, \log\left(\hat{p}(y=i)\right)$$

$$= -\log\left(\hat{p}(y=1)\right)$$

How do you train?



$$\text{CrossEntropy}(p, \hat{p}) = -\sum_{i=1}^{k} p(y = i) \log\left(\hat{p}(y = i)\right)$$

Ground Truth

Logit

$$= -\log\left(\hat{p}(y = 1)\right)$$

$$= -\log\left(\frac{\exp\left(\vec{\theta}_i^\top x\right)}{\sum_{j=1}^{k} \exp\left(\vec{\theta}_j^\top x\right)}\right)$$

Train with Gradient Descent!