CIS 5200: MACHINE LEARNING

LINEAR AND LOGISTIC REGRESSION

Surbhi Goel

Content here draws from material by Vassal Sharan (USC), Christopher De Sa and Kilian Weinberger (Cornell)

Spring 2023

LOGISTICS - UPCOMING

Homework:

- HW0 due on **Friday, Jan 20, 2023** end of day
- For those on waitlist, email your HW0 to Keshav and Wendi (head TAs)
- HW1 will be out on Monday, Jan 23, 2023

Recitation:

- Sign up link will be posted on Ed this Friday
- Math background recitation next week

Instructor OH:

- Eric and I will run joint office hours after class on Tuesdays 3:30-4:30

OUTLINE - TODAY

- Quick Review of Perceptron
- Logistic Regression
  - MLE perspective
- Linear Regression
  - Least squares solution
  - MLE perspective
- Regularization

PERCEPTRON - SUMMARY

Input space: $\mathcal{X} \subseteq \mathbb{R}^d$

Output space: $\mathcal{Y} = \{-1, 1\}$

Hypothesis Class: $\mathcal{F} := \{x \mapsto \text{sign}(w^\top x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$

Loss function: $\ell(f(x), y) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise.} \end{cases}$

Assumption: Linearly separable data

Guarantee: Zero-error on training data after $1/\gamma^2$ iterations for margin γ

PERCEPTRON - FAILURES

XOR:

Led to the AI winter till mid 1980s

Minsky and Papert in a 1969 book "Perceptrons" showed that Perceptron fails on XOR problems

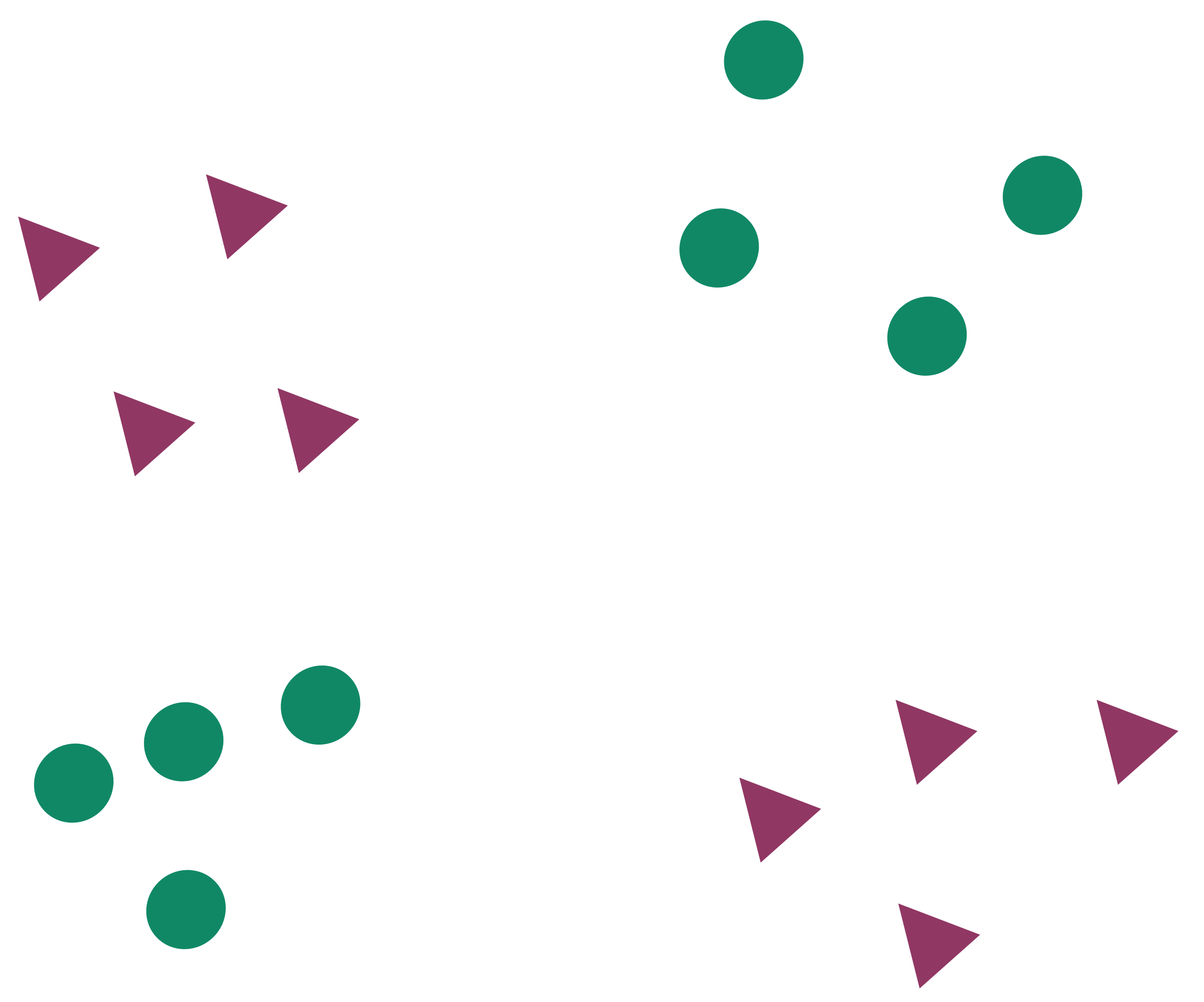
Non-linearly separable data:

Kernels (later in class)

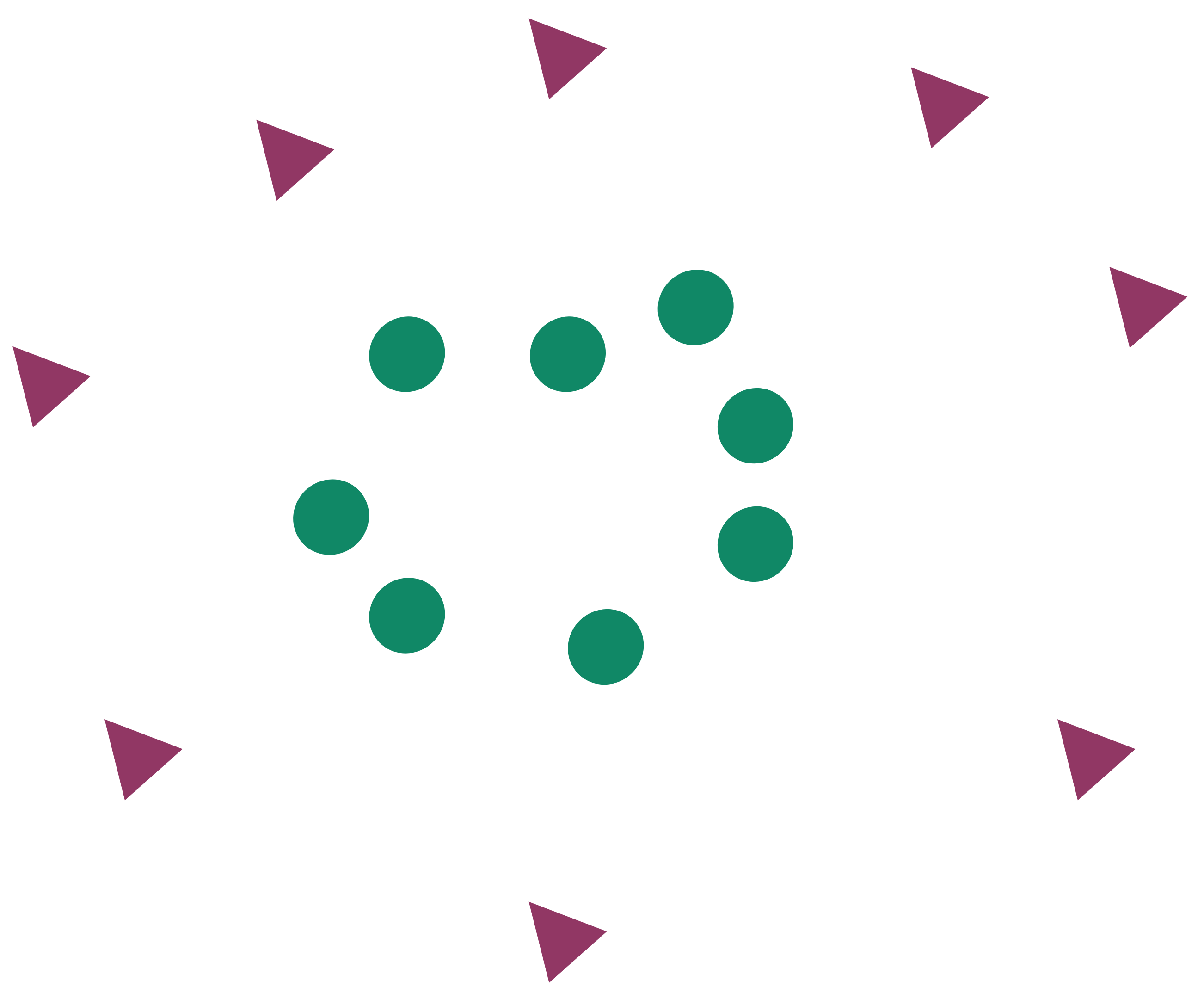
Separable in a lifted space

Noise:

Hard classifier, cannot model inherent noise



XOR

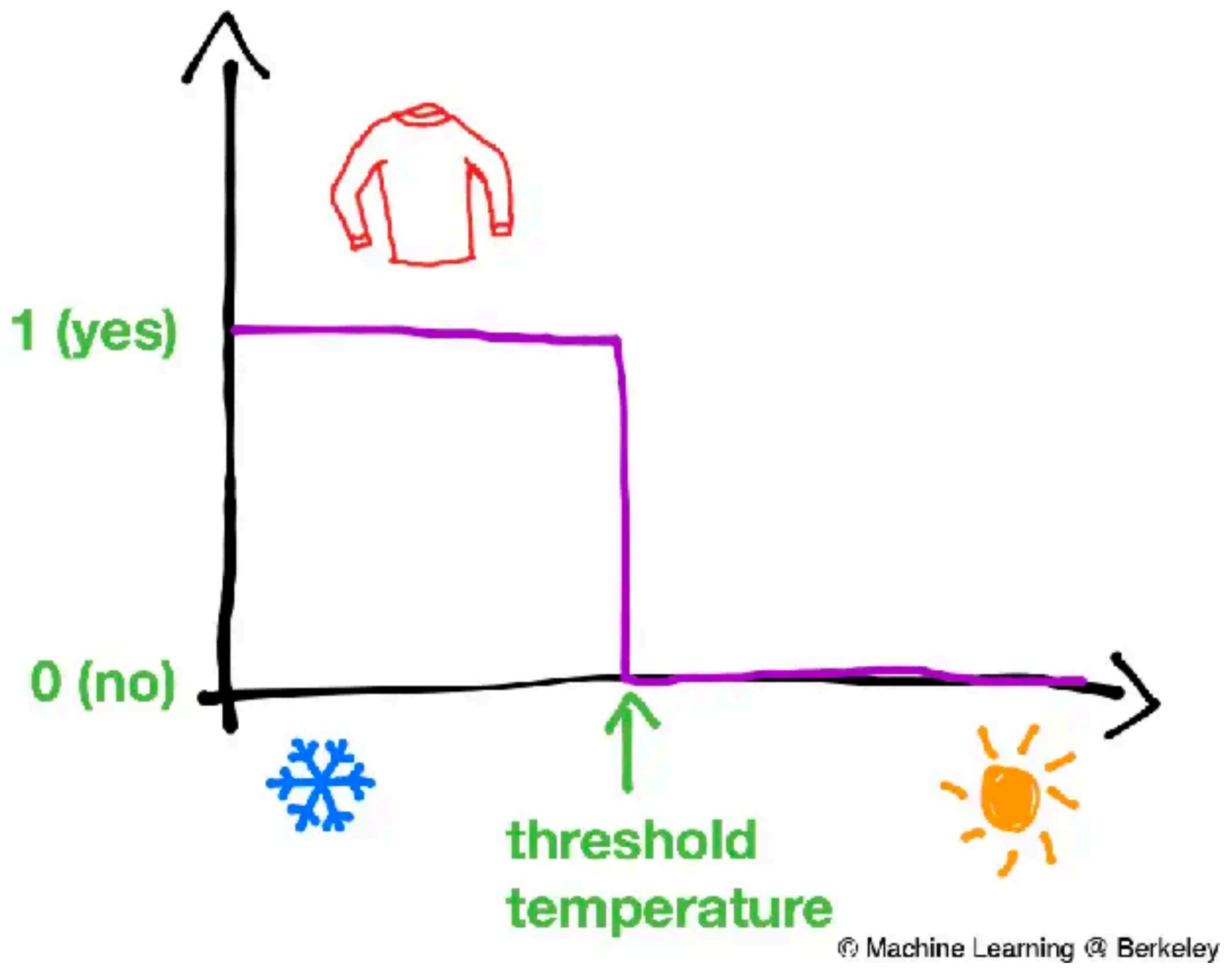


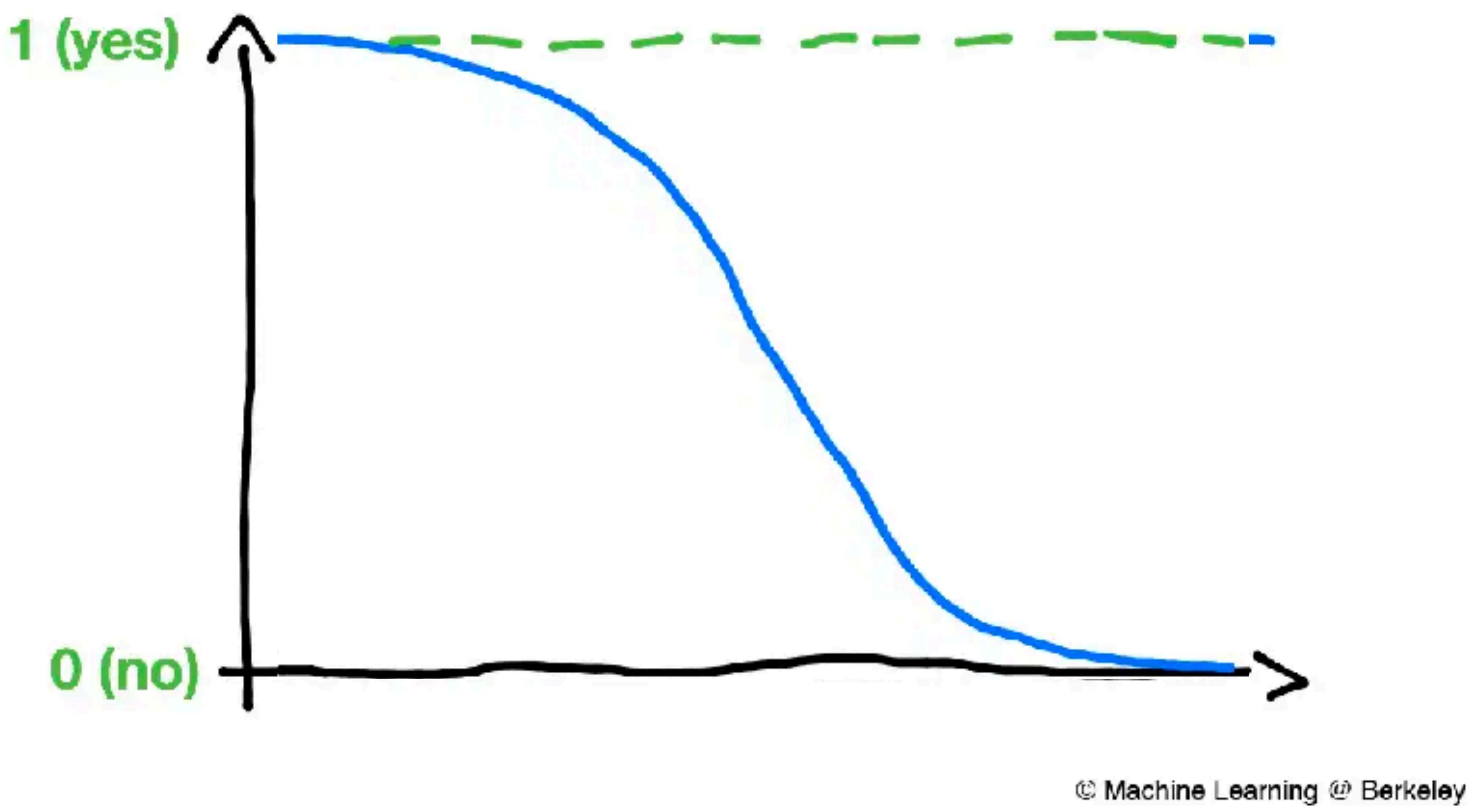
Non-separable Data

NON-DETERMINISTIC INPUTS

Perceptron assumed deterministic labels

But there may be inherent uncertainty in the label





We can model this uncertainty using some function $\eta(x) = P(y = 1 \mid x)$

LOGISTIC FUNCTION

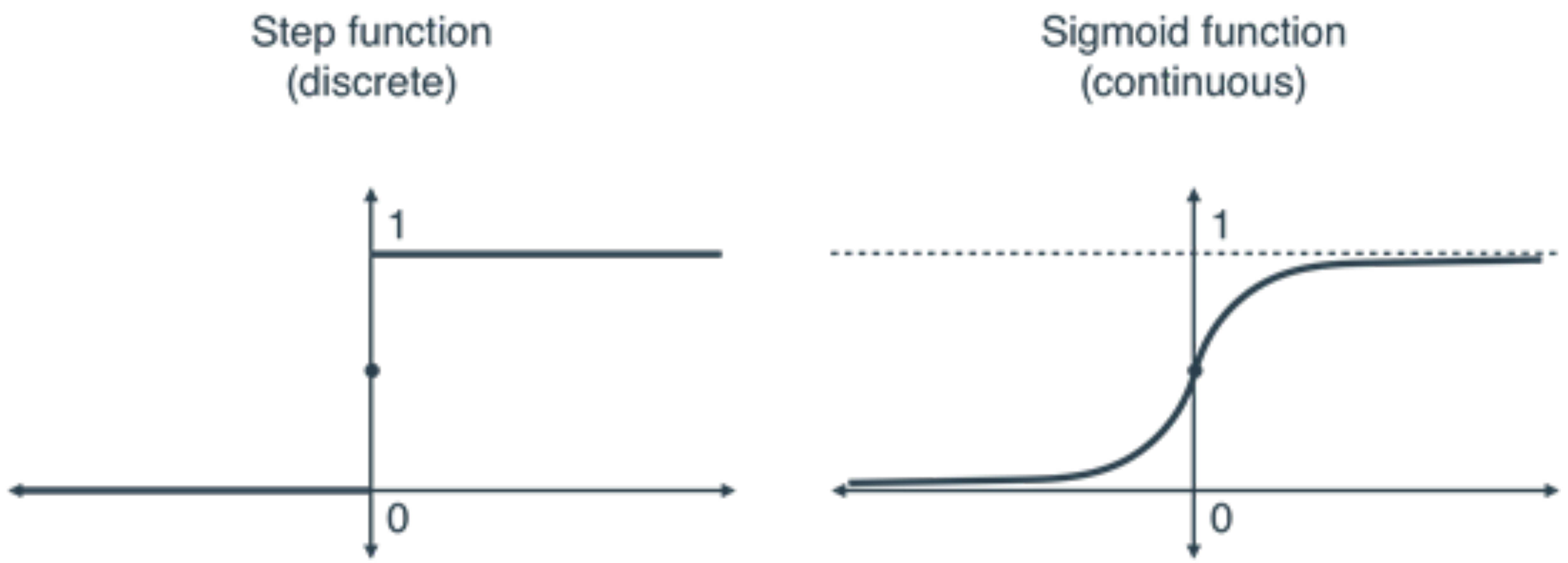
We can model $\eta(x) = P(y = 1 \mid x)$ using different functions

$$\mathrm{sign}_{0/1}(a) = \begin{cases} 1 & \text{if } a \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Step function

$$\mathrm{sigmoid}(a) = \frac{1}{1 + \exp(-a)}$$

Sigmoid function



$$P(y = 1 \mid x) = \eta(x) = \mathrm{sigmoid}(w^\top x) = \frac{1}{1 + \exp(-w^\top x)}$$

More unsure near the decision boundary

$$P(y = -1 \mid x) = 1 - \eta(x) = 1 - \mathrm{sigmoid}(w^\top x) = \frac{1}{1 + \exp(w^\top x)}$$

Like perceptron away from the decision boundary

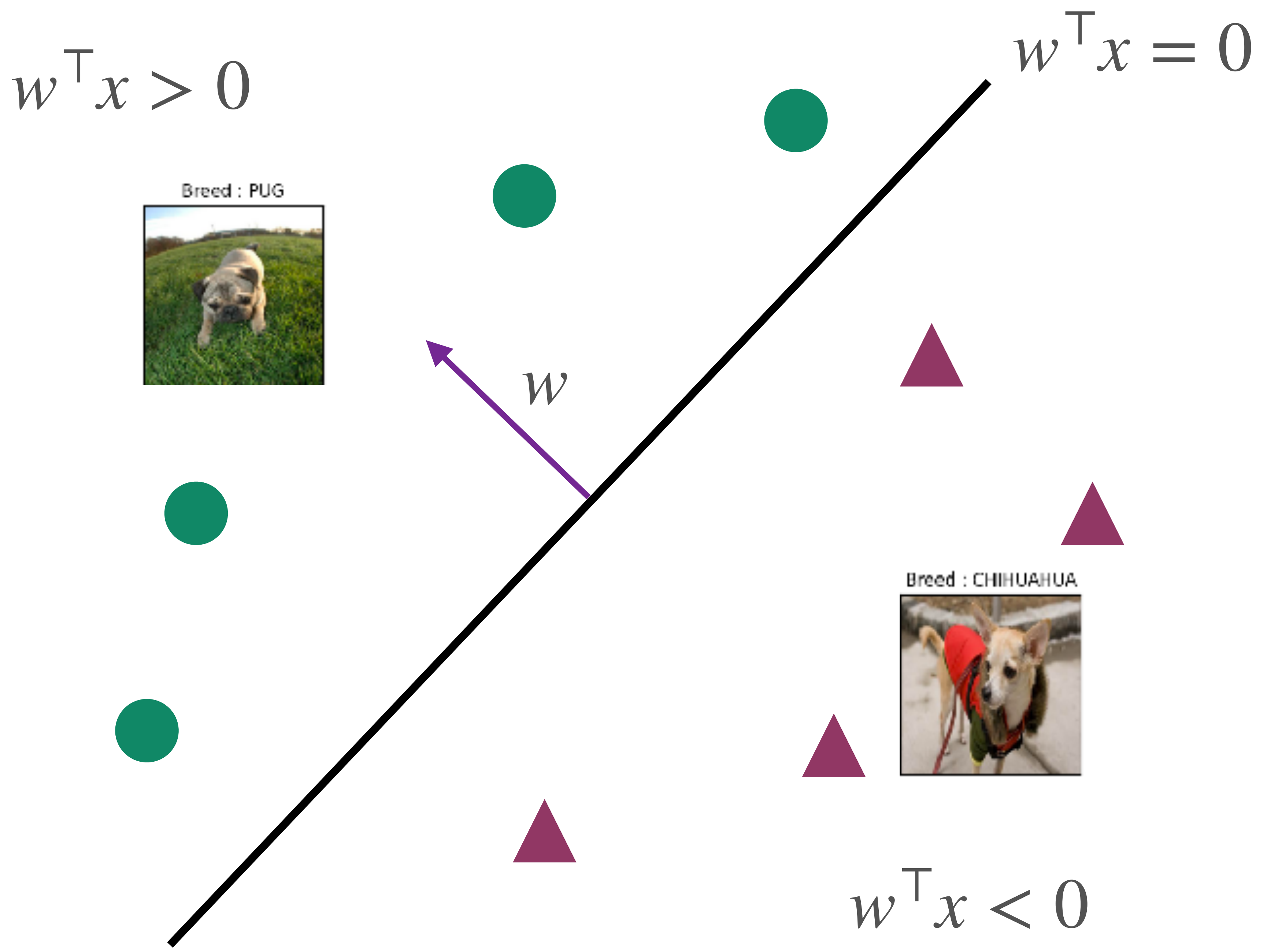
DECISION BOUNDARY

How do we decide the label given the logistic model?

$$\frac{P(y=+1\,|\,x)}{P(y=-1\,|\,x)} = \frac{1+\exp(w^\top x)}{1+\exp(-w^\top x)} = \exp(w^\top x)$$

$= 1$ *when* $w^\top x = 0$

Linear decision boundary



LOSS FUNCTION

Logistic Loss

$$\ell(f(x), y) = \begin{cases} -\log(f(x)) & \text{if } y = 1 \\ -\log(1 - f(x)) & \text{otherwise} \end{cases}$$

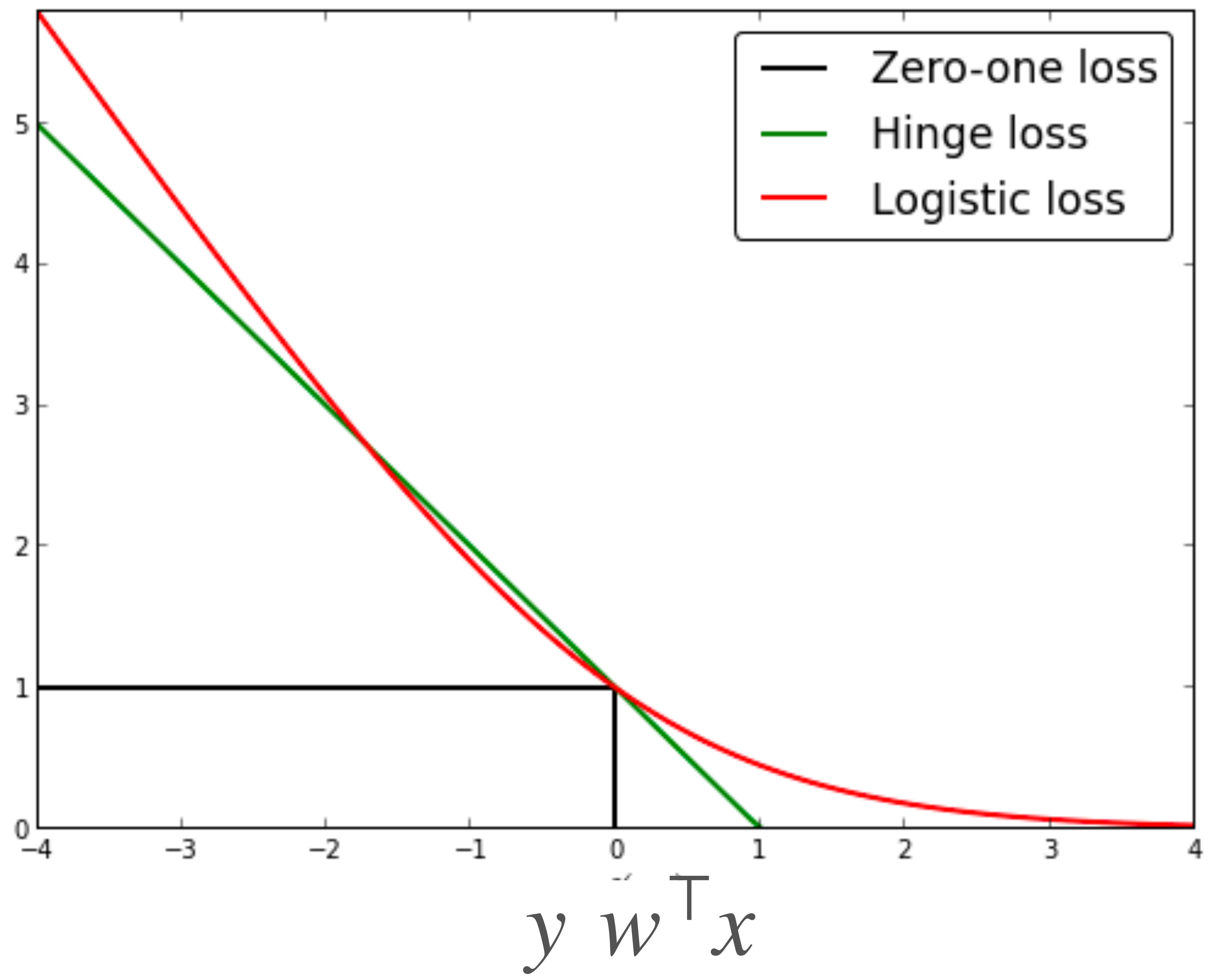
For our setting logistic loss is $\log\left(1 + \exp(-y\, w^\top x)\right)$

0/1 Loss

$$\ell_{0/1}(f(x), y) = 1[f(x) \neq y]$$

For linear classifier this is $1[\text{sgn}(w^\top x) \neq y] = 1[y\, w^\top x < 0]$

Why this loss?



Logistic loss is an upper bound of 0/1 loss

PROBABILISTIC VIEW - MAXIMUM LIKELIHOOD ESTIMATOR

Another way to view the supervised learning task is to maximize the probability of seeing the training data

- Make an explicit modeling condition on the data distribution
- Find parameters that maximize the probability of seeing the data

Suppose the parameters of the model are denoted by θ

$$\hat{\mathscr{L}}(\theta) = P(S \mid \theta)$$

$\boldsymbol{S}$ is the training data

$$= \prod_{i=1}^{m} P(x_i, y_i \mid \theta)$$

Training data is i.i.d.

MAXIMUM (CONDITIONAL) LOG LIKELIHOOD

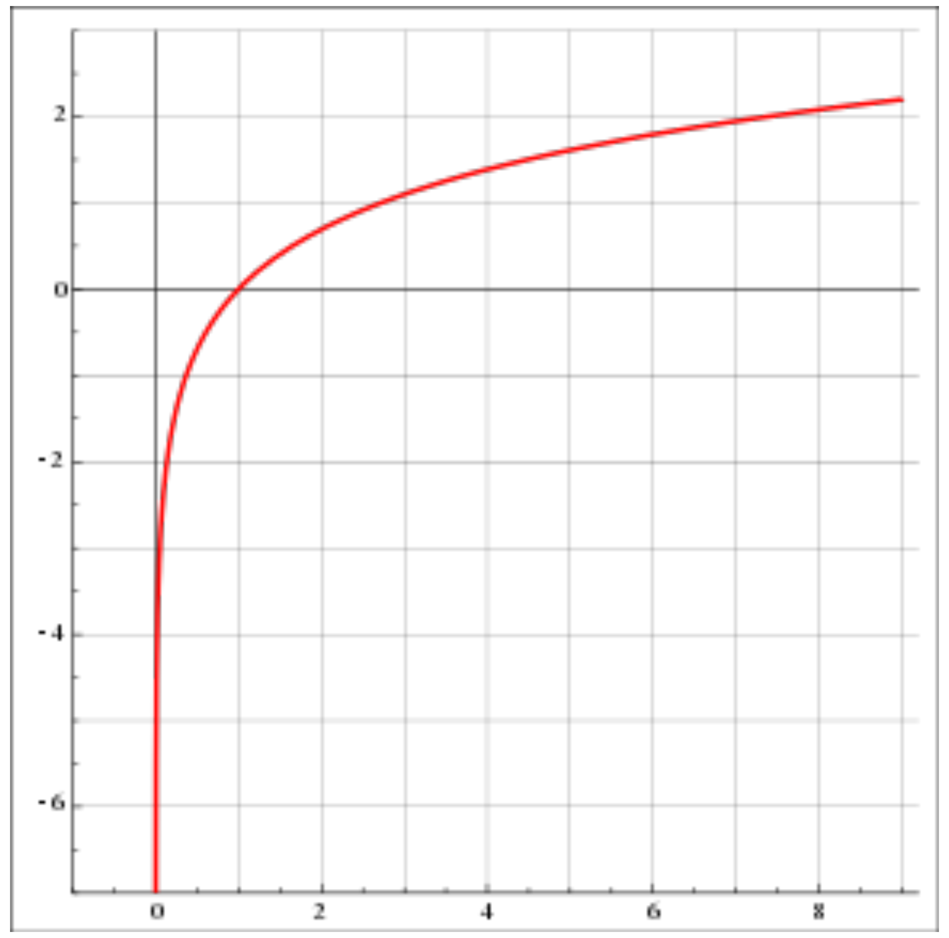
Suppose we don't have any assumption on the generation process of x, then we can maximize a conditional likelihood

$$\hat{\mathscr{L}}(\theta) = \prod_{i=1}^{m} P(y_i \mid x_i, \theta)$$

The log-likelihood is then equivalent to:

$$\log \hat{\mathscr{L}}(\theta) = \log \left(\prod_{i=1}^{m} P(y_i \mid x_i, \theta) \right)$$

$$= \sum_{i=1}^{m} \log \left(P(y_i \mid x_i, \theta) \right)$$



$\log$ is an increasing function
Maximizers of both are identical

M(C)LE - LOGISTIC REGRESSION

We have the model for $P(y\,|\,x, w)$, substituting it gives us

$$\log \hat{\mathscr{L}}(w) = \sum_{i=1}^{m} \log\left(P(y_i \mid x_i, w)\right)$$

$$= \sum_{i=1}^{m} \log\left(\frac{1}{1 + \exp(-y_i w^\top x_i)}\right)$$

$$= -\sum_{i=1}^{m} \log\left(1 + \exp(-y_i w^\top x_i)\right)$$

This is the negative of the logistic loss!

$$\max_{w} \log \hat{\mathscr{L}}(w) = \min_{w} \hat{R}(w)$$

LOGISTIC REGRESSION - TRAINING

Training Dataset: $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$,
$x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Empirical Risk Minimization: Find $\hat{w}$ that minimizes

$$\widehat{R}(w) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp(-y_i\, w^\top x_i)\right)$$

How do we solve this minimization problem?

The problem is convex so we can use convex optimization (will discuss in later lectures)

LOGISTIC REGRESSION - SUMMARY

Input space: $\mathcal{X} \subseteq \mathbb{R}^d$ Perceptron

Output space: $\mathcal{Y} = [0,1]$ $\mathcal{Y} = \{-1,1\}$

Hypothesis Class: $\mathcal{F} := \{x \mapsto \text{sigmoid}(w^\top x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$

$$\mathcal{F} := \{x \mapsto \text{sign}(w^\top x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

Loss function: $\ell(f(x), y) = \begin{cases} -\log(f(x)) & \text{if } y = 1 \\ -\log(1 - f(x)) & \text{otherwise} \end{cases}$

$$\ell(f(x), y) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise.} \end{cases}$$

SUPERVISED LEARNING

Predict future outcomes based on past outcomes

Inputs $x \in \mathcal{X}$

Labels $y \in \mathcal{Y}$

($\mathcal{Y}$ = Breeds)
"Pug"
"Chihuahua"

Classification

Discrete labels

($\mathcal{Y}$ = Stock prices)
"$130.02"

Regression

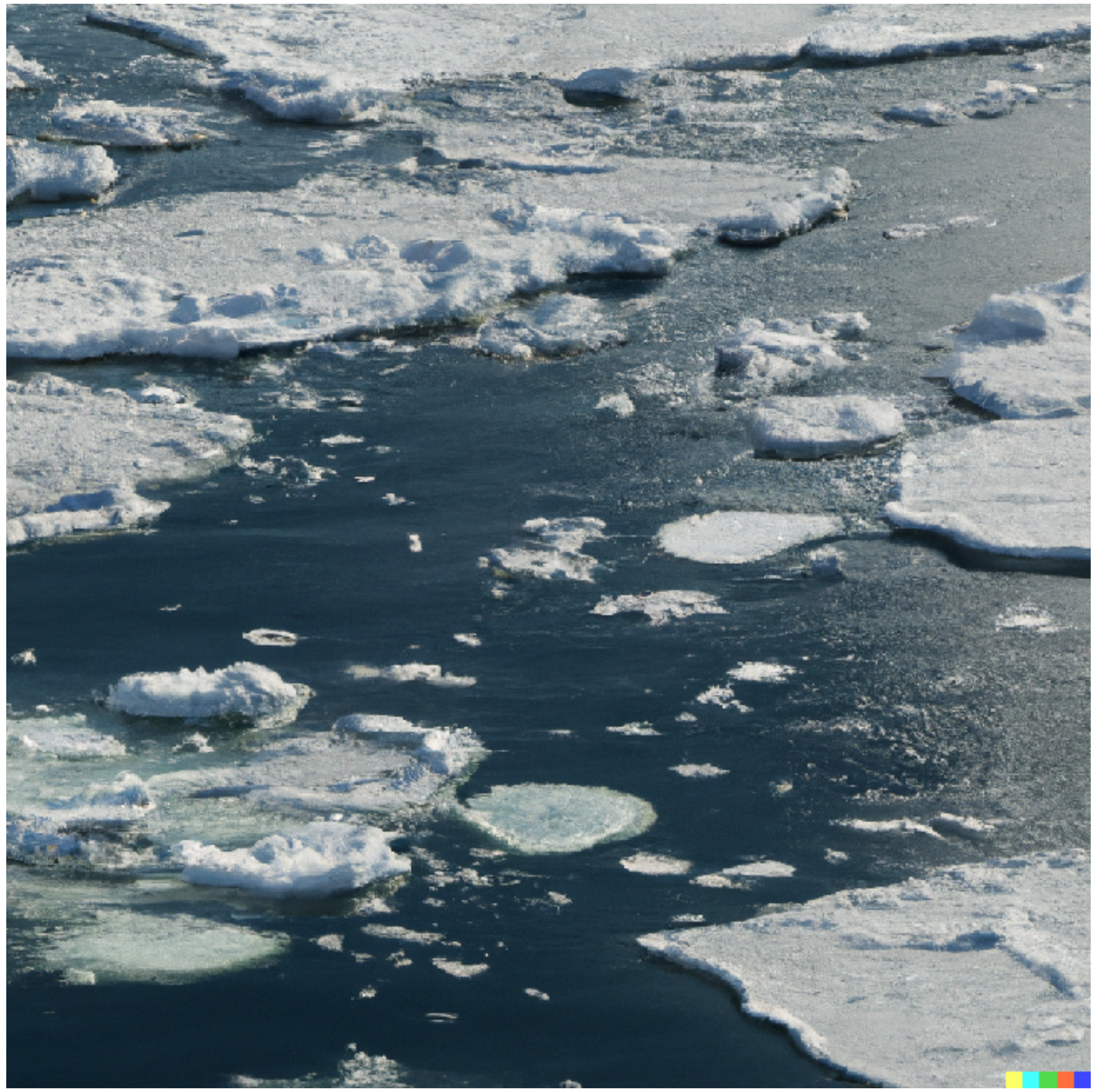
Continuous labels

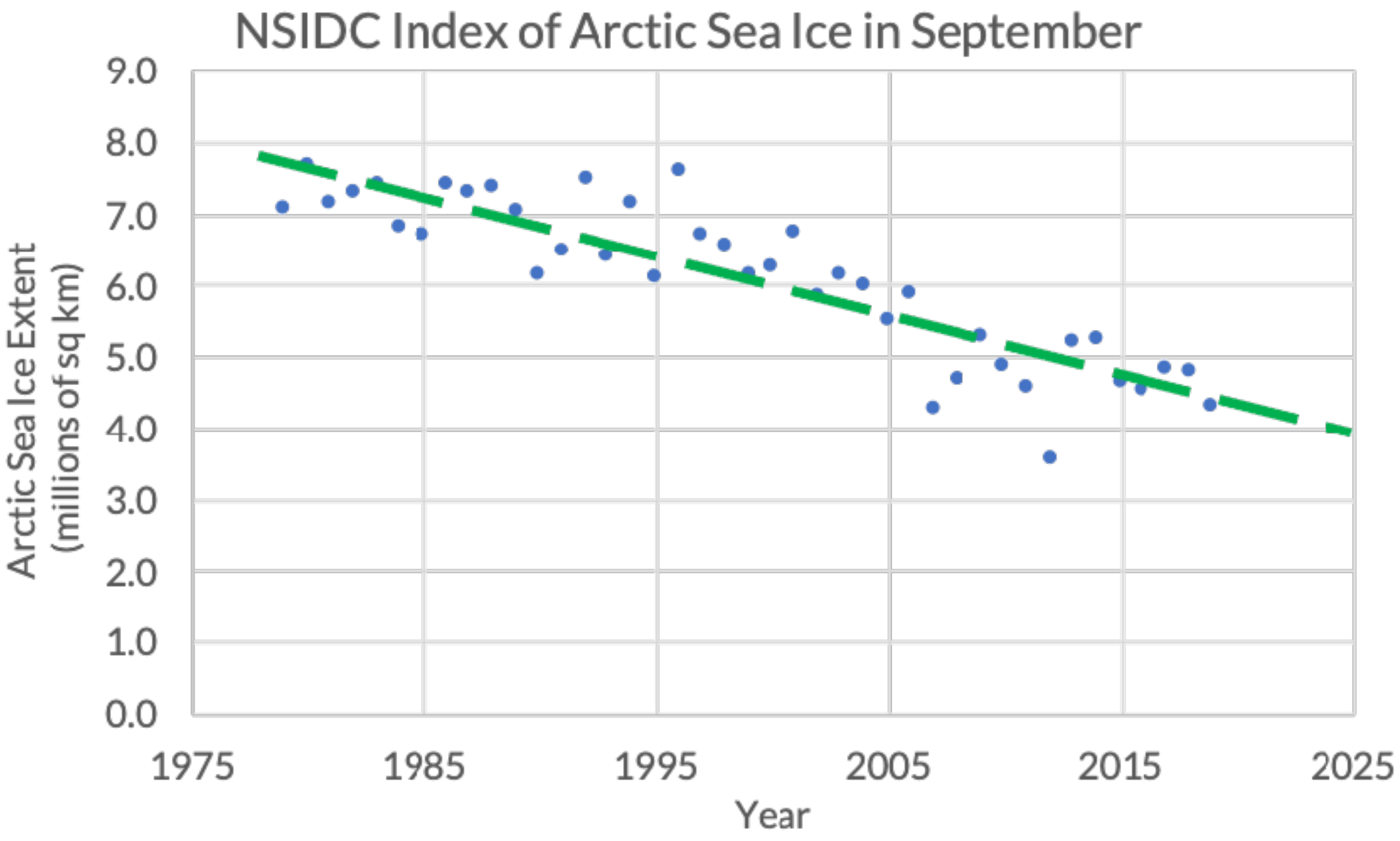
Task: Learn predictor $f : \mathcal{X} \to \mathcal{Y}$

HYPOTHESIS CLASS - LINEAR REGRESSORS

Similar to perceptron, can ignore bias

Linear regressors $\mathcal{F} := \{x \mapsto w^\top x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$





Data from https://nsidc.org/arcticseaicenews/sea-ice-tools/

LOSS FUNCTION

Square Loss

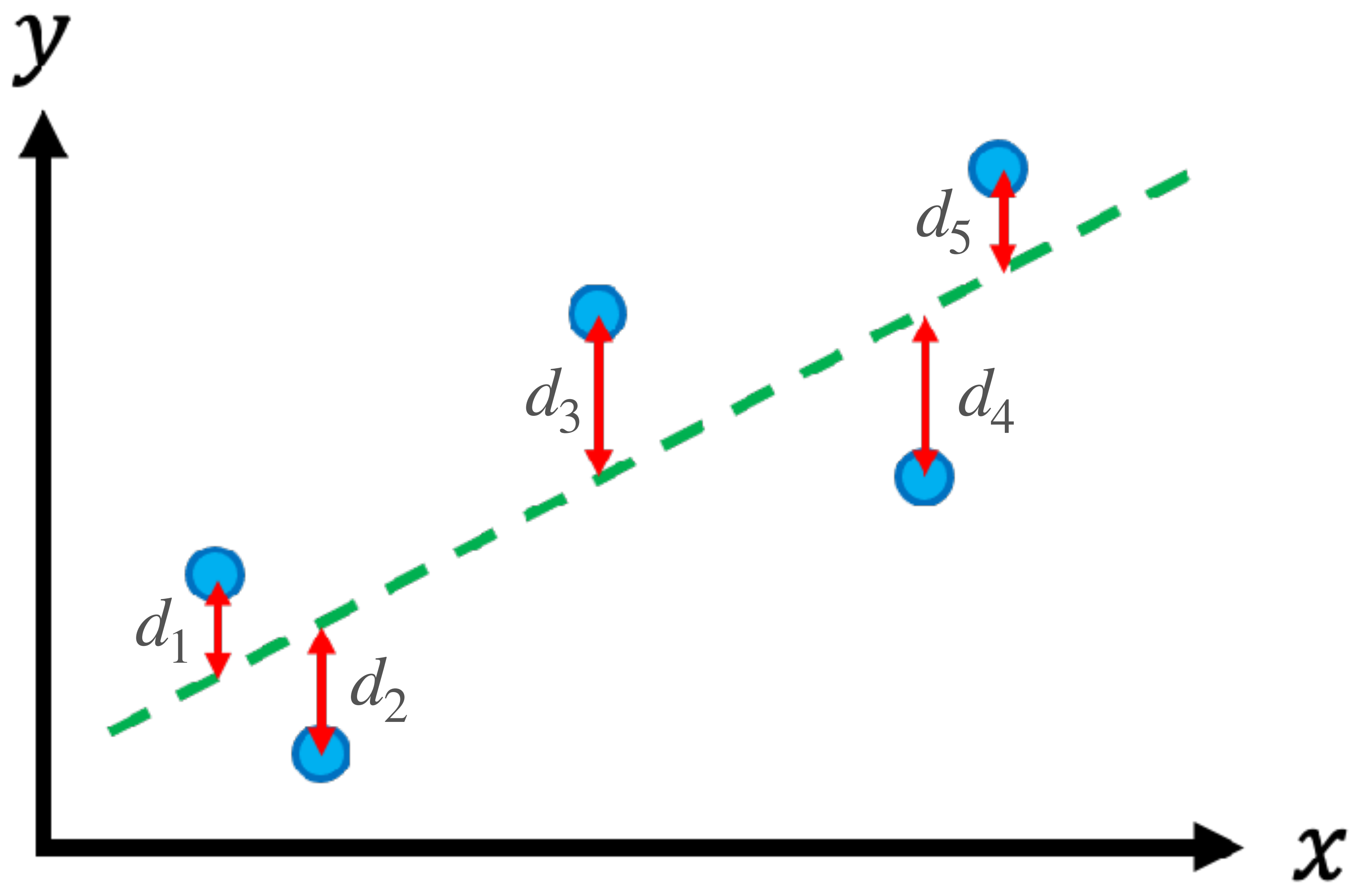
$$\ell(f(x), y) = (f(x) - y)^2$$

$$\text{Square-loss} = \frac{d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2}{5}$$

Absolute Loss

$$\ell(f(x), y) = |f(x) - y|^2$$

$$\text{Absolute-loss} = \frac{|d_1| + |d_2| + |d_3| + |d_4| + |d_5|}{5}$$



How does square loss behave on outliers?

LINEAR REGRESSION - TRAINING

Training Dataset: $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Empirical Risk Minimization: Find $\hat{w}$ that minimizes

$$\widehat{R}(w) = \frac{1}{m} \sum_{i=1}^{m} (y_i - w^\top x_i)^2$$

How do we solve this minimization problem?

The problem is convex, in fact we can get a closed form solution

LEAST SQUARES

Loss is convex $\implies$ differentiate to find minimizer

$$\widehat{R}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2$$

$\xrightarrow{\text{Take derivative and set to 0}}$

$$\frac{2}{m}\sum_{i=1}^{m}(w^\top x_i - y_i)x_i = 0$$

$$\implies \left(\sum_{i=1}^{m} x_i x_i^\top\right) w = \sum_{i=1}^{m} y_i x_i$$

$\downarrow$ Matrix notation

$$X^\top X w = X^\top Y$$

Normal Equations for Least Squares Regression

Let $X = \begin{bmatrix} - & x_1^\top & - \\ - & x_2^\top & - \\ & \vdots & \\ - & x_m^\top & - \end{bmatrix} \in \mathbb{R}^{m\times d}$, $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^{m\times 1}$

SOLVING THE SYSTEM

$$X = \begin{bmatrix} - & x_1^\top & - \\ - & x_2^\top & - \\ & \vdots & \\ - & x_m^\top & - \end{bmatrix} \in \mathbb{R}^{m \times d}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m$$

Normal Equations for Least Squares Regression

$$XX^\top w = X^\top Y$$

If $X^\top X$ is invertible, then

$$\hat{w} = (X^\top X)^{-1} X^\top Y$$

$\hat{Y} = X\hat{w}$ is the projection of Y onto the subspace spanned by $x_1, \ldots, x_m$

Recall that $X(X^\top X)^{-1}X$ is the projection matrix on to this subspace

What is the computational cost of computing this?

LINEAR REGRESSION - REGULARIZATION

What if $X^\top X$ is very close to being singular?

This can lead to large values for $\hat{w}$ which might overfit

$$\widehat{G}(w) = \widehat{R}(w) + \lambda\psi(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda\psi(w)$$

$\psi(w)$ is chosen to be some function that penalizes complexity of w

Common examples include: $\psi(w) = \|w\|_2^2$ or $\phi(w) = \|w\|_1$

RIDGE REGRESSION

$$X = \begin{bmatrix} - & x_1^\top & - \\ - & x_2^\top & - \\ & \vdots & \\ - & x_m^\top & - \end{bmatrix} \in \mathbb{R}^{m\times d}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^{m\times 1}$$

$$\widehat{G}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda\|w\|_2^2$$

Take derivative and set to 0

$$\frac{2}{m}\sum_{i=1}^{m}(w^\top x_i - y_i)x_i + 2\lambda w = 0$$

$$\implies \left(\sum_{i=1}^{m} x_i x_i^\top + \lambda I\right) w = \sum_{i=1}^{m} y_i x_i$$

Matrix notation

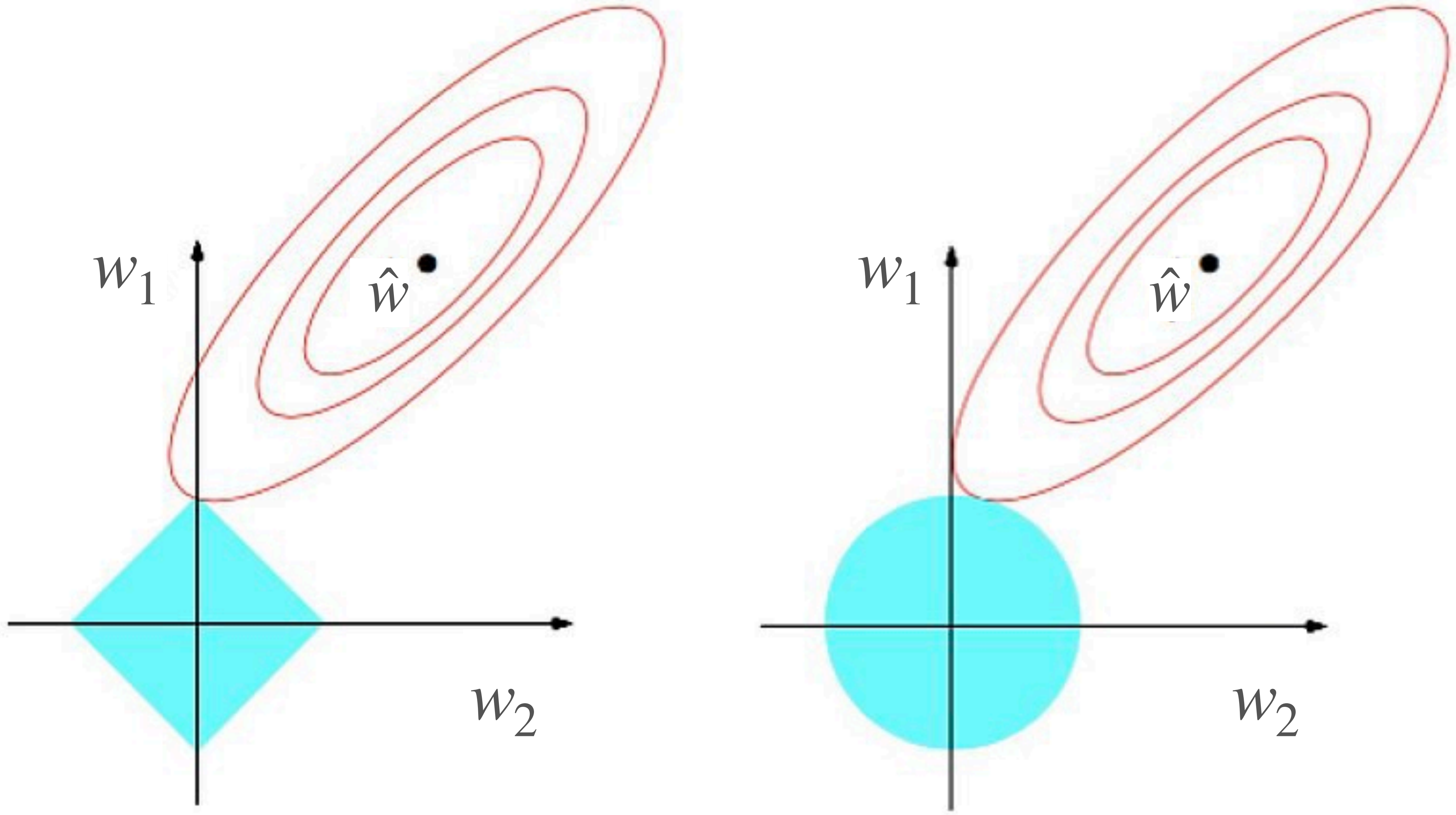
$$(X^\top X + \lambda I)w = X^\top Y$$

$$\hat{w}_\lambda = (X^\top X + \lambda I)^{-1} X^\top Y$$

Always invertible, eigenvalues are $\geq \lambda$

LASSO REGRESSION

$$\widehat{G}(w) = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda\|w\|_1$$



Leads to sparsity in the weights!

LINEAR REGRESSION - SUMMARY

Input space: $\mathcal{X} \subseteq \mathbb{R}^d$

Output space: $\mathcal{Y} = \mathbb{R}$

Hypothesis Class: $\mathcal{F} := \{x \mapsto w^\top x + b \,|\, w \in \mathbb{R}^d, b \in \mathbb{R}\}$

Loss function: $\ell(f(x), y) = (f(x) - y)^2$

Least Squares solution: $\hat{w} = (X^\top X)^{-1} X^\top Y$