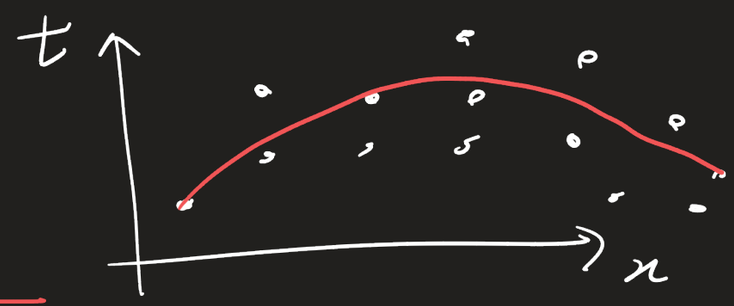
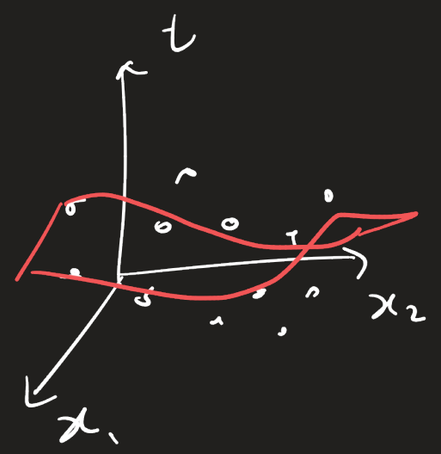
Supervised Learning for Regression

Tasks





Linear Regression

① INPUT SPACE

Input data: $\{x_i\}_{i=1}^{N}$, $N \equiv$ # samples, $x_i \in \mathbb{R}$

TARGET values: $\{t_i\}_{i=1}^{N}$

DATASET: $\{(x_i, t_i)\}_{i=1}^{N}$

[in some cases $x_i \in \mathbb{R}^D$]

$t_i \in \mathbb{R}$

② FEATURE SPACE

Polynomial basis function:

$$\phi(x_i) = [x_i^0, x_i^1, x_i^2, \dots, x_i^M]^T \in \mathbb{R}^{(M+1)}$$

③ MODEL / MAPPER :

$$y_i = f(\phi(x_i), w)$$

$w \equiv$ generalized notation for model parameters

$$y_i = w_0 \cdot x_i^0 + w_1 \cdot x_i^1 + \cdots + w_M \cdot x_i^M$$

$$= w^T \cdot \phi(x_i)$$

$$= \phi(x_i)^T \cdot w$$

$f \equiv$ Linear regression model

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}$$

$$\phi(x_i) = \begin{bmatrix} x_i^0 \\ x_i^1 \\ \vdots \\ x_i^M \end{bmatrix}$$

$y_i \equiv$ model prediction for i-th sample

$$
\begin{cases}
y_1 = w_0 \cdot x_1^0 + w_1 \cdot x_1^1 + \dots + w_M \cdot x_1^M \\
y_2 = w_0 \cdot x_2^0 + w_1 \cdot x_2^1 + \dots + w_M \cdot x_2^M \\
\vdots \qquad\qquad \vdots \\
y_N = w_0 \cdot x_N^0 + w_1 \cdot x_N^1 + \dots + w_M \cdot x_N^M
\end{cases}
$$

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_1^0 & x_1^1 & \dots & x_1^M \\ x_2^0 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^2 & \dots & x_N^M \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}
$$

output vector $\equiv y$

FEATURE MATRIX $\equiv X$

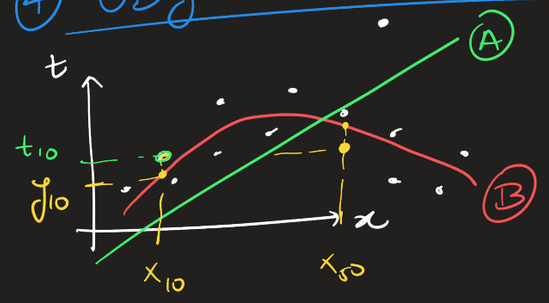
$N \times (M+1)$

Model parameters $\equiv w$

MODEL / MAPPER:

$$\boxed{y = X \cdot w}$$

④ OBJECTIVE FUNCTION:



$$\varepsilon_i = t_i - y_i$$

$$\boxed{J(w)} = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} (t_i - y_i)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} \left(t_i - w^T \phi(x_i)\right)^2$$

$$= \boxed{\frac{1}{N}(t - Xw)^T (t - Xw)}$$

$$= \frac{1}{N}\|t - Xw\|_2^2$$

MEAN least squares (MSE)

Reminder:

$\|x\|_2 = \|x\|$

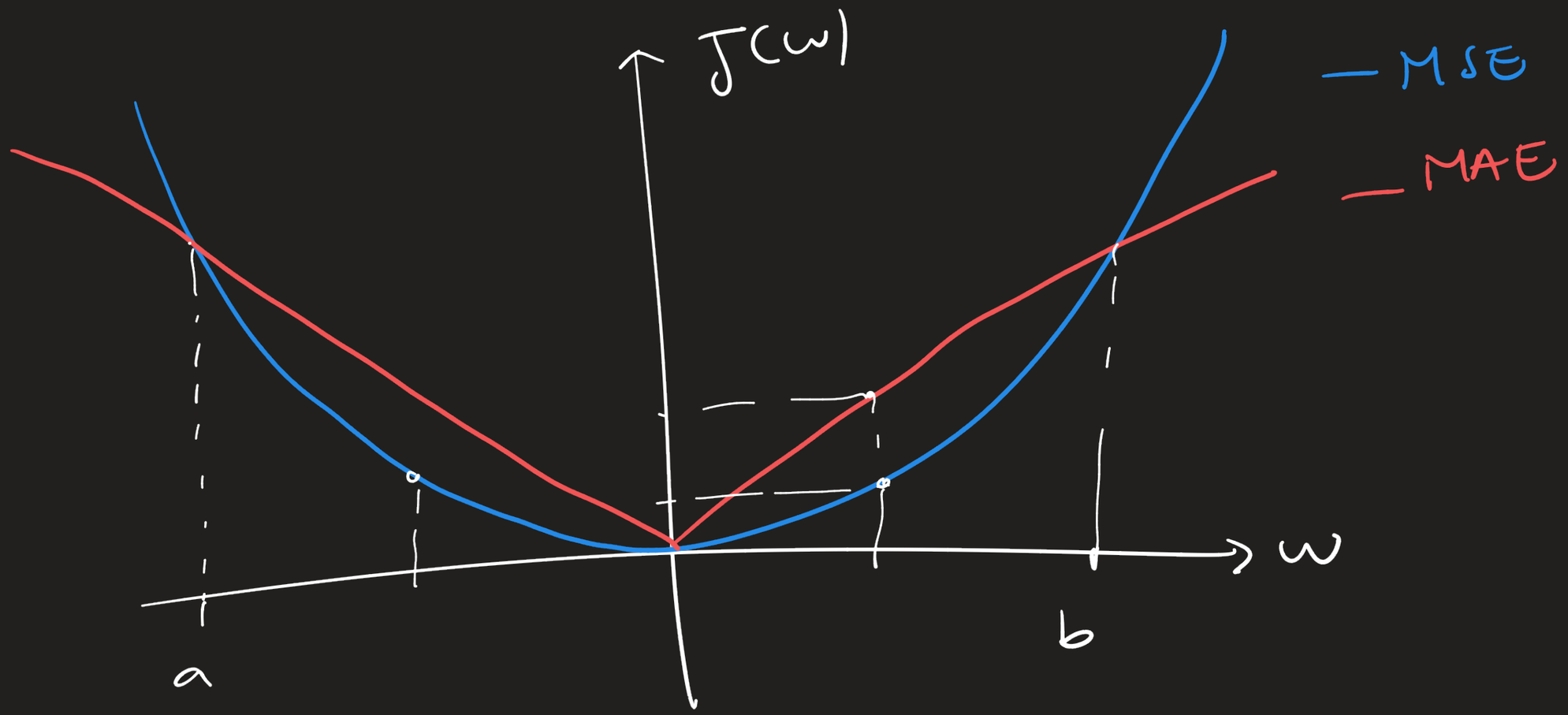
$= (x_1^2 + x_2^2 + \cdots + x_N^2)^{1/2}$

$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_N|$

An alternative, consider sum of absolute error for $\forall i$

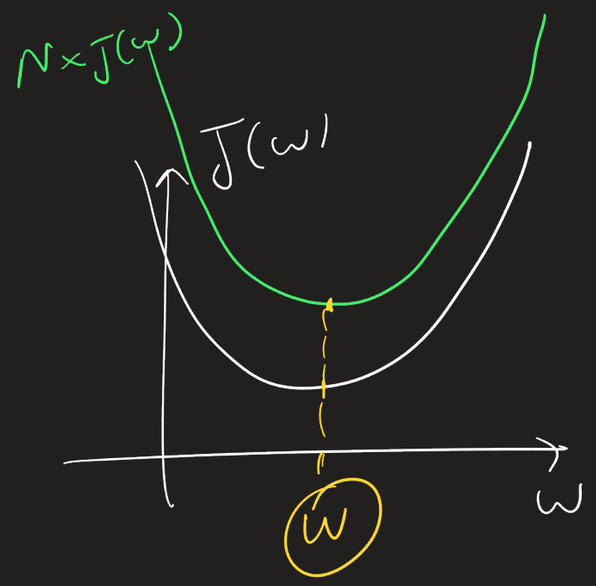
$$J(w) = \frac{1}{N}\|t - Xw\|_1$$

MEAN ABSOLUTE ERROR (MAE)


J(w)
MSE
MAE
w
b
a

⑤ Learning Algorithm

Finds the best solution for model parameters, w.



$$\frac{\partial J}{\partial w} = 0$$

$$J(w) = \frac{1}{N}(t - Xw)^T(t - Xw)$$

$$= \frac{1}{N}(t^T - w^T X^T)(t - Xw)$$

$$= \frac{1}{N}(t^T t - t^T X w - w^T X^T t + w^T X^T X w)$$

$$\frac{\partial J}{\partial w} = -t^T X - (X^T t)^T + (X^T X w)^T + w^T X^T X$$

$$= -t^T X - t^T X + w^T X^T X + w^T X^T X$$

$$= -2t^T X + 2 w^T X^T X = 0$$

$$-2\,t^T X + 2\,w^T X^T X = 0$$

$$w^T X^T X = t^T X$$

$$X^T X\, w = X^T t$$

Reminder

cyclic operation

$$(X^T a)^T = a^T (X^T)^T$$

$$\boxed{w = (X^T X)^{-1} X^T t}$$

if $(X^T X)^{-1}$ (matrix inverse) exists.

If X^TX is invertible:

1) $\det(X^TX) \neq 0$

2) Full rank matrix

3) Eigenvalues of X^TX they are all $\neq 0$.

4) Features are all linearly independent.

$$X = \left[\begin{array}{c|c|c} f_1 & f_2 & f_3 \end{array}\right]$$

$$= \begin{bmatrix} 1 & 0 & 3.5 \\ 0 & 1 & 4.2 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

$f_1 \equiv$ male $\in \{0,1\}$

$f_2 \equiv$ female $\in \{0,1\}$

$f_3 \equiv$ length $\in \mathbb{R}$

If we have 3 features

then we want them to

Exist in a 3-D basis.

$$B = \{(1,0,0), (0,1,0), (0,0,1)\}$$

NEW

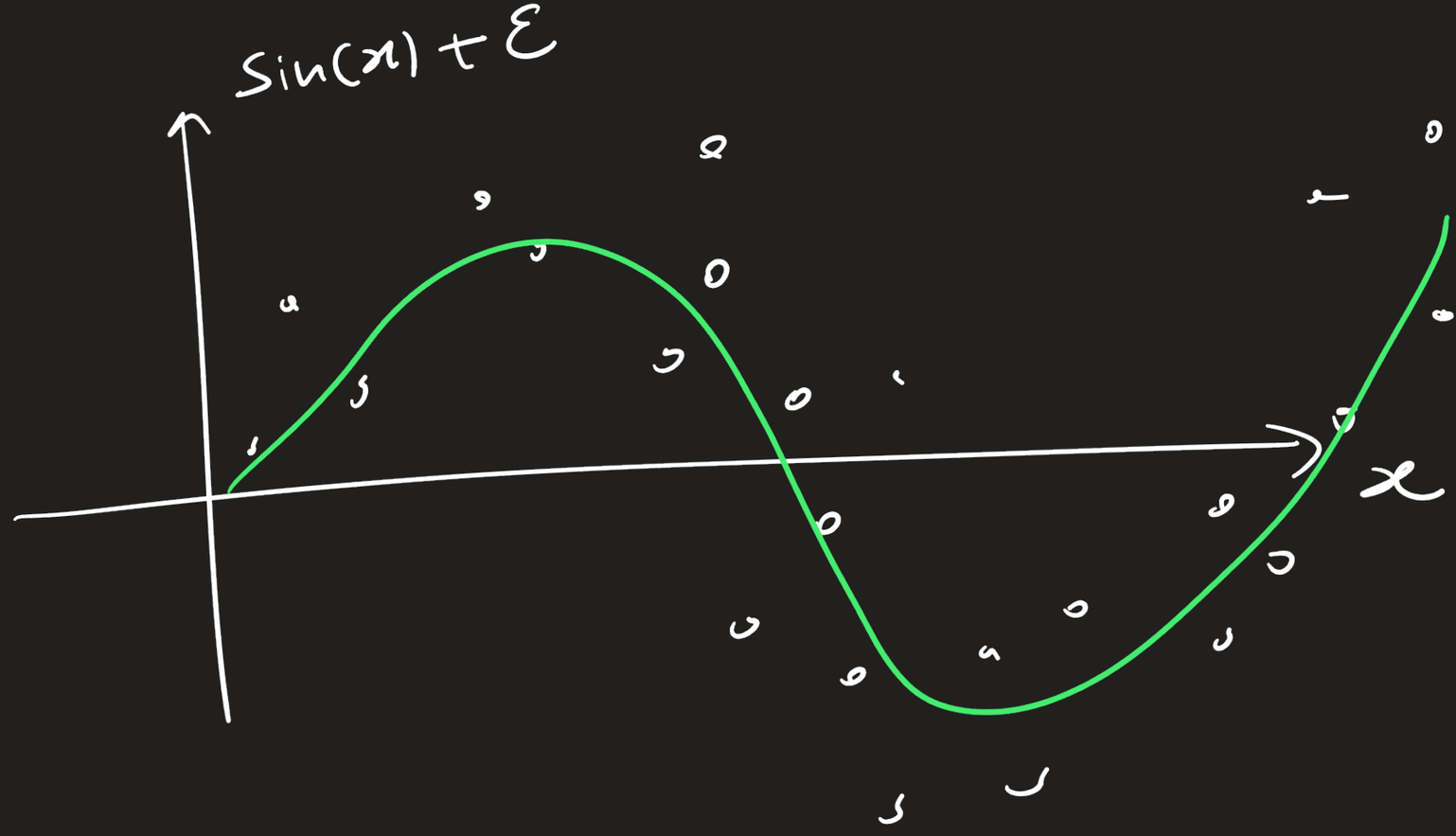
$$X^TX = X^TX + \lambda \cdot I$$

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Diagonally - loading

Hyperparameter ⟵ provided by the user.

M ≡ degree of polynomial



— Sin(x)
TRUE MAPPING function

• training samples

OVERfitting

① Model is NOT able to generalize to new and unseen data.

② The parameters ↑ in magnitude

Regularization

$$R_{L_2}(w) = \sum_{j=0}^{M} w_j^2$$

$$= w^T w$$

$$= \|w\|_2^2$$

L2-REGULARIZER OR RIDGE REGULARIZER

→ "prefers" to have small values for all w's.

$$R_{L_1}(w) = \sum_{j=0}^{M} |w_j|$$

$$= \|w\|_1$$

L_1 – REGULARIZER OR LASSO REGULARIZER

→ promotes sparsity of parameters vector w

Include the Regularizer in the objective function.

$$J(w) = \|t - Xw\|_2^2 + \lambda \|w\|_2^2$$

Squared Error

Regularizer

Ridge Regularizer

λ Regularizer hyperparameter

$\lambda \to 0$: does NOT regularize model

$\lambda \to \infty$: it will, in the limit, only consider 1 range of values for w.

$$J(w) = (t - Xw)^T (t - Xw) + \lambda \cdot w^T w$$

$$= t^T t - t^T Xw - w^T X^T t + w^T X^T X w + \lambda w^T w$$

$$\frac{\partial J}{\partial w} = -2 t^T X + 2 w^T X^T X + 2 \lambda \cdot w^T = 0$$

$$\lambda w^T + w^T X^T X = t^T X$$

$$\lambda w \cdot I + X^T X w = X^T \cdot t \qquad \left| w \cdot I = w \right.$$

$$(\lambda \cdot I + X^T X) \cdot w = X^T \cdot t$$

$$\boxed{w = (\lambda \cdot I + X^T X)^{-1} \cdot X^T \cdot t}$$

$$\underset{1 \times 1}{\lambda} + \underset{(n+1) \times (n+1)}{X^T X}$$

diagonally loading $X^T X$