DATASET : $\{(x_i, t_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R} \equiv$ samples
$t_i \in \mathbb{R} \equiv$ labels

FEATURES

Polynomial Basis fct. :

$$\phi : \mathbb{R} \longrightarrow \mathbb{R}^{M+1}$$
$$x \longmapsto \begin{bmatrix} x^0 \\ x^1 \\ \vdots \\ x^M \end{bmatrix}$$

MAPPER fct :

$$y(x_i) = \sum_{j=0}^{M} w_j \cdot x_i^j \quad , \forall i$$

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

FEATURE MATRIX : X is $N \times (M+1)$

$$X = \begin{bmatrix} x_1^0 & x_1^1 & x_1^2 & \cdots & x_1^M \\ x_2^0 & x_2^1 & x_2^2 & \cdots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & x_N^2 & \cdots & x_N^M \end{bmatrix}$$

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

OBJECTIVE fet : $J(w) = \frac{1}{2} \sum_{i=1}^{N} e_i^2$

where $e_i = t_i - y(x_i) = t_i - y_i$

$$J(w) = \frac{1}{2} \| t - X.w \|_2^2$$

t is the labels vector, $N \times 1$, $t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$

Solution for parameters w :

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

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

$(M+1) \times 1$

X is $N \times (M+1)$

Assuming the inverse of $X^T X$ exists

PSEUDO-CODE for LINEAR REGRESSION:

TRAINING FUNCTION

INPUT: input DATA $\{x_i\}_{i=1}^N = x$

target values $\{t_i\}_{i=1}^N = t$

hyperparameter $M \in \mathbb{R}$

(1) Compute the feature matrix X

(2) Compute the parametric solution for w:

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

(3) Predict labels for samples x

$$y = X \cdot w$$

(4) Compute the residual ERROR

$$\varepsilon = t - y$$

RETURN w, y, ε

DURING test ...

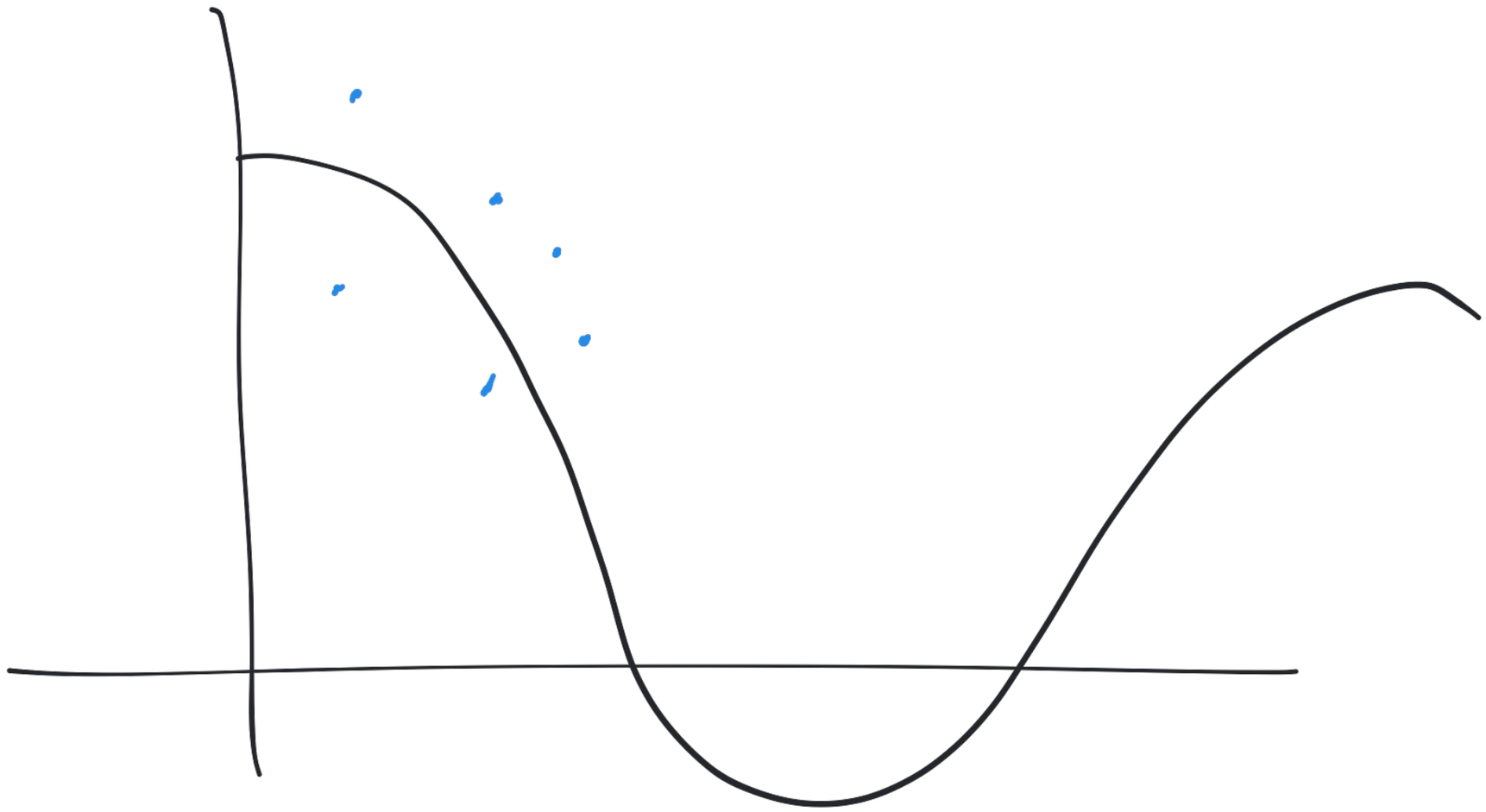
INPUT: TEST samples $\{z_i\}_{i=1}^{N} = z$

(trained) parameters w.

① Construct feature matrix for test samples

$$Z_1 = \begin{bmatrix} z_1^0 & z_1^1 & \cdots & z_1^M \\ z_2^0 & z_2^1 & \cdots & z_2^M \\ \vdots & \vdots & \ddots & \vdots \\ z_N^0 & z_N^1 & \cdots & z_N^M \end{bmatrix}$$

② $y_{TEST} = Z_1 \cdot w$



OBJECTIVE function:

will have 2 terms

ERROR term + REGULARIZATION term

$$J(w) = \frac{1}{2}\|t - Xw\|_2^2 + \lambda \sum_{j=0}^{N} w_j^2$$

λ: hyperparameter

Ridge regularizer

$\lambda \to 0$: no penalty on the parameters w

$\lambda \to \infty$: forces the parameters to be small.

Lasso regularizer: $R^{(L1)}(w) = \sum_{j=0}^{M} |w_j|$

or

(L1-penalty)

Ridge regularizer: $R^{(L2)}(w) = \sum_{j=0}^{M} w_j^2$

or

(L2-penalty)

Elastic Net: $R^{(EN)}(w) = \beta \sum_{j=0}^{M} |w_j| + (1-\beta) \cdot \sum_{j=0}^{M} w_j^2$

β is an additional hyperparameter for Elastic Net

Regularizer is also known as a penalty term.

OBSERVATIONS:

Lasso: ① Drive the parameters to exactly 0 much faster than ridge

② promotes sparsity
↳ a form of feature selection

Ridge:

① will apply a stronger penalty to outliers compared to Lasso.

② will "force" parameters to be smaller but never 0.

① $w = [0.5, 0.5, 1]$

$\|w\|_2^2 \equiv \text{ridge} = 0.5^2 + 0.5^2 + 1^2 = 1.5$

$\|w\|_1 \equiv \text{lasso} = |0.5| + |0.5| + |1| = 2$

② $w = [0, 0, 2]$

$\|w\|_2^2 = 0^2 + 0^2 + 2^2 = 4 \equiv \text{ridge}$

$\|w_1\| = |0| + |0| + |2| = 2 \equiv \text{lasso}$

$$y = w_0 + \cancel{w_1 x}^{=0} + w_2 x^2 + \cancel{w_3 x^3}^{=0}$$

$$y = w_0 + w_2 x^2$$

OBJective function w/ Ridge Regularization

$$J(w) = \frac{1}{2} \| t - Xw \|_2^2 + \frac{\lambda}{2} \cdot \| w \|_2^2$$

optimization problem

$$\arg\min_w J(w)$$

$$J(w) = \frac{1}{2}(t - Xw)^T(t - Xw) + \frac{\lambda}{2} w^T w$$

$$= \frac{1}{2}(t^T - w^T X^T)(t - Xw) + \frac{\lambda}{2} w^T w$$

$$= \frac{1}{2}(t^T t - t^T X w - w^T X^T t + w^T X^T X w) + \frac{\lambda}{2} w^T w$$

$$\frac{\partial J(w)}{\partial w} = 0 \Leftrightarrow \frac{1}{2}(-t^T X - (X^T . t)^T + (X^T X w)^T + w^T X^T X) + \frac{\lambda}{2} . w^T + \frac{\lambda}{2} . w^T = 0$$

$$\Leftrightarrow -t^T . X + w^T X^T X + \lambda . w^T = 0$$

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

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

$\Leftrightarrow$ applying transpose on both sides

identity matrix $(M+1) \times (M+1)$

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

if inverse exists

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

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

If $(X^T X)^{-1}$ does not exist

① $\det(X^T X) = 0$

② $X^T X$ is NOT full rank

③ $X^T X$ has linearly dependent columns

④ this will happen when $M > N$.

⑤ solution: diagonally-load $X^T X$ before inverting it

↳ this is the solution for LEAST SQUARES w/ RIDGE REG.

PSEUDO-CODE:

INPUT: samples $x = \{x_i\}_{i=1}^{N}$

target $t = \{t_i\}_{i=1}^{N}$

hyperparameters M and λ

① Compute feature matrix Φ

② Compute solution for w:

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

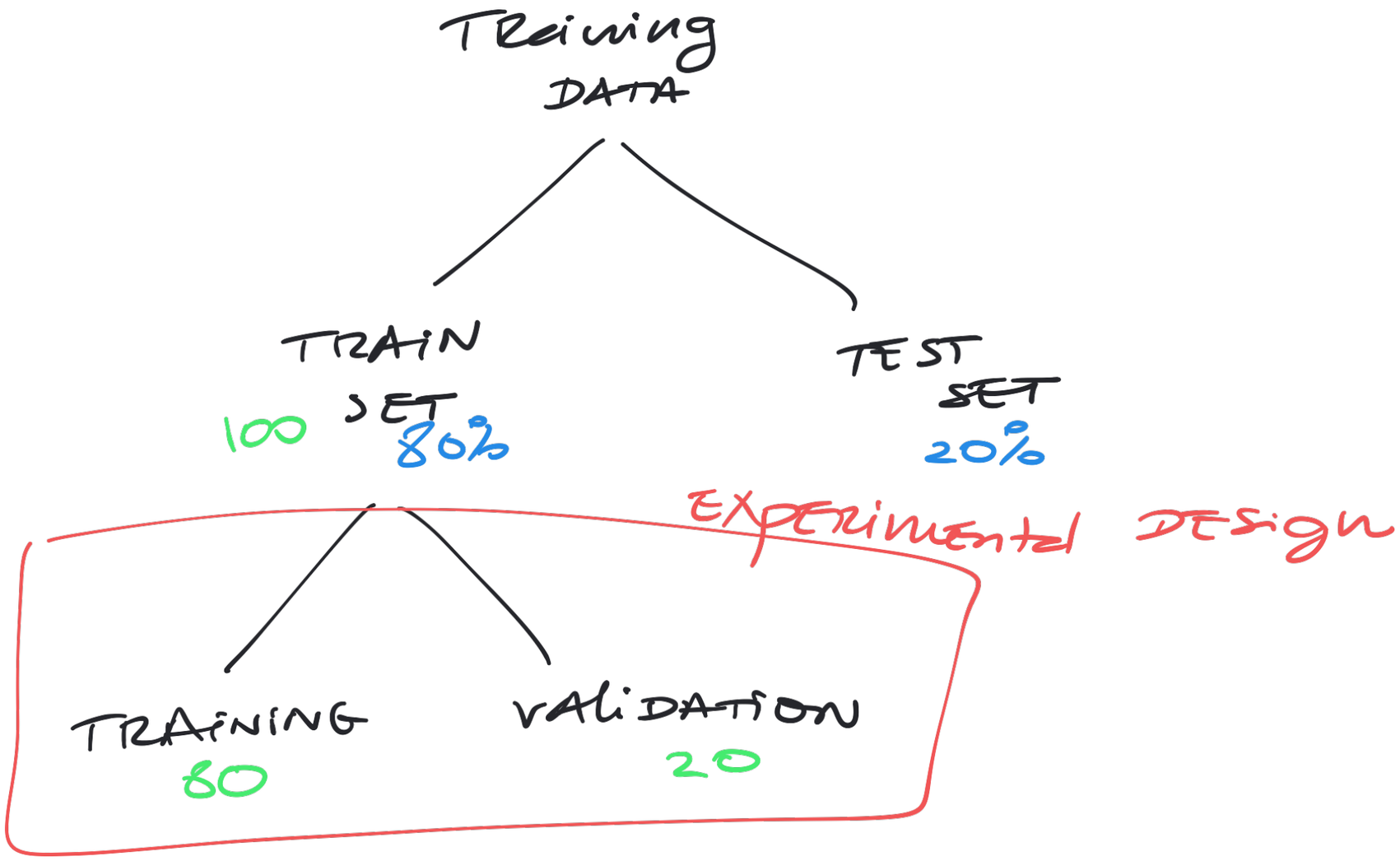
③ Model prediction:

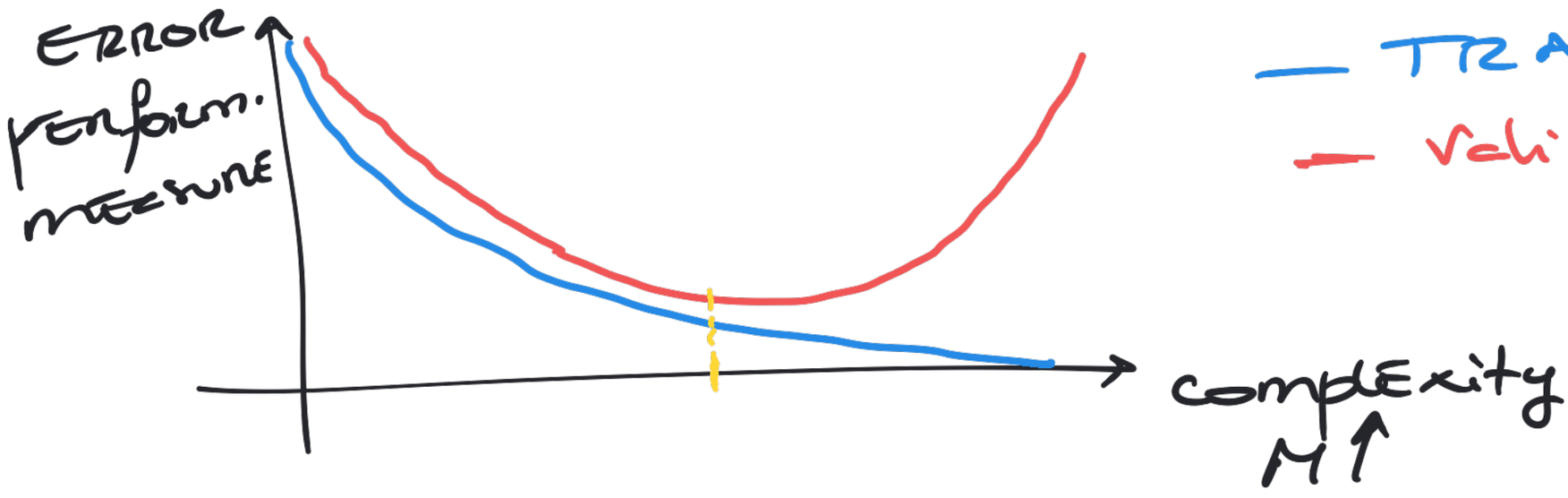
$$y = \Phi \cdot w$$

RETURN w, y

Ways to avoid overfitting

1. Apply Occam's Razor principle
2. Add more data
3. Use Regularization term
4. Cross - Validation





— TRAINING (80)
— Validation (20)