Artificial Intelligence
CE-417, Group 2
Computer Eng. Department
Sharif University of Technology

Falls 2020

By Mohammad Hossein Rohban, Ph.D.

Courtesy: Most slides are adopted from CSE-573 (Washington U.), original slides for the textbook, and CS-188 (UC. Berkeley).

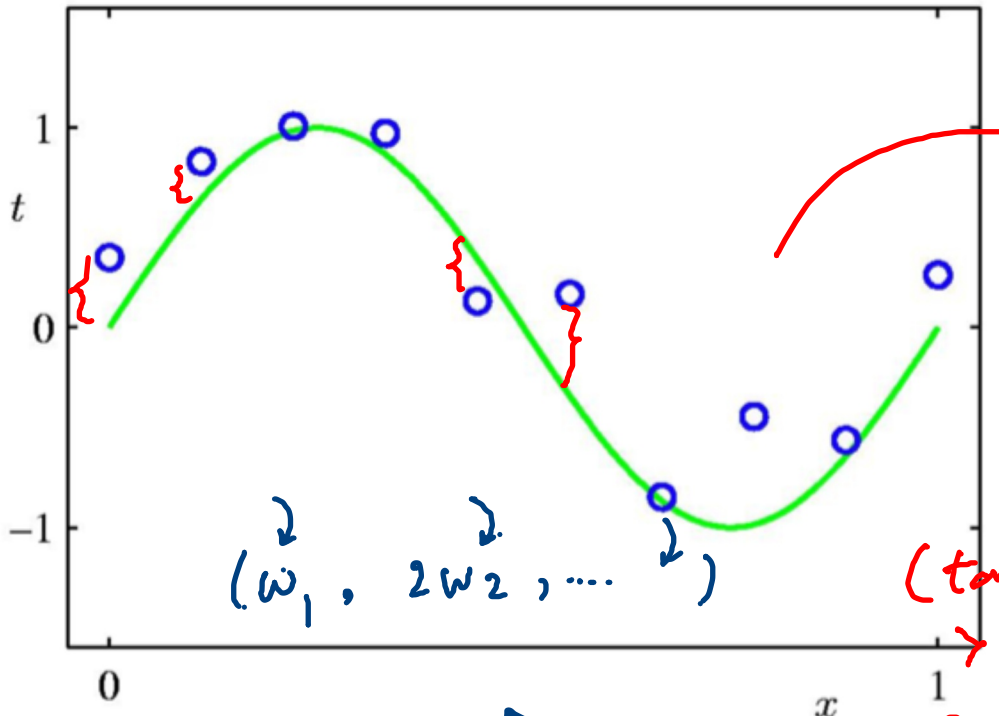
Regression

real-valued
(Output)

VS.

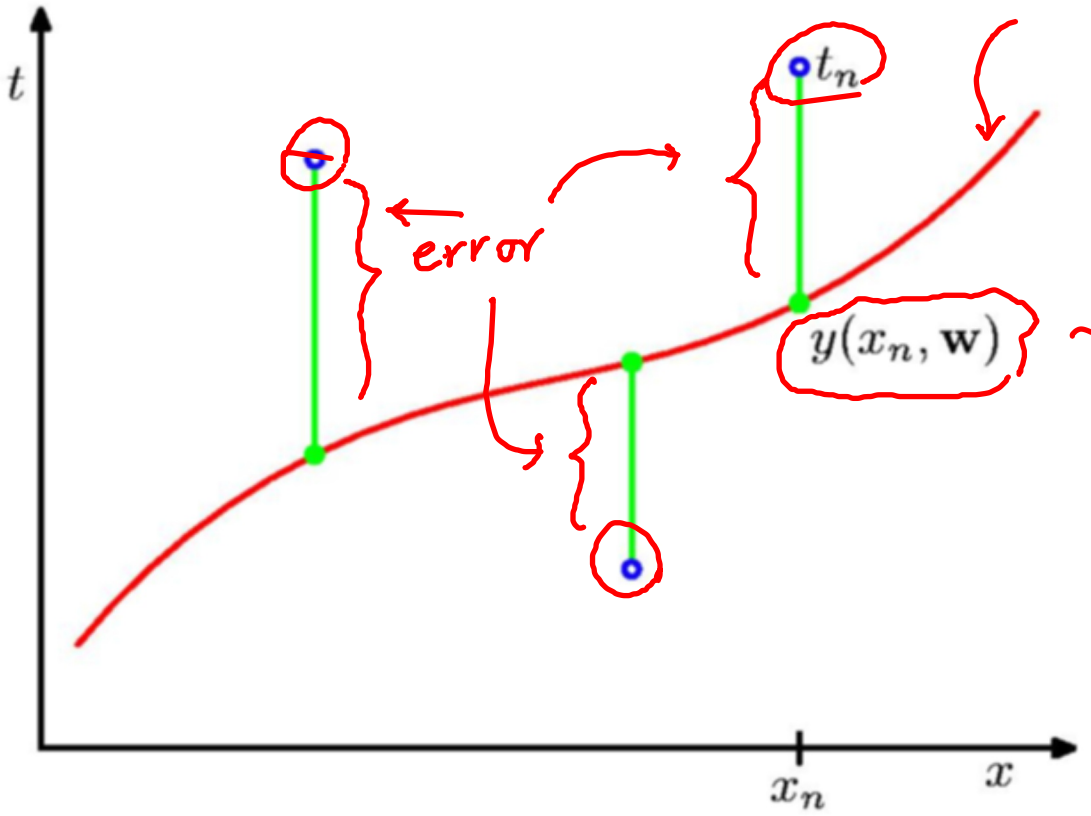
Classification
(K- Classes)

Polynomial Curve Fitting



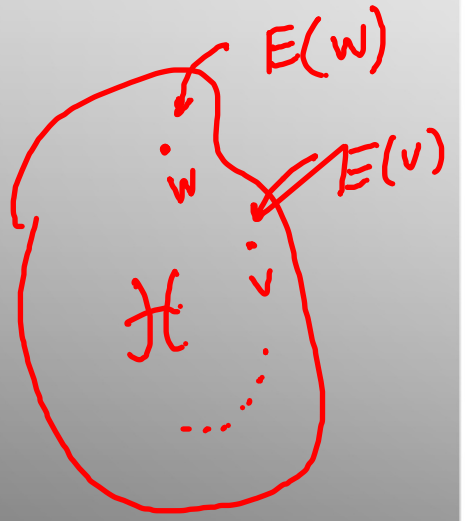
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function



→ the func. to be ranked

$\min_{\mathbf{w} \in \mathbb{R}^M} E(\mathbf{w})$



→ MSE

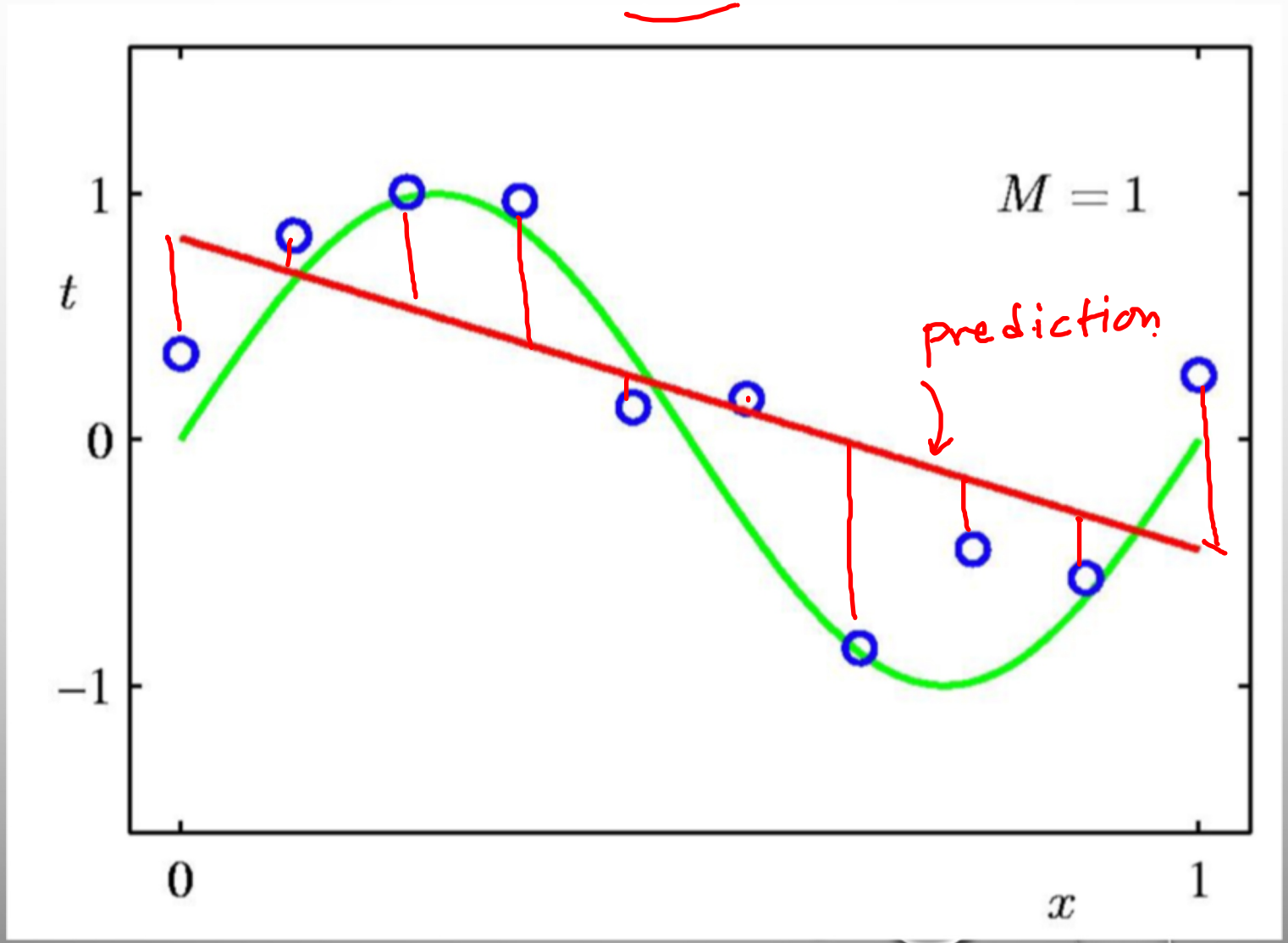
$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

error @ the n-th point

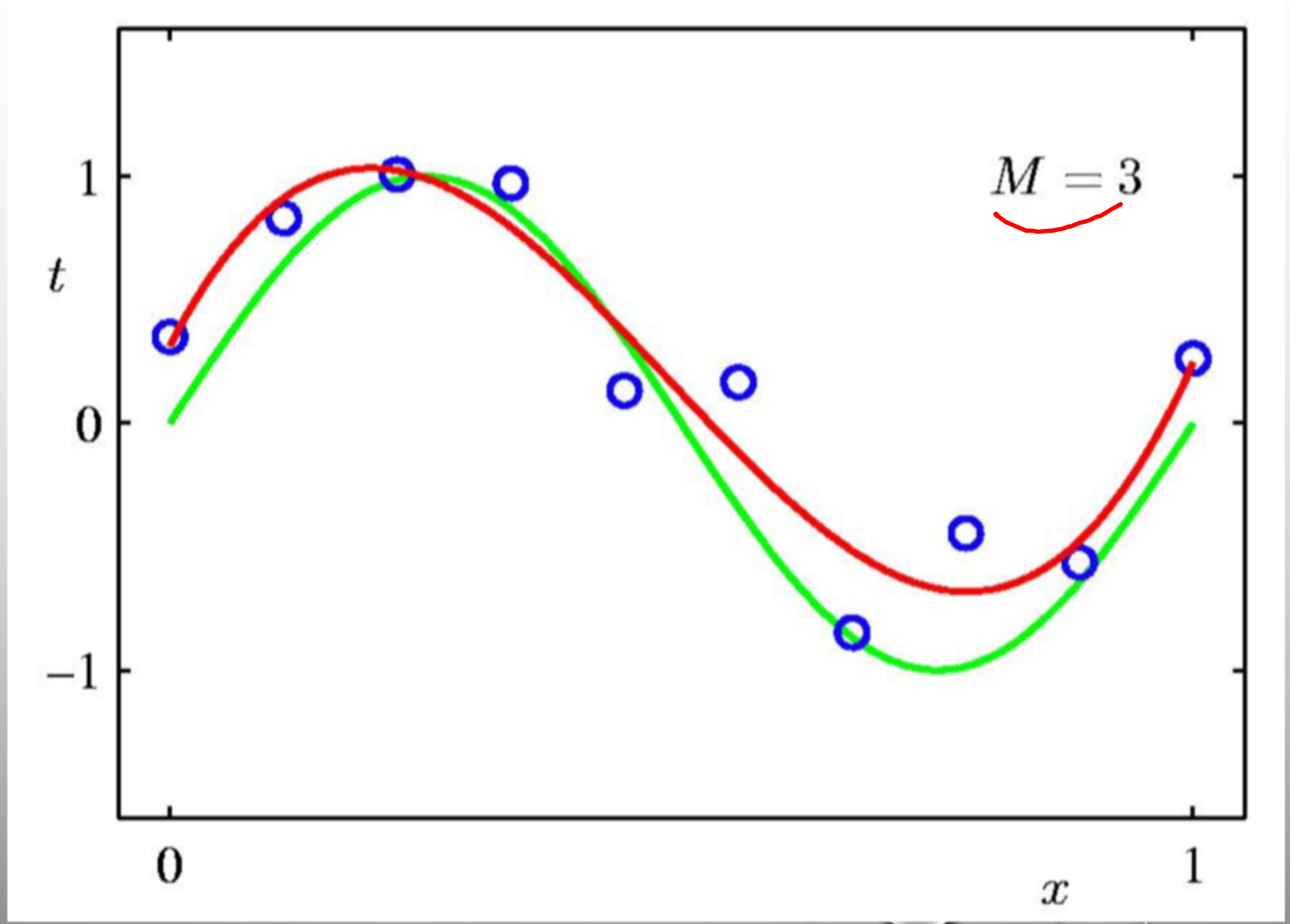
1st Order Polynomial

M = 1

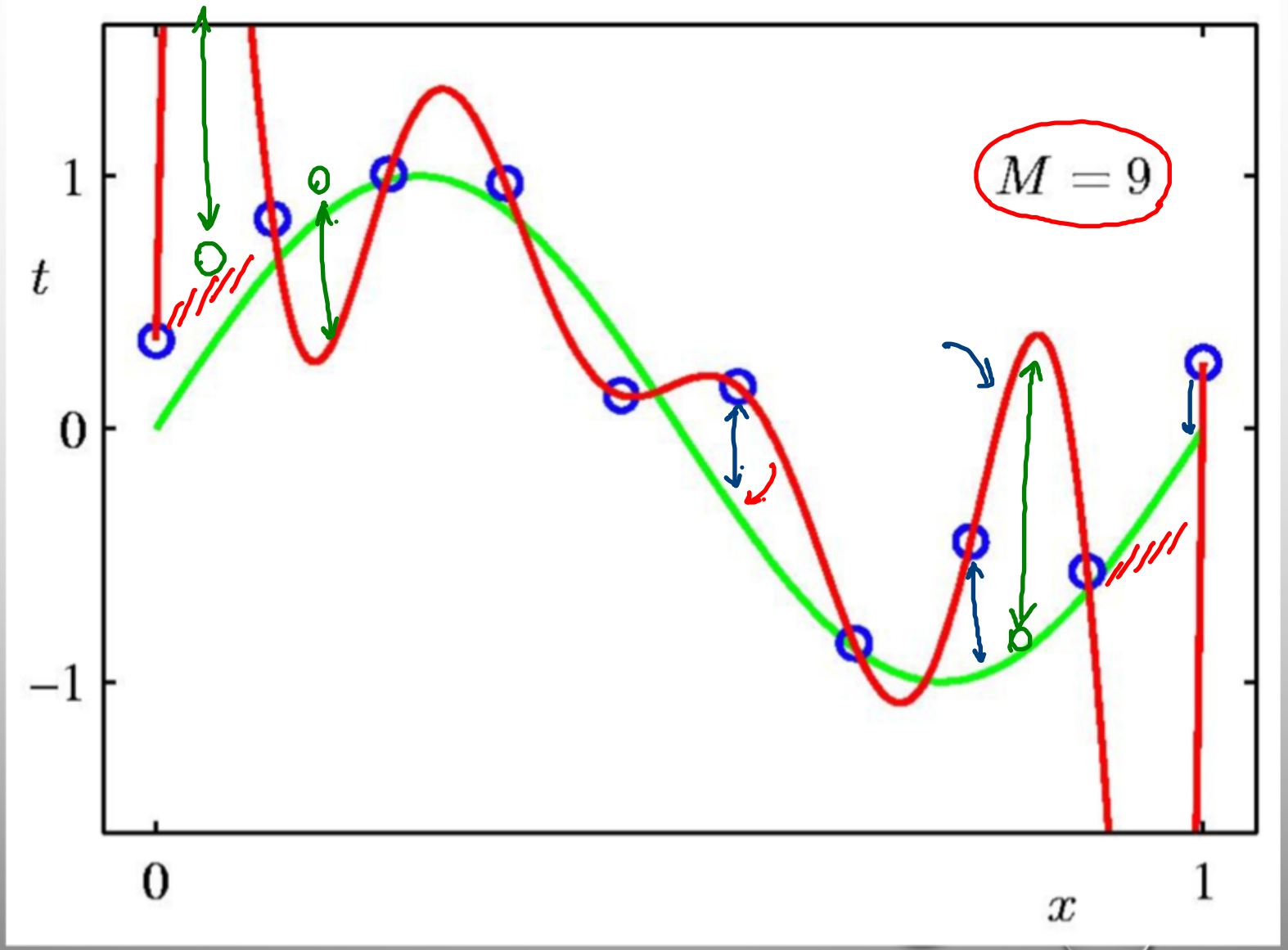
Gradient Descent



3rd Order Polynomial



9th Order Polynomial



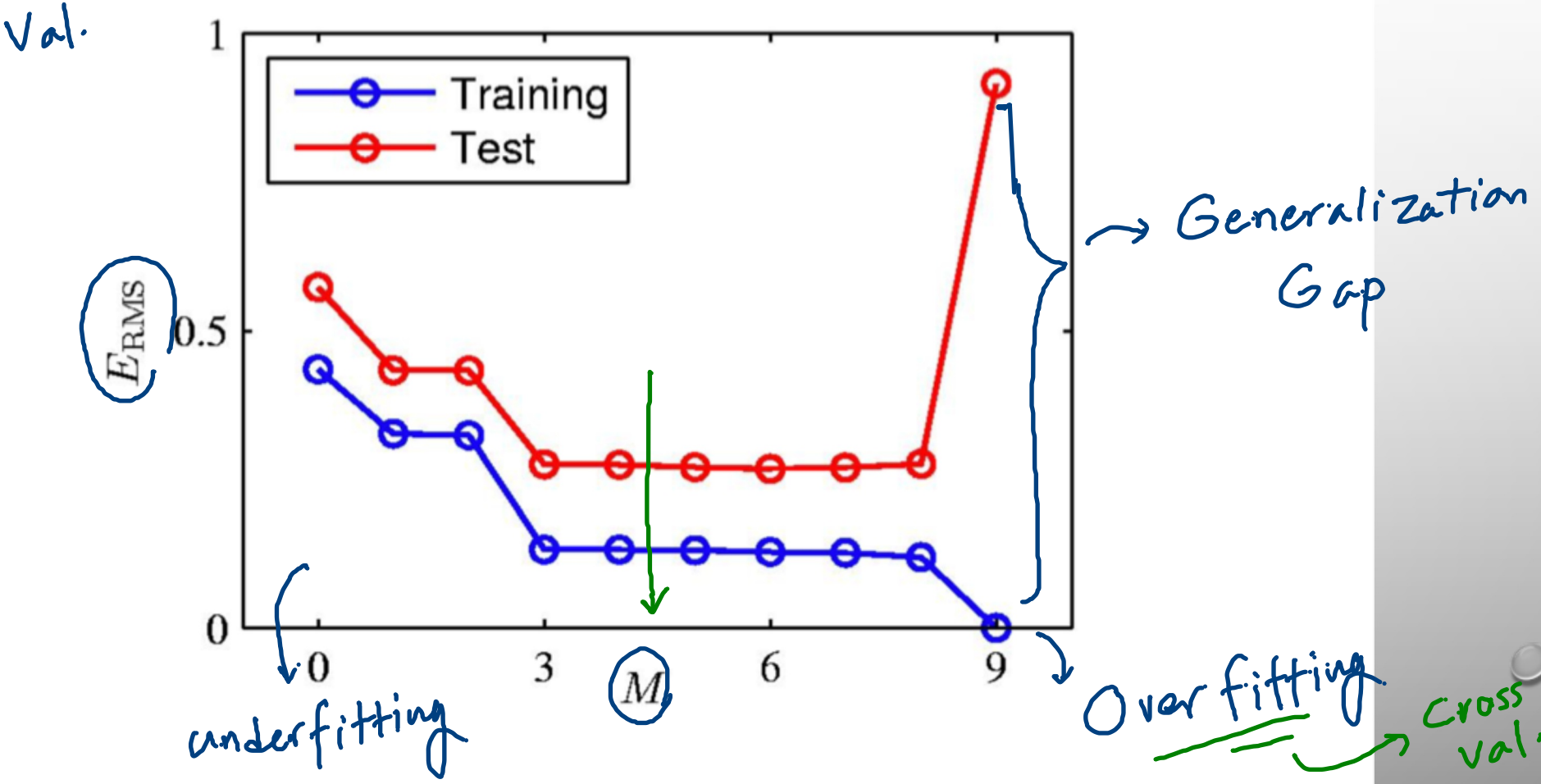
Training error = 0

Test error >> 0

→ Overfitting

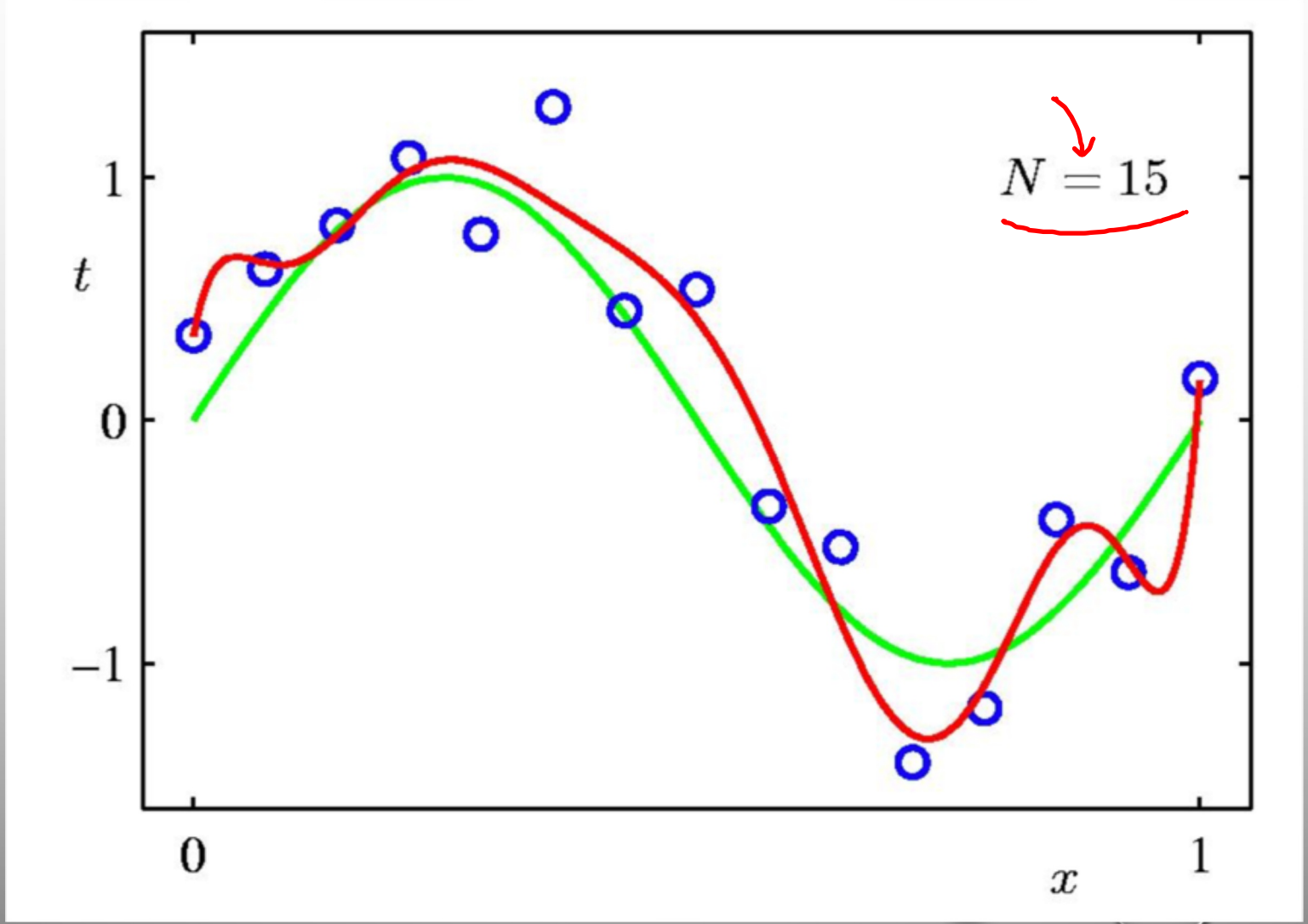
Over-fitting

P Chance Max



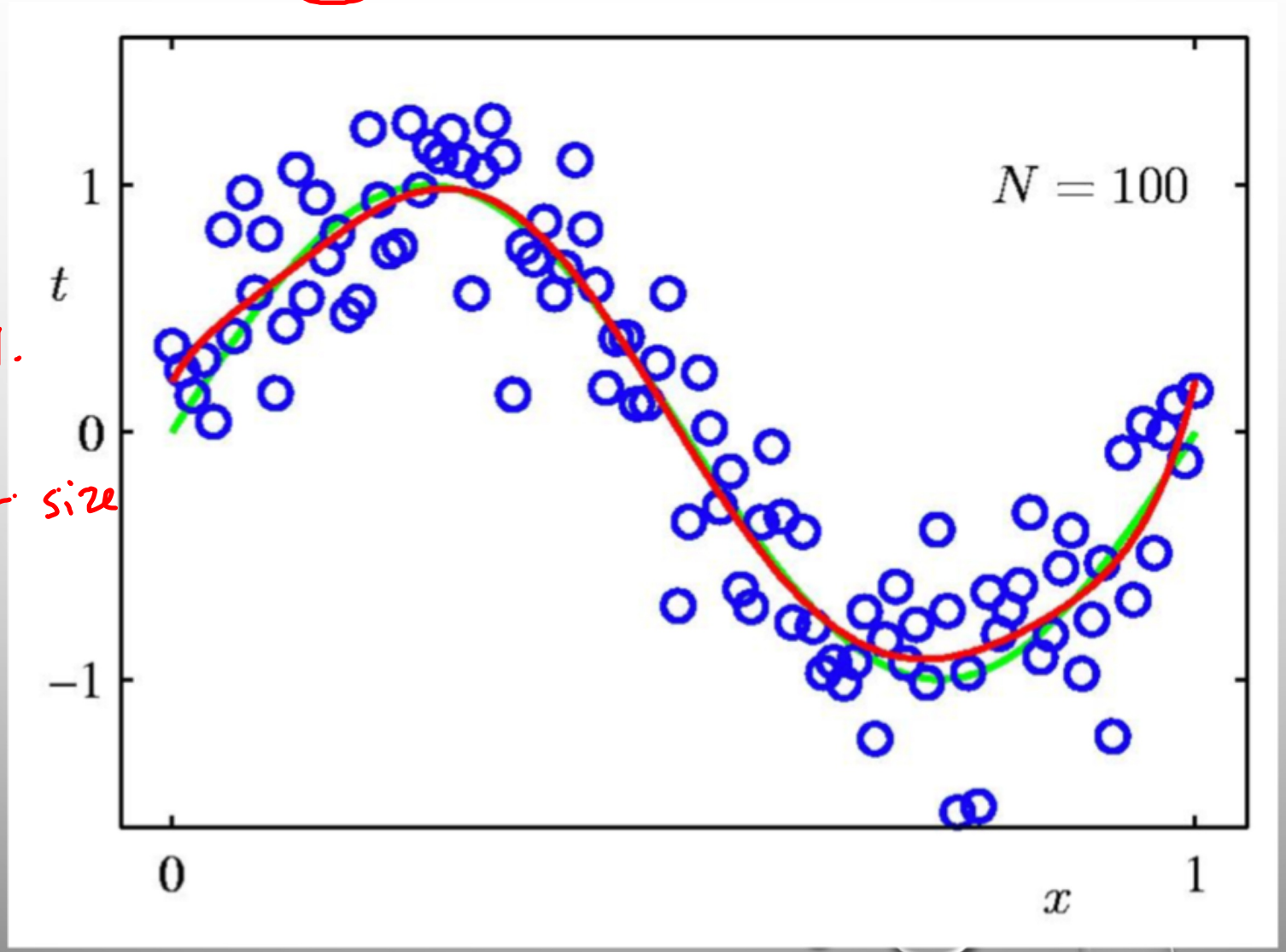
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Data Set Size:
9th Order Polynomial



Data Set Size:

9th Order Polynomial ⇒ Overfit is not only dependent on the Complexity of our hypothesis but also depends on trn. set size.



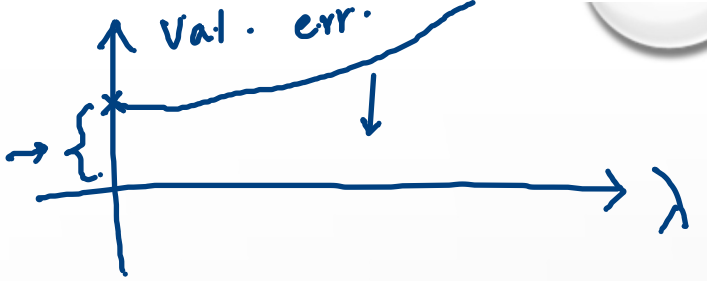
① Cross-val.

② Increase trn. set size

Polynomial Coefficients

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

③ Regularization



$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$\lambda > 0$

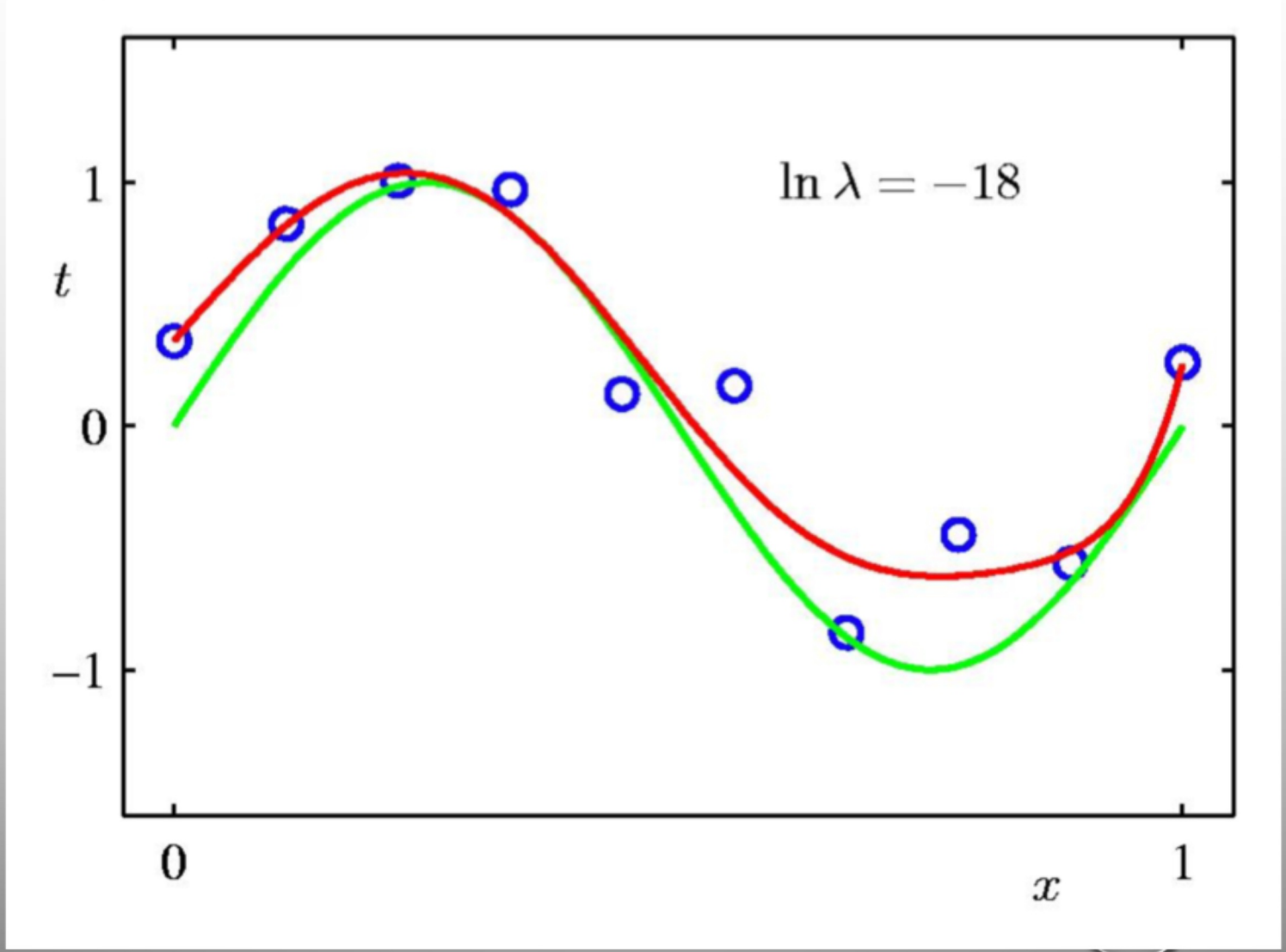
Empirical Err.

$\int_{\mathcal{X}} \|\nabla_x^k y\|_2^2 \, dx$

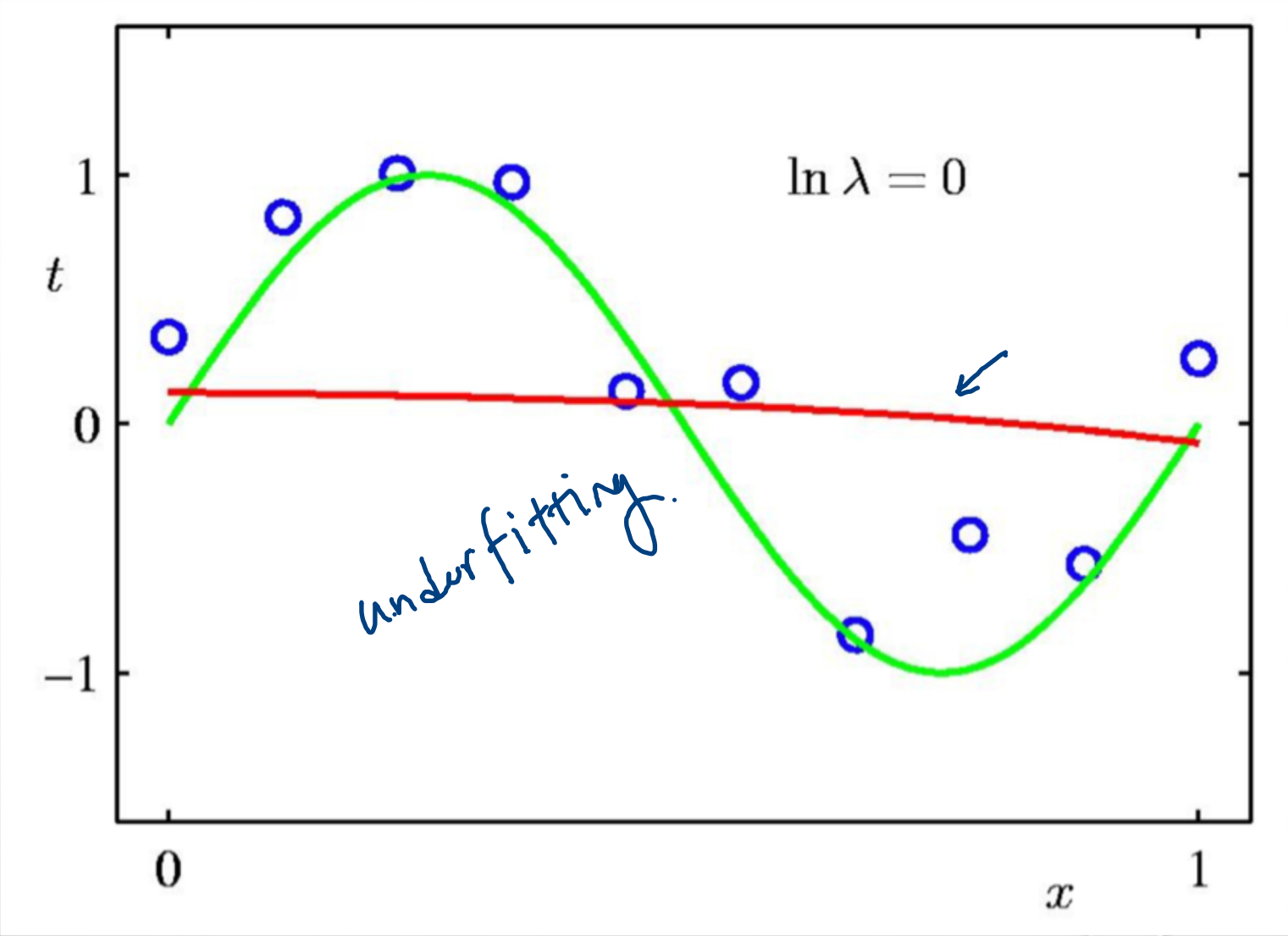
- Penalize large coefficient values

enforcing Smoothness ⟶ Derivative of any order
↳ the poly. Coeffs.

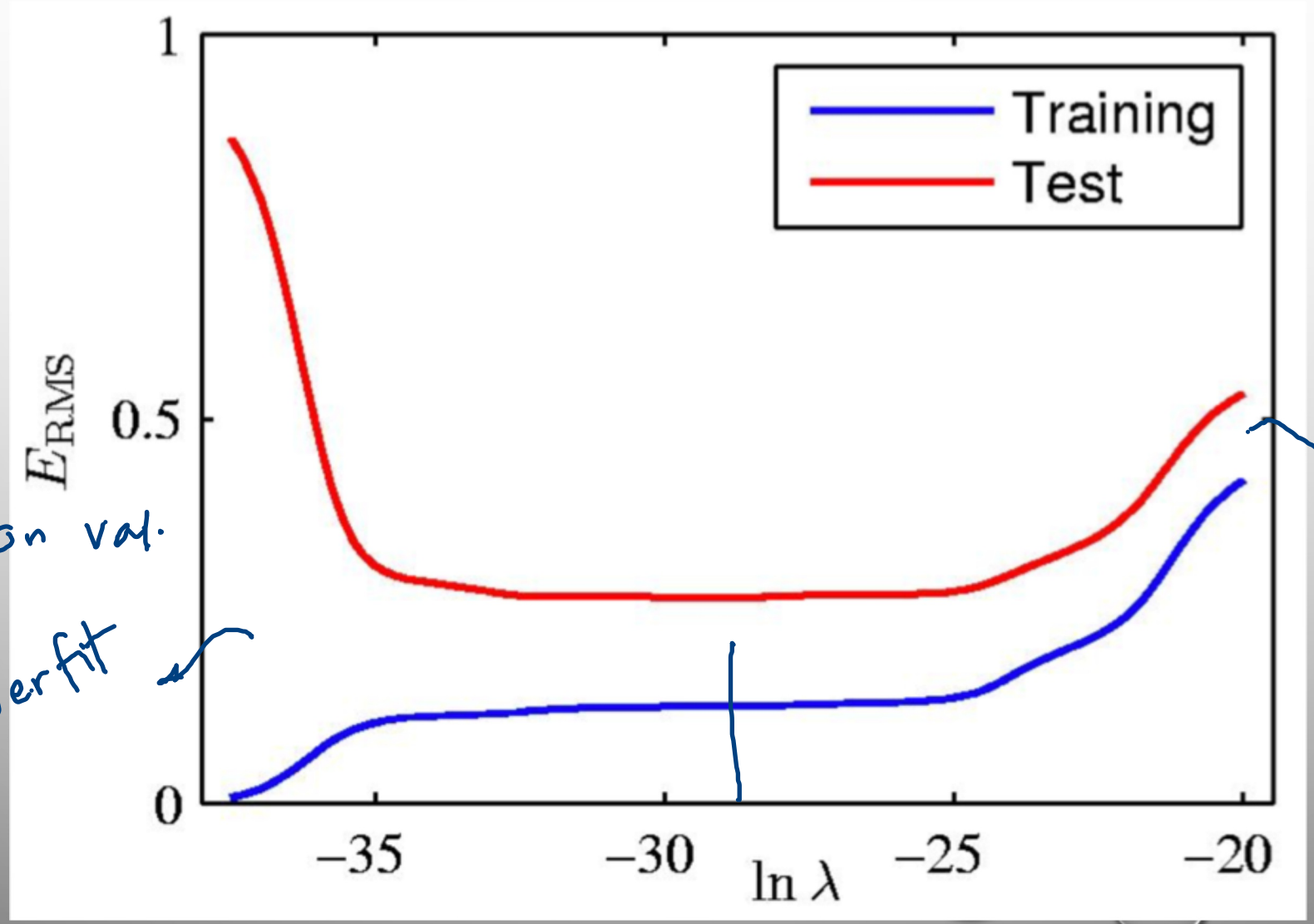
Regularization:
ln λ = -18



Regularization: $\ln \lambda = 0$



Regularization : E_{RMS} vs. ln λ



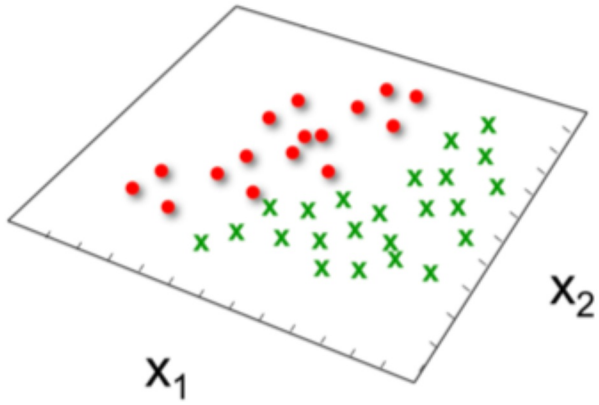
Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Logistic Regression

$$P(Y = 1|X =< X_1, ...X_n >) = \frac{1}{1 + exp(w_0 + \sum_i w_i X_i)}$$

implies



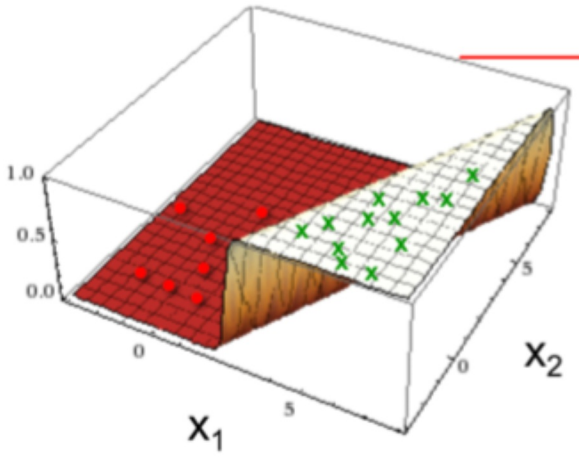
$$\ln \frac{P(Y = 0|X)}{P(Y = 1|X)} = w_0 + \sum_i w_i X_i$$

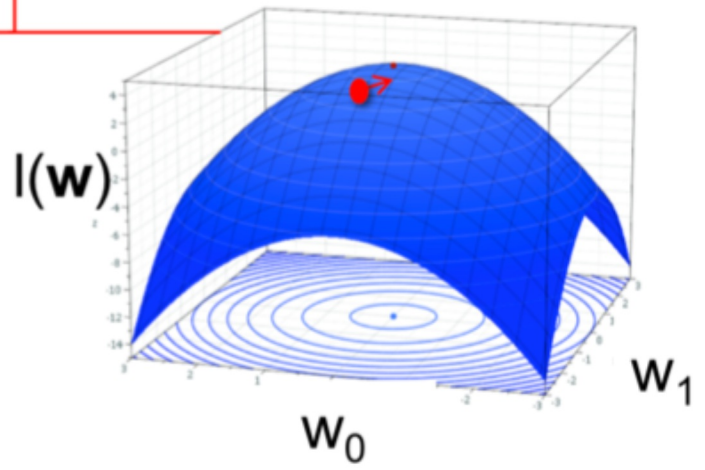
linear classification rule!

Gradient Ascent

w_0=40 w_1= -10 w_2 =5

Maximize $l(\mathbf{w}) = \ln P(D_Y \mid D_x, H_w)$





Update rule:

$$\Delta \mathbf{w} = \eta \nabla_{\mathbf{w}} l(\mathbf{w})$$

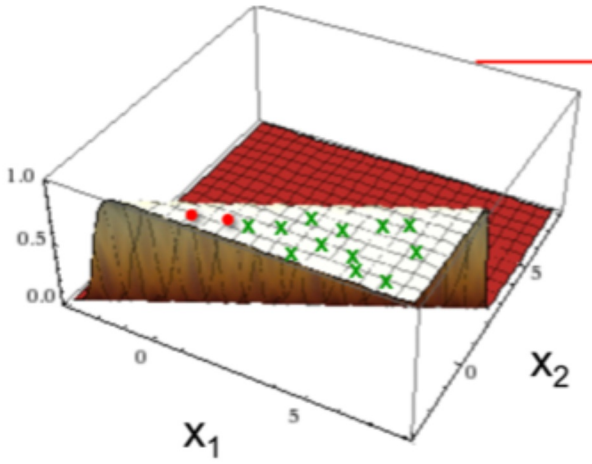
$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \frac{\partial l(\mathbf{w})}{\partial w_i}$$

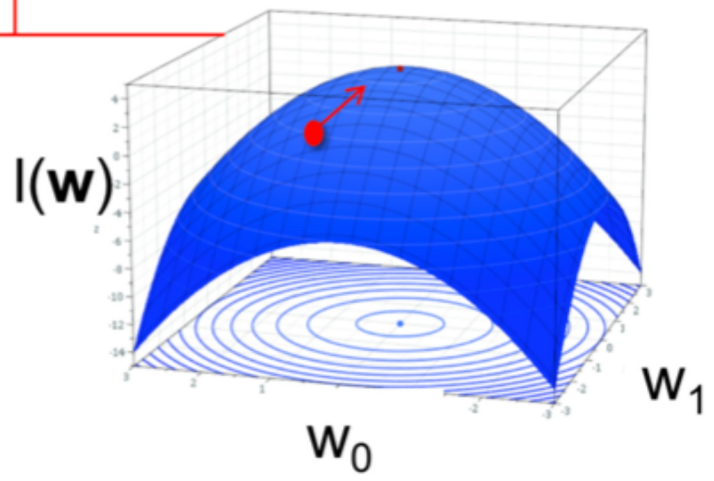
Logistic w/ Initial Weights

w_0=20 w_1= -5 w_2 =10

Loss(H_w) = Error(H_w, data)

Minimize Error → Maximize $l(\mathbf{w}) = \ln P(D_Y \mid D_X, H_w)$





Update rule:

$$\Delta \mathbf{w} = \eta \nabla_{\mathbf{w}} l(\mathbf{w})$$

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \frac{\partial l(\mathbf{w})}{\partial w_i}$$

Step size

x | y

Weight, Ch., Ins. ... | Hearth Attack → Trn. set $(\underline{x}_1, 1), (\underline{x}_2, -1) \cdots (\underline{x}_N, 1)$

age, Gender ...

↳ You can only predict its risk ⇝ real not the actual outcome

$\underline{x} \in \mathbb{R}^d, \quad \underline{w} \in \mathbb{R}^d$

$$\mathbb{P}(Y=1 \mid \underline{x}, \underline{w}) = \frac{1}{1 + e^{\langle \underline{w}, \underline{x} \rangle}} < 1$$

$e^{\langle \underline{w}, \underline{x} \rangle} > 0$

Outcome: 0/1

(prob. of the outcome)

$$\mathbb{P}(Y=-1 \mid \underline{x}, \underline{w}) = 1 - \frac{1}{1 + e^{\langle \underline{w}, \underline{x} \rangle}}$$

$$= \frac{1 + e^{\langle \underline{w}, \underline{x} \rangle} - 1}{1 + e^{\langle \underline{w}, \underline{x} \rangle}} = \frac{1}{1 + e^{-\langle \underline{w}, \underline{x} \rangle}}$$

$$\mathbb{P}(Y=y \mid \underline{x}, \underline{w}) = 1/(1 + e^{y \langle \underline{w}, \underline{x} \rangle})$$

* What fitness func. to use in this scenario ?!

Max. Likelihood.

$D = \{ (\underline{x}_i, y_i) \mid 1 \le i \le N \} \rightsquigarrow$ i.i.d

$\mathbb{P}(Y = y \mid \underline{x}, \underline{w})$

→ which $\underline{w}$ would make the prob. model to assign the highest prob. to D ?

$$\max_{\underline{w}} \mathbb{P}(D \mid \underline{w})$$

$$\mathbb{P}(D \mid \underline{w}) \underset{(i.i.d.)}{=} \prod_{i=1}^{N} \mathbb{P}((\underline{x}_i, y_i) \mid \underline{w})$$

$$\mathbb{P}((\underline{x}_i, y_i) \mid w) = \mathbb{P}(Y = y_i \mid \underline{x}_i, \underline{w}) \; \mathbb{P}(\underline{x}_i \mid \underline{w})$$

$$\equiv \min_{\underline{w}} -\sum_{i=1}^{N} \log \mathbb{P}(Y = y_i \mid \underline{x}_i, \underline{w})$$

$$\mathbb{P}(Y = y_i \mid \underline{x}_i, \underline{w}) \rightarrow \frac{1}{1 + e^{y_i \langle \underline{w}, \underline{x}_i \rangle}}$$

$$\equiv \min_{\underline{w}} \underbrace{\sum_{i=1}^{N} \log\left(1 + e^{\underbrace{y_i \langle \underline{w}, \underline{x}_i \rangle}_{A}}\right)}_{L} \rightsquigarrow \text{GD}$$

$\rightsquigarrow$ Convex.

$$\nabla_{\underline{w}} L = \sum_{i=1}^{N} \frac{y_i \underline{x}_i A}{1 + A} = \sum_{i=1}^{N} y_i \underline{x}_i \underbrace{\left(\frac{1}{1 + A^{-1}}\right)}_{\mathbb{P}(Y \neq y_i \mid \underline{x}_i, \underline{w})}$$

$$\underline{w}^{(t+1)} \leftarrow \underline{w}^{(t)} - \eta \sum_{i=1}^{N} y_i \underline{x}_i \underbrace{\mathbb{P}(Y \neq y_i \mid \underline{x}_i, \underline{w})}_{\text{if 1, the model did poorly on } \underline{x}_i, \text{ so rotate } \underline{w} \text{ in the } \underline{x}_i \text{ dir.}}$$

if 0, the model has done well for $\underline{x}_i$

$\Rightarrow$ $\underline{x}_i$ will not contribute to $\Delta \underline{w}$

$$\nabla_w^2 L = \nabla_w \left(\sum_{i=1}^{N} \frac{(y_i x_i)}{1 + A^{-1}} \right)$$

$$= \sum_{i=1}^{N} \frac{1/A^2 \cdot y_i^2 \, x_i x_i^T}{\underbrace{(1 + A^{-1})^2}_{B}} = \sum_{i=1}^{N} \underbrace{\left(\frac{C}{B}\right)}_{\alpha_i > 0} x_i x_i^T$$

$$= \sum_{i=1}^{N} \alpha_i \underbrace{x_i x_i^T}_{\text{rank-1}} \succeq 0$$

$\succeq 0$

$\Rightarrow$ L is convex.

$\Rightarrow$ GD on L converges to the global minimum if η is suff. small.