AAI 695: Applied Machine Learning

Lecture 1-2 : Polynomial Fitting

Dr. Shucheng Yu, Associate Professor
Department of Electrical and Computer Engineering
Stevens Institute of Technology

Elements for the learning problems

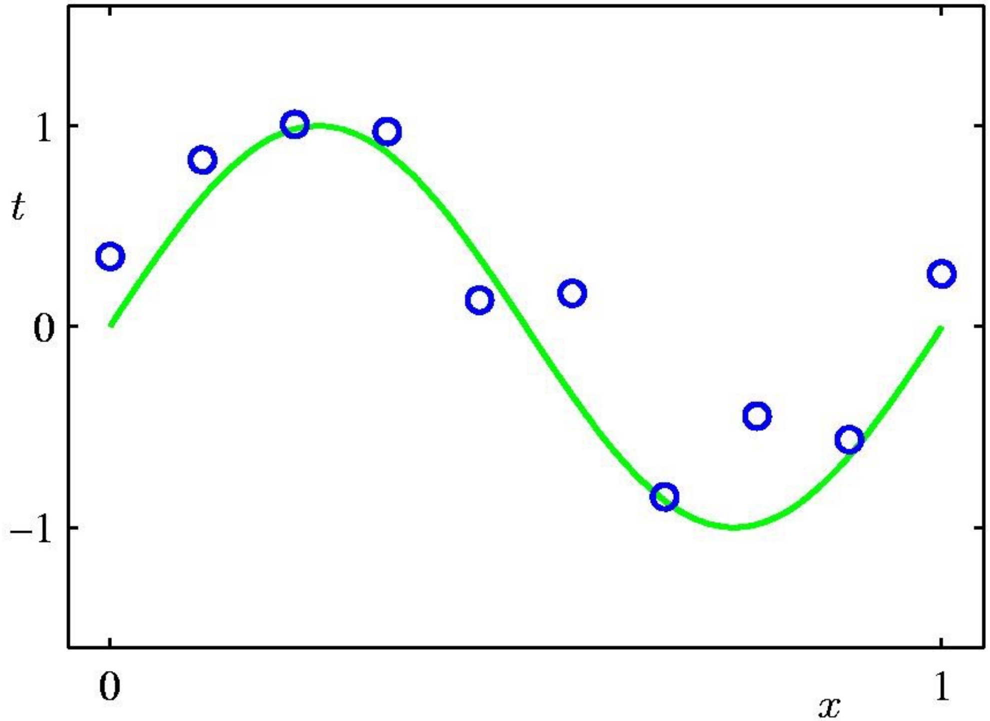
Learning = Improving with experience at some task

- Improve over task T,
- with respect to performance measure P,
- based on experience E.

E.g., Learn to play checkers

- T: Play checkers
- P: % of games won in world tournament
- E: opportunity to play against self

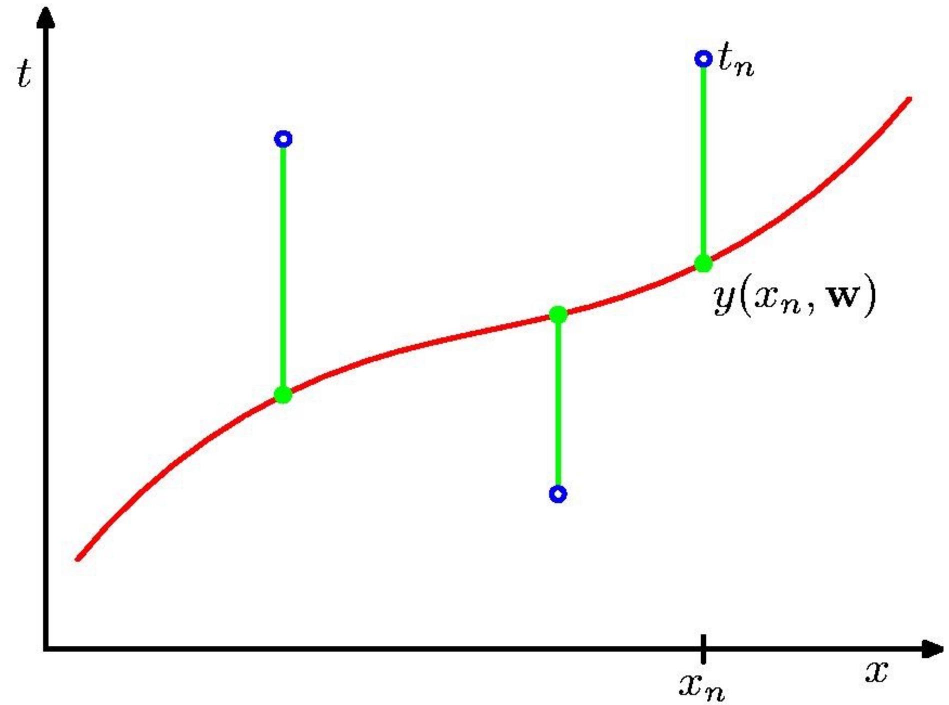
Polynomial Curve Fitting



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

How to measure the fit between model and training data?

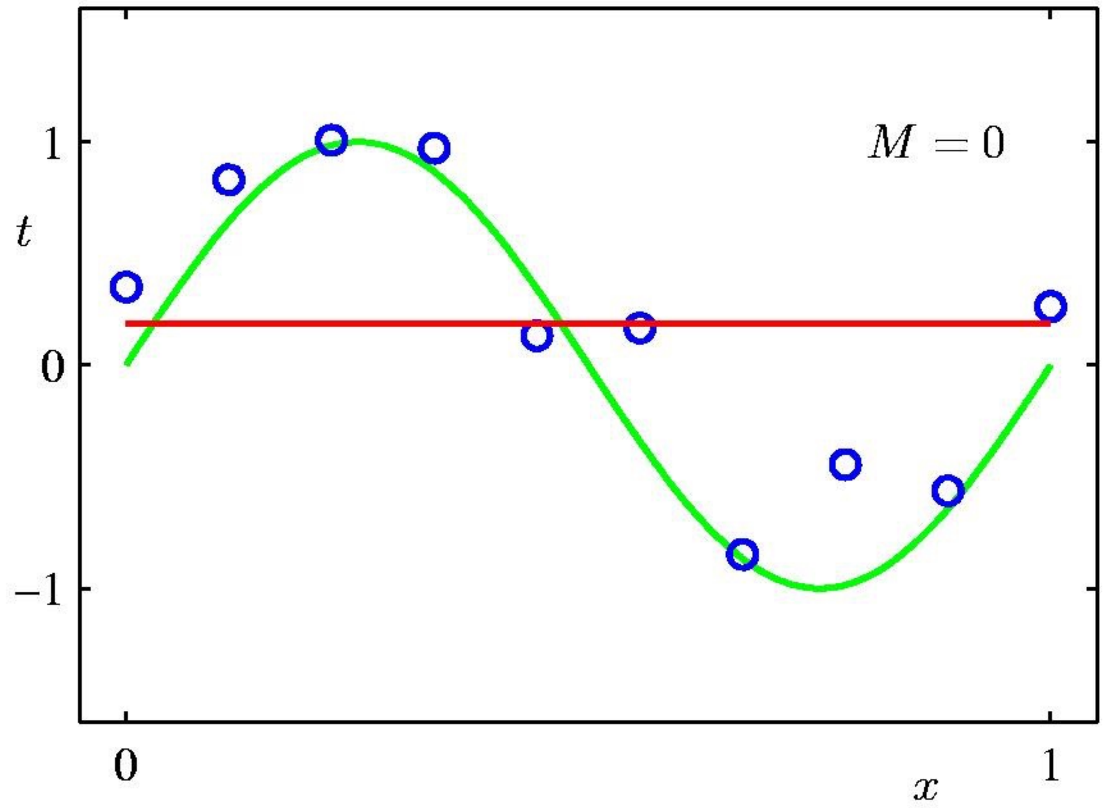
Sum-of-Squares Error Function



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

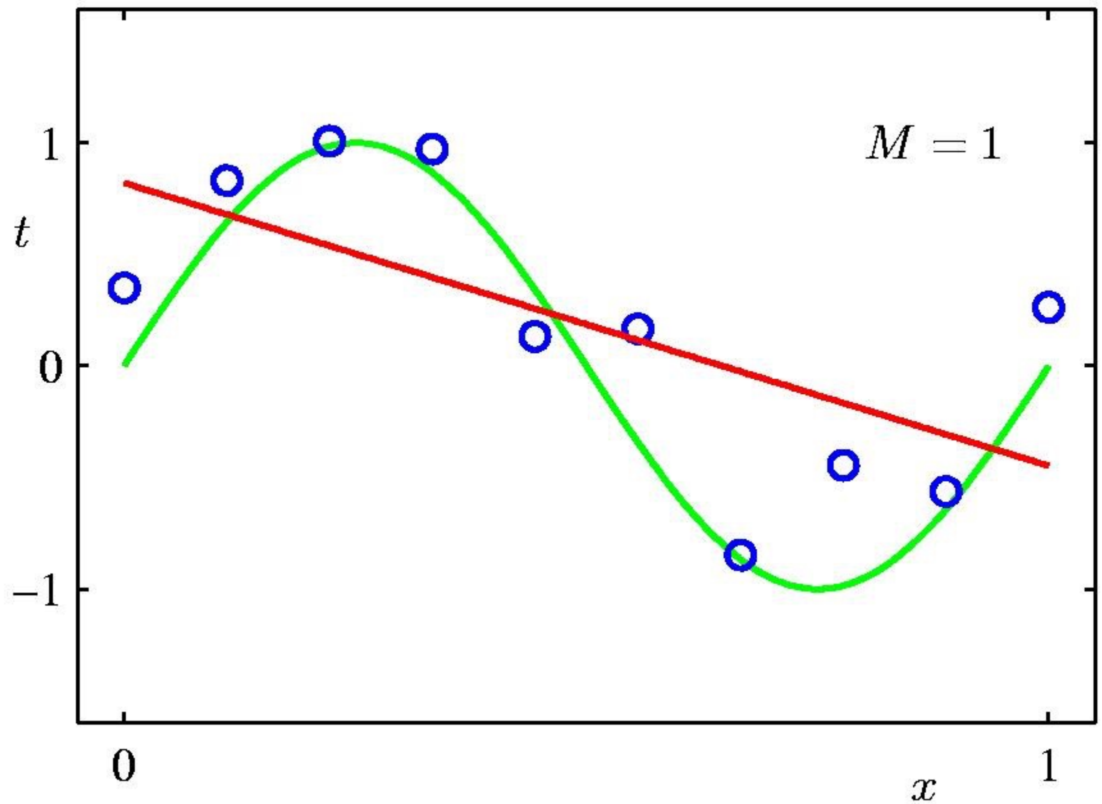
Minimize E(w) for unknown w. (maximum likelihood)

0^{th} Order Polynomial

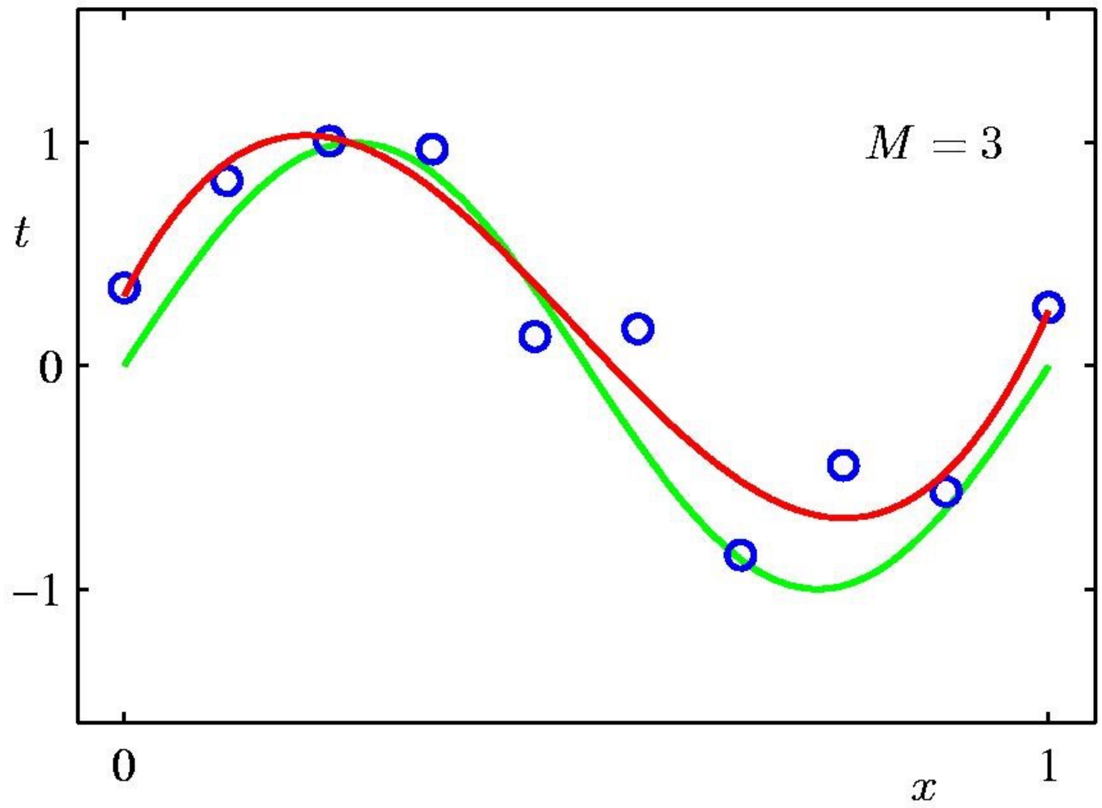


Model selection: how to choose the order M?

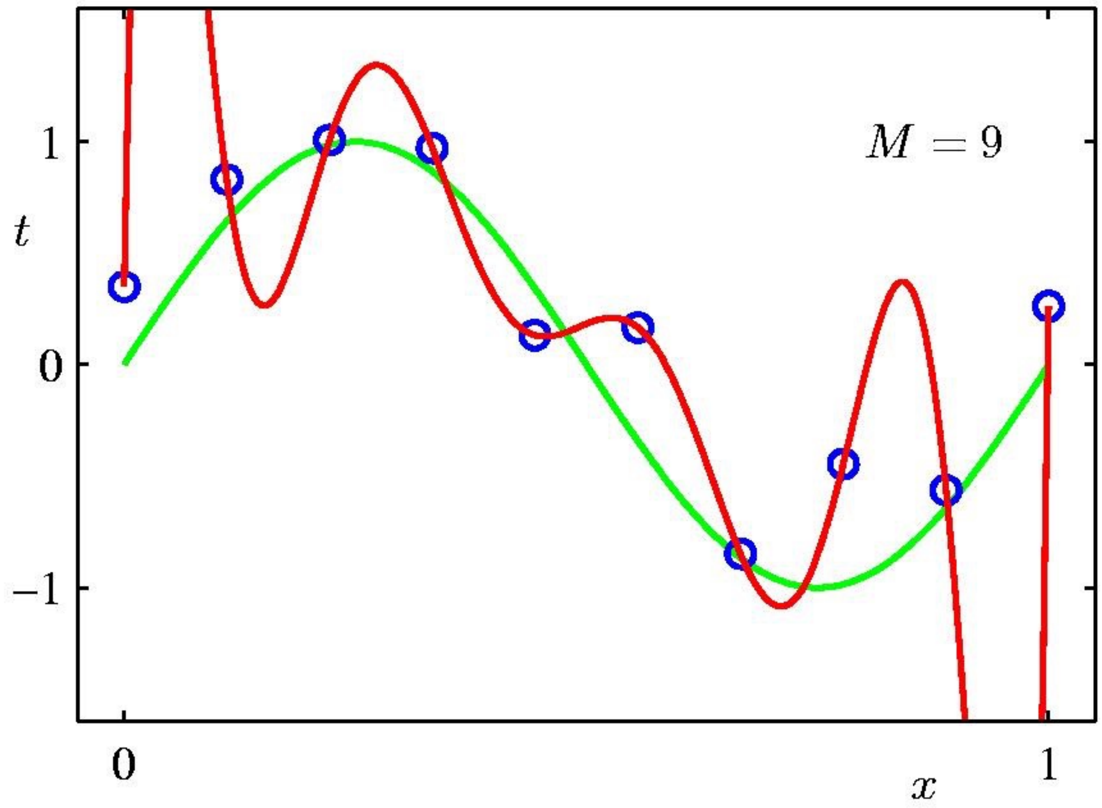
1^{st} Order Polynomial



3rd Order Polynomial

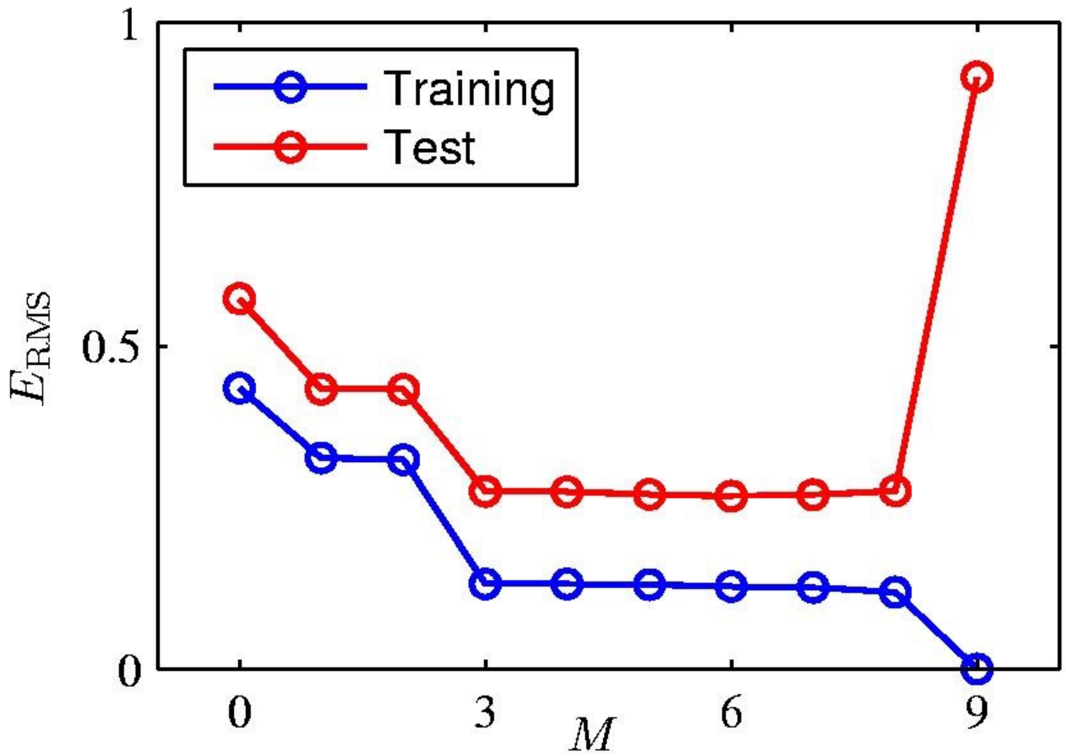


9th Order Polynomial



M=9: perfectly fit for training data set. Question: the larger M the better?

Over-fitting



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

M=9: good for training data, not for test data. What is under-fitting?

Bias-Variance Problem

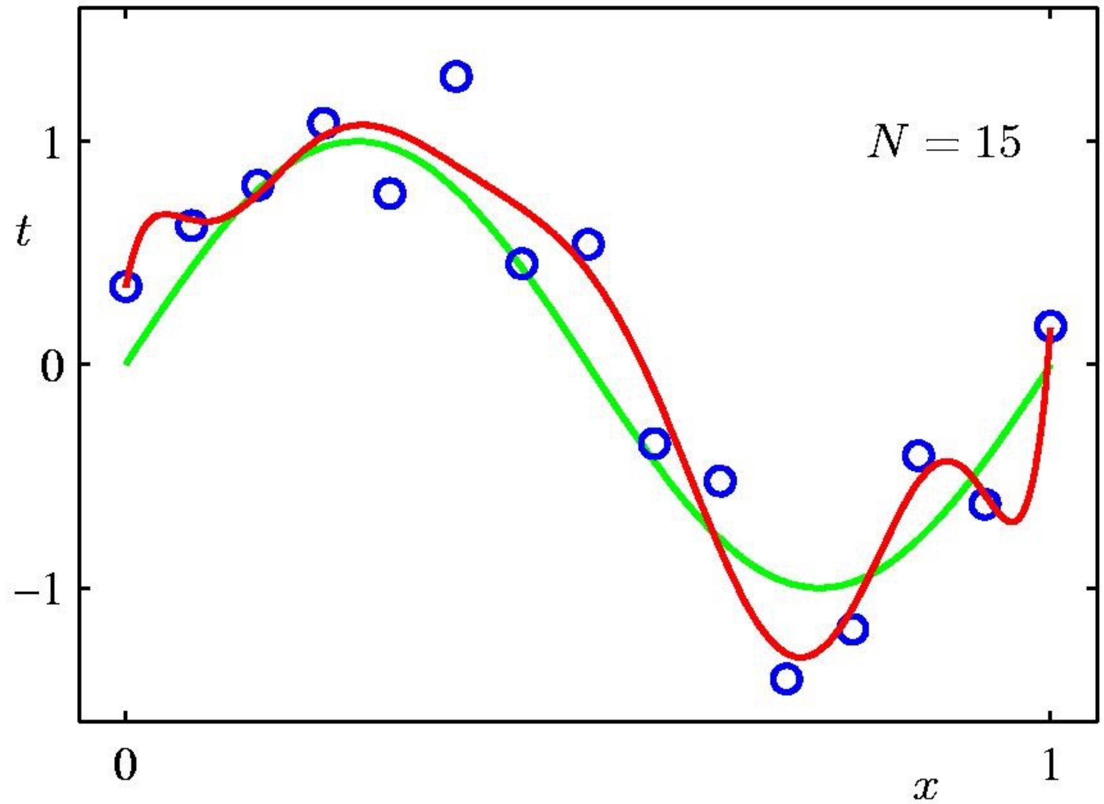
Polynomial Coefficients

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	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

coefficients increases as M getting larger (larger oscillations).

Data Set Size: $N = 15$

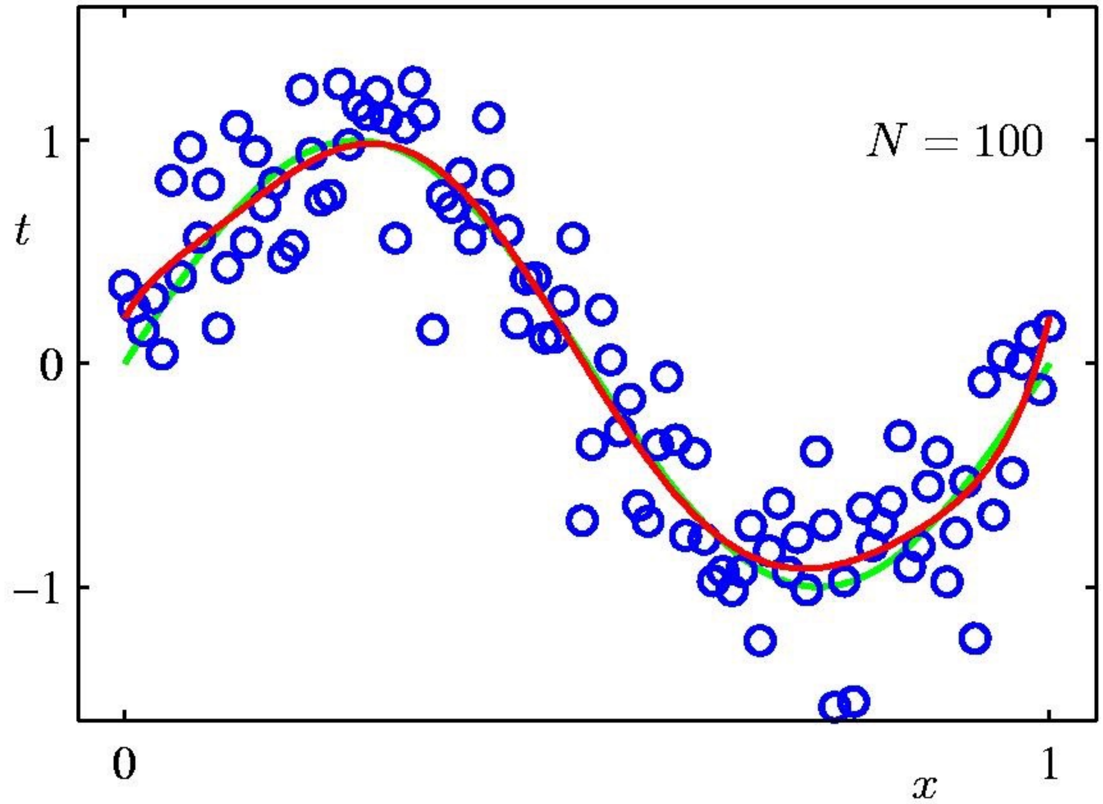
9th Order Polynomial



Overfitting less severe as data set size increases

Data Set Size: $N = 100$

9th Order Polynomial



The **larger** the data set, the more **complex** model we can afford

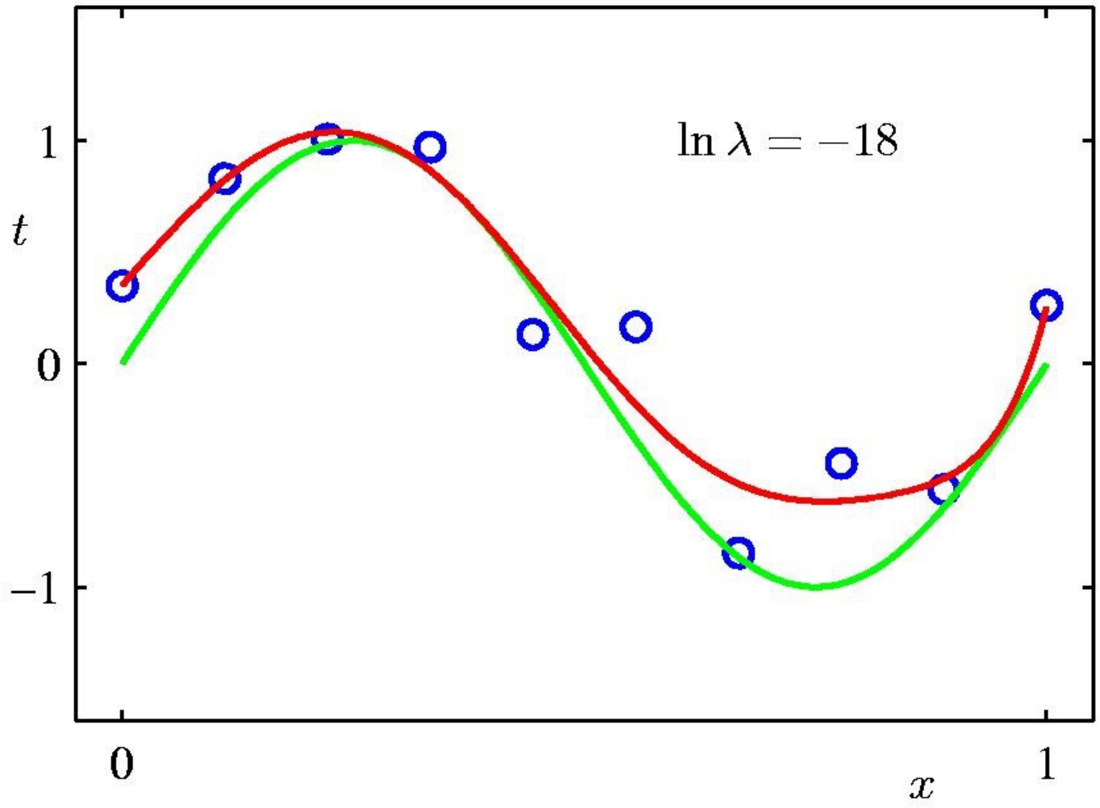
Regularization

- Penalize large coefficient values

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

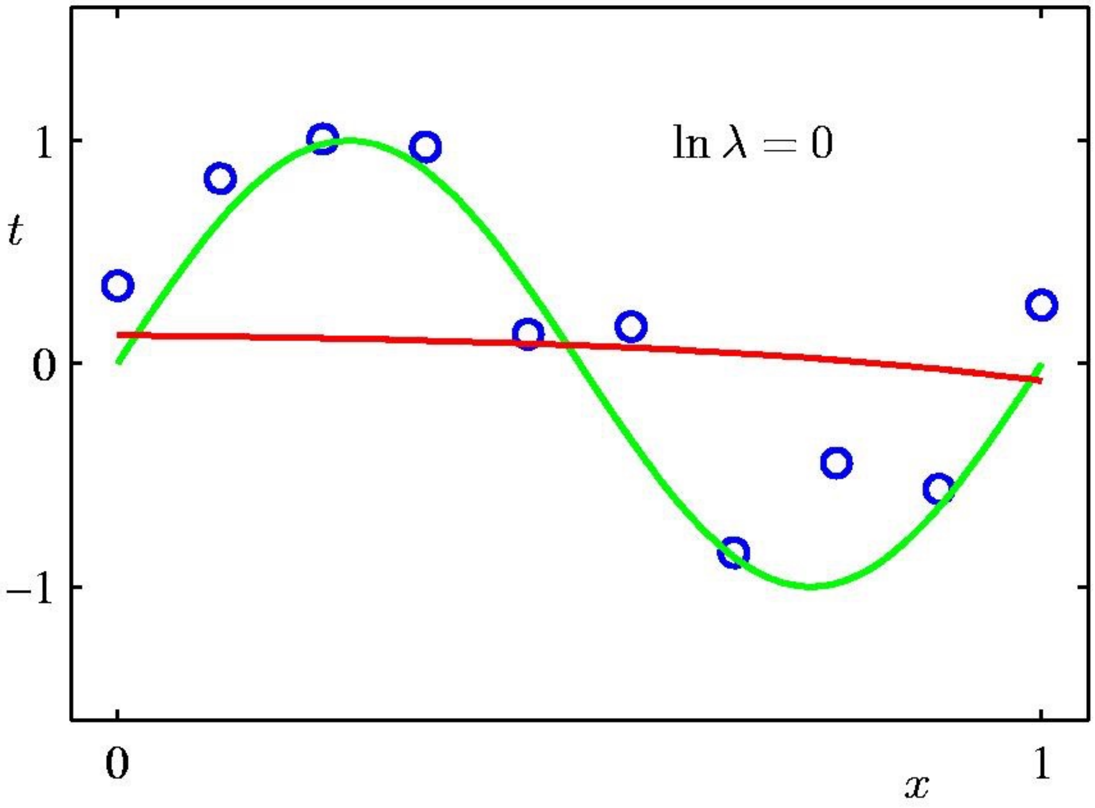
To support complex model under limited data size for maximum likelihood approach.

Regularization: $\ln \lambda = -18$



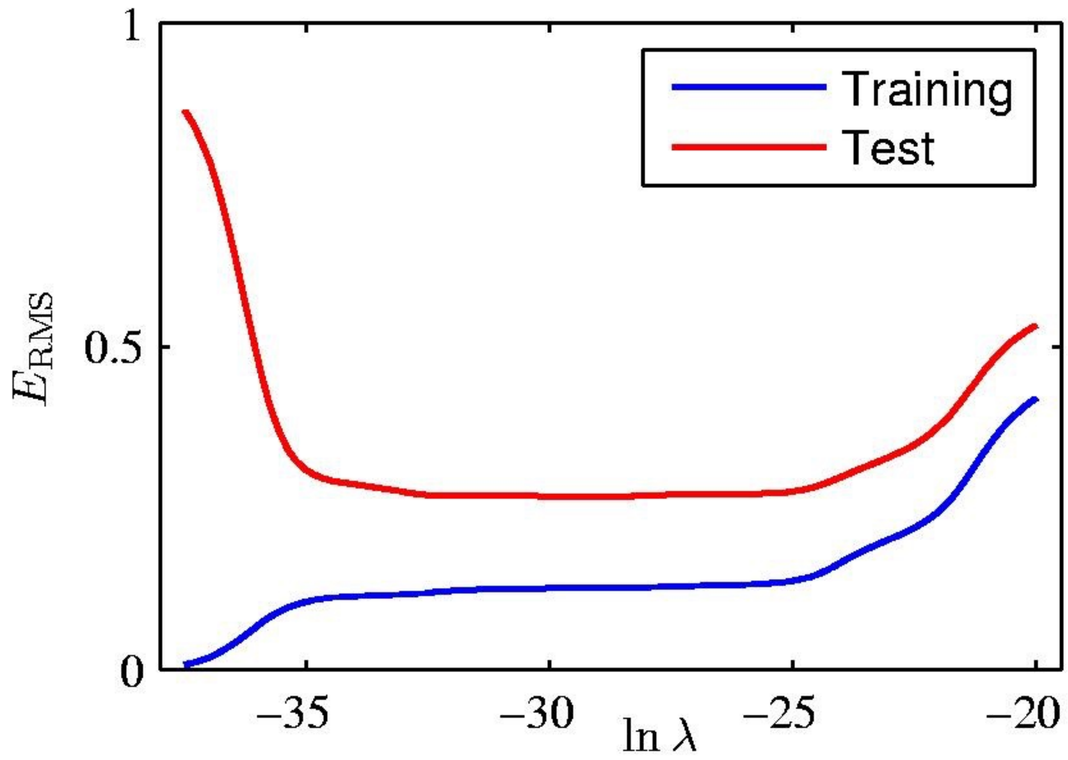
M = 9

Regularization: $\ln \lambda = 0$



M = 9

Regularization: E_{RMS} vs. $\ln \lambda$



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

Acknowledgement

The slides were largely borrowed from Dr. Christopher M. Bishop's material at https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book?from=http%3A%2F%2Fresearch.microsoft.com%2F%7Ecmbishop%2Fprml;

For details please read Chapter 1 of textbook "Pattern Recognition and Machine Learning", by Christopher Bishop.

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

stevens.edu