ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 5: Model Selection, CV, Regularization

Clayton T. Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

Special Thanks to Rev. Dawson

6 September 2017

Next Topics

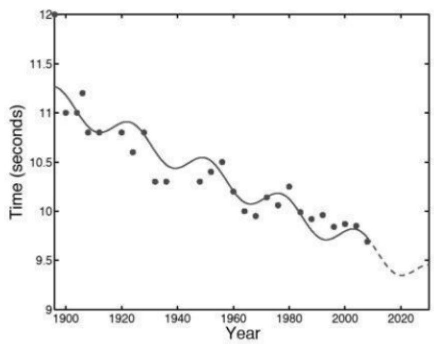
- Model Selection
  - Generalization and Overfitting
  - Method 1: Cross Validation
- Regularized Least Squares

- Probability Review
  - Definitions and Probability Calculus
  - Expectation
  - Continuous probability
  - Distributions
  - Likelihood

Linear Combination of *Basis Functions* (not just polynomials)

$$\mathbf{X} = \begin{bmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_K(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_K(x_2) \\ \vdots & \vdots & \cdots & \vdots \\ h_1(x_N) & h_2(x_N) & \cdots & h_K(x_N) \end{bmatrix}$$

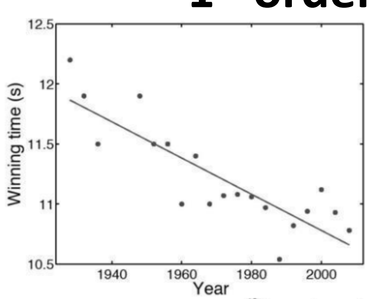
$$h_1(x) = 1$$
$$h_2(x) = x$$
$$h_3(x) = \sin\left(\frac{x-a}{b}\right)$$
$$f(x;\mathbf{w}) = w_0 + w_1 x + w_2 \sin\left(\frac{x-a}{b}\right).$$

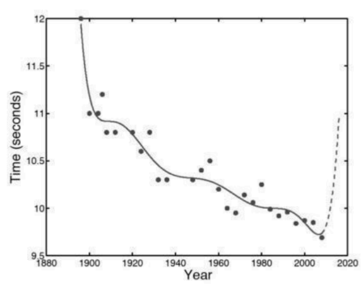


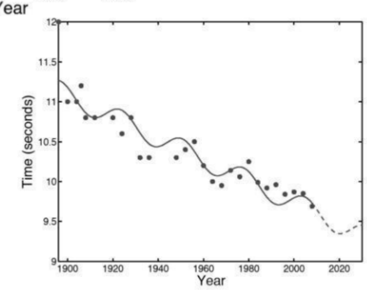
Careful !!
a and *b* must be **constants**

All parameters (as variables) must be ***linearly*** combined

Which Model is better: 1^{st} order, 8^{th} order ?

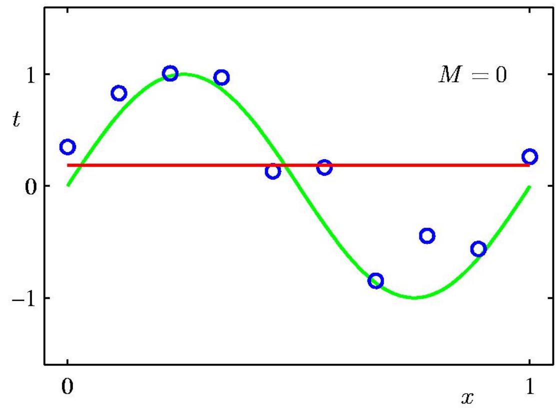




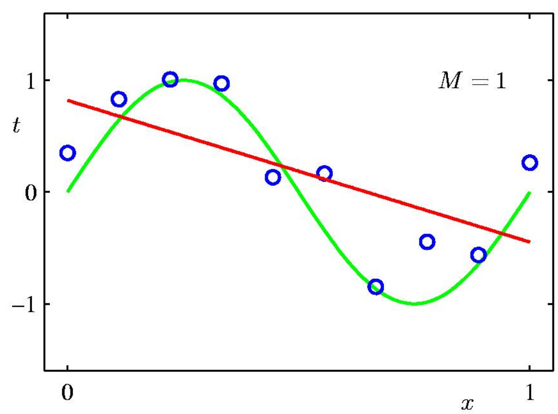


… periodic?

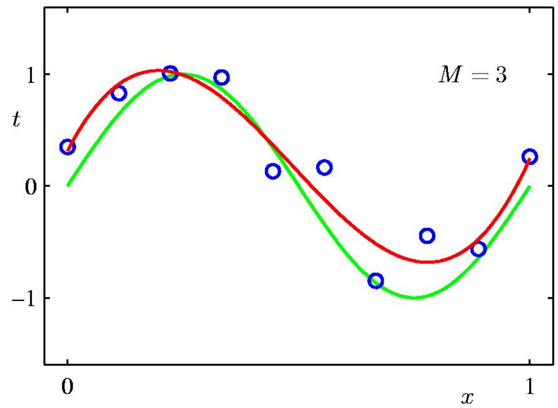
0^{th} Order Polynomial



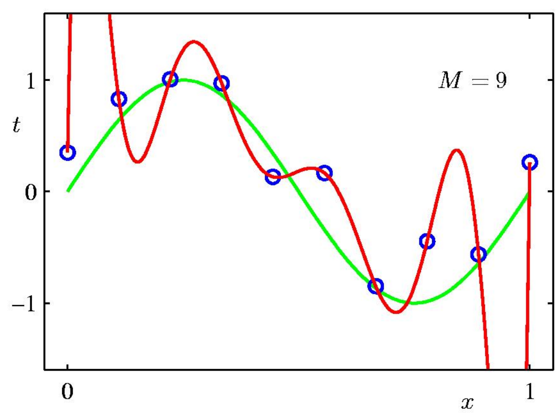
1^{st} Order Polynomial



3rd Order Polynomial



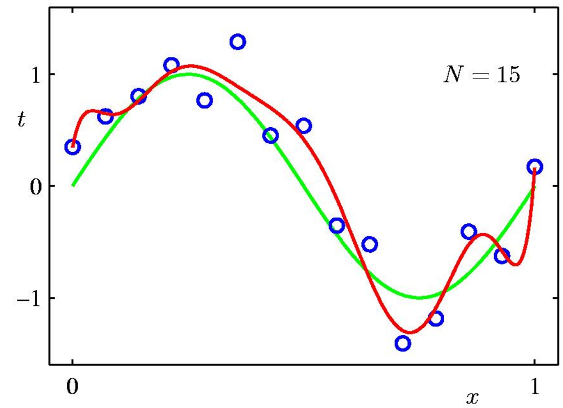
9th Order Polynomial



Overfitting

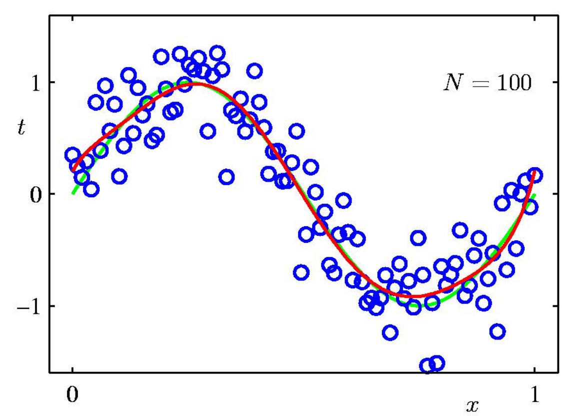
Data Set Size:
$N = 15$

9th Order Polynomial

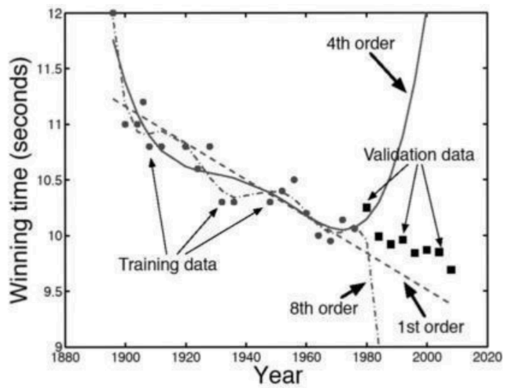


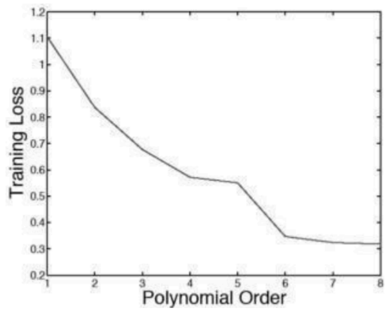
Data Set Size:
$N = 100$

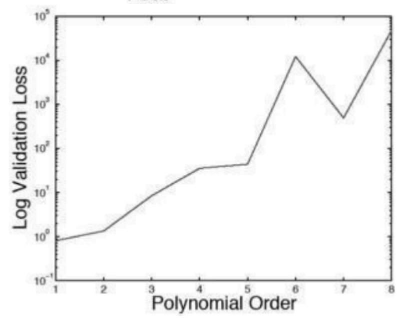
9th Order Polynomial



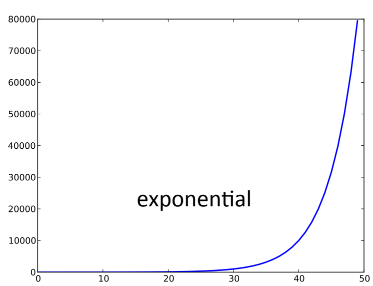
Training versus Testing

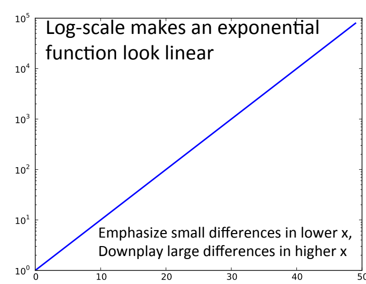


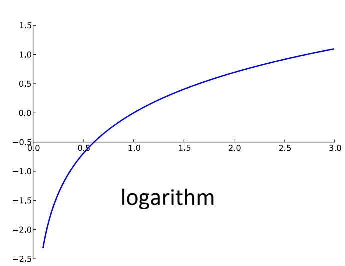




Sidenote: Log scale





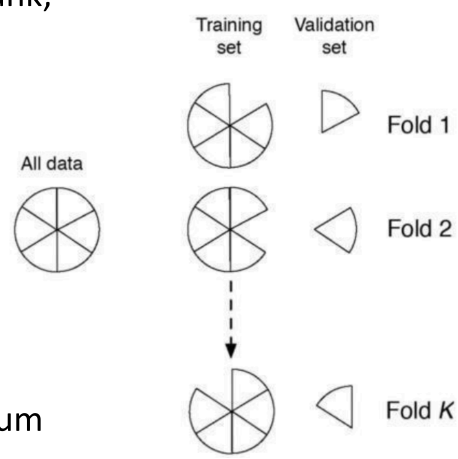


Cross-Validation

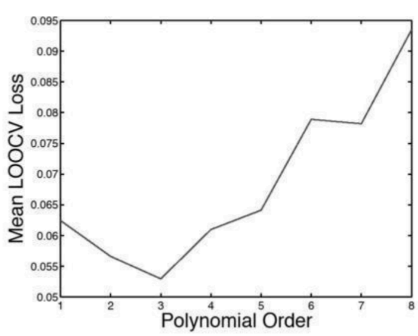
Randomly split your data into a set of *k* chunks
"hold out" a chunk of the data set;
train on everything but that chunk;
test with the chunk
Repeat this for all chunks

What this does:
Estimates the error
Of a number of possible
Models trained on data subsets

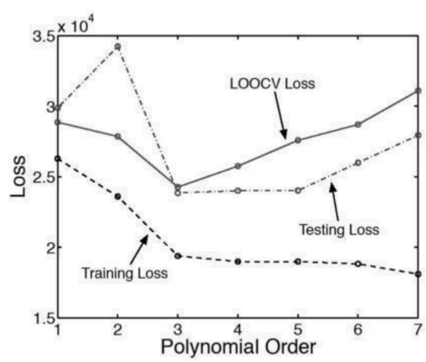
Leave-one-out-CV (LOOCV)
… same thing, but chunk = 1 datum



LOOCV for Model Selection

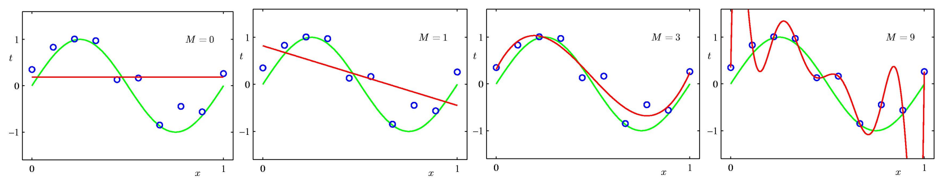


On Men's 100 meter data
Trying different orders of polynomials
for the models



Study with artificial data (3rd order poly)
Sample size: 50
Test error based on 1000 indep samples

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

SISTA 15

Regularization

- Penalize large coefficient values: add magnitude of all of the weights (e.g., their sum) as part of the loss.

$$\sum_i w_i^2 = \mathbf{w}^\top \mathbf{w} \qquad \mathcal{L}' = \mathcal{L} + \lambda \mathbf{w}^\top \mathbf{w}$$

$$\begin{aligned}
\mathcal{L}' &= \mathcal{L} + \lambda \mathbf{w}^\top \mathbf{w} \\
&= \frac{1}{N}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \frac{2}{N}\mathbf{w}^\top \mathbf{X}^\top \mathbf{t} + \lambda \mathbf{w}^\top \mathbf{w} \\
\frac{\partial \mathcal{L}'}{\partial \mathbf{w}} &= \frac{2}{N}\mathbf{X}^\top \mathbf{X}\mathbf{w} - \frac{2}{N}\mathbf{X}^\top \mathbf{t} + 2\lambda \mathbf{w}
\end{aligned}$$

Note: We've already removed $\mathbf{t}^\top \mathbf{t}$ from $\mathcal{L}$ because we'll be taking the derivative with respect to **w**.

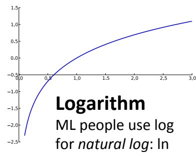
$$\begin{aligned}
\frac{2}{N}\mathbf{X}^\top \mathbf{X}\mathbf{w} - \frac{2}{N}\mathbf{X}^\top \mathbf{t} + 2\lambda \mathbf{w} &= 0 \\
\left(\mathbf{X}^\top \mathbf{X} + N\lambda \mathbf{I}\right)\mathbf{w} &= \mathbf{X}^\top \mathbf{t} \\
\hat{\mathbf{w}} &= \left(\mathbf{X}^\top \mathbf{X} + N\lambda \mathbf{I}\right)^{-1}\mathbf{X}^\top \mathbf{t}
\end{aligned}$$

Including a regularization term also ensures the inverse matrix is non-singular (which happens when X^TX has some columns that are colinear, or nearly so (leading to very large magnitude w values); near colinearity is not uncommon in real data).

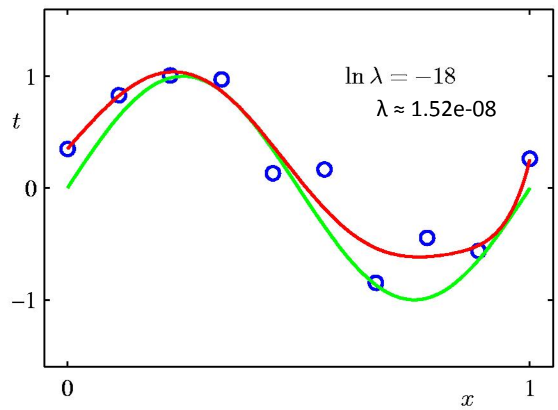
SISTA 16

Regularization:

$$\ln \lambda = -18$$



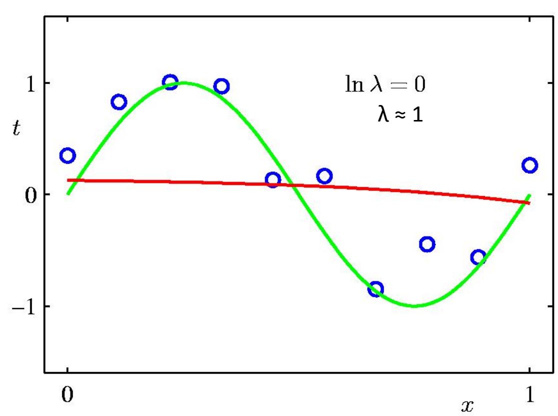
Logarithm
ML people use log
for *natural log*: ln



17

Regularization:

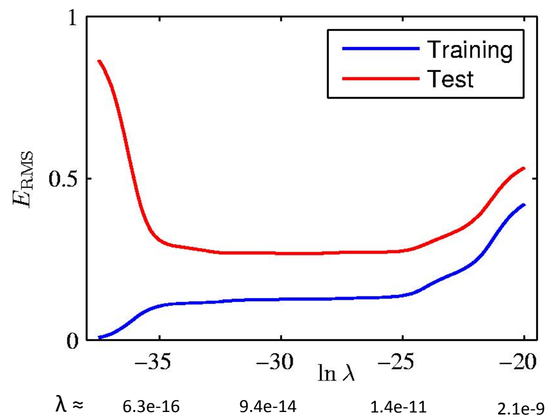
$$\ln \lambda = 0$$



18

Regularization:

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

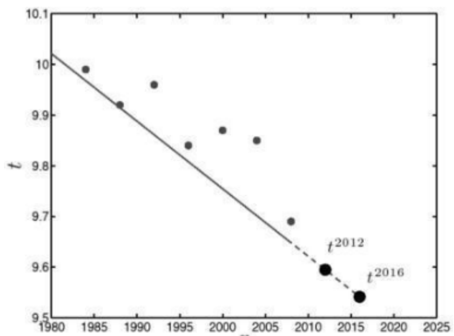


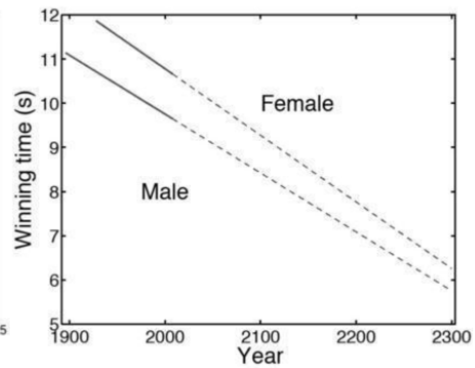
Polynomial Coefficients

	λ = 0 $\ln \lambda = -\infty$	λ = very small $\ln \lambda = -18$	λ = 1 $\ln \lambda = 0$
$w_0^\star$	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

Predicting with a learned model

Prediction: $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new} = \sum_{i=0}^{k} x_{new,i} w_i$





2592: look out boys!

3000: -3.5 seconds ??!

SISTA 21