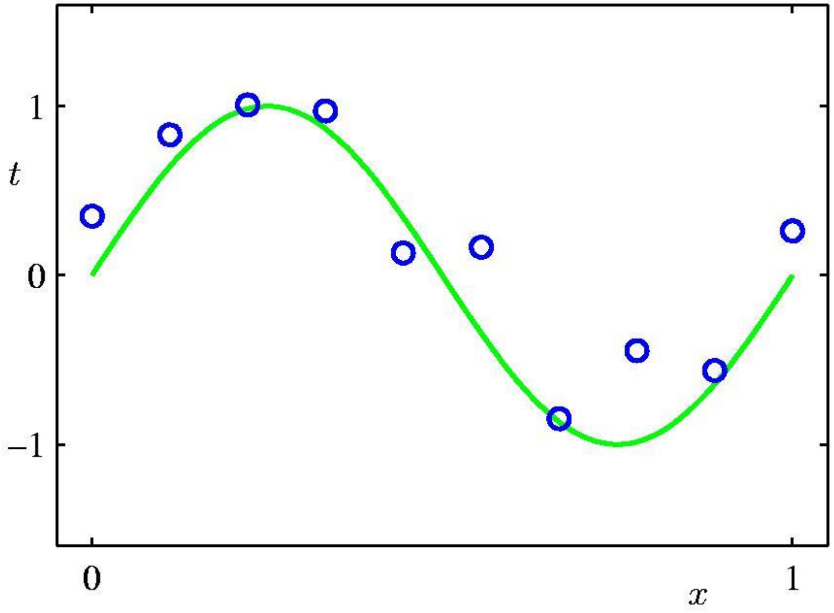
Overfitting and regularization

Case Study: Polynomial Curve Fitting

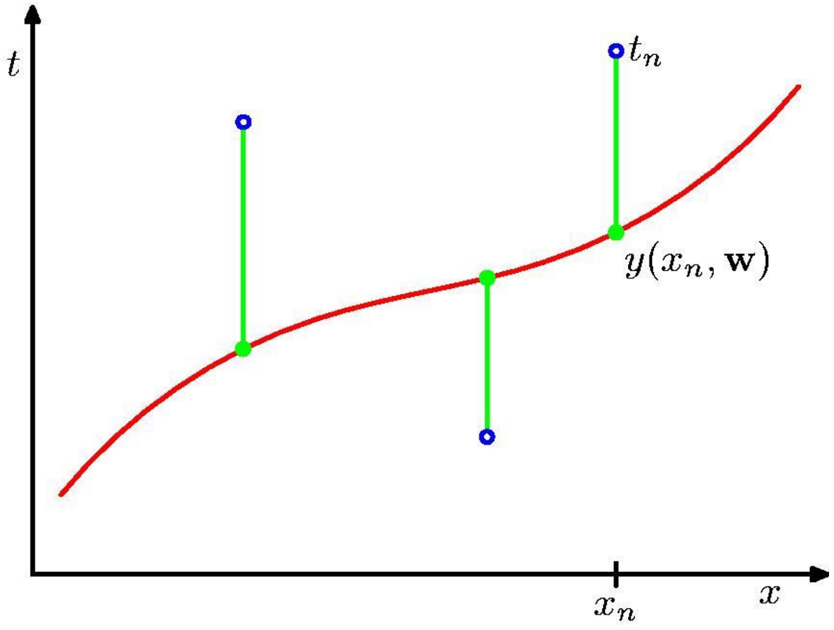
Suppose we observe a real-valued input variable x and we wish to use this observation to predict the value of a real-valued target variable t.



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

Sum-of-Squares Error Function

The values of the coefficients will be determined by fitting the polynomial to the training data. This can be done by minimizing an error function that measures the misfit between the function y(x,w), for any given value of w, and the training set data points.

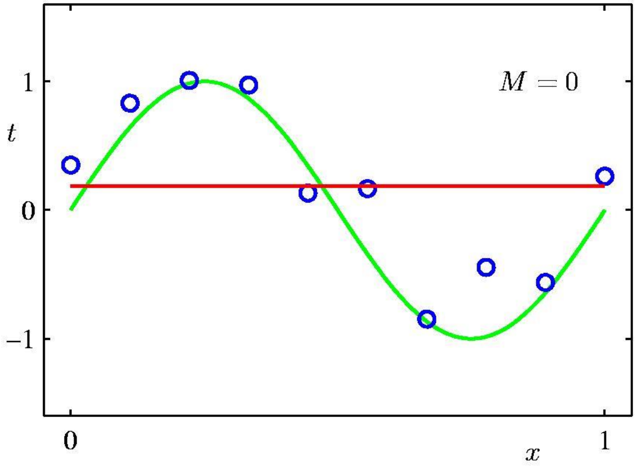


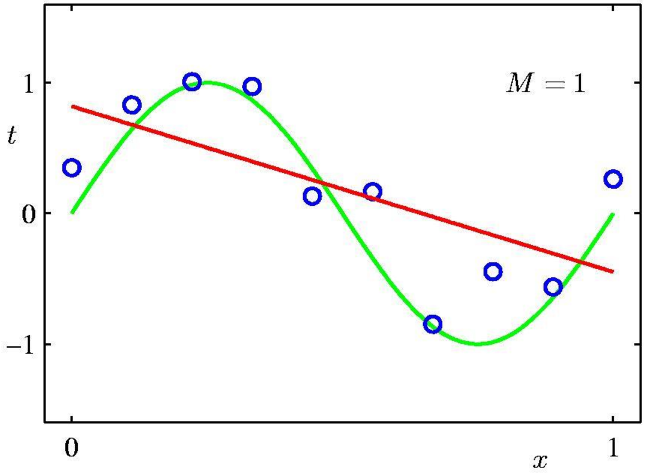
The sum of the squares error function:

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

How to choose the order M?

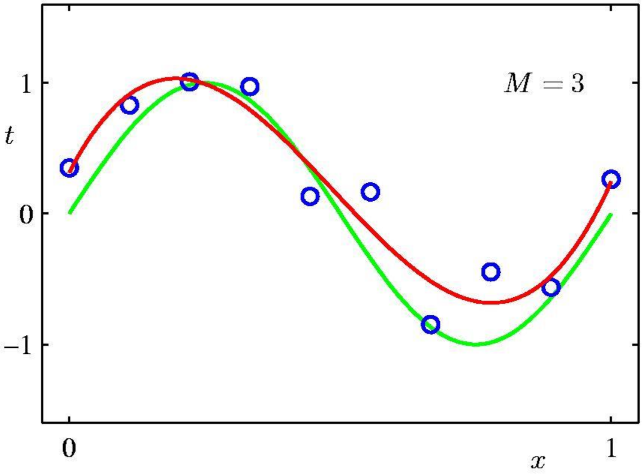
M=0

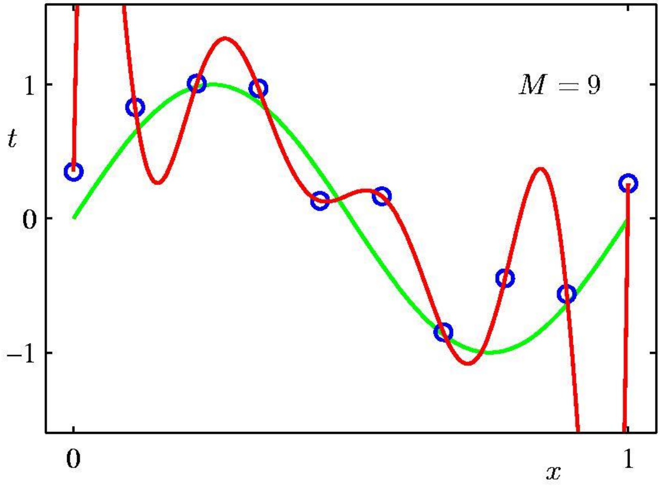




M=1

M=3





M=9

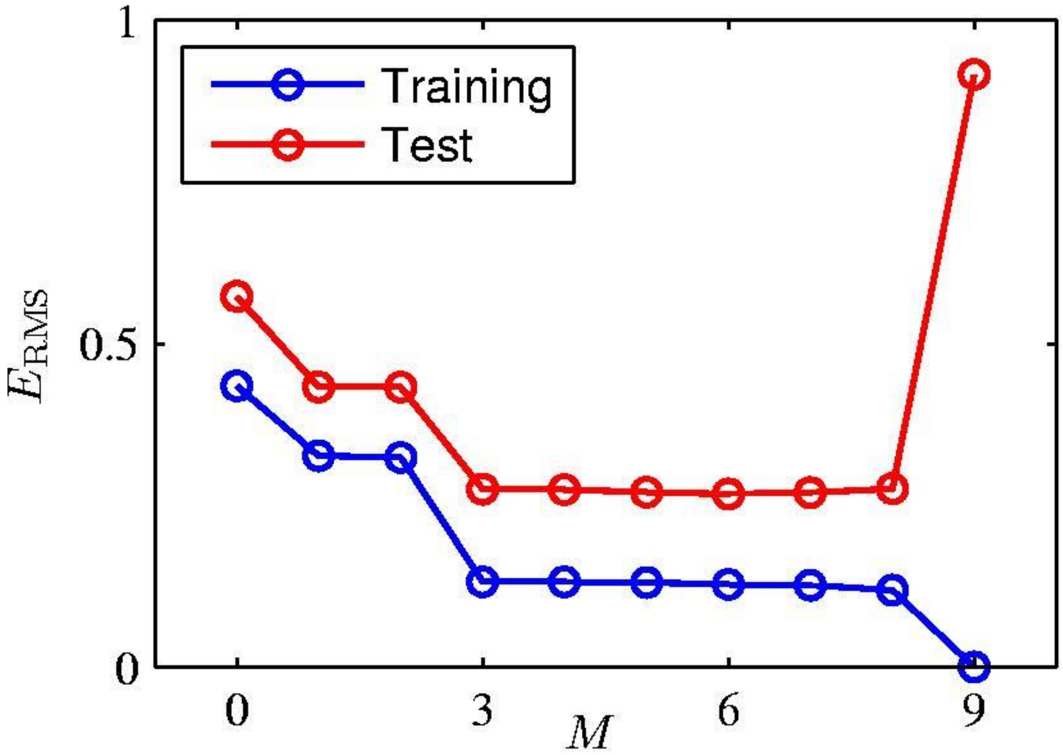
Observations

- The constant (M = 0) and first order (M = 1) polynomials give rather poor fits to the data.

- The third order (M = 3) polynomial seems to give the best fit to the data.

- Using a much higher order polynomial (M = 9), we obtain an excellent fit to the training data. However, the fitted curve oscillates wildly and gives a very poor representation. This leads to over-fitting.

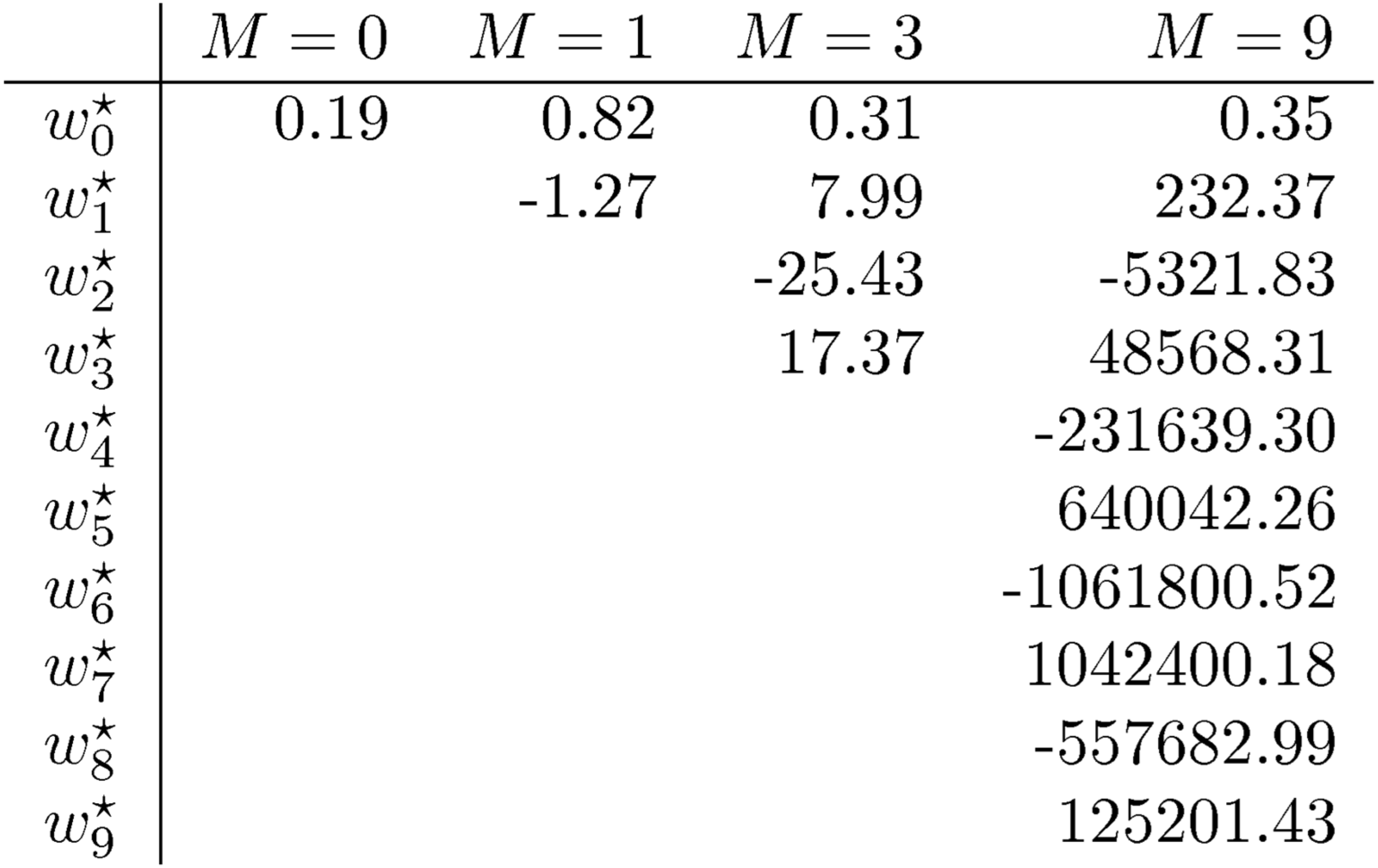
Over-fitting



poor generalization

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

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

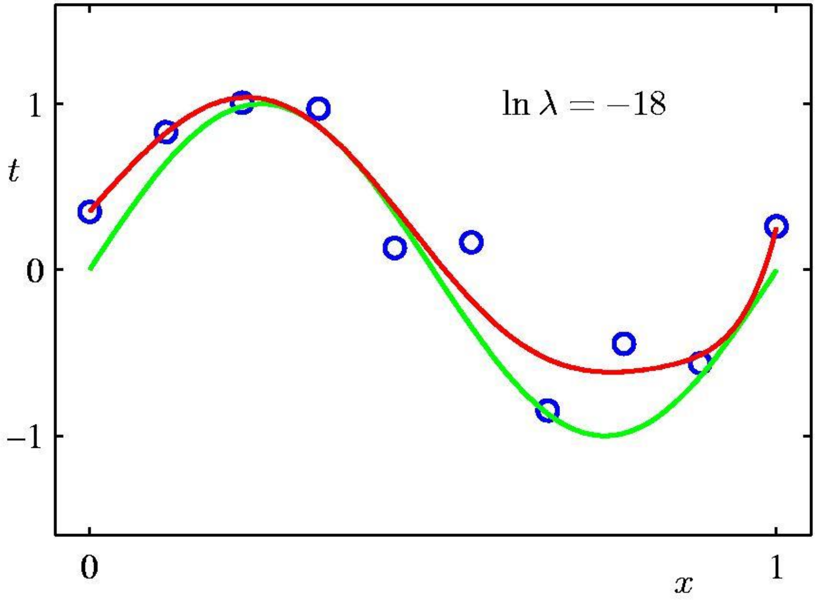
Regularization

- ❑ One technique that is often used to control the over-fitting phenomenon in such cases is that of regularization, which involves adding a penalty term to the error function in order to discourage the coefficients from reaching large values.

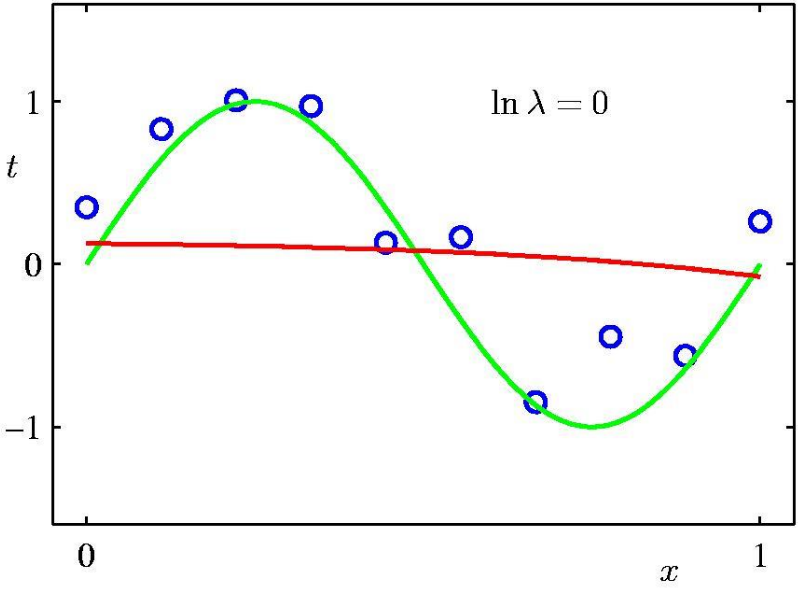
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- ❑ The coefficient λ governs the relative importance of the regularization term compared with the sum-of-squares error term.

Effect of Regularization Parameter

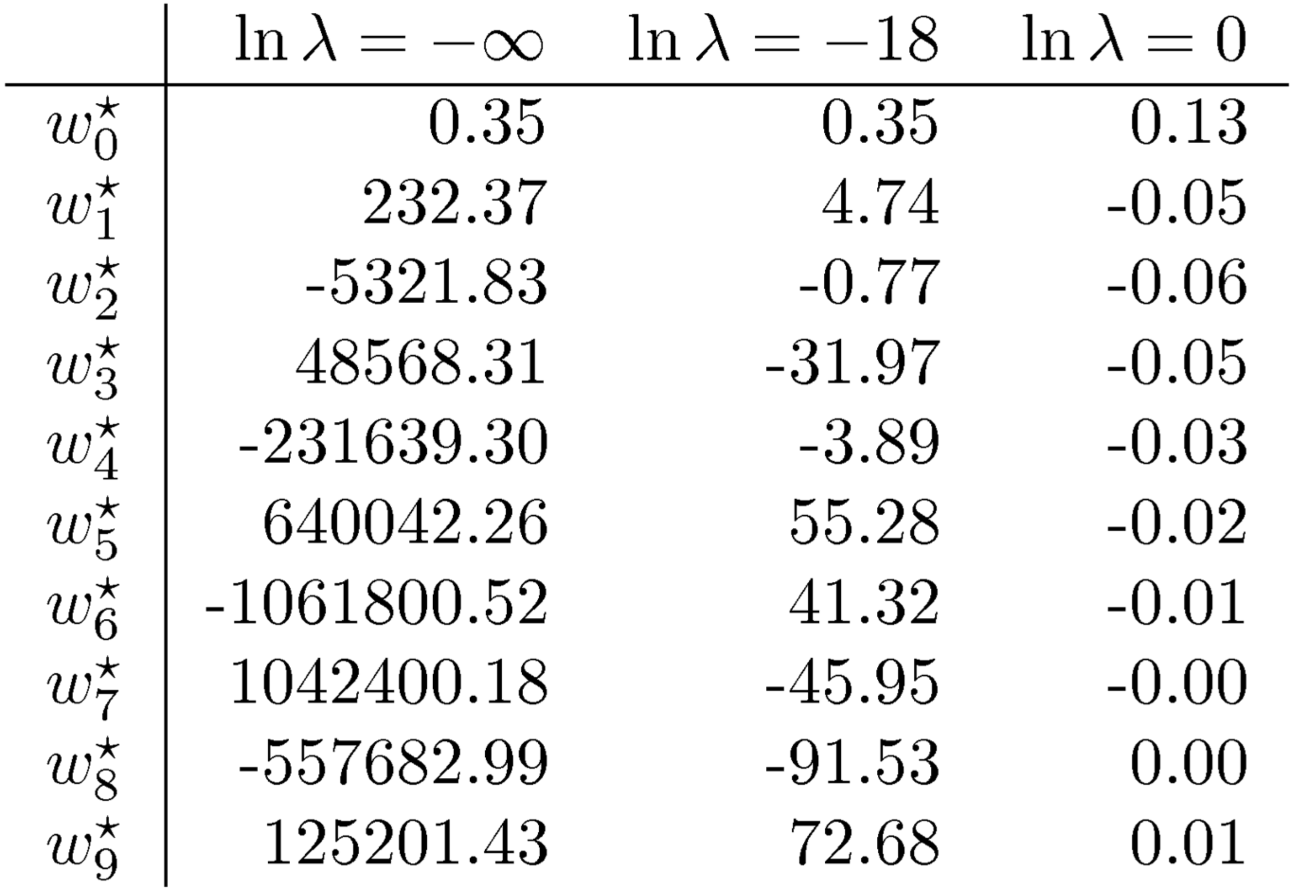


$$\ln \lambda = -18$$



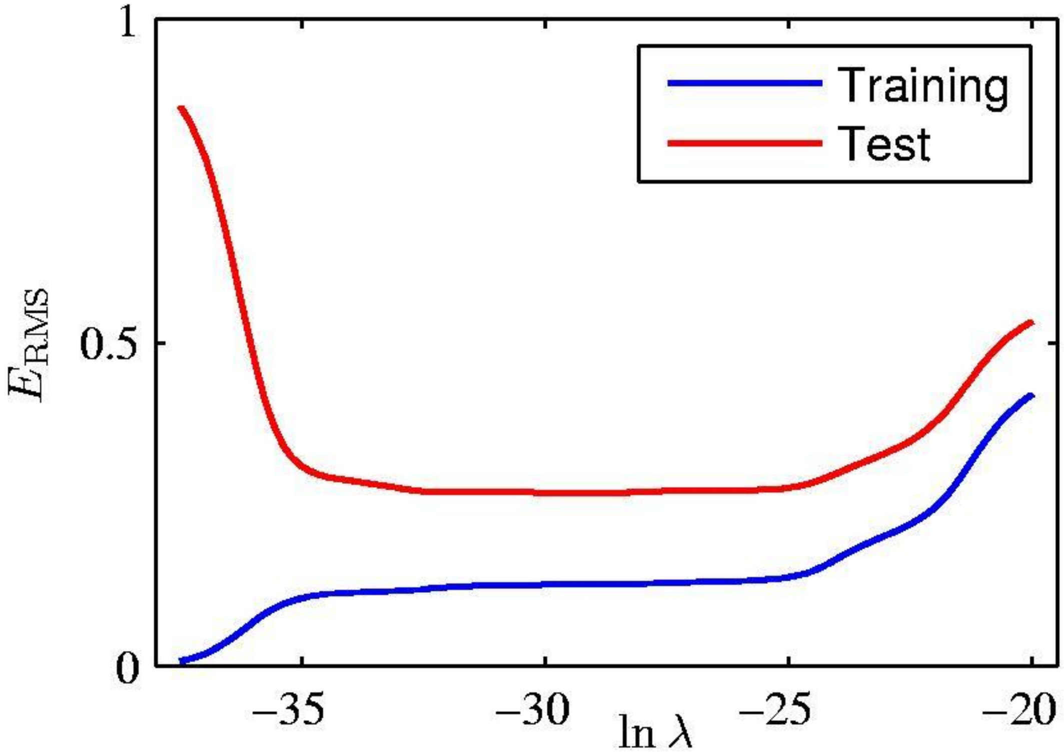
$$\ln \lambda = 0$$

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

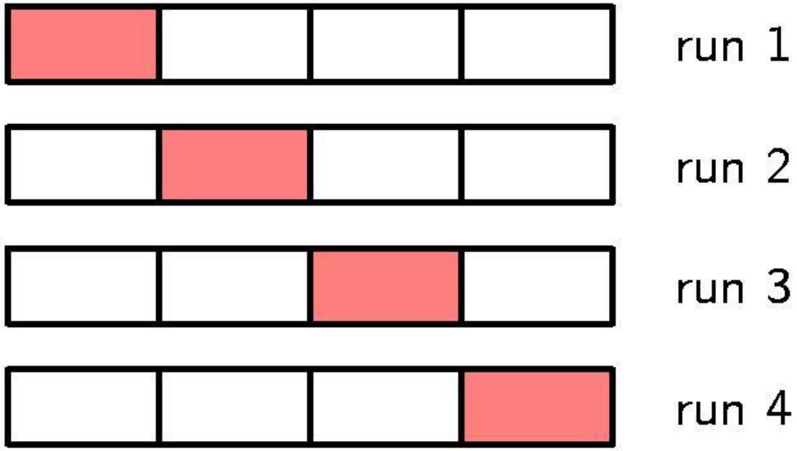
Regularization: E_{RMS} vs. $\ln \lambda$



Model selection: Estimation of the optimal value of the regularization parameter. In practice, cross validation is commonly applied for model selection.

Model Selection

Cross-Validation



S-fold cross-validation*:*

Partition the data into S groups of equal size. Then S – 1 of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is repeated for all S possible choices for the held-out group (red blocks), and the performance scores from the S runs are then averaged.