Bias and Variance

Bias of ML Estimate of Variance

• For a Gaussian distribution, maximum likelihood estimates for mean and variance are

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2$$

Systematically underestimates the variance

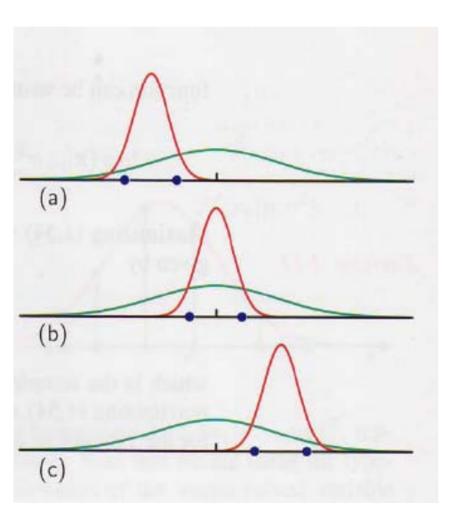
$$E[\mu_{ML}] = \mu$$

$$E[\sigma_{ML}^2] = \left(\frac{N-1}{N}\right)\sigma^2$$

• Following estimate for variance parameter is unbiased

$$\sigma_{ML}^{2} = \frac{1}{N-1} \sum_{n=1}^{N} (x_{n} - \mu_{ML})^{2}$$

Bias of Maximum Likelihood Estimate of Variance



- Green curve shows true Gaussian distribution
- Red curves shows three Gaussian distributions for three data sets (2 points each) obtained by the maximum likelihood estimates
- Averaged across 3 data sets mean is correct but variance is underestimated
- Because it is estimated relative to sample mean not relative to true mean

Bias and Variance in Statistics

- Bias is error that cannot be corrected by repeated experiments
- Bias-variance decomposition states that expected squared error is equal to the bias plus the random error
- You can reduce the variance but not the bias
- True value of the parameter is a constant
- Experimental estimate is a probilistic variable
- Bias is the systematic or average difference between these two variables and variance is the probabilistic component

Height of Emperor of China

True height is 200cm

- Every American believes it is 180cm
- Poll a random American and ask "How tall is the emperor of China?" The answer is always 180
- The error is always -20cm
- Average squared error is 400

- If Americans have normally distributed beliefs with mean 180cm and standard deviation 10cm
- Poll two Americans. One says 190cm and other says 170cm
- Errors are -10cm and -30cm
- Squared errors are 100 and 900
- Average error is -20
- Average squared error is 500

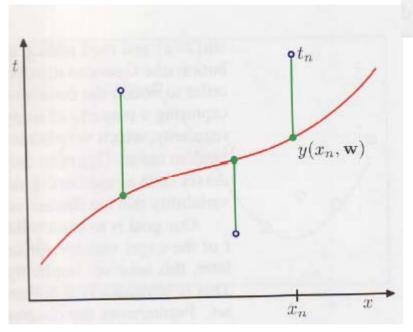
Average Squared Error (500) = Square of bias (-20) + Square of Random error (standard deviation of 10cm)

Simple Regression Problem

- Real valued input variable x
- Real valued target variable t

- N observations of x
- $\bullet \ \ X = (x_1, \dots x_N)$
- Corresponding observations of values of t
- $t = (t_1, \dots t_N)$

Polynomial Curve Fitting



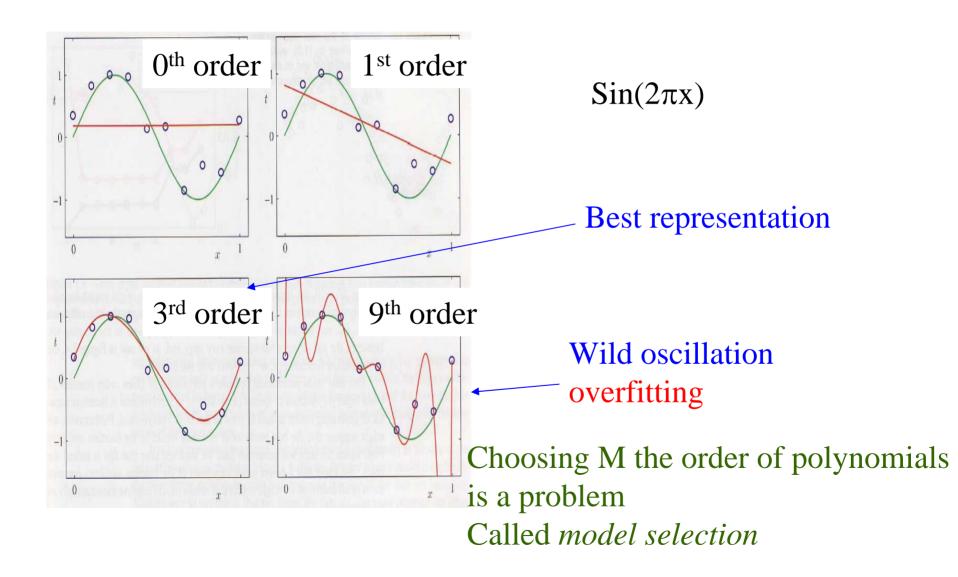
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

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

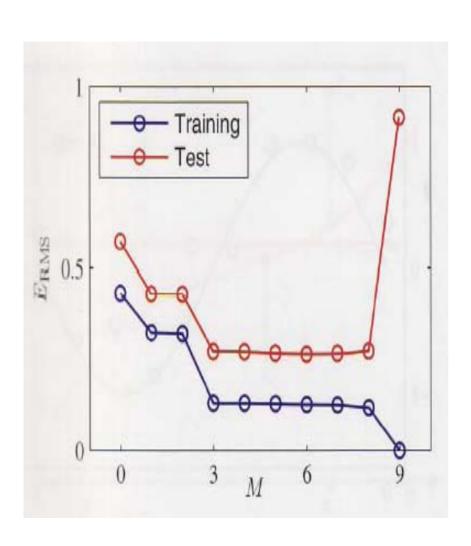
Error function is one half of the sum of squares of the displacements Shown by vertical green bars of each data point from the function y(x,w)

There is a unique solution obtained by computing derivative. Resulting polynomial is $y(x, w^*)$

Polynomials of Various Orders of t



Training and Test Set Error Rates



 More convenient measure is RMS error

$$E_{RMS} = \sqrt{2E(w^*)/N}$$

• Division by N allows different sizes of data sets to be compared

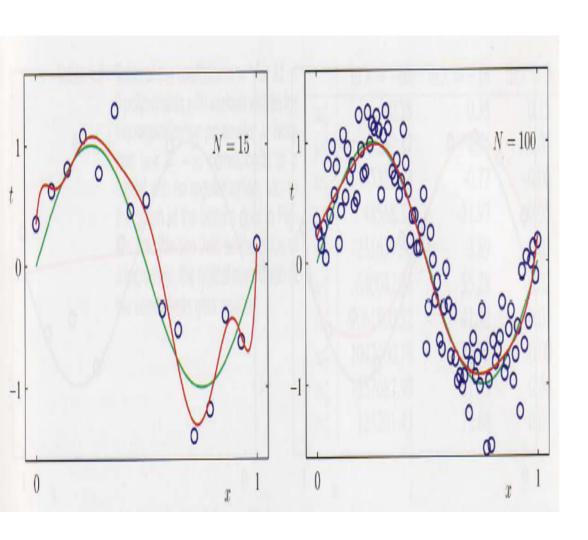
Table of coefficients

	M=0	M = 1	M = 6	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
v_3^*			17.37	48568.31
v_4^{\star}				-231639.30
v_5^{\star}				640042.26
v_6^{\star}				-1061800.52
U7*				1042400.18
v_8^{\star}				-557682.99
v_9^*				125201.43

Magnitude of coefficients

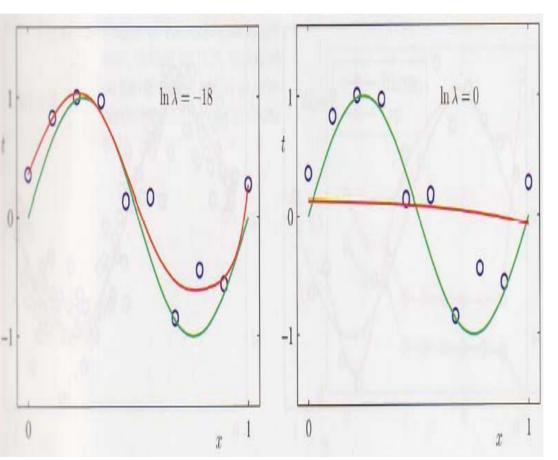
 increases dramatically as the order of polynomial increases

Behavior as size of data set is varied



- Increasing size of data set reduces overfitting problem
- Number of data points should be no less than 5 or 10 times the number of adaptive parameters in the model
- Unsatisfying to limit no of parameters based on data set rather than complexity of problem
- Bayesian approach solves these problems

Regularized Error Function



Adding a penalty term to the error function to discourage **Importance** Of regularization coefficients term reaching large values $\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{ y(x_n, \mathbf{w}) - t_n \right\}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2$

 $\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$

Error function can be minimized in closed form Methods called in statistics as *shrinkage methods*. In neural networks called as *weight decay*

Table of coefficients with regularization

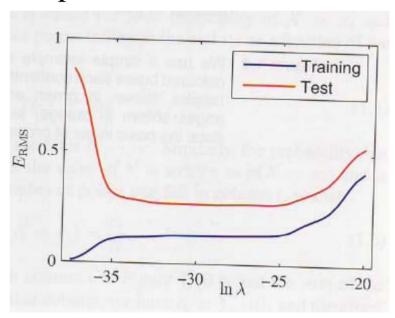
No regularization

As λ gets larger magnitude of coefficient

		geis	gets sman	
		*	*	
	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$	
	0.35	0.35	0.13	
	232.37	4.74	-0.05	
	-5321.83	-0.77	-0.06	
	48568.31	-31.97	-0.05	
	-231639.30	-3.89	-0.03	
1	640042.26	55.28	-0.02	
	-1061800.52	41.32	-0.01	
	1042400.18	-45.95	-0.00	
	-557682.99	-91.53	0.00	
	125201.43	72.68	0.01	

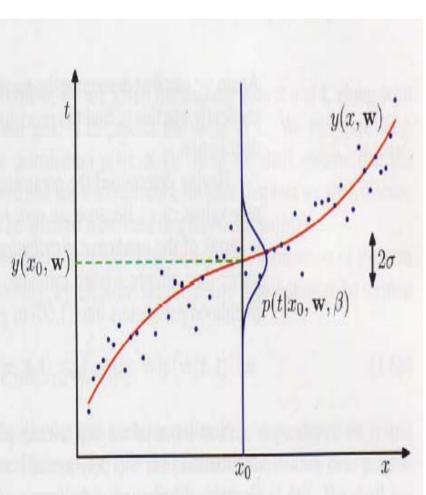
Impact of Regularization

Graph of RMS error versus $ln \lambda$



 λ controls the effective complexity of model and hence overfitting

Curve Fitting from a Probabilistic Perspective



$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \left\{ y(x_n, \mathbf{w}) - t_n \right\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

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

Results in a probability distribution

For the answer rather than a single point

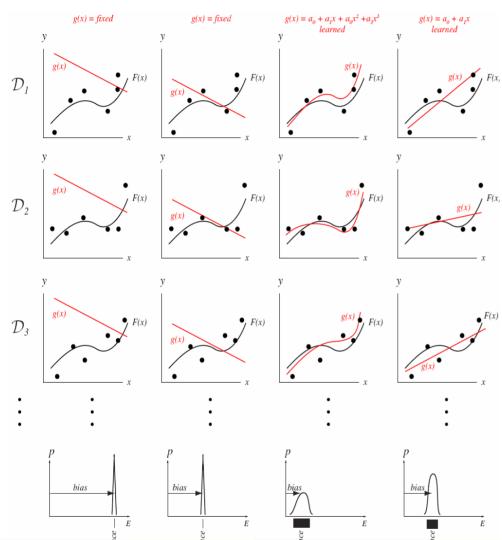
Bias Variance Decomposition

- True but unknown function F(x)
- Estimate F(.) based on n samples in set D
- Regression function estimated is denoted g(x;D)
- Effectiveness of the estimator is expressed as the mean squared deviation from the desired optimal

$$\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x}))^{2}\right]$$

$$= \underbrace{(\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x})])^{2}}_{bias^{2}} + \underbrace{\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^{2}\right]}_{variance}$$

Bias Variance Dilemma



- Each column is a different model
- Each row is a different set of dataD
- Probability

 function of mean
 squared error
 shown

Poor Model High bias zero variance Better Model Lower bias zero variance

Cubic
Model
Low bias
moderate variance

Linear Model to fit data intermediate bias intermediate variance