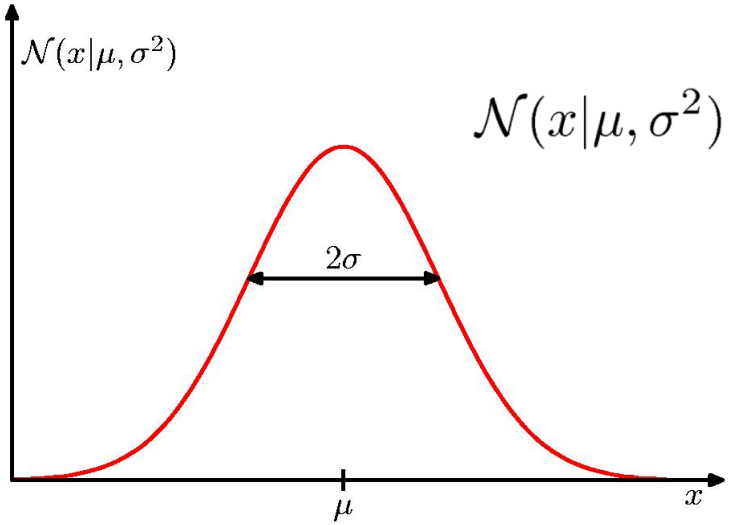
Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

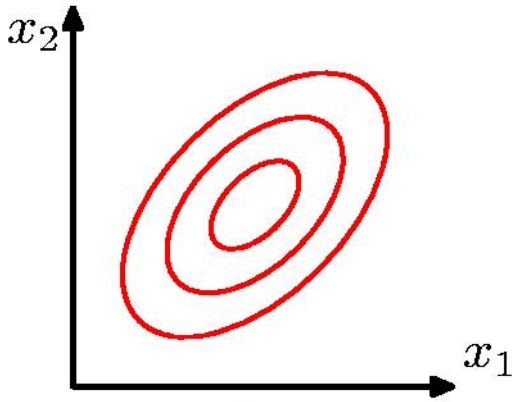
Lecture 8:
Linear Regression & Least Squares
Bayesian Linear Regression & Prediction

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Gaussian Distributions



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Simplest joint distribution that can capture arbitrary mean & covariance
- Justifications from *central limit theorem* and *maximum entropy* criterion
- Probability density above assumes covariance is *positive definite*
- ML parameter estimates are *sample mean* & *sample covariance*

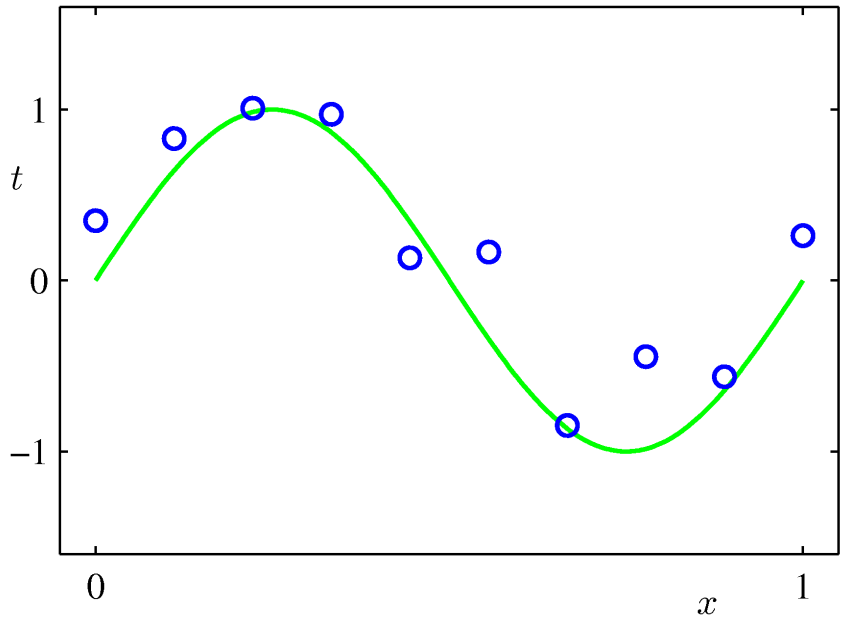
A Change in Direction

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

- GOAL: Predict label/response y from feature x
- Generative classification: Apply Bayes' rule to learned p(x,y)
- Discriminative or conditional regression & classification: directly learn a model of p(y | x), assuming x always given

Linear Basis Function Models (1)

- Example: Polynomial Curve Fitting



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

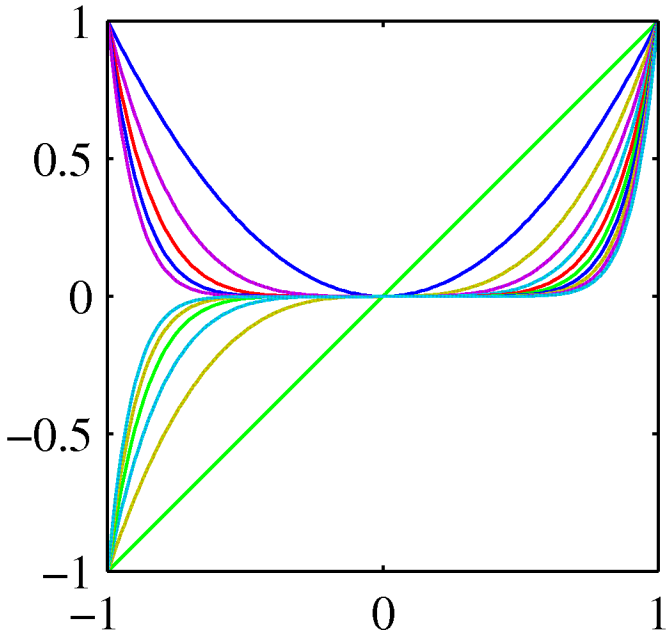
Slides adapted from Bishop's "Pattern Recognition and Machine Learning"
Notation differs from Murphy's "Machine Learning: A Probabilistic Perspective"

Linear Basis Function Models (2)

- Polynomial basis functions:

$$\phi_j(x) = x^j$$

- These are global: a small change in x affects all basis functions.

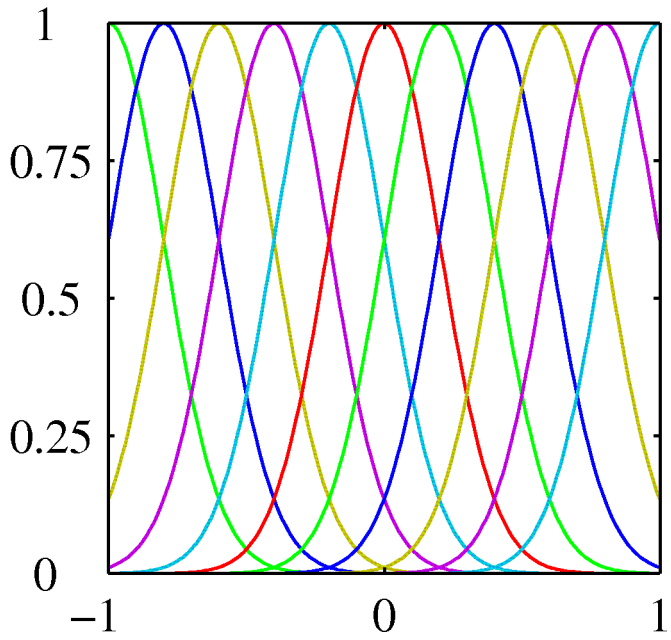


Linear Basis Function Models (3)

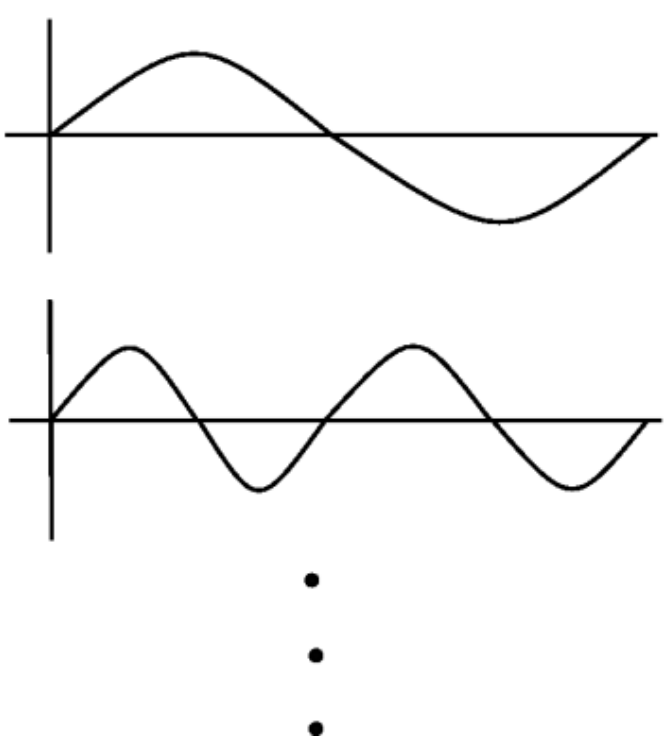
- Radial basis functions:

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

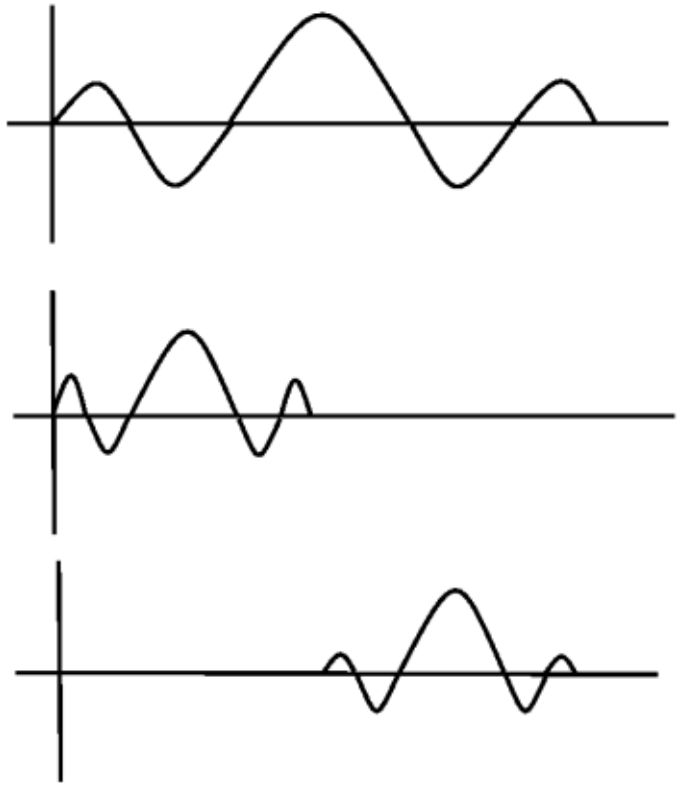
- These are local: a small change in x only affects nearby basis functions. Parameters control location and scale (width).



Linear Basis Function Models (4)

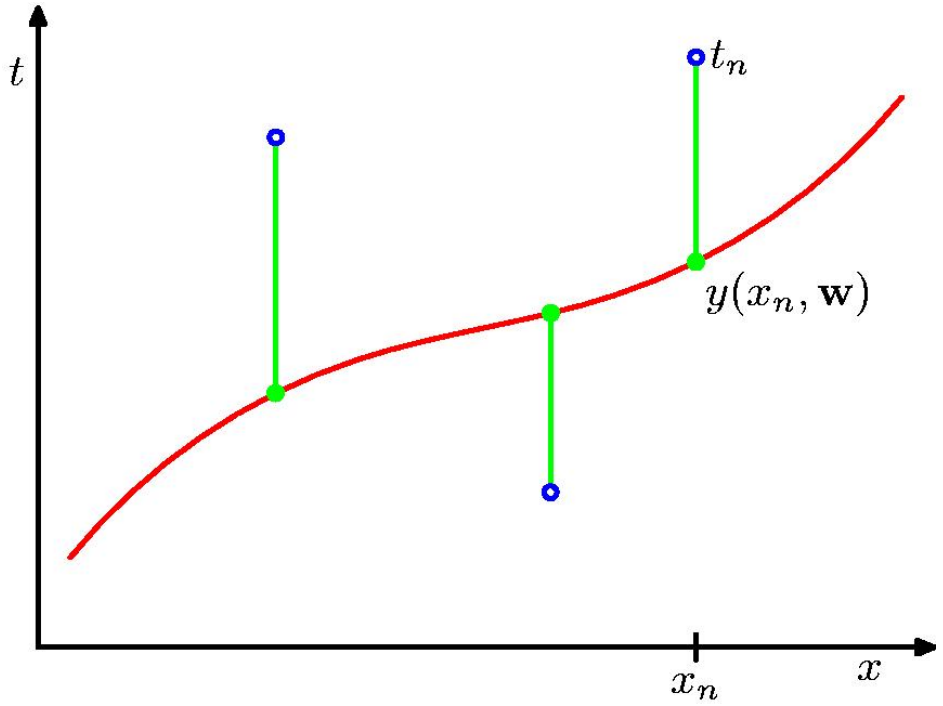


Fourier Basis



Wavelet Basis

Sum-of-Squares Error Function



$N \longrightarrow$ *number of examples*

$M \longrightarrow$ *number of features*

$t_n \longrightarrow$ *output or response*

$x_n \longrightarrow$ *input or covariates*

$$y(x_n, w) = \phi(x_n)^T w$$

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \phi(x_n)^T w)^2 = \frac{1}{2}||t - \Phi w||^2$$

Equivalent to maximum likelihood (ML) estimation under a Gaussian model:

$$p(t_n \mid w, x_n) = \mathcal{N}(t_n \mid \phi(x_n)^T w, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left\{-\frac{\beta}{2}(t_n - \phi(x_n)^T w)^2\right\}$$

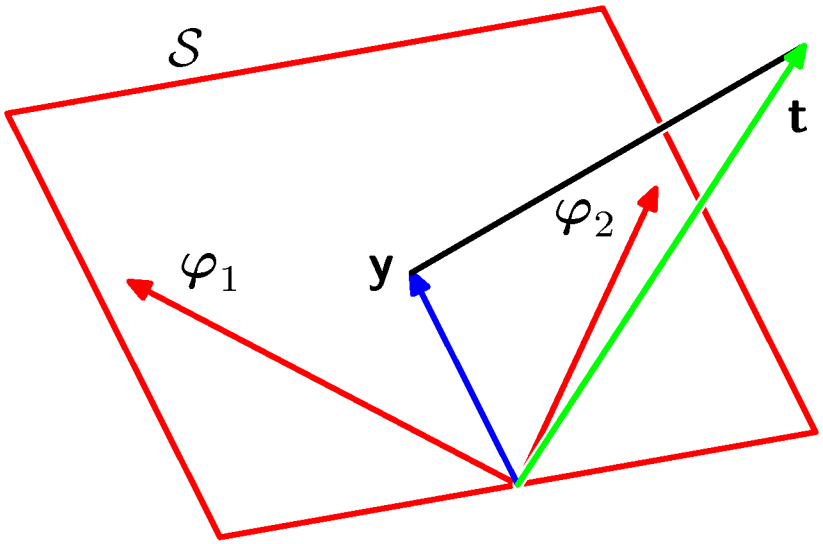
Geometry of Least Squares

- Consider

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w}_{\mathrm{ML}} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M]\,\mathbf{w}_{\mathrm{ML}}.$$

$$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T} \qquad \mathbf{t} \in \mathcal{T}$$

N-dimensional
M-dimensional



- S is spanned by $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M$.
- w_{ML} minimizes the distance between t and its orthogonal projection on S, i.e. y.

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \phi(x_n)^T w)^2 = \frac{1}{2}||t - \Phi w||^2$$

Finding Least Squares Solution

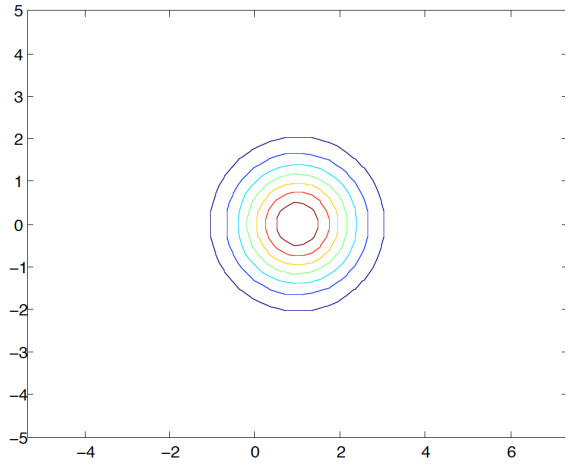
$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \phi(x_n)^T w)^2 = \frac{1}{2}||t - \Phi w||^2$$

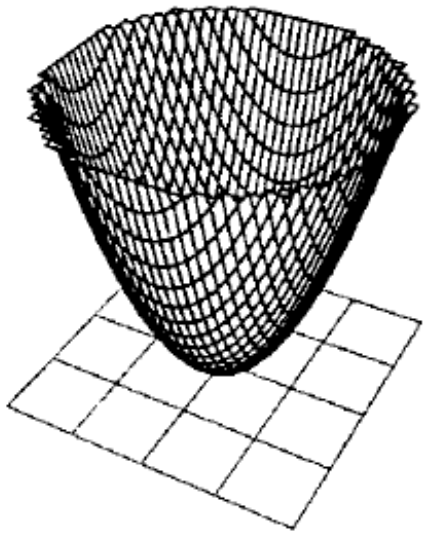
Gradient vectors:

$$f : \mathbb{R}^M \rightarrow \mathbb{R}$$

$$\nabla_w f : \mathbb{R}^M \rightarrow \mathbb{R}^M$$

$$(\nabla_w f(w))_k = \frac{\partial f(w)}{\partial w_k}$$





Gradient identities:

$$f(w) = \frac{1}{2}w^T A w + b^T w$$

$$\nabla_w f(w) = \frac{1}{2}(A + A^T)w + b$$

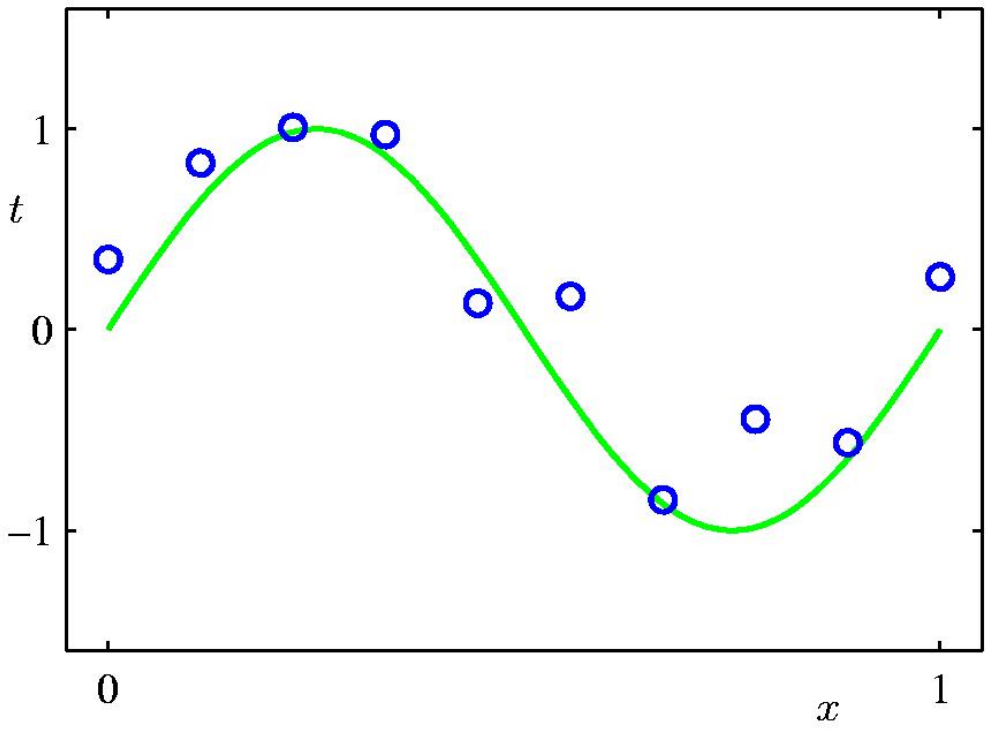
Normal equations:

$$\Phi^T \Phi \hat{w} = \Phi^T t$$

there is a unique solution if and only if

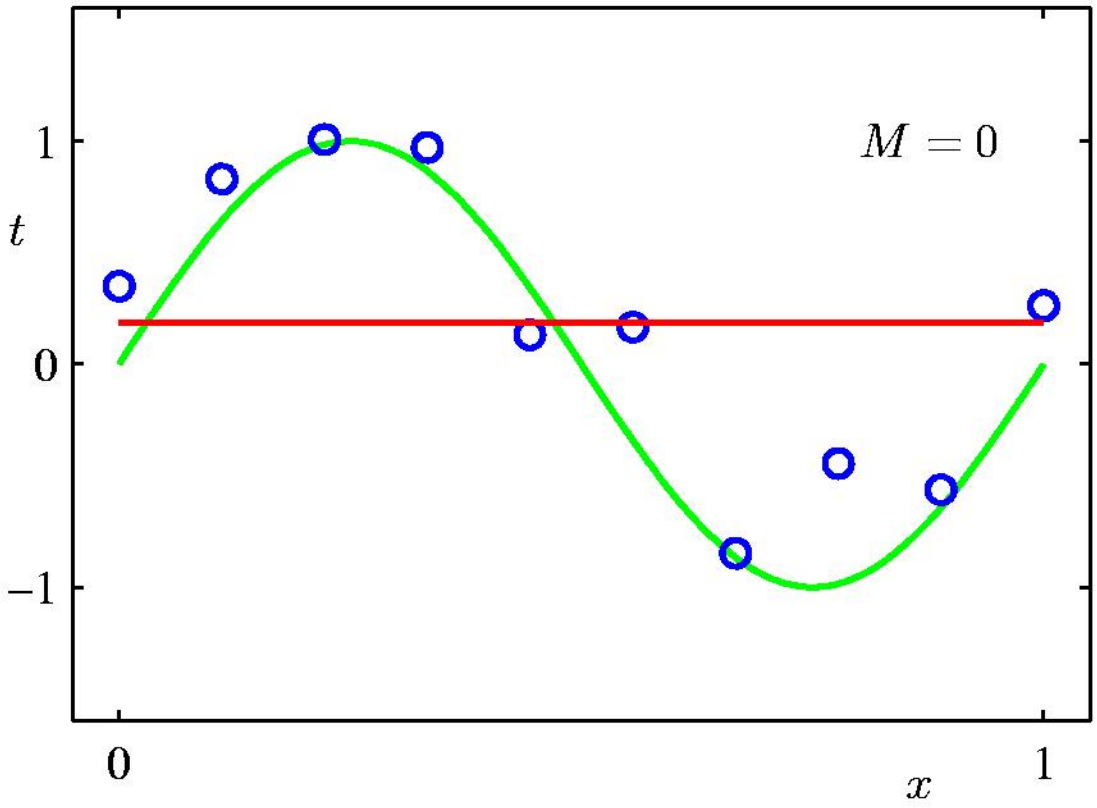
$\mathrm{rank}(\Phi) = M$ *(requires* $N \geq M$*)*

Polynomial Curve Fitting

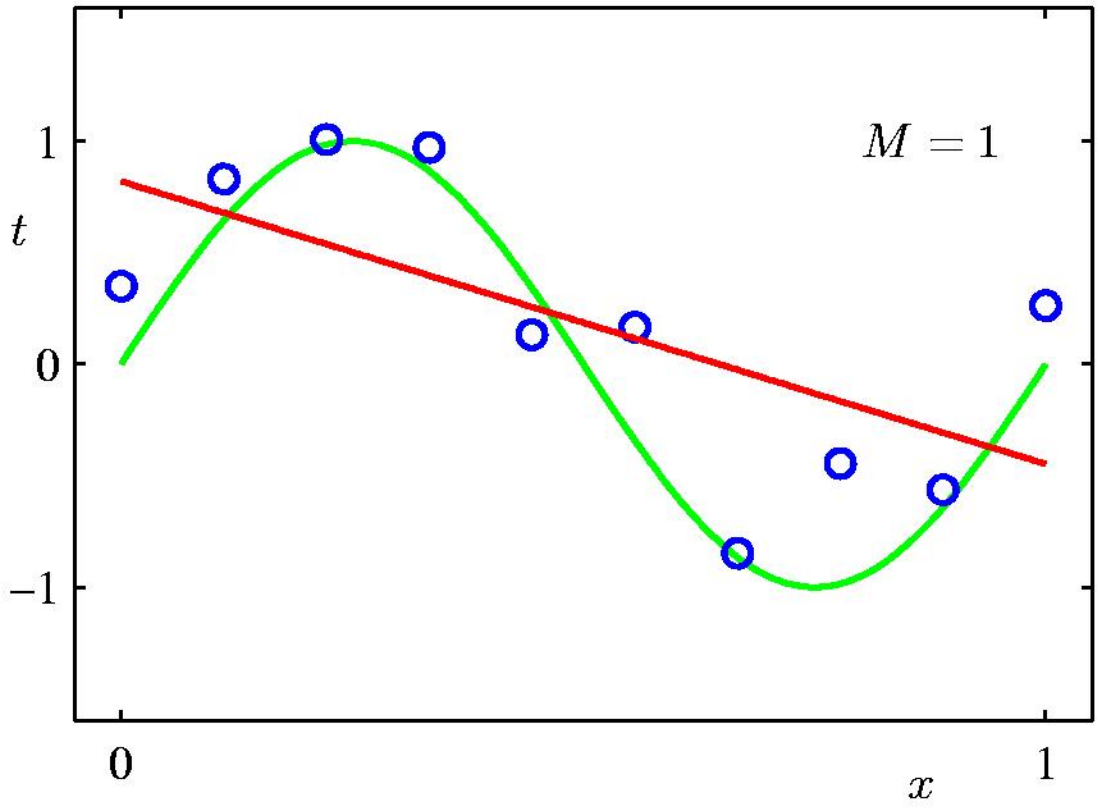


$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

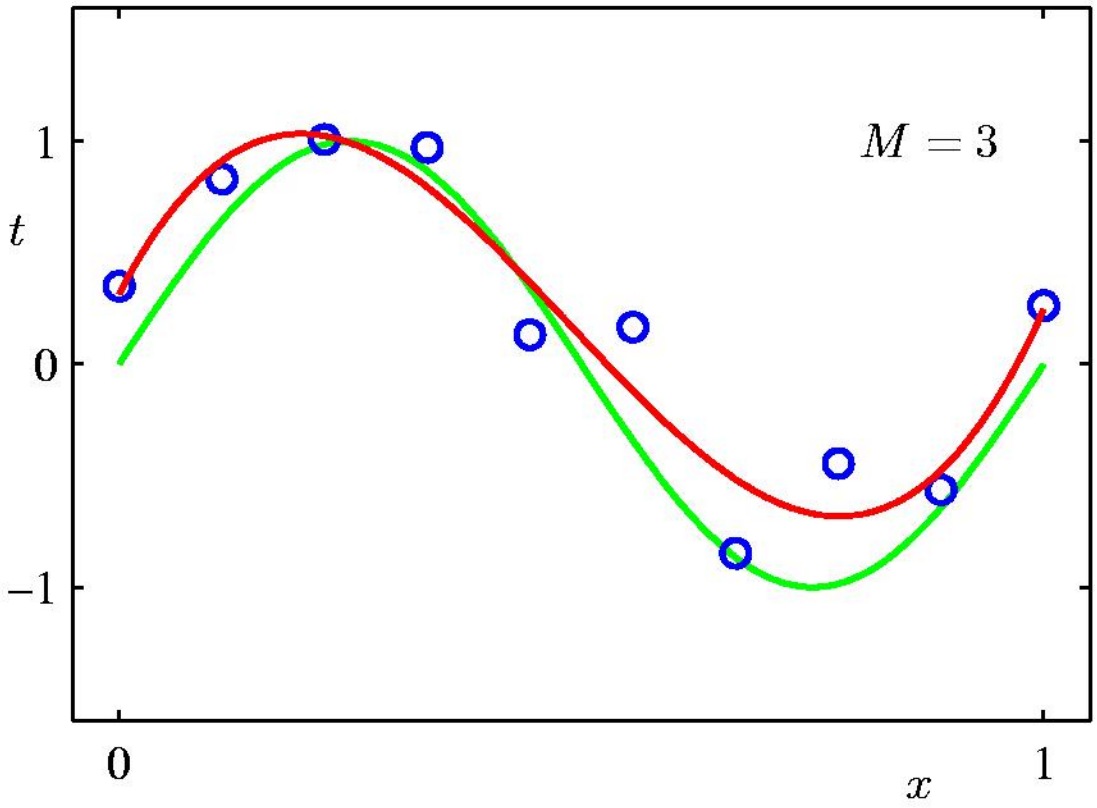
0^{th} Order Polynomial



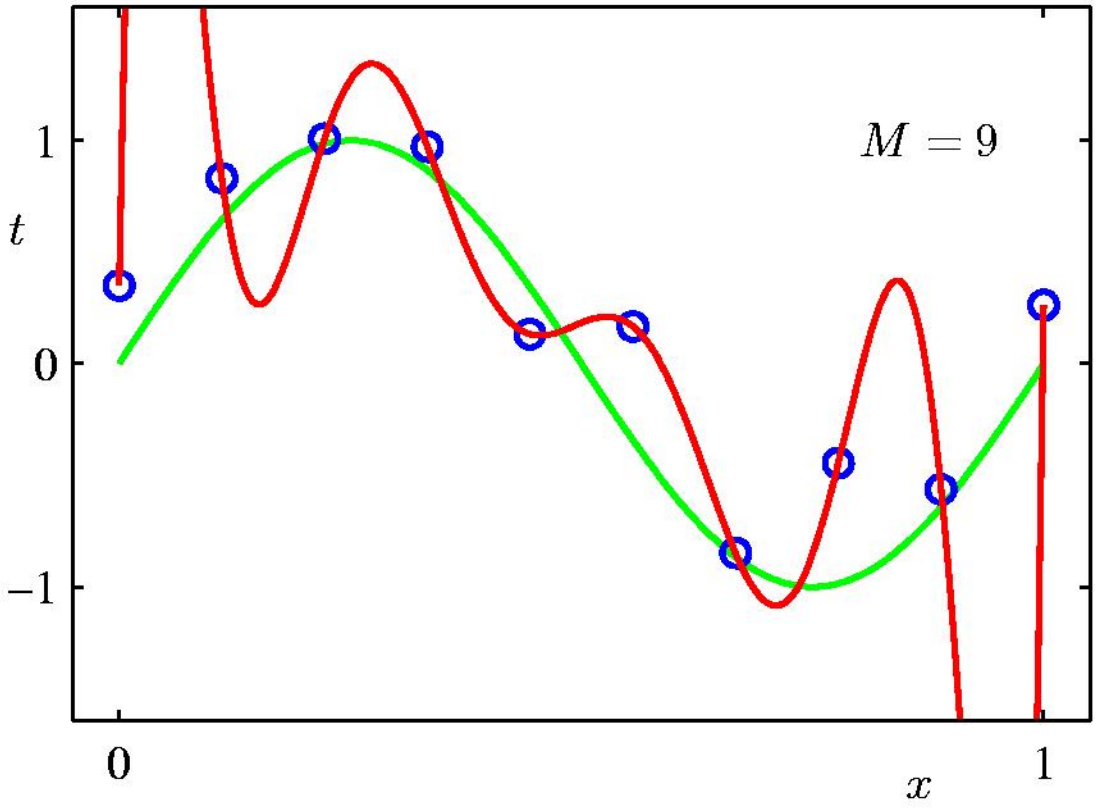
1st Order Polynomial



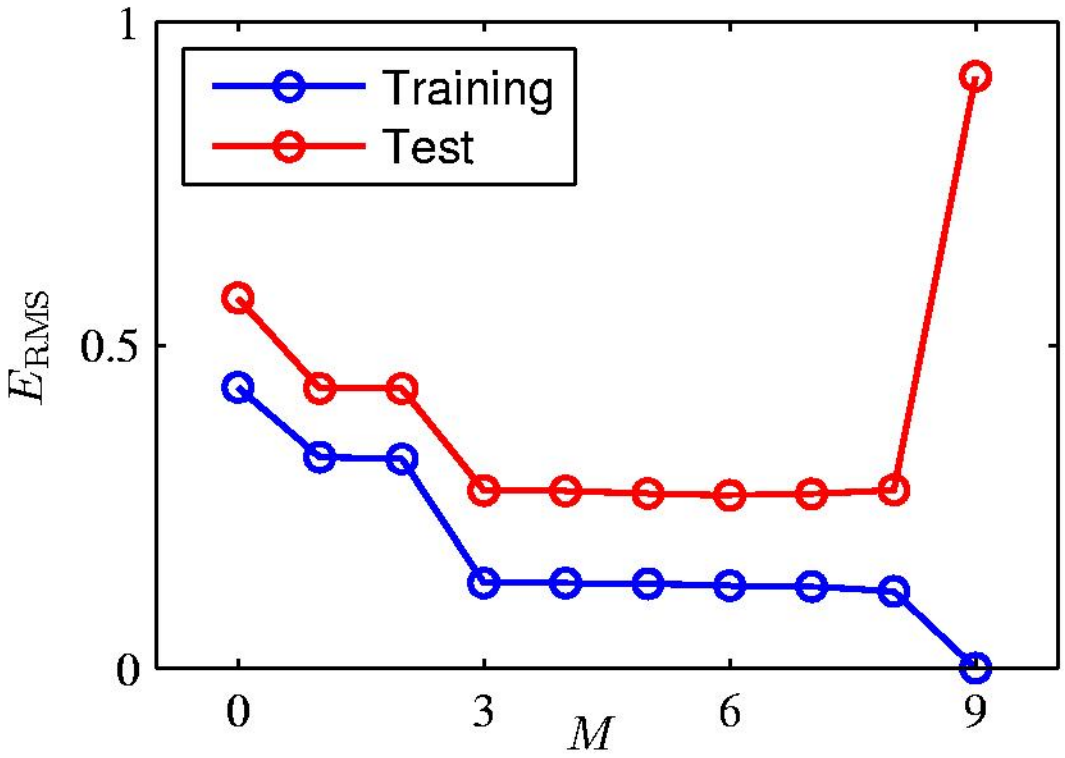
3rd Order Polynomial



9th Order Polynomial



Over-fitting



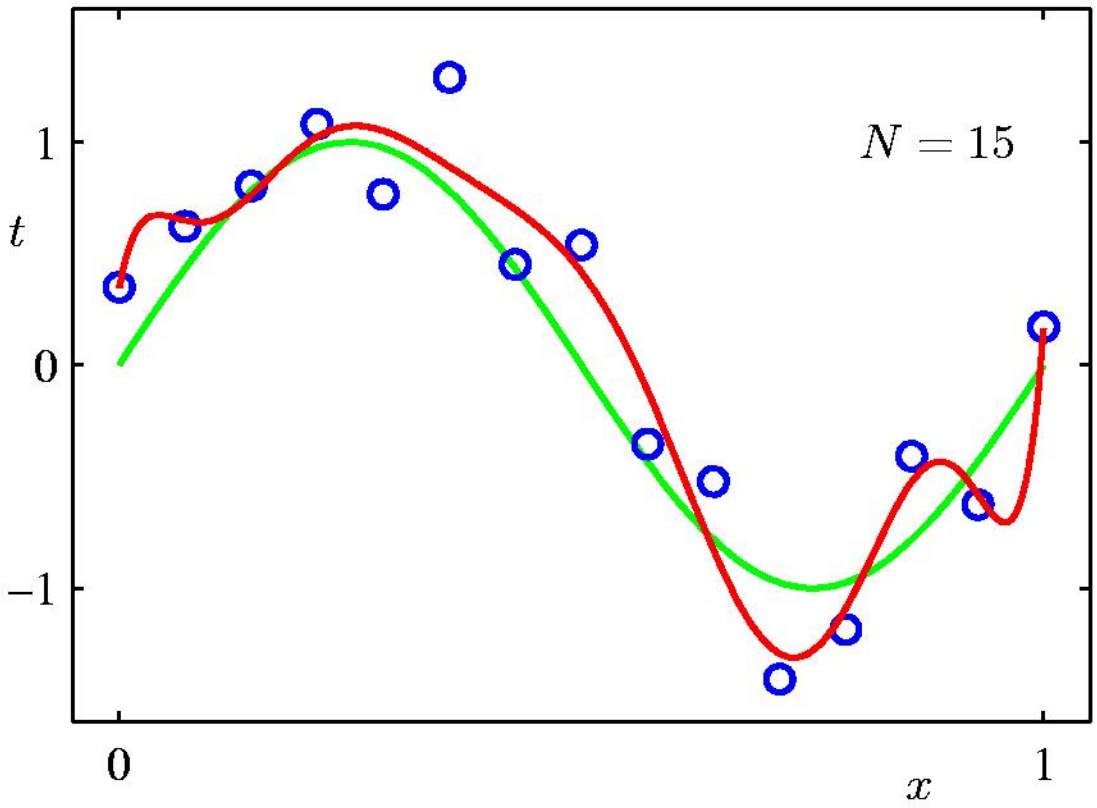
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

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

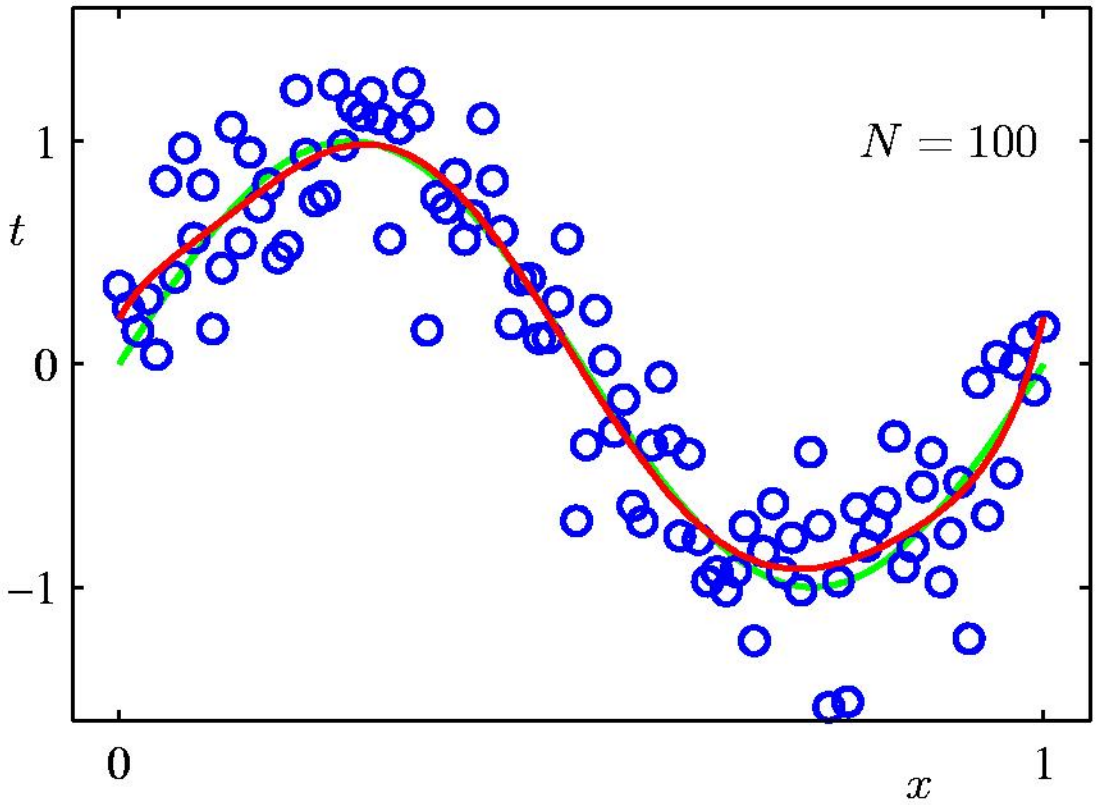
Data Set Size: $N = 15$

9th Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial



Regularized Least Squares

- Consider the error function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term

- With the sum-of-squares error function and a quadratic regularizer, we get

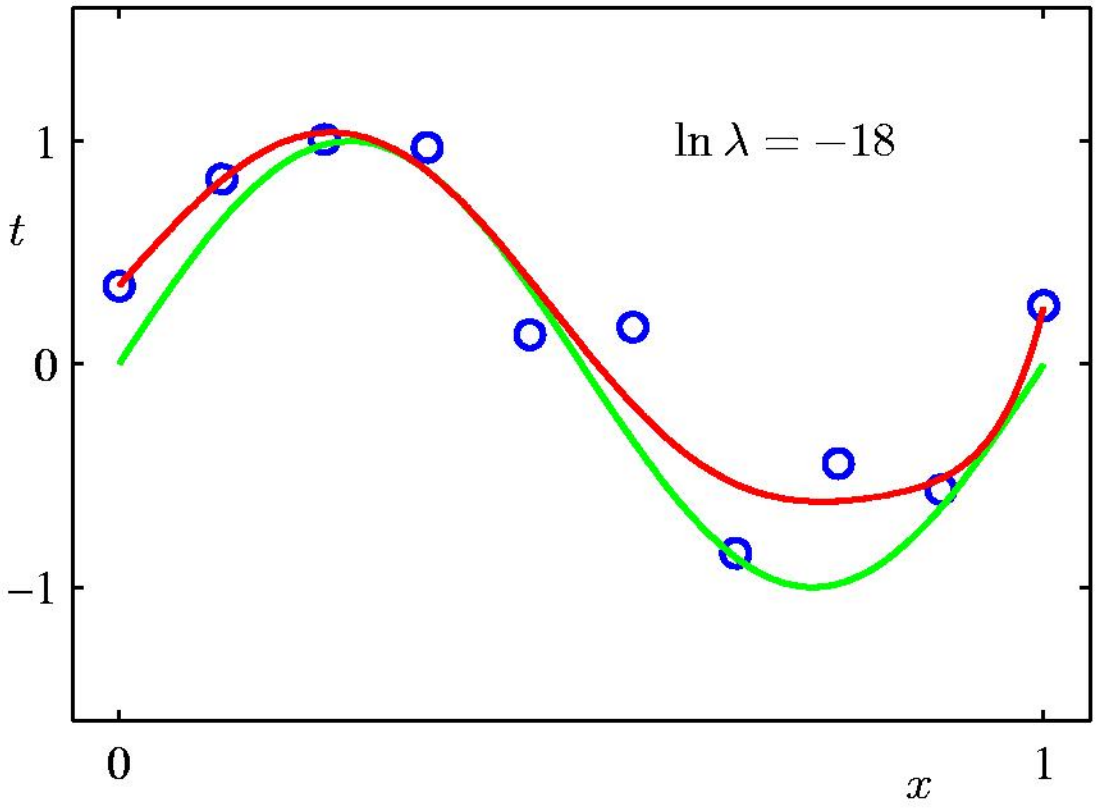
$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

- which is minimized by

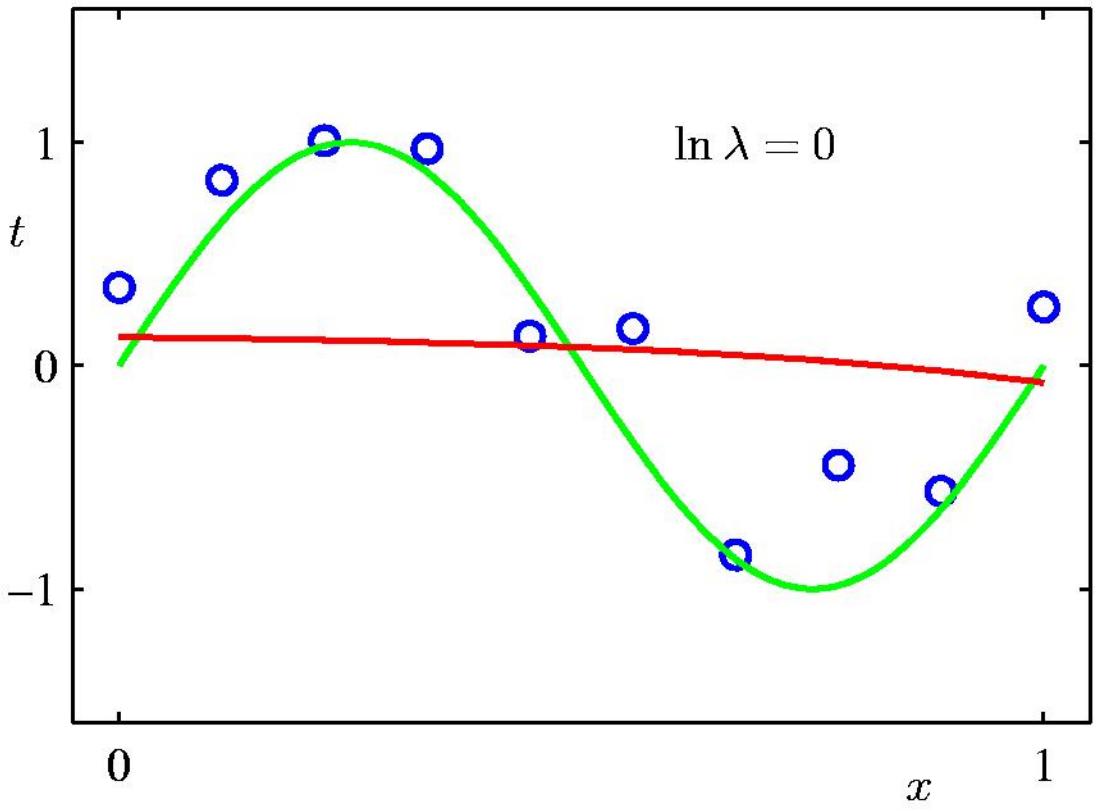
$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$$

- The matrix above is *always* invertible. Why?
- What is the probabilistic interpretation of this regularizer?

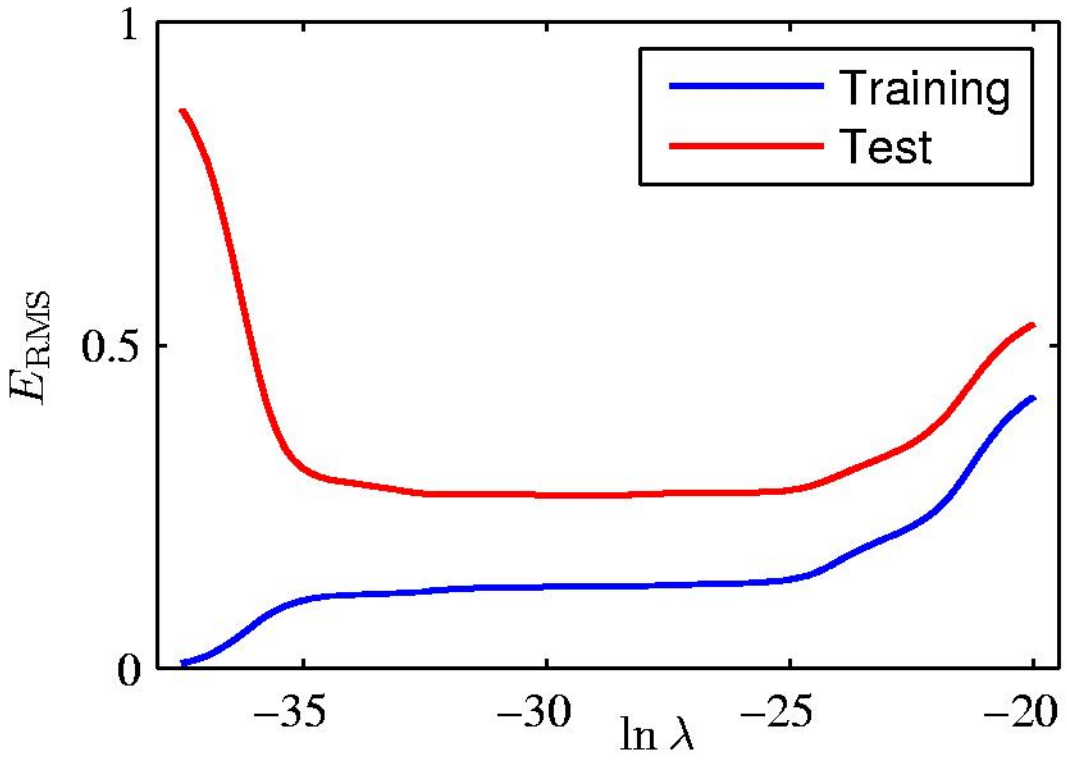
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



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

Bayesian Linear Regression

- Define a conjugate prior over w

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).$$

- Combining this with the likelihood function and using results for marginal and conditional Gaussian distributions, gives the posterior

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

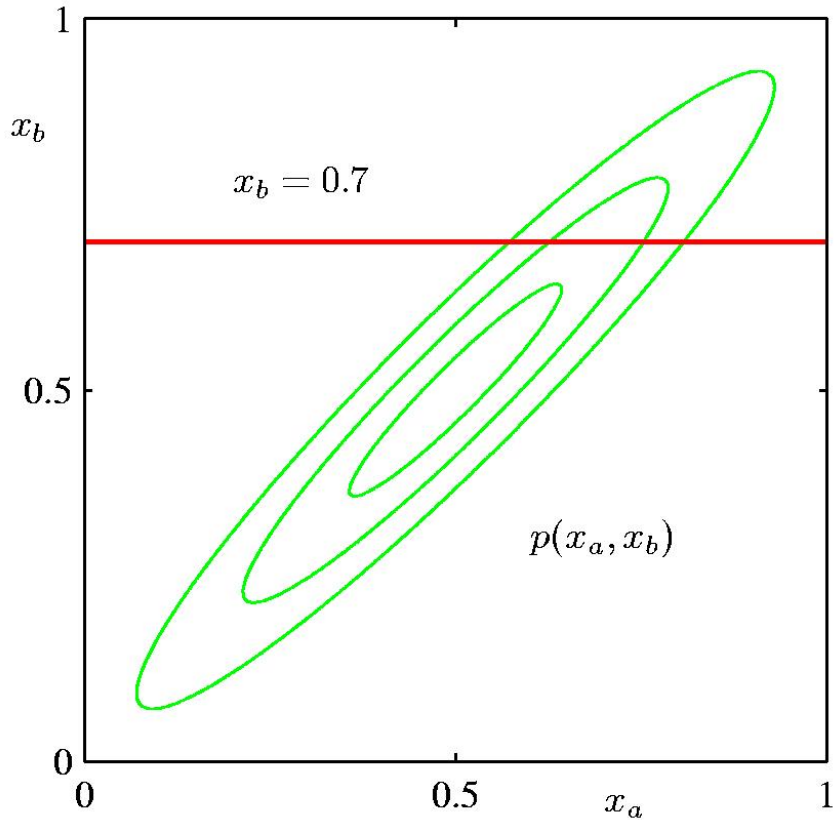
$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}. \end{aligned}$$

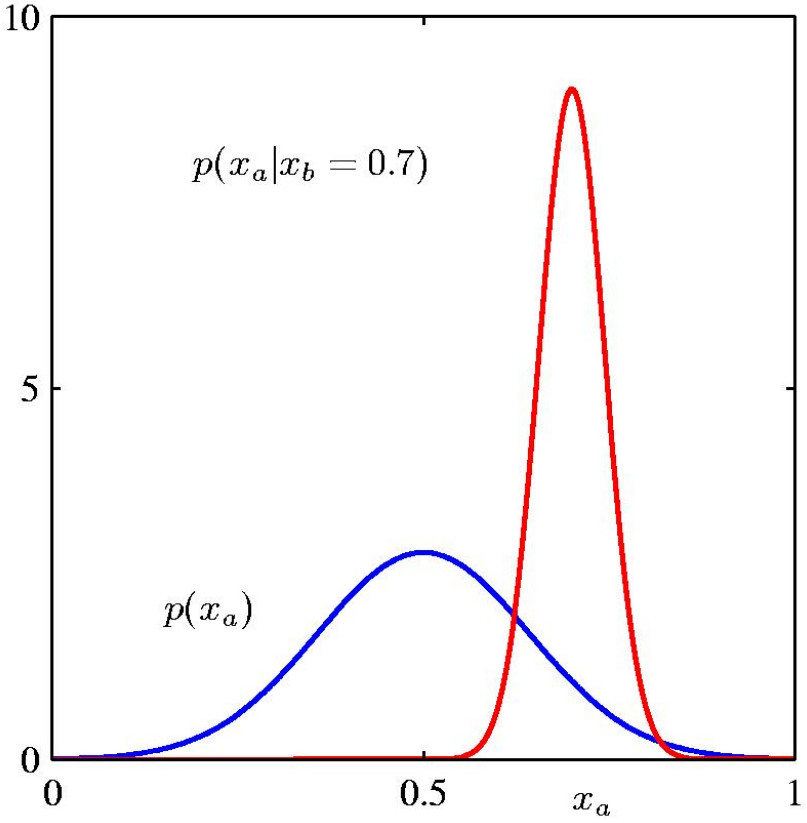
- A common choice for the prior is:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}. \end{aligned}$$

Gaussian Conditionals & Marginals





$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$p(x_1) = \mathcal{N}(x_1|\mu_1, \sigma_1^2)$$

$$p(x_1|x_2) = \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$$

Linear Gaussian Systems

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \qquad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_y)$$

Marginal Likelihood:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}_x + \mathbf{b}, \boldsymbol{\Sigma}_y + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$

Posterior Distribution:

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y})$$

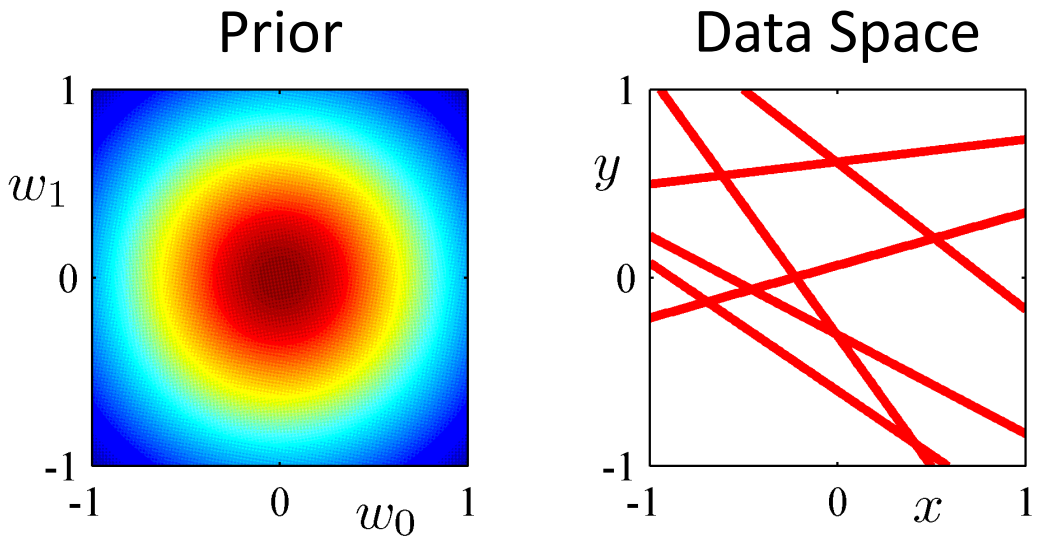
$$\boldsymbol{\Sigma}_{x|y}^{-1} = \boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A}$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}[\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x]$$

Board: Specialization to linear regression model

Bayesian Regression Example

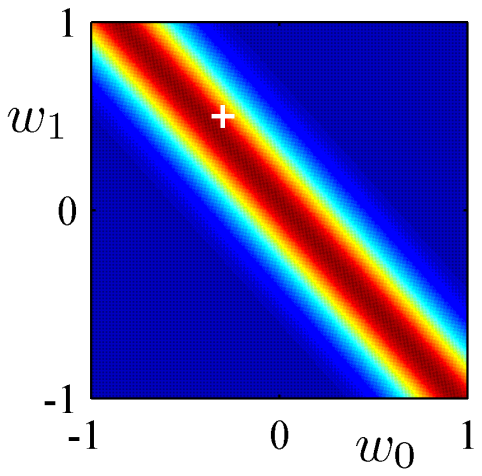
0 data points observed



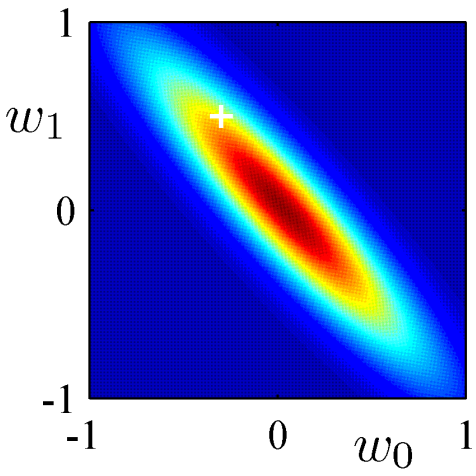
Bayesian Regression Example

1 data point observed

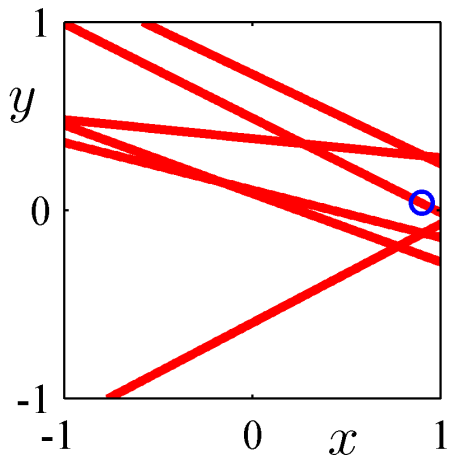
Likelihood



Posterior



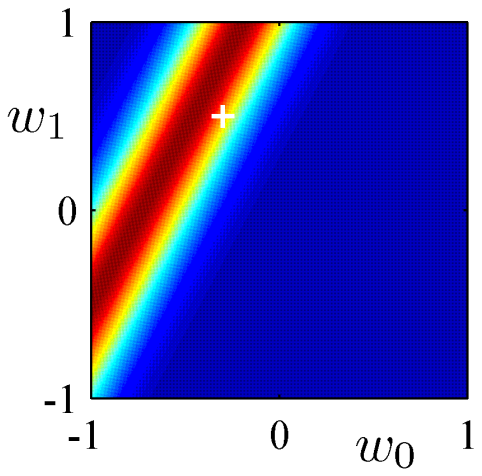
Data Space



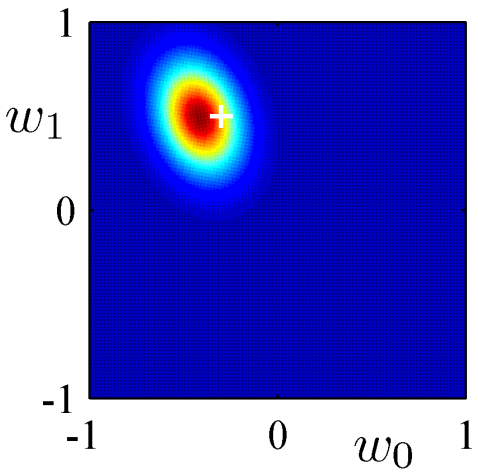
Bayesian Regression Example

2 data points observed

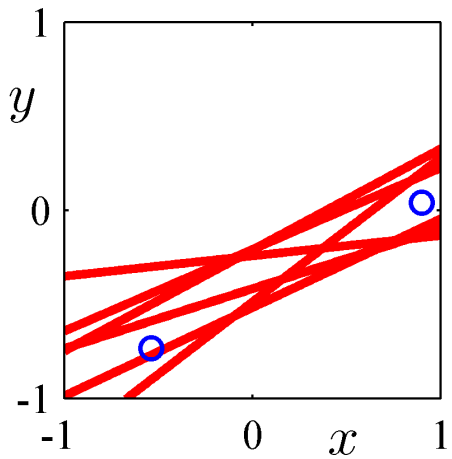
Likelihood



Posterior



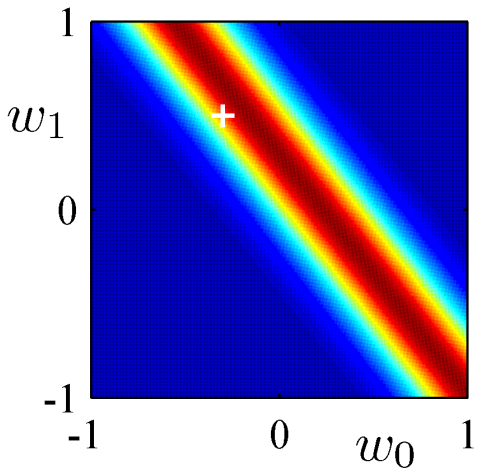
Data Space



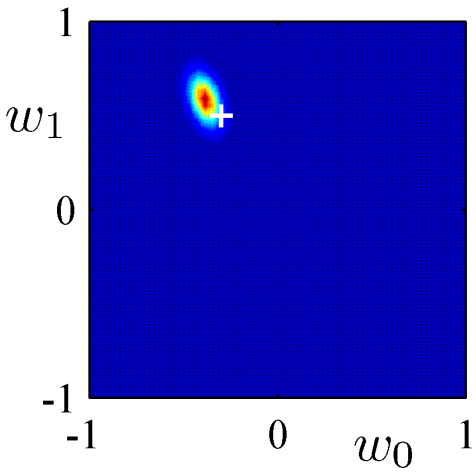
Bayesian Regression Example

20 data points observed

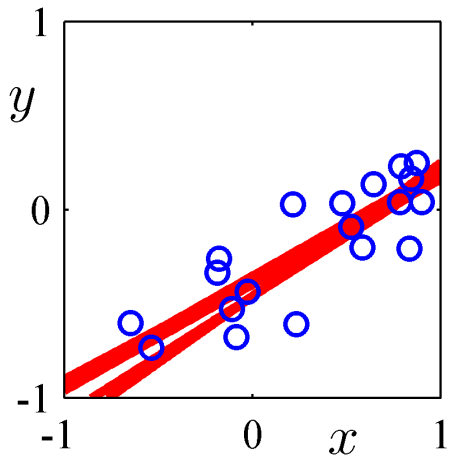
Likelihood



Posterior



Data Space



Predictive Distribution (1)

- Predict t for new values of x by integrating over w:

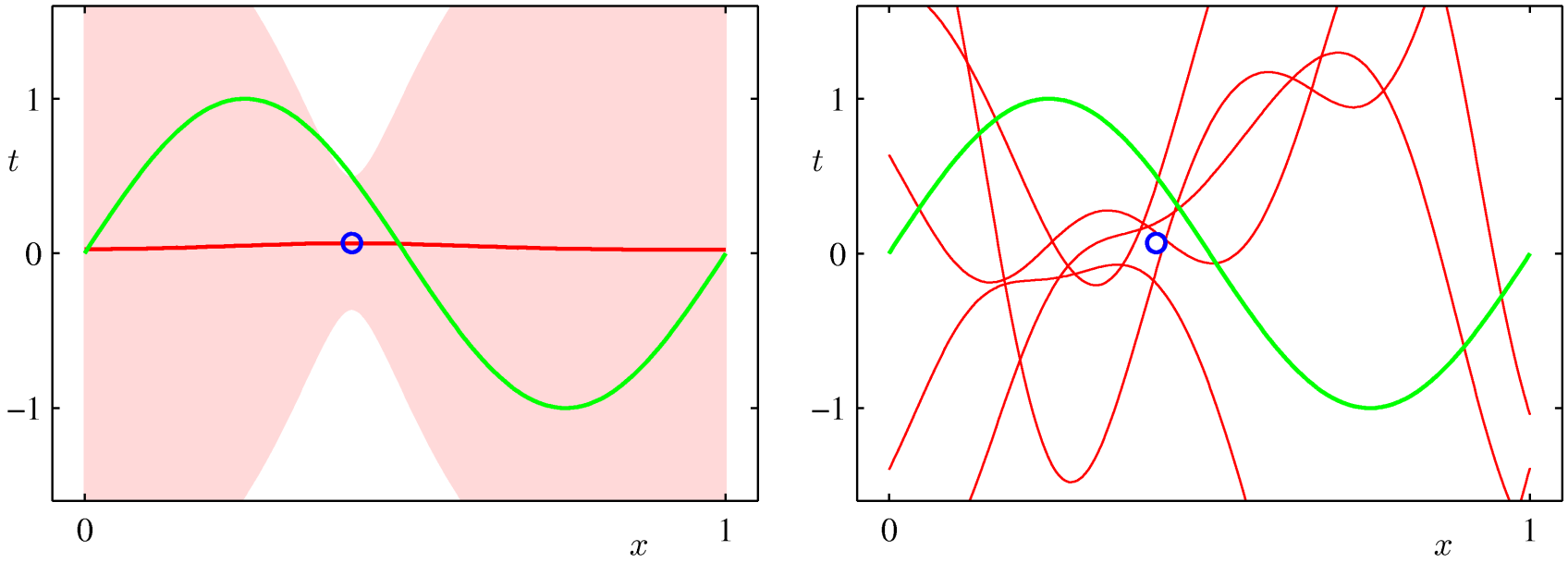
$$
\begin{aligned}
p(t|\mathbf{t}, \alpha, \beta) &= \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta)\, \mathrm{d}\mathbf{w} \\
&= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))
\end{aligned}
$$

- where

$$
\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}).
$$

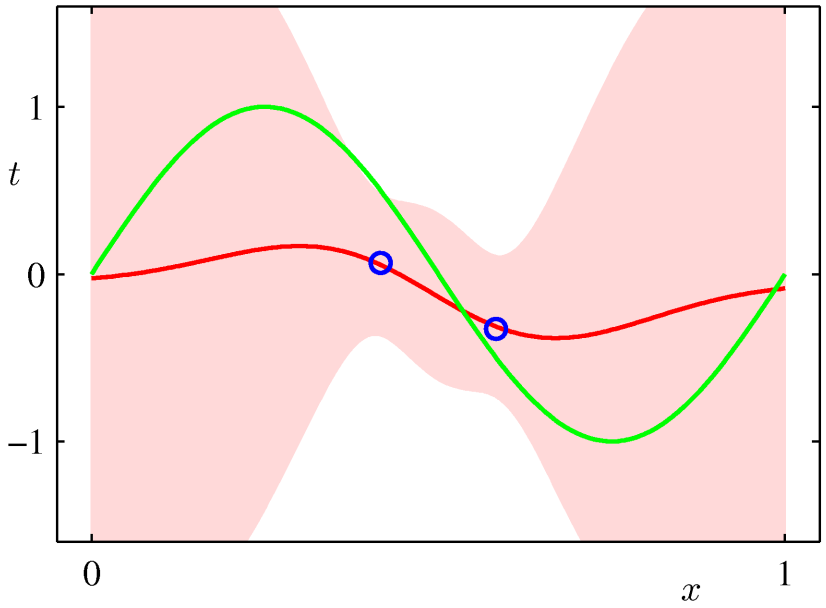
Predictive Distribution (2)

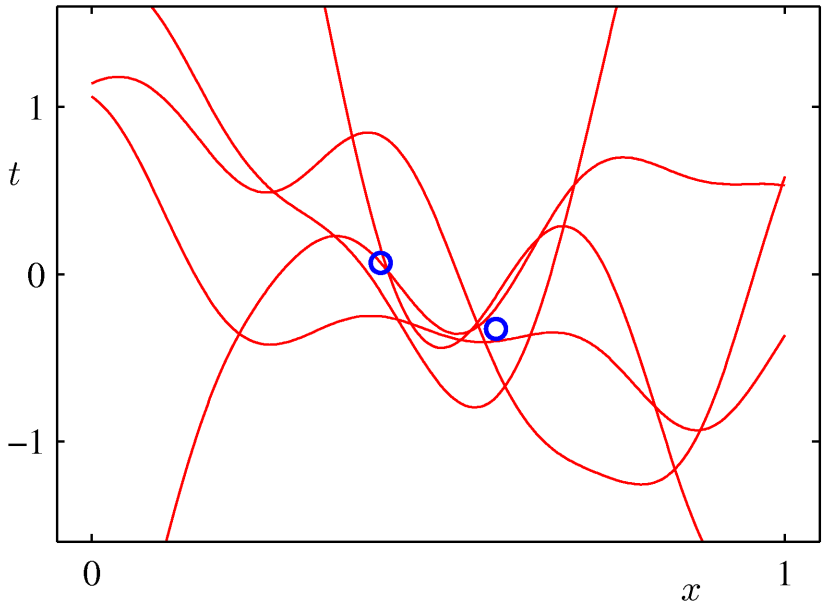
- Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point



Predictive Distribution (3)

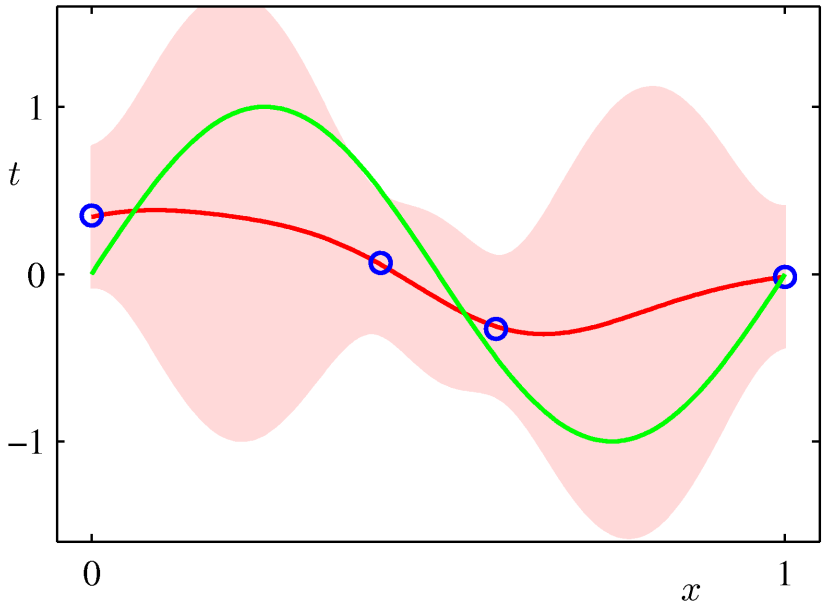
- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points

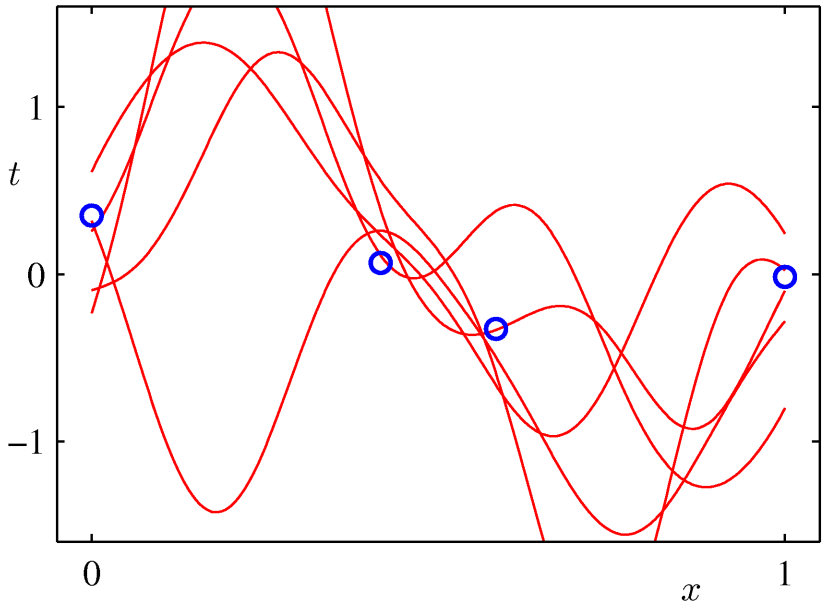




Predictive Distribution (4)

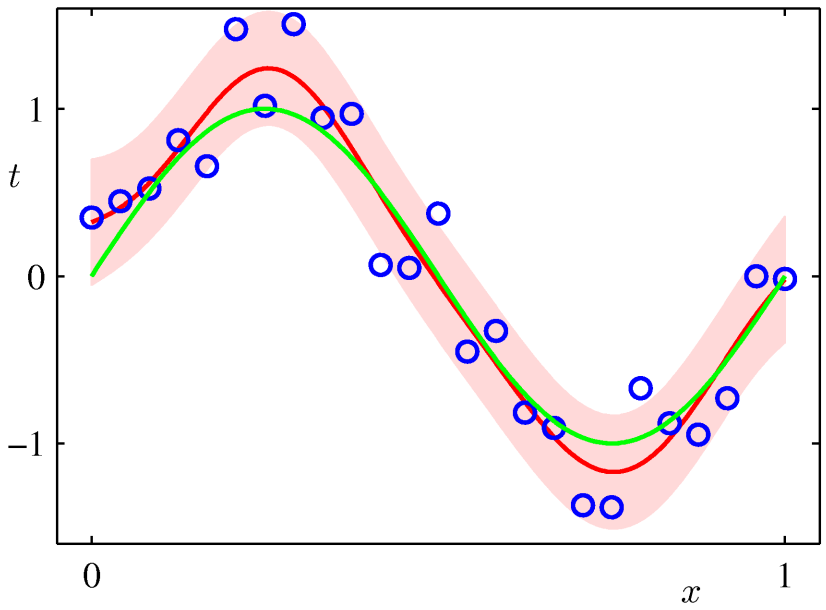
- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points

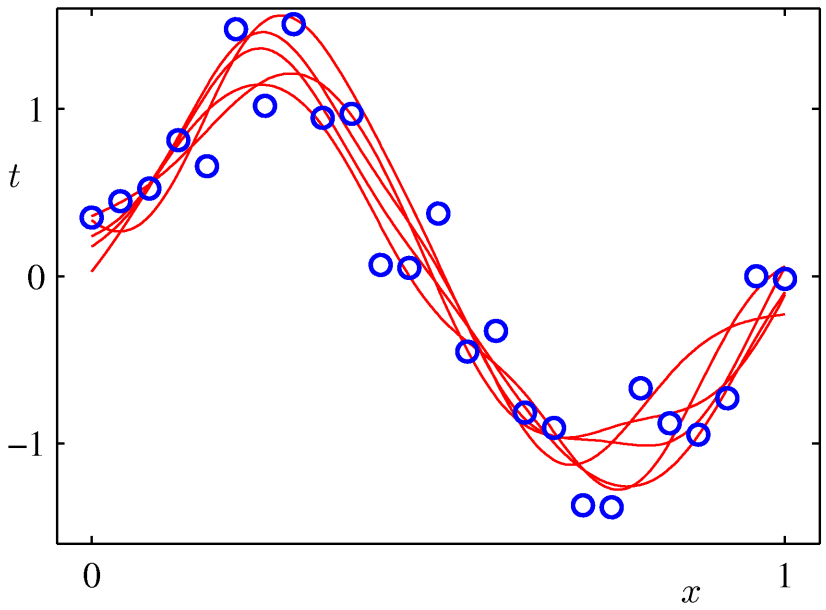




Predictive Distribution (5)

- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points





Estimation vs. Predictive Distributions