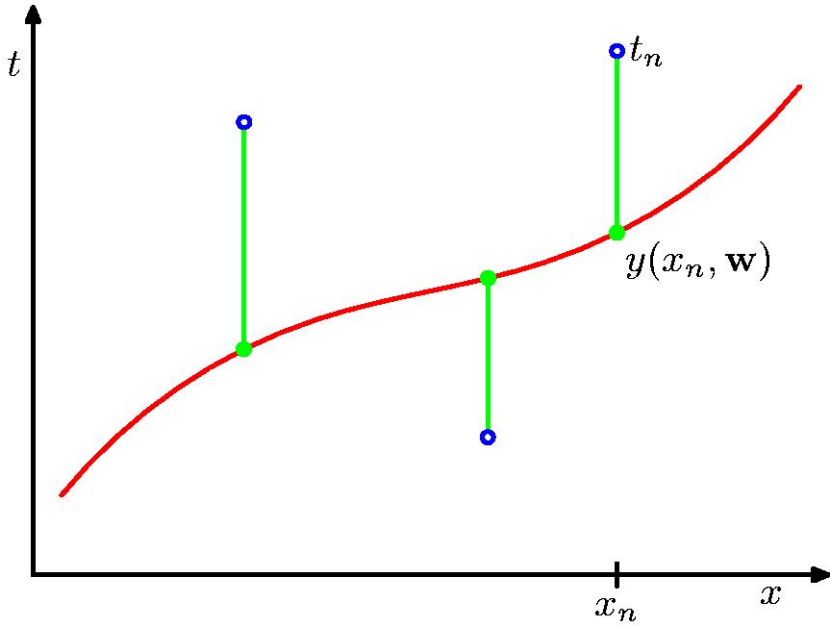
CSE 847: Statistical Machine Learning

Linear Models for Regression II

Jiayu Zhou
Computer Science & Engineering
Michigan State University

Polynomial curve fitting

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:

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

Maximum likelihood and least squares (1)

- We assume that the target variable t is given by a deterministic function y(x,w) with additive Gaussian noise so that

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

where ϵ is a zero mean Gaussian random variable with precision (inverse variance) β. Thus we can write

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

- If we assume a squared loss function, then the optimal prediction, for a new value of **x**, will be given by the conditional mean of the target variable.

Maximum likelihood and least squares (2)

- Consider a data set of inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with corresponding target values $t_1, \ldots, t_N$. The likelihood function is:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

- Log likelihood:

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}) \end{aligned}$$

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

Maximum likelihood and least squares (3)

- Maximize the likelihood function w.r.t. w

$$\frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$$

is equivalent to minimizing the squared error term:

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

- Maximization of the likelihood function under a conditional Gaussian noise distribution for a linear model is equivalent to minimizing a sum-of-squares error function

Maximum likelihood and least squares (4)

- The gradient of the log likelihood function takes the form:

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}$$

- Setting this gradient to zero gives

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \right)$$

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

Here $\boldsymbol{\Phi}$ is called the design matrix: $\Phi_{nj} = \phi_j(\mathbf{x}_n)$

Maximum likelihood and least squares (5)

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

$$\boldsymbol{\Phi}^{\dagger} \equiv \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}$$

Moore-Penrose pseudo-inverse

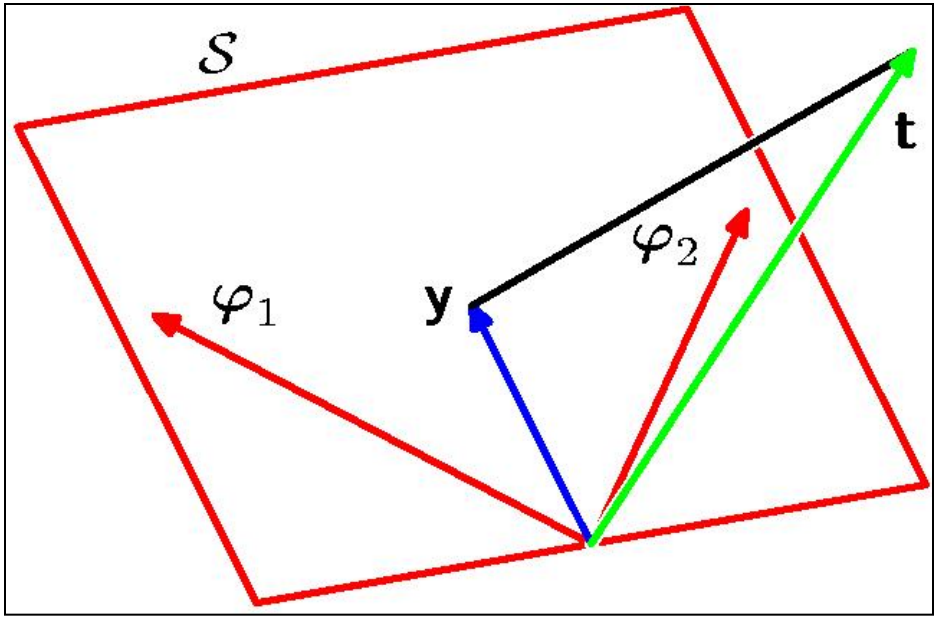
Here **ϕ** is called the design matrix:

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

Geometry of least squares

- The solution corresponds to the orthogonal projection of **t** onto the subspace *S*, which is spanned by the basis functions $\phi_j(\mathbf{x})$.

the j-th column of **ϕ**



Regularized least squares (1)

- Add a regularization term to an error function in order to control over-fitting:

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \qquad \mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$$

- Regularization allows complex models to be trained on data sets of limited size without severe over-fitting, essentially by limiting the effective model complexity.

Regularized least squares (2)

- Introduce a prior probability distribution over the model parameters w. Assume the noise precision β is known.
- The likelihood function $p(\mathbf{t}|\mathbf{w})$ is the exponential of a quadratic function of $\mathbf{w}$:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

- The corresponding conjugate prior is $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$
- The posterior is: $p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$

where

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

Regularized least squares (3)

- We consider a zero-mean isotropic Gaussian governed by a single precision parameter α:

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

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

- The log of the posterior distribution is given by:

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} + \text{const}$$

Regularized least squares vs MAP

- Maximization of this posterior distribution with respect to **w** is therefore equivalent to the minimization of the sum-of-squares error function with the addition of a quadratic regularization term.

- Introducing a prior on **w** is equivalent to adding a regularization term in least square problems.

Maximum likelihood ⟷ Least squares

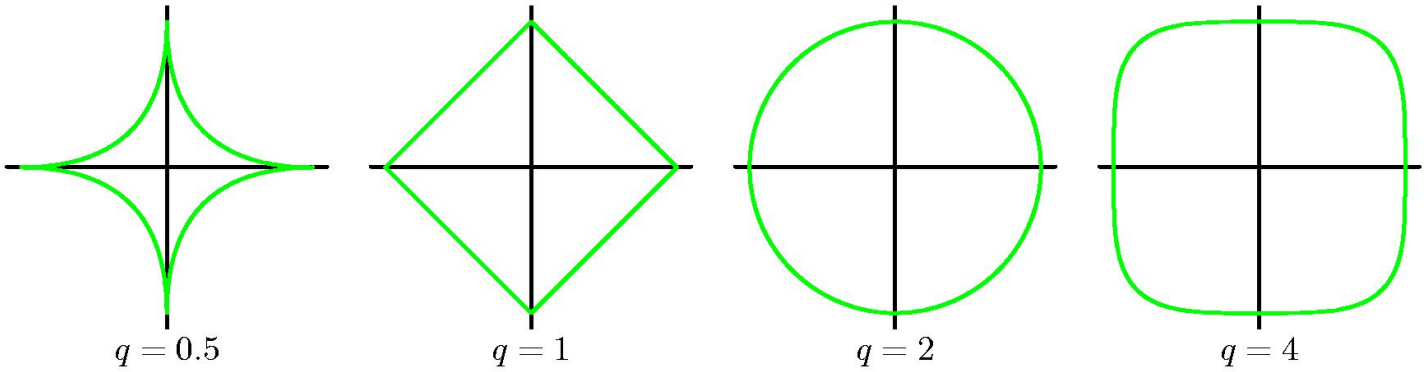
Maximum posterior ⟷ Regularized least squares

Bridge Regression

- A more general regularizer is sometimes used:

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

  - Ridge regression: q=2
  - Lasso: q=1



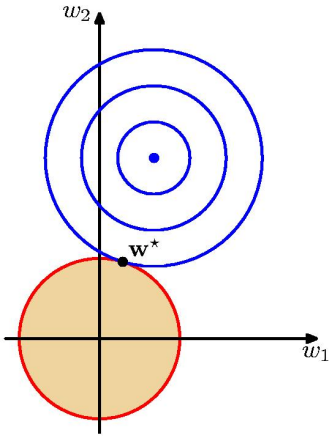
Why does L_1 Induce Sparsity? (1)

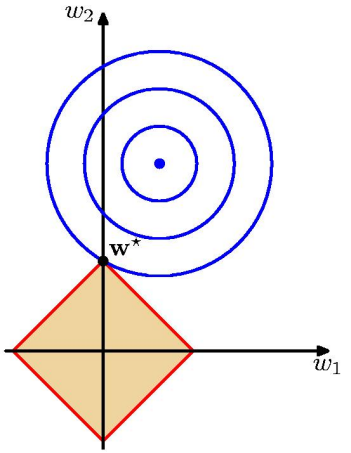
- The case of q = 1 is know as the lasso . It has the property that if λ is sufficiently large, some of the coefficients are driven to zero.
- Minimizing

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

is equivalent to minimizing the first term subject to the constraint

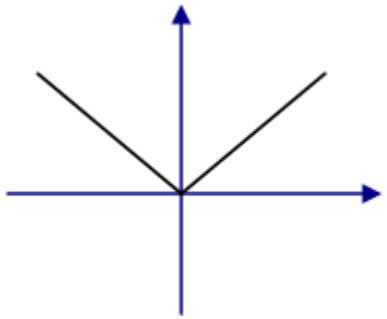
$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta$$



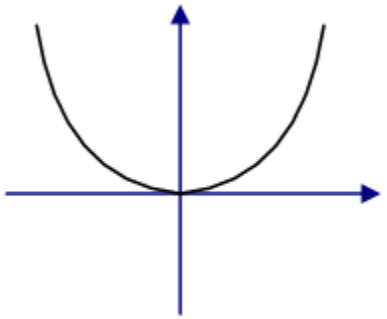


Why does L_1 Induce Sparsity? (2)

Analysis in 1D (comparison with L_2)



$0.5\times(x-v)^2 + \lambda|x|$



$0.5\times(x-v)^2 + \lambda x^2$

If $v \geq \lambda$, $x=v-\lambda$
If $v \leq -\lambda$, $x=v+\lambda$
Else, $x=0$

$x=v/(1+2\lambda)$

Nondifferentiable at 0

Differentiable at 0

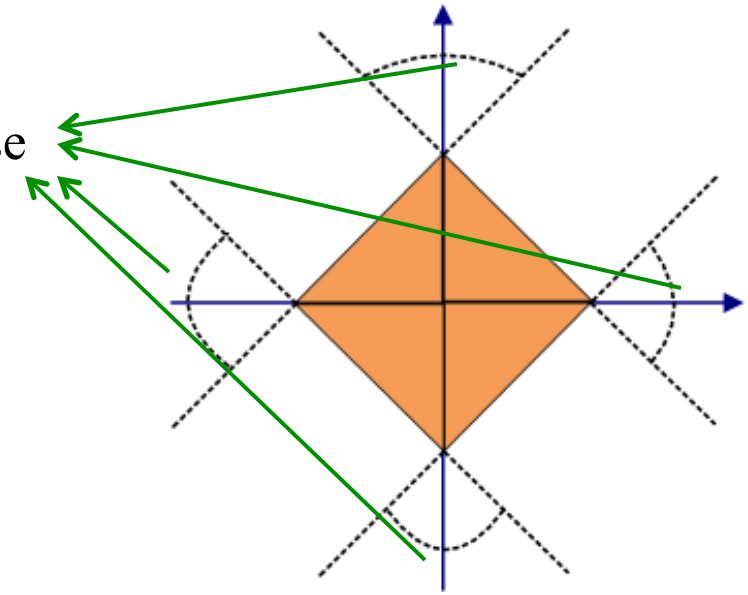
Why does L_1 Induce Sparsity? (3)

Understanding from the projection

min loss(x)
s.t. $||x||_1 \leq 1$

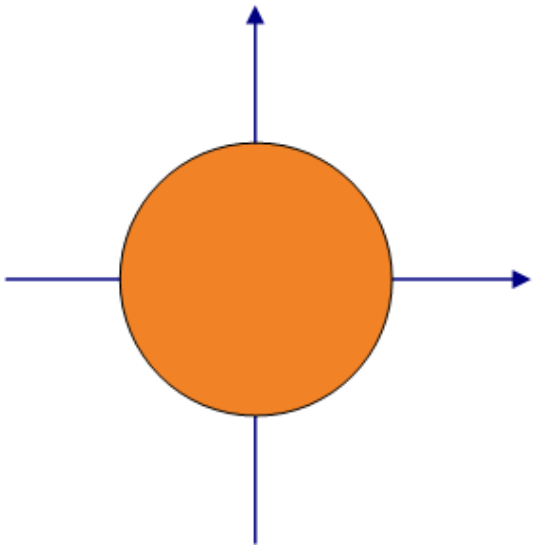
min $0.5||x-v||^2$
s.t. $||x||_1 \leq 1$

Sparse



min loss(x)
s.t. $||x||_2 \leq 1$

min $0.5||x-v||^2$
s.t. $||x||_2 \leq 1$



Regularized least squares (4)

- If data points arrive sequentially, then the posterior distribution at any stage acts as the prior distribution for the subsequent data point, such that the new posterior distribution is updated sequentially.

$$p(w \mid \{t_1, \mathrm{L}\ , t_{n+1}\}) \propto p(\{t_1, \mathrm{L}\ , t_{n+1}\} \mid w) p(w)$$

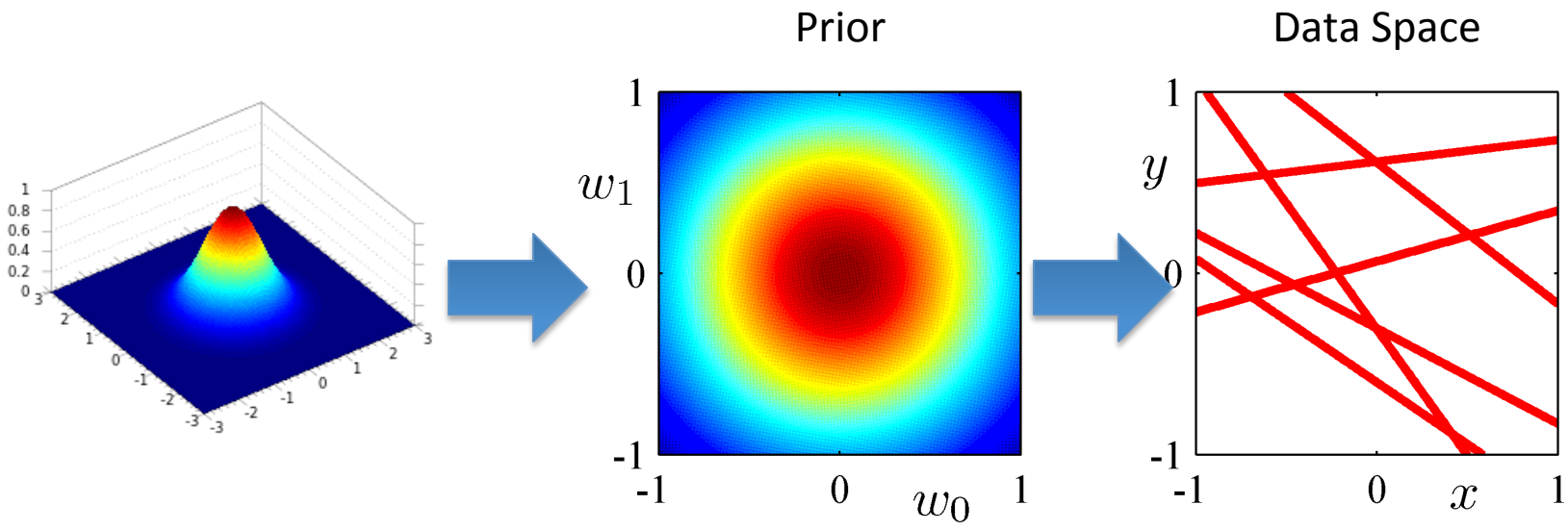
$$= \prod_{i=1}^{n+1} p(t_i \mid w) p(w)$$

$$= p(t_{n+1} \mid w) \left[\prod_{i=1}^{n} p(t_i \mid w) p(w) \right]$$

likelihood prior

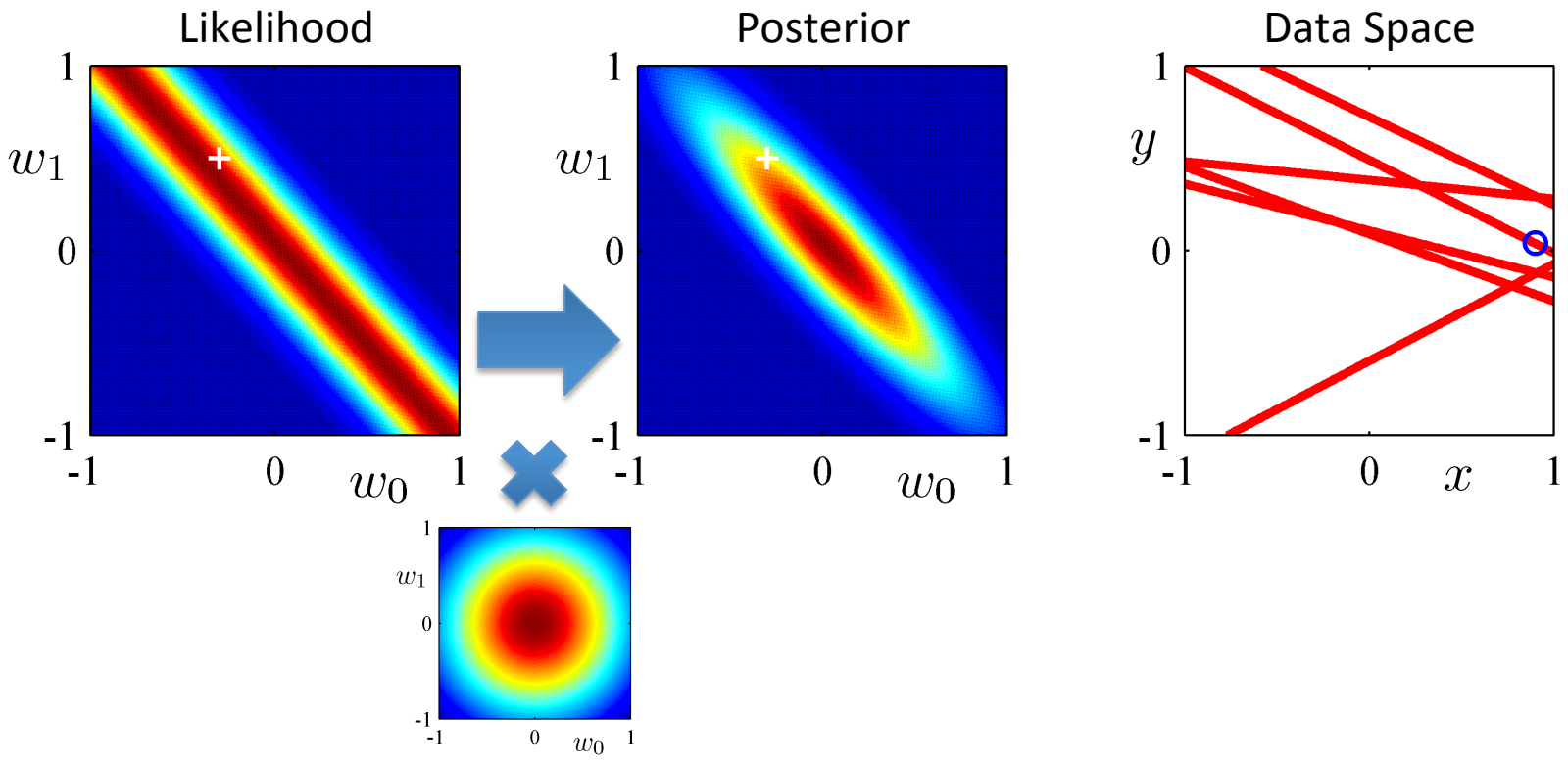
An Illustrative Example (1)

0 data points observed



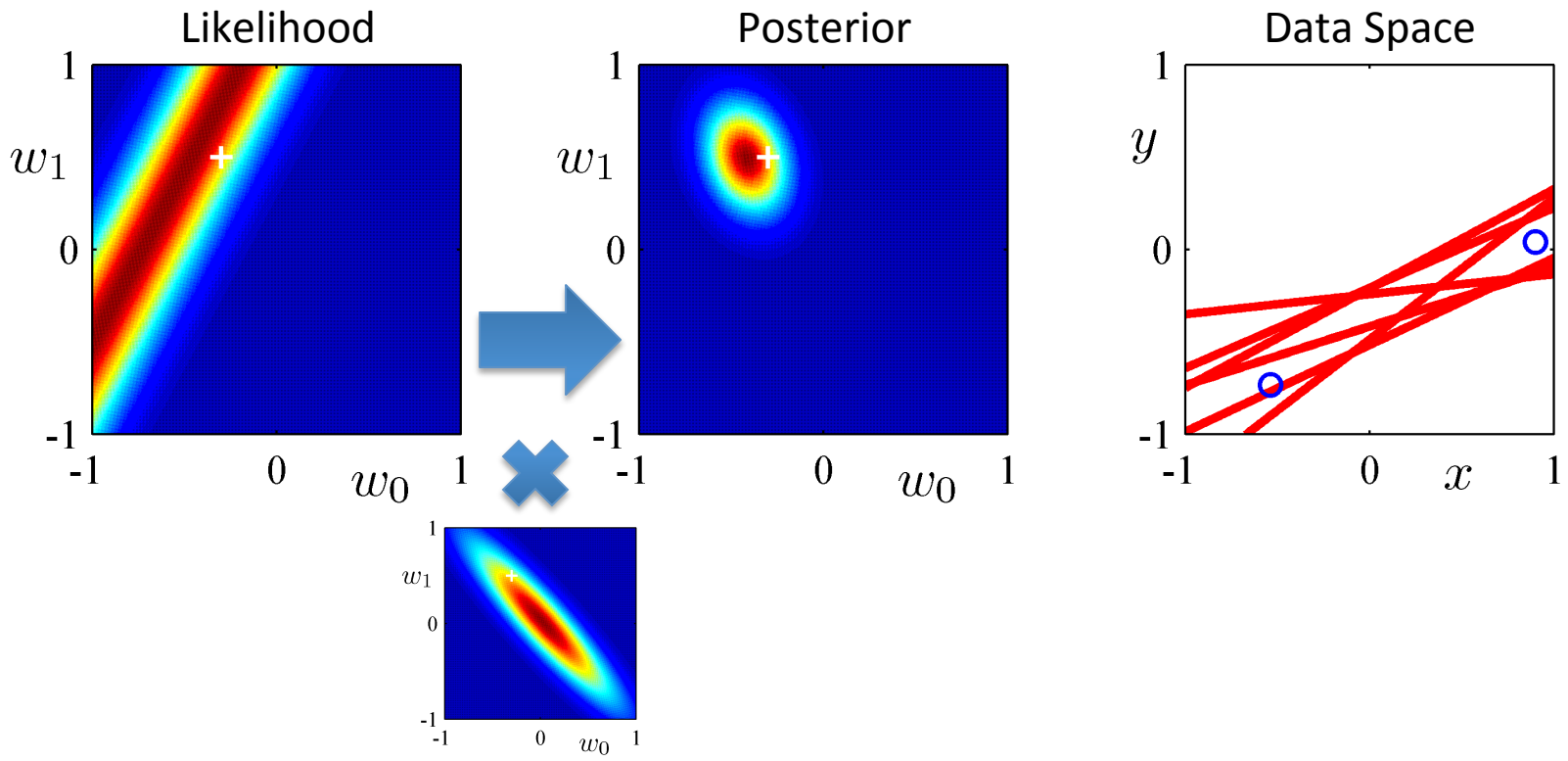
An Illustrative Example (2)

1 data point observed



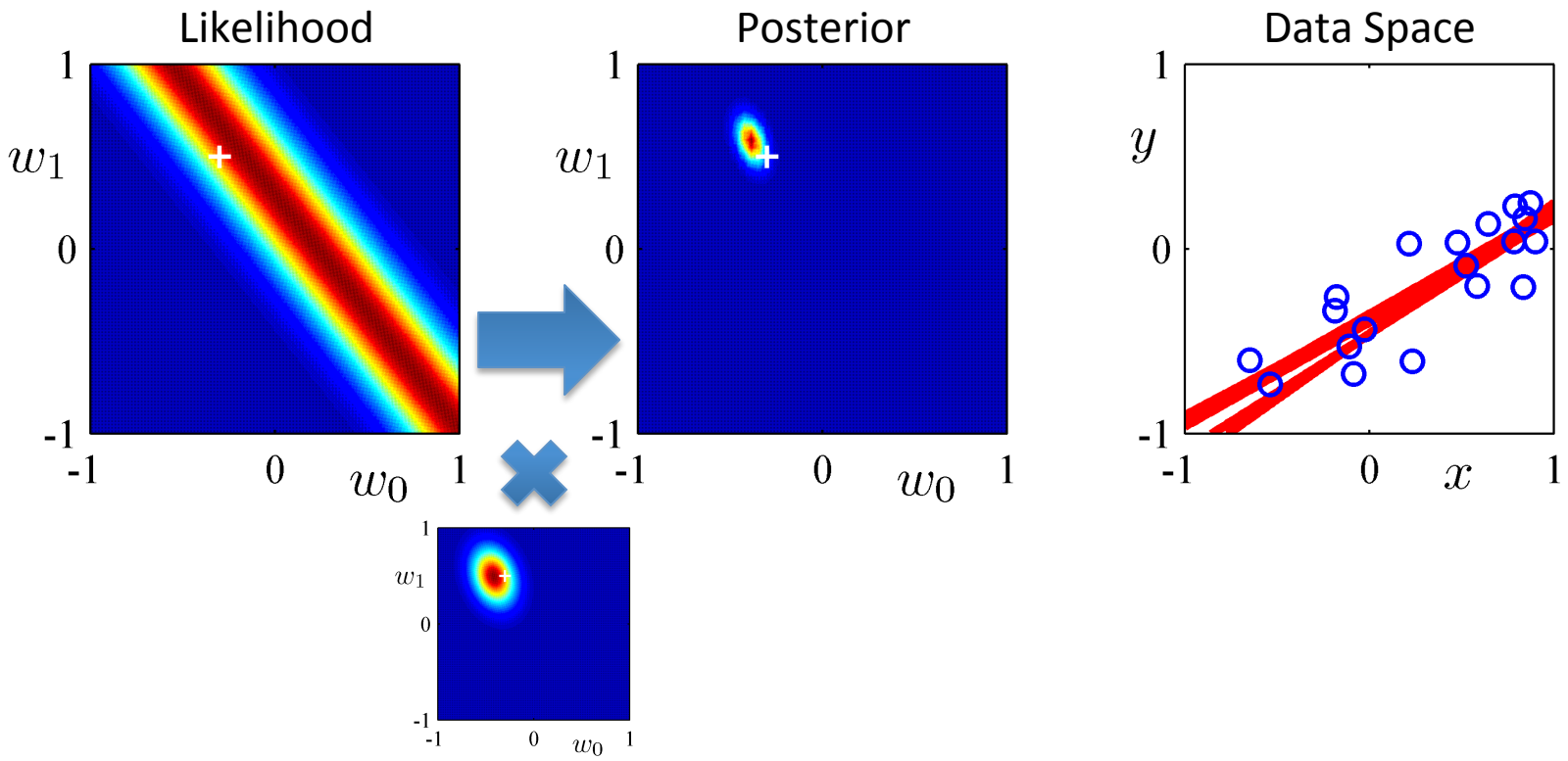
An Illustrative Example (3)

2 data points observed



An Illustrative Example (4)

20 data points observed



Predictive Distribution (1)

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

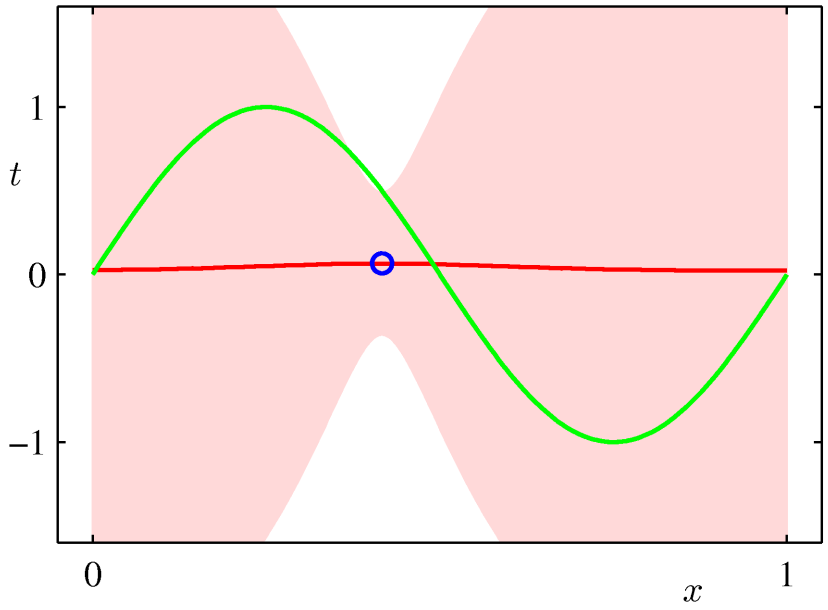
$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w}$$
$$= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))$$

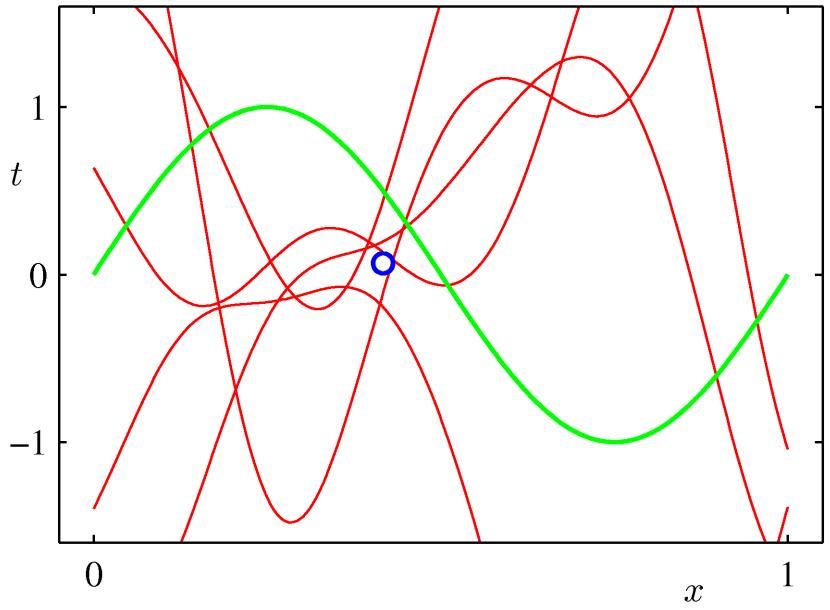
where

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

Predictive Distribution (2)

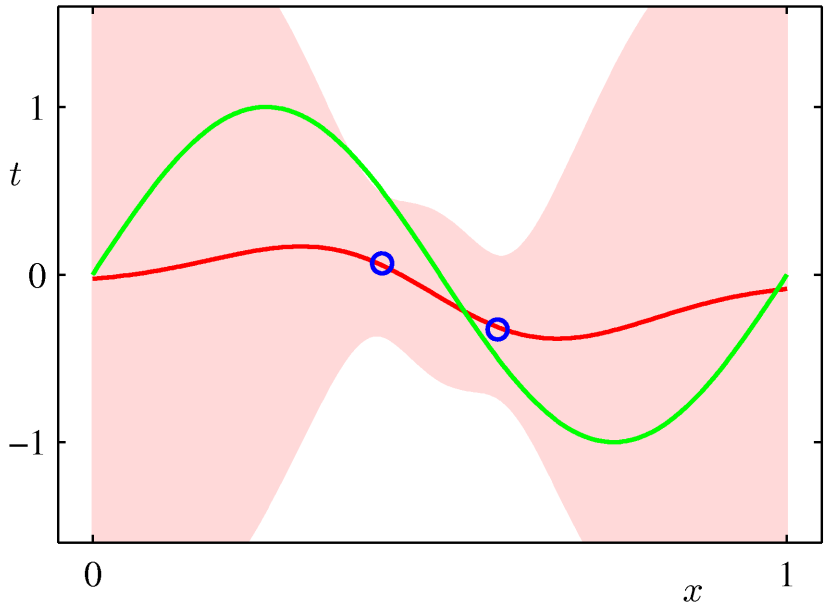
Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point

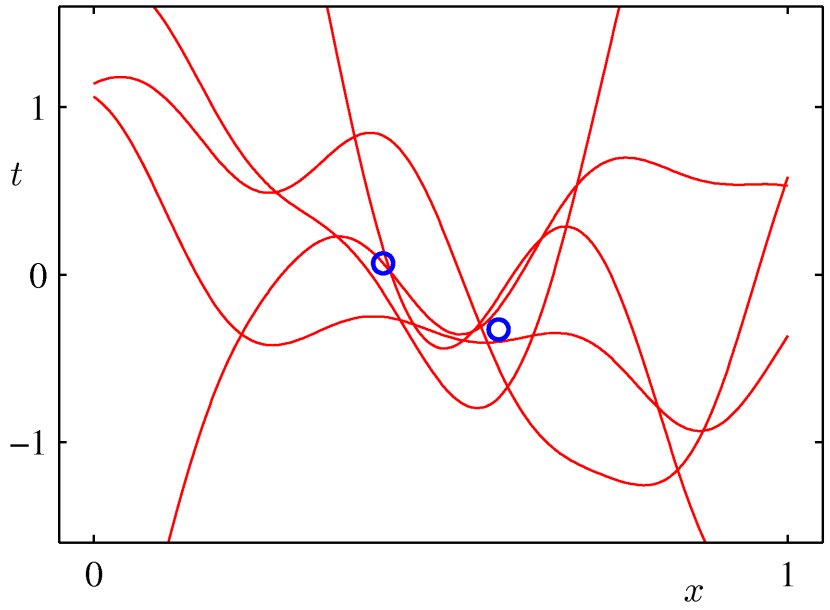




Predictive Distribution (3)

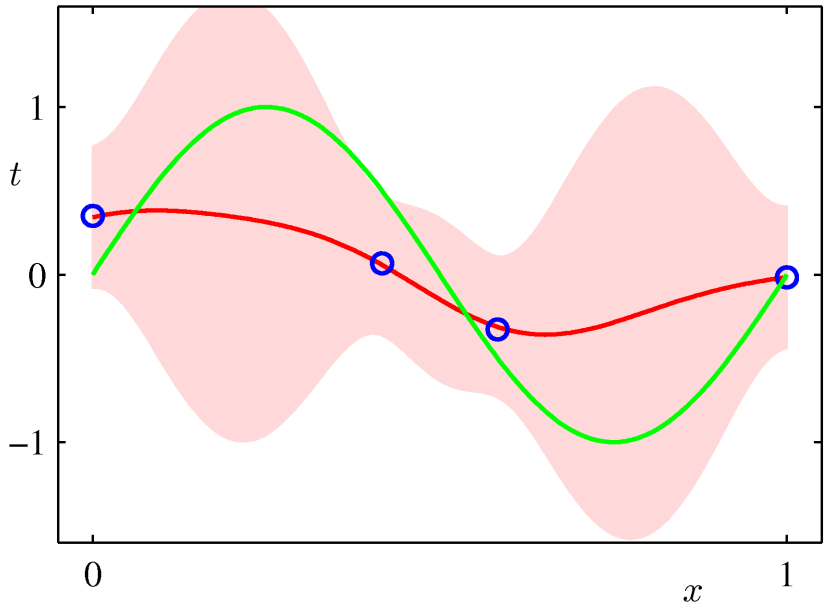
Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points

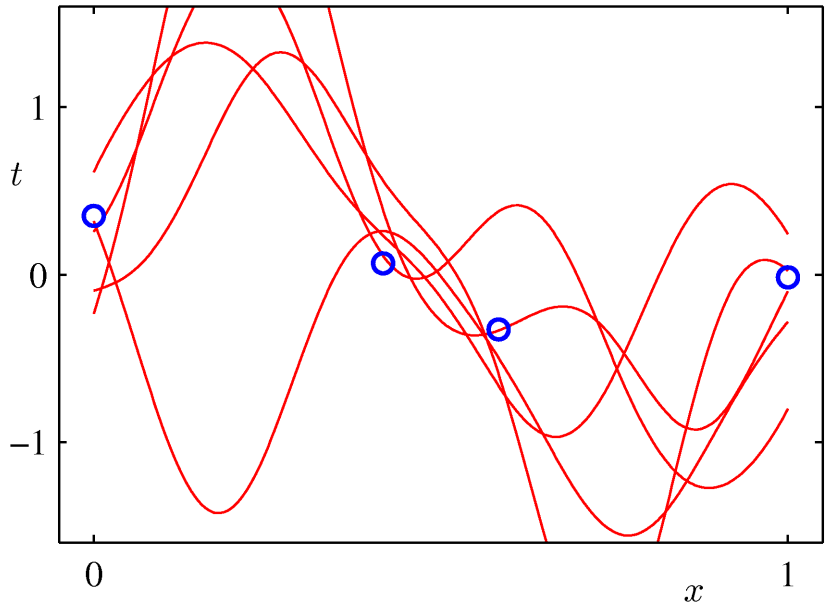




Predictive Distribution (4)

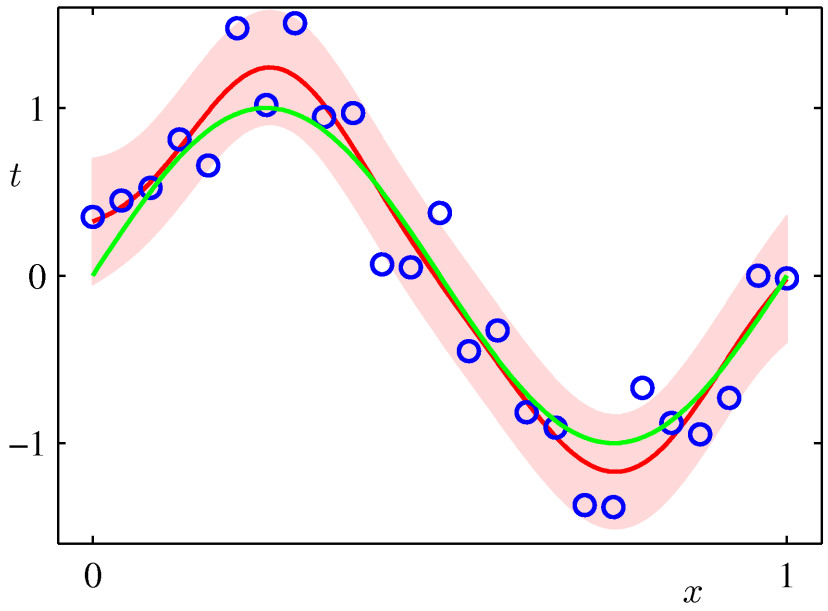
Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points

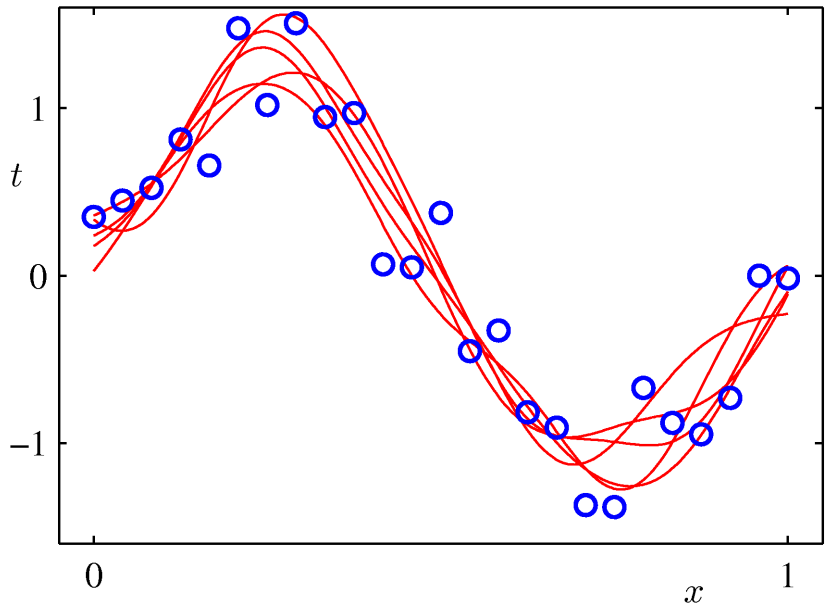




Predictive Distribution (5)

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





Next Class

- Topic
  - Linear Models for Classification

- Reading
  - Sections 4.1-4.3