PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 3: LINEAR MODELS FOR REGRESSION

Learning Objectives

1、How to achieve linear regression using basis functions?

2、What are the relationships between maximum likelihood and least squares, between maximum a posterior and regularization, and among expected loss, bias, variance, and noise?

3、What are the common regularization methods for regression?

4、How to achieve Bayesian linear regression?

5、What is the kernel for regression?

6、How to choose the model complexity?

7、What are the evidence approximation and maximization?

Outlines

- Linear Basis Function Models
- Maximum Likelihood and Least Squares
- Bias Variance Decomposition
- Bayesian Linear Regression
- Predictive Distribution
- Equivalent Kernel
- Bayesian Model Comparison
- Evidence Approximation and Maximization

Bayesian Linear Regression (1)

- Define a conjugate prior over $\mathbf{w}$

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

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

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

where

$$\boxed{\mathbf{w}_{\mathrm{MAP}}} \rightarrow \mathbf{S}_N{}^{-1}\mathbf{m}_N = \beta\Phi^T\mathbf{t} + \mathbf{S}_0{}^{-1}\mathbf{m}_0$$

$$\mathbf{S}_N{}^{-1} = \beta\Phi^T\Phi + \mathbf{S}_0{}^{-1}$$

Bayesian Linear Regression (2)

$$-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^T\ \mathbf{S}_N{}^{-1}(\mathbf{w}-\mathbf{m}_N)\ \propto\ -\frac{1}{2}(\mathbf{t}-\Phi\mathbf{w})^T\beta(\mathbf{t}-\Phi\mathbf{w})$$

$$-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^T\ \mathbf{S}_0{}^{-1}(\mathbf{w}-\mathbf{m}_0)$$

Quadratic terms of **w** are equal: (**w**^{T}****w**)

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

1st order terms of **w** are also equal: (**w**^{T} **)

Bayesian Linear Regression (3)

- A common choice for the prior is

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

for which

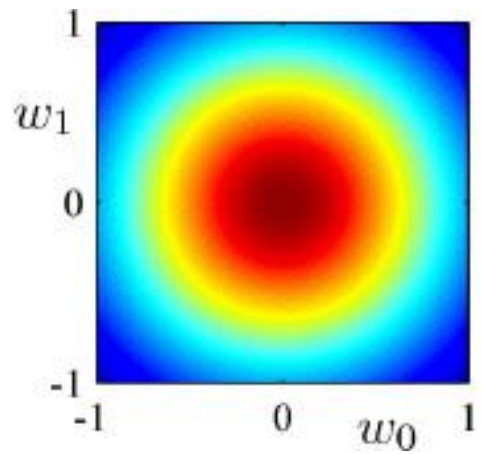
$$\begin{aligned} \boxed{\mathbf{w}_{\mathrm{MAP}}} \longrightarrow \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}. \end{aligned}$$

- Next we consider an example …

Bayesian Linear Regression (4)

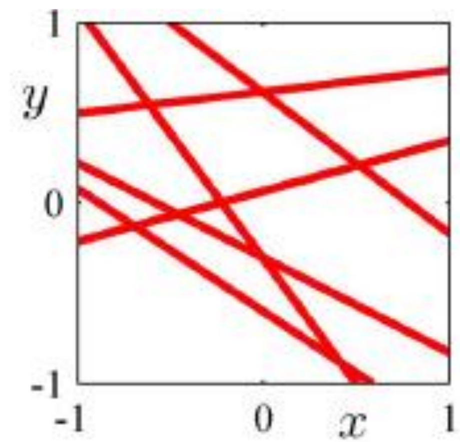
- 0 data points observed

Prior



$p(w_1, w_0)$

Data Space

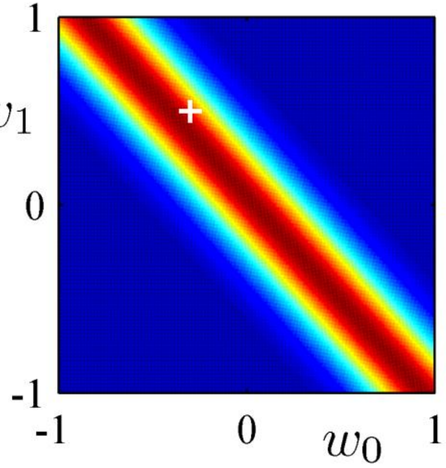


$y = w_1 x + w_0$ samples

Bayesian Linear Regression (5)

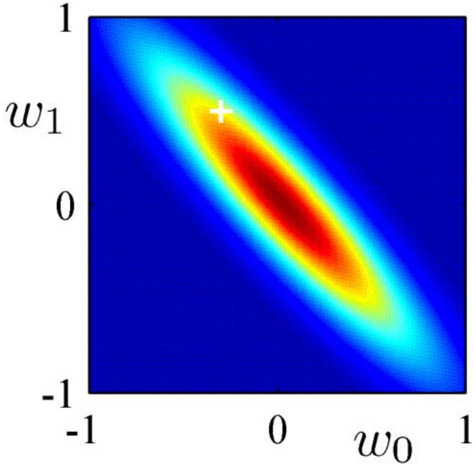
- 1 data point observed

Likelihood



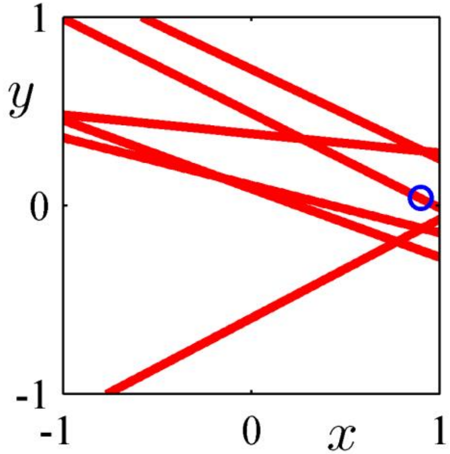
$p(\mathrm{t}|w_1, w_0)$

Posterior



$p(w_1, w_0|\mathrm{t})$

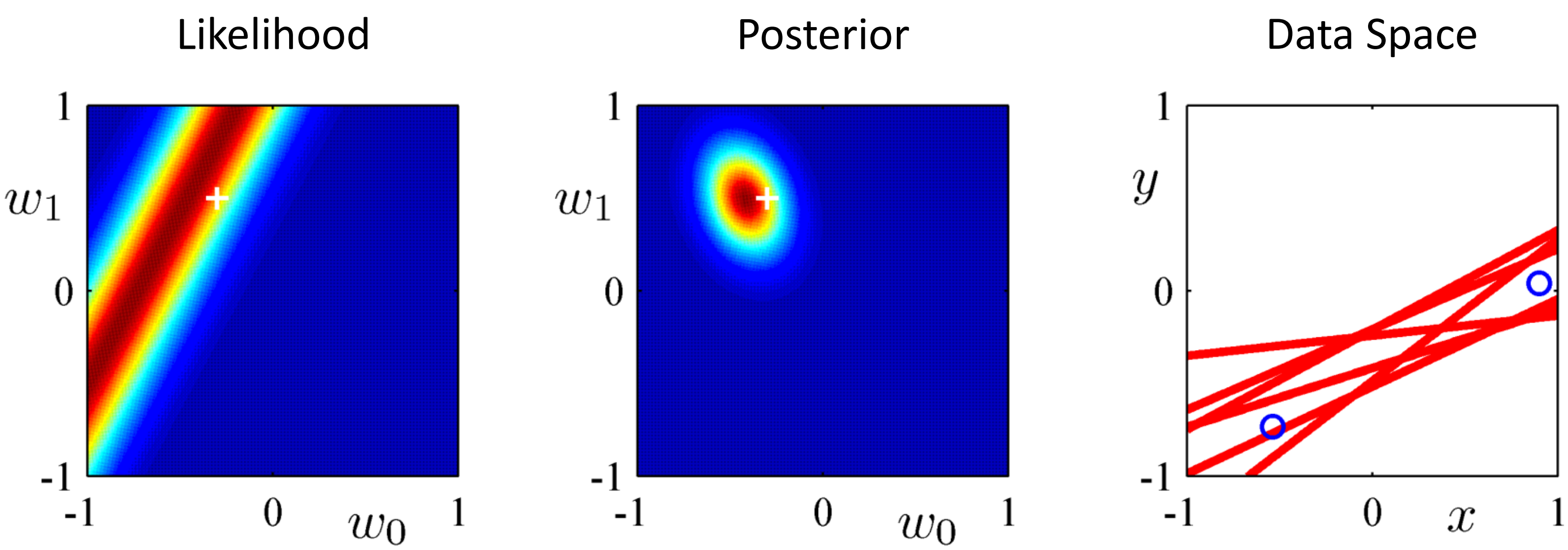
Data Space



$y = w_1 x + w_0$ samples

Bayesian Linear Regression (6)

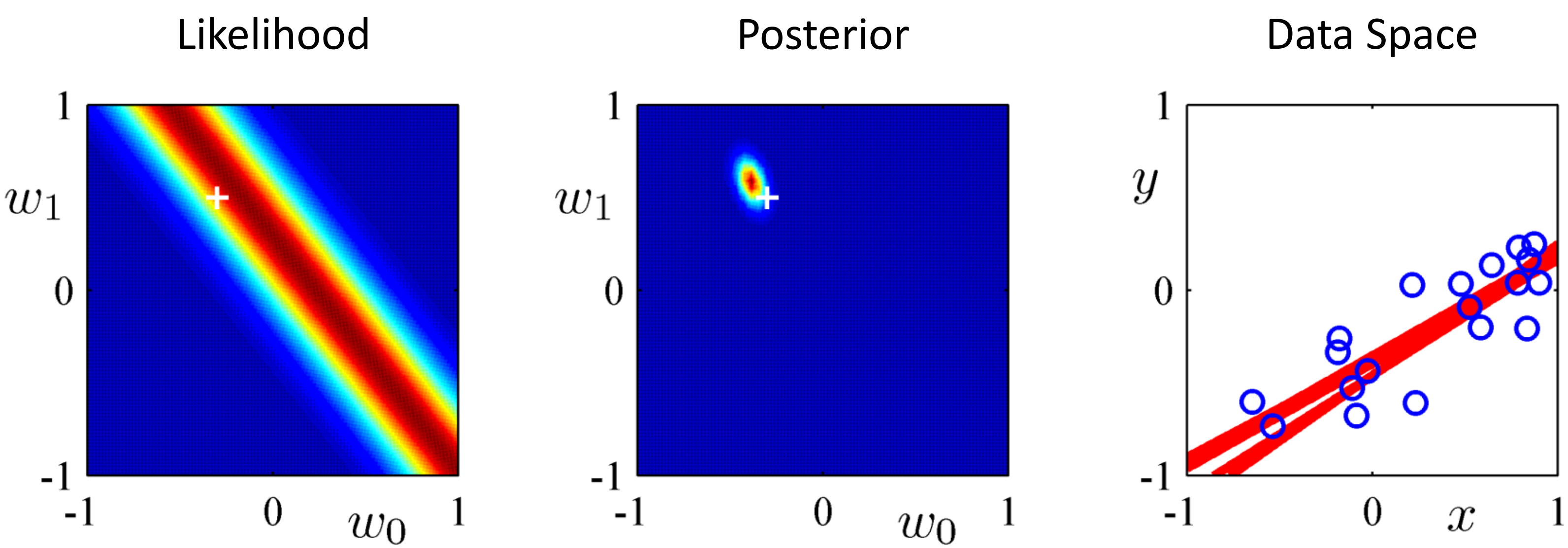
- 2 data points observed



$p(\mathbf{t}|w_1, w_0)$ $p(w_1, w_0|\mathbf{t})$ $y = w_1 x + w_0$ samples

Bayesian Linear Regression (7)

- 20 data points observed



$p(\mathbf{t}|w_1, w_0)$ $\quad$ $p(w_1, w_0|\mathbf{t})$ $\quad$ $y = w_1 x + w_0$ samples

Outlines

- Linear Basis Function Models
- Maximum Likelihood and Least Squares
- Bias Variance Decomposition
- Bayesian Linear Regression
- Predictive Distribution
- Equivalent Kernel
- Bayesian Model Comparison
- Evidence Approximation and Maximization

Predictive Distribution (1)

- Predict t for new values of $\mathbf{x}$ by integrating over $\mathbf{w}$:

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w}$$

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

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)

- Predict t for new values of $\mathbf{x}$ by expecting over $\mathbf{w}$ and ϵ:

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

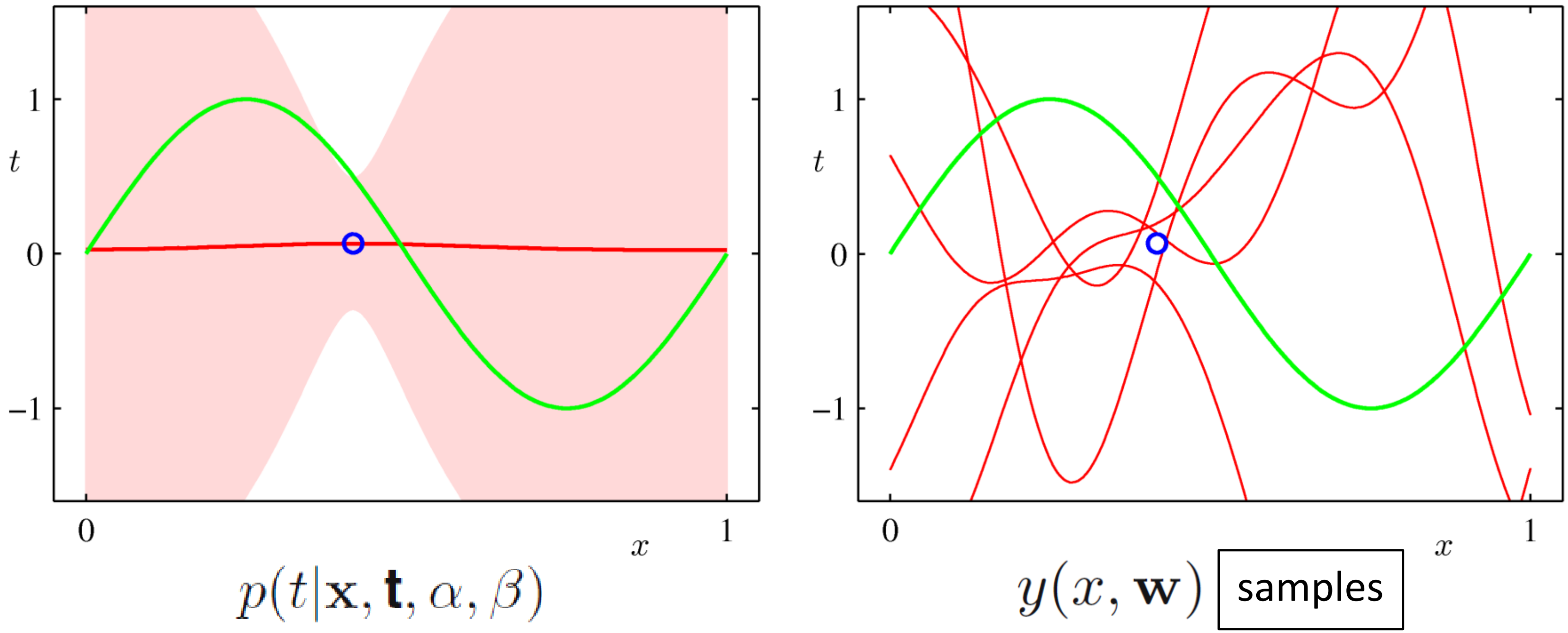
where

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \qquad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

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

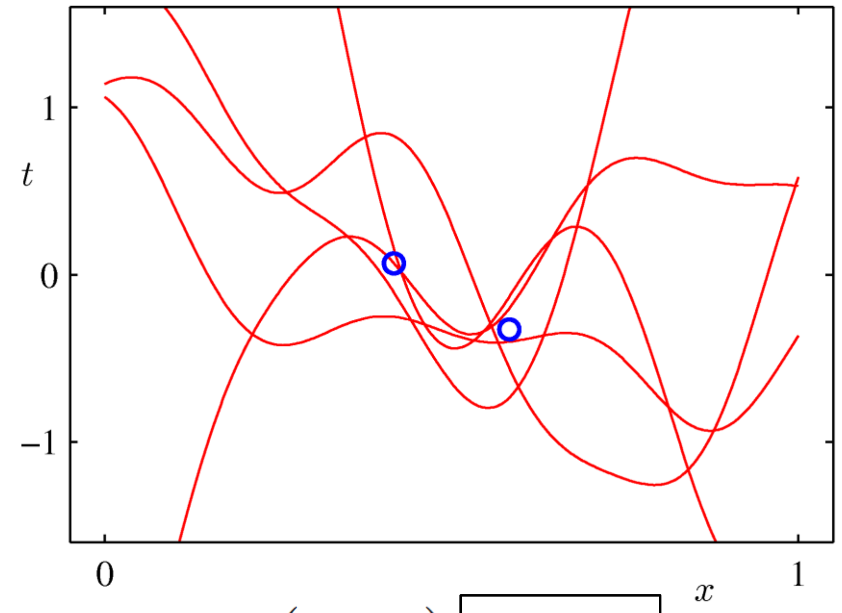
Predictive Distribution (3)

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



Predictive Distribution (4)

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

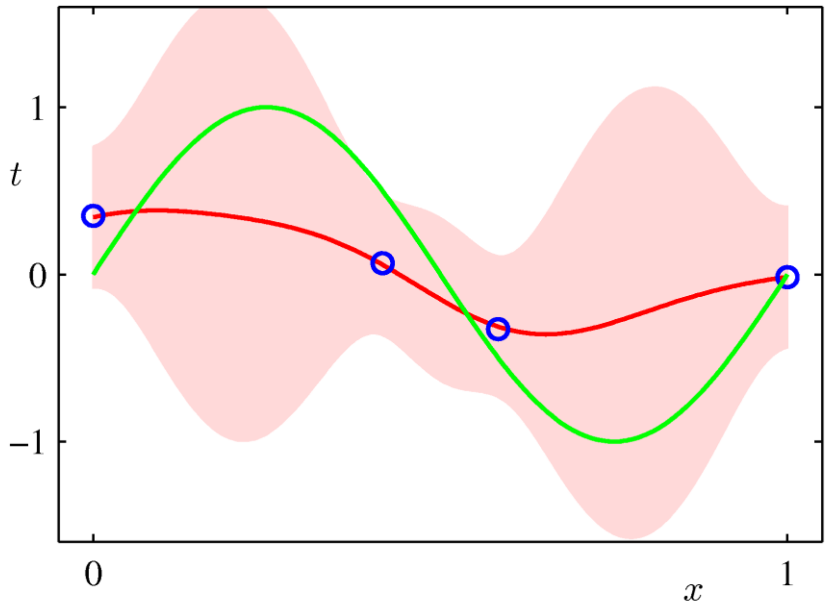


$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta)$

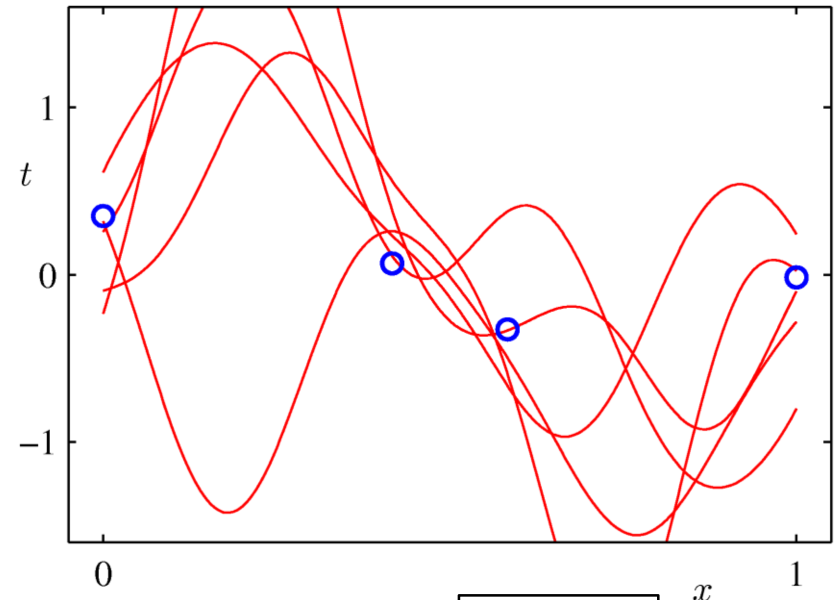
$y(x, \mathbf{w})$ samples

Predictive Distribution (5)

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



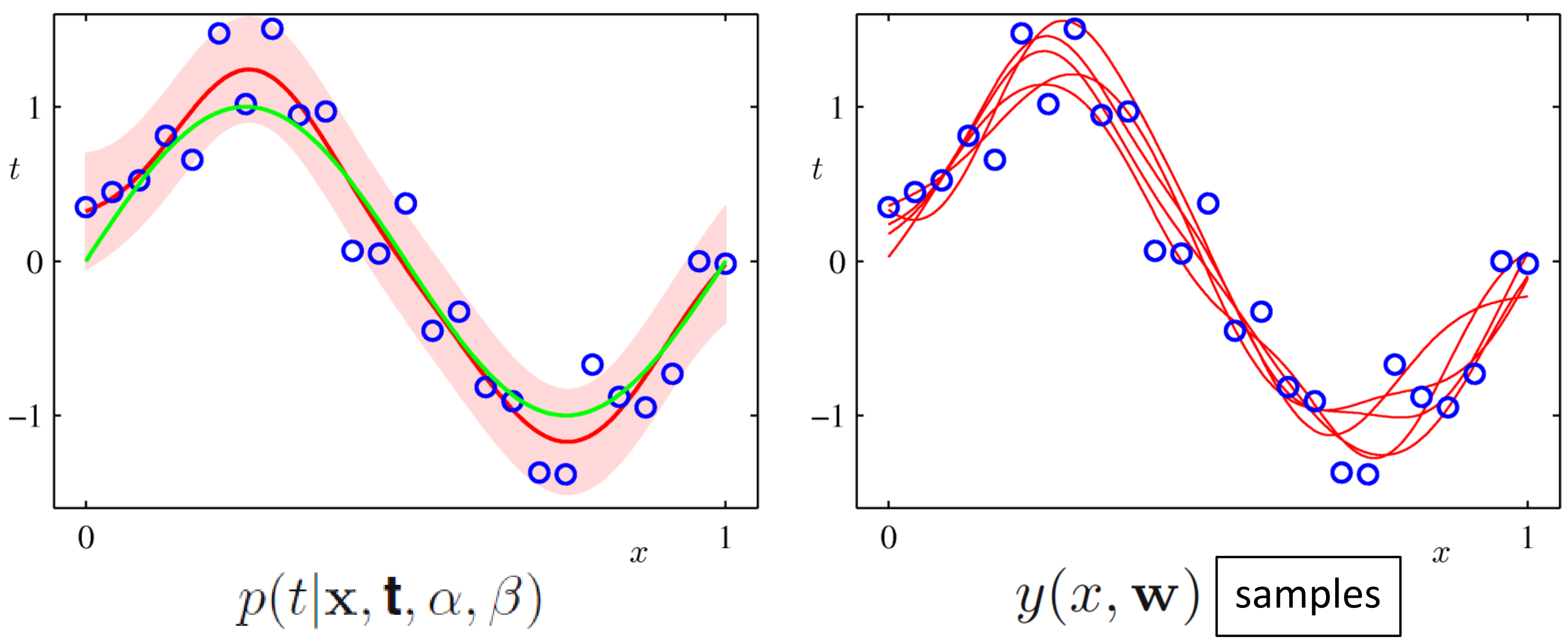
$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta)$



$y(x, \mathbf{w})$ samples

Predictive Distribution (6)

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



Outlines

- Linear Basis Function Models
- Maximum Likelihood and Least Squares
- Bias Variance Decomposition
- Bayesian Linear Regression
- Predictive Distribution
- Equivalent Kernel
- Bayesian Model Comparison
- Evidence Approximation and Maximization

Equivalent Kernel (1)

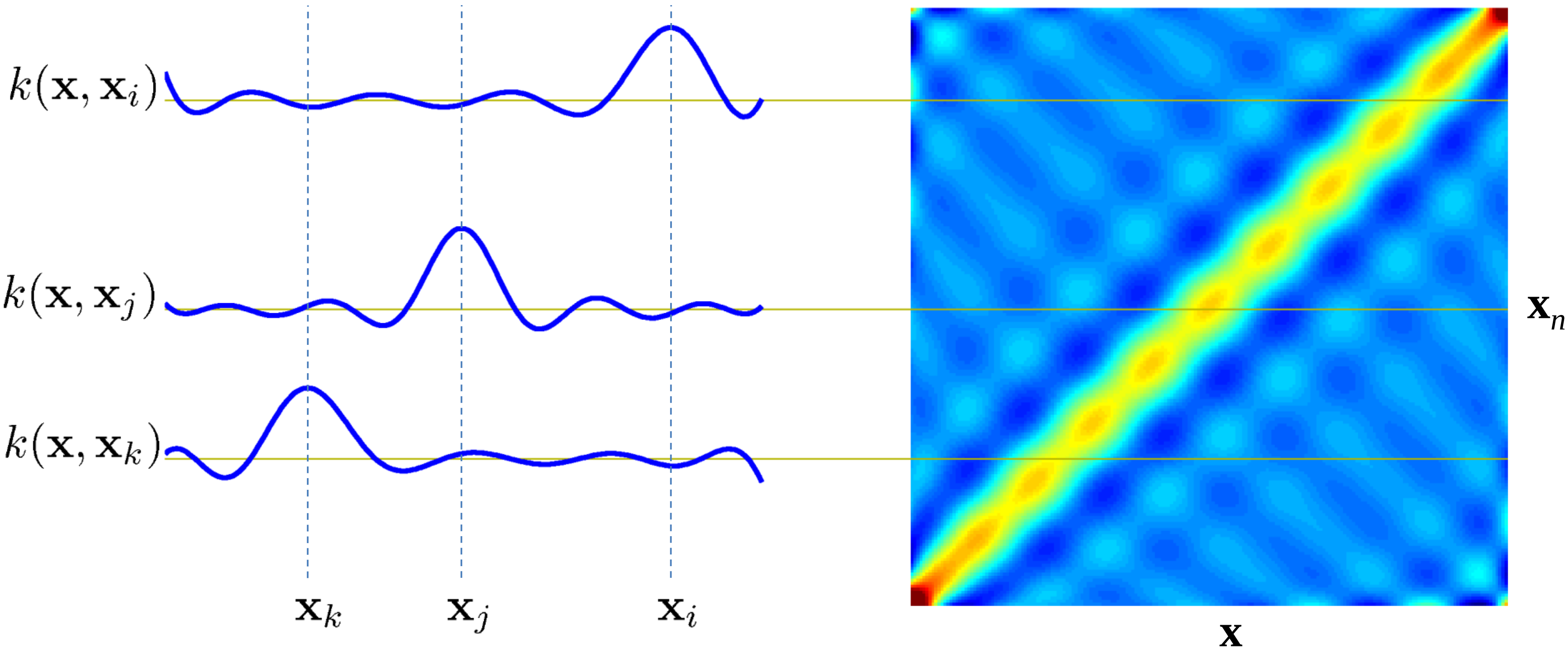
- The predictive mean can be written

$$
\begin{aligned}
y(\mathbf{x}, \mathbf{m}_N) &= \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \\
&= \sum_{n=1}^{N} \underbrace{\beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_n)} t_n \\
&= \sum_{n=1}^{N} \overbrace{k(\mathbf{x}, \mathbf{x}_n)} t_n .
\end{aligned}
$$

Equivalent kernel or *smoother matrix*.

- This is a weighted sum of the training data target values, t_n.

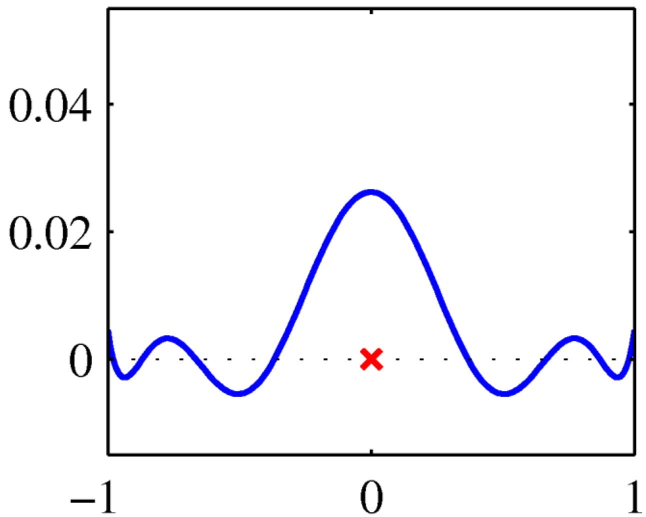
Equivalent Kernel (2)



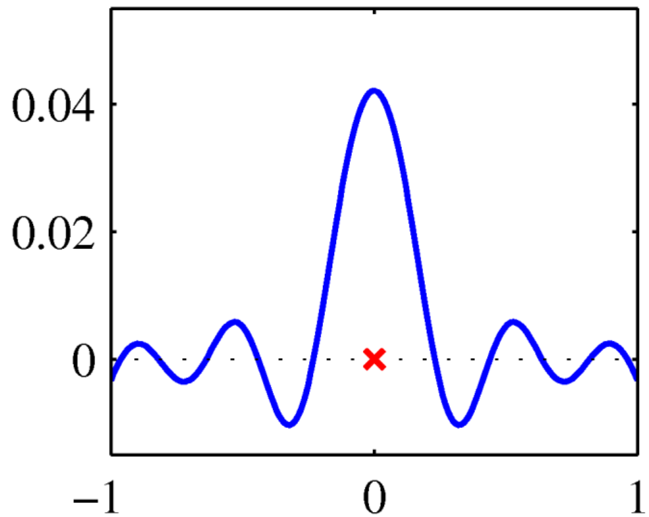
The weight of t_n depends on distance between $\mathbf{x}$ and $\mathbf{x}_n$; nearby $\mathbf{x}_n$ carry more weight.

Equivalent Kernel (3)

Non-local basis functions have local equivalent kernels:



Polynomial



Sigmoidal

Equivalent Kernel (4)

- The kernel as a covariance function: consider

$$
\begin{aligned}
\operatorname{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \operatorname{cov}[\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')] \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') = \beta^{-1}k(\mathbf{x}, \mathbf{x}').
\end{aligned}
$$

- We can avoid the use of basis functions and define the kernel function directly, leading to *Gaussian Processes* (Chapter 6).

Equivalent Kernel (5)

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1$$

for all values of $\mathbf{x}$; however, the equivalent kernel may be negative for some values of $\mathbf{x}$.

Like all kernel functions, the equivalent kernel can be expressed as an inner product:

$$k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\psi}(\mathbf{z})$$

where $\psi(\mathbf{x}) = \beta^{1/2} \mathbf{S}_N^{1/2} \boldsymbol{\phi}(\mathbf{x})$.