Radial basis functions

\[ \sum_{j=0}^{6} \phi_j z_j(x) \]

\[ z_i = \begin{bmatrix} 1 \\ \exp \left[ -\frac{(x_i - \alpha_1)^2}{\lambda} \right] \\ \exp \left[ -\frac{(x_i - \alpha_2)^2}{\lambda} \right] \\ \exp \left[ -\frac{(x_i - \alpha_3)^2}{\lambda} \right] \\ \exp \left[ -\frac{(x_i - \alpha_4)^2}{\lambda} \right] \\ \exp \left[ -\frac{(x_i - \alpha_5)^2}{\lambda} \right] \\ \exp \left[ -\frac{(x_i - \alpha_6)^2}{\lambda} \right] \end{bmatrix} \]
Bayesian regression

(a) 

(b) 

(c) 

(d) $Pr(w^*|x^*)$
Predictive Distribution (1)

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

$$p(t|t, \alpha, \beta) = \int p(t|w, \beta)p(w|t, \alpha, \beta) \, dw$$

$$= \mathcal{N}(t|m_N^T \phi(x), \sigma_N^2(x))$$

where

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x).$$
Bayesian Linear Regression

Likelihood

\[ p(t|X, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | w^T \phi(x_n), \beta^{-1}). \]

A common choice for the prior is

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

for which the posterior is

\[ p(w|t) = \mathcal{N}(w | m_N, S_N) \]

\[ m_N = \beta S_N \Phi^T t \]

\[ S_N^{-1} = \alpha I + \beta \Phi^T \Phi. \]
Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point

\[ \mathcal{N}(t|m_N^T \phi(x), \sigma_N^2(x)) \]

\[ y(x, w) \]
Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points

\[ \mathcal{N}(t|\mathbf{m}_N^T \phi(x), \sigma^2_N(x)) \]  
\[ y(x, w) \]
Predictive Distribution (4)

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

\[ \mathcal{N}(t | \mathbf{m}_N^T \phi(x), \sigma_N^2(x)) \]

\[ y(x, \mathbf{w}) \]
Predictive Distribution (5)

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

\[ \mathcal{N}(t | \mathbf{m}_N^T \phi(x), \sigma^2_N(x)) \]

\[ y(x, \mathbf{w}) \]
Predictive Distribution

\[ p(t|x, w_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, w_{ML}), \beta_{ML}^{-1}) \]
Bayesian Predictive Distribution

\[ p(t|x, x_t) = \mathcal{N}(t|m(x), s^2(x)) \]
Equivalent Kernel (1)

The predictive mean can be written

\[ y(x, m_N) = m_N^T \phi(x) = \beta \phi(x)^T S_N \Phi^T t \]

\[ = \sum_{n=1}^{N} \beta \phi(x)^T S_N \phi(x_n) t_n \]

\[ = \sum_{n=1}^{N} k(x, x_n) t_n. \]

This is a weighted sum of the training data target values, \( t_n \).
Equivalent Kernel (2)

Weight of $t_n$ depends on distance between $x$ and $x_n$; nearby $x_n$ 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

\[
\text{cov}[y(x), y(x')] = \text{cov}[\phi(x)^T w, w^T \phi(x')]
\]

\[
= \phi(x)^T S_N \phi(x') = \beta^{-1} k(x, x').
\]

We can avoid the use of basis functions and define the kernel function directly, leading to \textit{Gaussian Processes}.
Equivalent Kernel (5)

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

for all values of \( x \); however, the equivalent kernel may be negative for some values of \( x \).

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

\[ k(x, \mathbf{z}) = \psi(x)^T \psi(\mathbf{z}) \]

where \( \psi(x) = \beta^{1/2} S_N^{1/2} \phi(x) \).
Bayesian Model Comparison (1)

How do we choose the ‘right’ model?
Assume we want to compare models $M_i$, $i=1, \ldots, L$, using data $D$; this requires computing

$$p(M_i | D) \propto p(M_i) p(D | M_i).$$

Posterior Prior Model evidence or marginal likelihood

**Bayes Factor**: ratio of evidence for two models

$$\frac{p(D | M_i)}{p(D | M_j)}$$
Having computed $p(M_i|D)$, we can compute the predictive (mixture) distribution

$$p(t|x, D) = \sum_{i=1}^{L} p(t|x, M_i, D)p(M_i|D).$$

A simpler approximation, known as *model selection*, is to use the model with the highest evidence.
Bayesian Model Comparison (3)

For a model with parameters $w$, we get the model evidence by marginalizing over $w$

$$p(D|\mathcal{M}_i) = \int p(D|w, \mathcal{M}_i)p(w|\mathcal{M}_i) \, dw.$$ 

Note that

$$p(w|D, \mathcal{M}_i) = \frac{p(D|w, \mathcal{M}_i)p(w|\mathcal{M}_i)}{p(D|\mathcal{M}_i)}.$$
Bayesian Model Comparison (4)

For a given model with a single parameter, $w$, consider the approximation

$$p(D) = \int p(D|w)p(w) \, dw$$

$$\simeq p(D|w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

where the posterior is assumed to be sharply peaked.
Bayesian Model Comparison (5)

Taking logarithms, we obtain

\[
\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + \ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right).
\]

With M parameters, all assumed to have the same ratio \( \Delta w_{\text{posterior}} / \Delta w_{\text{prior}} \), we get

\[
\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + M \ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right).
\]

Negative and linear in M.
Bayesian Model Comparison (6)

Matching data and model complexity

$p(D)$

$\mathcal{M}_1$

$\mathcal{M}_2$

$\mathcal{M}_3$

$D_0$
The Evidence Approximation (1)

The fully Bayesian predictive distribution is given by

\[ p(t|t) = \int \int \int p(t|w, \beta)p(w|t, \alpha, \beta)p(\alpha, \beta|t) \, dw \, d\alpha \, d\beta \]

but this integral is intractable. Approximate with

\[ p(t|t) \simeq p\left(t|t, \hat{\alpha}, \hat{\beta}\right) = \int p\left(t|w, \hat{\beta}\right) p\left(w|t, \hat{\alpha}, \hat{\beta}\right) \, dw \]

where \((\hat{\alpha}, \hat{\beta})\) is the mode of \(p(\alpha, \beta|t)\), which is assumed to be sharply peaked; a.k.a. *empirical Bayes, type II* or *generalized maximum likelihood*, or evidence approximation.
The Evidence Approximation (2)

From Bayes’ theorem we have

\[ p(\alpha, \beta | t) \propto p(t | \alpha, \beta) p(\alpha, \beta) \]

and if we assume \( p(\alpha, \beta) \) to be flat we see that

\[ p(\alpha, \beta | t) \propto p(t | \alpha, \beta) \]
\[ = \int p(t | w, \beta) p(w | \alpha) \, dw. \]

General results for Gaussian integrals give

\[ \ln p(t | \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(m_N) + \frac{1}{2} \ln |S_N| - \frac{N}{2} \ln(2\pi). \]
The Evidence Approximation (3)

Example: sinusoidal data, $M^{\text{th}}$ degree polynomial,

$$\alpha = 5 \times 10^{-3}$$

$\ln p(t|\alpha, \beta)$