Bayesian linear regression

Models for linear regression

Francesco Corona

Bayesian linear regression
Models for linear regression

Bayesian linear regression

In a maximum likelihood approach for setting parameters in a linear model for regression, we tune effective model complexity, the number of basis functions

- We control it based on the size of the data set

Adding a regularisation term to the log likelihood function means that the effective model complexity can be controlled by the regularisation coefficient

- The choice of the number and form of the basis functions is still important in determining the overall behaviour of the model

This leaves the issue of setting appropriate model complexity for the problem

- It cannot be decided simply by maximising the likelihood function
- This always leads to excessively complex models and over-fitting

Independent hold-out data can be used to determine model complexity

- This can be both computationally expensive and wasteful of valuable data

Bayesian linear regression (cont.)

We therefore turn to a Bayesian treatment of linear regression

- Avoids the over-fitting problem of maximum likelihood
- Leads to automatic methods of setting model complexity

We again focus on the case of a single target variable t

Outline

Parameter distribution
Bayesian linear regression

Parameter distribution

The Bayesian treatment of linear regression starts by introducing a prior probability distribution over the model parameters $\mathbf{w}$ [1]

The likelihood function $p(\mathtt{t}|\mathbf{w})$ is the exponential of a quadratic function of $\mathbf{w}$

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

The corresponding conjugate prior is thus a Gaussian distribution of the form

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{1}$$

- Mean $\mathbf{m}_0$ and covariance $\mathbf{S}_0$

[1] There also is the noise precision parameter β, we first assume it is a known constant

Parameter distribution (cont.)

The posterior distribution is $\propto$ to the product of likelihood function and prior

- Due to the choice of a conjugate prior, the posterior is Gaussian too[2]

$$p(\mathbf{w}|\mathbf{t}) \propto \left(\prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})\right)\mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

$$\propto \exp\left(-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi})^T(\mathbf{t}-\boldsymbol{\Phi})\right)\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right)$$

The posterior distribution can be thus written directly in the form

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{2}$$

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{t}) \tag{3}$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} - \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{4}$$

[2] We derived something similar when discussing Bayes' theorem for Gaussian variables. This distribution is calculated by completing the square in the exponential and finding the normalisation coefficient using the result for a normalised Gaussian

Parameter distribution (cont.)

Because the posterior distribution is Gaussian, its mode coincides with its mean

- The maximum posterior weight vector is given by $\mathbf{w}_{\text{MAP}} = \mathbf{m}_N$

If we consider an infinitely broad prior $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ with $\alpha \to 0$, the mean $\mathbf{m}_N$ of the posterior distribution reduces to the maximum likelihood value

$$\mathbf{w}_{\text{ML}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

Similarly, if $N = 0$, then again the posterior distribution reverts to the prior

Parameter distribution (cont.)

We consider a simple form of the Gaussian distribution, the zero-mean isotropic

- Only a single precision parameter α characterises it

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{5}$$

The corresponding posterior distribution over $\mathbf{w}$ is then $p(\mathbf{w}|\mathtt{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathtt{t} \tag{6}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi} \tag{7}$$

Parameter distribution (cont.)

The log of the posterior distribution is given by the sum of the log likelihood and the log of the prior

- As a function of $\mathbf{w}$, it takes the form

$$\ln p(\mathbf{w}|\alpha) = -\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \text{const} \qquad (8)$$

Maximisation of this posterior distribution with respect to $\mathbf{w}$ is equivalent to

$$\frac{1}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\right)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}, \quad \text{with } \lambda = \alpha/\beta$$

- the minimisation of the sum-of-squares error function
- with the addition of a quadratic regularisation term

Parameter distribution (cont.)

To illustrate Bayesian learning in a linear basis function model, together with the sequential update of a posterior distribution, we consider line fitting

Consider a single input variable x, a single target variable t and linear model

$$y(x, \mathbf{w}) = w_0 + w_1 x$$

We generate a synthetic set of data from function $f(x, \mathbf{a}) = a_0 + a_1 x$

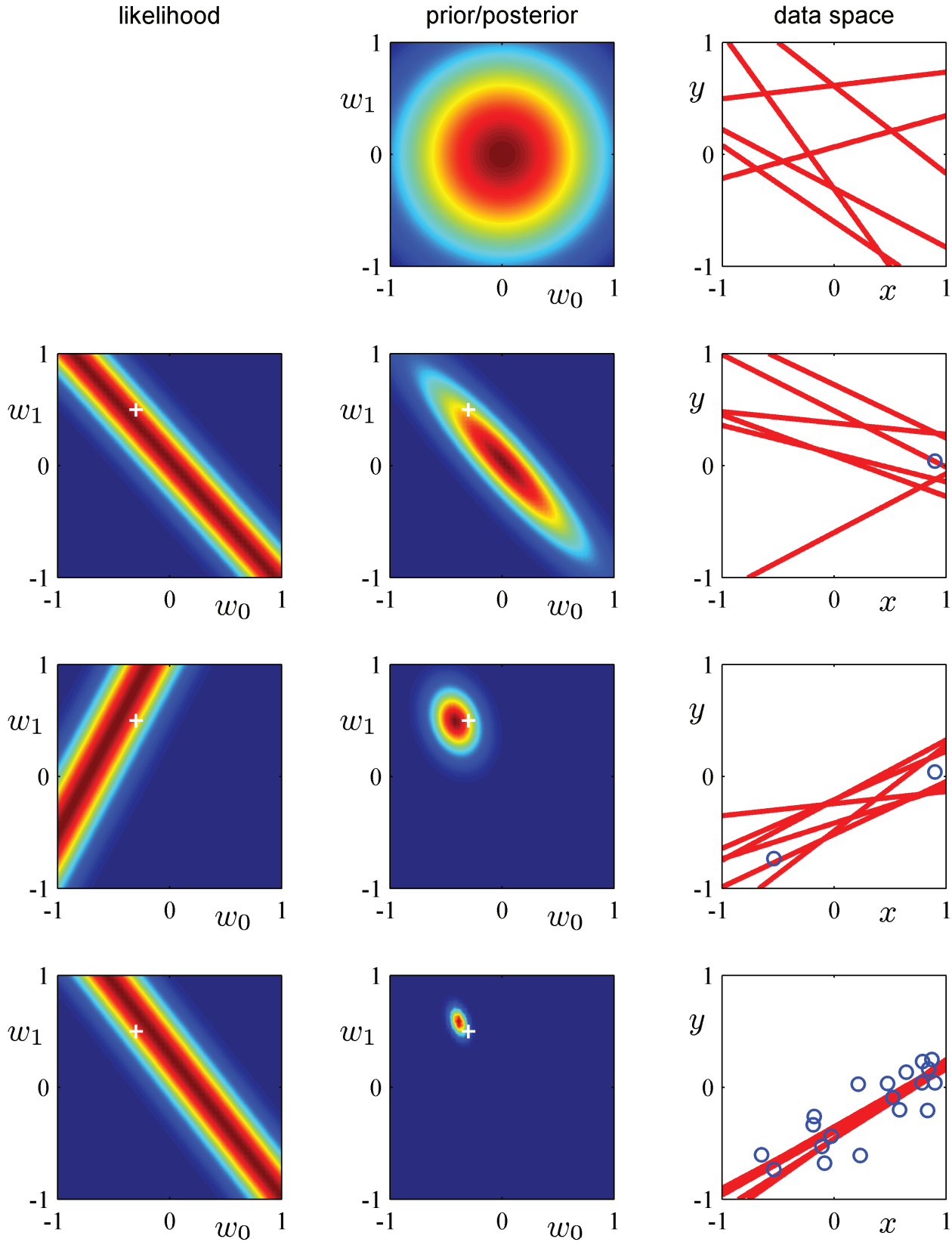
- with $a_0 = -0.3$ and $a_1 = 0.5$

For a selection of input points $x_n \sim \mathcal{U}(-1, +1)$, we first evaluate $f(x_n, \mathbf{a})$ and then we add Gaussian noise $\varepsilon \sim \mathcal{N}(0, 0.2^2)$ to get the target values t_n

- The goal is to recover the values of a_0 and a_1 (thru w_0 and w_1)
- Under the assumption that the variance of the noise is known

$$\beta = \left(\frac{1}{0.2}\right)^2 = 25$$

- We fix $\alpha = 2.0$ in the Gaussian prior $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$



Because **w** is bi-dimensional, we can plot the prior and posterior distribution

Parameter distribution (cont.)

The plain Gaussian is not the only available form of prior over the parameters

- The Gaussian can be generalised

$$p(\mathbf{w}|\alpha) = \left(\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right)^{M}\exp\left(-\frac{\alpha}{2}\sum_{j=0}^{M-1}|w_j|^q\right) \tag{9}$$

- It is not a conjugate prior to the likelihood function, unless $q = 2$

Finding the maximum of the posterior distribution over the parameters corresponds to the minimisation of a regularised error function

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

Predictive distribution
Bayesian linear regression

Predictive distribution

In practice, we are not usually interested in the value of $\mathbf{w}$ itself

- We want to predictions of t for new values of $\mathbf{x}$

This requires that we evaluate the **predictive distribution** defined by

$$p(t|\mathtt{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathtt{t}, \alpha, \beta) d\mathbf{w} \tag{10}$$

where $\mathtt{t}$ is the vector of target values from the training set[3]

- The conditional distribution of the target is $p(t|\cancel{\mathbf{x}}, \mathbf{w}, \beta)$ (can be omitted)
- The posterior distribution of the weights is $p(\mathbf{w}|\mathtt{t}, \alpha, \beta)$

[3] We omit the corresponding input vectors $\mathtt{X}$ from the rhs of the conditioning to simplify notation

Predictive distribution (cont.)

Calculating the predictive distribution involves the convolution of two Gaussians

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

- The conditional distribution of the target

$$p(t|\mathbf{w},\beta) = p(t|\mathbf{x},\mathbf{w},\beta) = \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}) \qquad \text{with} \begin{cases} y(\mathbf{x},\mathbf{w}) = \phi(\mathbf{x})^T\mathbf{w} \\ \beta^{-1} \end{cases}$$

- The posterior distribution of the weights

$$p(\mathbf{w}|\mathbf{t},\alpha,\beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N) \qquad \text{with} \begin{cases} \mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t}) \\ \mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} - \beta\mathbf{\Phi}^T\mathbf{\Phi} \end{cases}$$

The mean of the convolution is the sum of the mean of the two Gaussians, and the covariance of the convolution is the sum of their covariances

Predictive distribution (cont.)

Using old results (Eq. 2.115, $\star$), the predictive distribution takes the form

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})) \tag{11}$$

where the variance $\sigma_N^2(\mathbf{x})$ if the predictive distribution is

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}) \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}) \tag{12}$$

- the first term $1/\beta$ represents the noise on the data
- the second term reflects uncertainty associated with $\mathbf{w}$

The noise process and the distribution of $\mathbf{w}$ are independent Gaussians

- their variances are additive

Predictive distribution (cont.)

As additional points are observed, the posterior distribution becomes narrower $\star$

- As a consequence, it can be shown that $\sigma^2_{N+1}(\mathbf{x}) \leq \sigma^2_N(\mathbf{x})$

In the limit $N \to \infty$, the second term in $\sigma^2_N(\mathbf{x}) = \dfrac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})$ goes to zero

- The variance of the predictive distribution arises solely from the additive noise governed by the parameter β

Predictive distribution (cont.)

Illustration of the predictive distribution for Bayesian linear regression

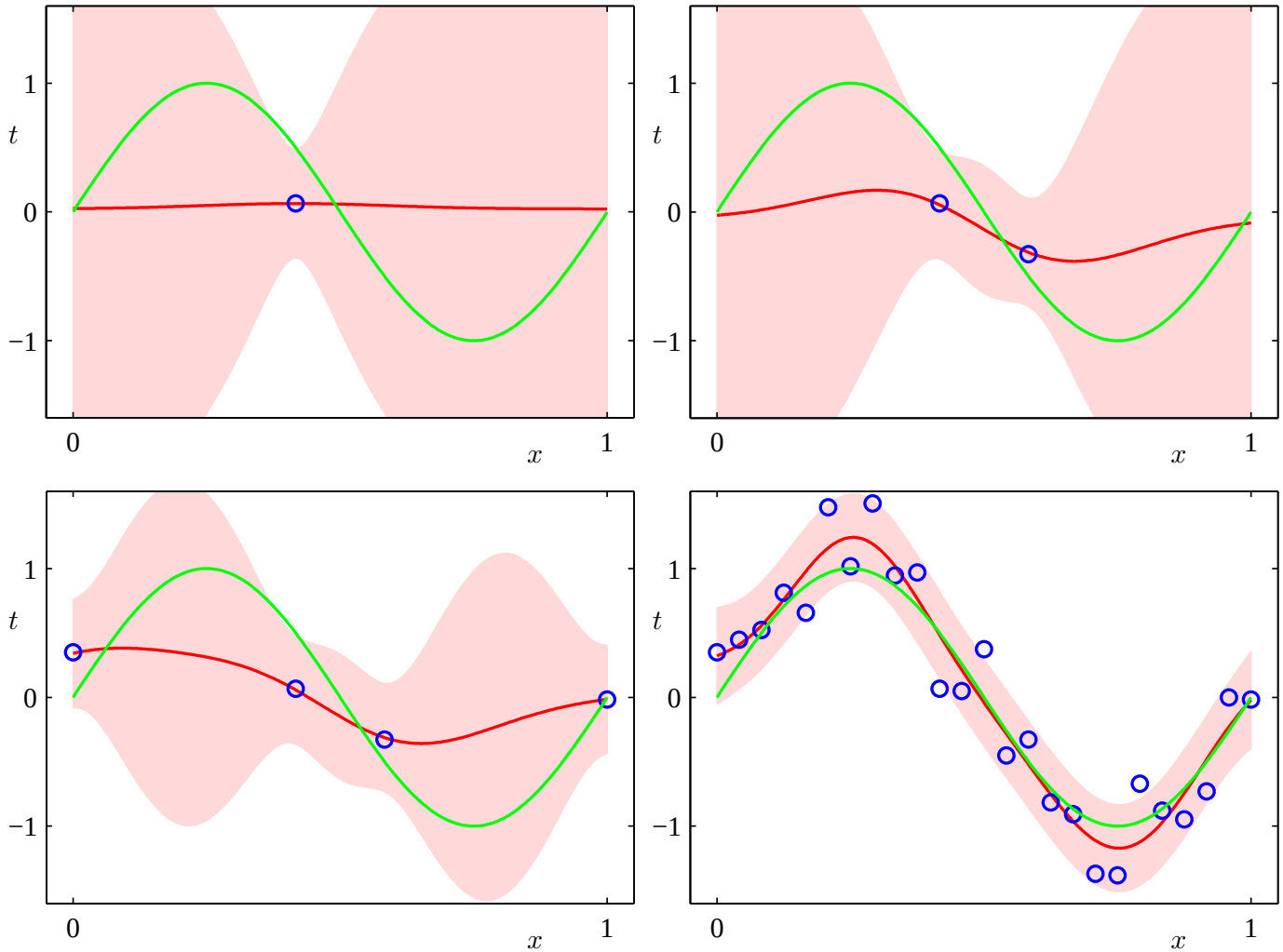
- The sinusoidal data with additive Gaussian noise

Model fitted to data, linear combination of 9 Gaussian basis functions

- Different datasets of different sizes
- $N = 1$, $N = 2$, $N = 4$ and $N = 25$

The red curve (one per N) is the mean of the Gaussian predictive distribution

- The red shaded region spans one standard deviation either side the mean



The predictive uncertainty (the variance) depends on x, it is smallest in the neighbourhood of the points and it decreases as more points are observed

Predictive distribution (cont.)

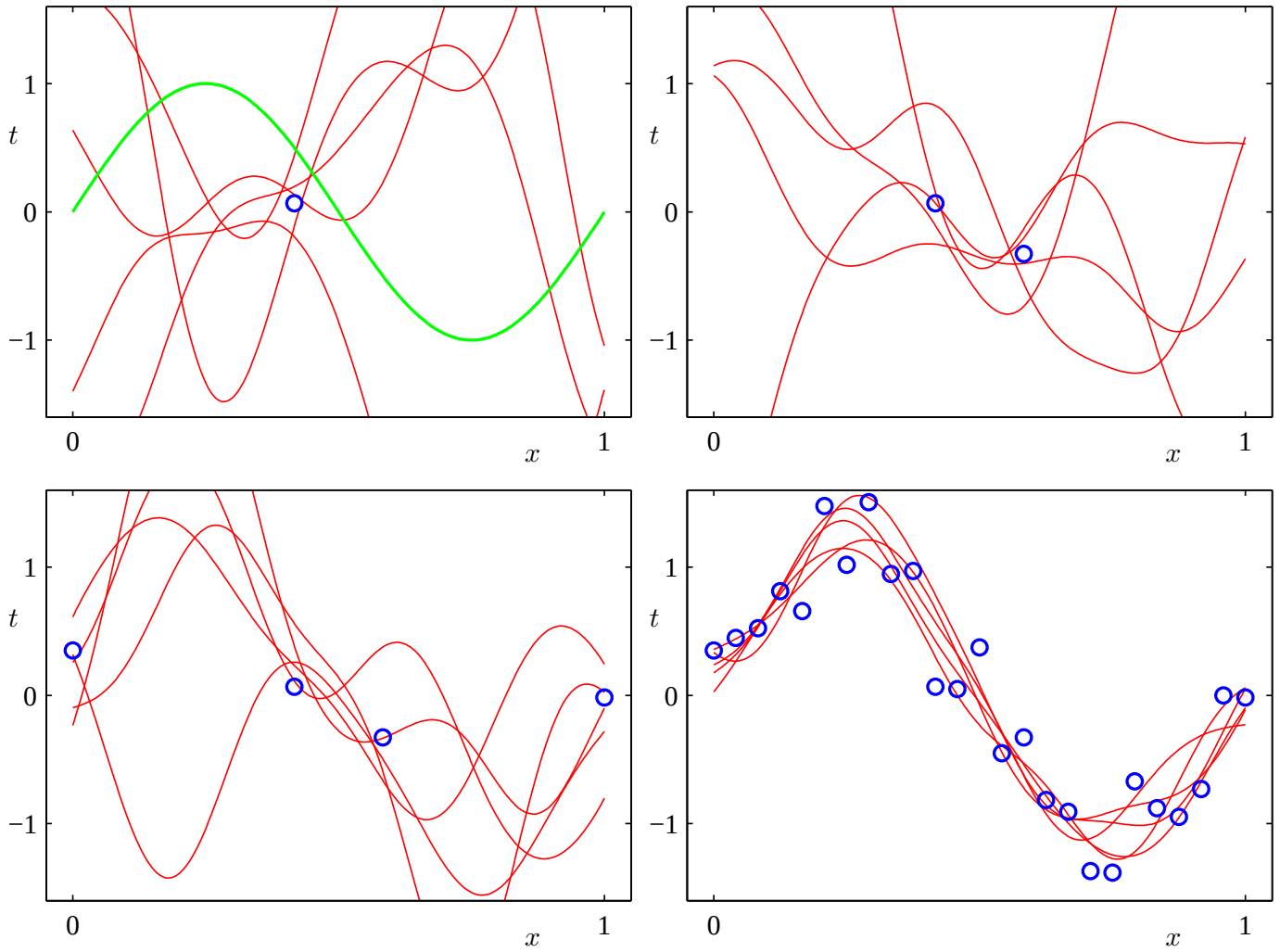
So far, we showed only the point-wise predictive variance as a function of x

In order to gain insight into the covariance between predictions at different values of x, we can draw samples from the posterior distribution over $\mathbf{w}$

- We have a probabilistic model and we can generate new data

Predictive distribution (cont.)

Plots of the functions $y(x, \mathbf{w})$, with sampled $\mathbf{w}$s from the posterior distribution



Predictive distribution (cont.)

If both $\mathbf{w}$ and β are treated as unknowns, we can introduce a conjugate prior distribution $p(\mathbf{w}, \beta)$ which will be given by a Gaussian-gamma distribution

- The resulting predictive distribution is a Student's t-distribution

Equivalent kernel
Bayesian linear regression

Equivalent kernel

The posterior mean solution $\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$ for the linear basis function model has an interesting interpretation that will set the stage for kernel methods

Substituting $\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$ into $y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$, we get

$$y(\mathbf{x}, \mathbf{m}_N) = \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \mathbf{\Phi}^t \mathbf{t} = \sum_{n=1}^{N} \beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_n) t_n \qquad (13)$$

A new expression for the predictive distribution, where $\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} - \beta \mathbf{\Phi}^T \mathbf{\Phi}$

- The mean of the predictive distribution at a point $\mathbf{x}$ is a linear combination of the training set target variables t_n

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} \underbrace{\beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_n)}_{k(\mathbf{x}, \mathbf{x}_n)} t_n$$

Equivalent kernel (cont.)

The function $k(\mathbf{x}, \mathbf{x}')$ is known as the **smoother matrix** or **equivalent kernel**

$$k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}') \tag{14}$$

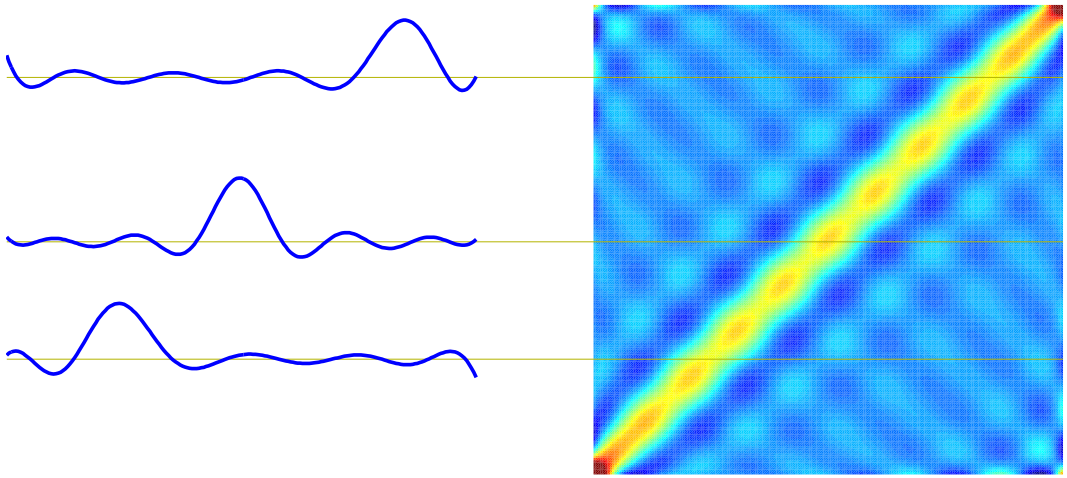
Regression functions that make predictions by taking linear combinations of the target values t_n in the training set are known as **linear smoothers**

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n \tag{15}$$

The dependence on the input values $\mathbf{x}_n$ in the training set are through $\mathbf{S}_N$

Equivalent kernel (cont.)

The kernel functions $k(x, x')$ are collected in the smoother matrix
They can be plotted as a function of x' for different (3) values of x



Localised around x, so the mean $y(x, \mathbf{m}_N)$ of the predictive distribution at x

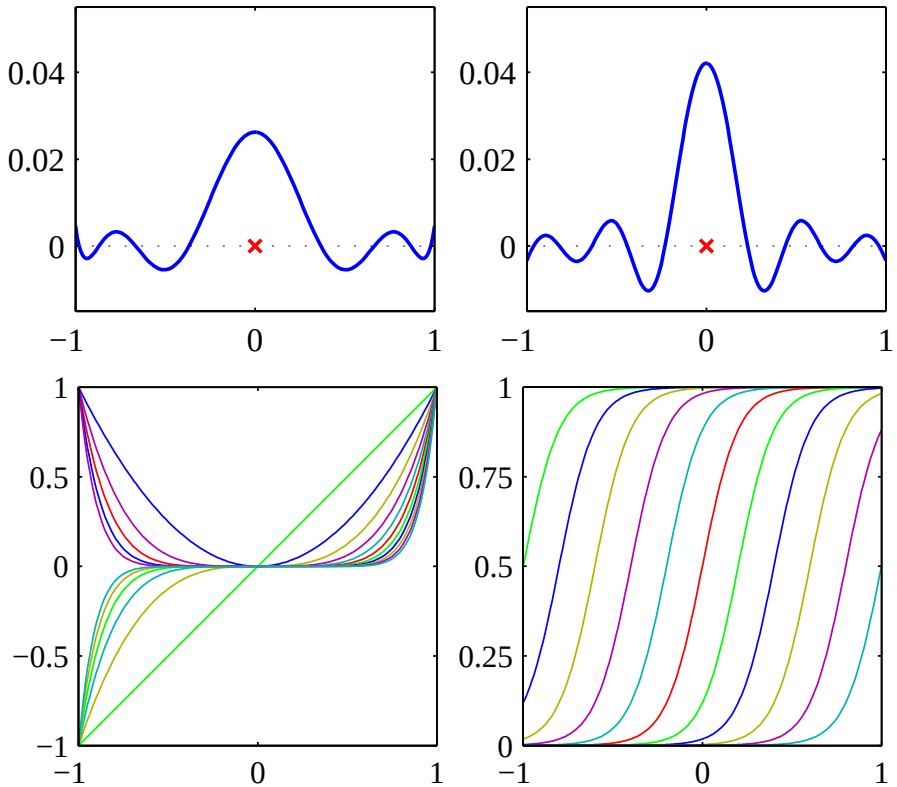
- is a weighted combination of the target values
- points close to x are given higher weight

Intuitively, local evidence is weighted more strongly that distant evidence

Equivalent kernel (cont.)

Examples of equivalent kernels $k(x, x')$ for $x = 0$ plotted as a function of x'

- Polynomial basis functions (left) and sigmoidal basis functions (right)



k is a localised function of x', though the corresponding basis function is not

Equivalent kernel (cont.)

Further insight into the role of the equivalent kernel can be obtained by considering the covariance between $y(\mathbf{x})$ and $y(\mathbf{x}')$, which is given[4] by

$$\begin{aligned} \text{cov}\Big[y(\mathbf{x}), y(\mathbf{x}')\Big] &= \text{cov}\Big[\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}, \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}')\Big] \\ &= \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') \\ &= \beta^{-1}k(\mathbf{x}, \mathbf{x}') \end{aligned} \tag{16}$$

From the form of the equivalent kernel, we see that the predictive mean at nearby points will be highly correlated, whereas for more distant pairs of points the correlation will be smaller

[4] We used $p(\mathbf{w}|\mathrm{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ and $k(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}')$

Equivalent kernel (cont.)

The formulation of linear regression in terms of a kernel function suggests an alternative approach to regression

Instead of introducing a set of basis functions, which implicitly determines an equivalent kernel, we can instead define a localised kernel directly and use this to make predictions for new input vectors $\mathbf{x}$, given the observed training set

This leads to a practical framework for regression (and classification) called Gaussian processes

Equivalent kernel (cont.)

The effective kernel defines the weights by which the training set target values are combined in order to make a prediction at a new value of $\mathbf{x}$

It can be shown that these weights sum to one, in other words

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}') = 1, \quad \forall \mathbf{x} \tag{17}$$

It can also be shown that the kernel function can be written

$$k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\psi}(\mathbf{x})^T \boldsymbol{\psi}(\mathbf{z}) \tag{18}$$

This is an inner product with respect to vector $\boldsymbol{\psi}(\mathbf{x})$ of a set of nonlinear functions, with

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