

3F8: Inference

Bayesian Linear Regression

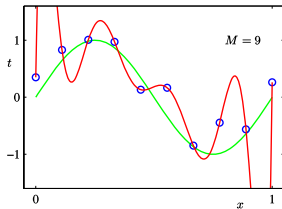
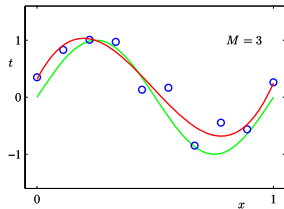
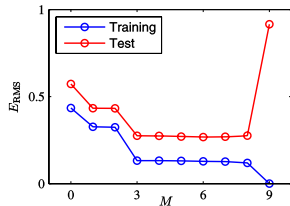
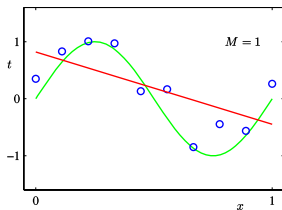
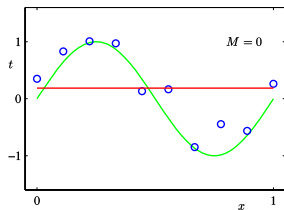
José Miguel Hernández–Lobato and Richard E. Turner

Department of Engineering
University of Cambridge

Lent Term

Motivation

A large number of basis functions can lead to **over-fitting** of the maximum likelihood estimate: the model fits the **training data** well but it performs poorly on new **test data**.



	$M=0$	$M=1$	$M=6$	$M=9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

Instead, favor **smooth solutions** by using Bayes rule with **priors** that enforce \mathbf{w} to be small.

Figures and table: C. Bishop. *Pattern Recognition and Machine Learning*, 2006.

Bayesian inference

Given data $\mathcal{D} = \{(\tilde{\mathbf{x}}_n, y_n)\}_{n=1}^N$, we assume the linear regression model

$$y_n = \mathbf{w}^T \tilde{\mathbf{x}}_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2),$$

with unknown \mathbf{w} . We assume σ^2 **is known** to simplify inference.

We also assume a **prior distribution** $p(\mathbf{w})$ on the model coefficients.

The **posterior distribution** for \mathbf{w} given \mathcal{D} is obtained by Bayes rule:

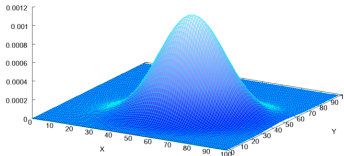
$$p(\mathbf{w} | \mathbf{y}, \tilde{\mathbf{X}}) = \frac{p(\mathbf{y} | \tilde{\mathbf{X}}, \mathbf{w}) p(\mathbf{w})}{p(\mathbf{y} | \tilde{\mathbf{X}})} . \quad \text{Model}$$

The **predictive distribution** for y_* given a new corresponding \mathbf{x}_* is

$$p(y_* | \tilde{\mathbf{x}}_*, \mathbf{y}, \tilde{\mathbf{X}}) = \int p(y_* | \tilde{\mathbf{x}}_*, \mathbf{w}) p(\mathbf{w} | \mathbf{y}, \tilde{\mathbf{X}}) d\mathbf{w} . \quad \text{Inference}$$

Exact inference is possible if prior and noise distributions are **Gaussian**.

Multivariate Gaussian distribution



The density of a D -dimensional vector \mathbf{x} is

$$\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^\top \mathbf{V}^{-1} (\mathbf{x} - \mathbf{m}) \right\}.$$

The density is proportional to the **exponential of a quadratic function** of \mathbf{x} :

$$\begin{aligned} \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) &= \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} \mathbf{x} \mathbf{V}^{-1} \mathbf{x}^\top + \mathbf{m}^\top \mathbf{V}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m} \mathbf{V}^{-1} \mathbf{m}^\top \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \mathbf{V}^{-1} \mathbf{x} + \mathbf{m}^\top \mathbf{V}^{-1} \mathbf{x} \right\}, \end{aligned} \quad (1)$$

with **normalization constant** $\sqrt{(2\pi)^D |\mathbf{V}|} \exp\{1/2 \mathbf{m}^\top \mathbf{V}^{-1} \mathbf{m}\}$.

\mathbf{m} is the D -dimensional mean vector and \mathbf{V} is the $D \times D$ covariance matrix:

$$\mathbf{m} = \mathbf{E}[\mathbf{x}], \quad \mathbf{V} = \mathbf{E}[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top] = \mathbf{E}[\mathbf{x}\mathbf{x}^\top] - \mathbf{m}\mathbf{m}^\top.$$

$\hat{\mathbf{x}} \sim \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V})$ can be obtained using the **matrix square root** $\mathbf{V} = \mathbf{V}^{1/2}(\mathbf{V}^{1/2})^\top$:

$$\hat{\mathbf{x}} = \mathbf{m} + \mathbf{V}^{1/2} \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Effect of parameters on density function

Assuming

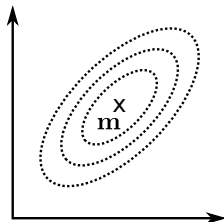
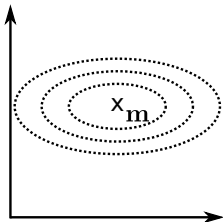
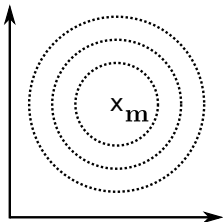
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^\top \mathbf{V}^{-1} (\mathbf{x} - \mathbf{m}) \right\}.$$

The parameter \mathbf{m} determines **mode location** and \mathbf{V} **scales and rotates** the space.

What can you say about

$$\mathbf{V} = \begin{bmatrix} v_1 & \text{cov} \\ \text{cov} & v_2 \end{bmatrix}$$

given the following contour plots of $p(\mathbf{x})$?



Effect of parameters on density function

Assuming

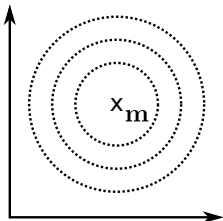
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^\top \mathbf{V}^{-1} (\mathbf{x} - \mathbf{m}) \right\}.$$

The parameter \mathbf{m} determines **mode location** and \mathbf{V} **scales and rotates** the space.

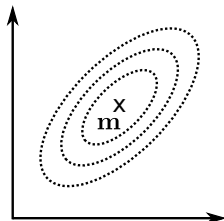
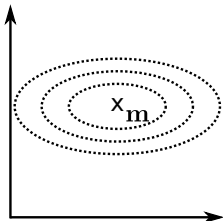
What can you say about

$$\mathbf{V} = \begin{bmatrix} v_1 & \text{cov} \\ \text{cov} & v_2 \end{bmatrix}$$

given the following contour plots of $p(\mathbf{x})$?



$$v_1 = v_2, \text{cov} = 0.$$



Effect of parameters on density function

Assuming

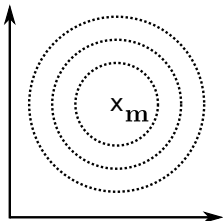
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^\top \mathbf{V}^{-1} (\mathbf{x} - \mathbf{m}) \right\}.$$

The parameter \mathbf{m} determines **mode location** and \mathbf{V} **scales and rotates** the space.

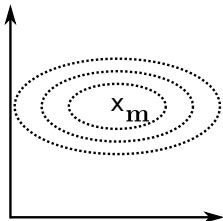
What can you say about

$$\mathbf{V} = \begin{bmatrix} v_1 & \text{cov} \\ \text{cov} & v_2 \end{bmatrix}$$

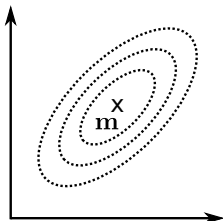
given the following contour plots of $p(\mathbf{x})$?



$$v_1 = v_2, \text{cov} = 0.$$



$$v_1 > v_2, \text{cov} = 0.$$



Effect of parameters on density function

Assuming

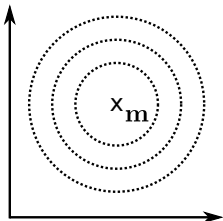
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^\top \mathbf{V}^{-1} (\mathbf{x} - \mathbf{m}) \right\}.$$

The parameter \mathbf{m} determines **mode location** and \mathbf{V} **scales and rotates** the space.

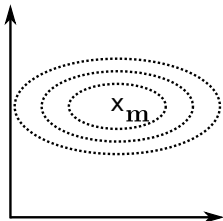
What can you say about

$$\mathbf{V} = \begin{bmatrix} v_1 & \text{cov} \\ \text{cov} & v_2 \end{bmatrix}$$

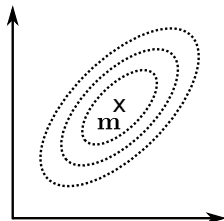
given the following contour plots of $p(\mathbf{x})$?



$$v_1 = v_2, \text{cov} = 0.$$



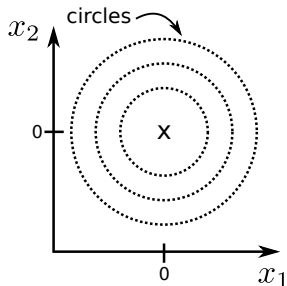
$$v_1 > v_2, \text{cov} = 0.$$



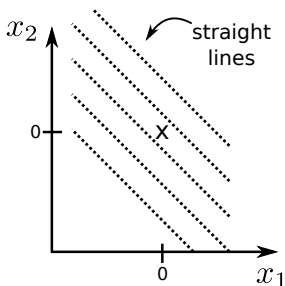
$$\text{cov} > 0.$$

Gaussian quiz

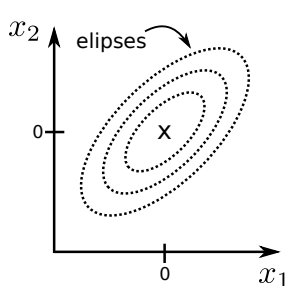
A)



B)



C)



Match each exponential of quadratic function with its contour plot.

i) $\exp \left\{ -\frac{1}{2}x_1^2 a - \frac{1}{2}x_2^2 a \right\} ,$

ii) $\exp \left\{ -\frac{1}{2}x_1^2 a - \frac{1}{2}x_2^2 b + x_1 x_2 c \right\} ,$

iii) $\exp \left\{ -\frac{1}{2}(y - x_1 - x_2)^2 \right\} ,$

Linear combination of Gaussian random variables

Let $p(\mathbf{x}) = \mathbf{N}(\mathbf{x}|\mathbf{0}, \mathbf{V}_1)$ and $p(\mathbf{e}) = \mathbf{N}(\mathbf{x}|\mathbf{0}, \mathbf{V}_2)$ and assume that, for a matrix \mathbf{W} ,

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{e}.$$

What is $p(\mathbf{y})$? Linear combinations of Gaussian random variables are Gaussian.

Therefore, $p(\mathbf{y})$ is Gaussian with mean vector

$$\mathbf{m}_3 = \mathbf{E}[\mathbf{y}] = \mathbf{E}[\mathbf{W}\mathbf{x} + \mathbf{e}] = \mathbf{W}\mathbf{E}[\mathbf{x}] + \mathbf{E}[\mathbf{e}] = \mathbf{0}$$

and covariance matrix

$$\begin{aligned}\mathbf{V}_3 &= \mathbf{E}[\mathbf{y}\mathbf{y}^T] - \mathbf{E}[\mathbf{y}]\mathbf{E}[\mathbf{y}]^T \\ &= \mathbf{E}[\mathbf{y}\mathbf{y}^T] \\ &= \mathbf{E}[(\mathbf{W}\mathbf{x} + \mathbf{e})(\mathbf{W}\mathbf{x} + \mathbf{e})^T] \\ &= \mathbf{E}[\mathbf{W}\mathbf{x}\mathbf{x}^T\mathbf{W}^T + \mathbf{e}\mathbf{x}^T\mathbf{W}^T + \mathbf{W}\mathbf{x}\mathbf{e}^T + \mathbf{e}\mathbf{e}^T] \\ &= \mathbf{W}\mathbf{V}_1\mathbf{W}^T + \mathbf{V}_2.\end{aligned}$$

What if $p(\mathbf{x}) = \mathbf{N}(\mathbf{x}|\mathbf{m}_1, \mathbf{V}_1)$?

Completing the square

Let

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^\top \mathbf{V}^{-1} (\mathbf{x} - \mathbf{m}) \right\} ,$$

$$q(\mathbf{x}) = \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \mathbf{P} \mathbf{x} + \mathbf{a}^\top \mathbf{x} \right\} .$$

Assume that $p(\mathbf{x}) \propto q(\mathbf{x})$.

Write \mathbf{m} and \mathbf{V} in terms of \mathbf{P} and \mathbf{a} .

We have that

$$p(\mathbf{x}) \propto \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \mathbf{V}^{-1} \mathbf{x} + \mathbf{m}^\top \mathbf{V}^{-1} \mathbf{x} \right\} \quad \text{from equation (1) ,}$$

$$q(\mathbf{x}) = \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \mathbf{P} \mathbf{x} + \mathbf{a}^\top \mathbf{x} \right\} .$$

Therefore, $\mathbf{V} = \mathbf{P}^{-1}$ and $\mathbf{m} = \mathbf{V} \mathbf{a}$.

What is the normalization constant of $q(\mathbf{x})$?

Product of Gaussian densities

Let $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_1, \mathbf{V}_1)$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_2, \mathbf{V}_2)$. What is $t(\mathbf{x}) \propto p(\mathbf{x})q(\mathbf{x})$?

$t(\mathbf{x})$ is Gaussian $\mathcal{N}(\mathbf{x}|\mathbf{m}_3, \mathbf{V}_3)$ because the product of exponentials of quadratic functions is also the exponential of a quadratic function. What are \mathbf{m}_3 and \mathbf{V}_3 ?

$$\begin{aligned} p(\mathbf{x})q(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}_1, \mathbf{V}_1)\mathcal{N}(\mathbf{x}|\mathbf{m}_2, \mathbf{V}_2) \\ &= \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}_1|}} \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \mathbf{V}_1^{-1} \mathbf{x} + \mathbf{m}_1^\top \mathbf{V}_1^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m}_1^\top \mathbf{V}_1^{-1} \mathbf{m}_1 \right\} \\ &\quad \frac{1}{\sqrt{(2\pi)^D |\mathbf{V}_2|}} \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \mathbf{V}_2^{-1} \mathbf{x} + \mathbf{m}_2^\top \mathbf{V}_2^{-1} \mathbf{x} - \frac{1}{2} \mathbf{m}_2^\top \mathbf{V}_2^{-1} \mathbf{m}_2 \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \underbrace{(\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1})}_{\mathbf{V}_3^{-1}} \mathbf{x} + \underbrace{(\mathbf{m}_1^\top \mathbf{V}_1^{-1} + \mathbf{m}_2^\top \mathbf{V}_2^{-1})}_{\mathbf{m}_3^\top \mathbf{V}_3^{-1}} \mathbf{x} \right\}. \end{aligned}$$

Therefore, $\mathbf{V}_3 = (\mathbf{V}_1^{-1} + \mathbf{V}_2^{-1})^{-1}$ and $\mathbf{m}_3 = \mathbf{V}_3(\mathbf{m}_1^\top \mathbf{V}_1^{-1} + \mathbf{m}_2^\top \mathbf{V}_2^{-1})^\top$.

What is the normalization constant of $p(\mathbf{x})q(\mathbf{x})$?

Bayesian linear regression

Consider a regression model in which σ^2 is known: the only unknown is \mathbf{w} .

Recall that the likelihood function under Gaussian noise is

$$p(\mathbf{y}|\tilde{\mathbf{X}}, \mathbf{w}) = \prod_{n=1}^N \mathcal{N}(y_n | \mathbf{w}^T \tilde{\mathbf{x}}_n, \sigma^2) \propto \exp \left\{ -\frac{\mathbf{w}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{w}}{2\sigma^2} + \frac{\mathbf{y}^T \tilde{\mathbf{X}} \mathbf{w}}{\sigma^2} \right\}.$$

We choose the prior for \mathbf{w} to be a zero-mean isotropic Gaussian:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \lambda^{-1} \mathbf{I}) \propto \exp \left\{ -\frac{1}{2} \mathbf{x}^T \lambda \mathbf{x} \right\}.$$

The posterior is then Gaussian:

$$\begin{aligned} p(\mathbf{w} | \mathbf{y}, \tilde{\mathbf{X}}, \sigma^2) &\propto p(\mathbf{y} | \tilde{\mathbf{X}}, \mathbf{w}) p(\mathbf{w}) \\ &\propto \exp \left\{ -\frac{1}{2} \mathbf{w}^T \underbrace{\left(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \sigma^{-2} + \lambda \mathbf{I} \right)}_{\mathbf{V}^{-1}} \mathbf{w} + \underbrace{\mathbf{y}^T \tilde{\mathbf{X}} \sigma^{-2}}_{\mathbf{m}^T \mathbf{V}^{-1}} \mathbf{w} \right\}. \end{aligned}$$

Therefore, $p(\mathbf{w} | \mathbf{y}, \tilde{\mathbf{X}}, \sigma^2) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{V})$ where

$$\mathbf{V} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \sigma^{-2} + \lambda \mathbf{I})^{-1}, \quad \mathbf{m} = \mathbf{V} \sigma^{-2} \tilde{\mathbf{X}}^T \mathbf{y}.$$

The Bayesian predictive distribution

The predictive distribution for the y_* of a given new corresponding $\tilde{\mathbf{x}}_*$ is

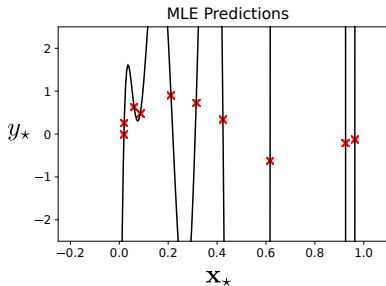
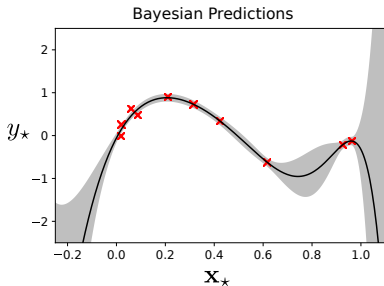
$$p(y_*|\tilde{\mathbf{x}}_*, \mathbf{y}, \tilde{\mathbf{X}}) = \int p(y_*|\tilde{\mathbf{x}}_*, \mathbf{w})p(\mathbf{w}|\mathbf{y}, \tilde{\mathbf{X}}) d\mathbf{w} = \int \mathcal{N}(y_*|\mathbf{w}^T\tilde{\mathbf{x}}_*, \sigma^2)\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{V}) d\mathbf{w}.$$

We have that $y_* = \mathbf{w}^T\tilde{\mathbf{x}}_* + e_*$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$ and $e_* \sim \mathcal{N}(0, \sigma^2)$. Thus,

$$p(y_*|\tilde{\mathbf{x}}_*, \mathbf{y}, \tilde{\mathbf{X}}) = \mathcal{N}(y_*|m_*, v_*),$$

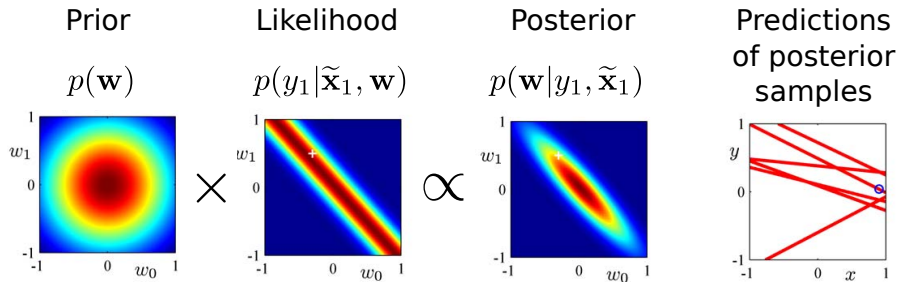
where $m_* = \mathbf{m}^T\tilde{\mathbf{x}}_*$ and $v_* = \tilde{\mathbf{x}}_*^T\mathbf{V}\tilde{\mathbf{x}}_* + \sigma^2$.

Example with polynomial basis functions, $M = 10$, $\lambda = 10^{-5}$, $\sigma^2 = 0.005$:



We reduce **overfitting** and obtain **confidence bands** $m_* \pm v_*^{1/2}$ in our predictions!

Example



Another example with Gaussian basis functions

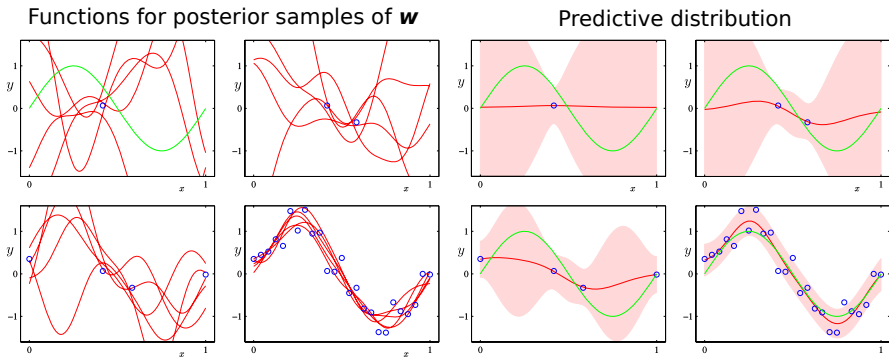


Figure: C. Bishop. *Pattern Recognition and Machine Learning*, 2006.

Maximum a posteriori (MAP) inference

Assumes that the posterior is well approximated by a point mass at its mode:



In particular,

$$\begin{aligned} p(y_\star|\tilde{\mathbf{x}}_\star, \mathbf{y}, \tilde{\mathbf{X}}) &= \int p(y_\star|\tilde{\mathbf{x}}_\star, \mathbf{w}) p(\mathbf{w}|\mathbf{y}, \tilde{\mathbf{X}}) d\mathbf{w} & \mathbf{w}_{\text{MAP}} &= \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}, \tilde{\mathbf{X}}) \\ p(y_\star|\tilde{\mathbf{x}}_\star, \mathbf{y}, \tilde{\mathbf{X}}) &\approx \int p(y_\star|\tilde{\mathbf{x}}_\star, \mathbf{w}) \delta(\mathbf{w} - \mathbf{w}_{\text{MAP}}) d\mathbf{w} & &= \arg \max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w}, \tilde{\mathbf{X}}) p(\mathbf{w}) \\ &\approx p(y_\star|\tilde{\mathbf{x}}_\star, \mathbf{w}_{\text{MAP}}), & &= \arg \max_{\mathbf{w}} \left\{ \log p(\mathbf{y}|\mathbf{w}, \tilde{\mathbf{X}}) + \log p(\mathbf{w}) \right\}. \end{aligned}$$

MAP inference is a form of **regularized** MLE. For $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})$, we obtain

$$\mathbf{w}_{\text{MAP}} = \arg \max_{\mathbf{w}} \left\{ \log p(\mathbf{y}|\mathbf{w}, \tilde{\mathbf{X}}) - \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \right\} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \sigma^{-2} + \lambda \mathbf{I})^{-1} \sigma^{-2} \tilde{\mathbf{X}}^T \mathbf{y}.$$

MAP inference fails to generate **confidence bands** in the resulting predictions!