

Probabilistic & Unsupervised Learning

Gaussian Processes

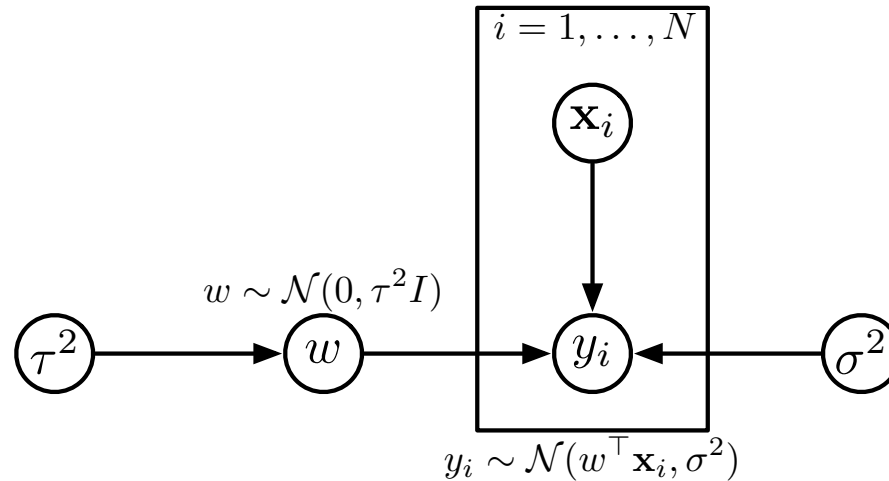
Maneesh Sahani

`maneesh@gatsby.ucl.ac.uk`

**Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London**

Term 1, Autumn 2011

Bayesian Linear Regression



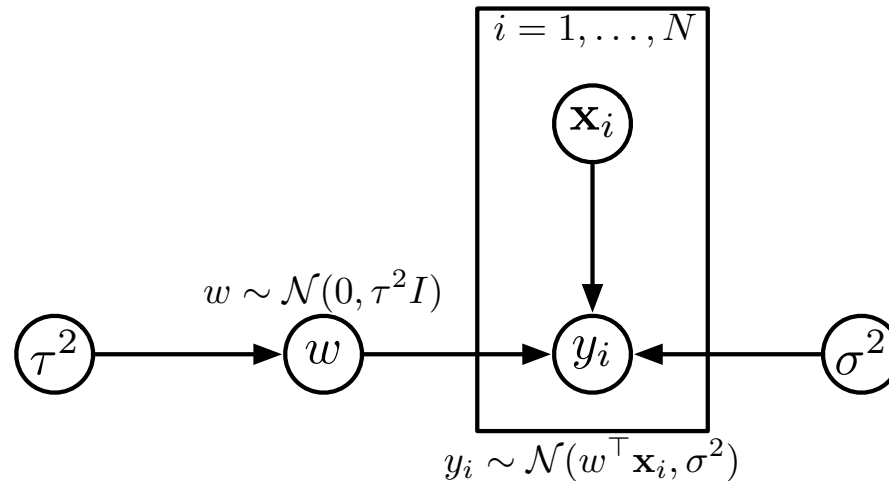
Given observed data $\mathcal{D} = \{X = [\mathbf{x}_1 \dots \mathbf{x}_N], Y = [y_1 \dots y_N]\}$, the **posterior** on \mathbf{w} is:

$$\mathbf{w}|\mathcal{D} \sim \mathcal{N}\left(\underbrace{\frac{1}{\sigma^2} \Sigma_{\mathbf{w}} X Y^\top}_{\mu_{\mathbf{w}}}, \underbrace{\left(\frac{1}{\sigma^2} X X^\top + \frac{1}{\tau^2} I\right)^{-1}}_{\Sigma_{\mathbf{w}}}\right)$$

The **Bayesian** predictive distribution for $y'|\mathbf{x}'$ is obtained by integrating out \mathbf{w} :

$$\begin{aligned} p(y'|\mathbf{x}', \mathcal{D}) &= \int d\mathbf{w} p(y'|\mathbf{w}, \mathbf{x}') p(\mathbf{w}|\mathcal{D}) \\ &= \int d\mathbf{w} \mathcal{N}(y'|\mathbf{w}^\top \mathbf{x}', \sigma^2) \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}) \\ &= \mathcal{N}(\mu_{\mathbf{w}}^\top \mathbf{x}', \mathbf{x}'^\top \Sigma_{\mathbf{w}} \mathbf{x}' + \sigma^2). \end{aligned}$$

Alternative View of Linear Regression



Integrating out \mathbf{w} , the joint distribution of y_1, \dots, y_N given $\mathbf{x}_1, \dots, \mathbf{x}_N$ is Gaussian.

The means and covariances are:

$$E[y_i] = E[\mathbf{w}^\top \mathbf{x}_i] = 0^\top \mathbf{x}_i = 0$$

$$E[(y_i - \bar{y}_i)^2] = E[(\mathbf{x}_i^\top \mathbf{w})(\mathbf{w}^\top \mathbf{x}_i)] + \sigma^2 = \tau^2 \mathbf{x}_i^\top \mathbf{x}_i + \sigma^2$$

$$E[(y_i - \bar{y}_i)(y_j - \bar{y}_j)] = E[(\mathbf{x}_i^\top \mathbf{w})(\mathbf{w}^\top \mathbf{x}_j)] = \tau^2 \mathbf{x}_i^\top \mathbf{x}_j$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \Bigg| \mathbf{x}_1, \dots, \mathbf{x}_N \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 \mathbf{x}_1^\top \mathbf{x}_1 + \sigma^2 & \tau^2 \mathbf{x}_1^\top \mathbf{x}_2 & \cdots & \tau^2 \mathbf{x}_1^\top \mathbf{x}_N \\ \tau^2 \mathbf{x}_2^\top \mathbf{x}_1 & \tau^2 \mathbf{x}_2^\top \mathbf{x}_2 + \sigma^2 & & \tau^2 \mathbf{x}_2^\top \mathbf{x}_N \\ \vdots & & \ddots & \vdots \\ \tau^2 \mathbf{x}_N^\top \mathbf{x}_1 & \tau^2 \mathbf{x}_N^\top \mathbf{x}_2 & \cdots & \tau^2 \mathbf{x}_N^\top \mathbf{x}_N + \sigma^2 \end{bmatrix} \right)$$

$$Y^\top | X \sim \mathcal{N}(0_N, \tau^2 X^\top X + \sigma^2 I_N)$$

Alternative View of Linear Regression

Now, include the test input vector \mathbf{x}' and test output y' :

$$\begin{bmatrix} Y^\top \\ y' \end{bmatrix} \Big| X, \mathbf{x}' \sim \mathcal{N} \left(\begin{bmatrix} 0_N \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 X^\top X + \sigma^2 I & \tau^2 X^\top \mathbf{x}' \\ \tau^2 \mathbf{x}'^\top X & \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 \end{bmatrix} \right)$$

We can find $y'|Y$ by the standard multivariate Gaussian result:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix} \right) \Rightarrow \mathbf{b}|\mathbf{a} \sim \mathcal{N} (C^\top A^{-1} \mathbf{a}, B - C^\top A^{-1} C)$$

So

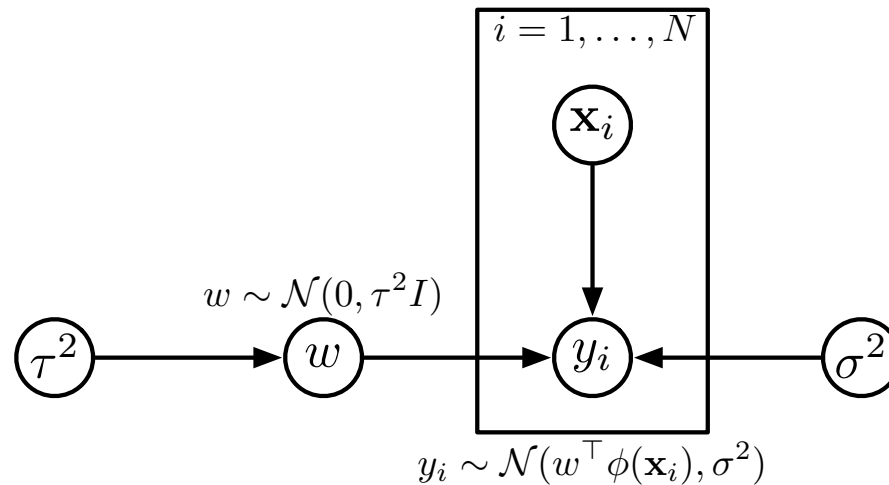
$$\begin{aligned} y'|Y, X, \mathbf{x}' &\sim \mathcal{N} \left(\tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} Y^\top, \tau^2 \mathbf{x}'^\top \mathbf{x}' + \sigma^2 - \tau^2 \mathbf{x}'^\top X (\tau^2 X^\top X + \sigma^2 I)^{-1} \tau^2 X^\top \mathbf{x}' \right) \\ &\sim \mathcal{N} \left(\frac{1}{\sigma^2} \mathbf{x}'^\top \Sigma X Y^\top, \mathbf{x}'^\top \Sigma \mathbf{x}' + \sigma^2 \right) \quad \Sigma = \left(\frac{1}{\sigma^2} X X^\top + \frac{1}{\tau^2} I \right)^{-1} \end{aligned}$$

Same answer as when we integrated posterior over \mathbf{w} to obtain predictive distribution over y' .

Similarly, evidence $P(Y|X)$ is just probability under Gaussian, and reduces to previous expression.

The point: Bayesian regression can be derived from a **joint, parameter-free** distribution on the outputs conditioned on the inputs.

Nonlinear Regression



What if we introduce a nonlinear mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$? Each element of $\phi(\mathbf{x})$ is a (nonlinear) feature extracted from \mathbf{x} . May be many more features than elements on \mathbf{x} .

The regression function $f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x})$ is nonlinear, but outputs Y still jointly Gaussian!

$$Y^\top | X \sim \mathcal{N}(0_N, \tau^2 \Phi^\top \Phi + \sigma^2 I_N)$$

where the i^{th} column of matrix Φ is $\phi(\mathbf{x}_i)$.

Proceeding as before, the predictive distribution over y' on a test input \mathbf{x}' is:

$$y' | Y, X, \mathbf{x}' \sim \mathcal{N}(\tau^2 \phi(\mathbf{x}')^\top \Phi K^{-1} Y^\top, \tau^2 \phi(\mathbf{x}')^\top \phi(\mathbf{x}') + \sigma^2 - \tau^4 \phi(\mathbf{x}')^\top \Phi K^{-1} \Phi^\top \phi(\mathbf{x}'))$$

$$K = \tau^2 \Phi^\top \Phi + \sigma^2 I$$

The Covariance Kernel

$$Y^\top | X \sim \mathcal{N}(\mathbf{0}_N, \tau^2 \Phi^\top \Phi + \sigma^2 I_N)$$

The covariance of the output vector Y plays a central role in the development of the theory of Gaussian processes.

Define the **covariance kernel** function $K : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ such that if $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$ are two input vectors with corresponding outputs y, y' , then

$$K(\mathbf{x}, \mathbf{x}') = \text{Cov}[y, y'] = E[yy'] - E[y]E[y']$$

In the nonlinear regression example we have $K(\mathbf{x}, \mathbf{x}') = \tau^2 \phi(\mathbf{x})^\top \phi(\mathbf{x}') + \sigma^2 \delta_{\mathbf{x}=\mathbf{x}'}$.

The covariance kernel has two properties:

- **Symmetric**: $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$ for all \mathbf{x}, \mathbf{x}' .
- **Positive semidefinite**: the matrix $[K(\mathbf{x}_i, \mathbf{x}_j)]$ formed by any finite set of input vectors $\mathbf{x}_1, \dots, \mathbf{x}_N$ is positive semidefinite.

Theorem: A covariance kernel $K : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is symmetric and positive semidefinite if and only if there is a feature map $\phi : \mathbb{X} \rightarrow \mathbb{H}$ such that

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$$

The feature space \mathbb{H} can potentially be infinite dimensional.

Regression using the Covariance Kernel

For non-linear regression, all operations depended on $K(\mathbf{x}, \mathbf{x}')$ rather than explicitly on $\phi(\mathbf{x})$.

So we can define the joint in terms of K *implicitly* using a (potentially infinite-dimensional) feature map $\phi(\mathbf{x})$.

$$Y|X, K \sim \mathcal{N}(0_N, K(X, X))$$

where the i, j entry in the covariance matrix $K(X, X)$ is $K(\mathbf{x}_i, \mathbf{x}_j)$.

This is called the **kernel trick**.

Prediction: compute the predictive distribution of y' conditioned on Y :

$$y'|x', X, Y, K \sim \mathcal{N}\left(\underbrace{K(x', X)K(X, X)^{-1}Y^\top}_{\text{mean}}, \underbrace{K(x', x') - K(x', X)K(X, X)^{-1}K(X, x')}_{\text{variance}}\right)$$

Evidence: this is just the Gaussian likelihood:

$$P(Y|X, K) = |2\pi K(X, X)|^{-\frac{1}{2}} e^{-\frac{1}{2}YK(X, X)^{-1}Y^\top}$$

Evidence optimisation: the covariance kernel K often has parameters, and these can be optimized by gradient ascent in $\log P(Y|X, K)$.

The Gaussian Process

A **Gaussian process** (GP) is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.

In our regression setting, corresponding to each input vector \mathbf{x} we have an output $f(\mathbf{x})$. Given $X = [\mathbf{x}_1, \dots, \mathbf{x}_N]$, the joint distribution of the outputs $F = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)]$ is:

$$F|X, K \sim \mathcal{N}(0, K(X, X))$$

Thus the random function $f(\mathbf{x})$ (as a collection of random variables, one $f(\mathbf{x})$ for each \mathbf{x}) is a Gaussian process.

In general, a Gaussian process is parametrized by a **mean function** $m(\mathbf{x})$ and **covariance kernel** $K(\mathbf{x}, \mathbf{x}')$, and we write

$$f(\cdot) \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Posterior Gaussian process: on observing X and F , the conditional joint distribution of $F' = [f(\mathbf{x}'_1), \dots, f(\mathbf{x}'_M)]$ on another set of input vectors $\mathbf{x}'_1, \dots, \mathbf{x}'_M$ is still Gaussian:

$$F'|X', X, F, K \sim \mathcal{N}(K(X', X)K(X, X)^{-1}F^\top, K(X', X') - K(X', X)K(X, X)^{-1}K(X, X'))$$

thus the posterior over functions $f(\cdot)|X, F$ is still a Gaussian process!

Regression with Gaussian Processes

We wish to model the joint distribution of outputs y_1, \dots, y_N given inputs $\mathbf{x}_1, \dots, \mathbf{x}_N$.
Use a GP prior over functions:

$$f(\cdot) \sim \mathcal{GP}(0, K(\cdot, \cdot))$$

Usually, instead of treating y_i as direct observation of the function value $f(\mathbf{x}_i)$, we add Gaussian observation noise:

$$y_i | \mathbf{x}_i, f(\cdot) \sim \mathcal{N}(f(\mathbf{x}_i), \sigma^2)$$

Evidence: again this is just a multivariate Gaussian likelihood,

$$P(Y|X) = |2\pi(K(X, X) + \sigma^2 I)|^{-\frac{1}{2}} e^{-\frac{1}{2} Y (K(X, X) + \sigma^2 I)^{-1} Y^\top}$$

Posterior: the posterior function is still a GP,

$$f(\cdot) | X, Y \sim \mathcal{GP}(K(\cdot, X)(K(X, X) + \sigma^2 I)^{-1} Y^\top, K(\cdot, \cdot) - K(\cdot, X)(K(X, X) + \sigma^2 I)^{-1} K(X, \cdot))$$

Prediction: the predictive distribution is just posterior plus observation noise:

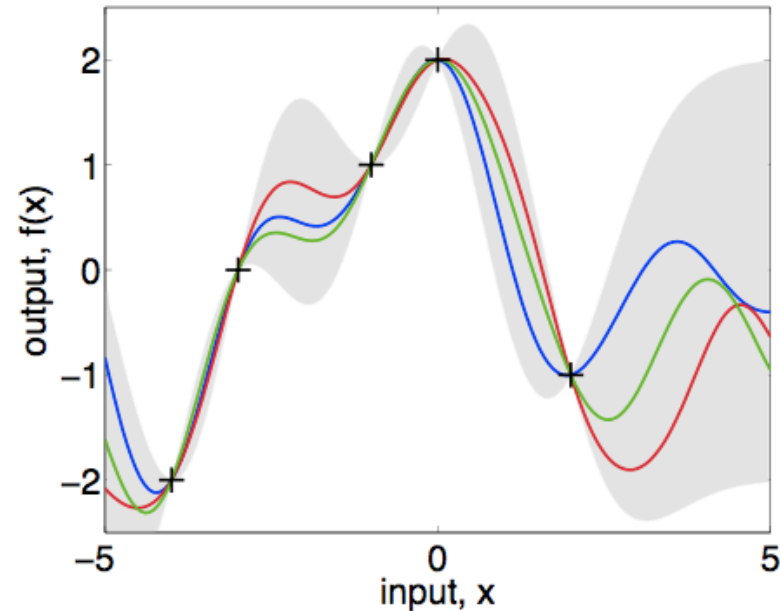
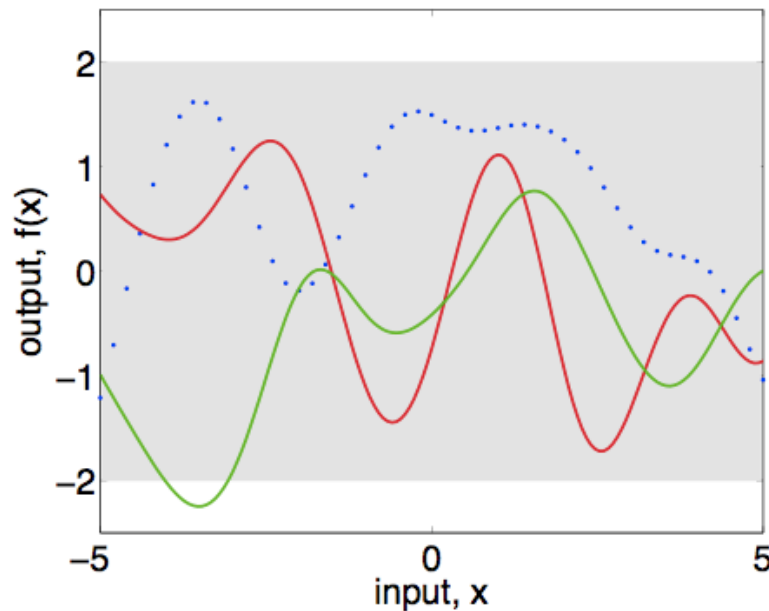
$$y' | X, Y, \mathbf{x}' \sim \mathcal{N}(E[f(\mathbf{x}') | X, Y], \text{Var}[f(\mathbf{x}') | X, Y] + \sigma^2)$$

Evidence Optimisation: we can do this by gradient ascent in $\log P(Y|X)$.

Samples from a Gaussian Process

We can draw sample functions from a GP by fixing a set of input vectors $\mathbf{x}_1, \dots, \mathbf{x}_N$, and drawing a sample $f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)$ from the corresponding multivariate Gaussian. This can then be plotted.

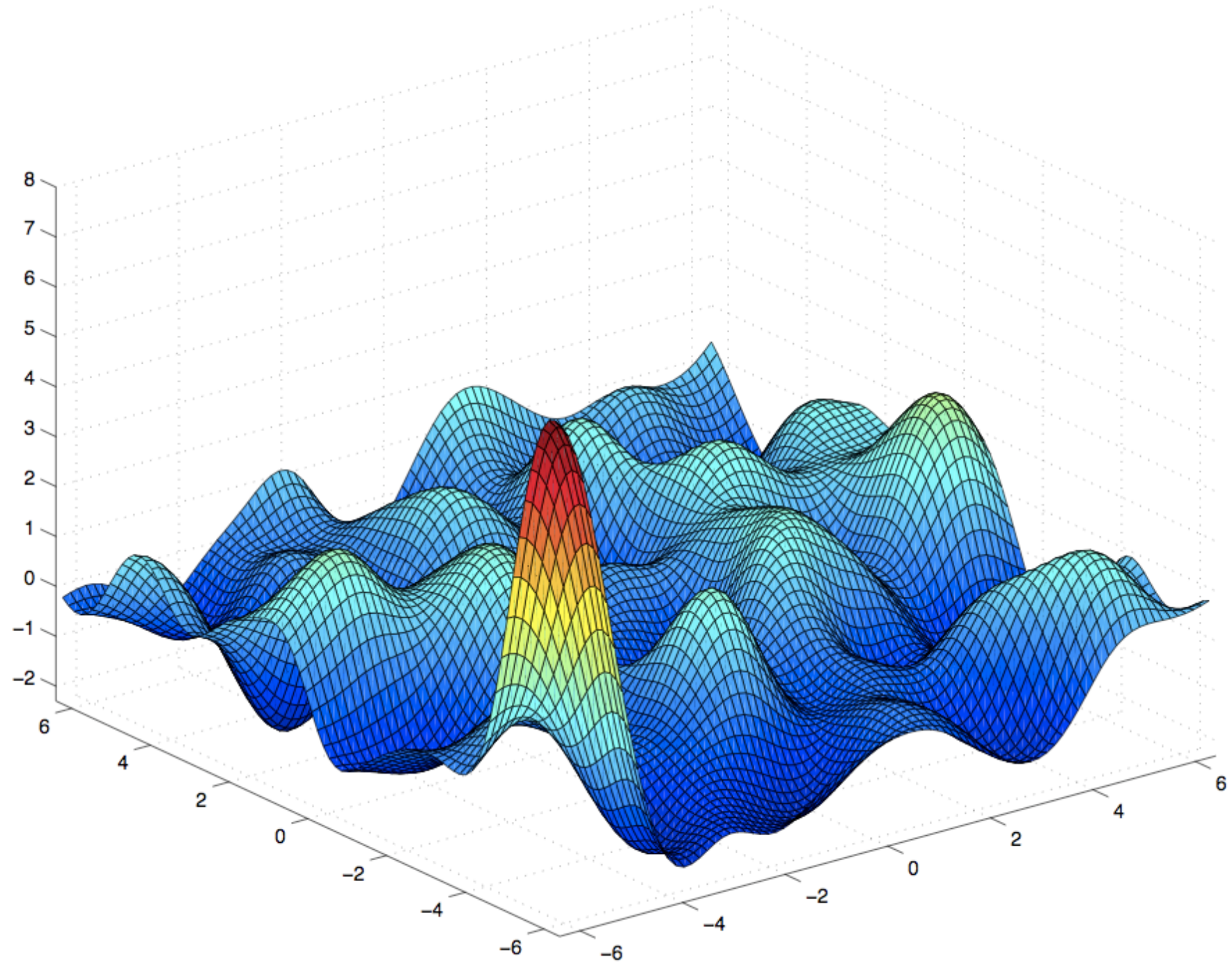
Below we plot samples from an example prior and corresponding posterior GP.



Another approach is to

- sample $f(\mathbf{x}_1)$ first,
- then $f(\mathbf{x}_2)|f(\mathbf{x}_1)$,
- and generally $f(\mathbf{x}_n)|f(\mathbf{x}_1), \dots, f(\mathbf{x}_{n-1})$ for $n = 1, 2, \dots$

Sample from a 2D Gaussian Process



Covariance Kernels

Examples of covariance kernels:

- Polynomial:

$$K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^\top \mathbf{x}')^m \quad m = 1, 2, \dots$$

- Squared-exponential:

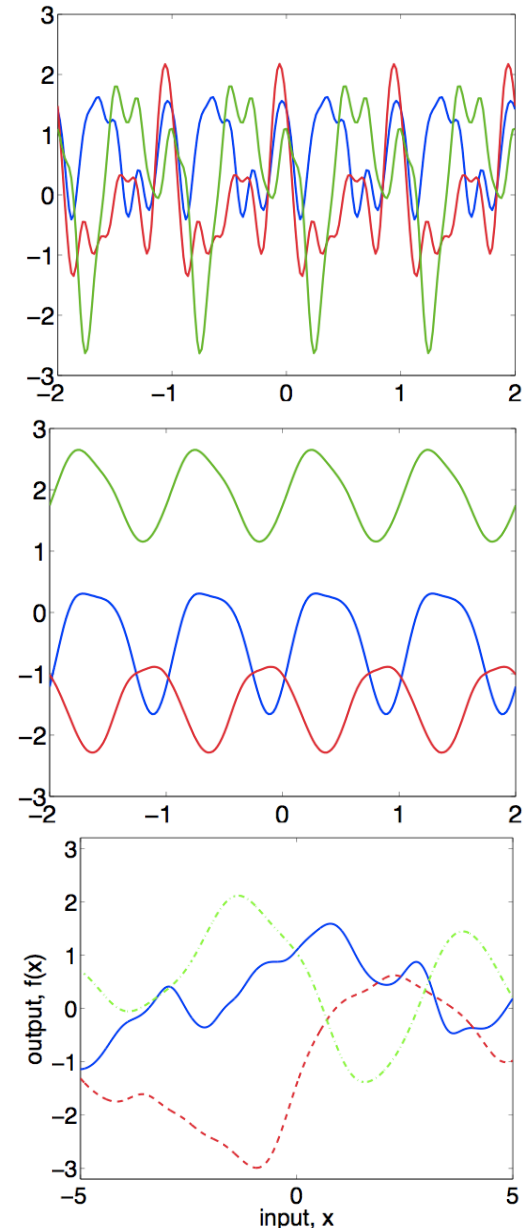
$$K(\mathbf{x}, \mathbf{x}') = \theta^2 e^{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\eta^2}}$$

- Periodic:

$$K(x, x') = \theta^2 e^{-\frac{2 \sin^2(\pi(x-x')/\tau)}{\eta^2}}$$

- Rational Quadratic:

$$K(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\alpha\eta^2}\right)^{-\alpha} \quad \alpha > 0$$



Covariance Kernels

If K_1 and K_2 are covariance kernels, then so are:

- Rescaling: αK_1 for $\alpha > 0$.
- Addition: $K_1 + K_2$
- Elementwise product: $K_1 K_2$
- Mapping: $K_1(\phi(\mathbf{x}), \phi(\mathbf{x}'))$ for some function ϕ .

We say a covariance kernel is translation-invariant if

$$K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$$

A GP with a translation-invariant covariance kernel is stationary: if $f(\cdot) \sim \mathcal{GP}(0, K)$, then so is $f(\cdot - \mathbf{x}) \sim \mathcal{GP}(0, K)$ for each \mathbf{x} .

We say a covariance kernel is radial if

$$K(\mathbf{x}, \mathbf{x}') = h(\|\mathbf{x} - \mathbf{x}'\|)$$

A GP with a radial covariance kernel is stationary with respect to translations, rotations, and reflections of the input space.

Nonparametric Bayesian Models and Occam's Razor

Overparameterised models can **overfit**.

But the Bayesian treatment integrates parameters out, so they cannot be adjusted to overfit the data! In the GP, the parameter is the function $f(\mathbf{x})$ which can be infinite-dimensional.

The Gaussian process is an example of a larger class of **nonparametric Bayesian models**.

- Infinite number of parameters.
- Often constructed as the infinite limit of a nested family of finite models (sometimes equivalent to infinite model averaging).
- Parameters integrated out, so effective number of parameters to overfit is zero or small (hyperparameters).
- No need for model selection. Bayesian posterior on parameters will concentrate on “sub-model” with largest integral automatically.
- No explicit need for Occam's razor, validation or added regularisation penalty.

End Notes

Automatic relevance determination appeared in MacKay (1993) [Bayesian Methods for Back-propagation Networks](#) and Neal (1993) [Bayesian Learning for Neural Networks](#).

Gaussian processes can also be used in classification and latent variable models. We will consider classification in the second half of course.

Many of the figures have been copied from a Gaussian process tutorial by Carl Rasmussen (MLSS 2007) at <http://agbs.kyb.tuebingen.mpg.de/wikis/mlss07/CarlERasmussen>

An excellent text book on Gaussian processes is [Gaussian processes for Machine Learning](#) by Rasmussen and Williams, available online at <http://www.gaussianprocess.org/gpml/>

The original paper on Gaussian process latent variable models is by Neil Lawrence (NIPS 2004) at <http://www.cs.man.ac.uk/~neill/>

End Notes