Introduction to Gaussian Processes

David Dalton

July 22, 2020

Department of Mathematics and Statistics
University of Glasgow
d.dalton.1@research.gla.ac.uk

Overview of Probabilistic Modelling

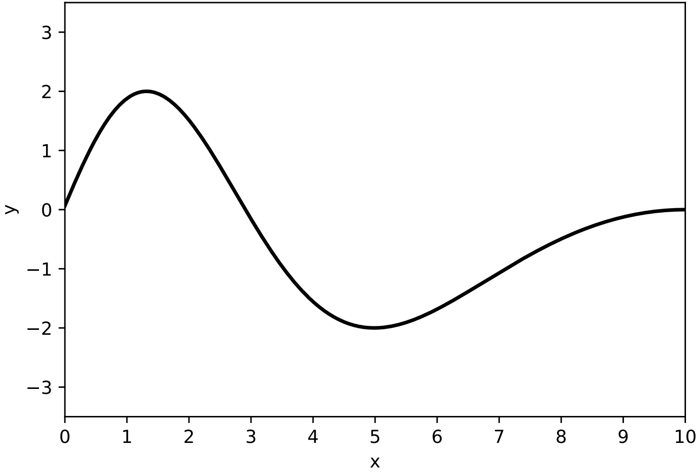
The probabilistic, or Bayesian, approach to modelling is conceptually simple and consists of two primary steps:

1. We first specify a joint distribution over all variables we are interested in modelling. This includes latent variables $\mathbf{z}$, which we cannot directly observe, and variables $\mathbf{x}$ for which we can collect data. We write this distribution as $p(\mathbf{z}, \mathbf{x})$.
2. Once we observe data for $\mathbf{x}$, the next step is to perform *inference*:

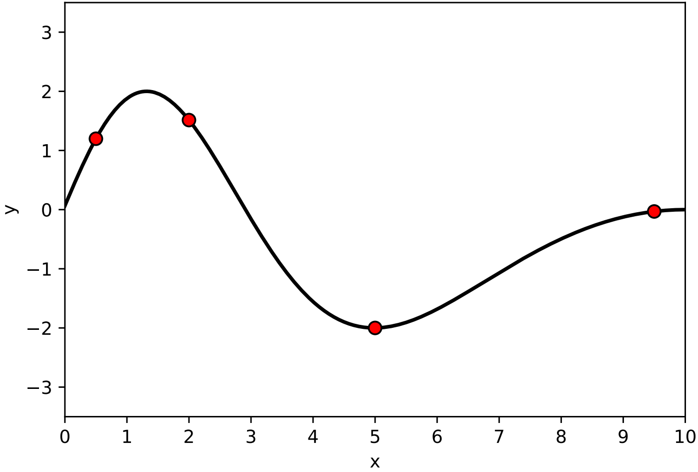
$$p(\mathbf{z} \,|\, \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{z}, \mathbf{x})}{\int p(\mathbf{z}, \mathbf{x}) d\mathbf{z}}$$

For many models, the integral above is not available in closed form and obtaining a numerical solution is infeasible. Instead, approximate inference methods, such as Variational Inference or Markov Chain Monte Carlo must be used.

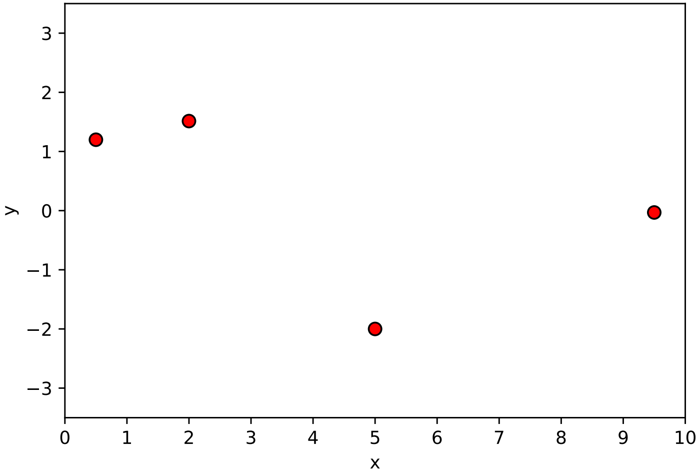
Regression Example



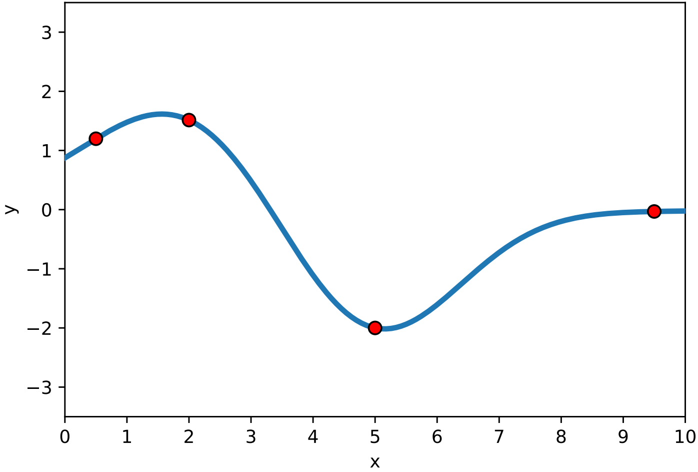
Regression Example



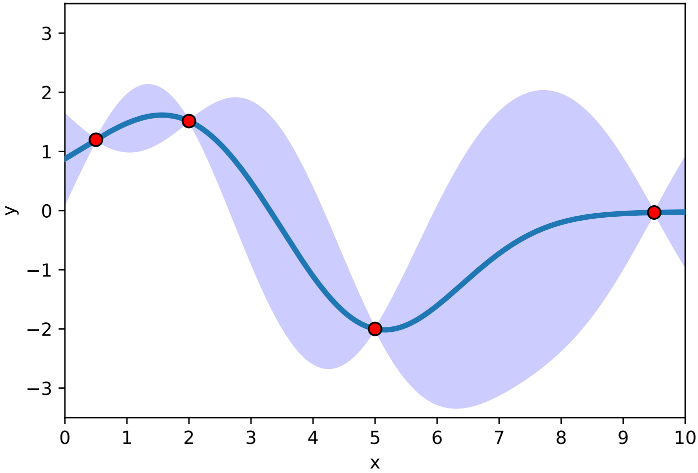
Regression Example



Regression Example



Regression Example



Gaussian Process Definition

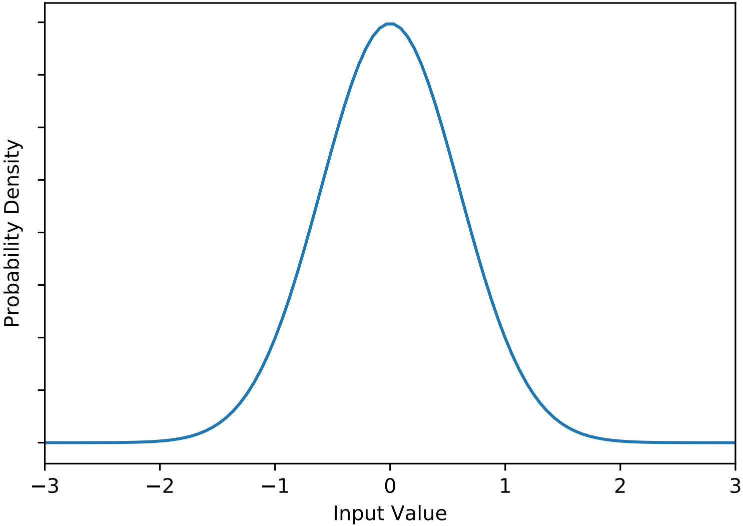
A *Gaussian process* (GP) is a collection of random variables, indexed by some set $\mathcal{X}$, such that any finite number of the variables follow a (multivariate) Gaussian distribution.

A GP is completely specified by its mean function $m(x)$, and covariance function, $k(x, x')$, where $x, x' \in \mathcal{X}$. We denote the GP as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

A Gaussian process is a distribution over *functions*; that is if, we sample from a Gaussian process, we get a function.

The Bell Curve



The Gaussian Distribution

A univariate Gaussian distribution is fully specified by two numbers: its mean μ, and its variance, σ^2:

$$p(x) = (2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

The Gaussian Distribution

A univariate Gaussian distribution is fully specified by two numbers: its mean μ, and its variance, σ^2:

$$p(x) = (2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

A multivariate Gaussian distribution is fully specified by its mean vector $\mathbf{m}$ and covariance matrix $\mathbf{K}$:

$$p(\mathbf{f}) = (2\pi)^{-\frac{D}{2}}|\mathbf{K}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{f}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{f}-\mathbf{m})\right]$$

$$\mathbf{f} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

Gaussian Processes versus Gaussian Distributions

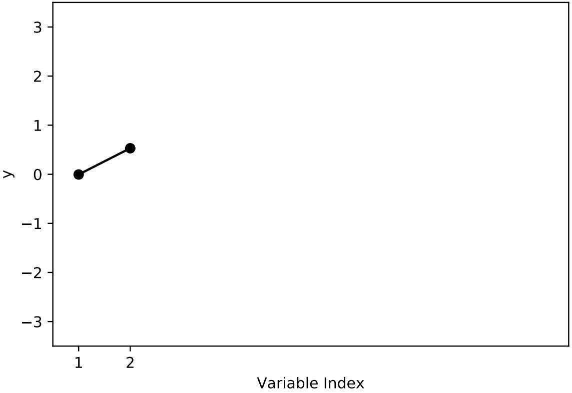
A Gaussian *process* is completely specified by its mean function $m(x)$, and covariance function, $k(x, x')$. We denote the GP as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

A multivariate Gaussian *distribution* is fully specified by its mean vector $\mathbf{m}$ and covariance matrix $\mathbf{K}$. We denote the Gaussian distribution as:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

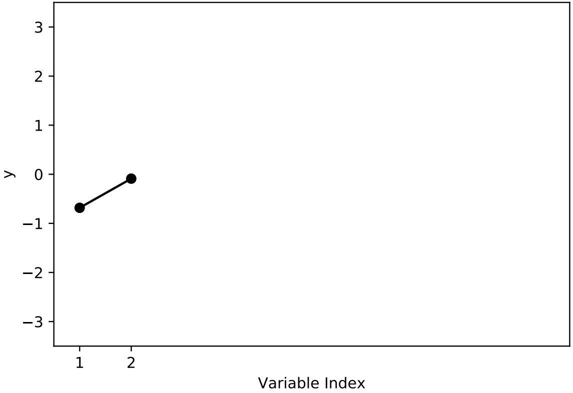
Visualising Multivariate Gaussian Samples



$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

$$\mathbf{K} = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}$$

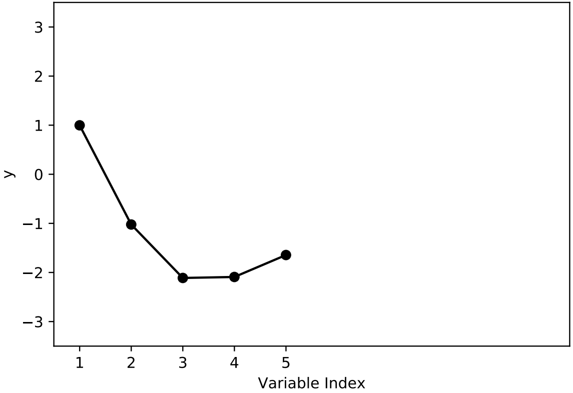
2D Gaussian Samples



$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

$$\mathbf{K} = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}$$

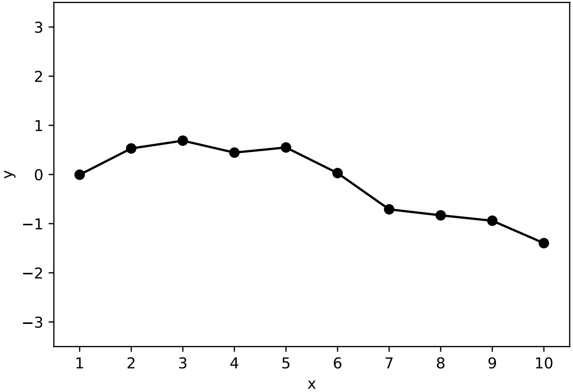
5D Gaussian Samples



$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

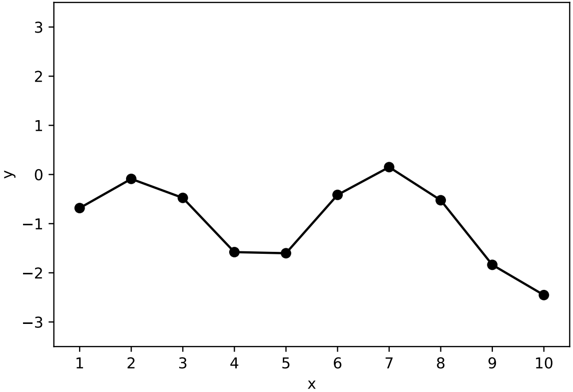
$$\mathbf{K} = \begin{bmatrix} 1 & .8 & .4 & .2 & .1 \\ .8 & 1 & .8 & .4 & .2 \\ .4 & .8 & 1 & .8 & .4 \\ .2 & .4 & .8 & 1 & .8 \\ .1 & .2 & .4 & .8 & 1 \end{bmatrix}$$

10D Gaussian Samples



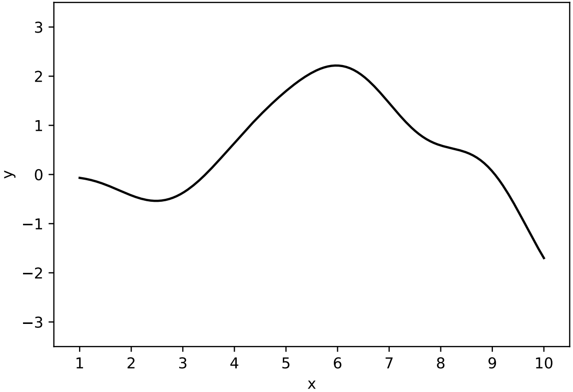
$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

10D Gaussian Samples



$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

Gaussian Process Samples



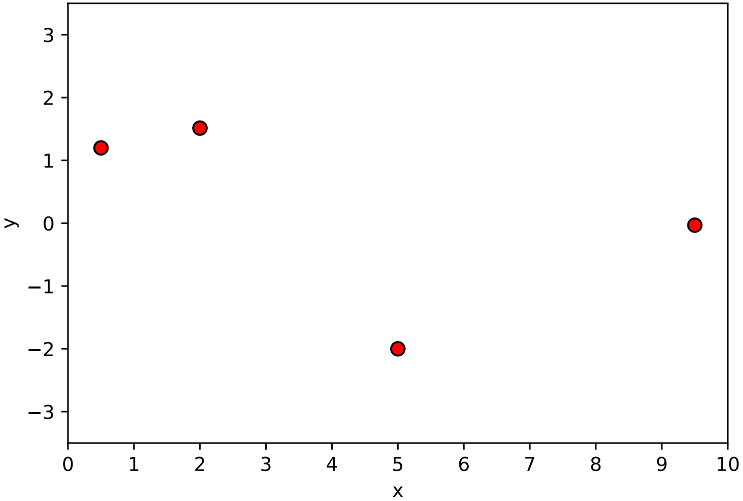
$$f(x) \sim \mathcal{GP}(0, k(x, x'))$$

Regression with Gaussian Processes

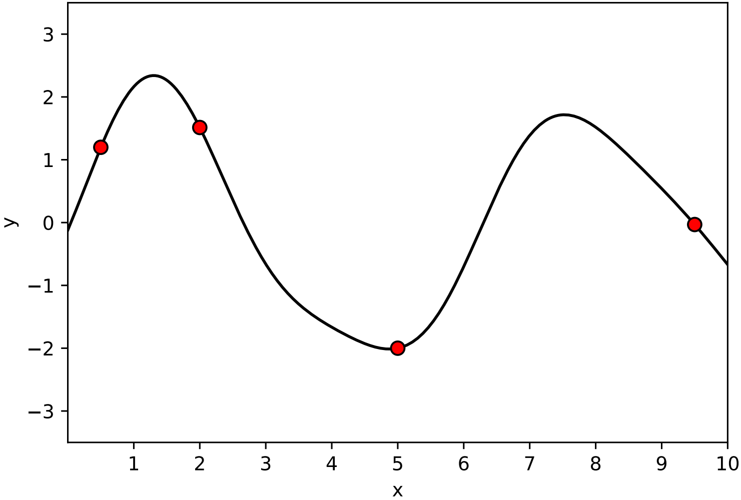
We can use Gaussian processes for regression by making use of the following very useful property of the multivariate Gaussian distribution:

If we observe the values of some of the variables of the multivariate Gaussian, the conditional distribution of the unobserved variables, given our observations, is again a Gaussian with known mean and known covariance.

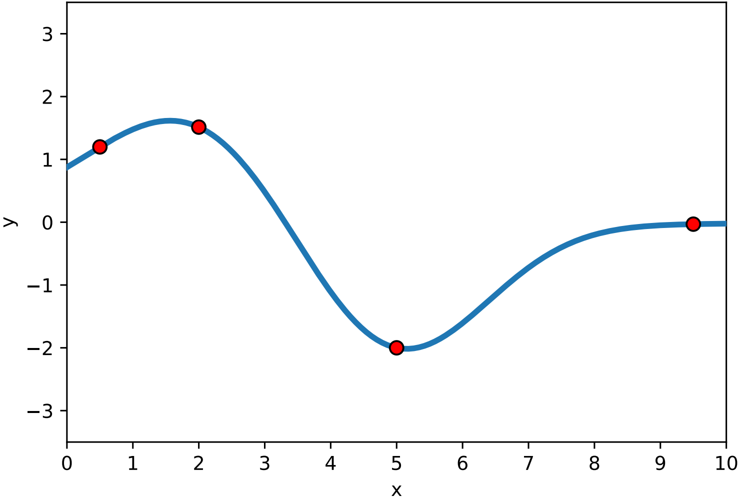
Regression with Gaussian Processes



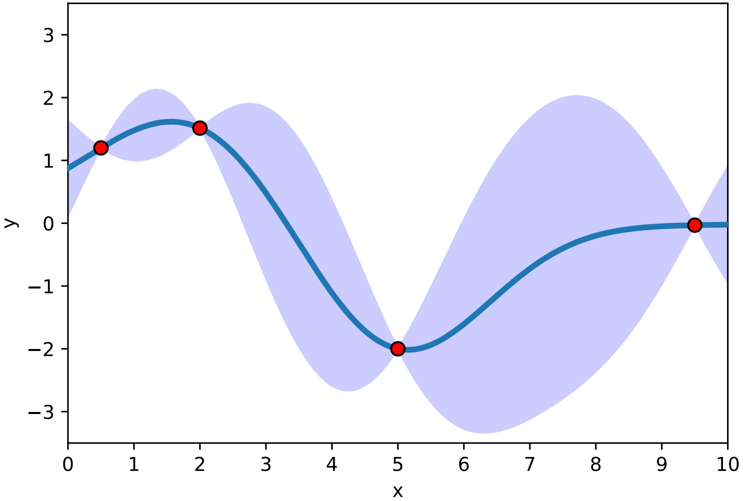
Regression with Gaussian Processes



Regression with Gaussian Processes



Regression with Gaussian Processes



Gaussian Process Definition Revisited

A *Gaussian process* (GP) is a collection of random variables, indexed by some set $\mathcal{X}$, such that any finite number of the variables follow a (multivariate) Gaussian distribution.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

Gaussian Process Definition Revisited

A *Gaussian process* (GP) is a collection of random variables, indexed by some set $\mathcal{X}$, such that any finite number of the variables follow a (multivariate) Gaussian distribution.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

Suppose you observe a finite collection of variables from the process $\mathcal{D} = \{(x_i, f_i)_{i=1}^N\}$. Then, the joint distribution over $\mathbf{f} = (f_1, f_2, ..., f_N)^T$ marginalises to a multivariate Gaussian distribution:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

Gaussian Process Definition Revisited

A *Gaussian process* (GP) is a collection of random variables, indexed by some set $\mathcal{X}$, such that any finite number of the variables follow a (multivariate) Gaussian distribution.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

Suppose you observe a finite collection of variables from the process $\mathcal{D} = \{(x_i, f_i)_{i=1}^{N}\}$. Then, the joint distribution over $\mathbf{f} = (f_1, f_2, ..., f_N)^T$ marginalises to a multivariate Gaussian distribution:

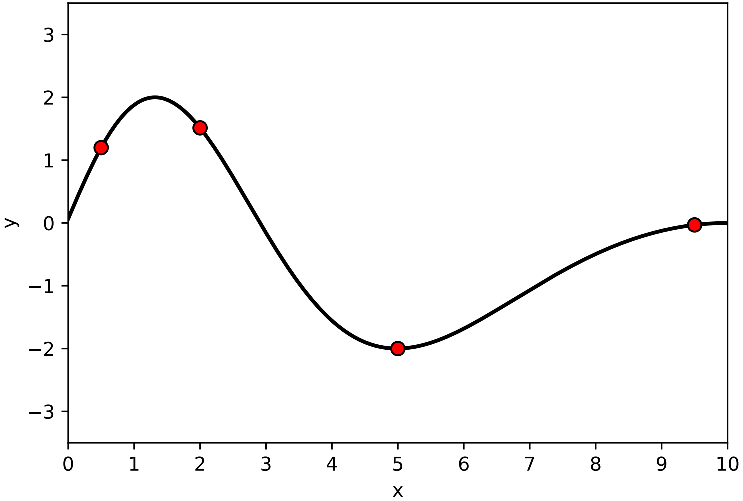
$$\mathbf{f} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

where $\mathbf{m}_i = \mathbb{E}(f_i) = m(x_i)$ and $\mathbf{K}_{ij} = \mathrm{Cov}(f_i, f_j) = k(x_i, x_j)$

Regression with Gaussian Processes

Suppose we have a latent function $f(x)$, for which we have observed some values $\mathbf{f}$, and we wish to predict the function values $\mathbf{f}^{*}$ at some test points of interest.

Regression with Gaussian Processes



Regression with Gaussian Processes

Suppose we have a latent function $f(x)$, for which we have observed some values $\mathbf{f}$, and we wish to predict the function values $\mathbf{f}^*$ at some test points of interest.

We can do this by assuming that our underlying function was drawn from a Gaussian process, with specified mean function $m(x)$ and covariance function $k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

With this assumption, we can then follow the two step probabilistic modelling framework from the first slide to get a probability distribution over $\mathbf{f}^*$.

It is important that we select and tune our mean and covariance functions respectively to best model our underlying function.

Regression with Gaussian Processes

1. We first write down the joint distribution over **f** and **f***:

$$p(\mathbf{f}^*, \mathbf{f}) = p\begin{pmatrix} \mathbf{f}^* \\ \mathbf{f} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{**} & \mathbf{K}_{*N} \\ \mathbf{K}_{N*} & \mathbf{K}_{NN} \end{bmatrix}\right)$$

Regression with Gaussian Processes

1. We first write down the joint distribution over **f** and **f***:

$$p(\mathbf{f}^*, \mathbf{f}) = p\begin{pmatrix} \mathbf{f}^* \\ \mathbf{f} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{**} & \mathbf{K}_{*N} \\ \mathbf{K}_{N*} & \mathbf{K}_{NN} \end{bmatrix}\right)$$

2. The second step is to perform inference:

$$p(\mathbf{f}^* \,|\, \mathbf{f}) = \frac{p(\mathbf{f}^*, \mathbf{f})}{p(\mathbf{f})} = \frac{p(\mathbf{f}^*, \mathbf{f})}{\int p(\mathbf{f}^*, \mathbf{f}) d\mathbf{f}^*}$$

$$p(\mathbf{f}) = \int p(\mathbf{f}^*, \mathbf{f}) d\mathbf{f}^* = \mathcal{N}(\mathbf{0}, \mathbf{K}_{NN})$$

Regression with Gaussian Processes

1. We first write down the joint distribution over **f** and $\mathbf{f}^*$:

$$p(\mathbf{f}^*, \mathbf{f}) = p\begin{pmatrix} \mathbf{f}^* \\ \mathbf{f} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{**} & \mathbf{K}_{*N} \\ \mathbf{K}_{N*} & \mathbf{K}_{NN} \end{bmatrix}\right)$$

2. The second step is to perform inference:

$$p(\mathbf{f}^* \mid \mathbf{f}) = \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$$

$$\begin{aligned} \boldsymbol{\mu}_* &= \mathbf{K}_{N*}^{\mathrm{T}} \mathbf{K}_{NN}^{-1} \mathbf{f} \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_{N*}^{\mathrm{T}} \mathbf{K}_{NN}^{-1} \mathbf{K}_{N*} \end{aligned}$$

Modelling Noisy Data

In most regression applications, we observe a dataset of input-output pairs $\mathcal{D} = \{(x_i, y_i)_{i=1}^N\}$ where the outputs are "noisy" in some sense, and therefore we do not want to interpolate their values.

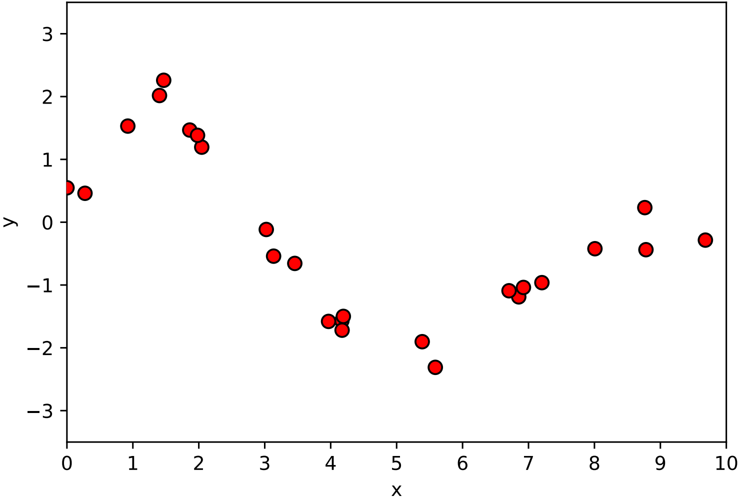
In the GP framework, we model noisy data by assuming that the observed outputs were generated by a latent function plus a random noise term

$$y_i = f(x_i) + \epsilon_i$$

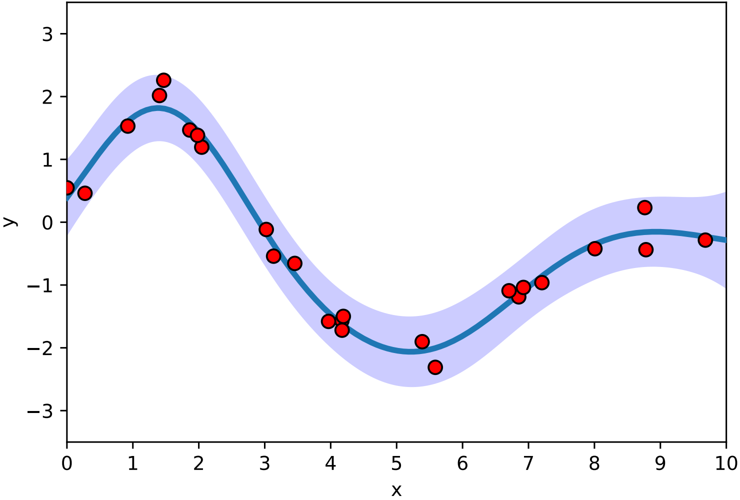
We model the latent function with a GP, and then assume that the errors follow some specified probability distribution.

For regression, errors are commonly assumed to be independently and identically Gaussian distributed. In this case, because the latent function is Gaussian, and the errors are Gaussian, our posterior predictive distribution over a finite set of test outputs of interest is again Gaussian.

Modelling Noisy Data



Modelling Noisy Data



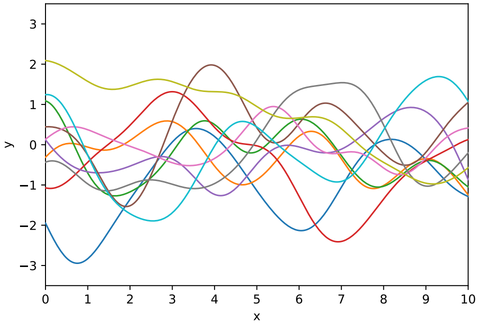
Choice of Covariance Function

We need to select a covariance function in order to perform Gaussian process regression.

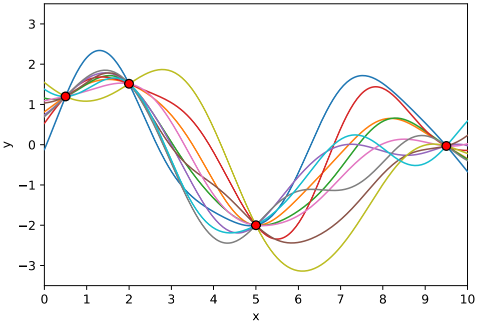
The choice of covariance function allows us to specify our beliefs about the properties of the true underlying function f(x) we wish to model.

Squared Exponential Covariance Function

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$



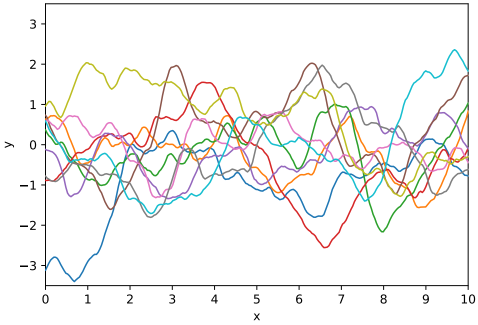
Prior Samples



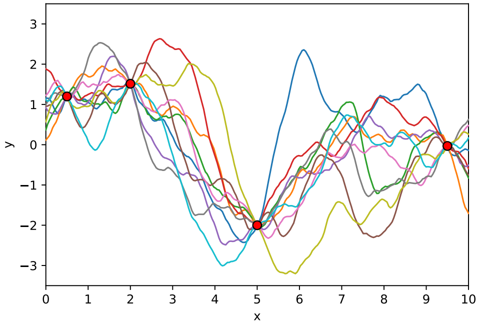
Posterior Samples

Matern 3/2 Covariance Function

$$k(x, x') = \sigma_f^2 \left(1 + \frac{\sqrt{3(x - x')}}{\sigma_l}\right) \exp\left(-\frac{\sqrt{3(x - x')}}{\sigma_l}\right)$$



Prior Samples



Posterior Samples

Hyperparameter Selection

Having chosen a covariance function, we must then select its *hyperparameters*.

For example, the squared-exponential covariance function has two hyperparameters, σ^2 and ℓ^2:

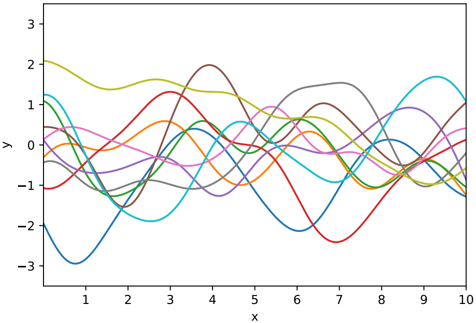
$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

The most common approach is to find a point estimate of these hyperparameters by maximising the marginal likelihood of the training data, $p(\mathbf{f})$.

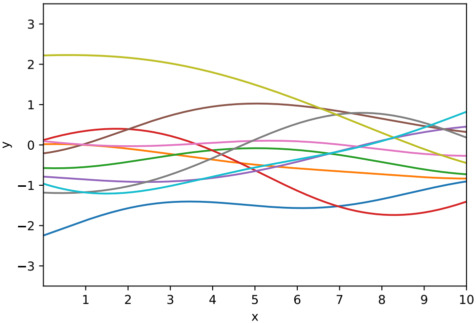
It turns out that tuning the hyperparameters in this way allows us to balance fitting the training data against model complexity.

Impact of Covariance Function Hyperparameters

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$



$\sigma^2 = 1,\ \ell^2 = 1$



$\sigma^2 = 1,\ \ell^2 = 4$

Implementing Gaussian Process Regression

The standard approach to GP regression shown here is not appropriate for use on large datasets, because training times grow cublicly with the number of training data points, N.

To see why this is the case, recall the equations for the posterior mean $\boldsymbol{\mu}_*$ and posterior covariance $\boldsymbol{\Sigma}_*$ of the GP:

$$\boldsymbol{\mu}_* = \mathbf{K}_{N*}^{\mathrm{T}} \mathbf{K}_{NN}^{-1} \mathbf{f}$$

$$\boldsymbol{\Sigma}_* = \mathbf{K}_{**} - \mathbf{K}_{N*}^{\mathrm{T}} \mathbf{K}_{NN}^{-1} \mathbf{K}_{N*}$$

These equations require the inversion of the NxN data covariance matrix $\mathbf{K}_{NN}$. Matrix inversion is an $\mathcal{O}(N^3)$ operation.

References / Further Reading

Talks by David McKay, Richard Turner.

Carl Rasmussen and Chris Williams, 2006: *Gaussian processes for machine learning*. Chapter 2 on regression covers the material of this presentation.

Chris Bishop, 2006: *Pattern Recognition and Machine Learning*. Chapter 2 contains proofs of many of the useful analytical properties of the Gaussian distribution.

Bui et. al., "A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation", *Journal of Machine Learning Research 18*, (2017)