Gaussian Processes for Regression

1 Introduction

Guassian Processes (GPs) are Bayesian probabilistic models commonly applied to regression and classification problems. Bayesian methods are based on Bayes' theorem, which given an underlying model $\mathcal{M}$, with parameters $\boldsymbol{\theta}$, and observed data $\mathcal{D}$, is written as

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})} \tag{1}$$

The quantity $p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$ is known as the likelihood and expresses the probability of the observed data for the specific value of $\boldsymbol{\theta}$. Different values of $\boldsymbol{\theta}$ will produce different likelihoods, indicating which parameter choices appear to best describe the observed data. Note that the likelihood is not a probability distribution when viewed as a function of $\boldsymbol{\theta}$ for a given $\mathcal{D}$. The distribution $p(\boldsymbol{\theta}|\mathcal{M})$ is known as the prior over $\boldsymbol{\theta}$ for the model $\mathcal{M}$, and contains all our assumptions and beliefs about $\boldsymbol{\theta}$ before any observations of the data have occurred. The denominator is called the marginal likelihood, or the evidence, and is defined as

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})\, d\boldsymbol{\theta} \tag{2}$$

It is the result of integrating the product of the likelihood and the prior over the parameters ($\boldsymbol{\theta}$ is marginalised out from the likelihood, hence the name), and acts as a normalisation constant, ensuring the RHS of the theorem integrates to one, as a probability density must. The conditional probability $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ is the posterior distribution of $\boldsymbol{\theta}$ after combining our prior intuition with the current data observations and normalising by the overall evidence. The posterior is a compromise between the prior and the likelihood, with the exact combination influenced by the properties of the prior and the quality of the data used to compute the likelihood.

To make predictions for unobserved data points, a posterior predictive distribution (see Appendix) can be formed from the posterior distribution over $\boldsymbol{\theta}$.

A GP defines a prior distribution over functions, from which the posterior predictive distribution can be estimated once some data has been observed. Conveniently, the GP formalism allows inference to be performed in closed form. Apart from learning the kernel hyperparameters, there is no training; we are able to write down the exact equations that we will use to make predictions.

2 Weight-Space View

Consider the training data $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, ..., N\}$ of N observations.

In the standard Bayesian linear regression model we assume that the outputs are a linear function of the inputs with additional Gaussian noise:

$$y = f(\mathbf{x}) + \epsilon \tag{3}$$
$$= \mathbf{x}^T\mathbf{w} + \epsilon \tag{4}$$

where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is a weights vector and $\epsilon \sim \mathcal{N}(0, \sigma_N^2)$. The likelihood of the observations given the parameters is

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{w}) \tag{5}$$
$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\mathbf{w})^2}{2\sigma_N^2}\right) \tag{6}$$
$$= \frac{1}{(2\pi\sigma_N^2)^{N/2}} \exp\left(-\frac{|\mathbf{y} - \mathbf{X}^T\mathbf{w}|^2}{2\sigma_N^2}\right) \tag{7}$$
$$= \mathcal{N}(\mathbf{X}^T\mathbf{w}, \sigma_N^2\mathbf{I}) \tag{8}$$

where we have assumed that the data is independent and used the fact that the likelihood of all the observations is the product of the likelihoods for each independent observation.

We assume that the prior distribution over the weights is Gaussian with zero mean:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{9}$$

Equation (4) defines a function of $\mathbf{x}$ for a given $\mathbf{w}$. Therefore, the probability distribution over $\mathbf{w}$ induces a probability distribution over functions y, i.e., a GP.

With the prior distribution we can employ Bayes' theorem to form the posterior distribution over the weights (ignoring the normalising constant denominator):

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}) \tag{10}$$
$$= \mathcal{N}(\mathbf{X}^T\mathbf{w}, \sigma_N^2\mathbf{I})\,\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{11}$$
$$\propto \exp\left(-\frac{(\mathbf{y} - \mathbf{X}^T\mathbf{w})^T(\mathbf{y} - \mathbf{X}^T\mathbf{w})}{2\sigma_N^2}\right) \exp\left(-\frac{\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}}{2}\right) \tag{12}$$

It can then be shown that

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) \propto \exp\left[-\frac{1}{2}(\mathbf{w}-\mathbf{z})^T\left(\frac{\mathbf{X}\mathbf{X}^T}{\sigma_N^2}+\mathbf{\Sigma}^{-1}\right)(\mathbf{w}-\mathbf{z})\right] \tag{13}$$

where $\mathbf{z} = \sigma_N^{-2}(\sigma_N^{-2}\mathbf{X}\mathbf{X}^T + \mathbf{\Sigma}^{-1})\mathbf{X}\mathbf{y}$. We recognise this posterior distribution to be Gaussian with mean $\mathbf{z}$ and covariance matrix $\mathbf{A}^{-1}$:

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) \sim \mathcal{N}\left(\mathbf{z} = \frac{\mathbf{A}^{-1}\mathbf{X}\mathbf{y}}{\sigma_N^2}, \mathbf{A}^{-1}\right) \tag{14}$$

where $\mathbf{A} = \sigma_N^{-2}\mathbf{X}\mathbf{X}^T + \mathbf{\Sigma}^{-1}$.

For the test input $\mathbf{x}_*$, we obtain the posterior predictive distribution for $f_* = f(\mathbf{x}_*)$ by averaging the output of all possible linear models, with respect to the posterior distribution $p(\mathbf{w}|\mathbf{X},\mathbf{y})$, by integrating over $\mathbf{w}$. It can be shown that the result is

$$p(f_*|\mathbf{x}_*,\mathbf{X},\mathbf{y}) = \int p(f_*|\mathbf{x}_*,\mathbf{w})p(\mathbf{w}|\mathbf{X},\mathbf{y})\,d\mathbf{w} \tag{15}$$

$$= \mathcal{N}\left(\frac{1}{\sigma_N^2}\mathbf{x}_*^T\mathbf{A}^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}_*^T\mathbf{A}^{-1}\mathbf{x}_*\right) \tag{16}$$

To make a prediction of $\mathbf{y}_*$ we can just take the mean of this predictive distribution, which is the mean of the posterior distribution in (14) multiplied by $\mathbf{x}_*$, as would be expected from equation (4).

To model nonlinear relationships, we map the input $\mathbf{x}$ into a higher-dimension feature space using the function $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), ..., \phi_N(\mathbf{x})]^T$, so that the model becomes

$$y = \boldsymbol{\phi}(\mathbf{x})^T\mathbf{w} + \epsilon \tag{17}$$

The above equations for the linear model still hold, except that everywhere $\mathbf{X}$ is replaced with $\Phi(\mathbf{X})$, which represents $\boldsymbol{\phi}(\mathbf{x})$ for all $\mathbf{x}$ in $\mathbf{X}$ (see appendix). The corresponding covariance function is defined as

$$k(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\mathbf{\Sigma}\boldsymbol{\phi}(\mathbf{x}') \tag{18}$$

3 Function-Space View

A GP is essentially the generalisation of the multivariate Gaussian distribution to potentially infinitely many variables.

Definition: A GP is a stochastic process (a collection of random variables indexed by some mathematical set) such that any finite collection of the random variables has a joint Gaussian distribution, i.e., the marginalisation of a GP results in a joint Gaussian distribution.

When we sample from an n-dimensional Gaussian distribution, the result is an n-dimensional vector, with each component representing the value of the random variable in that particular dimension. For example, when sampling from a two-dimensional Gaussian distribution, a two-dimensional vector is obtained, say $[0.42, 0.56]^T$. The first component, 0.42, is the value of the random variable in the first dimension, and 0.56 is the value of the random variable in the second dimension.

In theory, the result of sampling from a GP would be an infinite-dimensional vector, i.e., a function. The i-th component of this vector is the value of the random variable in the i-th dimension. Therefore, a GP describes a distribution over functions.

For the input $\mathbf{x}$, the GP is written as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{19}$$

where the mean function $m(\mathbf{x})$, and the covariance (or kernel) function $k(\mathbf{x}, \mathbf{x}')$ are given by

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{20}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \tag{21}$$

The covariance function must be positive semi-definite (see Kernels), and specifies the covariance between pairs of random variables

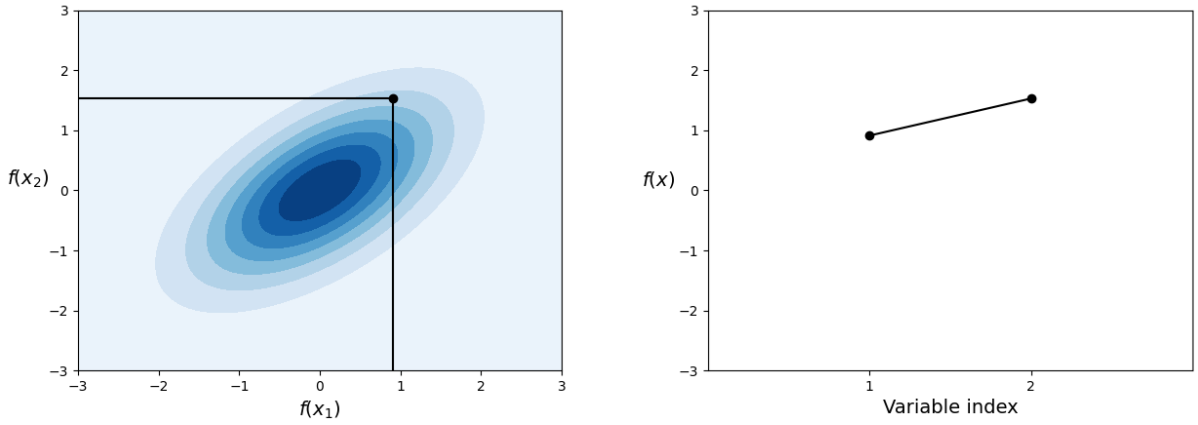
$$\text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j) \tag{22}$$

As mentioned above, for any finite set of points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, the GP defines the multivariate Gaussian distribution on the associated function values $\mathbf{f} = f(\mathbf{X})$:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}) \tag{23}$$

where $\mathbf{K}$ is the covariance matrix (a Gram matrix) corresponding to the covariance function.

Alternately, this can be written as



(a) (b)

Figure 1: (a) A contour plot of a two-dimensional Gaussian distribution and a single sample (black dot). (b) The same sample plotted on new axes.

$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{x}_1) \\ \mu(\mathbf{x}_2) \\ \vdots \\ \mu(\mathbf{x}_n) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & k(\mathbf{x}_n, \mathbf{x}_2) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right) \tag{24}$$

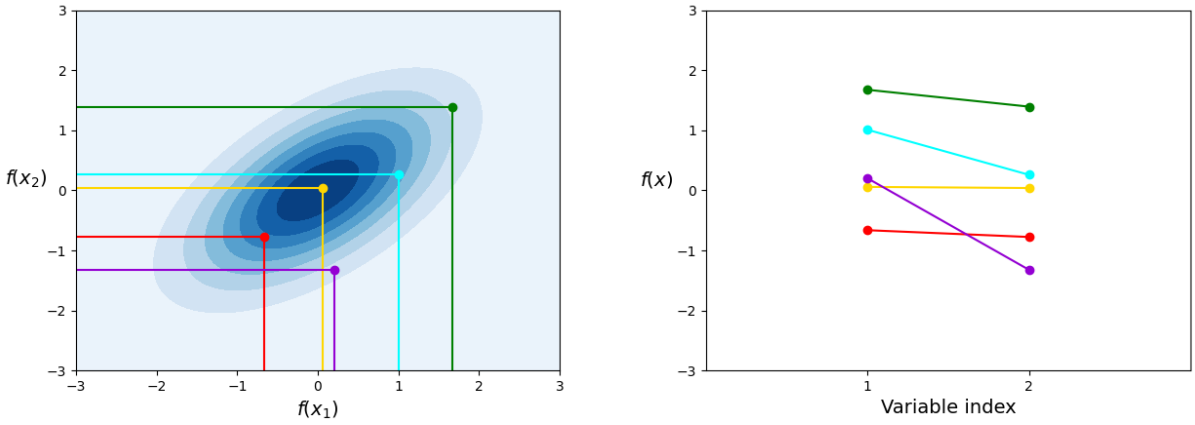
3.1 Intuition

Consider the contour plot of the two-dimensional Gaussian distribution in Figure 1(a). The two random variables forming the joint distribution are $f(x_1)$ and $f(x_2)$. A sample taken from the distribution is represented with a black dot. The same sample is represented on new axes in Figure 1(b). The random variable value lies along the y-axis, and the x-axis indicates the random variable index. There are now two black dots, one representing the sample value for $f(x_1)$ and the other the value for $f(x_2)$.

Figure 2(a) is identical to Figure 1(a), except that five samples have now been taken from the distribution and are represented by coloured dots.Figure 2(b) represents these samples on new axes.

Figure 3 represents samples taken from a four-dimensional Gaussian distribution. At this stage the plotted lines would be of limited practical use for regression tasks since they are not sufficiently smooth.

Figures 4(a) and 4(b) display ten samples taken from 15-variate and 50-variate Gaussian distributions, respectively. The plotted curves are now smoother since there are more random variables sharing stronger correlations than in the previous plots, i.e., the off-diagonal elements in the corresponding covariance matrix are generally closer to one. The curves in these two plots were generated using the Radial Basis Function (RBF) kernel.



(a) (b)

Figure 2: (a) The same contour plot as Figure 1(a), but with five samples taken from the distribution. (b) The five samples represented on new axis.

As the number of random variables in the joint Gaussian distribution increases, so does the smoothness of the curves representing the samples. If there are an infinite number of random variables the curves are perfectly smooth. The variable index consisting of a finite number of integer values is replaced with the continuous variable x, since now there is a random variable associated with every possible position on this axis.

Figure 5 displays sample functions drawn from a GP prior. There are an infinite number of functions to choose from, but only 10 are shown for the sake of clarity. The GP prior is typically assumed to have a mean of zero (black line), i.e., the mean value over the sample functions is zero for all values of x. The ten sample functions shown in Figure 5 do not have zero mean, but for a large number of sample functions this would be the case. A GP prior is said to describe a distribution over functions, since it defines a joint distribution over the infinite number of random variables that cover every possible value of x along its axis in Figure 5.

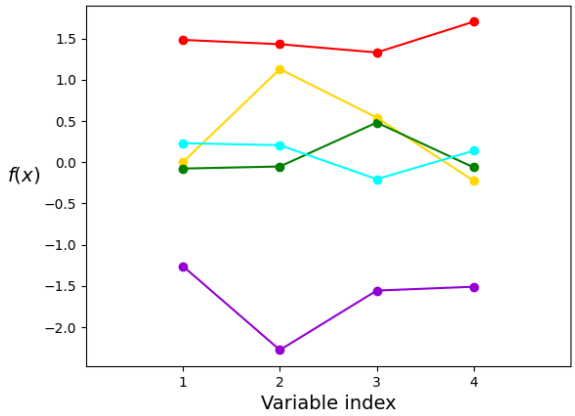
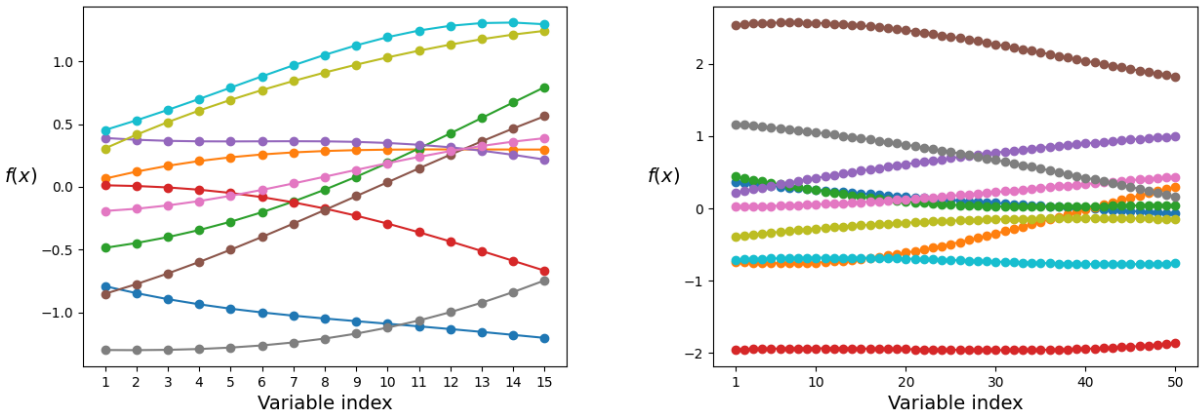


Figure 3: Five samples taken from a four-dimensional Gaussian distribution.



(a) (b)

Figure 4: (a) Ten samples from a 15-variate multivariate Gaussian distribution.(b) Ten samples from a 50-variate multivariate Gaussian distribution.

If we wish to obtain samples from the GP prior, it is not practically feasible to draw infinite-dimensional vectors. Instead, the GP prior is typically sampled over a finely-spaced, finite set of input (or index) points $\mathbf{X}$, resulting in the distribution over the function values $\mathbf{f} = f(\mathbf{X})$

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}) \tag{25}$$

The "functions" displayed in Figure 5, have actually been sampled over a set of 1000 input points, i.e., the function values have been sampled from a 1000-variate Gaussian distribution.

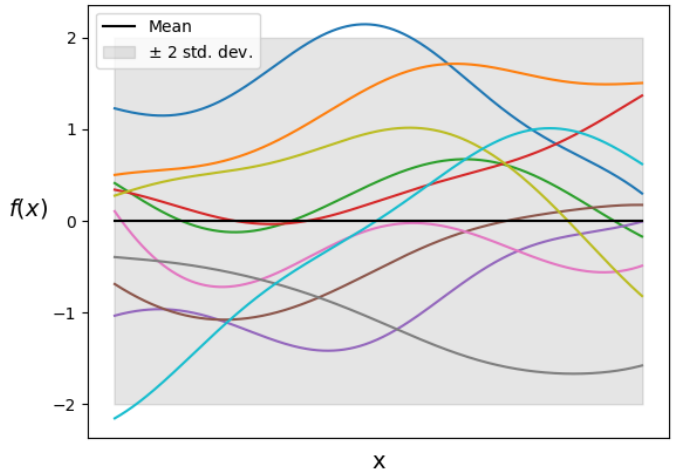


Figure 5: Ten sample functions from a GP prior.

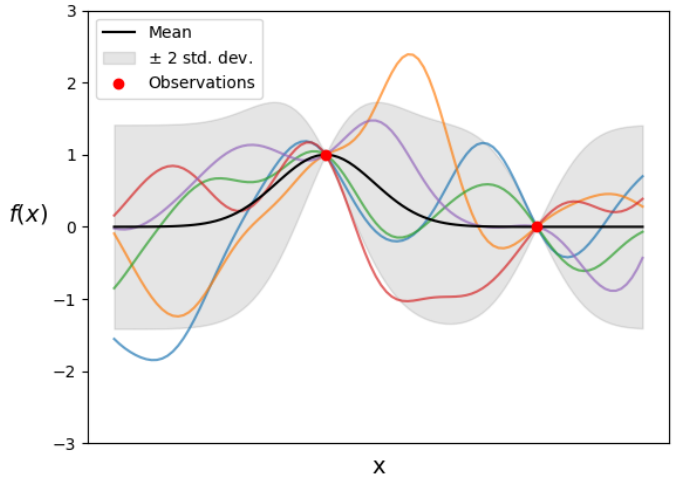


Figure 6: Sample functions from the GP posterior after two data points have been observed.

Suppose there is a regression problem for a dataset of two noise-free observations: $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$. Firstly, a GP prior capable of generating functions with properties suitable for the task is specified (via the mean and covariance functions). Firstly, a GP prior capable of generating functions with properties suitable for the task is specified (via the mean and covariance functions). Given this GP prior and the observed data, the GP posterior over functions can then be inferred.

The situation is displayed in Figure 6, in which five sample functions (sampled over a set of 100 equally-spaced test points) from the GP posterior are shown. Since there is no uncertainty in these observations (they are noise-free), the sample functions pass through the observed points. There are an infinite number of functions that pass through the observed data points but only five are displayed for the sake of clarity. The mean of these functions, i.e., the mean function, is represented by the black line. This is often used to make predic-

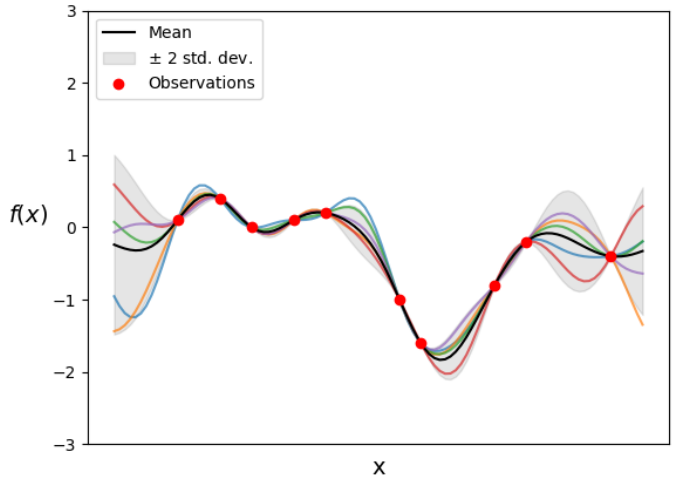


Figure 7: Sample functions from the GP posterior after ten data points have been observed.

tions. The shaded region represents the mean plus two standard deviations, corresponding to the 95% credible interval (the interval which is 95% likely to contain the ground truth function).

In Figure 7, more noise-free observations have been added to the dataset $\mathcal{D}$. As before, the sample and mean functions travel through the observed data points. Because there is more data, the model uncertainty is generally reduced, particularly close to the observations.

3.2 Predictions with Noise-free Observations

Suppose we have the training data $\mathcal{D} = (\mathbf{X}, \mathbf{f}) = \{(\mathbf{x}_i), f_i) \,|\, i = 1, 2, ..., N\}$ of N observations, where $f_i = f(\mathbf{x}_i)$. Given a test set $\mathbf{X}_*$ of size N_*, the objective is to predict the function outputs $\mathbf{f}_* = f(\mathbf{X}_*)$.

Using the properties of multivariate Gaussian distributions (see Appendix) we can perform marginalisation on the GP prior to obtain the joint Gaussian distribution between $\mathbf{f}$ and $\mathbf{f}_*$. This can be expressed as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \tag{26}$$

where $K(\mathbf{X}, \mathbf{X})$ is an $N \times N$ matrix of the covariance function evaluated for all possible pairs of training points, $K(\mathbf{X}, \mathbf{X}_*)$ is an $N \times N_*$ matrix of the covariance function evaluated for all possible pairs of training and test points, and similarly for the matrices $K(\mathbf{X}_*, \mathbf{X})$ and $K(\mathbf{X}_*, \mathbf{X}_*)$.

Employing the rules of Gaussian conditioning (see Appendix), the posterior predictive distribution of $\mathbf{f}_*$ is obtained as

$$p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{f}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}_*|\mathbf{f}}, \boldsymbol{\Sigma}_{\mathbf{f}_*|\mathbf{f}}) \tag{27}$$

$$\boldsymbol{\mu}_{\mathbf{f}_*|\mathbf{f}} = \boldsymbol{\mu}_* + K(\mathbf{X}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{f} - \boldsymbol{\mu}) \tag{28}$$

$$\boldsymbol{\Sigma}_{\mathbf{f}_*|\mathbf{f}} = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{X}_*) \tag{29}$$

This is the predictive distribution for the specific set of test cases $\mathbf{X}_*$. Sample functions from two such distributions are displayed in Figures 6 and 7, for a set of 100 equally-spaced test points, and with two and ten training observations, respectively. The posterior mean is represented by the black line.

If we consider only a single test point, $\mathbf{x}_*$, we can rewrite equation (28) as

$$\mu_{f_*|f} = \mu_* + K(\mathbf{x}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{f} - \boldsymbol{\mu}) \tag{30}$$

$$= \mu_* + \sum_i^N \alpha_i k(\mathbf{x_i}, \mathbf{x}_*) \tag{31}$$

where $\alpha_i = K(\mathbf{X}, \mathbf{X})^{-1}(f(\mathbf{x}_i) - \mu(\mathbf{x}_i))$.

Therefore, the posterior mean can be interpreted as a correction to the prior mean, consisting of a weighted combination of kernel functions, one for each training observation. When making new predictions, the kernel weights every output $f(\mathbf{x}_i)$ by the similarity its corresponding input $\mathbf{x}_i$ has to the test point.

The corresponding posterior process to (27) is

$$f_p(\mathbf{x}) \sim \mathcal{GP}(m_p(\mathbf{x}), k_p(\mathbf{x}, \mathbf{x}')) \tag{32}$$
$$m_p(\mathbf{x}) = m(\mathbf{x}) + K(\mathbf{x}, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{f} - \mathbf{m}) \tag{33}$$
$$k_p(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{x}') \tag{34}$$

where $k(\mathbf{x}, \mathbf{X})$ is a vector of covariances between every training observation and $\mathbf{x}$.

Notice that the posterior covariance $k_p(\mathbf{x}, \mathbf{x}')$ is equal to the prior covariance $k(\mathbf{x}, \mathbf{x}')$ minus a positive term which depends on the training observations. Therefore, the additional information provided by the training observations means the posterior variance is always smaller than the prior variance.

To make predictions from $p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{f})$, we can either sample from the distribution, or simply use the mean $\boldsymbol{\mu}_{\mathbf{f}_*|\mathbf{f}}$ as the prediction, i.e., we marginalise the distribution on each test point to extract the mean and standard deviation.

We see in equation (27) that although GPs are defined in an infinite-dimensional space, to make predictions we only need to know the Gaussian distribution of a finite set (the training and test points) of the possible inputs. Similarly, equation (31) shows that the posterior mean is predicted with a finite sum over the training points only.

3.3 Predictions with Noisy Observations

In real-world scenarios it is likely that we will not observe noise-free observations of function values, but rather noisy versions $y = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_N^2)$. In this case, the covariance of the noisy observations is

$$\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_N^2 \delta_{ij} \tag{35}$$

Following the noise-free case, we can take advantage of the properties of multivariate Gaussians to marginalise the GP prior and obtain the joint Gaussian distribution between $\mathbf{y}$ and $\mathbf{f}_*$:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \tag{36}$$

Making use of the rules of Gaussian conditioning, the posterior predictive distribution is then

$$p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}_*|\mathbf{f}}, \boldsymbol{\Sigma}_{\mathbf{f}_*|\mathbf{f}}) \tag{37}$$

$$\boldsymbol{\mu}_{\mathbf{f}_*|\mathbf{f}} = \boldsymbol{\mu}_* + K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}]^{-1}(\mathbf{y} - \boldsymbol{\mu}) \tag{38}$$

$$\boldsymbol{\Sigma}_{\mathbf{f}_*|\mathbf{f}} = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}]^{-1}K(\mathbf{X}, \mathbf{X}_*) \tag{39}$$

As with the noise-free observations, there is a corresponding posterior process.

4 Gaussian Processes as Non-Parametric Models

A GP will always fit the (noise-free) data since it is a non-parametric model. Parametric models assume that the observed data can be modelled by a finite and fixed number of parameters. For example, the linear regression model $y = \theta_0 x + \theta_1$ utilises the parameters θ_0 and θ_1. Non-parametric models, still have parameters, but their number is not fixed beforehand. In GP regression, the noise-free latent function values at the training observations can be regarded as the parameters of the model. Therefore, the parameters scale with the number of training observations. In the weight-space view, the parameters can be thought of as the weights of the linear model which uses the basis functions ϕ (see equation (17)).

5 Covariance Functions

Covariance functions control the general properties of GPs; we have already seen in 2.2 that a GP is completely specified by its mean and covariance functions. Covariance functions determine how the correlations between points are modelled by the GP, and therefore manifest our assumptions about the functions we wish to learn. A basic supposition in supervised learning, and in particular regression, is that inputs $\mathbf{x}$ which are close, are likely to have similar target values, $f(\mathbf{x})$. In GP regression, this notion of similarity is defined by the covariance function, which as a result enforces smoothness of the random functions across the input set of points.

It is possible to add or multiply kernel functions together to create new kernels. Presented below are some of the covariance functions most commonly used with GPs.

5.1 Radial Basis Function Kernel

One of the most popular covariance functions for GPs (and in other areas of machine learning) is the Radial Basis Function (or squared exponential function). The RBF is stationary, meaning that it is a function of $\mathbf{x} - \mathbf{x}'$, and is therefore translationally invariant.

For the inputs $\mathbf{x}$ and $\mathbf{x}'$, the RBF kernel is defined as

$$k(\mathbf{x}, \mathbf{x}')_{RBF} = \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|^2}{2l^2}\right) \tag{40}$$

where l is the characteristic length-scale. Often when used for GPs, the exponential is multiplied by a^2, where a is an amplitude parameter.

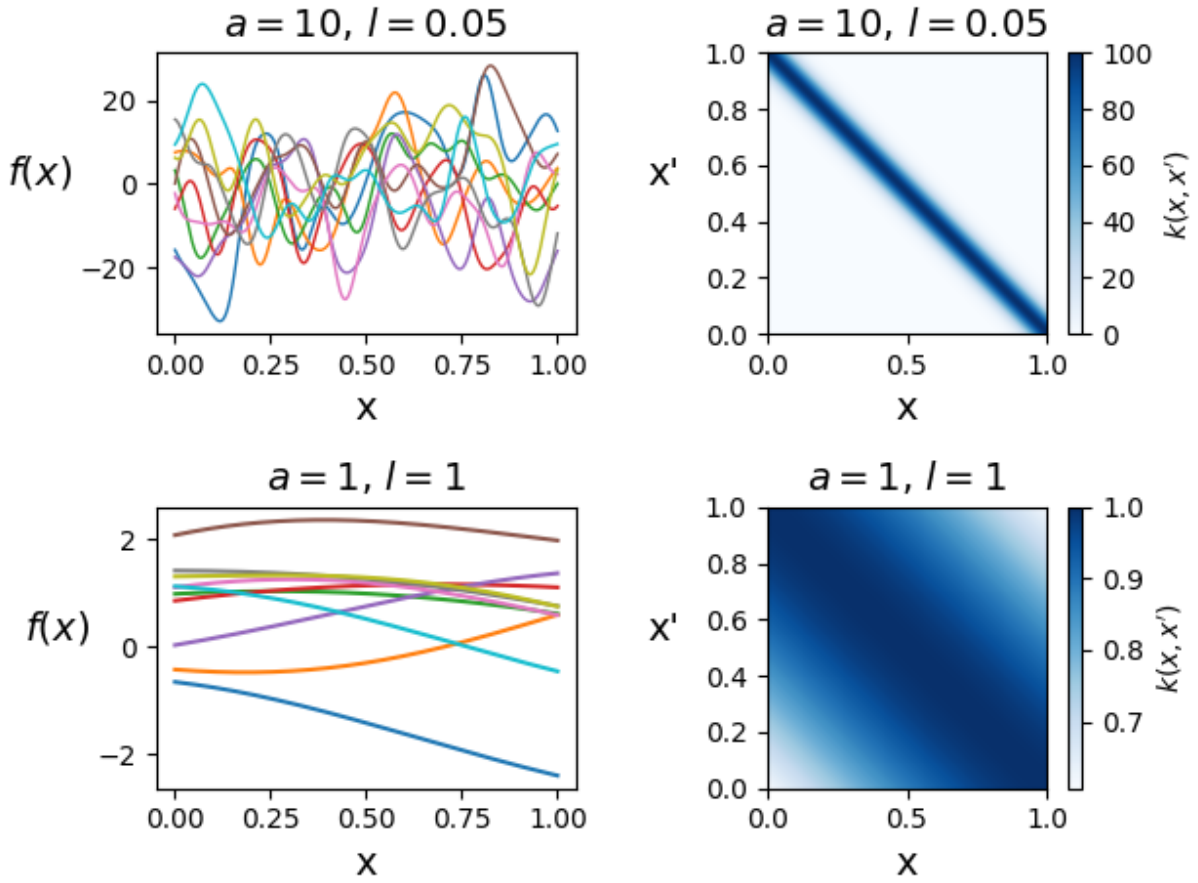


Figure 8: Sample functions (left) drawn from GP priors using a RBF covariance matrix with different values of a and l (right).

Since the value of the RBF kernel is largest when the squared Euclidean distance between the inputs, $|\mathbf{x}-\mathbf{x}'|^2$, is small, and smallest when this quantity is large, the kernel can clearly be interpreted as a similarity measure.

The RBF kernel is infinitely differentiable, and consequently generates extremely smooth functions.

In Figure 8 we observe sample functions drawn from GP priors for which a RBF covariance matrix (a matrix since practically we cannot work with infinite-dimensional objects) has been utilised with different values of a and l. When $a = 10$ and $l = 0.05$, note how the covariance is only large when $\mathbf{x}$ and $\mathbf{x}'$ are very similar, meaning inputs further apart have little correlation with one another. This is because with small values of l, unless we also have small values of $|\mathbf{x}-\mathbf{x}'|$, the term within the brackets in equation (40) will be a large negative number, and the kernel will produce a small number. When $a = 1$ and $l = 1$, the smaller a value leads to smaller deviations of $f(\mathbf{x})$ from the mean, whilst the larger l value results in functions that vary at a slower rate. This is because with a larger value of l, unless $|\mathbf{x}-\mathbf{x}'|$ is very large, the term within the brackets in equation (40) will be small, and the kernel will

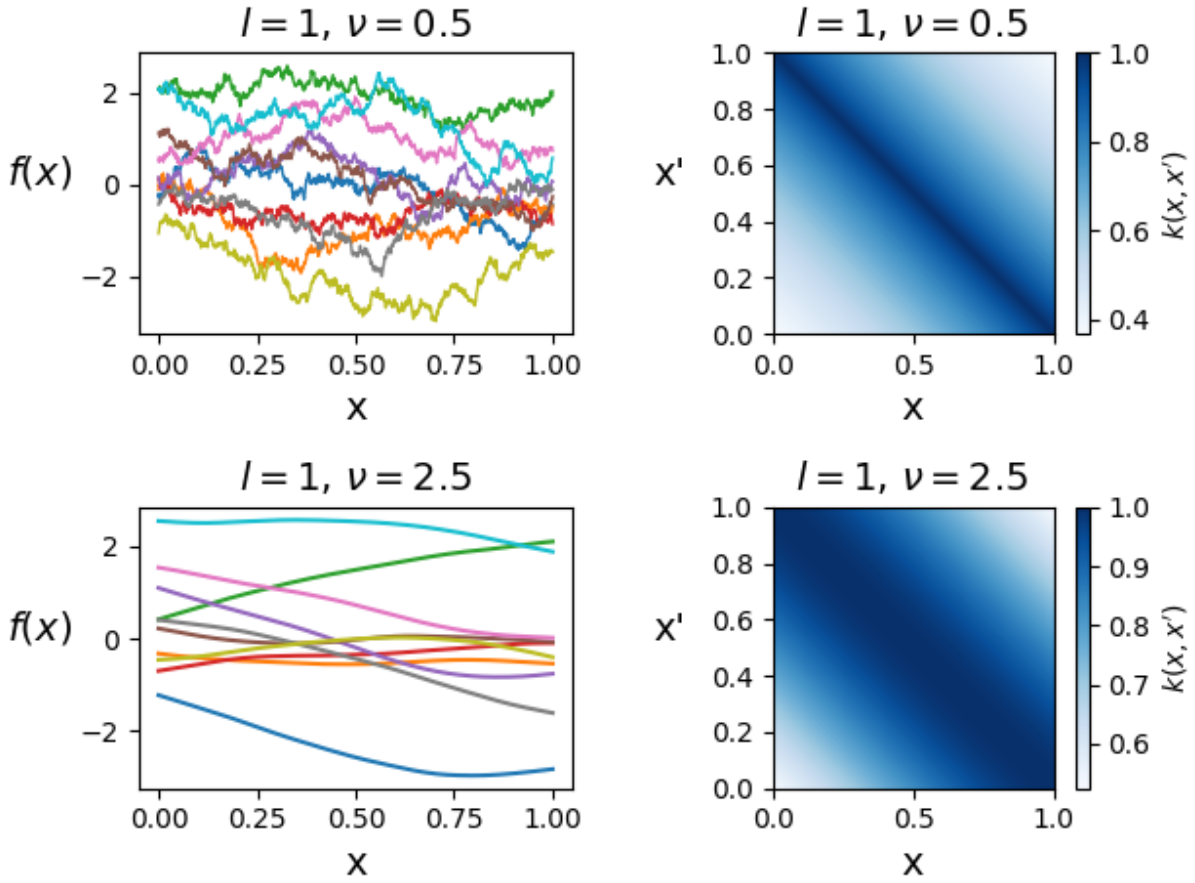


Figure 9: Sample functions (left) drawn from GP priors using a Matérn covariance matrix with different values of ν (right).

produce a large number (closer to one, that is). The kernel produces relatively large values for the entire covariance matrix, as opposed to the previous example, where larger values are restricted to the diagonal elements and those in the vicinity.

5.2 Matérn Kernel

Another popular covariance function for GP regression is the Matérn kernel, which can be used for problems where the data is not smooth, such as those found throughout the physical sciences. It is a generalisation of the RBF kernel, with an additional parameter v, that along with l, regulates the smoothness of the generated functions; a GP making use of the Matérn kernel produces functions that are $\lceil \nu - 1 \rceil$ times differentiable, where $\lceil x \rceil$ returns the smallest integer greater than or equal to x. The Matérn kernel has the form

$$k_M(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|\mathbf{x} - \mathbf{x}'|}{l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|\mathbf{x} - \mathbf{x}'|}{l} \right) \tag{41}$$

where Γ is the gamma function, and K_ν is a modified Bessel function. It can be shown that as $\nu \to \infty$, this exactly results in the RBF kernel. Whilst the parameter l is learnt, the parameter v is typically fixed beforehand, since the extreme changes that occur in the functions as ν is varied make learning computationally challenging. With half-integer values of ν, the Matérn kernel can be decomposed into a product of an exponential and a polynomial of order $\nu - 1/2$. Three popular choices of ν are $\frac{1}{2}$, $\frac{3}{2}$ and $\frac{5}{2}$

$$k_{M,\nu=1/2}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|}{l}\right) \tag{42}$$

$$k_{M,\nu=3/2}(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{\sqrt{3}|\mathbf{x} - \mathbf{x}'|}{l}\right) \exp\left(-\frac{\sqrt{3}|\mathbf{x} - \mathbf{x}'|}{l}\right) \tag{43}$$

$$k_{M,\nu=5/2}(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{\sqrt{5}|\mathbf{x} - \mathbf{x}'|}{l} + \frac{5|\mathbf{x} - \mathbf{x}'|^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}|\mathbf{x} - \mathbf{x}'|}{l}\right) \tag{44}$$

Figure 9 shows some sample functions drawn from GP priors for which a Matérn covariance matrix has been employed with different values of ν. With $\nu = 1/2$ the sample functions do not demonstrate smoothness, but with $\nu = 5/2$ the sample functions are already starting to resemble those generated by the RBF kernel.

5.3 Rational Quadratic Kernel

The Rational Quadratic (RQ) covariance function is expressed as

$$k_{RQ}(\mathbf{x}, \mathbf{x}') = \exp\left(1 + \frac{|\mathbf{x} - \mathbf{x}'|^2}{2\alpha l^2}\right)^{-\alpha} \tag{45}$$

It is described as a scale mixture, or infinite sum, of RBF kernels with different characteristic length-scales, l. If we let $\tau = l^{-2}$, we can express the superposition of RBF kernels with a gamma distribution of length scales, $p(\tau|\alpha, \beta) \propto \tau^{\alpha-1} \exp(-\alpha\tau/\beta)$, as the following integral

$$\int \tau^{\alpha-1} \exp\left(-\frac{\alpha\tau}{\beta}\right) \exp\left(-\frac{\tau|\mathbf{x} - \mathbf{x}'|^2}{2}\right) d\tau \propto \exp\left(1 + \frac{|\mathbf{x} - \mathbf{x}'|^2}{2\alpha l^2}\right)^{-\alpha} \tag{46}$$

$$= k_{RQ}(\mathbf{x}, \mathbf{x}') \tag{47}$$

The RQ covariance function can be expanded as

$$k_{RQ}(\mathbf{x}, \mathbf{x}') = 1 - \frac{\alpha|\mathbf{x} - \mathbf{x}'|^2}{2\alpha l^2} + \frac{\alpha(\alpha+1)}{2}\frac{|\mathbf{x} - \mathbf{x}'|^4}{(2\alpha l^2)^2} + ... \tag{48}$$

$$= 1 - \frac{|\mathbf{x} - \mathbf{x}'|^2}{2l^2} + \frac{|\mathbf{x} - \mathbf{x}'|^4}{8l^4} + \frac{|\mathbf{x} - \mathbf{x}'|^4}{8\alpha l^4} + ... \tag{49}$$

Likewise, from equation (40), the expansion of the RBF kernel is

$$k_{RBF}(\mathbf{x}, \mathbf{x}') = 1 - \frac{|\mathbf{x} - \mathbf{x}'|^2}{2l^2} + \frac{|\mathbf{x} - \mathbf{x}'|^4}{8l^4} - \frac{|\mathbf{x} - \mathbf{x}'|^6}{48l^4} + ... \tag{50}$$

We can see that as $\alpha \to \infty$ the RQ kernel is equal to the RBF kernel.

Figure 10 displays sample functions drawn from GP priors for which the RQ covariance matrix with two values of α has been used. With $\alpha = 1000$, the sample functions resemble those produced by the RBF kernel.

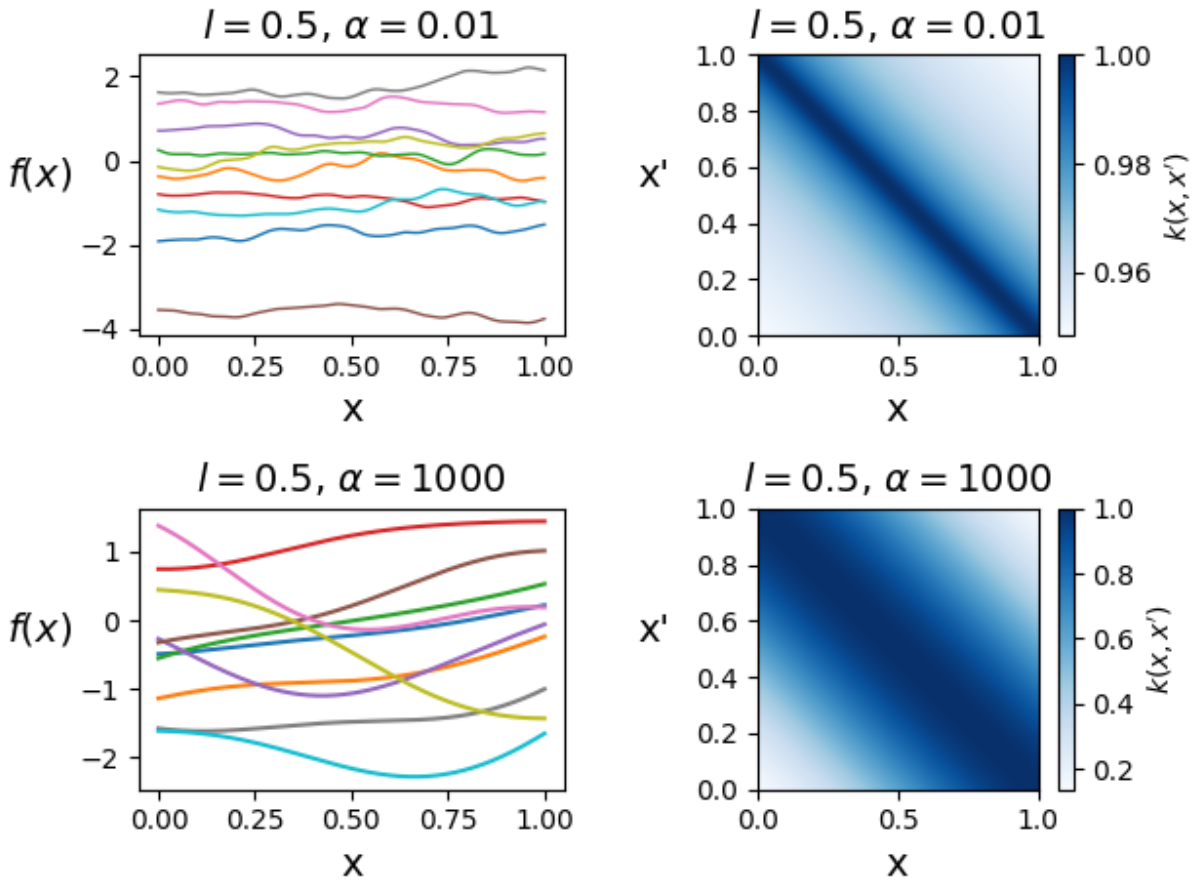


Figure 10: Sample functions (left) drawn from GP priors using a RQ covariance matrix with different values of α (right).

6 Training a Gaussian Process

Kernel functions typically contain trainable hyperparameters. For example, the RBF kernel has the hyperparameters length-scale, l, and amplitude, α. One method to optimise the kernel hyperparameters is to conduct an exhaustive search over a discrete grid of values, with

validation loss as an objective. Another approach is to obtain the posterior distribution over the hyperparameters, but in practice the computations required are often intractable. Therefore, common practice is to obtain point estimates of the hyperparameters by maximising the marginal likelihood, which is the integral of the likelihood multiplied by the prior. Given the data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ and kernel hyperparameters $\boldsymbol{\theta}$, the marginal likelihood is

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) \, d\mathbf{f} \tag{51}$$

where marginalisation occurs over the function values $\mathbf{f}$, i.e., the function values are integrated out, leaving a function of the hyperparameters only.

The marginal likelihood represents the probability of the observed data given the model and its hyperparameters. Since the prior, $p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{X})$, and the likelihood, $p(\mathbf{y}|\mathbf{f}, \mathbf{X}) = \prod_i \mathcal{N}(y_i|f_i, \sigma_i^2)$ are both Gaussians, we can use Gaussian identities to calculate the integral and obtain

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) &= \frac{1}{(2\pi)^{N/2}|\mathbf{K}_y|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{K}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\} && (52) \\ &= \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{K}_y) && (53) \end{aligned}$$

where $\mathbf{K}_y = K(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I}$, and N is the number of training observations.

Primarily due to issues with numeric stability, it is preferred to work with the log marginal likelihood

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{K}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{2} \log |\mathbf{K}_y| - \frac{N}{2} \log 2\pi \tag{54}$$

The first term of the log marginal likelihood, the only one involving the training observations, is interpreted as a data-fit measure, whilst the second term, dependent only upon the covariance function and the inputs, is a complexity penalty. The final term is a normalisation constant.

The log marginal likelihood can be maximised using gradient-based optimisers, as we know its partial derivatives with respect to the hyperparameters

$$\begin{aligned} \frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) &= \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{K}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{2} \mathrm{tr}\left(\mathbf{K}_y^{-1} \frac{\partial \mathbf{K}_y}{\partial \theta_j} \right) && (55) \\ &= \frac{1}{2} \mathrm{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T - \mathbf{K}_y^{-1}) \frac{\partial \mathbf{K}_y}{\partial \theta_j} \right) && (56) \end{aligned}$$

where $\boldsymbol{\alpha} = \mathbf{K}_y^{-1} \mathbf{y}$.

Since the log marginal likelihood is not convex, care must be taken with regards to local optima when optimising. Parameter optimisation also increases the vulnerability of the GP to overfitting, particularly if the training set is small and the number of hyperparameters is large.

To choose the functional form of the covariance (and potentially the mean) function itself for the problem at hand, expert knowledge and experience is invaluable. Alternately, marginal likelihoods obtained from the candidates can be compared to determine which one is most likely to produce the observed data. Similarly, cross-validation can be performed; the training observations are separated into training and validation sets, with the performance on the latter with respect to some metric used to measure the generalisation error and inform on model choice. Work is ongoing to develop techniques for automating the kernel choice in GPs.

7 Performance

A significant drawback of GP regression is that for N observations there is a computational complexity of $\mathcal{O}(N^3)$ for inference and training due to inversion of the matrix $K(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I}$ in equations (38), (39) and (54). For smaller datasets, direct inversion of this matrix is possible, but for larger datasets it proves too costly.

Methods to improve the performance of GP regression are an ongoing area of research. One popular method is to employ Cholesky decomposition to compute the inverse. There are also numerous approximation techniques that adopt a variety of approaches to tackle this issue.

References

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Do, C.B. (2008). *Gaussian processes*. Stanford CS229 Machine Learning. Stanford University.

Engelhardt, B. (2013). *Gaussian Processes*. STA561: Probabilistic machine learning. Princeton University.

Martin, O.A., Kumar, R., Lao, J. (2021). *Bayesian Modelling and Computation in Python*. CRC Press, Boca Raton, Florida.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts.

Murphy, K.P. (2022). *Probabilistic Machine Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts.

Rasmussen, C.E. (2003). *Gaussian Processes in Machine Learning*. Advanced Lectures on Machine Learning, Lecture Notes in Computer Science (3176), Springer, Berlin.

Rasmussen, C.E., Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts.

Schulz, E., Speekenbrink, M., Krause, A. (2018). *A tutorial on Gaussian Process regression: Modelling, exploring, and exploiting functions.* Journal of Mathematical Psychology (85), pp. 1-16.

Speagle, J.S. (2020). *A Conceptual Introduction to Markov Chain Monte Carlo Methods*. arXiv:1909.12313.

Wang, J. (2024). *An Intuitive Tutorial to Gaussian Process Regression*. arXiv:2009.10862v5.

Zhang, A., Lipton, Z.C., Li, M., Smola, A.J. (2023). *Dive into Deep Learning*. arXiv:2106.11342v5.

Zhu, X. (2011). *Bayesian Nonparametrics*. CS731 Spring 2011 Advanced Artificial Intelligence. University of Wisconsin-Madison.

Appendix

A1. Marginal and Conditional Distributions

Given a joint distribution, the marginal distribution of a random variable, or a subset of random variables, is the probability of the variables contained in the subset, without reference to the values of the other variables.

Suppose the random vector (a vector whose components are scalar-valued random variables) $\mathbf{x} \in \mathcal{R}^D$ is partitioned into $\mathbf{x}_a \in \mathcal{R}^p$ and $\mathbf{x}_b \in \mathcal{R}^q$, where $p + q = D$:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix} \tag{57}$$

The joint distribution is given by

$$p(\mathbf{x}) = p(x_1, x_2, ..., x_D) = p(\mathbf{x}_a, \mathbf{x}_b) \tag{58}$$

The marginal distribution of $\mathbf{x}_a$ is then

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\, d\mathbf{x}_b = \int p(\mathbf{x}_a|\mathbf{x}_b)p(\mathbf{x}_b)\, d\mathbf{x}_b \tag{59}$$

If $\mathbf{x}_a$ consists of only a single variable, the marginal distribution will be of only that single variable. However, if $\mathbf{x}_a$ contains multiple variables, the marginal distribution will itself be a joint distribution.

Table 1 shows the joint and marginal distributions of two discrete, random variables x and y. The values of the joint distributions are contained within the inner 3×3 grid, whilst the values of the marginal distributions are shown along the right and bottom margins. For discrete-valued variables, the integral in equation (59) is replaced with a sum:

$$p(\mathbf{x}_a) = \sum_b p(\mathbf{x}_a, \mathbf{x}_b) \tag{60}$$

The marginal distribution of x is given by

$$\begin{aligned} p(x_1) &= 0.1 + 0 + 0.1 = 0.2 \\ p(x_2) &= 0.1 + 0.1 + 0.2 = 0.4 \\ p(x_3) &= 0.2 + 0.1 + 0.1 = 0.4 \end{aligned} \tag{61}$$

	x_1	x_2	x_3	Total
y_1	0.1	0.1	0.2	0.4
y_2	0	0.1	0.1	0.2
y_3	0.1	0.2	0.1	0.4
Total	0.2	0.4	0.4	1

Table 1: Joint and marginal distributions of two discrete, random variables x and y. The values of the joint distribution are contained within the inner 3×3 grid, whilst the values of the marginal distributions are along the right and bottom margins.

and that of y

$$\begin{aligned} p(y_1) &= 0.1 + 0.1 + 0.2 = 0.4 \\ p(y_2) &= 0 + 0.1 + 0.1 = 0.2 \\ p(y_3) &= 0.1 + 0.2 + 0.1 = 0.4 \end{aligned} \tag{62}$$

Given a joint distribution of random variables, the conditional distribution of a variable, or a subset of variables, is its distribution when the other variables are known to be particular values.

The conditional distribution of $\mathbf{x}_a$ is defined as

$$p(\mathbf{x}_a|\mathbf{x}_b) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_b)} \tag{63}$$

where $p(\mathbf{x}_b) > 0$ is the marginal distribution of $\mathbf{x}_b$.

Returning to the two discrete, random variables in Table 1, the conditional distribution of, for example, x given $y = y_3$, i.e., $p(x|y = y_3)$, is

$$\begin{aligned} p(x_1|y_3) &= \frac{0.1}{0.4} = 0.25 \\ p(x_2|y_3) &= \frac{0.2}{0.4} = 0.5 \\ p(x_3|y_3) &= \frac{0.1}{0.4} = 0.25 \end{aligned} \tag{64}$$

Likewise, the conditional distribution of say, y given $x = x_2$, i.e., $p(y|x = x_2)$, is

$$
\begin{aligned}
p(y_1|x_2) &= \frac{0.1}{0.4} = 0.25 \\
p(y_2|x_2) &= \frac{0.1}{0.4} = 0.25 \\
p(y_3|x_2) &= \frac{0.2}{0.4} = 0.5
\end{aligned}
\tag{65}
$$

A2. Prior Predictive and Posterior Predictive Distributions

The posterior predictive distribution is the distribution of possible unobserved values conditional on the observed values $\mathbf{X}$.

For a fixed parameter θ, the unobserved test point $\mathbf{x}_*$ follows the distribution $p(\mathbf{x}_*|\theta)$. However, since the true value of θ is unknown, we have to average (by integration) over all the possible values of $\theta \in \Theta$, where Θ is the parameter space, according to the posterior distribution $p(\theta|\mathbf{X})$.

Therefore, the posterior predictive distribution is given by

$$p(\mathbf{x}_*|\mathbf{X}) = \int_\Theta p(\mathbf{x}_*|\theta)p(\theta|\mathbf{X})\, d\theta \tag{66}$$

The prior predictive distribution is defined similarly, but with the posterior distribution in the integrand replaced by the prior distribution $p(\theta)$:

$$p(\mathbf{x}_*) = \int_\Theta p(\mathbf{x}_*|\theta)p(\theta)\, d\theta = \int_\Theta p(\mathbf{x}_*, \theta)\, d\theta \tag{67}$$

It is the distribution of possible unobserved values without any conditioning on observations. This equation looks very similar to the equation for the marginal likelihood, which in this case would be

$$p(\mathbf{X}) = \int_\Theta p(\mathbf{X}|\theta)p(\theta)\, d\theta \tag{68}$$

The difference is that the marginal likelihood is conditioned on the observations, meaning the result is a number and not a distribution.

A3. The Multivariate Gaussian Distribution

For a D-dimensional random vector, $\mathbf{x} = [x_1, x_2, ..., x_D]^T$, the multivariate Gaussian distribution is

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{69}$$

where $\boldsymbol{\mu}$ is the D-dimensional mean vector, and $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix. This is often written as

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{70}$$

The multivariate Gaussian distribution is the generalisation of the univariate Gaussian distribution to two or more variables. There is a univariate Gaussian distribution over each component of the random vector $\mathbf{x}$. When the covariance matrix is equal to the identity matrix there are no correlations between the variables, and each component of $\mathbf{x}$ is independent (known as the standard multivariate Gaussian distribution).

The covariance matrix is defined as

$$\boldsymbol{\Sigma}_{ij} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] \tag{71}$$

The elements of the matrix represent the covariances between each possible pair of variables in $\mathbf{x}$. The diagonal elements contain the variances of the variables (the covariance of a variable with itself). Since $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ji}$, the covariance matrix is square, and also symmetric.

Marginals and Conditionals

Suppose the jointly Gaussian vector $\mathbf{x} \in \mathcal{R}^D$ is partitioned into $\mathbf{x}_a \in \mathcal{R}^p$ and $\mathbf{x}_b \in \mathcal{R}^q$, where $p + q = D$:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix} \tag{72}$$

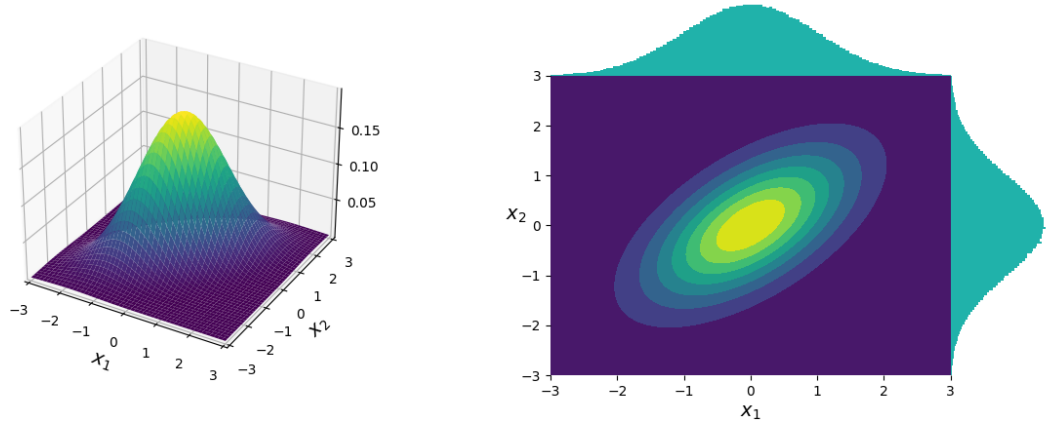
The mean vector and covariance matrix can be similarly partitioned:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \tag{73}$$

The marginal distributions of $\mathbf{x}_a$ and $\mathbf{x}_b$ can be shown to be

$$p(\mathbf{x}_a) = \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{74}$$

$$p(\mathbf{x}_b) = \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb}) \tag{75}$$



(a) (b)

Figure 11: (a) A three-dimensional plot of the two-dimensional Gaussian distribution specified in (78). (b) A contour plot of the same distribution. The marginal distributions of x_1 and x_2 are plotted along their respective margins.

and the conditional

$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned} \tag{76}$$

Therefore, both the marginal and conditional distributions of a multivariate Gaussian are themselves Gaussian distributions. This is a key result in the theory of Gaussian distributions, and by extension, GPs.

Two-Dimensional Gaussian Distribution

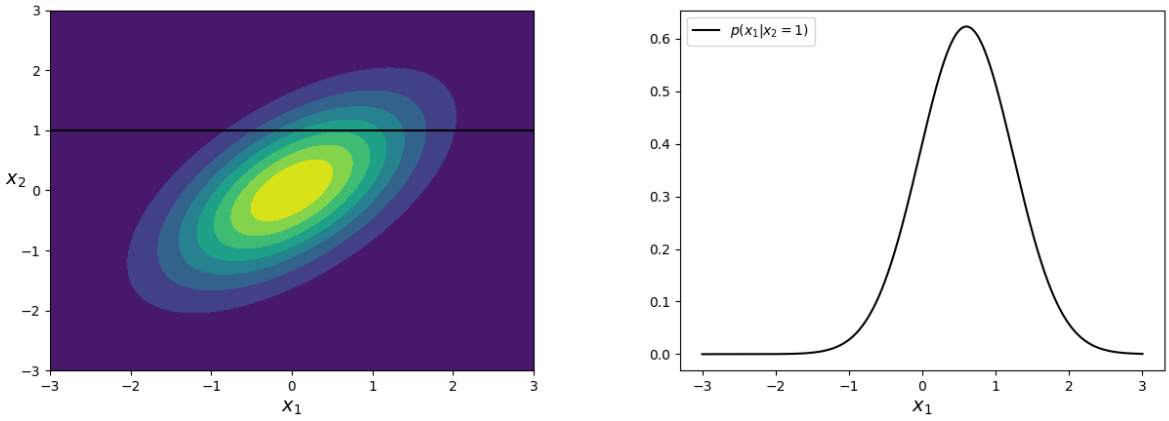
A two-dimensional Gaussian distribution can be expressed as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \tag{77}$$

Consider the two-dimensional Gaussian distribution given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \right) \tag{78}$$

A three-dimensional plot of this distribution in presented in Figure 11(a). Figure 11(b) shows a contour plot of the same distribution. The marginal distributions of x_1 and x_2 are shown along their respective margins. They are given by



(a) (b)

Figure 12: (a) A contour plot of the two-dimesional Gaussian distribution specified in (78), with a line indicating $x_2 = 1$. (b) The conditional distribution of x_1, given $x_2 = 1$.

$$p(x_1) = \mathcal{N}(0, 1) \tag{79}$$

$$p(x_2) = \mathcal{N}(0, 1) \tag{80}$$

Figure 12 provides a visualisation of the probability distribution of x_1 conditional on $x_2 = 1$. Using equations (76) this conditional distribution can be calculated:

$$\mu_{x_1|x_2=1} = 0 + 0.6 \times 1(1 - 0) = 0.6$$
$$\Sigma_{x_1|x_2=1} = 1 - 0.6 \times 1 \times 0.6 = 0.64$$
$$p(x_1|x_2 = 1) \sim \mathcal{N}(0.6, 0.64)$$

A4. The Equivalence of Gaussian Process Regression and Bayesian Linear Regression

Recall the linear regression model

$$y = \boldsymbol{\phi}(\mathbf{x})^T\mathbf{w} + \epsilon, \qquad \mathbf{w} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \qquad \epsilon \sim \mathcal{N}(0, \sigma_N^2)$$

Following the discussion in section 2 and using equation (16), the posterior predictive distribution of $f_* = f(\mathbf{x}_*)$ for a single test point $\mathbf{x}_*$ is

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\frac{1}{\sigma_N^2}\boldsymbol{\phi}(\mathbf{x}_*)^T\mathbf{A}^{-1}\boldsymbol{\Phi}\mathbf{y}, \boldsymbol{\phi}(\mathbf{x}_*)^T\mathbf{A}^{-1}\boldsymbol{\phi}(\mathbf{x}_*)\right) \tag{81}$$

where $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{X})$ and $\mathbf{A} = \sigma_N^{-2}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}^{-1}$.

Utilising $K(\mathbf{X}, \mathbf{X}) = \boldsymbol{\Phi}^T\boldsymbol{\Sigma}\boldsymbol{\Phi}$, we can write

$$\frac{1}{\sigma_N^2}\boldsymbol{\Phi}[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}] = \frac{1}{\sigma_N^2}\boldsymbol{\Phi}[\boldsymbol{\Phi}^T\boldsymbol{\Sigma}\boldsymbol{\Phi} + \sigma_N^2\mathbf{I}] \tag{82}$$

$$= \mathbf{A}\boldsymbol{\Sigma}\boldsymbol{\Phi} \tag{83}$$

Multiply from the left by $\boldsymbol{A}^{-1}$ to obtain

$$\frac{1}{\sigma_N^2}\boldsymbol{A}^{-1}\boldsymbol{\Phi}[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}] = \boldsymbol{\Sigma}\boldsymbol{\Phi} \tag{84}$$

Multiplying from the right by $[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}]^{-1}$ yields

$$\frac{1}{\sigma_N^2}\boldsymbol{A}^{-1} = \boldsymbol{\Sigma}\boldsymbol{\Phi}[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}]^{-1} \tag{85}$$

Therefore, denoting $\boldsymbol{\phi}_* = \boldsymbol{\phi}(\mathbf{x}_*)$, we can write the mean from equation (86) as

$$\frac{1}{\sigma_N^2}\boldsymbol{\phi}_*^T\mathbf{A}^{-1}\boldsymbol{\Phi}\mathbf{y} = \boldsymbol{\phi}_*^T\boldsymbol{\Sigma}\boldsymbol{\Phi}[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}]^{-1}\mathbf{y} \tag{86}$$

$$= K(\mathbf{x}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_N^2\mathbf{I}]^{-1}\mathbf{y} \tag{87}$$

where $K(\mathbf{x}_*, \mathbf{X}) = \boldsymbol{\phi}_*^T\boldsymbol{\Sigma}\boldsymbol{\Phi}$.

The Woodbury matrix identity states that

$$(A + BCD^T)^{-1} = A^{-1} - A^{-1}B(C^{-1} + D^TA^{-1}B)^{-1}D^TA^{-1} \tag{88}$$

If we set $A^{-1} = \mathbf{\Sigma}$, $C^{-1} = \sigma_N^2 \mathbf{I}$, and $B = D = \mathbf{\Phi}$, we can write the covariance of equation (86) as

$$\begin{aligned}
\boldsymbol{\phi}_*^T \mathbf{A} \boldsymbol{\phi} &= \boldsymbol{\phi}_*^T (\sigma_N^{-2} \mathbf{\Phi} \mathbf{\Phi}^T + \mathbf{\Sigma}^{-1})^{-1} \boldsymbol{\phi}_* && (89) \\
&= \boldsymbol{\phi}_*^T \mathbf{\Sigma} \boldsymbol{\phi}_* - \boldsymbol{\phi}_*^T \mathbf{\Sigma} \mathbf{\Phi} [\mathbf{\Phi}^T \mathbf{\Sigma} \mathbf{\Phi} + \sigma_N^2]^{-1} \mathbf{\Phi}^T \mathbf{\Sigma} \boldsymbol{\phi}_* && (90) \\
&= K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{X}) [K(\mathbf{X}, \mathbf{X} + \sigma_N^2]^{-1} K(\mathbf{X}, \mathbf{x}_*) && (91)
\end{aligned}$$

where $K(\mathbf{x}_*, \mathbf{x}_*) = \boldsymbol{\phi}_*^T \mathbf{\Sigma} \boldsymbol{\phi}_*$ and $K(\mathbf{X}, \mathbf{x}_*) = \mathbf{\Phi}_*^T \mathbf{\Sigma} \boldsymbol{\phi}_*$.

To summarise, we can write the mean and covariance of the posterior predictive distribution as

$$\begin{aligned}
\mu_*' &= K(\mathbf{x}_*, \mathbf{X}) [K(\mathbf{X}, \mathbf{X}) + \sigma_N^2 \mathbf{I}]^{-1} \mathbf{y} && (92) \\
\Sigma_*' &= K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{X}) [K(\mathbf{X}, \mathbf{X} + \sigma_N^2]^{-1} K(\mathbf{X}, \mathbf{x}_*) && (93)
\end{aligned}$$

These results are a match with equations (38) and (39) in the case of a single test point, and assuming $\boldsymbol{\mu} = \boldsymbol{\mu}_* = 0$.

Therefore, we have derived a GP from Bayesian linear regression. However, note that linear regression assumes that $\boldsymbol{\phi}(\mathbf{x})$ is a finite length vector, whereas a GP operates with kernels which may correspond to infinite length vectors, i.e., a GP operates in function space.

Radial Basis Function Kernel

The RBF kernel corresponds to a Bayesian linear regression model in which the inputs have been projected into feature space with an infinite number of basis functions.

Starting in the weight-space view, consider the function of the scalar input x

$$f(x) = \sum_{i=1}^{N} w_i \phi_i(x) \tag{94}$$

where

$$\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{N}\mathbf{I}\right), \qquad \phi_i(x) = \exp\left(-\frac{(x-c_i)^2}{2l^2}\right) \tag{95}$$

$f(x)$ is a sum of weighted radial basis functions, with width l, centred at c_i. We recognise $f(x)$ as having the form $\boldsymbol{\phi}(x)^T\mathbf{w}$, where $\boldsymbol{\phi}(x) = [\phi_1(x), \phi_2(x), ..., \phi_N(x)]^T$. Since there is a normal distribution over $\mathbf{w}$, $f(x)$ is a GP.

Using the weight-space definition of the covariance function, we obtain

$$\begin{aligned} k(x,x')_{ws} &= \boldsymbol{\phi}(x)^T\boldsymbol{\Sigma}\boldsymbol{\phi}(x') \\ &= \boldsymbol{\phi}(x)^T\frac{\sigma^2}{N}\mathbf{I}\boldsymbol{\phi}(x') \\ &= \frac{\sigma^2}{N}\sum_{i=1}^{N}\phi_i(x)\phi_i(x') \end{aligned} \tag{96}$$

If we allow an infinite number of basis functions centred everywhere on an interval and note that as $N \to \infty$, $1/N \propto \triangle c \to 0$, the covariance function becomes

$$\begin{aligned} k(x,x')_{ws} &= \lim_{N\to\infty}\frac{\sigma^2}{N}\sum_{i=1}^{N}\phi_i(x)\phi_i(x') \\ &\propto \lim_{N\to\infty}\sigma^2\sum_{i=1}^{N}\phi_i(x)\phi_i(x')\triangle c \\ &= \sigma^2\int_{c_{min}}^{c_{max}}\phi_c(x)\phi_c(x')\,dc \end{aligned}$$

By setting $c_{max} = \infty$ and $c_{min} = -\infty$ we spread the infinitely many basis functions across the entire real line and obtain

$$k(x, x')_{ws} \propto \sigma^2 \int_{\infty}^{\infty} \exp\left(-\frac{(x-c)^2}{2l^2}\right) \exp\left(-\frac{(x'-c)^2}{2l^2}\right) dc \tag{97}$$

$$= \sigma^2 \exp\left[-\frac{(x^2+x'^2)}{2l^2}\right] \int_{\infty}^{\infty} \exp\left(-\frac{c^2}{l^2}\right) \exp\left[-\frac{c(-x-x')}{l^2}\right] dc \tag{98}$$

$$= \sigma^2 \exp\left[-\frac{(x^2+x'^2)}{2l^2}\right] \sqrt{\pi} l \exp\left(\frac{x^2+x'^2-2xx'}{4l^2}\right) \tag{99}$$

$$= \sigma^2 \sqrt{\pi} l \exp\left(\frac{-x^2-x'^2-2xx'}{4l^2}\right) \tag{100}$$

$$= \sigma^2 \sqrt{\pi} l \exp\left[-\frac{(x-x')^2}{2(\sqrt{2}l)^2}\right] \tag{101}$$

$$\propto k(x, x')_{RBF} \tag{102}$$

Therefore, with an infinite number of radial basis functions, the kernel of the GP in the weight-space view is equivalent (proportional to) the RBF kernel in the function-space view.

For simplicity we considered a scalar input, but the result can be readily generalised to the multivariate input $\mathbf{x}$.