Master's Thesis in requirement for partial completion

The University of Melbourne

Jacob Cumming

May 2024

Contents

List of Figures

Chapter 1

Gaussian Processes

- Prove all pos-sem is cov
- prove exponential quad kern is pos semi-def

1.1 Motivation and Definitions

If we knew the distribution of $D(\boldsymbol{\theta})$ exactly, then in order to do approximate Bayesian computation, running the model becomes superfluous to obtain a 'sample' from $D(\boldsymbol{\theta})$. Instead a sample $D(\boldsymbol{\theta})$ could be drawn directly from it's distribution. It is highly improbable that the distribution of $D(\boldsymbol{\theta})$ is known in practice, and so this chapter describes a method of approximating the distribution. Furthermore, since $\Pr(D(\boldsymbol{\theta}) < \varepsilon)$ is approximately proportional to the true likelihood, sampling from the approximation of $D(\boldsymbol{\theta})$ can be used to for more efficient approximation of the likelihood. The approximation considered is achieved by modelling $D(\boldsymbol{\theta})$ as a realisation of a Gaussian process.

Definition 1.1 (Gaussian Process)**.** *A collection of random variables $\{f(x)\}_{x\in\mathcal{X}}$ (where x may be a vector) is a* Gaussian process *if any finite subset of the collection of random variables is multivariate normal distributed. That is, there is a function $m : \mathcal{X} \to \mathbb{R}$ and symmetric kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all finite sets $\mathbf{x} := \{x_1, x_2, \dots, x_n\} \subset \mathcal{J}$, with $f(\mathbf{x}) := [f(x_1), f(x_2), \dots, f(x_n)]^T$*

$$f(\mathbf{x}) \sim \mathrm{MVN}\left(\begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \dots & k(x_1,x_n) \\ k(x_2,x_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(x_n,x_1) & \cdots & \cdots & k(x_n,x_n) \end{bmatrix} \right).$$

Definition 1.2 (Mean and Covariance Function)**.** *The the* mean function *and* covariance kernel *are*

$$m(x_i) := \mathbb{E}\left[f(x_i)\right]$$

and

$$k(x_i, x_{i'}) := \mathrm{cov}\left(f(x_i), f(x_{i'})\right).$$

Although Gaussian processes are simultaneously realised over the whole space $\mathcal{X}$ (for example $\mathbb{R}^d$) and are hence collections of (uncountably infinite) random variables, the choice of covariance function $\mathrm{corr}(x, x') \to 1$ as $||x - x'|| \to 0$ induces continuity in x almost surely. Therefore they can be thought of as realisations of continuous functions. (Should I prove this?)

Some common examples of Gaussian processes include

1. Brownian motion on $\mathbb{R}$:
$$m \equiv 0, \quad \text{and} \quad k(s,t) = \min(s,t)$$
2. Ornstein Uhlenbeck process with parameters θ and σ:
$$m \equiv 0, \quad \text{and} \quad k(s,t) = \frac{\sigma_k^2}{2\theta}\left(e^{-\theta|t-s|} - e^{-\theta(t+s)}\right)$$

Properties such as the smoothness and undulation of the realised functions are also determined by the covariance kernel k, and associated hyperparameters. Before exploring different kernel options, we begin by defining a valid kernel function, and formalising 'smoothness.'

Definition 1.3 (Positive Semi-Definite Matrix). *An $n \times n$ matrix $\mathbf{A}$ is* positive semi-definite *if $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^n$.*

Theorem 1.1 (Sufficient Condition for Positive Semi-Definite). *A symmetric matrix $\mathbf{A}$ is positive semi-definite, if (and only if) it's eigenvalues are non-negative.*

Proof. □

Definition 1.4 (Positive Semi-Definite Kernel). *A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is* positive semi-definite *if the matrix*

$$\mathbf{K} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(x_n, x_1) & \dots & \dots & k(x_n, x_n) \end{bmatrix}$$

is positive semi-definite for any collection of $x_i \in \mathcal{X}$

Theorem 1.2. *All symmetric positive semi-definite matrices are covariance matrices for some set of random variables*

Proof. □

Definition 1.5 (Mean Square Continuous). *A function $f : \mathbb{R}^d \to \mathbb{R}$ is* mean square continuous *at $\mathbf{x}$ in the ith direction at if $\mathbb{E}(|f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})|^2) \to 0$ as $|h| \to 0$, where $\mathbf{e}_i$ is the unit vector with a 1 in the ith coordinate.*

Definition 1.6 (Mean Square Differentiable). *A function $f : \mathbb{R}^d \to \mathbb{R}$ is* mean square differentiable *at $\mathbf{x}$ in the ith direction with derivative $\frac{\partial f(\mathbf{x})}{\partial x_i}$ if*

$$\mathbb{E}\left[\left|\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \frac{\partial f(\mathbf{x})}{\partial x_i}\right|^2\right] \to 0$$

as $|h| \to 0$, where $\mathbf{e}_i$ is the unit vector in the direction of the ith coordinate.

The concept of mean square differentiability and continuity are analogous to differentiability and continuity in the non-random function case.

Theorem 1.3. *Brownian motion is mean square continuous, but not mean square differentiable.*

Proof. $(B_{t+h} - B_t)^2 \sim (\sqrt{|h|}Z)^2$ where $Z \sim N(0, 1)$. Therefore $(B_{t+h} - B_t)^2 \sim |h|\chi_1^2 \to 0$ almost surely as $|h| \to 0$, and hence $\mathbb{E}[(B_{t+h} - B_t)^2 = 0]$. Since $\frac{B_{t+h} - B_t}{h} \sim N(0, 1/|h|)$, $\frac{B_{t+h} - B_t}{h}$ does not converge to any valid probability distribution as $|h| \to 0$, as the variance approaches $+\infty$. □

Some common

Theorem 1.4 (Positive Semi-Definiteness of RBF). *The radial basis function $\sigma_k^2 \exp(-\frac{(x-x')^2}{2\gamma^2})$ is a positive semi-definite kernel.*

Theorem 1.5 (Bochner's Theorem). *Let k be a stationary kernel function such that $k(x, x') = f(d)$. A function $k : \mathbb{R}^d \to \mathbb{C}$ is the covariance function of a weakly stationary mean square continuous complex-valued random process of $\mathbb{R}^d$ if and only if it can be represented as*

$$k(\tau) = \int_{\mathbb{R}^d} \exp(2\pi i \mathbf{s} \cdot \tau)$$

(Rasmussen and Williams 2008, p. 82)

1.2 Families of Kernel Function

The two most common families of kernel functions are the squared exponential and Matérn families.

Matérn Family

The Matérn exponential kernel is of the form

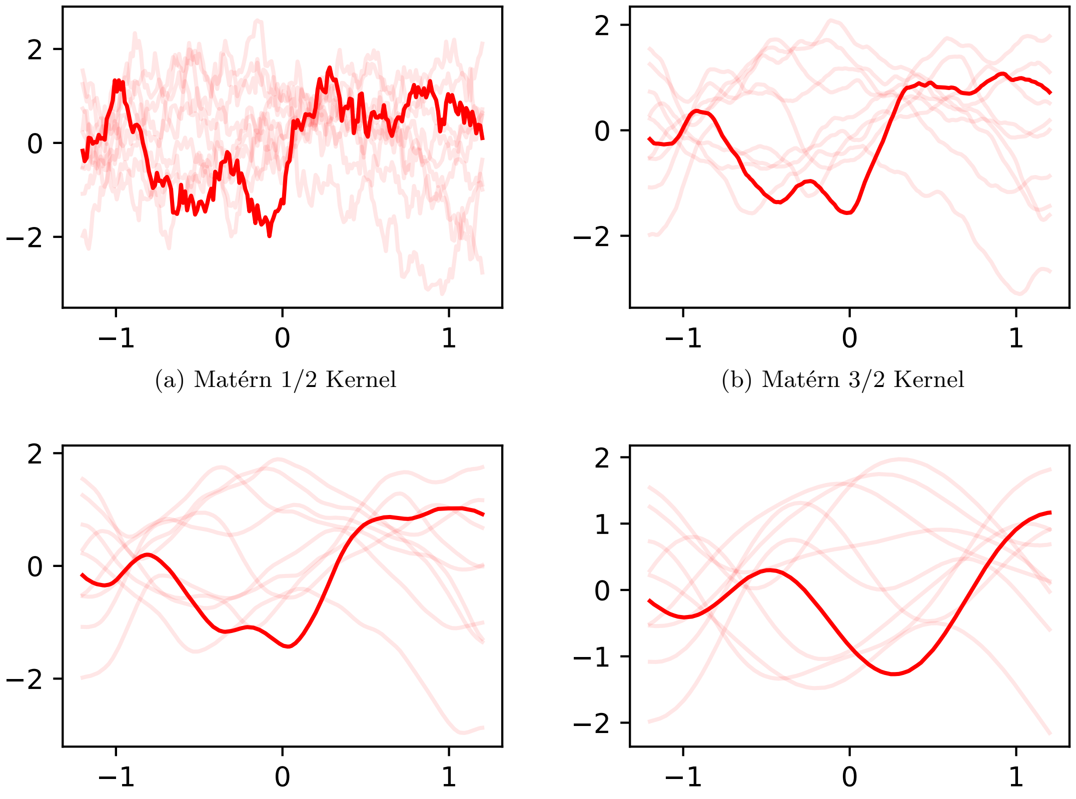
$$k_\nu(x,x') = \sigma_k^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||x-x'||}{\ell} \right)^\nu K_\nu \left(-\frac{\sqrt{2\nu}||x-x'||}{\ell} \right)$$

where K_ν is a modified Bessel function (defined in Abramowitz and Stegun 2013, p. 374). The general form is not very insightful, however for $\nu = 1/2, 3/2$ and $5/2$, (the most common values used) the kernel can be written as:

$$k_{1/2}(x,x') = \sigma_k^2 \exp\left(-\frac{||x-x'||}{\ell} \right)$$

$$k_{3/2}(x,x') = \sigma_k^2 \left(1 + \frac{\sqrt{3}||x-x'||}{\ell} \right) \exp\left(-\frac{\sqrt{3}||x-x'||}{\ell} \right)$$

$$k_{5/2}(x,x') = \sigma_k^2 \left(1 + \frac{\sqrt{5}||x-x'||}{\ell} + \frac{5||x-x'||}{3\ell^2} \right) \exp\left(-\frac{||x-x'||^2}{2 * \ell^2} \right)$$



(a) Matérn 1/2 Kernel

(b) Matérn 3/2 Kernel

(c) Matérn 5/2 Kernel

(d) Squared Exponential Kernel

Figure 1.1: Ten sample realisations from 4 different kernels with hyperparameters $\ell = 1$, and $\sigma_o^2 = 1$. One realisation is bolded. Samples for each kernel were generated from the same seed.

Zero mean Gaussian processes with a Matérn kernel are n times mean square differentiable, for all $n < \nu$. As seen in Figure 1.1, this means that this kernel allows for flexibility in how smooth realised functions are. As $\nu \to \infty$, with appropriate rescaling, the limit of the Matérn kernel is the squared exponential kernel.Rasmussen and Williams 2008, p. 85 Proof in CHAPTER 4 SKOROKHOD STOCHASTIC I?

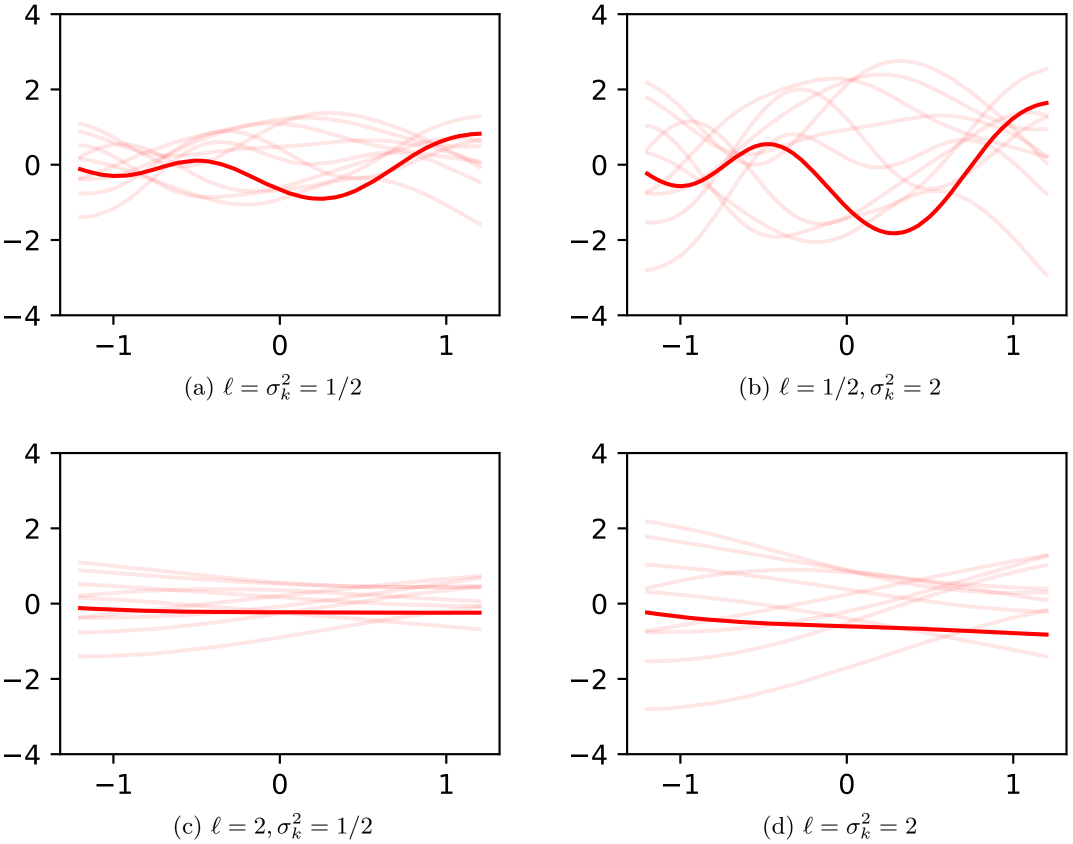
Squared Exponential Kernel

The squared exponential kernel is of the form

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

As the limit of Matérn kernels, the squared exponential kernel is infinitely mean square differentiable. Despite this being the 'default' kernel in much of the literature, infinite differentiability is a very strong condition on functions which are very smooth, which can be seen in Figure 1.1d

Length and Amplitude Hyperparameters



(a) $\ell = \sigma_k^2 = 1/2$

(b) $\ell = 1/2, \sigma_k^2 = 2$

(c) $\ell = 2, \sigma_k^2 = 1/2$

(d) $\ell = \sigma_k^2 = 2$

Figure 1.2: Ten realisations of zero mean Gaussian processes with the squared exponential kernel, varying the length and amplitude parameters. The samples were generated using the same seed

In both the Matérn and squared quadratic kernels (as well as most other common kernels choices), there are two hyperparameters ℓ and σ_k^2 which are referred to as length and amplitude hyperparameters. ℓ determines how close two points need to be to be highly correlated. Larger values of ℓ generates functions with higher correlation within a larger neighbourhood, as seen in Figure 1.2. σ_k^2 does not impact the correlation between x and x', but scales the correlation matrix. In other words, larger σ_k^2 increase the size but not rate of fluctuations. This can be seen comparing Figure 1.2a to Figure 1.2b.

Chapter 2

Gaussian Process Regression

Given the set of observations $f(\mathbf{x}_*)$ for the set of indices $\mathbf{x}_*$, it is often desirable to infer information about the function values at unobserved values. By choosing a Gaussian process with fixed kernel and hyperparameters, we can condition the process on the observed data, limiting the family of possible functions that we assume truly describe the model. Under the assumption that the function is a realisation of a Gaussian process predicting unseen function values reduces to elementary linear algebra. This is because a conditional multivariate normal distribution is still multivariate normal, and so the distribution of unobserved points will be multivariate normal.

Consider

$$\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix} \right)$$

Theorem 2.1 (Conditional Multivariate Normal Distribution is Multivariate Normal)**.** *With $f(\mathbf{x})$ and $f(\mathbf{x}_*)$, the conditional distribution is*

$$f(\mathbf{x})|f(\mathbf{x}_*) \sim \mathcal{N}\left(m(\mathbf{x}) + K_*K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)),\ K - K_*K_{**}^{-1}K_*^T\right).$$

Proof. Since marginal distribution of the multivariate normal distribution, is also multivariate normal, $f(\mathbf{x}_*) \sim \mathcal{N}(m(\mathbf{x}_*), K)$. Let the inverse of $\begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix}$ be defined as

$$\begin{bmatrix} \tilde{K} & \tilde{K}_* \\ \tilde{K}_*^T & \tilde{K}_{**} \end{bmatrix} = \begin{bmatrix} (K - K_*K_{**}^{-1}K_*^T)^{-1} & -(K - K_*K_{**}^{-1}K_*^T)^{-1}K_*K_{**}^{-1} \\ -K_{**}^{-1}K_*^T(K - K_*K_{**}^{-1}K_*^T)^{-1} & K_{**}^{-1} + K_{**}^{-1}K_*^T(K - K_*K_{**}^{-1}K_*^T)^{-1}K_*K_{**}^{-1} \end{bmatrix}$$

by the inverse of a block matrix. Therefore

$$\begin{aligned}
p(f(\mathbf{x})|f(\mathbf{x}_*)) =& \frac{p(f(\mathbf{x}), f(\mathbf{x}_*))}{p(f(\mathbf{x}_*))} \\
\propto& \frac{\exp\left[-\frac{1}{2}\left(\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*)\end{bmatrix}\right)^{\mathrm{T}} \begin{bmatrix} K & K_* \\ K_*^T & K \end{bmatrix}^{-1} \left(\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*)\end{bmatrix}\right)\right]}{\exp\left[-\frac{1}{2}(f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}} K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right]} \\
=& \exp\left[-\frac{1}{2}\left(\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*)\end{bmatrix}\right)^{\mathrm{T}} \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix}^{-1} \left(\begin{bmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*)\end{bmatrix} - \begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*)\end{bmatrix}\right)\right. \\
& \qquad \left. + \frac{1}{2}(f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}} K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right] \\
=& \exp\left[-\frac{1}{2}\Big((f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}} \tilde{K}(f(\mathbf{x}) - m(\mathbf{x}))\right. \\
& \qquad + 2(f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}} \tilde{K}_*(f(\mathbf{x}_*) - m(\mathbf{x}_*)) \\
& \qquad + (f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}} \tilde{K}_{**}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\Big) \\
& \qquad \left. + \frac{1}{2}(f(\mathbf{x}_*) - m(\mathbf{x}_*))^{\mathrm{T}} K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right] \\
\propto& \exp\left[-\frac{1}{2}(f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}} \tilde{K}(f(\mathbf{x}) - m(\mathbf{x}))\right. \\
& \qquad \left. - \ (f(\mathbf{x}) - m(\mathbf{x}))^{\mathrm{T}} \tilde{K}_*(f(\mathbf{x}_*) - m(\mathbf{x}_*))\right]. \\
& \text{(by removing the terms independent of } f(\mathbf{x}))
\end{aligned}$$

Since

$$p(\mathbf{z}) \propto \exp\left(-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z} + \mathbf{z}^T \mathbf{c}\right) \implies \mathbf{z} \sim \mathcal{N}\left(\Sigma\mathbf{c}, \Sigma\right),$$

$f(\mathbf{x}) - m(\mathbf{x})|f(\mathbf{x}_*)$ is multivariate normal with mean

$$\begin{aligned}
-\tilde{K}^{-1}\tilde{K}_*(f(\mathbf{x}_*) - m(\mathbf{x}_*)) =& (K - K_* K_{**}^{-1} K_*^T) \\
& \times (K - K_* K_{**}^{-1} K_*^T)^{-1} K_* K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)) \\
=& K_* K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))
\end{aligned}$$

and covariance matrix

$$\tilde{K}^{-1} = K - K_* K_{**}^{-1} K_*^T$$

by the alternative parametrisation of the multivariate normal distribution as a member of the exponential family of distributions (see Wikipedia contributors 2024, Table of Distributions). Finally, by the linearity of the multivariate normal mean,

$$f(\mathbf{x})|f(\mathbf{x}_*) \sim \mathcal{N}\left(m(\mathbf{x}) + K_* K_{**}^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)),\ K - K_* K_{**}^{-1} K_*^T\right).$$

□

After observing the function at multiple indices, we update the predictive distribution of any unobserved points, and generate new paths. The more points that the Gaussian process is conditioned on, the more narrow the sample paths, as seen in Figure 2.1

2.1 Observation Variance

For most functions, model outputs, or processes desirable for approximating through Gaussian process regression, multiple observations (through model runs or an real life measurements) of the

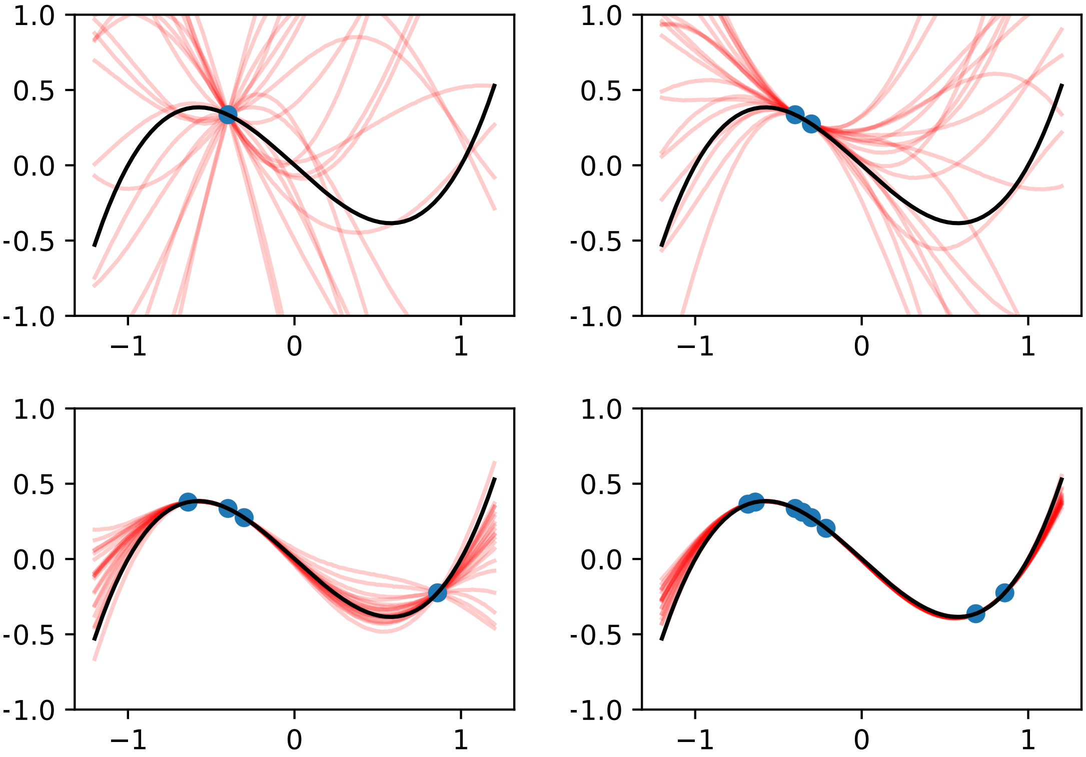


Figure 2.1: Sequence of Gaussian process regressions on the target function (black) $f(x) = x(x-1)(x+1)$, after 1, 2, 4, and 8 observations in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was zero mean and had a squared exponential kernel. The hyperparameters were fixed with $\ell = 2.7$ and $\sigma_k^2 = 1.1$

same point will result in different observations. This is in contrast to exact realisations as in Figure 2.1. The simplest assumption is that the observations are of the form

$$f_o(\mathbf{x}_*) = f(\mathbf{x}_*) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_o^2 I)$. Under these assumptions, $\mathrm{Cov}(f_o(\mathbf{x}_*), f_o(\mathbf{x}_*)) = K_{**} + \sigma_o^2 I$, where $K_{**} = \mathrm{Cov}(f(\mathbf{x}_*), f(\mathbf{x}_*))$ matrix of $f(\mathbf{x}_*)$ without noise. Therefore the conditional distribution of our unobserved function outputs given noisy observations

$$f(\mathbf{x})|f_o(\mathbf{x}_*) \sim \mathcal{N}\left(m(\mathbf{x}) + K_*(K_{**} + \sigma_o^2 I)^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)),\ K - K_*(K_{**} + \sigma_o^2 I)^{-1} K_*^T\right).$$

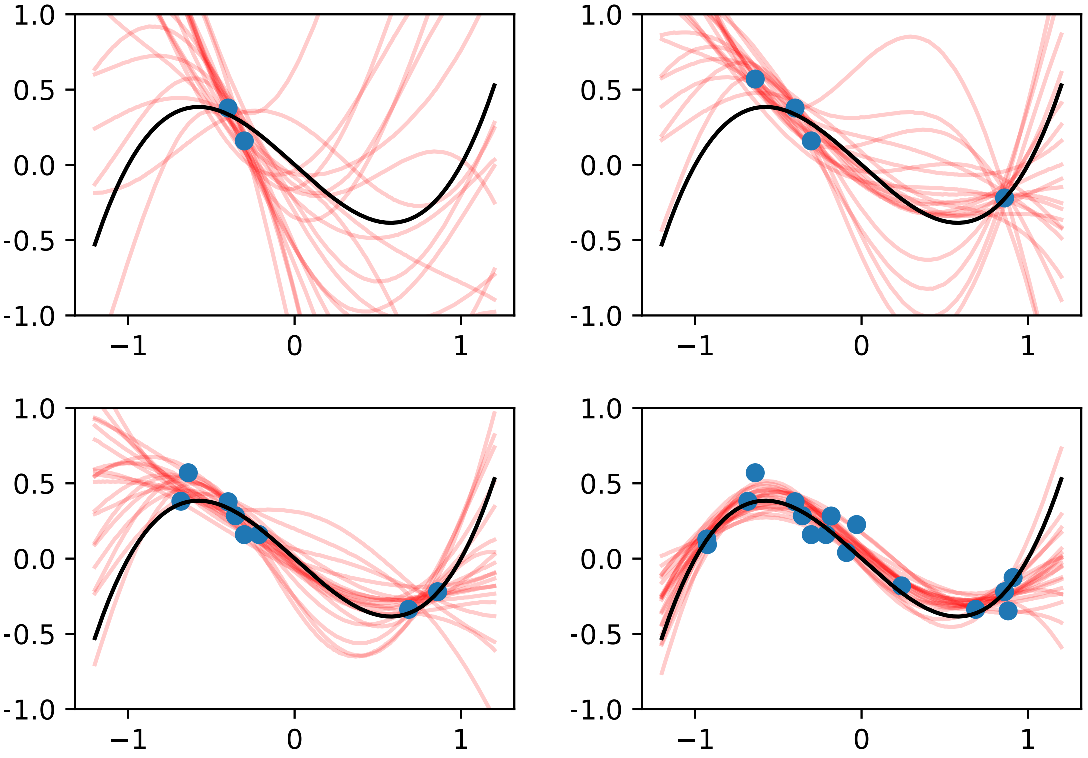


Figure 2.2: Sequence of Gaussian process regressions on the target function (black) $f(x) = x(x-1)(x+1)$, after 2, 4, 8, and 16 observations of $f(x_i) + \varepsilon_i$, where ε_i is i.i.d. $\mathcal{N}(0, \sigma_o^2)$ with $\sigma_o^2 = 0.01$ in blue. The red lines are new realisations from the conditioned Gaussian process. The Gaussian process was 0 mean and had a squared exponential kernel. The hyperparameters were fixed with $\ell = 2.7$ and $\sigma_k^2 = 1.1$

Adding noise to the observations makes the predictive distributions means that a single observation gives less information about the underlying function, and hence the predictive distributions of unseen data are much less stable, as empirically seen in Figure 2.2.

2.2 Model Selection

Kernel Family

The appropriate choice of kernel will depend on the properties behaviour of the target function to approximate. In the case of estimating an extremely stochastic distribution (such as the price of a stock over time), it is unlikely to be smooth, so no mean square differentiability is required, and a Matérn 1/2 kernel would be appropriate. If it is known that our target function is extremely smooth, such as a finite sum of infinitely differentiable functions (such as polynomials, sin, cos etc.) then the choice of squared exponential kernel is the most appropriate kernel. Realistically, the smoothness of the function will not be known a priori, and hence some sort of compromise (such as Matérn 5/2) kernel allows for flexibility.

Many other kernels exist that induce varying behaviours, such as periodic kernels find periodic kernel, and non-stationary kernels (where the covariance is dependent on x and x', not just $|x-x'|$) have i defined stationary?.

Hyperparameters

The hyperparameters ℓ and σ_k^2 for a choice of kernel have to be are not known beforehand, unless the function is actually a realisation from a Gaussian process. Similarly, the observation variance σ_o^2 hyperparameter may not be a priori known. There are two main (frequentist) ways to fit these hyperparameters: maximum likelihood estimation, and leave-one-out cross validation.

Defining the likelihood $\mathcal{L}(\ell, \sigma_k^2, \sigma_o^2) := p(f(\mathbf{x}_*)|\ell, \sigma_k^2, \sigma_o^2)$ in the usual way, the maximum likelihood estimates are

$$\{\hat{\ell}, \hat{\sigma}_k^2, \hat{\sigma}_o^2\} := \underset{\{\ell, \sigma_k^2, \sigma_o^2\}}{\arg\max}\, \mathcal{L}(\ell, \sigma_k^2, \sigma_o^2)$$

which is equivalent to minimising

$$-\ln(\mathcal{L}) = \frac{1}{2}\left[\ln(|K_{**}(\ell, \sigma_k^2) + \sigma_o^2|\,) + (f(\mathbf{x}_*) - m(\mathbf{x}_*))^T (K_{**}(\ell, \sigma_k^2) + \sigma_o^2)^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*)) + c\right].$$

The covariance matrix generated by the choice of kernel K_{**} is explicitly written with its dependence on ℓ and σ_k^2. c is a constant.

Leave-one-out cross validation aims to maximise the predictive log probability.

$$\{\tilde{\ell}, \tilde{\sigma}_k^2, \tilde{\sigma}_o^2\} := \underset{\ell, \sigma_k^2, \sigma_o^2}{\arg\max} \sum_i \ln p(f_i(\mathbf{x}_*)|f_{-i}(\mathbf{x}_*), \ell, \sigma_k^2, \sigma_o^2),$$

where $f_i(\mathbf{x}_*)|f_{-i}(\mathbf{x}_*)$ is the distribution of the ith element of $f(\mathbf{x}_*)$ conditioned on the rest of the observed data excluding that element (represented by $f_{-i}(\mathbf{x}_*)$). $f_i(\mathbf{x}_*)|f_{-i}(\mathbf{x}_*)$ can be found by Theorem 2.1. Computationally efficient methods for calculating the predictive log probability that avoid having to invert the covariance matrix for every summand element exist. In particular it can be shown that $f_i(\mathbf{x}_*)|f_{-i}(\mathbf{x}_*)$ has mean

$$f_i(\mathbf{x}_*) - m_i(\mathbf{x}_*) - [(K_{**} + \sigma_o^2 I)^{-1}(f(\mathbf{x}_*) - m(\mathbf{x}_*))]_i / [(K_{**} + \sigma_o^2 I)^{-1}]_{ii}$$

and variance $1/[(K_{**} + \sigma_o^2 I)^{-1}]_{ii}$, where both the mean and covariance are (surprisingly) independent of $f_i(\mathbf{x}_*)$ (Rasmussen and Williams 2008).

Both methods can be extended to include estimating hyperparameters given a family of possible mean functions. For example Gutmann and Cor 2016 use maximimum likelihood estimates for the amplitude and turning points of the quadratic mean functions.

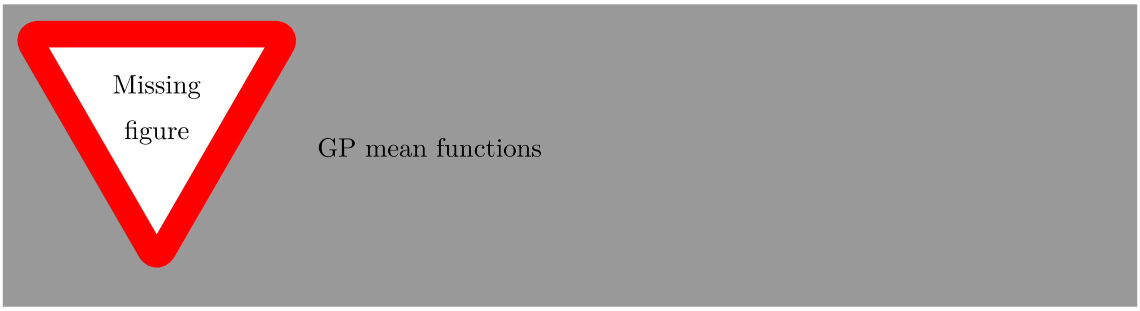
Recent work has shown that at least under specific conditions, the leave-one-out estimates for the scale hyperparameter are more robust to a larger family of target functions (Naslidnyk et al. 2024), and the broader literature seems to favor leave-one-out cross validation.

Finally there is scope for a Bayesian approach to model selection. By setting priors on the hyperparameters to be estimated and using the likelihood as described in the maximum likelihood estimation approach, a posterior distribution can be easily contrived. A set of samples from this posterior could then be taken, or a less prinicipled maximum a posteriori probability estimate could then be taken for a point estimate of the hyperparameters. This approach has some obvious benefits, particularly when taking a posterior sample. Most obviously is that since multiple values for each hyperparameter are sampled, the set of functions after

2.3 Differing mean functions

2.4 Bayesian Acquisition Functions

Under the assumptions that making observations from the underlying function is costly, and we care about regions of the function with high (or low) values, new observations should be taken



where there is high probability the function will be low. There also needs to be a trade-off between observing from areas with high predictive mean, and high predictive variance. These ideas are formalised by Bayesian acquisition functions $\mathcal{A}(x)$, with larger values corresponding to a higher 'desirability.' The target function is then sampled at the x which maximises this aquisition function. The new observation is then incorporated into the acquisition function.

Upper Confidence Bound

The upper confidence bound is one common way of exploring this trade off. The upper confidence bound

$$\mathcal{A}_{\text{UCB}}(x) := \mathbb{E}[f(x)|f(\mathbf{x}_*)] + \eta_t \sqrt{\text{Var}[f(x)|f(\mathbf{x}_*)]}$$

Similarly we can also define

- BOLFI paper uses
$$\mu(\boldsymbol{\theta}) - \eta_t \sqrt{\text{v}(\boldsymbol{\theta})}$$
  - $\eta_t := \sqrt{c + 2\ln(t^{d/2+2})}$, and c can be chosen
  - $\mu(\boldsymbol{\theta})$ and $\text{v}(\boldsymbol{\theta})$ are the posterior mean and variance
- Could use expected information
$$(\mu_{\text{min}} - \mu(\boldsymbol{\theta}))\mathit{\Phi}\left(\frac{\mu_{\text{min}} - \mu(\boldsymbol{\theta})}{\sqrt{\text{v}(\boldsymbol{\theta})}}\right) + \sqrt{\text{v}(\boldsymbol{\theta})}\phi\left(\frac{\mu_{\text{min}} - \mu(\boldsymbol{\theta})}{\sqrt{\text{v}(\boldsymbol{\theta})}}\right)$$
  - $\mu_{\text{min}} := \min_{\boldsymbol{\theta}} \mu(\boldsymbol{\theta})$
  - $\mathit{\Phi}, \phi$ CDF and PDF of standard normal

exploration parameter proven by Srinivas et al. 2010

Bibliography

Abramowitz, Milton and Irene A. Stegun, eds. (2013). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables.* 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover books on mathematics. New York, NY: Dover Publ. 1046 pp. ISBN: 978-0-486-61272-0.

Gutmann, Michael U. and Jukka Cor (2016). "Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models". In: *Journal of Machine Learning Research* 17.125, pp. 1–47. ISSN: 1533-7928. URL: `http://jmlr.org/papers/v17/15-017.html` (visited on 04/28/2024).

Naslidnyk, Masha et al. (2024). *Comparing Scale Parameter Estimators for Gaussian Process Interpolation with the Brownian Motion Prior: Leave-One-Out Cross Validation and Maximum Likelihood.* arXiv: `2307.07466 [math.ST]`.

Rasmussen, Carl Edward and Christopher K. I. Williams (2008). *Gaussian processes for machine learning.* 3. print. Adaptive computation and machine learning. Cambridge, Mass.: MIT Press. 248 pp. ISBN: 978-0-262-18253-9.

Srinivas, Niranjan et al. (2010). "Gaussian process optimization in the bandit setting: no regret and experimental design". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning.* ICML'10. Haifa, Israel: Omnipress, pp. 1015–1022. ISBN: 9781605589077.

Wikipedia contributors (2024). *Exponential family — Wikipedia, The Free Encyclopedia.* [Online; accessed 16-May-2024]. URL: `https://en.wikipedia.org/w/index.php?title=Exponential_family&oldid=1202463189`.