Kernels, Gaussian Processes and Applications

by

Michael Yiu

MA395 Essay

Supervised by Dr. Tim Sullivan

Mathematics Institute

WARWICK
THE UNIVERSITY OF WARWICK

Contents

1 Introduction

Consider the problem of performing regression on the following dataset:

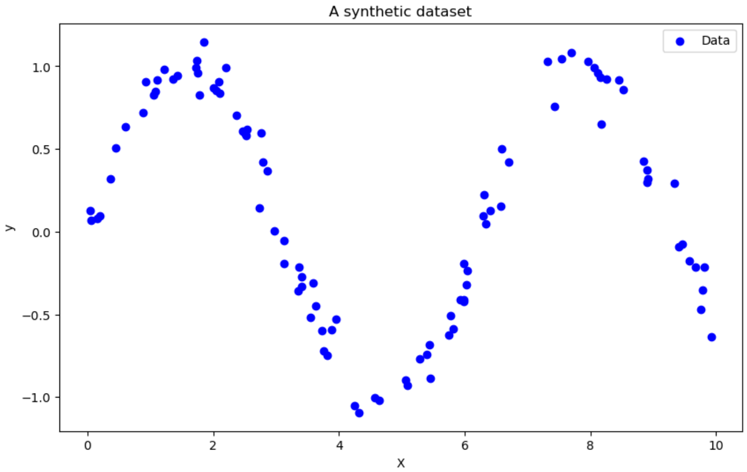

Figure 1: A simulated dataset

As a first attempt, we fit two models using linear regression: a simple linear regression model and a polynomial regression[1] with degree 4 (since there are 3 turning points in the data). The results are shown below:

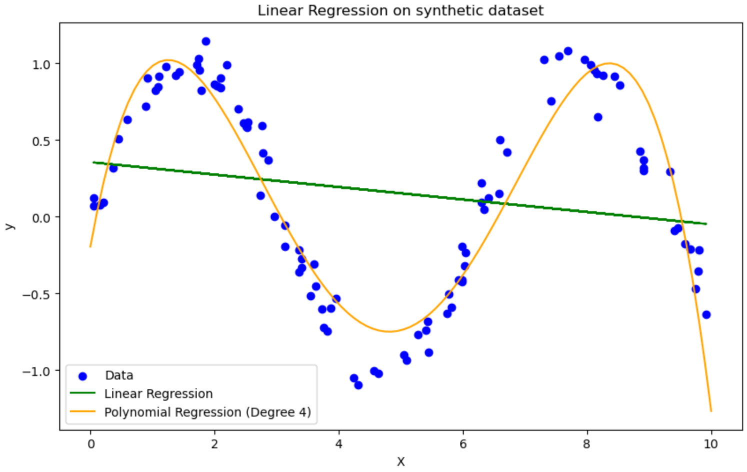

Figure 2: Linear regression on given dataset

As we can see, both model clearly do not fit well to the given data, and hence will not be able to perform predictions in a reliable manner. This is because we synthetically generated

[1] Polynomial regression is considered as a special case of linear regression since it is linear in parameters, even though the relationship between the input and output is non-linear – see Section 4.1

a sinusoidal dataset which cannot be accurately captured by linear models. So how can we deal with this non-linearity?

One approach is by mapping the data to a new (usually higher dimensional, potentially infinite-dimensional) vector space which allows techniques from linear regression to operate effectively. However, the main difficulty with this approach is that defining and computing the transformed features (or inputs) could be difficult, or even impractical when the feature space is infinite-dimensional. To address this problem, Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik introduced the *kernel trick* in their seminal paper on Support Vector Classification in 1992[2] – the idea is to introduce a *kernel function*, originated from functional analysis, which computes the inner product between features in the feature space without the need of evaluating the transformed features. This allows many supervised-learning algorithms such as ridge regression (Section 4.4) and support vector machines, which can be expressed purely in terms of inner products, to be kernelised and enable us to use these linear algorithms to learn non-linear functions or decision boundaries.

Going back to the dataset in Figure 1, if we use *kernel ridge regression* (a kernalised version of ridge regression, covered in Section 4.5) with a Gaussian kernel instead of linear models, we can see that it achieves the best fit among these 3 different regression models.

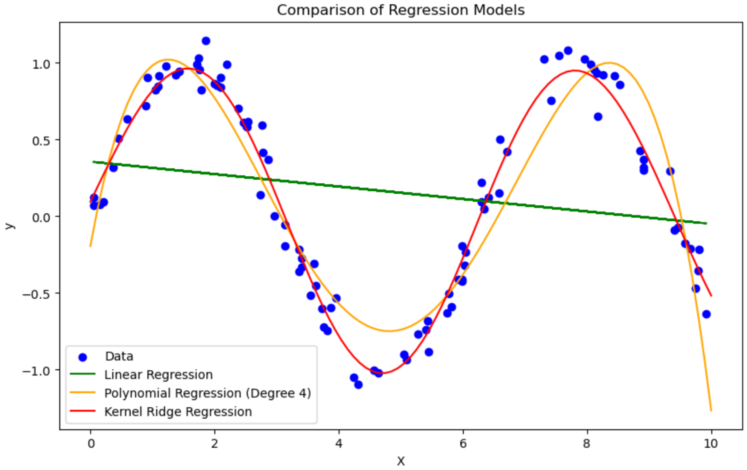

Figure 3: Kernel Ridge Regression

Due to this ability of extending linear techniques to non-linear settings, kernel methods have a wide range of applications in statistics and machine learning, including Kernel SVMs for classification, Kernel Ridge Regression, Kernel Principal Component Analysis (PCA)[3] for dimensionality reduction and Kernel Means Embedding for two-sample testing [4, Lecture 4-5 notes].

As we will see in Section 4.2, Kernel Ridge Regression is a Frequentist non-parametric approach to regression, and the Bayesian counterpart of this is *Gaussian Process regression*

[2] Boser, B, Guyon, I & Vapnik, V (1992). *A Training Algorithm for Optimal Margin Classifiers.* Association for Computing Machinery, New York, NY, USA. `https://doi.org/10.1145/130385.130401`

[3] Schölkopf, B., Smola, A., Müller, KR. (1997). *Kernel principal component analysis.* Lecture Notes in Computer Science, vol 1327. Springer, Berlin, Heidelberg. `https://doi.org/10.1007/BFb0020217`

(Section 5.2). In words, a Gaussian Process is a sequence of random variables (a stochastic process) in which every finite collection has a multivariate Gaussian distribution. Gaussian Processes and kernels are closely related in both theoretical and applicable aspects: In the theoretical aspect, there is a one-to-one correspondence between the Gaussian probability measure and Gaussian Processes in a function space induced by a kernel, known as the *Reproducing Kernel Hilbert Space* (Section 3.4). In the application side, under specific conditions, kernel ridge regression and Gaussian process regression would produce the same regression function (Section 5.3). It is worth noting that Gaussian Processes do not only have applications in regression, but also in classification[4] and in numerical integration, which is known as Bayesian quadrature[5], which views numerical integration as a Bayesian inference task.

In this essay, we will first explore the mathematical theory of Kernels and Gaussian Processes, including the construction of the Reproducing Kernel Hilbert Space and how Gaussian Process is related to the Gaussian measure, which allows us to view it as a random variable taking values in a Hilbert Space. We will then see the application of Kernels and Gaussian Processes in the context of regression, in which some abstract ideas in the theoretical part will be applied in action. This essay will be theoretical in nature, focusing on the mathematics behind these methods rather than practical implementation of algorithms.

Acknowledgements

I would like to thank Dr. Tim Sullivan for supervising this essay, and Dr. Roger Tribe, who had been my personal tutor for 2 years, for providing constructive feedback on my essay. I also made extensive use of my notes from the courses *ST221 Linear Statistical Modelling* lectured by Dr. Elke Thonnes, *MA3K1 Mathematics of Machine Learning* lectured by Dr. Clarice Poon and *ST342 Mathematics of Random Events* lectured by Dr. Karen Habermann.

[4] [8, Section 3.3]

[5] Osborne, M., Garnett, R., Roberts, S., Hart, C., Aigrain, S. &amp; Gibson, N.. (2012). Bayesian Quadrature for Ratios. ¡i¿Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics¡/i¿, in ¡i¿Proceedings of Machine Learning Research¡/i¿ 22:832-840 Available from `https://proceedings.mlr.press/v22/osborne12.html`

2 Kernels and the Reproducing Kernel Hilbert Space

2.1 Some Functional Analysis

To formally discuss kernels, we need to first introduce some basic notions from functional analysis[6]. For simplicity, we will only consider vector spaces over $\mathbb{R}$, i.e. we take the field $\mathbb{K} = \mathbb{R}$.

Definition 2.1 (Completeness and Banach Spaces, [1, Section 3, Definition 18]). Let $(V, ||\cdot||)$ be a normed space.

- A sequence $(x_n)_{n\in\mathbb{N}} \subseteq V$ is said to be **Cauchy** if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $m, n \geq N \implies ||x_m - x_n|| < \varepsilon$;
- We say that $(V, ||\cdot||)$ is **complete** if every Cauchy sequence $(x_n)_{n\in\mathbb{N}}$ in V converges to some $x \in V$;
- A complete normed space is called a **Banach space**.

Definition 2.2 (Dense set, [1, Section 3, Definition 19]). Let $(V, ||\cdot||)$ be a normed space. A subset $\mathcal{X} \subseteq V$ is said to be **dense** if for every $v \in V$ and $\varepsilon > 0$ there exists $x \in \mathcal{X}$ such that $||x - v|| < \varepsilon$.

Definition 2.3 (Inner product, [1, Section 4, Definition 23]). Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be a **real inner product** if

(i) (Linearity in its first argument) $\langle \alpha_1 f_1 + \alpha_2 f_2, g\rangle_{\mathcal{H}} = \alpha_1\langle f_1, g\rangle_{\mathcal{H}} + \alpha_2\langle f_2, g\rangle_{\mathcal{H}}$ for all $f_1, f_2, g \in \mathcal{H}$ and $\alpha_1, \alpha_2 \in \mathbb{R}$;

(ii) (Symmetry) $\langle f, g\rangle_{\mathcal{H}} = \langle g, f\rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$;

(iii) (Positive semi-definiteness) $\langle f, f\rangle_{\mathcal{H}} \geq 0$ and $\langle f, f\rangle_{\mathcal{H}}$ if and only if $f = 0$ for all $f \in \mathcal{H}$.

A vector space $\mathcal{H}$ endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is call an **inner product space**. In the rest of this essay, we will simply use "linearity" when referencing property (i) due to symmetry in real inner products (this is not true for complex inner products, where $\langle f, g\rangle_{\mathcal{H}} = \overline{\langle g, f\rangle_{\mathcal{H}}}$).

We can define a norm using the inner product [1, Section 4, Definition 24]:

Definition 2.4 (Natural norm induced by an inner product). Let $\mathcal{H}$ be an inner product space. Then for $f \in \mathcal{H}$, the expression

$$||f||_{\mathcal{H}} := \sqrt{\langle f, f\rangle_{\mathcal{H}}}$$

is said to be the **natural norm** induced by the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

Theorem 2.1 (Cauchy-Schwarz inequality, [1, Section 4, Theorem 14]). Any inner product space $\mathcal{H}$ satisfies the **Cauchy-Schwarz inequality**: For all $f, g \in \mathcal{H}$,

$$|\langle f, g\rangle| \leq ||f||_{\mathcal{H}}||g||_{\mathcal{H}}$$

[6] Here I have included some basic functional analysis since I did not take the module *MA3G7: Functional Analysis I.*

With the notion of a Banach space and an inner product space, we can now define *Hilbert Spaces* [1, Section 5, Definition 29].

Definition 2.5 (Hilbert Space)**.** A **Hilbert Space** $\mathcal{H}$ is a complete inner product space, that is, $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ is a Hilbert Space if $(\mathcal{H}, ||f||_{\mathcal{H}})$ is a Banach space.

We give an example of Hilbert Space which we will use later in Section 2.3 [4, Theory lecture notes, Example 11]:

Example 2.1. Let $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)$ be a measure space, $d \geq 1$ and $\mathcal{X} \subseteq \mathbb{R}^d$. Then the space

$$L^2(\mathcal{X}, \mu) := \left\{ f : \mathcal{X} \to \mathbb{R} \text{ measurable} : ||f|| = \left(\int_{\mathcal{X}} |f|^2 \,\mathrm{d}\mu \right)^{\frac{1}{2}} < \infty \right\} \subseteq L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)$$

equipped with the inner product

$$\langle f, g \rangle = \int_{\mathcal{X}} fg \,\mathrm{d}\mu$$

is a Hilbert Space. If μ is the Lebesgue measure on $\mathbb{R}^d$ then we simply write $L^2(\mathcal{X})$.

Proposition 2.1 (Pointwise convergence implies norm convergence in Hilbert Space)**.** Let $\mathcal{H}$ be a Hilbert Space. Then for $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}$ and $f \in \mathcal{H}$,

$$\lim_{n\to\infty} f_n = f \implies \lim_{n\to\infty} ||f_n - f||_{\mathcal{H}} = 0.$$

Proof.

$$\begin{aligned}
||f_n - f||_{\mathcal{H}} &= \langle f_n - f, f_n - f \rangle_{\mathcal{H}} \\
&= \langle f_n, f_n - f \rangle_{\mathcal{H}} - \langle f, f_n - f \rangle_{\mathcal{H}} \quad \text{(by linearity)} \\
&= \langle f_n, f_n \rangle_{\mathcal{H}} - \langle f_n, f \rangle_{\mathcal{H}} - \langle f, f_n \rangle_{\mathcal{H}} + \langle f, f \rangle_{\mathcal{H}} \quad \text{(by symmetry and linearity)} \\
&= ||f_n||_{\mathcal{H}} - 2\langle f_n, f \rangle_{\mathcal{H}} + ||f||_{\mathcal{H}} \quad \text{(by Definition 2.4)} \\
&\to ||f||_{\mathcal{H}} - 2\langle f, f \rangle_{\mathcal{H}} + ||f||_{\mathcal{H}} \quad \text{as } n \to \infty \\
&= 0.
\end{aligned}$$

□

Definition 2.6 (Orthogonal complement, [1, Section 5, Definition 32])**.** Let $\mathcal{H}$ be a Hilbert space and $X \subseteq \mathcal{H}$. We define the **orthogonal complement** of X as

$$X^{\perp} := \{ y \in \mathcal{H} : \forall\, x \in X,\ \langle x, y \rangle_{\mathcal{H}} = 0 \}.$$

Definition 2.7 (Projection, [1, Section 5, Definition 33])**.** A linear map $P : \mathcal{H} \to \mathcal{H}$ is called a **projection** if $P \circ P = P$, that is, for all $x \in \mathcal{H}$, $P(P(x)) = P(x)$. P is said to be an **orthogonal projection** if for all $x \in \mathcal{H}$ and $y \in \mathcal{H}$ such that $P(y) = 0$, we have $\langle P(x), y \rangle_{\mathcal{H}} = 0$.

Proposition 2.2 ([1, Section 5, Proposition 21])**.** Let $X \subseteq \mathcal{H}$. Then $X^{\perp}$ is a **closed linear subspace** of $\mathcal{H}$, i.e. it is a linear subspace that contains all its limit points.

Theorem 2.2 (1, Section 5, Theorem 19)**.** Let U be a closed linear subspace of a Hilbert space $\mathcal{H}$. Then every $x \in \mathcal{H}$ can be uniquely written as

$$x = u + u^* \ \text{ for } u \in U \text{ and } u^* \in U^{\perp}.$$

Definition 2.8 (Functional, [1, Section 6, Definition 34])**.** Let V be a vector space over $\mathbb{R}$. A **functional** is a map $F : V \to \mathbb{R}$. A functional is said to be **linear** if for every $x, y \in V$ and $\alpha, \beta \in \mathbb{R}$ we have

$$F(\alpha x + \beta y) = \alpha F(x) + \beta F(y).$$

If V is a normed spaced equipped with the norm $|| \cdot ||$, then a linear functional is said to be:

- **bounded** if there exists $M > 0$ such that $|F(x)| \leq M||x||$ for every $x \in V$;
- **continuous** if

  $$||x_k - x|| \to 0 \implies |F(x_k) - F(x)| \to 0$$

  as $k \to \infty$ for every $x \in V$ and $(x_k)_{k \in \mathbb{N}} \subseteq V$.

A linear functional is bounded if and only if it is continuous [1, Section 6, Proposition 22].

Definition 2.9 (Dual of a normed space, [1, Section 6, Definition 36])**.** Let V be a a vector space over $\mathbb{R}$. Then V^*, the space of all continuous linear functionals is a linear subspace of $\mathbb{R}^V$ (the space of functions from V to $\mathbb{R}$) and thus is a vector space, and it is called the **dual** of V.

2.2 Kernels and Examples

Now we are ready to formally define kernels. This definition is taken from [2, Section 4.1 Definition 4.1]:

Definition 2.10 (Kernel)**.** Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ over $\mathbb{R}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$, we have

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

We call ϕ a **feature map** and $\mathcal{H}$ the **feature space** of k.

We can think of kernels as a measure of similarity between elements (objects) in $\mathcal{X}$: If the inner product is large then these features are probably well-aligned, i.e. the objects are similar. If the inner product is small then these features are almost orthogonal, so the objects are different.

It is worth noting that neither ϕ nor $\mathcal{H}$ are uniquely defined by k: Consider $\mathcal{X} = \mathbb{R}$ and $k(x, x') = xx' \in \mathbb{R}$. Then k is a kernel: we can simply take $\phi : \mathcal{X} \to \mathbb{R}$ be the identity map $\phi(x) = x$ with feature space $\mathcal{H} = \mathbb{R}$ endowed with the Euclidean dot product. However, we can also define $\phi' : \mathcal{X} \to \mathbb{R}^2$ by

$$\phi'(x) = \left(\frac{x}{\sqrt{2}}, \frac{x}{\sqrt{2}} \right).$$

Then ϕ' and $\mathcal{H} = \mathbb{R}^2$ endowed with the Euclidean dot product is also a feature map and feature space of k respectively:

$$\begin{aligned} \langle \phi'(x), \phi'(x') \rangle_{\mathcal{H}} &= \left(\frac{x}{\sqrt{2}}, \frac{x}{\sqrt{2}} \right) \left(\frac{x'}{\sqrt{2}}, \frac{x'}{\sqrt{2}} \right)^T \\ &= \frac{xx'}{2} + \frac{xx'}{2} \\ &= xx'. \end{aligned}$$

However, for each kernel k there exists a special feature space (the Reproducing Kernel Hilbert Space mentioned in the introduction) and a special feature map that are indeed uniquely defined by k – we will see this in Section 2.3.

To show that a function k is a kernel, we need to show that at least one feature space exists. Finding this feature space, or at the very least showing that it exists, can be somewhat difficult, but thankfully we have a criterion that characterises kernels in terms of inequalities via the notion of *positive semi-definite functions* [2, Section 4.1, Definition 4.15][7]. We start by introducing the notion of symmetric function, taken from [2, Section 4.1, Definition 4.15]:

Definition 2.11 (Symmetric function)**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function. We say that k is **symmetric** if $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$.

All kernels are symmetric: this follows directly from the symmetric property of an inner product.

Definition 2.12 (Positive semi-definite function)**.** A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be **positive semi-definite** if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and for all $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

k is said to be **positive definite** if, for mutually distinct $x_1, \ldots, x_n \in \mathcal{X}$, equality holds if and only if $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$.

All kernels are positive semi-definite, as shown in the following proposition [4, Lecture 1-2 Notes, Section 3.2, Lemma 14]:

Proposition 2.3 (Kernels are positive semi-definite)**.** Let $\mathcal{H}$ be a Hilbert space and $\mathcal{X}$ a non-empty set. Then for all $x_i, x_j \in \mathcal{X}$, $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$ is a positive semi-definite function.

Proof. (Adapted from [4, Lecture 1-2 Notes, Section 3.2, Lemma 14])

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \langle \phi(x_j), \phi(x_i) \rangle_{\mathcal{H}} \quad \text{(by definition)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \langle \alpha_j \phi(x_j), \alpha_i \phi(x_i) \rangle_{\mathcal{H}} \quad \text{(by linearity)} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\rangle_{\mathcal{H}} \quad \text{(by linearity and symmetry)} \\
&\geq 0 \quad \text{(by positive-definiteness of inner products)}
\end{aligned}$$

□

This proposition is very useful when it comes to proving various properties of kernels (which we will see later), but it does not provide a criterion for us to check whether any given function is a kernel. However, it turns out that the converse of Proposition 2.3 is also true: Any symmetric and positive semi-definite function is a kernel. This is given by the following theorem [2, Section 4.1, Theorem 4.16], which we will state without proof:

[7] The authors use the term *positive definite* functions instead, but I decided to use the term positive semi-definite functions to align with the notion of positive semi-definite matrices.

Theorem 2.3 (Characterisation of kernels)**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function. k is a kernel if and only if it is symmetric and positive semi-definite.

The proof of this theorem requires the notion of a pre-RKHS in Section 2.4, and can be found in [2, Section 4.1, p.118].

Theorem 2.3 essentially tells us that we can show that a function k is a kernel by simply showing it is symmetric and positive semi-definite without specifying a feature map and feature space. Observe that the definition of a positive semi-definite functions looks somewhat similar to that of a positive semi-definite symmetric matrix: If we define the **Gram matrix** K with entries $(K)_{ij} = k(x_i, x_j)$, then K is symmetric (which follows from the symmetry of kernels), and

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

is simply another way of writing $\alpha^T K \alpha \geq 0$. Combining this with Theorem 2.3, we define kernels via positive semi-definite matrices [3, Section 1.4, Definition 1]:

Definition 2.13 (Kernel, revisited)**.** A kernel k is a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ in which for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, the matrix K with entries $(K)_{ij} = k(x_i, x_j)$ is positive semi-definite, i.e. for all $a \in \mathbb{R}^n, a^T K a \geq 0$. We call the matrix K the **Gram matrix** of the kernel k. If k is positive definite, then its corresponding Gram matrix is positive definite, i.e. $a^T K a > 0$ for all $a \in \mathbb{R}^n$.

This implies we can show that k is a kernel just by checking whether its corresponding Gram matrix is positive semi-definite.

Now we are ready to prove some properties of kernels:

Lemma 2.1 (Sums of kernels, [2, Section 4.1, Lemma 5])**.** Let $\mathcal{X}$ be a non-empty set, $\alpha \geq 0$ and k, k_1 and k_2 be kernels on $\mathcal{X}$. Then

(i) αk is a kernel;

(ii) $k_1 + k_2$ is a kernel.

Remark. When we say k is a kernel on $\mathcal{X}$, we mean $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel.

Proof.

(i)

$$\begin{aligned}\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \alpha k(x_j, x_i) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \alpha \langle \phi(x_j), \phi(x_i) \rangle_{\mathcal{H}} \\ &= \alpha \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{j=1}^{n} \alpha_j \phi(x_j) \right\rangle_{\mathcal{H}} \\ &\geq 0.\end{aligned}$$

So αk is positive semi-definite. Again, αk is symmetric by symmetry of inner products. Hence by Theorem 2.3, αk is a kernel.

(ii)

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j[k_1(x_j,x_i)+k_2(x_j,x_i)] = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_1(x_j,x_i)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_2(x_j,x_i)$$
$$\geq 0.$$

So k_1 and k_2 are positive semi-definite. Since k_1 and k_2 are both symmetric, $k_1 + k_2$ is symmetric, and the result follows by Theorem 2.3.

□

Lemma 2.2 (Products of kernels, [2, Section 4.1, Lemma 4.6])**.** Let $\mathcal{X}$ be a non-empty set and k_1, k_2 be kernels on $\mathcal{X}$. Then $k := k_1 \cdot k_2$ is a kernel on $\mathcal{X}$.

Proof. Let K_1 and K_2 be Gram matrices of k_1 and k_2 respectively. By Spectral Decomposition theorem for symmetric matrices, there exists orthogonal matrices U, V and diagonal matrices Σ, Λ such that

$$K_1 = U\Sigma U^T = \sum_{k=1}^{n}\sigma_k u_k u_k^T$$
$$K_2 = V\Lambda V^T = \sum_{m=1}^{n}\lambda_m v_m v_m^T,$$

where σ_k's and λ_m's are the eigenvalues of K_1 and K_2 respectively and the diagonal entries of Σ and Λ, u_k's and v_m's are orthogonal eigenvectors of K_1 and K_2 and are columns of U, V respectively. Note that

$$(K_1)_{ij} = \sum_{k=1}^{n}\sigma_k u_{ik}u_{jk}$$

and

$$(K_2)_{ij} = \sum_{m=1}^{n}\lambda_m v_{im}v_{jm},$$

so we have

$$\begin{aligned}
a^T K a &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i,x_j)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sum_{k=1}^{n}\sigma_k u_{ik}u_{jk}\sum_{m=1}^{n}\lambda_m v_{im}v_{jm}\\
&= \sum_{k=1}^{n}\sum_{m=1}^{n}\sigma_k\lambda_m\sum_{i=1}^{n}a_i u_{ik}v_{im}\sum_{j=1}^{n}a_j u_{jk}v_{jm}\\
&= \sum_{k=1}^{n}\sum_{m=1}^{n}\sigma_k\lambda_m\left(\sum_{i=1}^{n}a_i u_{ik}v_{im}\right)^2\\
&\geq 0
\end{aligned}$$

since $\sigma_k, \lambda_m \geq 0$ for all $k, m \in \{1, \ldots, n\}$ since K_1 and K_2 are positive semi-definite. This shows that k is positive semi-definite. Since k_1 and k_2 are symmetric, we have

$k(x_i, x_j) = k_1(x_i, x_j)k_2(x_i, x_j) = k_1(x_j, x_i)k_2(x_j, x_i) = k(x_j, x_i)$, so k is symmetric. Hence k is a kernel on $\mathcal{X}$ by Theorem 2.3. □

This result can be generalised to the case when k_1 and k_2 are kernels on $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are non-empty sets. In this case, $k := k_1 \cdot k_2$ is a kernel on $\mathcal{X} \times \mathcal{Y}$. The proof is more sophisticated and requires tensor products – see [2, Section 4.1, Proof of Lemma 4.6].

Lemma 2.3 (Mapping between spaces, [4, Lecture 1-2 Notes, Lemma 5])**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be non-empty sets, $A : \mathcal{X} \to \mathcal{Y}$ be a map and k be a kernel on $\mathcal{Y}$. Then k' defined by $k'(x, x') = k(A(x), A(x'))$ is a kernel on $\mathcal{X}$.

Proof. Define the Gram matrix of k' by $(K')_{ij} = k(A(x_i), A(x_j))$ for some $\{x_1, \dots, x_n\} \in \mathcal{X}$. Let $a = (a_1, \dots, a_n)^T \in \mathbb{R}^n$. Then

$$\begin{aligned} a^T K' a &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k'(x_i, x_j) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(A(x_i), A(x_j)) \\ &\geq 0 \quad (k \text{ is a kernel on } \mathcal{Y}) \end{aligned}$$

So k' is positive semi-definite. Since k is a kernel,

$$k'(x, x') = k(A(x), A(x')) = k(A(x'), A(x)) = k'(x', x),$$

so k' is symmetric. By Theorem 2.3, k' is a kernel. □

Lemma 2.4 (Limits of kernels, [2, Section 4.1, Corollary 4.17])**.** Let $(k_n)_{n\in\mathbb{N}}$ be a sequence of kernels on the set $\mathcal{X}$ that converges pointwise to a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e.

$$\lim_{n\to\infty} k_n(x, x') = k(x, x')$$

for all $x, x' \in \mathcal{X}$. Then k is a kernel on $\mathcal{X}$.

Proof. Consider the sequence of Gram matrices $(K_n)_{n\in\mathbb{N}}$ that correspond to each term in the sequence $(k_n)_{n\in\mathbb{N}}$. By Definition 2.10 and Proposition 2.2, for all $a \in \mathbb{R}^m$ and $n \in \mathbb{N}$, we have

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j k_n(x_i, x_j) \geq 0.$$

Taking limit as $n \to \infty$ gives:

$$\begin{aligned} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \left(\lim_{n\to\infty} k_n(x_i, x_j) \right) &\geq 0 \\ \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j k(x_i, x_j) &\geq 0. \end{aligned}$$

Hence k is positive semi-definite. To show that k is symmetric, note that

$$\begin{aligned} \lim_{n\to\infty} k_n(x', x) &= k(x', x) \quad \text{(by assumption)} \\ \lim_{n\to\infty} k_n(x', x) &= k(x, x') \quad \text{(by symmetry of } k_n\text{)} \end{aligned}$$

Hence k is symmetric, and k is a kernel on $\mathcal{X}$. □

Lemma 2.1 - Lemma 2.4 allow us to construct a wide variety of kernels. Here are some examples, taken from [3, Section 1.4.1]:

Example 2.2 (Linear kernels)**.** Let $x, x' \in \mathbb{R}^n$ for $n \geq 1$. Then $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $k(x, x') = x^T x'$ is the **linear kernel**.

Proof. Take $\mathcal{H} = \mathbb{R}^n$ endowed with the Euclidean dot product, and $\phi : \mathcal{X} \to \mathcal{H}$ be the identity map completes the proof. □

The linear kernel is perhaps the simplest type of kernel, and also provide the basis of the kernel trick seen in the introduction: If a learning algorithm can be written purely in terms of dot products, then we can rewrite them using the linear kernel. We will use this to motivate Kernel ridge regression from the standard ridge regression in Section 4.5.

Example 2.3 (Polynomial kernels)**.** Let $x, x' \in \mathbb{R}^n$ where $n \geq 1$, $m \in \mathbb{Z}^+$ and $c \in \mathbb{R}^+$. Then $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$k(x, x') = (c + x^T x')^m$$

is a kernel.

Proof. Define $k_1 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by $k_1(x, x') = c$. Then the Gram matrix K_1 is a matrix with entries $(K)_{ij} = c$ for all $i, j \in \{1, \dots, n\}$. Since $c \geq 0$, K_1 is positive semi-definite. k_1 is symmetric by construction, so k_1 is a kernel by Theorem 2.3.
Since $c + x^T x$ is a sum of kernels, it is a kernel by Lemma 2.1.
Since m is a positive integer, k is a finite product of kernels, hence it is a kernel by Lemma 2.2. □

Remark. The linear kernel is a special case of the polynomial kernel with $c = 0$ and $m = 1$.

The polynomial kernel is an example of Euclidean dot product between finitely many features. The following example, however, is an inner product between infinitely many features:

Example 2.4 (Gaussian kernel)**.** For $x, x' \in \mathbb{R}^n$, we define the **Gaussian / Radial basis function (RBF) kernel** by

$$k(x, x') = \exp\left(-\frac{||x - x'||_2^2}{2\sigma^2}\right),$$

where $\sigma \in \mathbb{R}$ is the *length-scale* parameter. This is indeed a kernel.

Proof. Observe that

$$\begin{aligned} ||x - x'||_2^2 &= \sum_{i=1}^{n} (x_i - x'_i)^2 \\ &= ||x||_2^2 + ||x'||_2^2 - 2x^T x' \\ k(x, x') &= \exp\left(-\frac{||x||_2^2}{2\sigma^2}\right) \exp\left(-\frac{||x'||_2^2}{2\sigma^2}\right) \exp\left(\frac{x^T x'}{\sigma^2}\right) \end{aligned}$$

Define $k_1(x, x') = \exp\left(-\frac{||x||_2^2}{2\sigma^2}\right) \exp\left(-\frac{||x'||_2^2}{2\sigma^2}\right)$ and $k_2(x, x') = \exp\left(\frac{x^T x'}{\sigma^2}\right)$. Then

- k_1 is a kernel: Take $\phi : \mathcal{X} \to \mathcal{H}$ by $\phi(x) = \exp\left(-\frac{||x||_2^2}{2\sigma^2}\right)$.
- $k_2(x, x') = \exp\left(\frac{x^T x'}{\sigma^2}\right) = \sum_{k=1}^{\infty} \frac{1}{k!}\left(\frac{x^T x}{\sigma^2}\right) = \lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{k!}\left(\frac{x^T x}{\sigma^2}\right)$. Since k_2 is a limit of sums of kernels, it is a kernel by Lemma 2.4.

Hence $k(x, x') = k_1(x, x')k_2(x, x')$ is a kernel since it a product of two kernels by Lemma 2.2. □

To show that the Gaussian kernel has infinitely many features, we will need the following proposition, taken from [4, Lecture 1-2 Notes, Lemma 9]:

Proposition 2.4. *Let $\ell^2(\mathbb{R})$ be the space of square-summable sequences. Given a sequence of functions $(\phi_n(x))_{n\in\mathbb{N}} \subseteq \ell^2(\mathbb{R})$ where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the i-th coordinate of the feature map ϕ. Then*

$$k(x, x') = \sum_{n=1}^{\infty} \phi_n(x)\phi_n(x') < \infty.$$

Proof. (Adapted from same source)

By Cauchy-Schwartz inequality,

$$\left|\sum_{n=1}^{\infty} \phi_n(x)\phi_n(x')\right| \leq ||\phi(x)||_{\ell^2}||\phi(x')||_{\ell}^2 < \infty,$$

where $||\phi(x)||_{\ell^2} = \sum_{n=1}^{\infty} \phi_n(x)^2$. Since $(\phi_n(x))_{n\in\mathbb{N}}$ is square-summable, the sequence defining the inner product converges for all $x, x \in \mathcal{X}$. □

Now we restrict ourselves to the case where $\sigma = 1$ and consider the sequence

$$(\phi_n(x))_{n\in\mathbb{N}} = \exp\left(-\frac{1}{2}x^2\right)\frac{x^n}{\sqrt{n!}}$$

and the inner product defined by

$$\langle u, v\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} u_i v_i.$$

Then

$$\begin{aligned}\langle \phi_n(x), \phi(x')\rangle_{\mathcal{H}} &= \exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}x'^2\right)\sum_{n=0}^{\infty}\frac{x^n x'^n}{n!}\\ &= \exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}x'^2\right)\exp(xx')\\ &= \exp\left(-\frac{1}{2}(x^2 - 2xx' + x'^2)\right)\\ &= \exp\left(-\frac{||x - x'||^2}{2}\right)\\ &= k(x, x').\end{aligned}$$

so the Gaussian kernel can be written as an infinite series of the features ϕ, which implies the feature space is infinite-dimensional.

Example 2.5 (Sobolev kernel)**.** Take $\mathcal{X} = [0, 1]$ and let $k : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be defined by $k(x, x') = x \wedge x' = \min\{x, x'\}$. Then k is a kernel.

Proof. Since $x \wedge x'$ is the covariance function of the standard Brownian motion (Section 3.5, Proposition 3.1), it must be positive semi-definite. □

2.3 The Reproducing Kernel Hilbert Space

As mentioned in Section 2.2, feature maps and feature spaces are not uniquely defined by a kernel k. The question one might ask is whether there exists a feature space and feature map that satisfy certain "nice" properties that are indeed uniquely defined by k. It turns out that there does exist such feature space, the so-called *Reproducing Kernel Hilbert Space* mentioned in the introduction and the corresponding feature map is known as the *canonical feature map*.

Before we formally introduce these concepts, we shall clarify some notation that we are going to use: We use $f(\cdot)$ to indicate we consider the function itself, and not just the function evaluated at a particular point. For the kernel $k(x_i, \cdot)$, we think of it as a function fixed at one argument x_i, and hence is a function that maps from $\mathcal{X}$ to $\mathbb{R}$.

Definition 2.14 (Reproducing kernels and RKHS, [2, Section 4.2, Definition 4.18])**.** Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert Space that consist of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if

(i) $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$;

(ii) (Reproducing property) $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ holds for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.

The space $\mathcal{H}$ is called a **Reproducing Kernel Hilbert Space** (RKHS) over $\mathcal{X}$ if for all $x \in \mathcal{X}$, the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by $\delta_x(f) := f(x)$ is continuous.

Remark. The reproducing property implies that we can evaluate functions in the RKHS at any point x by taking the inner product of f with the kernel $k(x, \cdot)$ fixed at x. In particular, for all $x, y \in \mathcal{X}$, we have $k(x, \cdot) \in \mathcal{H}$ and $k(\cdot, y) \in \mathcal{H}$ by (i). By the reproducing property,

$$\langle k(x, \cdot), k(\cdot, y)\rangle_{\mathcal{H}} = k(x, y).$$

We have seen earlier in Proposition 2.1 that in a Hilbert Space, pointwise convergence implied convergence in the Hilbert Space norm. The converse is not true in general (as shown in the remark below), but is true when the Hilbert Space is an RKHS, as shown in the following proposition [4, Theory lecture notes, Section 3.1, Corollary 26]

Proposition 2.5 (Norm convergence implies pointwise convergence)**.** Let $\mathcal{H}$ be an RKHS, $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}$ be a sequence of reproducing kernels and $f \in \mathcal{H}$ a reproducing kernel. If $||f_n - f||_{\mathcal{H}} \to 0$ as $n \to \infty$, then for all $x \in \mathcal{X}$ we have

$$\lim_{n\to\infty} f_n(x) = f(x).$$

Proof. (Adapted from the same source)

The Dirac functional is linear: $\delta_x(\alpha f + \beta g) = (\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x)$. Hence, by continuity,

$$||f_n - f||_{\mathcal{H}} \to 0 \implies |\delta_x(f_n) - \delta_x(f)| \to 0,$$

i.e. $\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) = \delta_x(f) = f(x)$. □

Remark. Note that the above proposition is not necessarily true if $\mathcal{H}$ is not an RKHS [5, Chapter 1 Introduction]: Let $\mathcal{H} = L^2([0,1])$ (Example 2.1) and consider the sequence of functions $(q_n)_{n\in\mathbb{N}}$ where $q_n(x) = x^n$. Then

$$\begin{aligned} ||q_n - 0||_{\mathcal{H}} &= \left(\int_0^1 x^{2n}\,\mathrm{d}x\right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{2n+1}} \\ &\to 0 \text{ as } n \to \infty, \end{aligned}$$

but $q_n(1) = 1 \neq 0$ for all $n \in \mathbb{N}$. Hence pointwise convergence is not true for all $x \in [0,1]$.

Reproducing kernels are actually kernels, as shown in the lemma below [2, Section 4.2, Lemma 4.19]:

Lemma 2.5 (Reproducing kernels are kernels)**.** Let $\mathcal{H}$ be a Hilbert Space of functions over a non-empty set $\mathcal{X}$ which has a reproducing kernel k. Then $\mathcal{H}$ is an RKHS and $\mathcal{H}$ is also a feature space of k where the feature map $\phi : \mathcal{X} \to \mathcal{H}$ is given by

$$\phi(x) = k(\cdot, x), \qquad x \in \mathcal{X}.$$

Proof. (Adapted from the same source)

Since k is a reproducing kernel, for all $f \in \mathcal{H}$, we have

$$f(x) = \langle f, k(\cdot, x)\rangle.$$

Take $f(x) = \delta_x(f)$, the Dirac functional, $x \in \mathcal{X}$ arbitrary. Then if $||f_n - f||_{\mathcal{H}} \to 0$ as $n \to \infty$, then

$$\begin{aligned} |\delta_x(f_n) - \delta_x(f)| &= |f_n(x) - f(x)| \\ &= |(f_n - f)(x)| \\ &= \langle (f_n - f), k(\cdot, x)\rangle_{\mathcal{H}} \\ &\le ||f_n - f||_{\mathcal{H}} ||k(\cdot, x)||_{\mathcal{H}} \qquad \text{(by Cauchy-Schwartz inequality)} \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

Hence $\delta_x(f)$ is continuous for all $x \in \mathcal{X}$, which implies $\mathcal{H}$ is an RKHS. To show that $\mathcal{H}$ is a feature space of k, fix some $x' \in \mathcal{X}$ and define $f := k(\cdot, x')$. Then for all $x \in \mathcal{X}$, by the reproducing property,

$$\begin{aligned} \langle \phi(x'), \phi(x)\rangle &= \langle k(\cdot, x'), k(\cdot, x)\rangle \\ &= \langle f, k(\cdot, x)\rangle \\ &= f(x) = k(x, x'). \end{aligned}$$

□

So, every Hilbert Space of functions with a reproducing kernel is an RKHS, and we call $\phi(x) = k(\cdot, x)$ the **canonical feature map** associated with the reproducing kernel k.

One of the reasons why RKHS is so important in both theoretical and applied context is that every positive semi-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is also a reproducing kernel of a unique corresponding RKHS, that is, it can specifically be written as an inner product in an RKHS such that the reproducing property holds. This is known as the **Moore-Aronszajn theorem**, which will be covered towards the end of Section 2.4.

2.4 Constructing RKHS from kernels

This section is heavily based on [4, Theory lecture notes, Section 4].

So far we have just shown some abstract definition and properties of RKHS, and we have yet to show what the corresponding inner product is, or even what kind of functions lie in an RKHS. We will do this now by deliberately constructing an RKHS that corresponds to a kernel k.

Starting with a kernel k, we will construct a pre-RKHS $\mathcal{H}_0$, from which we will form the RKHS $\mathcal{H}$. This pre-RKHS has to satisfy the following two axioms:

(i) Dirac functionals δ_x are continuous on $\mathcal{H}_0$. This ensures that if k is a reproducing kernel of $\mathcal{H}$ then $\mathcal{H}$ is an RKHS.

(ii) Any Cauchy sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ that converges pointwise to 0 also converges in $\mathcal{H}_0$-norm to 0. This implies that any Cauchy sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ that converges pointwise to f also converges to f in $\mathcal{H}_0$-norm, since $(f_n - f)_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ converges pointwise to 0.

A particular important pre-RKHS that will be used to construct RKHS is given as follows [4, Theory lecture notes, Section 4.5, Theorem 42]:

Proposition 2.6. The space

$$\mathcal{H}_0 := \operatorname{span}\{(k(\cdot, x)) : x \in \mathcal{X}\} = \left\{ f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : n \in \mathbb{N}, \alpha_1, \dots, \alpha_n \in \mathbb{R}, x_1, \dots, x_n \in \mathcal{X} \right\}$$

equipped with the inner product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(y_j, x_i),$$

where $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j k(\cdot, y_j)$, is a valid pre-RKHS.

Proof. We first show that $\langle f, g \rangle_{\mathcal{H}_0}$ is a valid inner product by verifying all three properties in Definition 2.3:

(i) Let $f_1 = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $f_2 = \sum_{p=1}^{q} \alpha_p k(\cdot, x_p)$. Then

$$\begin{aligned} f_1 + f_2 &= \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) + \sum_{p=1}^{q} \alpha_p k(\cdot, x_p) \\ &= \sum_{i=1}^{\min\{n,q\}} 2\alpha_i k(\cdot, x_i) + \sum_{i=\min\{n,q\}+1}^{\max\{n,q\}} \alpha_i k(\cdot, x_i) \\ &= \sum_{i=1}^{\max\{n,q\}} \left[\alpha_i k(\cdot, x_i) + \mathbf{1}_{\{i \leq \min\{n,q\}\}} \alpha_i k(\cdot, x_i)\right] \\ &= \sum_{i=1}^{\max\{n,q\}} (\alpha_i + \mathbf{1}_{\{i \leq \min\{n,q\}\}} \alpha_i) k(\cdot, x_i) \\ &= \sum_{i=1}^{z} \gamma_i k(\cdot, x_1), \end{aligned}$$

where $z = \max\{n, q\}$ and $\gamma_i = \alpha_i + \mathbf{1}_{\{i \leq \min\{n,q\}\}} \alpha_i$. This shows that $f_1 + f_2$ can be written in a similar form as f_1 and f_2. Now,

$$\begin{aligned} \langle f_1 + f_2, g\rangle_{\mathcal{H}_0} &= \sum_{i=1}^{z} \sum_{j=1}^{m} \gamma_i \beta_j k(y_j, x_i) \\ &= \sum_{j=1}^{m} \beta_j \sum_{i=1}^{z} \gamma_i k(y_j, x_i) \\ &= \sum_{j=1}^{m} \beta_j \left(\sum_{i=1}^{n} \alpha_i k(y_j, x_i) + \sum_{p=1}^{q} \alpha_p k(y_j, x_p) \right) \\ &= \langle f_1, g\rangle_{\mathcal{H}_0} + \langle f_2, g\rangle_{\mathcal{H}_0}. \end{aligned}$$

(ii) $\langle f, g\rangle_{\mathcal{H}_0} = \langle g, f\rangle_{\mathcal{H}_0}$ by symmetry of k and commutativity of addition and multiplication of real numbers.

(iii) $\langle f, f\rangle_{\mathcal{H}_0} \geq 0$ for all $f \in \mathcal{H}_0$ follows directly from Proposition 2.3. Now we need to show that $\langle f, f\rangle_{\mathcal{H}_0} \iff f = 0$.
$(\impliedby)$

$$\begin{aligned} \langle f, f\rangle_{\mathcal{H}_0} &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(y_j, x_i) \\ &= \sum_{i=1}^{n} \alpha_i f(x_i) \quad \text{(by construction of } f) \\ &= 0 \quad \text{if } f(x_i) = 0 \text{ for all } x_i \in \mathcal{X}. \end{aligned}$$

($\implies$) [12, Theorem 12, Proof] Observe that

$$\begin{aligned}\langle f, g\rangle_{\mathcal{H}_0} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j k(y_j, x_i)\\ &= \sum_{i=1}^{n}\alpha_i\sum_{j=1}^{m}\beta_j k(y_j, x_i)\\ &= \sum_{i=1}^{n}\alpha_i g(x_i).\end{aligned}$$

So by setting $g = k(\cdot, x)$ we have the reproducing property:

$$\langle f, g\rangle_{\mathcal{H}_0} = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\alpha_i k(x_i, x) = f(x).$$

Let $\varepsilon \in \mathbb{R}$ and define $g := f + \varepsilon k(\cdot, x)$. Then by Proposition 2.3, $\langle g, g\rangle_{\mathcal{H}_0} \geq 0$.

$$\begin{aligned}\langle f + \varepsilon k(\cdot, x), f + \varepsilon k(\cdot, x)\rangle_{\mathcal{H}_0} &= \langle f, f + \varepsilon k(\cdot, x)\rangle_{\mathcal{H}_0} + \langle \varepsilon k(\cdot, x), f + \varepsilon k(\cdot, x)\rangle_{\mathcal{H}_0}\\ &= \langle f, f\rangle_{\mathcal{H}_0} + 2\varepsilon\langle f, k(\cdot, x)\rangle_{\mathcal{H}_0} + \varepsilon^2 k(x, x) \quad \text{(by reproducing property)}\\ &\geq 0.\end{aligned}$$

This is a quadratic expression in ε, so we have

$$\begin{aligned}4\langle f, k(\cdot, x)\rangle_{\mathcal{H}_0}^2 - 4k(x, x)\langle f, f\rangle_{\mathcal{H}_0} &\leq 0\\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}_0}^2 &\leq 0\\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}_0} &= 0 \quad (\langle f, k(\cdot, x)\rangle_{\mathcal{H}_0}^2 \geq 0)\\ f(x) &= 0 \quad \text{(by reproducing property).}\end{aligned}$$

Since $x \in \mathcal{X}$ is arbitrary, $f(x) = 0$ for all $x \in \mathcal{X} \implies f = 0$.

Now we show that $\mathcal{H}_0$ satisfy pre-RKHS axioms (i) and (ii):

(i) Let $x \in \mathcal{X}$. For $f = \displaystyle\sum_{i=1}^{n}\alpha_i k(\cdot, x_i)$,

$$\begin{aligned}\langle f, k(\cdot, x)\rangle_{\mathcal{H}_0} &= \left\langle \sum_{i=1}^{n}\alpha_i k(\cdot, x_i), k(\cdot, x)\right\rangle_{\mathcal{H}_0}\\ &= \sum_{i=1}^{n}\alpha_i\langle k(\cdot, x_i), k(\cdot, x)\rangle_{\mathcal{H}_0} \quad \text{(by linearity)}\\ &= \sum_{i=1}^{n}\alpha_i k(x, x_i) \quad \text{(by reproducing property)}\\ &= f(x)\end{aligned}$$

hence by Cauchy-Schwarz inequality we have

$$\begin{aligned}|\delta_x(f)| &= |f(x)|\\ &\leq ||f||_{\mathcal{H}_0}||k(\cdot, x)||_{\mathcal{H}_0}\\ &= ||f||_{\mathcal{H}_0}\sqrt{k(x, x)}\end{aligned}$$

which shows that $\delta_x(f)$ is bounded, and hence by Definition 2.8, $\delta_x(f)$ is continuous on $\mathcal{H}_0$.

(ii) Let $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ be a Cauchy sequence such that $(f_n)_{n\in\mathbb{N}}$ converges pointwise to 0. By boundedness of Cauchy sequences, there exists $M > 0$ such that $||f_n||_{\mathcal{H}_0} < M$ for all $n \in \mathbb{N}$. Since $(f_n)_{n\in\mathbb{N}}$ is Cauchy, there exists $N_1 \in \mathbb{N}$ such that $||f_n - f_m||_{\mathcal{H}_0} < \frac{\varepsilon}{2M}$ for all $n, m \geq N_1$. Let $f_{N_1} = \displaystyle\sum_{i=1}^{k} \alpha_i k(\cdot, x_i)$. We have

$$\begin{aligned}
||f_n||^2_{\mathcal{H}_0} &= |\langle f_n, f_n\rangle_{\mathcal{H}_0}| \quad \text{(by definition of } \mathcal{H}_0\text{-norm)}\\
&= |\langle f_n, f_n\rangle_{\mathcal{H}_0} - \langle f_{N_1}, f_n\rangle_{\mathcal{H}_0} + \langle f_{N_1}, f_n\rangle_{\mathcal{H}_0}|\\
&\leq |\langle f_n - f_{N_1}, f_n\rangle_{\mathcal{H}_0}| + |\langle f_{N_1}, f_n\rangle_{\mathcal{H}_0}| \quad \text{(by linearity and Triangle inequality)}\\
&\leq ||f_n - f_{N_1}||_{\mathcal{H}_0} ||f_n||_{\mathcal{H}_0} + \left| \left\langle \sum_{i=1}^{k} \alpha_i k(\cdot, x_i), f_n \right\rangle_{\mathcal{H}_0} \right| \quad \text{(by Cauchy-Schwarz inequality)}\\
&= ||f_n - f_{N_1}||_{\mathcal{H}_0} ||f_n||_{\mathcal{H}_0} + \left| \sum_{i=1}^{k} \alpha_i f_n(x_i) \right| \quad \text{(by reproducing property and linearity)}
\end{aligned}$$

Since $(f_n)_{n\in\mathbb{N}}$ converges pointwise to 0, for all $x_i \in \mathcal{X} and i = 1, \dots, k$, there exists $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$, $|f_n(x_i)| < \frac{\varepsilon}{2k|\alpha_i|}$. This gives

$$\begin{aligned}
||f_n||_{\mathcal{H}_0} &\leq \frac{\varepsilon}{2A} \cdot A + k|\alpha_i| \cdot \frac{\varepsilon}{2k|\alpha_i|}\\
&= \varepsilon,
\end{aligned}$$

so f_n converges to 0 in $\mathcal{H}_0$-norm, and this completes the proof.

□

Now define $\mathcal{H}$ to be the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which there exists a $\mathcal{H}_0$-Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ converging pointwise to f. Since the limits of these Cauchy sequences might not lie in $\mathcal{H}_0$, $\mathcal{H}_0 \subseteq \mathcal{H}$. Our goal now is to show that given a pre-RKHS $\mathcal{H}_0$ that satisfy axioms (i) and (ii), $\mathcal{H}$ is an RKHS. This will be done in 4 steps:

1. Define the inner product between $f, g \in \mathcal{H}$ as the limit of an inner product of the Cauchy sequences $(f_n)_{n\in\mathbb{N}}, (g_n)_{n\in\mathbb{N}}$ converging to f and g respectively. Lemma 2.6 shows that this inner product is well-defined, i.e. it converges and only depends on the limit of the Cauchy sequences;

2. Lemma 2.7 shows that this is indeed an inner product;

3. Lemma 2.9 shows that Dirac functionals are continuous in $\mathcal{H}$;

4. Finally, Lemma 2.10 tells us that $\mathcal{H}$ is complete, i.e. it is a Hilbert Space.

Lemma 2.6. For $f, g \in \mathcal{H}$ and Cauchy sequences (with respect to the $\mathcal{H}_0$-norm) $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ converging pointwise to f, g, define

$$\alpha_n := \langle f_n, g_n\rangle_{\mathcal{H}_0}.$$

Then,

(i) $(\alpha_n)_{n\in\mathbb{N}}$ is convergent; and

(ii) The limit of $(\alpha_n)_{n\in\mathbb{N}}$ only depends on f and g, i.e. it is independent of the Cauchy sequences $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ chosen.

This lemma shows that the limit of $(a_n)_{n\in\mathbb{N}}$, say α, is well-defined. The following lemma shows that it is indeed an inner product:

Lemma 2.7. $\alpha = \langle f, g\rangle_{\mathcal{H}} := \lim\limits_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{H}_0}$ is an inner product.

Now we have an inner product structure for $\mathcal{H}$, but for it to be an RKHS, Dirac functionals have to be continuous in $\mathcal{H}$:

Lemma 2.8. The Dirac functionals are continuous on $\mathcal{H}$.

Continuity of Dirac functionals implies that if $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{H}$, then for any $t \in \mathcal{X}$, $(f_n(t))_{n\in\mathbb{N}}$ converges to some limit $f(t) \in \mathbb{R}$ by completeness of $\mathbb{R}$. The question remaining is whether the function $f(t)$ is still in $\mathcal{H}$. This is indeed true:

Lemma 2.9. $\mathcal{H}$ is complete.

This gives a construction of an RKHS $\mathcal{H}$ in terms of the completion of its pre-RKHS $\mathcal{H}_0$. To be more precise, $\mathcal{H}$ is given by [11, Section 2.3, Remark 2.8, p.10]

$$\mathcal{H} = \Bigg\{ f = \sum_{i=1}^{\infty} \alpha_i k(\cdot, x_i) : (a_i)_{i\in\mathbb{N}} \subseteq \mathbb{R}, (x_i)_{i\in\mathbb{N}} \subseteq \mathcal{X} \text{ such that}$$
$$||f||^2_{\mathcal{H}} := \lim_{n\to\infty} \left|\left| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \right|\right|^2_{\mathcal{H}_0} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \alpha_i \alpha_j k(x_i, x_j) < \infty \Bigg\}$$

Now we are ready to state and prove the Moore-Aronszajn theorem:

Theorem 2.4 (Moore-Aronszajn)**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive semi-definite. Then there exists a unique RKHS $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ (here $\mathbb{R}^{\mathcal{X}}$ is the set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$) with reproducing kernel k.

Proof. To show existence, take $\mathcal{H}_0$ as in Proposition 2.6, whose completion is an RKHS $\mathcal{H}$. Note that if $f \in \mathcal{H}$ and $(f_n)_{n\in\mathbb{N}}$ converges pointwise to f, then

$$\begin{aligned}
\langle f, k(\cdot, x)\rangle_{\mathcal{H}_0} &= \lim_{n\to\infty} \langle f, k(\cdot, x)\rangle_{\mathcal{H}_0} \qquad \text{(by definition of } \mathcal{H}) \\
&= \lim_{n\to\infty} f_n(x) \\
&= f(x),
\end{aligned}$$

so k is in fact a reproducing kernel of $\mathcal{H}$.

To show uniqueness, let $\mathcal{H}'$ be another RKHS with reproducing kernel k. Then for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}'$ by Definition 2.14 (i). Hence $\mathcal{H}_0 \subseteq \mathcal{H}'$. By Lemma 2.8, $\mathcal{H}_0$ is dense in $\mathcal{H}$, so the $\mathcal{H}'$-inner product on $\mathcal{H}_0$ is given by

$$\langle f, g\rangle_{\mathcal{H}'} = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(y_j, x_i),$$

which is the same as the $\mathcal{H}$-inner product on $\mathcal{H}_0$. Since $\mathcal{H}_0$ is dense in $\mathcal{H}$, by Lemma 2.8 there exists a Cauchy sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ such that $||f_n - f||_{\mathcal{H}} \to 0$. Since $\langle\cdot,\cdot\rangle_{\mathcal{H}} = \langle\cdot,\cdot\rangle_{\mathcal{H}'}$ on $\mathcal{H}_0$, $(f_n)_{n\in\mathbb{N}}$ is also Cauchy in $\mathcal{H}'$ and converges to some $f' \in \mathcal{H}'$. By Proposition 2.5, for all $x \in \mathcal{X}$,

$$\begin{aligned}\lim_{n\to\infty} f_n(x) &= f(x)\\ \lim_{n\to\infty} f_n(x) &= f'(x)\\ f(x) &= f'(x)\end{aligned}$$

and this implies $\mathcal{H} \subseteq \mathcal{H}'$. By symmetry, $\mathcal{H}' \subseteq \mathcal{H}$. Now we need to check that the inner products on $\mathcal{H}$ and $\mathcal{H}'$ are the same:

$$\langle f, g\rangle_{\mathcal{H}} = \lim_{n\to\infty}\langle f_n, g_n\rangle_{\mathcal{H}} = \lim_{n\to\infty}\langle f_n, g_n\rangle_{\mathcal{H}_0} = \lim_{n\to\infty}\langle f_n, g_n\rangle_{\mathcal{H}'} = \langle f, g\rangle_{\mathcal{H}'}.$$

□

This theorem essentially tells us that every symmetric, positive semi-definite function is a reproducing kernel of some RKHS. The converse of Moore-Aronszajn theorem [9, Section 4.2] also holds:

Theorem 2.5. Every RKHS $\mathcal{H}$ has a unique corresponding reproducing kernel.

Proof. (Adapted from the same source)

We proof by contradiction. Suppose k_1 and k_2 are two reproducing kernels of $\mathcal{H}$. Then for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$\begin{aligned}\langle f, k_1(\cdot,x) - k_2(\cdot,x)\rangle_{\mathcal{H}} &= \langle f, k_1(\cdot,x)\rangle_{\mathcal{H}} - \langle f, k_2(\cdot,x)\rangle_{\mathcal{H}} \quad \text{(by linearity)}\\ &= f(x) - f(x) \quad \text{(by reproducing property)}\\ &= 0.\end{aligned}$$

Now take $f = k_1(\cdot,x) - k_2(\cdot,x)$. Then $\langle f, f\rangle_{\mathcal{H}} = 0 \implies k_1(\cdot,x) = k_2(\cdot,x)$ by positive semi-definiteness of inner products. □

Combining these two theorems together we can see that there is a one-to-one correspondence between kernels and its corresponding RKHS. We have demonstrated in Section 2.2 that a kernel k could have multiple feature maps and feature spaces, and the special feature space and feature map that we referred to is indeed the canonical feature map $\phi : x \mapsto k(\cdot,x)$ and the RKHS with reproducing kernel k. We will see an application of RKHS in statistical learning theory in Section 4.3.

3 Gaussian Processes

As mentioned in the introduction, Gaussian Process is a stochastic process that is closely related to kernels and RKHS. In this chapter, we will try to construct Gaussian Processes in a rigorous manner via a very general result on constructing continuous-time stochastic processes, which shows that Gaussian Process actually exists and is unique.

3.1 Construction of Stochastic Processes

The measure-theoretic definition of a stochastic process is given as follows [6, Section 1, Definition 1.1]:

Definition 3.1 (Stochastic Process)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{S})$ be a measurable space. A **stochastic process** $(X_t)_{t\in\mathcal{T}}$ is a sequence of random variables $X_t : \Omega \to S$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $(S, \mathcal{S})$, which is usually called the **state space**. We also sometimes write $X(t)$ instead of X_t.

It is sensible to ask what the probability space that $(X_t)_{t\in\mathcal{T}}$ is defined on actually is. To answer this, we need to first define *sample paths* [6, Section 1.1, Definition 1.1 & Definition 1.5]:

Definition 3.2 (Sample paths)**.** For a fixed $\omega \in \Omega$, the mapping $t \mapsto X_t(\omega)$ represents a realisation of the stochastic process $(X_t)_{t\in\mathcal{T}}$ and is called a **sample path** of $(X_t)_{t\in\mathcal{T}}$. $(X_t)_{t\in\mathcal{T}}$ is said to have **continuous** sample paths if

$$\mathbb{P}(\{\omega \in \Omega : t \mapsto X_t(\omega) \text{ is continuous for all } t \geq 0\}) = 1.$$

The canonical measurable space for a stochastic process [6, Section 1.1, Definition 1.1, Remark] is given by:

- $\Omega_0 := S^{\mathcal{T}}$, the path space, i.e. space of functions that maps from $\mathcal{T}$ to $\mathcal{S}$;
- $\mathcal{F}_0 := \sigma(\{\omega \in \Omega_0 : \omega(t_i) \in A_i \in \mathcal{S}, 1 \leq i \leq n\}, t_1, \ldots, t_n \in \mathcal{T}, n = 1, 2, \ldots)$, the smallest σ-algebra such that all canonical projections $\omega \mapsto \omega(t), t \in \mathcal{T}$, are measurable.

We will discuss what the canonical probability measure is in Theorem 3.1.

In probability, one often describes random variables and random vectors via their law or distribution, and we would like to ask whether we can do the same thing for a stochastic process: we would like to know the joint distribution for all the X_t's. However, that is quite a lot to ask for, and we will show that it is possible to characterise stochastic processes via the so-called *finite-dimensional distributions* [6, Section 1.1, Definition 1.1].

Definition 3.3 (Finite-dimensional distributions)**.** Let $(X_t)_{t\in\mathcal{T}}$ be a stochastic process and $t = (t_1, \ldots, t_n) \subseteq \mathcal{T}$ with $t_1 < \ldots < t_n$. The **finite-dimensional distributions** (FDDs) of $(X_t)_{t\in\mathcal{T}}$ is the probability measure μ_t on $\mathbb{R}^n$ characterised by

$$\mu_t(A_1 \times \ldots \times A_n) = \mathbb{P}(X_{t_1} \in A_1; \ldots ; X_{t_n} \in A_n), \qquad A_i \subseteq \mathcal{S},$$

i.e. the FDDs is the law (or distribution) of the random vector $(X_{t_1} \ldots , X_{t_n})$.

This suggests a strategy for constructing stochastic processes. First, we specify what properties we would like the FDDs of the process to have. Then, provided the FDDs satisfy certain consistency criteria below [6, Theorem 1.1], a result of Kolmogorov (Theorem 3.1) asserts the existence of a measure on the space $S^{\mathcal{T}}$ having precisely these FDDs, and hence the existence of the stochastic process.

Definition 3.4 (Consistency)**.** A collection of FDDs $\{\mu_t\}_{t\in\mathcal{T}}$ is said to be **consistent** if for all $t = (t_1, \dots, t_n) \in \mathcal{T}$ with $t_1 < \dots < t_n$ and $k \in \{1, \dots, n\}$ we have

$$\mu_t(A_1 \times \dots \times A_{k-1} \times S \times A_{k+1} \times \dots \times A_n) = \mu_{t\setminus\{t_k\}}(A_1 \times \dots \times A_{k-1} \times A_{k+1} \times \dots \times A_n).$$

This essentially means the FDDs marginalise (in this case we marginalise over t_k) consistently. To see this, consider the case where $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We can write the FDDs using cumulative distribution functions: If $F_t(x_1, \dots, x_n) = \mathbb{P}(X_{t_1} \leq x_1; \dots; X_{t_n} \leq x_n)$, then marginalising over t_k is essentially equal to setting $x_k = +\infty$. Hence the condition for consistency is equivalent to

$$F_t(x_1, \dots, x_{k-1}, +\infty, x_{k+1}, \dots, x_n) = F_{t\setminus\{t_k\}}(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n).$$

Remark. Here we have taken the definition of a cumulative distribution of a random variable X to be the function $F : \bar{\mathbb{R}} \to [0, 1]$ given by $F_X(x) = \mathbb{P}_X([-\infty, x])$, where $\mathbb{P}_X$ is the law of X and $\bar{\mathbb{R}}$ is the extended real numbers defined by $\mathbb{R} \cup \{-\infty, +\infty\}$.

The following result asserts that a stochastic process is uniquely characterised by the collection of all FDDs [6, Section 1.1, Theorem 1.1]:

Theorem 3.1 (Kolmogorov's Extension Theorem)**.** Let $\{\mu_t\}_{t\in T}$ be a consistent family of FDDs. Then there exists a unique probability measure $\mathbb{P}$ on $(\Omega_0, \mathcal{F}_0)$ such that

$$\mathbb{P}(\omega \in \Omega : \omega(t_1) \in A_i, \dots, \omega(t_n) \in A_n) = \mu_T(A_1 \dots \times A_n),$$

i.e. the stochastic process $(X_t)_{t\in\mathcal{T}}$ with $X_t(\omega) = \omega(t)$ is uniquely defined by its FDDs. This is known as the **canonical construction** of a stochastic process with given FDDs, and $(\Omega_0, \mathcal{F}_0, \mathbb{P})$ is the canonoical probability space of a stochastic process.

A proof of this theorem can be found in R. Durrett: *Probability: Theory and Examples*, 5th edition, Appendix A.3, page 464.

The canonical probability space implies that we can view stochastic processes as random variables defined on a space of functions. For a continuous-time stochastic process, where the index set $\mathcal{T}$ is continuous, it is essentially a random function of time which assigns a random variable to every point $t \in I$ within a continuous interval $I \subseteq \mathcal{T}$.

3.2 Gaussian Processes

By Kolmogorov's Extension Theorem, we can construct a unique stochastic process with a given FDD. Now we define a stochastic process by taking our FDDs to be multivariate Gaussian random variables, which we call a Gaussian Process [6, Section 1.1, Definition 1.3]:

Definition 3.5 (Gaussian Process)**.** A continuous-time stochastic process $(X_t)_{t\geq 0}$ with state space $\mathcal{S} = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be a **Gaussian process** if all FDDs are Gaussian, i.e. for all $n \in \mathbb{N}$ and all $t_1, \dots, t_n \geq 0$ the random vector $(X_{t_1}, \dots, X_{t_n})$ has a multivariate Gaussian distribution. Its **mean** and **covariance functions** are given by

$$\begin{aligned} m(t) &:= \mathbb{E}(X_t) \\ k(s, t) &:= \mathrm{Cov}(X_s, X_t), \quad s, t \geq 0. \end{aligned}$$

Since multivariate Gaussian random variables are completely characterised by its mean vector and covariance matrix, and all FDDs of a Gaussian Process is a multivariate Gaussian random variable, we can immediately conclude that Gaussian Processes are uniquely characterised by their mean and covariance functions. Note that $k(s,t)$ is the covariance between X_s and X_t, hence covariance functions have to be positive semi-definite, which by Theorem 2.3, is equivalent to it being a kernel. In fact, some authors use the term “covariance kernels” instead of covariance functions to emphasise this equivalence. This means we can, in modelling, specify a Gaussian process by choosing a mean function and covariance function, which could be the linear kernel, Gaussian kernel or any other type of kernels (e.g. Matérn kernels, see [8, Section 4.2]). We show 5 realisations of a mean-zero Gaussian process with the Gaussian kernel (also known as squared exponential kernel in GP context) with different length-scale parameters $\sigma = 0.1, 0.5, 1, 5$:

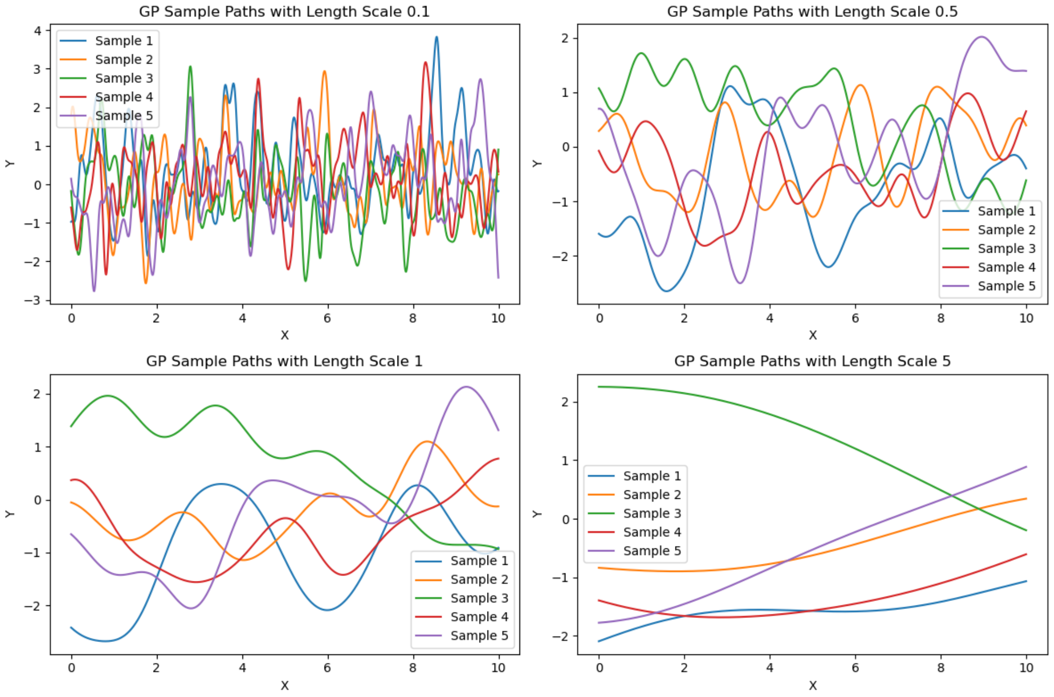

Figure 4: Draws from GP with Gaussian kernel

We can see that as the length-scale parameter gets smaller, the more ‘wiggle” behaviour the sample paths exhibit. Recall the Gaussian kernel from Example 2.4:

$$k(x,x') = \exp\left(-\frac{||x-x'||_2^2}{2\sigma^2}\right).$$

This kernel is positive for all $x, x' \in \mathcal{X}$, so all input points are positively correlated under the squared exponential kernel ($\text{Corr}(x,x') = \frac{k(x,x')}{\sqrt{k(x,x)}\sqrt{k(x',x')}} = k(x,x')$). As σ decreases, the more rapidly this correlation decreases as the separation between x and x' increases.

3.3 Gaussian Measures and Gaussian Processes

We have seen that Gaussian random variables form the backbone of Gaussian Processes, and the probability measure that defines Gaussian random variables is known as the *Gaussian measure*, which is defined as follows [7, Section 2.7, Definition 2.39]:

Definition 3.6 (Gaussian measure on $\mathbb{R}^d$)**.** Let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ be a symmetric and positive definite matrix. The **Gaussian measure** with mean μ and covariance Σ is denoted $\mathcal{N}(\mu, \Sigma)$ and is defined by

$$\mathcal{N}(\mu, \Sigma)(E) := \frac{1}{\det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}} \int_E \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \mathrm{d}x$$

for all measurable sets $E \subseteq \mathbb{R}^d$. The standard Gaussian measure $\gamma := \mathcal{N}_d(0, I)$ is called the standard Gaussian measure on $\mathbb{R}^d$.

Gaussian measures are theoretically important since it is well-defined on infinite-dimensional spaces, such as Banach and Hilbert spaces, unlike the Lebesgue measure [7, Section 2.7]. In $\mathbb{R}^d$, Lebesgue measure is characterised as the unique Borel measure that is

- locally finite, i.e. every point of $\mathbb{R}^d$ has an open neighbourhood of finite Lebesgue measure;
- strictly positive, and
- translation-invariant.

In addition, Lebesgue measure is also σ-finite. The following theorem [7, Section 2.7, Theorem 2.38] asserts that there does not exists such measure in infinite-dimensional spaces.

Theorem 3.2. Let μ be a Borel measure on an infinite-dimensional Banach space $\mathcal{V}$, and for $v \in \mathcal{V}$, let $T_v : \mathcal{T} \to \mathcal{T}$ be a translation map $T_v(x) := v + x$.

- If μ is locally finite and invariant under all translations, then μ is the trivial (zero) measure;
- If μ is a σ-finite and quasi-invariant under all translations (i.e. $(T_v)_*\mu$ is equivalent to μ), then μ is the trivial (zero) measure.

This implies that we can no longer define Gaussian measures in terms of a density with respect to the Lebesgue measure in infinite-dimensional spaces. In $\mathbb{R}^d$, one usually take the σ-algebra $\mathcal{F} = \mathcal{B}(\mathbb{R})$, the smallest σ-algebra containing $\mathbb{R}^d$. For a normed vector space $\mathcal{V}$, e.g. a Banach Space or Hilbert Space, one usually take $\mathcal{F}$ to be the smallest σ-algebra such that every continuous linear functional on $\mathcal{V}$ is measurable. This is known as the *cylindrical* σ-algebra and can be written as $\sigma(\ell^{-1}(B) : \ell \in \mathcal{V}^*, B \in \mathcal{B}(\mathbb{R}))$, where $\mathcal{V}^*$ is the dual space of $\mathcal{V}$.

The Gaussian measure on a normed vector space $\mathcal{V}$ is defined as follows [7, Definition 2.41]:

Definition 3.7 (Gaussian measure on normed vector spaces)**.** A Borel measure μ on a normed vector space $\mathcal{V}$ is said to be a **Gaussian measure**, if for every continuous linear functional $\ell : \mathcal{V} \to \mathbb{R}$, the push-forward measure $\ell_*\mu = \mu \circ \ell^{-1}$ is a Gaussian measure on $\mathbb{R}$. Equivalently, μ is Gaussian if for every linear map $T : \mathcal{V} \to \mathbb{R}^d$, $T_*\mu = \mathcal{N}_d(\mu_T, \Sigma_T)$ for some $\mu_T \in \mathbb{R}^d$ and some symmetric positive definite matrix $\Sigma_T \in \mathbb{R}^{d \times d}$.

Observe that the Radon-Nikodym derivative of $\mathcal{N}(\mu, \Sigma)(E)$ with respect to the Lebesgue measure is the probability density function of a Gaussian random vector on $\mathbb{R}^d$. One might ask whether Gaussian measures on $\mathcal{V}$ and Gaussian Processes are related in a similar way. As from above, continuous linear functionals play a crucial role when defining Gaussian measures. If we restrict ourselves to the case where $\mathcal{V} = \mathbb{R}^{\mathcal{T}}$, the space of functions mapping from $\mathcal{T} \to \mathbb{R}$, then $(X_t)_{t\in\mathcal{T}}$ is a Gaussian Process if for all $t \in \mathcal{T}, \delta_t((X_t)_{t\in\mathcal{T}})$ has a Gaussian distribution. This suggests that if $\delta_t \in \mathcal{V}^*$, i.e. δ_t is a continuous linear functional, then we can establish some sort of connection between Gaussian measures and Gaussian Processes. This is indeed a defining property of an RKHS (as seen in Definition 2.14), and it turns out that a Gaussian measure defined on an RKHS $\mathcal{H}$ is equivalent to a Gaussian Process whose sample paths lie in $\mathcal{H}$. We first establish a connection between a stochastic process and a random variable taking values in an RKHS [5, Section 4.2, Corollary 13]:

Lemma 3.1. A stochastic process $(X_t)_{t\in\mathcal{T}}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with sample paths in an RKHS $\mathcal{H}$ is equivalent to a random variable on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(\mathcal{H}, \mathcal{C}_{\mathcal{H}})$, where $\mathcal{C}_{\mathcal{H}}$ is the cylindrical σ-algebra on $\mathcal{H}$.

Now we can state the equivalence between the notion of a Gaussian process and the notion of a Gaussian measure [5, Section 4.3.1, Theorem 91]:

Theorem 3.3 (Equivalence of Gaussian measure and Gaussian process on RKHS)**.**

(i) If $(X_t)_{t\in\mathcal{T}}$ is a Gaussian process on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with sample paths $t \mapsto X_t(\omega), \omega \in \Omega$ belonging to an RKHS $\mathcal{H}$, then the push-forward measure $X_*\mathbb{P}$ on $(\mathcal{H}, \mathcal{C}_{\mathcal{H}})$ induced by the random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{C}_{\mathcal{H}})$ is a Gaussian measure.

(ii) Conversely, for any Gaussian measure μ on $(\mathcal{H}, \mathcal{C}_{\mathcal{H}})$, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which one can define a Gaussian process $(X_t)_{t\in\mathcal{T}}$ with sample paths in $\mathcal{H}$ such that $X_*\mathbb{P} = \mu$.

This theorem is essentially a consequence of the continuity of evaluation functionals on RKHS. For a mean-zero Gaussian Process, it is completely characterised by its covariance kernel k, and by the Moore-Aronszajn theorem (Theorem 2.4), there exists a unique RKHS $\mathcal{H}_k$ with reproducing kernel k. So in order for Theorem 3.3 to be relevant, we need to ask whether the sample paths of a Gaussian process lie in $\mathcal{H}_k$. It turns out that a Gaussian Process sample path almost surely does not belong to the corresponding RKHS $\mathcal{H}_k$ if it is infinite dimensional [11, Section 4.2, Corollary 4.10]. However, given a kernel k, one can construct a finer (larger) RKHS $\mathcal{H}_r$ with reproducing kernel r such that sample paths from GP almost surely lie in $\mathcal{H}_r$ [11, Section 4.3].

3.4 Brownian Motion as a Gaussian Process

One particularly important example of Gaussian processes in probability theory is Brownian motion, which is widely considered to be one of the most important stochastic processes: It is a Gaussian process, a Markov process [6, Section 2], a martingale [6, Section 4] and is central to the theory of stochastic differential equations (see MA482 Stochastic Analysis).

Consider a stochastic process $X = (X_t)_{t\geq 0}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 3.8 (Stationary and independent increments, [6, Section 1.2, Definition 1.4])**.**

- X has **stationary increments** if

$$X_t - X_s \sim X_{t-s} - X_0 \quad \forall\, 0 \leq s < t;$$

- X has **independent increments** if for any $n \in \mathbb{N}$ and any $0 \leq t_1 \leq \ldots \leq t_n$, the increments

$$X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$$

  are independent random variables.

The following proposition would allow us to define Brownian Motion [6, Section 1.2, Proposition 1.5]:

Proposition 3.1. For a real-valued stochastic process $(X_t)_{t\geq 0}$ with a fixed initial condition $X_0 = 0$, the following are equivalent:

(i) $(X_t)_{t\geq 0}$ has stationary, independent increments with

$$X_t - X_s \sim \mathcal{N}(0, t-s) \quad \text{for all } 0 \leq s < t,$$

(ii) $(X_t)_{t\geq 0}$ is a Gaussian process with zero mean $m(t) = 0$ for all $t \geq 0$ and covariance function

$$k(s,t) = s \wedge t = \min\{s,t\}.$$

A proof is provided in Appendix B.

Now we can define the standard Brownian motion as a Gaussian process [6, Section 1.2, Definition 1.6]:

Definition 3.9 (Standard Brownian motion)**.** The standard Brownian motion is a stochastic process $(B_t)_{t\geq 0}$ with $B(0) = 0$, continuous sample paths and satisfies the equivalent properties (i) and (ii) in Proposition 3.1.

4 Frequentist Regression

We have now explored the mathematical and theoretical side of kernels and Gaussian Processes. Whilst they are extensively studied in functional analysis and probability theory, kernel methods and Gaussian Processes have numerous applications in machine learning and statistics. In fact, if one google the terms “Kernel Methods” or “Gaussian Processes”, most results are often related to machine learning and statistics rather than pure mathematics.

4.1 Maximum Likelihood Estimation vs Empirical Risk Minimisation

One of the biggest difference between statistical inference and (supervised) machine learning can be demonstrated in the case of linear regression. In the context of regression, we are given observations or training data $\{(x_i, y_i)\}_{i=1}^n$, which we model $\{(x_i, y_i)\}_{i=1}^n$ to be i.i.d samples from an unknown probability distribution of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. We call X our input or feature vector and Y our output or response, and $\mathcal{X}$ and $\mathcal{Y}$ be the input and output space respectively. There are two distinct ways to come up with a regression function:

- The statisticians' approach: We assume the distribution of Y is dependent of the value of X. For $i = 1, \dots, n$, we let Y_i be an independent copy of Y conditional on the value x_j of X. We assume that our model has $p < n$ parameters and come up with a *normal linear model* for the observations

  $$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

  where $\mathbf{Y} = (Y_1, \dots, Y_n)$, $\mathbf{X}$ is the $n \times p$ design matrix of rank n with rows $(1, x_i^T)$, $i \in \{1, \dots, p-1\}$, β is the p-dimensional parameter vector and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ be the noise/random error with i.i.d. components in which $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$ and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Here *linear* means linearity in the parameters, and perhaps the most common examples of linear models are

  – Linear regression: $Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{1k} + \varepsilon_i$ $(p = k + 1)$;

  – Polynomial regression: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 \ldots + \beta_k x_{1k}^k + \varepsilon_i$.

  We then estimate the parameter vector β via the maximum likelihood estimator (MLE):

  $$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

  and given a test data x^*, we have the regression function, which is defined to be the conditional mean function

  $$y^* = \mathbb{E}(Y|X = x^*) = (x^*)^T \beta.$$

  Combining this estimate and its corresponding confidence intervals allows us to infer on the value of the unknown parameter vector, which is the ultimate goal of statistical inference.

- The machine learner's approach: Our goal is to learn a function $h : \mathcal{X} \to \mathcal{Y}$ (in the linear regression case $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}$), which we call a *hypothesis* or *estimator*, that allows us to make predictions on testing data that are unlabelled. To measure the quality of out hypothesis h, we introduce a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The goal is to choose a hypothesis h from the hypothesis class $\mathcal{H}$ (the set of all functions that could potentially be a hypothesis) that minimises the *true risk* or expected loss

$$\mathcal{R}(h) = \mathbb{E}_{X,Y}[\ell(h(X), Y)].$$

This is an expectation with respect to the true joint probability measure of X and Y, $\mathbb{P}_{X,Y}$, hence $\mathcal{R}$ is a deterministic function, and this is known as the **population problem**.

However, $\mathbb{P}_{X,Y}$ is often unknown in practice, so we could estimate this true risk by the so-called *empirical risk*: Choose $h \in \mathcal{H}$ that minimises the empirical risk

$$\hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i).$$

Since the training data are random samples from $\mathbb{P}_{X,Y}$, $\hat{\mathcal{R}}$ is random. It is in fact an unbiased estimator for the true risk:

$$\begin{aligned}
\mathbb{E}(\hat{\mathcal{R}}(h)) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X_i,Y_i}[\ell(h(X_i), Y_i)] \quad \text{(by linearity of expectation)} \\
&= \frac{n}{n} \cdot \mathbb{E}_{X,Y}[\ell(h(X), Y)] \quad \text{(by i.i.d. assumption)} \\
&= \mathcal{R}(h).
\end{aligned}$$

It is also strongly consistent, i.e. as $n \to \infty$, the empirical risk converges almost surely to the true risk. This is a consequence of Kolmogorov's Strong Law of Large Numbers. This approach is known as **empirical risk minimisation**, and our goal is not to do inference on the parameters but to come up with a function that allows us to make predictions.

A very popular loss function in regression is by choosing the least-squares loss $\ell(h(x), y) = (y - h(x))^2$, which corresponds to the least-squares estimate of the linear model. Now we can formulate the ERM problem as follows: Take $\mathcal{X} = \mathbb{R}^{p-1}, \mathcal{Y} = \mathbb{R}$, loss function $\ell(h(x), y) = (y - h(x))^2$ and hypothesis class $\mathcal{H} = \{x \mapsto x^T\beta + b\}$. The empirical risk minimiser is $\hat{h}(x) = x^T\hat{\beta} + \hat{b}$ where

$$(\hat{\beta}, \hat{b}) = \underset{\beta \in \mathbb{R}^{p-1}, b \in \mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta - b)^2.$$

Putting this into matrix form we have

$$(\hat{\beta}, \hat{b}) = \underset{\beta \in \mathbb{R}^{p-1}, b \in \mathbb{R}}{\arg\min} \frac{1}{n}\left|\left|Y - X\begin{pmatrix}\beta \\ b\end{pmatrix}\right|\right|_2^2,$$

where X is the $n \times p$ design matrix. This is basically equivalent to the least-squares problem in the linear model (the only difference is the intercept term $\hat{b}$ is included in β in formulation of the linear model) which can be written as:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, ||Y - X\beta||_2^2 = (X^TX)^{-1}X^TY.$$

Note that this is the same as the MLE of the linear model: If we define the *log-likelihood loss* [13]

$$\ell(h(X), Y)) = -2 \log l(\beta, \sigma^2),$$

where $l(\beta, \sigma^2)$ is the log-likelihood of the normal linear model, then we have

$$\begin{aligned}
\ell(h(X), Y)) &= -2 \log l(\beta, \sigma^2) \\
&= -2 \log \left[\frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det(\sigma^2 I_n)}} \exp\left(-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) \right) \right] \\
&= -2 \log \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_i - X_i^T \beta)^2}{2\sigma^2} \right) \right] \qquad \text{(by independence of } Y_i\text{)} \\
&= -2 \left[\sum_{i=1}^{n} -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y_i = X_i^T \beta)^2 \right] \\
&= n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 \right)
\end{aligned}$$

This shows that maximising the likelihood is the same as minimising the empirical risk using log-likelihood loss, which in turns is equivalent to minimising empirical risk using the least-squared loss – this gives a reason to use the least-squares loss over other loss functions. Another reason why least-squares loss (or equivalently, least-squares estimators) are so popular is that it exhibits some optimality property (best linear unbiased estimator, see Section 4.4) by Gauss-Markov theorem.

4.2 Different views of regression

The above example is just linear regression, and as we have seen in the introduction, it does not work well in certain scenarios. In general, there are 4 different ways to model the hypothesis h [9, Section 6.1]:

- **Frequentist parametric approach**: Model h as h_θ for some parameter vector θ and estimate θ by minimising the empirical risk. An example of this would be linear regression with least-squares loss. This is a parametric approach since $(\hat{\beta}, \hat{b}) \in \Theta$ (the parameter space) is finite-dimensional.

- **Frequentist non-parametric approach**: Instead of modelling h as h_θ for some parameter vector θ, we model h as some unknown parameter taking values in an infinite-dimensional space of functions. Fit h by minimising the regularised empirical risk. An example of this would be *kernel ridge regression*, which will be covered in Section 4.3.

- **Bayesian parametric approach**: Model h as h_θ for some parameter vector θ. Model θ as a random variable and put a prior $p(\theta)$ on it, and obtain the posterior distribution using Bayes' rule. An example of this would be *Bayesian linear regression*, which will be briefly covered in Section 5.1.

- **Bayesian non-parametric approach**: Model h as a random variable taking value in an infinite-dimensional function space $\mathcal{H}$ (i.e. a stochastic process by Lemma 3.1). Put a prior over functions $h \in \mathcal{H}$, and compute the posterior distribution using Bayes' rule. An example of this would be *Gaussian process regression*, covered in Section 5.2.

4.3 Regularisation and the Representer theorem

In frequentist approaches, ultimately we want to find an h that minimises the true risk. However, since $\mathbb{P}_{XY}$ is unknown, we minimise the empirical risk, an unbiased estimator of the true risk, instead. This means we are now considering an optimisation problem over the training data (our observations), a subset of $\mathcal{X} \times \mathcal{Y}$. This implies that h might not generalise well from our training data to test data, hence prone to overfitting: When $\mathcal{H}$ is too large and flexible, we can easily choose $h \in \mathcal{H}$ that fits well to the training data but do not generalise to the test data and the generalisation risk will be large.

One way of avoiding overfitting is to try to not fit the training data too closely, and instead by adding a penalty term to limit the flexibility of $\mathcal{H}$. Mathematically, this means instead of solving the ERM problem, we solve the following *regularised* ERM problem:

$$\hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i) + \lambda ||\theta||^2,$$

where θ is the parameter, $||\cdot||$ is some norm and $\lambda \geq 0$ is known as the *tuning parameter* which controls the severity of the penalty and hence the complexity of the model. This approach is known as *Tikhonov regularisation* which originates in the theory of inverse problems to solve ill-posed problems [7, Section 6.1].

Before solving the (regularised) ERM problem we first need to define our hypothesis class $\mathcal{H}$. Is it possible that we choose $\mathcal{H}$ to be an RKHS? It turns out that not only can we do this, but choosing RKHS as a hypothesis class has significant implications on the representation of the solution. This result is known as the *representer theorem* [9, Section 4.4, Theorem 13]:

Theorem 4.1 (Representer Theorem)**.** Consider the problem of solving the regularised ERM problem:

$$\hat{h} = \underset{h \in \mathcal{H}_k}{\arg\min}\, \hat{R}(h) + \Omega(||h||^2_{\mathcal{H}_k}),$$

where $\mathcal{H}_k$ is an RKHS with corresponding kernel k and $\Omega : [0, \infty) \to \mathbb{R}$ be any non-decreasing function. Then there always exists a solution which takes the form

$$\hat{h} = \sum_{i=1}^{n} a_i k(\cdot, x_i)$$

for all $a_i \in \mathbb{R}$. In other words, $\hat{h}$ is a linear combination of the feature mappings of our training data. Moreover, if Ω is strictly increasing, then all solutions can be written in this form.

Proof. (Adapted from the same source)

Denote by h_s the projection of h onto the subspace $\text{span}\{k(\cdot, x_i) : i = 1, \ldots, n\} \subseteq \mathcal{H}_k$ such that $h = h_s + h^{\perp}$, where $h_s = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $h^{\perp}$ is orthogonal to span$\{k(\cdot, x_i) : i = 1, \ldots, n\}$ by Theorem 2.2. For the regularisation term, we have

$$\begin{aligned}
||h||^2_{\mathcal{H}_k} &= ||h_s + h^\perp||^2_{\mathcal{H}_k} \\
&= ||h_s||^2_{\mathcal{H}_k} + 2\langle h_s, h^\perp\rangle_{\mathcal{H}_k} + ||h^\perp||^2_{\mathcal{H}_k} \\
&= ||h_s||^2_{\mathcal{H}_k} + ||h^\perp||^2_{\mathcal{H}_k} \qquad (h_s \text{ is orthogonal to } h^\perp) \\
&\geq ||h_s||^2_{\mathcal{H}_k} \\
\Omega(||h||^2_{\mathcal{H}_k}) &\geq \Omega(||h_s||^2_{\mathcal{H}_k}) \qquad (\Omega \text{ is non-decreasing}).
\end{aligned}$$

On the other hand, the individual terms in the loss is given by

$$\begin{aligned}
h(x_i) &= \langle h, k(\cdot, x_i)\rangle_{\mathcal{H}_k} \\
&= \langle h_s + h^\perp, k(\cdot, x_i)\rangle_{\mathcal{H}_k} \\
&= \langle h_s, k(\cdot, x_i)\rangle_{\mathcal{H}_k} + \langle h^\perp, k(\cdot, x_i)\rangle_{\mathcal{H}_k} \qquad (\text{by linearity in first argument}) \\
&= \langle h_s, k(\cdot, x_i)\rangle_{\mathcal{H}_k} \qquad (h^\perp \text{ is orthogonal to } k(\cdot, x_i)) \\
&= h_s(x_i).
\end{aligned}$$

Hence we have

$$\ell(Y_i, h(X_i)) = \ell(Y_i, h_s(X_i)) \;\; \forall\, i = 1, \ldots, n.$$

This implies that the empirical risks are the same: $\hat{\mathcal{R}}(h) = \hat{\mathcal{R}}(h_s)$. Thus h_s is also a minimiser of the regularised ERM problem. If Ω is strictly increasing, then to minimise $\Omega(||h_s + h^\perp||_{\mathcal{H}})$ we simply set $h^\perp = 0$. □

The representer theorem essentially tells us that the solution to this optimisation problem lies in the subspace of $\mathcal{H}$ spanned by our input data points, which is finite-dimensional. This reduces an infinite-dimensional optimisation problem to a finite-dimensional one, since our optimisation objective now becomes finding the α_i's that minimises the regularised empirical risk. This can lead to efficient algorithms even when the RKHS itself is infinite dimensional. We will see an application of the representer theorem in Section 4.5.

4.4 Ridge Regression

Consider the statistical setting of a normal linear model

$$Y = X\beta + \varepsilon.$$

Under the normality assumption, the mean and variance of $\hat{\beta}$ are given by

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \beta \qquad (\hat{\beta} \text{ is unbiased}) \\
\text{Var}(\hat{\beta}) &= \sigma^2(X^TX)^{-1}.
\end{aligned}$$

For unbiased estimators, the variance of $\hat{\beta}$ gives a way to compare their quality in terms of the least-sqaured loss. For a biased estimator, we use the mean-squared error (MSE) instead [3, Chapter 1, p.1]:

$$\begin{aligned}
\mathbb{E}_{\hat{\beta}}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] &= \mathbb{E}_{\hat{\beta}}[(\hat{\beta} - \mathbb{E}(\hat{\beta}) + \mathbb{E}(\hat{\beta}) - \beta)(\hat{\beta} - \mathbb{E}(\hat{\beta}) + \mathbb{E}(\hat{\beta}) - \beta)^T] \\
&= \mathbb{E}_{\hat{\beta}}\{[(\hat{\beta} - \mathbb{E}(\hat{\beta})) + (\mathbb{E}(\hat{\beta}) - \beta)][(\hat{\beta} - \mathbb{E}(\hat{\beta}))^T + (\mathbb{E}(\hat{\beta}) - \beta)^T]\} \\
&= \text{Var}(\hat{\beta}) + [\mathbb{E}(\hat{\beta} - \beta)][\mathbb{E}(\hat{\beta} - \beta)]^T
\end{aligned}$$

which is a sum of variance and squared bias terms and illustrates the bias-variance trade-off when choosing an estimator.

Apart from avoiding overfitting, there are two other advantages when choosing to minimise the regularised empirical risk instead of the empirical risk:

- Variance reduction: By Gauss-Markov theorem, $\hat{\beta}(Y) = (X^T X)^{-1} X^T Y$ is the unique linear unbiased estimator for β with minimum variance. The regularised ERM problem provides a way to reduce the variance of this estimator at the cost of introducing some bias, as a consequence of bias-variance trade-off.
- Addressing ill-posed problems: The least-squares problem above is well-posed when $n > p$: it has a unique closed-form solution. However, when $p > n$, which is often the setting for high-dimensional statistical problems, this becomes an ill-posed problem: The maximum rank X could have is n and hence the columns are not linearly independent, which implies $X^T X$ is not invertible. This can be addressed by adding a penalty term to the ERM problem which introduces a "perturbation" term to $X^T X$, making it invertible.

In this context, the norm in Tikhonov regularisation is simply the ℓ^2 (Euclidean) norm, and this regression technique is also known as **ridge regression**. It discourages large regression coefficients by imposing a penalty on their size, therefore acts as a constraint on the flexibility of $\mathcal{H}$.

Definition 4.1 (Ridge regression)**. Ridge regression** solves the following regularised ERM problem [10, Section 2.2, p.15]: $\hat{h}(x) = x^T \hat{\beta} + \hat{b}$ where

$$(\hat{\beta}, \hat{b}) = \underset{\beta \in \mathbb{R}^{p-1}, b \in \mathbb{R}}{\arg\min} \frac{1}{n} \left(\sum_{i=1}^{n} (Y_i - X_i^T \beta - b)^2 \right) + \lambda ||\beta||_2^2.$$

This is equivalent to solving ([3, Section 1.1, p.1])

$$(\hat{\beta}, \hat{b}) = \underset{\beta \in \mathbb{R}^{p-1}, b \in \mathbb{R}}{\arg\min} ||Y - X\beta - b\mathbf{1}||_2^2 + \lambda ||\beta||_2^2,$$

where $\mathbf{1}$ is a vector of 1s, X is the $n \times (p-1)$ design matrix with the initial column of 1s removed. In this case we have explicitly stated the intercept term which is usually integrated into X by adding a column of 1s.

Ridge regression is also known as ℓ^2-regularisation since the penalty term is proportional to the squared ℓ^2 norm of the parameter β. Note that here we are not penalising the intercept term, since penalising it would make the results depend on the choice of origin of Y: For instance, if the parameter is the price of some asset, we do not want our fitted values to change if we choose to measure the monetary value using dollars instead of pound sterling. The regularisation parameter λ here controls the amount of shrinkage: If $\lambda = 0$ then the ridge estimator is equal to the least-squares estimator; if λ is large then the penalty term will dominate the optimisation problem and β will shrink to zero. The choice of λ is important and is usually done via cross-validation (see [3, Section 1.2]).

Ridge regression only makes sense if each entry of β have the same order of magnitude, or else the penalty will be affected by the scale of the predictors. Hence, all inputs should be scaled before using ridge regression (see [3, Section 1.1, p.2]). The solutions to the ridge regression problem is given as follows [3, Section 1.1, p.2]:

Theorem 4.2. The solution to the ridge regression problem is given by

$$\hat{b} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y.$$

Observe that that the solution adds a positive constant to the diagonal of $X^T X$ before inversion, which makes the problem well-posed, even if $X^T X$ is singular, and was the main motivation for ridge regression when it was first proposed in 1970 by Hoerl and Kennard.

Remark. Note that ridge regression addressed the ill-posedness of the least-squares problem when $p > n$: In this case, $X^T X$ is singular, implying that at least one of its eigenvalues, say λ_i, equals zero. Adding λI to $X^T X$ adds λ to all eigenvalues (by Spectral Decomposition theorem), and since $X^T X$ is positive semi-definite, all of it eigenvalues are non-negative. Hence if $\lambda > 0$, $X^T X + \lambda I$ is always invertible as all of its eigenvalues are non-zero.

Apart from ridge regularisation, there are also other forms of regularisation: One could also use LASSO (least absolute shrinkage and selection operator) [3, Section 2.2, p.22] which uses an ℓ^1 norm for the penalty term, or elastic-net regularisation [9, Section 2.1, p.15], which uses a mixed ℓ^1/ℓ^2 penalty.

4.5 Kernel Ridge Regression

The following motivation of kernel ridge regression is adapted from [3, Section 1.3]:

The fitted values of the ridge regression model is given by

$$\hat{Y} = X\hat{\beta} = X(X^T X + \lambda I)^{-1} X^T Y.$$

We begin with the following observation:

$$(X^T X + \lambda I)X^T = X^T X X^T + \lambda X^T = X^T (X X^T + \lambda I).$$

Multiplying both sides by inverses gives:

$$\begin{aligned} X^T &= (X^T X + \lambda I)^{-1} X^T (X X^T + \lambda I) \\ X^T (X X^T + \lambda I)^{-1} &= (X^T X + \lambda I)^{-1} X^T \\ X X^T (X X^T + \lambda I)^{-1} Y &= X (X^T X + \lambda I)^{-1} X^T Y \end{aligned}$$

so we can rewrite the fitted values in the form

$$\hat{Y} = X X^T (X X^T + \lambda I) Y.$$

Here XX^T is an $n \times n$ matrix where $(XX^T)_{ij} = (x_i)^T x_j = k(x_i, x_j)$ where k is the linear kernel, so XX^T is equal to the Gram matrix of the linear kernel, and solving the ridge regression problem in this form only requires us to compute the inner products between observations. So we can write the fitted values as

$$\hat{Y} = K(K + \lambda I)^{-1} Y,$$

where K is the Gram matrix corresponding to the linear kernel. Wit this formulation, we can simply replace the linear kernel with any chosen kernel k. This regression technique is called *kernel ridge regression*, and when k is the linear kernel it reduces to the usual ridge regression problem.

Remark. By using linear kernels we are implicitly assuming a linear relationship between the predictor variables and the response variables. For modelling non-linear relationships, a typical approach is to transform the predictor variables in the context of linear models. In kernel ridge regression, we perform this transformation by choosing a different kernel: For example, we can choose the quadratic kernel (polynomial kernel with degree 2) to model quadratic relationships, or the Gaussian kernel, which has infinitely many features, to model complex non-linear relationships.

Lemma 4.1. The RKHS $\mathcal{H}$ with linear kernel $k(x, x') = x^T x'$ is the set of all linear functions $f(x) = \beta^T x$, with $||f||_{\mathcal{H}}^2 = ||\beta||_2^2$.

Proof. The pre-RKHS is

$$\mathcal{H}_0 = \left\{ f : \mathbb{R}^p \to \mathbb{R} \,|\, n \in \mathbb{N}, f(x) = \sum_{i=1}^{n} \alpha_i x_i^T x \right\}.$$

Since $\mathcal{H}_0$ is finite-dimensional, it is complete. Hence the RKHS $\mathcal{H} = \mathcal{H}_0$. This is in fact the same as

$$\mathcal{H} = \left\{ f : \mathbb{R}^p \to \mathbb{R} \,|\, n \in \mathbb{N}, f(x) = \beta^T x \right\}.$$

If we let $f(x) = \beta^T x$, then $||f||_{\mathcal{H}}^2 = \langle k(\beta, \cdot), k(\beta, \cdot) \rangle_{\mathcal{H}} = k(\beta, \beta) = ||\beta||_2^2$. □

This lemma implies that we can rewrite the usual ridge regression optimisation problem as

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}} \frac{1}{n} \left(\sum_{i=1}^{n} (Y_i - h(x_i))^2 \right) + \lambda ||h||_{\mathcal{H}}^2.,$$

where $\mathcal{H}$ is the RKHS induced by linear kernels and $|| \cdot ||_{\mathcal{H}}$ is its corresponding RKHS norm. This gives rise to the formal definition of kernel ridge regression as an optimisation problem [12, Section 3.2, p.18]

Definition 4.2 (Kernel Ridge Regression)**. Kernel Ridge Regression** solves the following regularised ERM problem: We are interested in finding $\hat{h} \in \mathcal{H}_k$ such that

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}_k} \frac{1}{n} \left(\sum_{i=1}^{n} (Y_i - h(x_i))^2 \right) + \lambda ||h||_{\mathcal{H}_k}^2,$$

where $\mathcal{H}_k$ is the RKHS induced by some chosen kernel k and $\lambda > 0$ is the regularisation parameter.

The representer theorem (Theorem 4.1) tells us that by choosing $\Omega(\cdot)$ to be $\lambda(\cdot)$, all solutions to the kernel ridge regression problem can be written in the form

$$\hat{h}(x) = \sum_{i=1}^{n} a_i k(x, x_i),$$

for some $a_i \in \mathbb{R}$ for $i \in \{1, \ldots, n\}$. Now the remaining questions are how do we find these a_i's, and whether the solution is unique. These are addressed in the following theorem [11, Section 3.2, Theorem 3.4]:

Theorem 4.3. If $\lambda > 0$, then the solution to the kernel ridge regression problem is unique as a function, and is given by

$$\hat{h}(x) = k_{xX}(K + n\lambda I)^{-1}Y = \sum_{i=1}^{n} \alpha_i k(x, x_i) \quad \text{for } x \in \mathcal{X},$$

where

$$\begin{aligned} (\alpha_1, \dots, \alpha_n)^T &= (K + n\lambda I_n)^{-1}Y \in \mathbb{R}^n; \\ K &\in \mathbb{R}^{n\times n} \text{ is the Gram matrix;} \\ k_{xX} &= [k(x, x_1), \dots, k(x, x_n)]^T. \end{aligned}$$

If we assume further that K is invertible, then the coefficients $\alpha_1, \dots, \alpha_n$ are unique.

Proof. The representer theorem asserts that all solutions are of the form

$$\hat{h} = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

for some $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$. Let $\alpha = (\alpha_1, \dots, \alpha_n)^T$. Then $\hat{h} = \alpha^T k_{xX}$ and $||\hat{h}||^2_{\mathcal{H}_k} = \alpha^T K \alpha$. Substituting this into the regularised ERM problem gives:

$$\min_{a \in \mathbb{R}^n} \frac{1}{n}[(a^T K^2 \alpha) - 2\alpha^T KY + ||Y||^2] + \lambda \alpha^T K \alpha \qquad (*)$$

Differentiating with respect to α and setting the derivative to be zero yields:

$$\begin{aligned} \frac{1}{n}[(K^2 + (K^2)^T)\alpha - 2KY] + 2\lambda K\alpha = 0 \\ 2K^2\alpha + 2\lambda n K\alpha &= 2KY \qquad \text{(K is symmetric)} \\ K(K + n\lambda I)\alpha &= KY \end{aligned}$$

This implies $\alpha = (K + n\lambda I)^{-1}Y$ is a solution to $(*)$. Since the objective function of $(*)$ is convex, α is a global minimiser, and hence

$$\hat{h} = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

with $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ is a global minimiser of the regularised ERM problem. As the square loss is convex, the above solution is unique as a function (i.e. α might not be unique as there might be different representations of $\hat{h}$). If K is invertible, then we have

$$(K + n\lambda I)\alpha = Y \implies \alpha = (K + n\lambda I)^{-1}Y,$$

hence the coefficients are unique. □

In practice, the kernel and its corresponding hyperparameters has to be chosen and tuned specifically, and the choice is not always obvious. One way to find the optimal choice is by performing cross-validation on different candidate kernels, and find the optimal hyperparameter by grid search.

5 Bayesian Regression

We have just seen how frequentist methods, i.e. estimating the parameter vector θ or function h by minimising the empirical risk (from a statistical learning perspective). In this section we introduce another approach to these regression problems – the *Bayesian* approach. We will mainly focus on Gaussian Process regression, but before that we will give a brief introduction to Bayesian linear regression.

5.1 Bayesian Linear Regression

In frequentist linear regression, we treat the parameters β and σ^2 as fixed, unknown objects which has to be estimated from the observations, and we aim to construct an estimator which converges in probability to the true parameter (consistent), and gives rise to a confidence interval that quantifies the uncertainty of our estimate. However, in the Bayesian approach, we treat β and σ^2 as random variables with some prior distribution $p(\beta)$ and $p(\sigma^2)$, the probability distribution that we "believe" that these parameters have prior to observing the data. After seeing the data, we then update the distribution of the parameters using Bayes' rule.

Given observed training data $\{x_i, y_i\}_{i=1}^n$, under the setting of $y_i = x_i^T\beta + \varepsilon_i$, where $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$, the Bayesian approach is [14]:

(i) Specify some prior distributions on β and σ^2.

(ii) Compute the posterior distribution using Bayes' rule:

$$p(\beta, \sigma^2|y, x) = \frac{p(y|x, \beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2)}{p(y|x)},$$

where $p(y|\beta, \sigma^2)$ is the likelihood function, $p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2)$ is the prior and $p(y|x)$ is the marginal likelihood given by

$$p(y|x) = \int_{\beta,\sigma^2} p(y|\beta, \sigma^2)p(\beta|\sigma^2)p(\sigma^2)\,\mathrm{d}\beta\,\mathrm{d}\sigma^2.$$

(iii) Given a new test data x^*, the *posterior predictive distribution* is given by

$$p(y^*|x^*, y, x) = \int_{\beta,\sigma^2} p(y^*|x^*, \beta, \sigma^2)p(\beta, \sigma^2|y, x)\,\mathrm{d}\beta\,\mathrm{d}\sigma^2,$$

and the prediction value y^* is given by the mean of this posterior predictive distribution.

(iii) shows a distinctive difference between frequentist and Bayesian approaches in regression: rather than obtaining only an output y^*, the Bayesian framework gives an entire probability distribution over possible outputs, which is useful for quantifying the uncertainty of the predicted outputs.

One can also perform regularisation using Bayesian methods: If we define the *maximum-a-posteriori estimator* (MAP) $\hat{\beta}$ which maximises the posterior distribution, then ridge regression is equivalent to using MAP with a Gaussian prior, and LASSO is equivalent to using MAP with a Laplace prior [14].

5.2 Gaussian Process Regression

In frequentist and Bayesian linear regression, we assume that $\varepsilon \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$, i.e. the observations $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I)$ are independent. But what if our observations are not necessarily independent? A natural extension of the linear model would be

$$Y \sim \mathcal{N}_n(\mu, \Sigma),$$

where μ is the mean vector of and Σ is the covariance matrix of Y. The natural question to ask is how to do we model μ and Σ, or equivalently, how do we parameterise our model in the absence of σ^2 to fit in the 4 different views of regression?

One way to model this is by using a non-parametric Bayesian approach: By Lemma 3.1, a stochastic process on some probability space is equivalent to a random variable on the same probability space taking values in an infinite-dimensional Hilbert Space. So instead of specifying a prior on the parameter vector θ, we can directly specify a prior over $h \in \mathcal{H}$ and compute its posterior distribution using Bayes' rule.

But why would we choose the stochastic process to be Gaussian?

(i) Gaussian Processes are completely characterised by its mean function and covariance kernel. If we think of the features of the observed training data $\{x_1, \ldots, x_n\}$ as draws from the finite-dimensional distribution $(X_1, \ldots, X_n)$, then we can completely characterise the distribution of the output vector Y by our choice of mean function and covariance kernel.

(ii) The conditional and marginal distributions of a multivariate Gaussian random variable is also Gaussian. This gives rise to analytic and tractable solutions to the posterior distribution, which we will see in Lemma 5.1 and Theorem 5.1.

In the Gaussian Process regression model, given training data $\{x_i, y_i\}_{i=1}^n$, we have

$$y_i = f(x_i) + \varepsilon_i,$$

where $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. We also assume a prior over functions $f(\cdot)$, which is a Gaussian process with mean function m and covariance function k:

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

Since this GP serves as a prior, the mean m and kernel k should be chosen so that they reflect one's prior knowledge or belief about the regression function f.

In Bayesian linear regression we would typically compute the posterior distribution using Bayes' rule. For Gaussian Process regression, however, one does not need to use Bayes's rule to find the posterior distribution. We start with a technical lemma on Gaussian conditional distributions [8, Appendix A2]:

Lemma 5.1 (Conditioning Gaussian random variables)**.** Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be Gaussian random vectors such that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}\right),$$

where $\mu_X \in \mathbb{R}^n$ and $\mu_Y \in \mathbb{R}^m$ are the mean vectors, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ are the covariance matrices of X and Y respectively and $C \in \mathbb{R}^{n\times m}$. Then, under the assumption that A is positive definite (which implies A^{-1} exists), the conditional distribution of Y given $X = x$ is

$$Y|X = x \sim \mathcal{N}_m(\mu_Y + C^T A^{-1}(x - \mu_X), B - C^T A^{-1} C).$$

This lemma, together with the assumption of Gaussian noise and Gaussian process prior, allows us to find a closed-form expression of the posterior distribution without explicitly using Bayes' rule [11, Section 3, Theorem 3.1]:

Theorem 5.1 (Posterior distribution of GP regression)**.** Let $X = (x_1, \dots, x_n)^T \in \mathcal{X}^n$ and $Y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$. The the posterior distribution of the Gaussian process regression problem is a Gaussian process given by

$$f|Y, X \sim \mathcal{GP}(m', k'),$$

where $m' : \mathcal{X} \to \mathbb{R}$ is the posterior mean function and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the posterior covariance kernel given by

$$\begin{aligned} m'(x) &= m(x) + K_{xX}(K_{XX} + \sigma^2 I_n)^{-1}(Y - m_X) \qquad \text{for } x \in \mathcal{X}; \\ k'(x, x') &= k(x, x') - K_{xX}(K_{XX} + \sigma^2 I_n)^{-1} K_{Xx'} \qquad \text{for } x, x' \in \mathcal{X}; \\ m_X &= [m(x_1), \dots, m(x_n)]^T \in \mathbb{R}^n. \end{aligned}$$

Proof. (Adapted from the same source)

Let $m \in \mathbb{N}$ and let $Z = (z_1, \dots, z_m)^T \in \mathcal{X}^m$ be any set of finite points. Under the Gaussian prior f, the observations $Y \in \mathbb{R}^n$ and GP-function values $f_z = [f(z_1), \dots, f(z_m)]^T \in \mathbb{R}^m$ are jointly Gaussian with

$$\begin{pmatrix} Y \\ f_Z \end{pmatrix} \sim \mathcal{N}_{n+m}\left(\begin{pmatrix} m_X \\ m_Z \end{pmatrix}, \begin{pmatrix} k_{XX} + \sigma^2 I_n & k_{XZ} \\ k_{ZX} & k_{ZZ} \end{pmatrix} \right).$$

By Lemma 5.1, $f_Z|Y = y \sim \mathcal{N}_m(\mu', \Sigma')$, where

$$\begin{aligned} \mu' &= m_Z + K_{XZ}^T(K_{XX} + \sigma^2 I_n)^{-1}(y - m_X) \\ \Sigma' &= K_{ZZ} - K_{ZX}(K_{XX} + \sigma^2 I_n)^{-1} K_{XZ}. \end{aligned}$$

Observe that the mean vector and covariance matrix can be rewritten as

$$\begin{aligned} \mu' &= m'_Z = [m'(z_1), \dots, m(z_m)]^T \\ \Sigma' &= K'_{ZZ} \text{ where } (K_{ZZ})_{ij} = k(z_i, z_j) \text{ for } i, j \in \{1, \dots, m\}. \end{aligned}$$

This implies $f_Z|Y = y \sim \mathcal{N}_m(m'_Z, K'_{ZZ})$. Since Z is arbitrary, this defines a collection of finite dimensional distributions which is consistent since marginal distributions of a Gaussian random vector is also Gaussian. Therefore, by Kolmogorov's Extension theorem (Theorem 3.1), the stochastic process $f \sim \mathcal{GP}(m, k)$ conditioned on the observed training data X, Y, i.e. the posterior distribution of f, is a draw from $\mathcal{GP}(m', k')$. □

Remark. The fact that we can obtain the posterior distribution of Gaussian Process regression without the use of Bayes' rule is remarkable for two reasons [11, Section 3.1]:

(i) In measure-theoretic terms, Bayes' rule defines the posterior probability measure $\mathbb{P}(\theta|y,x)$ in terms of the likelihood function $\mathbb{P}(y|\theta,x)$ and the prior probability measure $\mathbb{P}(\theta)$:

$$\mathbb{P}(\theta|y,x) = \frac{\mathbb{P}(y|\theta,x)\mathbb{P}(\theta)}{\mathbb{P}(y|x)}.$$

If the prior and posterior measure are defined on finite-dimensional Euclidean spaces, one can replace the probability measure with the corresponding probability density, since the probability density function is the Radon-Nikodym derivative of the probability measure with respect to the Lebesgue measure (which exists). However, in GP regression, since both the prior and posterior are defined on infinite-dimensional function spaces, the corresponding densities cannot be obtained by Theorem 3.2. One could use the *generalised Bayes' rule* [7, Section 6.2, Theorem 6.6] instead, but this does not produce the above expressions directly.

(ii) When dealing with the noise-free setting (i.e. we have access to the function values of the function we are trying to learn) where $\sigma^2 = 0$, Bayes' rule cannot be used because the likelihood function is degenerate.

Recall that the regression function is equal to the conditional mean function of the response given the observed data set. In Gaussian Process regression, given a testing data x^*, predicting the output value is done via evaluating the posterior conditional mean function, which is $m'(x^*)$ by definition. To find the posterior predictive distribution, one can simply take Z in the proof of Theorem 5.1 as our testing data set and see that it is equal to $\mathcal{N}_m(\mu', \Sigma')$.

Similar to kernel ridge regression, one need to choose the mean function and the kernel appropriately, as well as the hyperparameters of the kernel. This is covered in detail in [8, Section 5.4].

5.3 Links with Kernel Ridge Regression

If we set our Gaussian Process prior to have mean zero (which is very common for notational simplicity but not necessary, see [8, Section 2.7]), then our process is completely characterised by our choice of kernel, which implies our regression model depends only on the chosen kernel. This is also the case for kernel ridge regression, and one might ask whether there are any equivalences or connections between the two kernel-based regression methods. It turns out that under certain conditions, the posterior mean function of GP regression is equal to the solution of kernel ridge regression if the same kernel is chosen. This is given by the following corollary of Theorem 4.3 and Theorem 5.1 [12, Section 3.3, Proposition 3.6]:

Corollary 5.2. Let k be a positive semi-definite kernel on a non-empty set $\mathcal{X}$ and $(x_i, y_i)_{i=1}^n \subseteq \mathcal{X} \times \mathbb{R}$ be training data. Define $X = (x_1, \ldots, x_n)^T \in \mathcal{X}^n$ and $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Then we have $m' = \hat{h}$ if $\sigma^2 = n\lambda$, where

- m' is the posterior mean function of Gaussian Process regression based on (X, Y) with prior $f \sim \mathcal{GP}(0, k)$ and the modelling assumption $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$;
- $\hat{h}$ is the solution to the kernel ridge regression problem based on (X, Y), the hypothesis class $\mathcal{H}_k$ and regularisation parameter $\lambda > 0$.

If $\sigma^2 = 0$, i.e. in the noise-free setting, then $m' = \hat{h}$.

We will illustrate this equivalence using the dataset in the introduction where $n = 100$. For the kernel ridge regression model, we have used a regularisation parameter $\lambda = 0.1$ and length-scale parameter $\sqrt{5}$. By Corollary 5.2, we need to set the error variance $\sigma^2 = 100(0.1) = 10$ and use the same kernel for GP regression to demonstrate this equivalence. The result for kernel ridge regression is shown below:

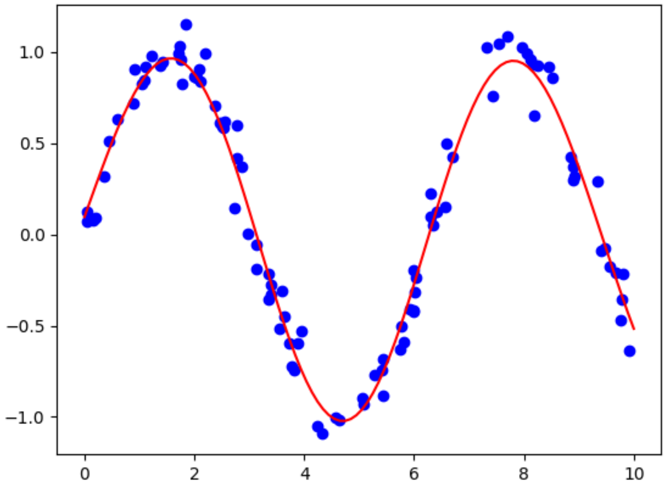

Figure 5: Kernel Ridge regression results

And this is the result for Gaussian Process regression:

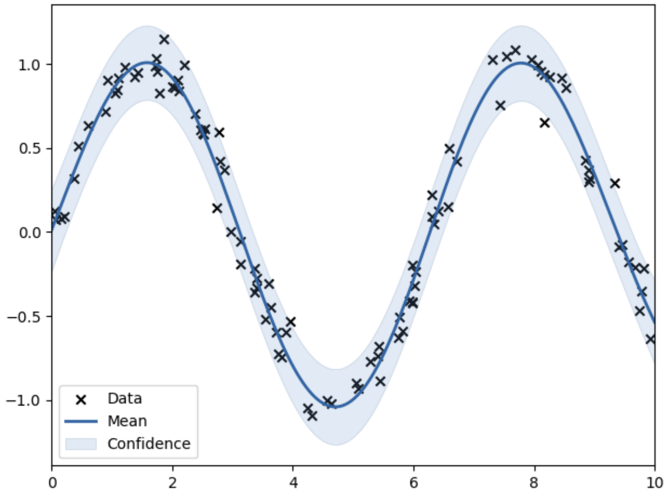

Figure 6: Gaussian Process regression results

As we can see from the figures, the regression functions look very similar, but Gaussian Process regression also provides a 95% confidence interval for predictions using this regression line, due to the probabilistic nature of Bayesian regression methods.

A Mathematical Background

In this essay, apart from second-year core mathematics modules (in particular MA244 Analysis III and MA259 Multivariable Calculus), I have also assumed knowledge from various optional modules offered by the Mathematics and Statistics departments which I have taken, as detailed below:

(i) ST220 Introduction to Mathematics Statistics:

- Definition of a positive semi-definite matrix;
- Definition of Multivariate Normal distribution (MVN), MVNs are uniquely characterised by its mean and covariance, marginal and conditional distributions of MVNs are MVNs;
- Definition of bias and consistent estimator, Maximum Likelihood Estimators (MLE);
- Defintion of mean-squared error (MSE) and bias-variance decomposition of MSE;

(ii) ST221 Linear Statistical Modelling:

- Linear model, the least-squares estimator, Gauss-Markov theorem and MLE for normal linear model;

(iii) MA3K1 Mathematics of Machine Learning:

- Definition of a hypothesis class, loss functions and empirical risk minimisation;
- Empirical risk minimisation formulation of linear regression and Bias-variance trade-off;
- Convex Optimisation.

(iv) ST342 Mathematics of Random Events:

- Definition of a measure, measurable functions, Lebesgue measure and integrals with respect to a measure;
- Measure-theoretic formulation of probability
- Definition of Radon-Nikodym derivative / density;
- Lebesgue measure is translation-invariant and strictly positive;
- Definition of a push-forward measure;
- Kolmogorov's Strong Law of Large Numbers.

B Proofs of Lemma 2.6-Lemma 2.9

The proofs of Lemma 2.6 - Lemma 2.9 reveal some key properties of Reproducing Kernel Hilbert Spaces, so they are included below. This appendix is heavily based on [4, Theory lecture notes, Section 4].

Lemma 2.6. For $f, g \in \mathcal{H}$ and Cauchy sequences (with respect to the $\mathcal{H}_0$-norm) $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ converging pointwise to f, g, define

$$\alpha_n := \langle f_n, g_n \rangle_{\mathcal{H}_0}.$$

Then,

(i) $(\alpha_n)_{n\in\mathbb{N}}$ is convergent; and

(ii) The limit of $(\alpha_n)_{n\in\mathbb{N}}$ only depends on f and g, i.e. it is independent of the Cauchy sequences $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ chosen.

Proof.

(i) For $n, m \in \mathbb{N}$,

$$\begin{aligned}|\alpha_n - \alpha_m| &= |\langle f_n, g_n\rangle_{\mathcal{H}_0} - \langle f_m, g_m\rangle_{\mathcal{H}_0}| \\ &= |\langle f_n, g_n\rangle_{\mathcal{H}_0} - \langle f_m, g_n\rangle_{\mathcal{H}_0} + \langle f_m, g_n\rangle_{\mathcal{H}_0} - \langle f_m, g_m\rangle_{\mathcal{H}_0}| \\ &= |\langle f_n - f_m, g_n\rangle_{\mathcal{H}_0} + \langle f_m, g_n - g_m\rangle_{\mathcal{H}_0}| \quad \text{(by linearity)} \\ &\leq |\langle f_n - f_m, g_n\rangle_{\mathcal{H}_0}| + |\langle f_m, g_n - g_m\rangle_{\mathcal{H}_0}| \quad \text{(by Triangle inequality)} \\ &\leq ||f_n - f_m||_{\mathcal{H}_0}||g_n||_{\mathcal{H}_0} + ||f_m||_{\mathcal{H}_0}||g_n - g_m||_{\mathcal{H}_0} \quad \text{(by Cauchy-Schwarz).}\end{aligned}$$

Since $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ are Cauchy, there exists $F, G \in \mathbb{R}$ such that $||f_m||_{\mathcal{H}_0} \leq F$ and $||g_n||_{\mathcal{H}_0} \leq G$ for all $m, n \in \mathbb{N}$. Moreover, for all $\varepsilon > 0$, there exists $N_1, N_2 \in \mathbb{N}$ such that $||f_n - f_m||_{\mathcal{H}_0} < \varepsilon$ for all $n, m > N_1$ and $||g_n - g_m||_{\mathcal{H}_0} < \varepsilon$ for all $n, m > N_2$. Hence choose $N_1 \in \mathbb{N}$ such that $||f_n - f_m|| < \frac{\varepsilon}{2G}$, $N_2 \in \mathbb{N}$ such that $||f_n - f_m|| < \frac{\varepsilon}{2F}$ and $N = \max\{N_1, N_2\}$ gives

$$\begin{aligned}|\alpha_n - \alpha_m| &\leq G\left(\frac{\varepsilon}{2G}\right) + F\left(\frac{\varepsilon}{2F}\right) \\ &= \varepsilon \ \ \forall n, m \geq N.\end{aligned}$$

(ii) Assume that there exists Cauchy sequences $(f_n')_{n\in\mathbb{N}}$ and $(g_n')_{n\in\mathbb{N}}$ that also converge pointwise to f and g respectively, and let $\alpha_n' = \langle f_n', g_n'\rangle_{\mathcal{H}_0}$. Using a similar argument as above, we have

$$|a_n - a_n'| \leq ||f_n - f_n'||_{\mathcal{H}_0}||g_n||_{\mathcal{H}_0} + ||f_n'||_{\mathcal{H}_0}||g_n - g_n'||_{\mathcal{H}_0}.$$

Since both $(f_n)_{n\in\mathbb{N}}$ and $(f_n')_{n\in\mathbb{N}}$ converge pointwise to f, $(f_n - f_n')_{n\in\mathbb{N}}$ converges pointwise to 0 and so does $(g_n - g_n')_{n\in\mathbb{N}}$. By pre-RKHS axiom (ii), both sequences converge to 0 in $\mathcal{H}_0$-norm. Hence $|\alpha_n - \alpha_n'| \to 0$ as $N \to \infty$, i.e.

$$\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \alpha_n'.$$

□

Lemma 2.7 $\alpha = \langle f, g\rangle_{\mathcal{H}} := \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{H}_0}$ is an inner product.

Proof.

(i) Let $f, f' \in \mathcal{H}$ such that $f = \lim_{n\to\infty} f_n$ and $f' = \lim_{n\to\infty} f'_n$. Then

$$\begin{aligned}\langle f + f', g\rangle_{\mathcal{H}} &= \lim_{n\to\infty} \langle f_n + f'_n, g_n\rangle_{\mathcal{H}_0} \\ &= \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{H}_0} + \lim_{n\to\infty} \langle f'_n, g_n\rangle_{\mathcal{H}_0} \\ &= \langle f, g\rangle_{\mathcal{H}} + \langle f', g\rangle_{\mathcal{H}}.\end{aligned}$$

(ii) $\langle f, g\rangle_{\mathcal{H}} = \lim_{n\to\infty} \langle f_n, g_n\rangle_{\mathcal{H}_0} = \lim_{n\to\infty} \langle g_n, f_n\rangle_{\mathcal{H}_0} = \langle g, f\rangle_{\mathcal{H}}$.

(iii) $\langle f_n, g_n\rangle_{\mathcal{H}_0} \geq 0$ for all $n \in \mathbb{N} \implies \langle f, g\rangle_{\mathcal{H}} \geq 0$.

To show that $\langle f, f\rangle_{\mathcal{H}} = 0 \implies f = 0$, observe that

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n).$$

By pre-RKHS axiom (i), δ_x is continuous, and since it is linear (as seen in the proof of Proposition 2.5), it is bounded, i.e. there exists $M > 0$ such that $|\delta_x(f_n)| \leq M||f_n||_{\mathcal{H}_0}$. Hence,

$$|f(x)| = \lim_{n\to\infty} |\delta_x(f_n)| \leq \lim_{n\to\infty} M||f_n||_{\mathcal{H}_0} = 0 \implies f(x) = 0 \,\forall\, x \in \mathcal{X}.$$

To show that $f = 0 \implies \langle f, f\rangle_{\mathcal{H}}$, note that $f = 0$ essentially means that its corresponding $\mathcal{H}_0$-Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise to 0. We have

$$\begin{aligned}\langle f, f\rangle_{\mathcal{H}} &= \lim_{n\to\infty} \langle f_n, f_n\rangle_{\mathcal{H}_0} \\ &= \lim_{n\to\infty} ||f_n||_{\mathcal{H}_0} \\ &= 0 \quad \text{(by Proposition 2.1)}.\end{aligned}$$

□

To prove Lemma 2.8 we first need the following Lemma:

Lemma. $\mathcal{H}_0$ is dense in $\mathcal{H}$, i.e. for all $f \in \mathcal{H}$ and $\varepsilon > 0$, there exists $f' \in \mathcal{H}_0$ such that $||f - f'||_{\mathcal{H}} < \varepsilon$.

Proof. Let $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ be a Cauchy sequence and $f \in \mathcal{H}$ such that $(f_n)_{n\in\mathbb{N}}$ converges pointwise to f. Such sequence exists by construction of $\mathcal{H}$. For all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $||f_m - f_n||_{\mathcal{H}_0} < \varepsilon$ for all $m, n \geq N$. Fix $k \geq N$. Then the sequence $(f_m - f_k)_{m\in\mathbb{N}}$ converges pointwise to $f - f_k$. Hence,

$$\begin{aligned}||f - f_k||_{\mathcal{H}} &= \langle f - f_k, f - f_k\rangle_{\mathcal{H}} \\ &= \lim_{m\to\infty} \langle f_m - f_k, f_m - f_k\rangle_{\mathcal{H}_0} \quad \text{(by definition of } \mathcal{H}) \\ &= \lim_{m\to\infty} ||f_m - f_k||_{\mathcal{H}_0} \\ &< \varepsilon\end{aligned}$$

This implies that $||f_n - f||_{\mathcal{H}} \to 0$, and hence the existence of such f'. □

Lemma 2.8 The Dirac functionals are continuous on $\mathcal{H}$.

Proof. Let $x \in \mathcal{X}$ and $\varepsilon > 0$. By pre-RKHS axiom (i), we know that δ_x is continuous on $\mathcal{H}_0$. Then there exists $\delta > 0$ such that

$$||g - 0||_{\mathcal{H}_0} = ||g||_{\mathcal{H}_0} < \delta \implies |\delta_x(g)| = |g(x)| < \frac{\varepsilon}{2}.$$

Take $f \in \mathcal{H}$ with $||f||_{\mathcal{H}} < \frac{\delta}{2}$. By the Lemma above, there exists a Cauchy sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ converging both pointwise to f and $||\cdot||_{\mathcal{H}}$ to f, so one can find $N \in \mathbb{N}$ such that

$$\begin{aligned} |f(x) - f_N(x)| &< \frac{\varepsilon}{2} \\ ||f - f_N||_{\mathcal{H}} &< \frac{\delta}{2} \end{aligned}$$

Since $f_N \in \mathcal{H}_0 \subseteq \mathcal{H}$, we have $||f_N||_{\mathcal{H}} = ||f_N||_{\mathcal{H}_0}$. Hence,

$$\begin{aligned} ||f_N||_{\mathcal{H}_0} &= ||f_N||_{\mathcal{H}} \\ &= ||f_N - f + f||_{\mathcal{H}} \\ &\leq ||f_N - f||_{\mathcal{H}} + ||f||_{\mathcal{H}} \qquad \text{(by Triangle inequality)} \\ &< \frac{\delta}{2} + \frac{\delta}{2} \\ &= \delta. \end{aligned}$$

Now set $g = f_N$. By continuity of δ_x on $\mathcal{H}_0$, we have

$$||f_N||_{\mathcal{H}_0} < \delta \implies |f_N(x)| < \frac{\varepsilon}{2}.$$

Thus,

$$|f(x)| = |f(x) - f_N(x) + f_N(x)| \leq |f(x) - f_N(x)| + |f(x)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This implies that δ_x is continuous at 0 in $||\cdot||_{\mathcal{H}}$, and hence by linearity, δ_x is continuous on $\mathcal{H}$. □

Lemma 2.9 $\mathcal{H}$ is complete.

The proof strategy is to define a sequence of functions $(g_n)_{n\in\mathbb{N}}$, where $g_n \in \mathcal{H}_0$, which is "close to" the Cauchy sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{H}$. We will then show that

(i) $(g_n)_{n\in\mathbb{N}}$ converges pointwise to f;

(ii) $(g_n)_{n\in\mathbb{N}}$ is Cauchy in $\mathcal{H}_0$.

The above implies that by our original construction of $\mathcal{H}$, we have $f \in \mathcal{H}$. Finally, we show that

(iii) $f_n \to f$ in $\mathcal{H}$-norm.

Proof. Define $f(x) := \lim_{n\to\infty} f_n(x)$. For $n \in \mathbb{N}$, choose $g_n \in \mathcal{H}_0$ such that $||g_n - f_n||_{\mathcal{H}} < \frac{1}{n}$. This can be done since $\mathcal{H}_0$ is dense in $\mathcal{H}$. Now we have

$$\begin{aligned}
|g_n(x) - f(x)| &= |g_n(x) - f_n(x) + f_n(x) - f(x)| \\
&\leq |g_n(x) - f_n(x)| + |f_n(x) - f(x)| \quad \text{(by Triangle inequality)} \\
&= |\delta_x(g_n - f_n)| + |f_n(x) - f(x)|
\end{aligned}$$

For the first term in the sum, by continuity of δ_x, we have $||g_n - f_n||_{\mathcal{H}} < \frac{1}{n} \to 0 \implies |\delta_x(g_n - f_n)| \to 0$ as $n \to \infty$, i.e. $\forall\, \varepsilon > 0$, $\exists\, N_1 \in \mathbb{N}$ such that $|\delta_x(g_n - f_n)| < \frac{\varepsilon}{2}$. For the second term, by definition of limit of sequences, $\forall\, \varepsilon > 0$, $\exists\, N_2 \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \frac{\varepsilon}{2}$. Hence by choosing $N = \max\{N_1, N_2\}$ we have $|g_n(x) - f(x)| < \varepsilon$, which proves (i). For (ii), we have

$$\begin{aligned}
||g_m - g_n||_{\mathcal{H}_0} &= ||g_m - g_n||_{\mathcal{H}} \\
&= ||g_m - f_m + f_m - f_n + f_n - g_n||_{\mathcal{H}} \\
&\leq ||g_m - f_m||_{\mathcal{H}} + ||f_m - f_n||_{\mathcal{H}} + ||f_n - g_n||_{\mathcal{H}} \quad \text{(by Triangle inequality)} \\
&= \frac{1}{m} + \frac{1}{n} + ||f_m - f_n||_{\mathcal{H}} \quad \text{(by construction)}
\end{aligned}$$

By choosing m and n such that $\frac{1}{m} + \frac{1}{n} < \frac{\varepsilon}{2}$, and combining this with definition of Cauchy sequences, we have, for sufficiently large m and n,

$$\begin{aligned}
||g_m - g_n||_{\mathcal{H}_0} &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&= \varepsilon
\end{aligned}$$

Hence $(g_n)_{n\in\mathbb{N}}$ is Cauchy in $\mathcal{H}_0$. Finally, f is a limit with respect to the $\mathcal{H}$-norm. This is because $\mathcal{H}_0$ is dense in $\mathcal{H}$ implies that $||g_n - f||_{\mathcal{H}} \to 0$, and hence

$$\begin{aligned}
||f_n - f||_{\mathcal{H}} &\leq ||f_n - g_n||_{\mathcal{H}} + ||g_n - f||_{\mathcal{H}} \quad \text{(by Triangle inequality)} \\
&= \frac{1}{n} + ||g_n - f||_{\mathcal{H}} \\
&\to 0 \text{ as } n \to \infty.
\end{aligned}$$

Hence $\mathcal{H}$ is complete. □

C Proof of Proposition 3.1

Proposition 3.1 tells us we can characterise the standard Brownian motion as a Gaussian process that satisfies certain properties.

Proposition 3.1. For a real-valued stochastic process $(X_t)_{t\geq 0}$ with a fixed initial condition $X_0 = 0$, the following are equivalent:

(i) $(X_t)_{t\geq 0}$ has stationary, independent increments with

$$X_t - X_s \sim \mathcal{N}(0, t-s) \quad \text{for all } 0 \leq s < t,$$

(ii) $(X_t)_{t\geq 0}$ is a Gaussian process with zero mean $m(t) = 0$ for all $t \geq 0$ and covariance function

$$k(s,t) = s \wedge t = \min\{s, t\}.$$

Proof. (Partially adapted from [6, Section 1.2, Proposition 1.5, Proof])

($\Longrightarrow$) Let $0 = t_0 < t_1 < \ldots < t_n$ be a set of ordered times. Writing X_{t_k} as a telescoping series gives

$$X_{t_k} = \sum_{i=1}^{k} [X_{t_i} - X_{t_{i-1}}] \qquad (X_0 = 0 = X_{t_0}).$$

Pick an arbitrary linear combination of $X_{t_0}, X_{t_1}, \ldots, X_{t_n}$, say $\sum_{i=1}^{n} a_i X_{t_i}$. Observe that

$$\begin{aligned}
a_1 X_{t_1} &= a_1(X_{t_1} - X_{t_0}) \\
a_1 X_{t_1} + a_2 X_{t_2} &= a_2(X_{t_2} - X_{t_1}) + (a_1 + a_2)(X_{t_1} - X_{t_0}) \\
a_1 X_{t_1} + a_2 X_{t_2} + a_3 X_{t_3} &= a_3(X_{t_3} - X_{t_2}) + (a_2 + a_3)(X_{t_2} - X_{t_1}) + (a_1 + a_2 + a_3)(X_{t_1} - X_{t_0}) \\
&\vdots \\
\sum_{i=1}^{n} a_i X_{t_i} &= \sum_{i=1}^{n} a_i (X_{t_1} - X_{t_0}) + \sum_{i=2}^{n} a_i (X_{t_2} - X_{t_1}) + \ldots + a_n(X_{t_n} - X_{t_{n-1}}).
\end{aligned}$$

Observe that the RHS is a sum of independent Gaussian random variables, and hence $(X_{t_0}, X_{t_1}, \ldots, X_{t_n})$ is a multivariate Gaussian random vector. This implies that all FDDs of $(X_t)_{t\geq 0}$ are Gaussian and so $(X_t)_{t\geq 0}$ is a Gaussian process. Now,

$$\begin{aligned}
m(t) &= \mathbb{E}(X_t) \\
&= \mathbb{E}(X_t - X_0) \qquad (X_0 = 0) \\
&= 0 \qquad (\text{mean of } \mathcal{N}(0, t) \text{ is zero}).
\end{aligned}$$

For the covariance function k, consider the case where $s < t$:

$$\begin{aligned}
k(s,t) &= \mathrm{Cov}(X_s, X_t) \\
&= \mathbb{E}[(X_s - \mathbb{E}(X_s))(X_t - \mathbb{E}(X_t))] \\
&= \mathbb{E}(X_s X_t) \quad (\mathbb{E}(X_s) = \mathbb{E}(X_t) = 0) \\
&= \mathbb{E}(X_s X_t - X_s^2 + X_s^2) \\
&= \mathbb{E}[(X_t - X_s)X_s] + \mathbb{E}(X_s^2) \\
&= \mathbb{E}[(X_t - X_s)(X_s - X_0)] + \mathbb{E}[(X_s - X_0)^2] \quad (X_0 = 0) \\
&= \mathrm{Cov}(X_t - X_s, X_s - X_0) + \mathrm{Var}(X_s - X_0) \\
&= 0 + s \quad (\text{by independence of increments and } X_s - X_0 \sim \mathcal{N}(0,s)) \\
&= s
\end{aligned}$$

Since $k(s,t)$ is symmetric, for $t < s$, we have $k(s,t) = t$. Hence $k(s,t) = s \wedge t$.

$(\impliedby)$ Since $(X_t)_{t \geq 0}$ is a Gaussian process, $X_t - X_s$ is a Gaussian random variable for $t > s \geq 0$.

$$\begin{aligned}
\mathbb{E}(X_t - X_s)) &= 0 \\
\mathrm{Var}(X_t - X_s) &= \mathrm{Var}(X_t) + \mathrm{Var}(X_s) - 2\mathrm{Cov}(X_t, X_s) \\
&= t + s - 2(t \wedge s) \\
&= t - s
\end{aligned}$$

Hence $X_t - X_s \sim \mathcal{N}(0, t-s)$. Since $X_{t-s} - X_0 \sim \mathcal{N}(0, t-s)$, $(X_t)_{t \geq 0}$ has stationary increments. To show that $(X_t)_{t \geq 0}$ has independent increments, we show that the increments are uncorrelated (since these increments follows a Gaussian distribution): Pick arbitrary $u < v \leq s < t$. Then

$$\begin{aligned}
\mathrm{Cov}(X_v - X_u, X_t - X_s) &= \mathbb{E}[(X_v - X_u)(X_t - X_s)] \\
&= \mathbb{E}(X_v X_t - X_u X_t - X_v X_s + X_u X_s) \\
&= \mathrm{Cov}(X_v X_t) - \mathrm{Cov}(X_u X_t) - \mathrm{Cov}(X_v X_s) + \mathrm{Cov}(X_u X_s) \\
&= v \wedge t - u \wedge t - v \wedge s + u \wedge s \\
&= v - u - v + s \\
&= 0
\end{aligned}$$

□

D Python code

- Figure 1: A simulated dataset

```
import numpy as np
import matplotlib.pyplot as plt

# Generate synthetic dataset
np.random.seed(100) #Set seed for reproducibility
X = np.random.rand(100, 1) * 10  # Generate 100 random feature values between 0 and 10
# Generate sinusoidal data with Gaussian noise
y = np.sin(X) + np.random.normal(0, 0.1, size=(100, 1))

plt.figure(figsize=(10, 6))
plt.scatter(X, y, color='blue', label='Data')
plt.xlabel('X')
plt.ylabel('y')
plt.title('A synthetic dataset')
plt.legend()
plt.show()
```

- Figure 2: Linear Regression on the dataset

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures

# Generate synthetic dataset
np.random.seed(100) #Set seed for reproducibility
X = np.random.rand(100, 1) * 10  # Generate 100 random feature values between 0 and 10
# Generate sinusoidal data with Gaussian noise
y = np.sin(X) + np.random.normal(0, 0.1, size=(100, 1))

# Fit simple linear regression
lr = LinearRegression()
lr.fit(X, y)

# Fit polynomial regression with degree 4
poly = PolynomialFeatures(degree=4)
X_poly = poly.fit_transform(X)
lr_poly = LinearRegression()
lr_poly.fit(X_poly, y)

# Plot the dataset
plt.figure(figsize=(10, 6))
plt.scatter(X, y, color='blue', label='Data')

# Plot linear regression
plt.plot(X, lr.predict(X), color='green', label='Linear Regression')
```

```
# Plot polynomial regression
X_poly_plot = np.linspace(0, 10, 100).reshape(-1, 1)
X_poly_plot = poly.transform(X_poly_plot)
plt.plot(np.linspace(0, 10, 100),
         lr_poly.predict(X_poly_plot),
         color='orange',
         label='Polynomial Regression (Degree 4)')

plt.xlabel('X')
plt.ylabel('y')
plt.title('Linear Regression on synthetic dataset')
plt.legend()
plt.show()
```

- Figure 3: Kernel Ridge Regression

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.kernel_ridge import KernelRidge

# Generate synthetic dataset
np.random.seed(100)
X = np.random.rand(100, 1) * 10  # Generate 100 random feature values between 0 and 10
# Generate sinusoidal data with Gaussian noise.
y = np.sin(X) + np.random.normal(0, 0.1, size=(100, 1))

# Fit simple linear regression
lr = LinearRegression()
lr.fit(X, y)

# Fit polynomial regression with degree 4
poly = PolynomialFeatures(degree=4)
X_poly = poly.fit_transform(X)
lr_poly = LinearRegression()
lr_poly.fit(X_poly, y)

# Fit Kernel Ridge Regression with RBF kernel
# Regularisation parameter = 0.1, length-scale parameter = sqrt(5)
krr = KernelRidge(alpha=0.1, kernel='rbf', gamma=0.1)
krr.fit(X, y)

# Plot the dataset and regression models
plt.figure(figsize=(10, 6))
plt.scatter(X, y, color='blue', label='Data')

# Plot linear regression
```

```
plt.plot(X, lr.predict(X), color='green', label='Linear Regression')

# Plot polynomial regression
X_poly_plot = np.linspace(0, 10, 100).reshape(-1, 1)
X_poly_plot = poly.transform(X_poly_plot)
plt.plot(np.linspace(0, 10, 100),
         lr_poly.predict(X_poly_plot),
         color='orange',
         label='Polynomial Regression (Degree 4)')

# Plot Kernel Ridge Regression
plt.plot(np.linspace(0, 10, 100),
        krr.predict(np.linspace(0, 10, 100).reshape(-1, 1)),
        color='red',
        label='Kernel Ridge Regression')

plt.xlabel('X')
plt.ylabel('y')
plt.title('Comparison of Regression Models')
plt.legend()
plt.show()
```

- Figure 4: Draws from GP with Gaussian kernel

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import RBF

# Set seed
np.random.seed(100)

# Length scale vector
length_scales = [0.1, 0.5, 1, 5]

# Input values
x = np.linspace(0, 10, 1000)[:, np.newaxis]

# Plotting sample paths
plt.figure(figsize=(12, 8))

for i, length_scale in enumerate(length_scales):
    kernel = RBF(length_scale=length_scale)
    gp = GaussianProcessRegressor(kernel=kernel)

    # generating sample paths
    y_samples = gp.sample_y(x, n_samples=5)

    # generate subplot grid
    plt.subplot(2, 2, i + 1)
    for j in range(5):
        plt.plot(x, y_samples[:, j], label=f'Sample {j + 1}')

    plt.title(f'GP Sample Paths with Length Scale {length_scale}')
    plt.xlabel('X')
    plt.ylabel('Y')
    plt.legend()

plt.tight_layout()
plt.show()
```

- Figures 5 and 6:

```
import GPy # GPy allows customisation of error variance
import numpy as np
import matplotlib.pyplot as plt
from sklearn.kernel_ridge import KernelRidge

# Generate synthetic dataset
np.random.seed(100)
X = np.random.rand(100, 1) * 10
y = np.sin(X) + np.random.normal(0, 0.1, size=(100, 1))

# Fitting Kernel Ridge Regression with RBF kernel
krr = KernelRidge(alpha=0.1, kernel='rbf', gamma=0.1)
krr.fit(X, y)

# Define a kernel function
kernel = GPy.kern.RBF(input_dim=1, variance=np.sqrt(10), lengthscale=np.sqrt(5))

# Create a Gaussian Process model
model = GPy.models.GPRegression(X, y, kernel)
model.optimize(messages=False)

# Ploting the original data
plt.scatter(X, y, color='blue', label='Original Data')

# Ploting Kernel Ridge Regression
plt.plot(np.linspace(0, 10, 100),
         krr.predict(np.linspace(0, 10, 100).reshape(-1, 1)),
         color='red',
         label='Kernel Ridge Regression')

# Plotting GP regression
model.plot(plot_limits=(0, 10), plot_data=True, label='GP Regression')
plt.legend()
plt.show()
```

References

[1] Hudson, T (2023). *MA3G7 Functional Analysis I*, Lecture Notes, Mathematics Institute, University of Warwick, Autumn term 2023.

[2] Steinwart, I. & Christmann, A. (2008). *Support Vector Machines*, 1st Edition, Springer, New York.

[3] Shah, R (2021). *Modern Statistical Methods*, Lecture Notes, University of Cambridge, Michaelmas Term 2021. Retrievable at: `http://www.statslab.cam.ac.uk/~rds37/teaching/modern_stat_methods/notes2.pdf`

[4] Gretton, A (2019). *Reproducing kernel Hilbert spaces in Machine Learning*, Lecture Material, Gatsby Computational Neuroscience Unit, University College London, Autumn Term 2023. Retrievable at: `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/rkhscourse.html`

[5] Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science+Business Media, LLC.

[6] Zaboronski, O. (2021). *MA4F7/ST403 Brownian Motion*, Lecture Notes, Mathematics Institute, University of Warwick, Autumn Term 2021.

[7] Sullivan, T. (2015). *Introduction to Uncertainty Quantification*, Texts in Applied Mathematics, Springer. `https://doi.org/10.1007/978-3-319-23395-6`

[8] Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* MIT Press.

[9] Rainforth, T. (2022). *Advanced Topics in Statistical Machine Learning*, Lecture Notes, Department of Statistics, University of Oxford, Hilary Term 2022. Retrievable at: `https://www.robots.ox.ac.uk/~twgr/assets/teaching/notes.pdf`

[10] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction (Second, correct 7th printing. ed.).* Springer. `https://doi.org/10.1007/978-0-387-84858-7`

[11] Kanagawa, M., Hennig, P., Sejdinovic, D., & Sriperumbudur, B.K. (2018). *Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences.* ArXiv, abs/1807.02582.

[12] Harchaoui, Z. (2020). *STAT538A: Statistical Learning: Lecture 2*, Lecture notes, Department of Statistics, University of Washington, Winter Term 2020. Retrievable at: `https://edenx.github.io/stat538lec2.pdf`.

[13] Everitt, R. (2021). *ST420: Statistical Learning and Big Data*, Lecture 13: Bayesian Statistics, Lecture recording, Department of Statistics, University of Warwick, Spring Term 2021.

[14] Everitt, R. (2021). *ST420: Statistical Learning and Big Data*, Lecture 15: Bayesian Linear Regression, Lecture recording, Department of Statistics, University of Warwick, Spring Term 2021.