The University of Southampton

Academic Year 2021/2022

Faculty of Social Sciences

Mathematical Sciences

MSc Dissertation

Probabilistic Numerical Methods for
Fredholm Integral Equations of the First Kind

Yakun Wang

A dissertation submitted in partial fulfilment of the MSc in Statistics

I am aware of the requirements of good academic practice and the potential penalties for any breaches. I confirm that this dissertation is all my own work.

Executive Summary

This dissertation examines Fredholm Integral Equation of the First Kind (FIEFK). FIEFK is first studied by Ivar Fredholm in the early twentieth century. Over the following decades, mathematical researches based on FIEFK produced many famous theories such as compact operator theory and spectral theory. In other words, FIEFK sets the stage for Functional Analysis and advanced the development of modern mathematics. FIEFK plays an important role in both certain mathematical problems and real-world problems. A simple example in the field of mathematics is solving initial or boundary problems of differential equations. These problems can be transformed into integral equations naturally without considering complex restrictions. As for practical application, it has been found that many engineering and scientific problems can be reduced to solving FIEFK such as atmospheric science, medical imaging, signal processing and so on.

However, FIEFK is essentially a class of ill-posed problems. This class of problems usually do not have a unique solution or the solution's behaviour changes continuously with the initial conditions in the sense that it is highly sensitive to changes in the final data. This prominent feature, ill-posedness, leads to little solvability of FIEFK. In fact, only a few FIEFKs can find their analytical solutions while most of this equation need to be solved numerically. Therefore, finding an effective numerical solution for FIEFK has become a research direction that mathematicians and engineers focus on.

So far, many studies have been carried out on the numerical solution of FIEFK. The developed numerical methods also achieved ideal results and desirable contraction. Among all these approaches, the regularisation method is generally regarded as the best way to solve FIEFK. The core of regularisation method is to use the prior information of the solution to modify the original problem by a well-posed optimisation problem to obtain a regular solution and then obtain a stable method to solve the original problem.

Notably, classical numerical methods are based on discretisation to some extent. Unavoidably, numerical errors were produced in discretising functions or continuous variables. Although this error can be reduced by using a more finely spaced lattice, the corresponding computational cost will be multiplied. Furthermore, these classical numerical methods also need to bear extra computational cost for error quantification.

Hence, a new low-cost calculation method is urgently needed under the constraint of limited computing resources.

In order to overcome this dilemma, we propose a probabilistic method for FIEFK. The output of this method is not a single evaluation (point estimate) but a distribution on entire solution space. This richer output contributes a lot to uncertainty quantification as it captures numerical/discretisation error through variance of distribution.

There is certainly evidence that probabilistic numerical method applies to ill-posed problem. It can be shown that there exists strong link between regularisation principle and probabilistic numerical method, which proves effectiveness of probabilistic numerical method from another perspective.

As for common numerical tasks, probabilistic numerical method is a better choice as well due to its various advantages. Probabilistic numerical method has been applied to several numerical problems such as linear algebra and optimisation and achieved ideal results.

Acknowledgements

I would like to thank my project supervisor Dr. Jon Cockayne for all his support, contributions and patience throughout this work; particular to his continued help with my work during his paternity leave.

I would also like to thank my parents. Without your support, I would not have the precious chance to this master's course. It was you who made this journey come true.

Contents

Abstract

The objective of this project is to develop a new Probabilistic Numerical Method for inferring the true value of Fredholm integral equation of the first kind. In detail, We will have two goals: introduce PNM and demonstrate how it can be used as a statistical solver in numerical tasks and derive a Probabilistic Numerical Method based on *Probabilistic Meshless Method* (PMM, [1]) in solving Fredholm Integral Equation of the first kind. Examples will be presented for the illustration of both objectives.

List of Notations and Abbreviations

The following list describes notation and abbreviations that are used frequently in this work:

FIEFK	Fredholm Integral Equation of the First Kind
PNM	Probabilistic Numerical Method
PMM	Probabilistic Meshless Method
RKHS	Reproducing Kernel Hilbert Space
SPD	Symmetric Positive Definite
GP	Gaussian Process
QoI	Quantity of Interest
MAP	Maximum A Posterior
$\mathcal{X}$	A non-empty set
$\mathcal{H}$	A Hilbert space
$k(\cdot,\cdot)$	Gaussian Process kernel; also used for the reproducing kernel and symmetric positive definite function
$h(\cdot,\cdot)$	Integral kernel
K_h	Integral operator w.r.t. integral kernel h
$K_h^{\dagger}$	The adjoint of operator K_h
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$	Inner product in a Hilbert space $\mathcal{H}$
$\|\cdot\|_{\mathcal{H}}$	Norm associated to inner product in a Hilbert space $\mathcal{H}$

Chapter 1

Introduction

1.1 Motivation

In this project we will focus on the use of probabilistic numerical method (PNM) as a statistically valid tool in inference for the analytical numerical solution of Fredholm Integral Equation of the First Kind (FIE of the first kind, **FIEFK**). In fact, most of these integral equations are difficult in finding their explicit solutions or even they do not have analytical and computable solutions, given that **FIEFK** is an ill-posed problem in the sense that it does not have a unique solution or the solution of this problem is not well-behaved. The regular approach to these problems is to solve them numerically, as there is no known way to assign an exact value to them by rule-based method. Henceforth, numerical methods used to calculate such numbers are essentially of an approximate nature. However, the results are not the exact one and we do not know how far they are from the true answer. For common numerical methods which are based on discretisation, they inevitably produce uncertainties which can not be ignored until some scalar bounds reached. Therein, [2, Hennig et al.] proposes *Probabilistic Numerical Method* as an improvement. PNM regards numerical computation as statistical inference by endowing measure on Hilbert space, in the sense that PNM returns a full probability measure of uncertainties arising from discretising while classical numerical methods can only return a sole point estimate [3, p. 1]. This characteristic of PNM allows uncertainty to propagate in computational workflow, and avoids the potentially serious consequences of accumulating numerical errors in subsequent statistical inferences [1].

In this work, we will investigate Fredholm Integral Equation of the first kind by means of PNM. To be more specific, we consider the following **FIEFK** on a compact domain $D \subset \mathbb{R}^d$ with a given integral domain E:

$$g(\boldsymbol{x}) = \int_E h(\boldsymbol{x}, \boldsymbol{t}) f(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}, \ \boldsymbol{x} \in D \tag{1.1}$$

where $h(x,t) : D \times E \to \mathbb{R}$ is said to be an **integral kernel**. In (1.1), given $h(x,t)$ and

LHS term $g(x)$, underlying function $f(x)$ will be seen as a Gaussian Process(**GP**) under the framework of PNM whose mean function and kernel need to be further interrogated. We shall assume in this work that domain E is finite dimensional and real-valued.

Remark. If the integral kernel $h(x, t)$ is a function only of the difference of its arguments, namely $h(x - t)$, and E is infinite, then (1.1) can be seen as as a convolution of the functions h and f. Therefore, it can be analytically solved by means of Fourier transformation.

For simplicity, (1.1) can be rewritten as an operator equation:

$$g(\boldsymbol{x}) = K_h(f)(\boldsymbol{x}), \forall \boldsymbol{x} \in D \tag{1.2}$$

where K_h is called a **integral operator** associated with its integral kernel $h(x, t)$. We will demonstrate that operator K_h is linear later. Furthermore, the linearity of the operator K_h underpins the viability of PNM.

The objective of this project is to develop a new Probabilistic Numerical Method for inferring the true value of Fredholm integral equation of the first kind. In detail, We will have two goals:

- Introduce PNM and demonstrate how it can be used as a statistical solver in numerical tasks.
- Consider the use of methods similar to the *Probabilistic Meshless Method* (PMM, [1]) in solving Fredholm Integral Equation of the first kind.

Some developed PNMs will be provided in the first case. In the second case several numerical experiments will be presented for illustration.

In the next section, we will show a couple of existing classical method in solving **FIEFK**. Numerical evaluation of such equations is vital in many physical models and engineer problems as it is often the case that atmospheric science, medical imaging, signal processing and so on can be reduced to solving the **FIEFK**s (see [4–7]).

1.2 Brief Overview of Common methods in Solving FIE of the First Kind

In most literature on solving **FIEFK**, its ill-posedness as a key point is often considered because that makes solvability of **FIEFK**s different from other integral equations. Generally, the existence of solutions of the first kind of Fredholm integral equation usually has infinite solutions and quite large computation resource would be consumed in solving them by regular methods. In response to this issue, many scholars have studied the numerical solution of the **FIEFK**s from different aspects, and achieved ideal results.

The most common method, proposed by Tikhonov (see [8]), for solving **FIEFK**s is regularisation method. [8] also shows that the Tikhonov regularisation method is an effective method for solving ill-posed problems. The core idea is to evaluate the true solution of ill-posed problem with an approximated appropriate well-posed equation given some prior information about exact solution. These well-posed equations can always be solved analytically by regular methods and the solution is showed to be feasible and stable. Based on this work, Tanana et al. [9] propose a variational regularisation method with a regularisation parameter from the residual principle and reducing the problem to a system of linear algebraic equations. Another two similar methods involving direct orthogonal and boundary integral also have been validated by Caldwell [10].

Other methods have been investigated in the literature. [11] provides an extensive overview. Herein we briefly cite these now. In [12], wavelet methods is proposed for solving **FIEFK**. Similarly, [13] and [14] present a method combining wavelet, collocation and Tikhonov regularisation together to transform integral equation into algebraic equation and solve it. Iterative approaches to ill-posed integral equation have also been considered. For instance, [15] and [16] first propose iterative solution by using Galerkin multiscale method and make use of sparse matrices to obtain a fast algorithm. In [17] Iterative method have also been used to find the fast solution of equations that appears in the discrete regularisation method.

In this project we will be considering a emergent technique for producing numerical solvers for **FIEFK** that are statistical in nature. Approaches mentioned above do achieve ideal result for solving FIE of the first kind. However, the epistemic uncertainty arising from discretisation, namely numerical error, has not been focused on in the analysis. Instead, we will develop a specific *Probabilistic Numerical Method* which return a complete probability distribution over the solution functional space of **FIEFK**. Before explaining PNM and its principle, we will first go through relevant background theory.

1.3 Background Theory

Herein we start with stating some vital properties in operator theory which will be used in eliciting **RKHS**. We will then present the definition of **RKHS** and introduce a several relevant result derived from that. We will end up with providing a brief overview of Gaussian Process before constructing the methodology of PNM.

1.3.1 Linear Operator

We now briefly describe definition of a linear operator. We will follow the exposition given in [18, p.4].

Definition 1.3.1 (Linear Operator)**.** A function $A : \mathcal{F} \to \mathcal{G}$, where $\mathcal{F}$ and $\mathcal{G}$ are both normed linear spaces over $\mathbb{R}$, is called a **linear operator** if and only if it satisfies the following properties:

- **Homogeneity:** $A(\alpha f) = \alpha(Af) \;\; \forall x \in \mathbb{R}, f \in \mathcal{F}$
- **Additivity:** $A(f+g) = Af + Ag \;\; \forall f, g \in \mathcal{F}$

Example 1.3.1 (Equation 1.2)**.** Let $\mathcal{G}, \mathcal{F}$ be inner product space. For $f \in \mathcal{F}$, operator $K_h : \mathcal{F} \to \mathcal{G}$, defined with $K_h := \int h(\boldsymbol{x}, \boldsymbol{t}) f(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}$ is clearly a linear operator.

Example 1.3.2 (from p.4 [18])**.** Given an inner product space $\mathcal{F}$, for $g \in \mathcal{F}$, operator $A_g : \mathcal{F} \to \mathcal{G},\; A_g := \langle f, g \rangle_{\mathcal{F}}$ is a linear operator. Note that the image of A_g is $\mathbb{R}$, trivially, a normed linear space with $|\cdot|$. Such scalar-valued operators are called **functionals** on $\mathcal{F}$.

Having defined what an linear operator is, we now introduce another property which is equivalent to linearity for operators given in [18, p.5].

Definition 1.3.2 (Operator Norm)**.** The operator norm of a linear operator $A : \mathcal{F} \to \mathcal{G}$ is defined as:

$$\|A\| = \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}.$$

Definition 1.3.3 (Bounded Operator)**.** The linear operator $A : \mathcal{F} \to \mathcal{G}$ is said to be a bounded operator if $\|A\| < \infty$.

These three definitions can be unified by the following theorem presented at [18, p.6]:

Theorem 1.3.1. *Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ and $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ be normed linear space. If L is a linear operator, then the following three conditions are equivalent:*

- *L is a bounded operator.*
- *L is continuous on $\mathcal{F}$.*
- *L is continuous at one point of $\mathcal{F}$.*

It can be proved simply where we will not describe the proof here. So far, we have possess an essential ingredient in defining an **RKHS**. Before moving on to **RKHS**, we need another important theorem to unify the form of linear functional. In [18, p.6], it is presented as follows:

Theorem 1.3.2 (Riesz representation)**.** *In a Hilbert space $\mathcal{F}$, all continuous linear functionals are of the form $\langle \cdot, f \rangle$ for some $f \in \mathcal{F}$.*

Readers can refer to Rudin (1987, Theorem 4.12) as an extensive proof. We will not prove it in this work. Riesz representation theorem plays a key role in finding reproducing kernel. We will mention it again shortly after.

Lastly, another notion about operator will be given here:

Definition 1.3.4 (Adjoint Operator)**.** Let $L : V \to V$ be a linear operator on a inner product space V. The **adjoint** of L is a transformation:

$$L^{\dagger} : V \to V \quad s.t. \quad \langle L(\boldsymbol{x}), \boldsymbol{y} \rangle = \left\langle \boldsymbol{x}, L^{\dagger}(\boldsymbol{y}) \right\rangle, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in V$$

.

Example 1.3.3. Consider a Hilbert space $\mathcal{F}$ with elements $f \in \mathcal{F}$, the adjoint operator of $K_h : \mathcal{F} \to \mathcal{F}$ in (1.2) is defined as:

$$K_h^{\dagger}(f)(\boldsymbol{x}) = \int_D h(\boldsymbol{t}, \boldsymbol{x}) f(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}.$$

It can be shown trivially following Definition 1.3.4.

This example is necessary for subsequent derivation in Chapter 2 and 3.

1.3.2 Reproducing Kernel Hilbert Space

In this work, reproducing kernel Hilbert space and its properties are mainly used in finding a proper kernel for the prior probability measure of solution function in **FIEFK**. We now present one possible definition of an **RKHS** given in [18, p.7]. An **RKHS** is a Hilbert space which possesses some special properties and that makes itself relatively well-behaved. Let $\mathcal{H}$ be a Hilbert space of functions mapping from some non-empty set $\mathcal{X}$ to $\mathbb{R}$, i.e., $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. [1] Then, in **RKHS** $\mathcal{H}$, for $f, g \in \mathcal{H}$ which are close in $\|\cdot\|_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$, $f(x)$ and $g(x)$ will be close for all $x \in \mathcal{X}$. Before given the formal definition of an **RKHS**, we need to define a very special functional which assign concrete value to each $f \in \mathcal{H}$ at x. Here we take the definition of evaluation functional from [18, p.7]:

Definition 1.3.5 (Evaluation Functional)**.** Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, map $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x := f \mapsto f(x)$ is called the (Dirac) evaluation functional at x.

Given the Definition 1.3.5, we can directly obtain the linearity of evaluation functional. In [18, p.7], an **RKHS** can be defined as follows.

Definition 1.3.6 (Reproducing Kernel Hilbert Space)**.** A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space(**RKHS**) if δ_x is continuous for $\forall x \in \mathcal{X}$.

[1] Broadly, the range space of functions in $\mathcal{H}$ can be extend to $\mathbb{C}$ equipped with definition of **conjugate symmetry**.

According to Theorem 1.3.1, we can know that $\delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded operator in such an **RKHS** for $\forall x \in \mathcal{X}$. Based on that, a useful consequence that **RKHS** are particular well-behaved can be easily derived [18, p.7].

Corollary 1.3.6.1 (Norm convergence in $\mathcal{H}$ implies pointwise convergence)**.** *If two functions converge in RKHS norm, then they converge at every point, i.e., if* $\lim_{n\to\infty}\|f_n - f\|_{\mathcal{H}}= 0$*, then* $\lim_{n\to\infty} f_n(x) = f(x), \forall x \in \mathcal{X}$.

Proof. Given δ_x is continuous in Definition 1.3.6, for any $x \in \mathcal{X} \subset \mathbb{R}$, in RKHS $\mathcal{H}$:

$$\begin{aligned} |f_n(x) - f(x)| &= |\delta_x f_n - \delta_x f| \\ &\leq \|\delta_x\|\|f_n - f\|_{\mathcal{H}} \end{aligned}$$

since δ_x is also bounded(Theorem 1.3.1) in RKHS $\mathcal{H}$. On the other hand, $\|f_n - f\|_{\mathcal{H}}\to 0$ when $n \to \infty$. Hence, $\lim_{n\to\infty} f_n(x) = f(x)$.

□

After the work above, we have built the concept of **RKHS** without mentioning kernels. We will then focus on describing how kernel fits in **RKHS**. We start by stating the definition of reproducing kernel given in [18, p.8].

Definition 1.3.7 (Reproducing kernel)**.** Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if it satisfies

- $\forall x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X},\ \forall f \in \mathcal{H}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

In particular, for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} \tag{1.3}$$

Having defined what an reproducing kernel is we are now in the position to consider its relationship between **RKHS**. Now, we will show that for every **RKHS** $\mathcal{H}$, there exists a unique reproducing kernel $k \in \mathcal{H}$. The relevant proofs come from [18, p.8-9].

Proposition 1.3.1 (Existence of the reproducing kernel)**.** *Given a Hilbert space* $\mathcal{H}$*,*

$$\mathcal{H} \textit{ is an RKHS} \quad \Longleftrightarrow \quad \mathcal{H} \textit{ has a reproducing kernel}$$

Proof. (Sufficiency) Suppose $\mathcal{H}$ is an RKHS. By Definition 1.3.6, for any $x \in \mathcal{X}$ we have evaluation functional δ_x is continuous at every x. Thus, by Riesz representation theorem (Theorem 1.3.2), we say that:

$$\exists f_{\delta_x} \in \mathcal{H}\ s.t.\ \delta_x(f) = \langle f, f_{\delta_x}\rangle_{\mathcal{H}}\ for\ \forall f \in \mathcal{H}.$$

Define $k(x', x) = f_{\delta_x}(x')$, $\forall x, x' \in \mathcal{X}$. Obviously, k is a reproducing kernel since $k(\cdot, x) = f_{\delta_x} \in \mathcal{H}$ and $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = \langle f, f_{\delta_x}\rangle_{\mathcal{H}} = \delta_x(f) = f(x)$.

(Necessity) Assume $\mathcal{H}$ has a reproducing kernel, i.e., $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$, then

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

where the Cauchy-Schwarz inequality and the reproducing property are used to obtain the bound of δ_x. As $\sqrt{k(x, x)} < \infty$, we have $\|\delta_x\|_{\mathcal{H}} < \infty$ and by Theorem 1.3.1 δ_x is continuous as RKHS required in Definition 1.3.6. Consequently, $\mathcal{H}$ is an **RKHS**. □

Proposition 1.3.2 (Uniqueness of the reproducing kernel)**.** *If it exists, reproducing kernel is unique.*

Proof. Assume that $\mathcal{H}$ has two reproducing kernels k_1 and k_2. Then,

$$\langle f, k_1(\cdot, x) - k_2(\cdot, x)\rangle_{\mathcal{H}} = f(x) - f(x) = 0$$

holds for any $f \in \mathcal{H}$ and $x \in \mathcal{X}$ due to conjugate symmetry and linearity of inner product.
In particular, if we take $f = k_1(\cdot, x) - k_2(\cdot, x)$, we have $\|k_1(\cdot, x) - k_2(\cdot, x)\|^2_{\mathcal{H}} = 0$ for any $x \in \mathcal{X}$. Hence, $k_1 = k_2$.

□

From above proofs, we actually see $k(\cdot, x)$ is the *representer of evaluation* at x. Henceforth, an **RKHS** is constructed by reproducing kernel.

We will end up this subsection with introducing Mercer Theorem, which provides an alternative construction of **RKHS**. Mercer Theorem is also the source of ”kernel tricks” in many machine learning literature, including Gaussian Process in next part. Previous to introduction we briefly enumerate a several definitions needed in Mercer Theorem by following [19, p.1-7]:

Definition 1.3.8 (L^p space)**.** Let $(\mathcal{X}, \Sigma, \mu)$ be a measurable space and $1 \leq p < \infty$. The space $L^p(\mathcal{X})$ consists of equivalence classes of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$\int |f|^p d\mu < \infty$$

where two measurable functions are equivalent if they are equal almost everywhere w.r.t. μ. The L^p-norm of $f \in L^p(\mathcal{X})$ is defined by

$$\|f\|_{L^p}= \left(\int |f|^p d\mu \right)^{1/p}$$

Example 1.3.4 (ℓ^p space)**.** If $\mathbb{N}$ is equipped with counting measure, then $L^p(\mathbb{N})$ consists of all sequences $\{x_n \in \mathbb{R} : n \in \mathbb{N}\}$ such that

$$\sum_{n=1}^{\infty} |x_n|^p < \infty.$$

We denote this sequence space as $\ell^p(\mathbb{N})$ w.r.t. norm

$$\|\{x_n\}\|_{\ell^p}= \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{1/p}$$

Definition 1.3.9 (Essential Supremum)**.** Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function on a measure space $(\mathcal{X}, \Sigma, \mu)$. The essential supremum of f on $\mathcal{X}$ is

$$\operatorname*{ess\,sup}_{\mathcal{X}} f = \inf\{a \in \mathbb{R} : \mu\{x \in \mathcal{X} : f(x) > a\}\}.$$

Eqivalently,

$$\operatorname*{ess\,sup}_{\mathcal{X}} f = \inf \left\{ \sup_{\mathcal{X}} g : g = f \text{ pointwise a.e.} \right\}.$$

Thus, the essential supremum of a function depends only on its μ-a.e. equivalence class. We say that f is *essentially bounded* on $\mathcal{X}$ if

$$\operatorname*{ess\,sup}_{\mathcal{X}} |f| < \infty$$

Definition 1.3.10 (L^∞ space)**.** Given a measurable space $(\mathcal{X}, \Sigma, \mu)$, the space $L^\infty(\mathcal{X})$ consists of pointwise a.e.-equivalence classes of essentially bounded measurable functions $f : \mathcal{X} \to \mathbb{R}$ with norm

$$\|\{x_n\}\|_{L^\infty}= \operatorname*{ess\,sup}_{\mathcal{X}} |f|.$$

With above definitions, we can formally present Mercer's Theorem following [20, p.37]:

Theorem 1.3.3 (Mercer)**.** *Suppose $k \in L^\infty(\mathcal{X} \times \mathcal{X})$ is a symmetric real-valued function* *such that the integral operator* (1.2): [2]

$$K_k : L^2(\mathcal{X}) \to L^2(\mathcal{X}) \tag{1.4}$$

$$K_k(f)(x) := \int_{\mathcal{X}} k(x,t)f(t)\mathrm{d}t,\ \ x,t \in \mathcal{X} \tag{1.5}$$

is positive definite; that is, $\forall f \in L^2(\mathcal{X})$*, we have*

$$\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,t)f(x)f(t)\mathrm{d}x\mathrm{d}t \geq 0. \tag{1.6}$$

Let $\psi_j \in L^2(\mathcal{X})$ *be the normalised orthogonal eigenfunctions of* K *associated with the eigenvalues* $\lambda_j > 0$*, sorted in non-increasing order. Then*

1. $(\lambda_j)_j \in \ell^1$

2. $k(x,t) = \sum_{j=1}^{N_{\mathcal{H}}} \lambda_j \psi_j(x)\psi_j(t)$ *holds for almost all* $(x,t) \in \mathcal{X} \times \mathcal{X}$*. Either* $N_{\mathcal{H}} \in \mathbb{N}$*, or* $N_{\mathcal{H}} = \infty$*; in the latter case, the series converges absolutely and uniformly for almost all* (x,t)*.*

Instantly, we have the following:

Proposition 1.3.3 (Mercer Kernel Map)**.** *Given a real Hilbert space* $\mathcal{H}$*, if* $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ *is a function satisfying the condition of Theorem 1.3.3, we can construct a map* $\phi : \mathcal{X} \to \mathcal{H}$ *such that,*

$$\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x,y) \tag{1.7}$$

for any $x, y \in \mathcal{X} \subset \mathbb{R}$*.*

Proof. ([20, see p.37]) From the statement 2 of Theorem 1.3.3 it follows that $k(x,t)$ corresponds to a dot product in Hilbert space $\mathcal{F} = L^2(\mathcal{X}^{N_{\mathcal{H}}})$, since $k(x,t) = \langle \phi(x), \phi(t) \rangle_{\mathcal{F}}$ with:

$$\phi : \mathcal{X} \to \mathcal{F}, \quad \phi := x \mapsto (\sqrt{\lambda_j}\psi_j(x))_{j=1,\cdots,N_{\mathcal{H}}}$$

□

Remark. Often in other literature, the function which can be written as inner product of maps above is also used in defining kernel (different from reproducing kernel). Such functions are called as Mercer kernel. The map ϕ is often referred to as the feature map and the space $\mathcal{H}$ as the feature space.

Next, We will present definition of positive definite kernel and show how it corresponds to Mercer kernel. To do so, herein we take the concept of positive definite kernel functions from both [18, p.9] and [20, p.30].

[2] L^∞ space is the normed Banach space of essentially bounded measurable functions with the essential supremum norm.

Definition 1.3.11 (Positive definite kernel)**.** A symmetric[3] function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1,\ \forall (x_1, \ldots, x_n) \in \mathcal{X}^n,\ \forall (a_1, \ldots, a_n) \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0. \tag{1.8}$$

The function $k(\cdot, \cdot)$ is *strictly* positive definite if for any set $X = [x_1, \ldots, x_n]$ s.t. x_i are mutually distinct, the equality holds only when all the a_i are zero. Such function is called a positive definite kernel. Often, we shall refer to it simply as a kernel.

Now we pose a lemma in [18, p.9] which clearly tells us Mercer kernels are positive definite kernels.

Lemma 1.3.11.1. *Let $\mathcal{F}$ be any Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{F}$. Then $k(x, y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$ is a positive definite kernel function.*

Proof. Using the linearity of inner product, we have

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \left\langle \phi(x_i), \phi(x_j) \right\rangle_{\mathcal{F}} \\ &= \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{i=1}^{n} a_j \phi(x_j) \right\rangle_{\mathcal{F}} \\ &= \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{F}}^{2} \geq 0 \end{aligned}$$

□

In addition, recall that reproducing kernel has reproducing property (1.3). Thus, we obtain the following corollary:

Corollary 1.3.11.1. *Reproducing kernels are positive definite kernels.*

Proof. For a reproducing kernel k in an **RKHS**, we can take $\phi : x \mapsto k(\cdot, x)$. Then, its reproducing property $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ by (1.3) satisfies Lemma 1.3.11.1.

□

Remark. In the proof above, we also see any reproducing kernel k in an **RKHS** $\mathcal{H}$ is also a Mercer kernel when the feature map is properly defined.

So far, we have seen that any **RKHS** uniquely determines its reproducing kernel k which is also a positive definite kernel and is also a Mercer kernel. Mercer kernel is a

[3] Namely, for real-valued function k, $k(x, y) = k(y, x)$.

positive definite kernel as well. In fact, these three concepts of reproducing kernel, Mercer kernel and positive definite kernel are equivalent due to Moore-Aronszajn Theorem [4] (see [18, p.12-18] for details). This consequence is of great importance both in Gaussian Process and solving **FIEFK**.

1.3.3 Gaussian Process Regression

In this subsection, we will motivate the Gaussian Process Regression(**GPR**) which can be related closely to the kernel under **RKHS** we have formally defined above. **GPR** is the key to probabilistic numerical methods. Before moving on to Gaussian Process, we will first start with so-called *Gaussian Algebra* given in [3, p.23-25].

A D-dimensional random variable follows Gaussian distribution, say $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, means it has following probability density function:

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{D/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1.9}$$

where the parameter vector $\boldsymbol{\mu} \in \mathbb{R}^D$ specifies the mean of the distribution, and the symmetric *positive definite* matrix[5] $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ defines the covariance of the distribution. We display the form of one element here:

$$\mu_i = \mathbb{E}_{N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}(X_i), \quad \boldsymbol{\Sigma}_{ij} = \mathrm{cov}_{N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}(X_i, X_j).$$

Gaussian distribution can be seen frequently in many disciplines as it arises from Central Limit Theorem. However, Gaussian Process does not involve this theorem. What makes Gaussian densities practical is that they have some good properties of linear algebra. Herein, we conclude them from [3, p.24]:

Proposition 1.3.4 (Sum rule)**.** *If a variable $\boldsymbol{x} \in \mathbb{R}^D$ follows Gaussian distribution, then every affine transformation of $\boldsymbol{x}$ is also normal distributed:*

$$\begin{aligned} &\textit{If } p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \textit{ and } \boldsymbol{y} := A\boldsymbol{x} + \boldsymbol{b} \textit{ for } A \in \mathbb{R}^{M\times D},\ b \in \mathbb{R}^M, \\ &\textit{then } p(\boldsymbol{y}) = N(\boldsymbol{y}; A\boldsymbol{\mu} + \boldsymbol{b}, A\boldsymbol{\Sigma}A^{\intercal}) \end{aligned} \tag{1.10}$$

Proposition 1.3.5 (Product rule)**.** *The product of two Gaussian probability density functions is another Gaussian probability distribution, scaled by a constant.*[6] *The value of that constant is itself given by the value of a Gaussian density function.*

$$\begin{aligned} &N(\boldsymbol{x}; \boldsymbol{a}, A)N(\boldsymbol{x}; \boldsymbol{b}, B) = N(\boldsymbol{x}; \boldsymbol{c}, C)N(\boldsymbol{a}; \boldsymbol{b}, A + B), \\ &\textit{where } C := (A^{-1} + B^{-1})^{-1}, \quad \boldsymbol{c} := C(A^{-1}\boldsymbol{a} + B^{-1}\boldsymbol{b}) \end{aligned} \tag{1.11}$$

[4] Briefly speaking, it tells us every pd kernel is a reproducing kernel which induces an unique **RKHS**.

[5] Similar to pd kernel in Definition 1.3.11: substitute function by matrix.

[6] This statement is about the product of two functions, not about two random variables.

These two propositions can be derived by some algebraic computation which we will not go over here. We only note the significance of them. These two properties means Gaussian densities are preserved under all linear operations. They constitute the fundamental mechanism for *Gaussian inference*[7](see [3, p.24]). We simply pose it here.

Proposition 1.3.6 (Gaussian inference)**.** *If the variable $\boldsymbol{x} \in \mathbb{R}^D$ is assigned a Gaussian prior, and observation $\boldsymbol{y} \in \mathbb{R}^M$ given $\boldsymbol{x}$,are Gaussian distributed:*

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \quad p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y}; A\boldsymbol{\mu} + \boldsymbol{b}, \Lambda),$$

then both the posterior and marginal distribution for $\boldsymbol{y}$ are Gaussian:[8]

$$\begin{aligned}
p(\boldsymbol{x} \mid \boldsymbol{y}) &= N(\boldsymbol{x}; \widetilde{\boldsymbol{\mu}}, \widetilde{\Sigma}), \quad \textit{with} && (1.12)\\
\widetilde{\Sigma} :&= (\Sigma^{-1} + A^\intercal \Lambda^{-1} A)^{-1} && (1.13)\\
&= \Sigma - \Sigma A^\intercal (A \Sigma A^\intercal + \Lambda)^{-1} A \Sigma && (1.14)\\
\widetilde{\boldsymbol{\mu}} :&= \widetilde{\Sigma}(A^\intercal \Lambda^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \Sigma^{-1}\boldsymbol{\mu}) && (1.15)\\
&= \boldsymbol{\mu} + \Sigma A^\intercal (A \Sigma A^\intercal + \Lambda)^{-1}(\boldsymbol{y} - (A\boldsymbol{\mu} + \boldsymbol{b})); && (1.16)\\
\textit{and} \quad p(\boldsymbol{y}) &= N(\boldsymbol{y}; A\boldsymbol{\mu} + \boldsymbol{b}, A\boldsymbol{\Sigma} A^\intercal + \Lambda). && (1.17)
\end{aligned}$$

Remark. Due to matrix inversion lemma (see [3, p.129]), (1.14) and (1.16) can be derived from (1.13) and (1.15). The reason for this specific step is to simplify the matrix computation. The former pair contains a matrix inverse of size $M \times M$, the latter on of size $D \times D$. In practical, we choose the smaller one between D and M.

We now present an important special example which from [3, p.25] and we will use it as motivation in discussing Gaussian Process:

Example 1.3.5 (Gaussian Algebra: Marginalisation & Conditioning)**.** Given a multidimensional Gaussian distributed random variable $\boldsymbol{x} \in \mathbb{R}^D$, consider a separation of $\boldsymbol{x} = [\boldsymbol{a}, \boldsymbol{b}]^\intercal$ into $\boldsymbol{a} \in \mathbb{R}^d$ and $\boldsymbol{b} \in \mathbb{R}^{D-d}$:

$$p(\boldsymbol{x}) = N\left(\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}, \begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}\right)$$

Recall that in sum rule (1.10), let $A = [I_d, \boldsymbol{0}_{D-d}]$. Then, the marginal of $p(\boldsymbol{x})$, i.e. $p(\boldsymbol{a}) = p(A\boldsymbol{x})$, is another Gaussian density, whose mean and covariance are a sub-vector and a sub-matrix of the full mean and covariance:

$$p(\boldsymbol{a}) = \int p(\boldsymbol{a}, \boldsymbol{b}) d\boldsymbol{b} = N(\boldsymbol{x}; \boldsymbol{\mu}_a, \Sigma_{aa}) \tag{1.18}$$

[7] Namely, Bayesian linear model regression: infer posterior given prior and likelihood using Gaussian densities.

[8] Often in the literature $\boldsymbol{y}$ here is referred as the *evidence*.

Further, by Gaussian inference (1.12), we see that the conditional of a subset conditioned on its complement is also a Gaussian:

$$p(\boldsymbol{a} \mid \boldsymbol{b}) = N(\boldsymbol{a}; \boldsymbol{\mu}_a + \Sigma_{ab}(\Sigma_{bb})^{-1}\boldsymbol{b}, \Sigma_{aa} - \Sigma_{ab}(\Sigma_{bb})^{-1}\Sigma_{ba}). \tag{1.19}$$

In this example, we see that the above two basic operations in statistical inference are enclosed to Gaussian distribution. Henceforth, we say that Gaussian properties map probability theory into matrix algebra, in the sense that *Gaussian Algebra*. It is vital to regression and we will frequently make use of Gaussian algebra in following discussion.

We now introduce **GPR**. **GPR** is essentially a non-parametric approach which is distinguished from common parametric inference. Herein we motivate it by parametric Gaussian regression based on feature map we discussed in Mercer Theorem 1.3.3 (see [3, p.27]).

To do so, we assume that any real-valued $f : \mathcal{X} \to \mathbb{R}$ can be written as a weighted sum over a finite number F of feature functions $[\phi_i : \mathcal{X} \to \mathbb{R}]_{i=1,\cdots,F}$:

$$f(x) = \sum_{i=1}^{F} \phi_i(x) w_i := \boldsymbol{\Phi}_x^\intercal \boldsymbol{w} \quad \text{with} \quad \boldsymbol{w} \in \mathbb{R}^F \tag{1.20}$$

where parameter vector $\boldsymbol{w}$ is called as weight and $\boldsymbol{\Phi}_x$ is a F-row vector. Generally, when input of f is a vector $X \subset \mathcal{X}$ with $[X]_j = x_j \in \mathcal{X}, j = 1, \cdots, N$, we will denote by $\boldsymbol{\Phi}_X \in \mathbb{R}^{F \times N}$ the feature matrix with element $[\boldsymbol{\Phi}_X]_{ij} = \phi_i(x_j)$. Similarly, we denote by $\boldsymbol{f}_X$ the vector of outcome $[f(x_1), \cdots, f(x_N)]$.

Note (1.20) is a linear function of weight $\boldsymbol{w}$. Thus, a linear regression model is constructed. In order to perform Gaussian inference we assign a Gaussian density for weight vector $\boldsymbol{w}$,[9]

$$p(\boldsymbol{w}) = N(\boldsymbol{w}; \boldsymbol{\mu}, \Sigma).$$

Let $\boldsymbol{y}$ be observations corrupted by Gaussian noise at the locations X: $\boldsymbol{y} = \boldsymbol{f}_X + \sigma$ with $\boldsymbol{y} \in \mathbb{R}^N,\ \sigma \sim N(0, \Lambda),\ \Lambda \in \mathbb{R}^{N \times N}$. Thus, We have the following likelihood function:

$$p(\boldsymbol{y} \mid f) = N(\boldsymbol{y}; \boldsymbol{f}_X, \Lambda). \tag{1.21}$$

[9] Also known as Bayesian linear regression in machine learning community.

By Gaussian algebra (1.19), the posterior over the weights $\boldsymbol{w}$ is:

$$\begin{aligned}
p(\boldsymbol{w} \mid \boldsymbol{y}) &= N(\boldsymbol{w}; \widetilde{\boldsymbol{\mu}}, \widetilde{\Sigma}), \quad \text{with} \\
\widetilde{\Sigma} &:= (\Sigma^{-1} + \boldsymbol{\Phi}_X \Lambda^{-1} \boldsymbol{\Phi}_X^{\intercal})^{-1} \\
&= \Sigma - \Sigma \boldsymbol{\Phi}_X (\boldsymbol{\Phi}_X^{\intercal} \Sigma \boldsymbol{\Phi}_X + \Lambda)^{-1} \boldsymbol{\Phi}_X^{\intercal} \Sigma \\
\widetilde{\boldsymbol{\mu}} &:= \widetilde{\Sigma}(\boldsymbol{\Phi}_X \Lambda^{-1} \boldsymbol{y} + \Sigma^{-1} \boldsymbol{\mu}) \\
&= \boldsymbol{\mu} + \Sigma \boldsymbol{\Phi}_X (\boldsymbol{\Phi}_X^{\intercal} \Sigma \boldsymbol{\Phi}_X + \Lambda)^{-1} (\boldsymbol{y} - \boldsymbol{\Phi}_X^{\intercal} \boldsymbol{\mu}).
\end{aligned}$$

This is known as weight-space view of regression in [21, p.12]. Alternatively, we can construct a function-space view, in the sense that a probability density directly over values of unknown $f(\boldsymbol{x}) = \boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \boldsymbol{w}$. Immediately, by sum rule (1.10), we have the posterior predictive distribution over function values $f_{\boldsymbol{x}}$ at a finite subset $\boldsymbol{x} \subset \mathcal{X}$:

$$p(f_{\boldsymbol{x}}) = N(f_{\boldsymbol{x}}; \boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \widetilde{\boldsymbol{\mu}}, \boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \widetilde{\Sigma} \boldsymbol{\Phi}_{\boldsymbol{x}}) \tag{1.22}$$

where

$$\begin{aligned}
\boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \widetilde{\boldsymbol{\mu}} &= \boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \boldsymbol{\mu} + \boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \Sigma \boldsymbol{\Phi}_X (\boldsymbol{\Phi}_X^{\intercal} \Sigma \boldsymbol{\Phi}_X + \Lambda)^{-1} (\boldsymbol{y} - \boldsymbol{\Phi}_X^{\intercal} \boldsymbol{\mu}), \\
\boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \widetilde{\Sigma} \boldsymbol{\Phi}_{\boldsymbol{x}} &= \boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \Sigma \boldsymbol{\Phi}_{\boldsymbol{x}} - \boldsymbol{\Phi}_{\boldsymbol{x}}^{\intercal} \Sigma \boldsymbol{\Phi}_X (\boldsymbol{\Phi}_X^{\intercal} \Sigma \boldsymbol{\Phi}_X + \Lambda)^{-1} \boldsymbol{\Phi}_X^{\intercal} \Sigma \boldsymbol{\Phi}_{\boldsymbol{x}}.
\end{aligned}$$

In fact, this consequence which we just obtained under a finite-dimensional Gaussian distribution is quite close to a **GP**. It can be extended to **GP** by enlarging the dimensionality of distribution to infinite. Recall that positive definite kernel(Definition 1.3.11) in last subsection. This kernel function is the key to describe the infinite limit of a covariance matrix which almost characterises a **GP**. We briefly showcase this technical from [3, p.31].

For general linear regression using features ϕ, the mean vector and covariance matrix of posterior on function values does not contain isolated, explicit form of the features. Instead, for finite subsets $a, b \subset \mathcal{X}$ it contains only projections and inner products:

$$m_a := \boldsymbol{\Phi}_a^{\intercal} \boldsymbol{\mu} \quad k_{ab} := \boldsymbol{\Phi}_a^{\intercal} \Sigma \boldsymbol{\Phi}_b = \sum_{i=1}^{F} \sum_{j=1}^{F} \phi_i(a) \phi_j(b) \Sigma_{ij}.$$

These two types of expressions are themselves functions, called the mean function $m : \mathcal{X} \to \mathbb{R}$ and the covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The mean function can be any tractable function because of linear shift[10] in likelihood. As covariance matrix Σ is positive definite, we can define $\Sigma^{1/2} = UD^{1/2}U^{\intercal}$ by singular value decomposition so that

[10] Recall that in (1.10), the linear shift is referred as to the intercept $\boldsymbol{b}$ in $\boldsymbol{y} := A\boldsymbol{x} + \boldsymbol{b}$.

$(\Sigma^{1/2})^\intercal(\Sigma^{1/2}) = UDU^\intercal = \Sigma$. Thus, we define a symmetric function:

$$k(a, b) = \langle\psi(a), \psi(b)\rangle := \psi(a) \cdot \psi(b), \quad \psi(a) = \Sigma^{1/2}\phi(a)$$

Clearly, we can see covariance function here is essentially a kernel function which we discussed in **RKHS**. Now we finally find an approach to define a infinite-dimensional Gaussian distribution without explicit stating a specific set of features: choose appropriate pair (m, k) instead of specifying mean and covariance matrix. This is called a non-parametric formulation of regression since parameters $\boldsymbol{w}$ are not explicitly represented in the computation.

Remark. The technical for constructing kernel here to simplify computation is often referred as *kernel trick* (see [20, p.34]). An extensive and more concrete example can be found in [3, p.32] which is the well-known radial basis function.

We now present one possible definition of Gaussian Process given in [3, p.34]:

Definition 1.3.12 (Gaussian Process)**.** Consider a function $m : \mathcal{X} \to \mathbb{R}$ and a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The Gaussian process $f \sim \mathbf{GP}(f; m, k)$ is the probability measure identified by the property that, for any finite subset $X := [x_1, \cdots, x_N] \subset \mathcal{X}$ the probability assigned to function values $f_X = [f(x_1), \cdots, f(x_N)]$ is given by the multivariate Gaussian probability density $p(f_X) = N(f_X; m(X), k(X, X))$.

The Gaussian process regression follows at once:

Proposition 1.3.7 (Gaussian Process Regression)**.** *Given a Gaussian process prior $f \sim \mathbf{GP}(f; m, k)$ over a unknown function $f : \mathcal{X} \to \mathbb{R}$ and the likelihood function $p(\boldsymbol{y} \mid f) = N(\boldsymbol{y}; f_X, \Lambda)$, the posterior over f is a Gaussian process $p(f \mid \boldsymbol{y}) = \mathbf{GP}(f; \mathfrak{m}, \mathfrak{v})$ with mean function and symmetric positive definite function $\mathfrak{m} : \mathcal{X} \to \mathbb{R}$ and $\mathfrak{v} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:*

$$\mathfrak{m}(x) = m(x) + k(x, X)[k(X, X) + \Lambda]^{-1}(\boldsymbol{y} - m(X)), \tag{1.23}$$

$$\mathfrak{v}(x, x') = k(x, x') - k(x, X)[k(X, X) + \Lambda]^{-1}k(X, x'). \tag{1.24}$$

Moreover, recall that linear projections keep Gaussian distributions closure. This property allows for an elegant extension of regression on functions from observed function values, which is of great importance for numerical uses of the Gaussian inference framework. We take the following from [3, p.39]:

Proposition 1.3.8 (Derivatives and Integrals)**.** *Given a $\mathbf{GP}(m, k)$ assigned to a real-valued function $f : \mathbb{R}^N \to \mathbb{R}$ where $m(x)$ and $k(x, x')$ are at least q-times continuously differentiable in all arguments and integrable. Then all partial derivatives and integrals are jointly Gaussian distributed, with mean functions:*

$$\mathbb{E}\left(\frac{\partial^\ell f(x)}{\partial x_i^\ell}\right) = \frac{\partial^\ell m(x)}{\partial x_i^\ell} \quad \textit{for } 0 \leq \ell \leq q \tag{1.25}$$

$$\mathbb{E}\left(\int_D f(x)\mathrm{d}x\right) = \int_D m(x)\mathrm{d}x, \tag{1.26}$$

and covariance functions:

$$cov\left(\frac{\partial^\ell f(x)}{\partial x_i^\ell}, \frac{\partial^m f(x)}{\partial x_j^m}\right) = \frac{\partial^\ell \partial^m k(x,x')}{\partial x_i^\ell \partial {x'}_j^m} \tag{1.27}$$

$$cov\left(\int_D f(x)\mathrm{d}x, \frac{\partial^m f(x)}{\partial x_j^m}\right) = \int_D \frac{\partial^m k(x,x')}{\partial {x'}_j^m}\mathrm{d}x, \tag{1.28}$$

and analogously for mixed partial derivatives and higher-order integrals.

Gaussian process makes inference feasible on infinite-dimensional hypothesis-spaces. It provides a statistical view in modeling numerical problems. **GP** together with regression forms the core of probabilistic numerical method and we will show this in next chapter. Now we conclude this subsection by making several remarks:

Remark. Under Gaussian framework, the probabilistic inference is closely connected to *kernel ridge regression* in notion of **RKHS**. Full explanation of this is in [3, p.37]. Therein, a theorem about worst-case approximation of ridge regression exhaustively displays that in the case of noise-free evaluations of f, the posterior variance of **GP** regression equals the worst-case approximation error if the true function is an element of the **RKHS**.

Remark. For Proposition 1.3.8 a more precisely description is that if L is a linear operator acting on f, then Lf, is a **GP** with mean function Lm and covariance function $LL^\dagger k$ given Lm and $Lk(\cdot, x')$ are bounded.[11] This can be seen as a pushforward of a **GP** through the linear operator. More details can be found in [22]. We will frequently utilise this property in next chapter.

[11] Note here L refers to action on first argument of the kernel function k, while the adjoint $L^\dagger$ refers to action on the second argument. This is in line with Definition 1.3.4.

Chapter 2

Probabilistic Numerical Method

In this chapter we will introduce the probabilistic numerical method and explain how it can be used to solve regular numerical tasks. We begin this chapter by establishing the formal definition of PNM. We will follow the procedure outlined in section 2 of [23].

2.1 What is a PNM

In this section we will assume the we are working on measurable functions. We denote a measurable space by $(\mathcal{X}, \Sigma_{\mathcal{X}})$ where $\mathcal{X}$ is a non-empty set and $\Sigma_{\mathcal{X}}$ is the Borel σ-algebra induced by $\mathcal{X}$. The shorthand $\mathcal{P}_{\mathcal{X}}$ will be used to denote the set of all probability measure on $(\mathcal{X}, \Sigma_{\mathcal{X}})$. $\rho \in \mathcal{P}_{\mathcal{X}}$ represents a probability distribution.

Before giving the formal definition of PNM, we motivate it by a concrete numerical approximation task which is from [23]. Consider a Lebesgue integral:

$$\int_D x(t)\nu(dt)$$

for some integrable function $x : D \to \mathbb{R}$, w.r.t. a measure ν on D. To evaluate this integral, an immediate idea is to investigate the integrand $x(t)$ at any $t \in D$. However, we can evaluate x at all t with a finite computational resource only if D is itself a finite set. In response to this situation, many integration algorithms were proposed based on information $[x(t_1), \ldots, x(t_n)]$ at locations $[t_1, \ldots, t_n]$.

To abstract the structure of this problem, we assume the state variable x exists in a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$. Let $A : \mathcal{X} \to \mathcal{A}$ be the *information operator* which provides information about x. By this notion, information operator in the Lebesgue integration problem is:

$$A(x) = \begin{bmatrix} x(t_1) \\ \vdots \\ x(t_n) \end{bmatrix} = a \in \mathcal{A}. \tag{2.1}$$

The space $\mathcal{X}$ here is a space of functions and its dimensionality can be high even infinite. In contrast, the space $\mathcal{A}$ of information is assumed to be finite-dimensional according to the finite computational resource. In addition, [23] poses a new concept, quantity of interest (QoI) $Q(x)$, defined by a map $Q : \mathcal{X} \to \mathcal{Q}$ into a measurable space $(\mathcal{Q}, \Sigma_{\mathcal{Q}})$. The function x itself may not be the QoI of a numerical problem. For instance, the QoI of Lebesgue integration is not x but $Q(x) = \int x(t)\nu(dt)$.

A successful algorithm means the information operator A is tailored to the QoI $Q(x)$. The common method to such numerical problem is to construct an algorithm which returns some approximation $\hat{q}(a)$ of $Q(x)$ given information $a \in \mathcal{A}$. These approximation usually stop at some point where the scalar bound of precision is satisfied. Additionally, their theoretical convergence order can be studied.

Distinguished from common method, the probabilistic numerical method start with the introduction of a random variable X on $(\mathcal{X}, \Sigma_{\mathcal{X}})$. [2] assumes the true state is fixed but unknown and the randomness is used an abstract device used to represent epistemic uncertainty about x prior to evolution of the information operator. On this basis [23] formalises PNM as follows:

Definition 2.1.1. (Belief Distribution) An element $\rho \in \mathcal{P}_{\mathcal{X}}$ is a *belief distribution* for x if it carries the formal semantics of belief about the true, unknown state variable x.

Definition 2.1.2. (Probabilistic Numerical Method) Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$, $(\mathcal{A}, \Sigma_{\mathcal{A}})$ and $(\mathcal{Q}, \Sigma_{\mathcal{Q}})$ be measurable spaces and let $A : \mathcal{X} \to \mathcal{A}$, $Q : \mathcal{X} \to \mathcal{Q}$ and $B : \mathcal{P}_{\mathcal{X}} \times \mathcal{A} \to \mathcal{A}$ where A and Q are measurable functions. The pair $M = (A, B)$ is called a *probabilistic numerical method* for estimation of a quantity of interest Q. The map A is called an *information operator*, and the map B is called a *belief update operator*.

[23] also gives that the output of a PNM is a distribution $B(\rho, a) \in \mathcal{P}_{\mathcal{Q}}$ as this holds the formal status of a belief distribution for the value of $Q(x)$ based on both the initial belief ρ about the value of x and information a that are input to the PNM. In next two sections, we will give specific examples of developed PNMs. We now conclude this section by two remarks on PNM:

Remark. An intuitive objection sometimes arises here: x itself is not random. Indeed x is fixed but in PNM notion we do not propose that x should be considered as such. Explanation can be found in [23] and [3, p.70]: X is a formal statistical device used to represent the epistemic uncertainty. For instance, in Lebesgue integral we evidently uncertain about its value on D because we cannot provide a correct value for it without further work. Hence, we replace its true value by a random variable.[1] Thus, there is no distinction from traditional statistics, in which x represents a fixed but unknown parameter and X embodies epistemic uncertainty about this parameter.

[1] A helpful but sloppy description is that under Dirac probability measure the true value itself can be seen as a random variable. In fact, the updated belief distribution $B(\rho, a)$ will converge to this Dirac distribution centred on true value if more and more information a is given(see [24, Figure.1]).

Remark. Bayesian theorem can be encoded into PNM framework as Bayesian probabilistic numerical method. It is well defined and thoroughly discussed in [23]. It is advocated due to its appealing Bayesian interpretation and ease of generalisation to pipelines of computation. We will not go into this in this work as we only work on single numerical computation. Nonetheless, the following two example in next two section are essentially Bayesian PNMs. Readers are invited to view [23] for further details.

2.2 Case 1: Integration

Numerical integration is not only one of the most common numerical tasks in real-life applications, but also forms the core of computational statistics such as computing marginal distribution and evidence in Bayesian approaches. In this section we will tackle integration from a probabilistic angle. Probabilistic integration, also known as Bayesian Quadrature, is a practical statistical tool for numerical integration. Briefly, Bayesian quadrature operates by evaluating the integrand at a set of states and returns a distribution over $\mathbb{R}$ that express belief about the true value of integration [24]. We will now summarise the key procedures of Bayesian quadrature from [3, p.75-85].

2.2.1 Probabilistic Integration

Here we consider the numerical approximation of Lebesgue integral discussed earlier. The generic form of an integral problem is as follows:

$$F = \int_D x(t)\nu(\mathrm{d}t), \quad t \in D$$

where ν is a measure. For simplicity, we assume its domain $D \subset \mathbb{R}$ is univariate and bounded, in the sense that we only consider univariate definite integration problem. Our goal is to compute F where the integrand x do not have a convenient closed form so that there exist epistemic uncertainty over F until x is actually evaluated at an input. [25] first propose a probabilistic model for this epistemic uncertainty.

Recall Gaussian inference model we discussed in Chapter 1, Gaussian models are powerful tool for inference due to their good properties (1.10) and (1.11). Furthermore, **GP** allows for an analytic formulation of integration as probabilistic inference by Proposition 1.3.8. Based on on that, [25] considered using **GP** in probabilistic integration.

Remark. Note that **GP** is not the only choice for prior in probabilistic integration. Other stochastic process such as Student-t process and Dirichlet process can also be considered. The crucial reason for using **GP** is that it is closed under linear projections. That guarantees the viability of algorithm for probabilistic integration. For other prior, their conjugate distribution may need to be investigated.

Probabilistic integration begins by assigning a **GP** prior to the integrand x, i.e., let $x(t) \sim \mathbf{GP}(m(t), k(t,t'))$. This Gaussian process prior amounts to a joint Gaussian measure over both function values $\boldsymbol{x} = x(T) = [x(t_1), \dots, x(t_n)]$ at a finite set $T = [t_1, \dots, t_n] \subset D$ and the integral F:

$$\begin{aligned} p(\boldsymbol{x}) &= N(\boldsymbol{x}; m(T), k(T,T)) \\ p(F) &= N\left(F; \int_D m(t)\nu(\mathrm{d}t), \iint_D k(t,t')\nu(\mathrm{d}t)\nu(\mathrm{d}t')\right). \end{aligned}$$

By sum rule (1.10), we have

$$p(F, \boldsymbol{x}) = N\left(\begin{bmatrix} \boldsymbol{x} \\ F \end{bmatrix}; \begin{bmatrix} m(T) \\ \int_D m(t)\nu(\mathrm{d}t) \end{bmatrix}, \begin{bmatrix} k(T,T) & \int_D k(T,t)\nu(\mathrm{d}t) \\ \int_D k(t,T)\nu(\mathrm{d}t) & \iint_D k(t,t')\nu(\mathrm{d}t)\nu(\mathrm{d}t') \end{bmatrix}\right).$$

Instantly, the posterior is also a Gaussian with mean $\mathfrak{m} \in \mathbb{R}$ and $\mathfrak{v} \in \mathbb{R}_+$. Denote functions $m(t)$ and $k(t,t')$ by m_t and $k_{tt'}$. Denote $1 \times n$ vector $k(t,T)$ by k_{tT}. Denote $n \times 1$ vectors $k(T,t')$ and $m(T)$ by $k_{Tt'}$ and m_T Also denote $n \times n$ matrix $k(T,T)$ by k_{TT} with its inverse k_{TT}^{-1}.

$$\begin{aligned} p(F \mid \boldsymbol{x}) &= N(F; \mathfrak{m}, \mathfrak{v}) && (2.2) \\ \mathfrak{m} &:= \int_D m_t + k_{tT}k_{TT}^{-1}(\boldsymbol{x} - m_T)\nu(\mathrm{d}t) && (2.3) \\ &= \mathfrak{m}_0 + \mathfrak{k}_T^{\intercal} k_{TT}^{-1}(\boldsymbol{x} - m_T) \\ \mathfrak{v} &:= \iint_D k_{tt'} - k_{tT}k_{TT}^{-1}k_{Tt'}\nu(\mathrm{d}t)\nu(\mathrm{d}t') && (2.4) \\ &= \mathfrak{K} - \mathfrak{k}_T^{\intercal} k_{TT}^{-1}\mathfrak{k}_T. \end{aligned}$$

Note that we must keep $p(F \mid \boldsymbol{x})$ tractable otherwise F still cannot be evaluated. Thus analytic forms for the following integrals are required:

$$\begin{aligned} \mathfrak{m}_0 &:= \int_D m(t)\nu(\mathrm{d}t) \in \mathbb{R}, && (2.5) \\ \mathfrak{k}(t_i) &:= \int_D k(t,t_i)\nu(\mathrm{d}t) \in \mathbb{R}, && (2.6) \\ \mathfrak{K} &:= \iint_D k(t,t')\nu(\mathrm{d}t)\nu(\mathrm{d}t') \in \mathbb{R}_+. && (2.7) \end{aligned}$$

When these expressions are tractable, (2.2) is essentially an closed form map of observations $\boldsymbol{x}$ onto a probability distribution over F. The maximum a posteriori(MAP)[2] value $\mathfrak{m}$ operates a point estimate of F whilst the rest of whole distribution plays the role of capturing uncertainty raised from finite inputs which are in line with the limited computational resources. Similarly, the posterior variance $\mathfrak{v}$ acts as an estimate of squared

[2] The MAP of a Gaussian distribution is its mean.

error of F, in the sense that a notion of uncertainty. Hence, probabilistic integration is now fully constructed for our goal that computing F via a statistical inference. We summarise it in following algorithm:

Algorithm 1 Probabilistic integration

```
Construct prior: let $x \sim \mathbf{GP}(m, k)$
Evaluate integrand at $T$: set $\boldsymbol{x} := x(T) = [x(t_1), \dots, x(t_n)]$
Infer posterior: $F \mid \boldsymbol{x} \sim N(F; \mathfrak{m}, \mathfrak{v})$
Output MAP estimator: $\mathfrak{m}$
```

In addition, there are several caveats for probabilistic integration, which we leave as remarks here.

Remark. Note the reason for keep (2.5), (2.6) and (2.7) tractable is that we need a practically useful method which minimises extra computation. One might consider some specific situations that these expressions are intractable. This will lead to an expensive algorithm albeit above three equations can be estimate. This reason also applies to all the PNMs that follow in this work.

Remark. In [24], the term quadrature rule is said to describe any functional L of the form:

$$L(x) = \sum_{i=1}^{n} w_i x(t_i),$$

for some sets $T = [t_1, \dots, t_n] \subset D$ and weights $\boldsymbol{w} = [w_1, \dots, w_n] \subset \mathbb{R}$. Intrinsically, (2.3) is a linear function of $\boldsymbol{x}$ if we use expression $\mathfrak{m} = u + \boldsymbol{w}^\intercal \boldsymbol{x}$ where $u = \mathfrak{m}_0 - \mathfrak{k}_T^\intercal k_{TT}^{-1} m_T$ and $w_i = \sum_{j=1}^{n} \mathfrak{k}_{t_j} [k_{TT}^{-1}]_{ji}$. That will be exactly same as quadrature rule in numerical analysis for integration. Furthermore, probabilistic integration is closely related to classical quadrature method by using different covariance functions in **GP** prior(see [3, p.87]).

Remark. In statistics and machine learning field the measure ν is frequently a probability measure. Typically, in Bayesian inference when calculating normalising constant, $\nu(x)$ is a prior distribution and $f(x)$ is likelihood function evaluated at some observed data point.

We now conclude this subsection by making the expression of probabilistic integration consistent with the Definition 2.1.2. Herein, our QoI is: $Q(x) = \int_D x(t)\mathrm{d}t = F$ and the state space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ is a Banach space of real-valued functions on $D \subset \mathbb{R}$ equipped with its Borel σ-algebra. The information $A(x) = a$ is $\boldsymbol{x} = [x(t_1), \dots, x(t_n)]$ and belief distribution ρ is a Gaussian measure(**GP**). Following [23], we define ρ^a to be the restriction of ρ to those functions which interpolate f at values X. The output of probabilistic integration $B(\rho, a) = p(F \mid \boldsymbol{x})$ is a pushforward[3] $Q_{\#}\rho^a$ of the probability measure ρ^a w.r.t. integration operator Q. Consequently, $Q_{\#}\rho^a$ is again a Gaussian distribution.

[3] Given a measurable operator $T : \mathcal{X} \to \mathcal{B}$, the pushforward $T_{\#}\rho$ of a distribution $\rho \in \mathcal{P}_{\mathcal{X}}$ is defined as $T_{\#}\rho(B) = \rho(T^{-1}(B))$ for all $B \in \Sigma_{\mathcal{B}}$

2.2.2 Prior Selection

After the introduction of probabilistic integration, a question arises naturally: Is prior selection free? The answer is no. Not all Gaussian process priors will be viable for probabilistic integration. To argue for one over another, we need to show that either quantity could make expression (2.5), (2.6) and (2.7) tractable to support this interpretation of MAP estimate and error estimate. To demonstrate this we follow [3, p.77] to construct a concrete probabilistic integration model.

Given a fixed set of nodes, probabilistic integration require two choice. One is the measure $\nu(x)$ which is often specified in problems. The rest is prior selection we pay primary attention to. Within Bayesian quadrature, the choice of **GP** prior leads to two further choice: that of the prior mean $m(t)$ and the prior covariance $k(t, t')$. As mentioned above, we need these choices to be made such that (2.5)-(2.7) are in closed form. The mean m carries a initial guess for the integrand , while the covariance should aim to capture both the deviation of the true integrand from m and knowledge about analytic structure of the integrand[4] apart from that represented by m.

In practice, a popular choice for mean function is zero: $m(t) = 0$ because it is rare that an integrand is known sufficiently well ahead of its evaluation to support any better-informed choice. Moreover, the zero mean function also makes $\mathfrak{m}$ from (2.5) trivially zero and hence simplify the posterior mean of probabilistic integration (2.3).

As for covariance function, a good choice is Gaussian kernel function(radial basis function) due to its enclosure under linear projections. Consider using a generic Gaussian kernel paired with Gaussian measure[5]:

$$k(t, t') = N(t; t', \lambda^2) \quad \nu(t) = N(t; \mu, \sigma^2).$$

Thus it yields tractable results for the integrals required for probabilistic integration:

$$\begin{aligned}
&p(F \mid \boldsymbol{x}) = N(F; \mathfrak{m}, \mathfrak{v}) \\
&\mathfrak{m} = \mathfrak{k}_T^\intercal k_{TT}^{-1} \boldsymbol{x} \\
&\mathfrak{v} = \mathfrak{K} - \mathfrak{k}_T^\intercal k_{TT}^{-1} \mathfrak{k}_T \quad \text{where} \\
&[\mathfrak{k}]_i = N(t_i; \mu, \lambda^2 + \sigma^2) \quad i = 1, \dots, n \\
&\mathfrak{K} = \frac{1}{\sqrt{2\pi(2\sigma^2 + \lambda^2)}}
\end{aligned}$$

This is an example of pairing (k, ν) which makes probabilistic integration work. No-

[4] For instance, periodicity and differentiability.

[5] By Radon-Nikodym Theorem, the Radon-Nikodym derivative of Gaussian measure w.r.t. Lebesgue measure is a Gaussian density function.

tably Gaussian kernel is not the only choice for tractable model. [3, Table 10.1] reviews other many pairings of prior measure and Gaussian process kernel that yield closed form expression for probabilistic integration. That also illustrates the interpretability of probabilistic integration to some extend.

Prior to introduce next application of PNM, there are a few more points to note:

Remark. Bayesian quadrature is readily extended to integral over $d > 1$ variables, namely multivariate integral. To do so, the only requirement is choice of a Gaussian process prior over the d-dimensional space that yields tractable result against measure. However, such multivariate probabilistic integration is unlikely to work well due to curse of dimensionality. In order to perform high-dimensional Bayesian quadrature we need number of evaluations grows exponentially. This essentially comes at the high cost of computational resources. Thus, [26] proposes a counter-prior and achieves good result in multivariate integral.

Remark. The probabilistic integration model engenders a natural rule for node selection according to minimisation of expected loss, one loss function suggests itself: the squared error $\mathfrak{v}$. we will not discuss the optimisation of grid design in this work. We direct the interested reader to [3, p.94] for a thorough discussion of this design rule.

2.3 Case 2: Partial Differential Equation

Aside from numerical integration, we may consider another common type of numerical problem, solving differential equations. Similar to integration, differentiation is also regarded as linear operator so that we may consider the use of PNM for this category. Differential equations is frequently used in dynamical systems as mathematical description for phenomena of interest such as liquid flow and heat transfer. In this section we will concentrate on Partial Differential Equation(PDE).

Most of PDEs do not have a closed form solution so we need to consider numerical solution. Traditional approaches such as finite element methods(FEM) and finite difference methods(FDM) have been established as numerical approximations based on the discretisation of continuous equations. These methods are theoretically analysed in literature and have been proved that ideal result can be attained via bounding approximation error as a function of the discretesation parameters. However, reductions in such approximation error become expensive to computational cost. In order to overcome this problem, [1] proposes a PNM, *probabilistic meshless method*(PMM), which computes PDE with a coarse discretisation and yields a meaningful solution. We will now demonstrate this method following the section 3 of [1].

2.3.1 Probabilistic Meshless Method

In this section we restrict all operators to be linear.[6] Given a compact domain $D \subset \mathbb{R}^d$ with Lipschitz boundary ∂D, here we consider a generic PDE written in operator equation form as follows:

$$A(u)(\boldsymbol{x}) = g(\boldsymbol{x}) \quad \boldsymbol{x} \in D \tag{2.8}$$

$$B(u)(\boldsymbol{x}) = b(\boldsymbol{x}) \quad \boldsymbol{x} \in \partial D \tag{2.9}$$

where $A : \mathcal{H}(D) \to \mathcal{H}_A(D)$ and $B : \mathcal{H}(D) \to \mathcal{H}_B(\partial D)$ are two measurable operators among Hilbert space of functions $\mathcal{H}(D)$, $\mathcal{H}_A(D)$ and $\mathcal{H}_B(D)$. To be more specific, we associate A with a PDE of interest and B with any initial or boundary conditions. Similarly,$g \in \mathcal{H}_A(D)$ and $b \in \mathcal{H}_B(D)$ are regarded as forcing and boundary terms for the PDE. We aim to compute the solution $u(x) \in \mathcal{H}(D)$ which often does not possess an analytic form in the setting above so that we are uncertain about u before we evaluate g and b at some data points. This is also formalised in [1] as the epistemic uncertainty.

Remark. In fact, many systems are associated with more than two operators. By [1], we generally focus on two-operator systems for simplicity as it is trivially to extend PMM to those systems of more than two operators which are restricted to subsets of D.

Similar to probabilistic integration, **GP** is used in PMM due to enclosure under linear projections (Propostion 1.3.8), say under the linear operator A and B above. PMM starts by positing a **GP** prior for u. Without loss of generality, we assume the mean function of this **GP** prior to be zero, as in Section 2.2.2 we have seen that zero mean function can simplify computation and it can also be relaxed trivially, i.e., $u \sim \mathbf{GP}(0, k)$ where $k : D \times D \to \mathbb{R}$ is left to be discussed. Thus, by Definition 1.3.12 the finite dimensional marginals of this prior: $[u(\boldsymbol{x}_1), \ldots, u(\boldsymbol{x}_n)]$ are Gaussian distributed for any $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \subset D$. The mean vector of this marginals are zero and their covariance matrix are identified by k.

In order to construct the posterior measure, let prior **GP** measure conditioned on $m_A \in \mathbb{N}$ evaluations of the forcing function at locations $X_0^A = [\boldsymbol{x}_{0,1}^A, \ldots, \boldsymbol{x}_{0,m_A}^A] \subset D$, and $m_B \in \mathbb{N}$ evaluations of the boundary function at locations $X_0^B = [\boldsymbol{x}_{0,1}^B, \ldots, \boldsymbol{x}_{0,m_B}^B] \subset \partial D$. In [1] these points are referred to as the *design* points. Thus, the solution function u and the above evaluations are connected:

$$A(u)(X_0^A) = g(X_0^A) = \boldsymbol{g} \tag{2.10}$$

$$B(u)(X_0^B) = b(X_0^B) = \boldsymbol{b}. \tag{2.11}$$

By this interpolation equation, we can formally posed a Gaussian inference framework

[6] This condition is relaxed to semi-linear operator in Section 5.2. of [1].

for PMM. To do so we first establish some notation. Define:

$$L := \begin{bmatrix} A \\ B \end{bmatrix}, \quad L^\dagger := [A^\dagger \; B^\dagger], \quad \boldsymbol{u} = \begin{bmatrix} u(X_0^A) \\ u(X_0^B) \end{bmatrix}, \quad X_0 = \begin{bmatrix} X_0^A \\ X_0^B \end{bmatrix}, \quad \boldsymbol{h} = \begin{bmatrix} \boldsymbol{g} \\ \boldsymbol{b} \end{bmatrix}.$$

Now, (2.8) and (2.9) can be rewritten as:

$$L(\boldsymbol{u}) = \boldsymbol{h}. \tag{2.12}$$

Analogous to probabilistic integration, by Proposition 1.3.8 , at $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \subset D$ we have:

$$\begin{aligned} p(\boldsymbol{u}) &= N(\boldsymbol{u}; 0, k(X, X)) \\ p(\boldsymbol{h}) &= N(\boldsymbol{h}; 0, LL^\dagger k(X_0, X_0)), \quad \text{with} \\ p(\boldsymbol{u}, \boldsymbol{h}) &= N\left(\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{h} \end{bmatrix}; \boldsymbol{0}, \begin{bmatrix} k(X, X) & L^\dagger k(X, X_0) \\ Lk(X_0, X) & LL^\dagger k(X_0, X_0) \end{bmatrix}\right). \end{aligned}$$

where

$$LL^\dagger k(X_0, X_0) = \begin{bmatrix} AA^\dagger k(X_0^A, X_0^A) & AB^\dagger k(X_0^A, X_0^B) \\ A^\dagger B k(X_0^B, X_0^A) & BB^\dagger k(X_0^B, X_0^B) \end{bmatrix}.$$

By conditioning (1.19) in Gaussian algebra, we can obtain the posterior measure $p(\boldsymbol{u} \mid \boldsymbol{h})$ immediately. We present this posterior measure in accordance with [1].

Proposition 2.3.1 (Probabilistic Meshless Method; PMM)**.** *Let* $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \subset D$*. Denote* $\boldsymbol{u}$ *the* $n \times 1$ *vector* $u(X)$*. Then under* $p(\boldsymbol{u} \mid \boldsymbol{g}, \boldsymbol{b})$ *we have*

$$p(\boldsymbol{u} \mid \boldsymbol{g}, \boldsymbol{b}) = N(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where the posterior mean an covariance are

$$\boldsymbol{\mu} = L^\dagger k(X, X_0)[LL^\dagger k(X_0, X_0)]^{-1}[\boldsymbol{g}^\intercal \; \boldsymbol{b}^\intercal]^\intercal \tag{2.13}$$

$$\boldsymbol{\Sigma} = k(X, X) - L^\dagger k(X, X_0)[LL^\dagger k(X_0, X_0)]^{-1} Lk(X_0, X). \tag{2.14}$$

Written in following algorithm

Algorithm 2 Probabilistic Meshless Method

```
Construct prior: let $u \sim \mathbf{GP}(0, k)$
Evaluate forcing and boundary terms at $X_0 = [X_0^A \; X_0^B]^\intercal$:
    set $\boldsymbol{g} = g(X_0^A), \boldsymbol{b} = b(X_0^B)$
Infer posterior: $\boldsymbol{u} \mid \boldsymbol{g}, \boldsymbol{b} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Output MAP estimator: $\boldsymbol{\mu}$
```

This PNM consists of the kernel of prior distribution and the underlying differential operator. The output of PMM is the posterior distribution which is over the solution function of the corresponding PDE. The nature of this quantification is statistical and as such it differs from the classical numerical approaches which only return a point estimate of true solution. Instead, the output of PMM not only operates this but also quantify the uncertainty over true solution due to finite evaluation of forcing and boundary terms. Additionally, Section 4 of [1] also provides a theoretical analysis which proves the validity of this quantification. We briefly remark some results here.

Remark. [1] shows that minimising value of entries in $\mathbf{\Sigma}$ leads to accurate estimate $\boldsymbol{\mu}$. This characteristic is consistent with probabilistic integration, and it reassure us that the posterior $p(\boldsymbol{u} \mid \boldsymbol{g}, \boldsymbol{b})$ over the solution space is locally well-behaved. [1] also proves contraction rate of PMM. Readers are invited to view detailed proofs and we will not cover it in this work.

Remark. Many inverse problems can reduce to inferring unknown parameters in differential equations. [1] shows PMM is well suited to solving inverse problems by considering the random fields of solution. Moreover, the output of PMM is ready for subsequent inference of those unknown parameters. Numerical error can be propagated in this pipeline so that keeps inference from failure causing by accumulation of errors.

Again, we end this subsection by making the expression of PMM consistent with the Definition 2.1.2. Herein, our QoI is $u(x)$ and the state space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ is separable Banach space of appropriately differentiable real-valued functions w.r.t. its Borel σ-algebra. (2.12) is regarded as information and belief distribution ρ is a Gaussian measure(**GP**). Equivalently, the belief update operator of PMM is $B(\rho, a) = \rho^a$, where ρ^a is the restriction of ρ to those functions for which (2.12) is satisfied. Consequently, ρ^a is again a Gaussian distribution.

2.3.2 Prior selection

We also need to discuss the choice of prior **GP** for implementation of PMM. The covariance term $\mathbf{\Sigma}$ plays a key role for error quantification so that the kernel k of prior **GP** must be selected appropriately. For similar reasons we explained in Probabilistic Integration, we cannot choose kernels at random. By employing the Green's function, [1] provides a natural prior which we present below.

This natural prior construction begins with assigning a **GP** to the forcing term g which is define on a Hilbert space $H_A(D)$. Let $g \sim \mathbf{GP}(0, \Lambda)$. Recall Moore–Aronszajn theorem in Chapter 1, we denote the unique **RKHS** induced by kernel Λ: $\mathcal{H}_\Lambda(D)$. It will be assumed that $\mathcal{H}_\Lambda(D) \subseteq \mathcal{H}_A(D)$. By this construction,all functions $g : D \to \mathbb{R}$ satisfying finite norm $\|g\|_\Lambda := \|A_\Lambda g\|_2 < \infty$ are included in $\mathcal{H}_\Lambda(D)$.[7]

[7] Here $\|\cdot\|_2$ and $\|\cdot\|_\Lambda$ are the norms associated with L^2 space and $\mathcal{H}_\Lambda(D)$ respectively.(Definition 1.3.8)

Next, we will display that forcing term g propagate uncertainty to the solution u of underlying PDE (2.8) and (2.9). [1] defines a specific inner product space $(\mathcal{H}_{\text{nat}}(D), \langle\cdot,\cdot\rangle_{\text{nat}})$ by:

$$\mathcal{H}_{\text{nat}}(D) := \{v \in \mathcal{H}(D) \mid A_\Lambda A v \in L^2(D), Bv = 0 \text{ on } \partial D \text{ and } B_\Lambda A v = 0 \text{ on } \partial D\}$$
$$\langle u, v\rangle_{\text{nat}} := \int_D [A_\Lambda A u(\boldsymbol{x})][A_\Lambda A v(\boldsymbol{x})]\mathrm{d}\boldsymbol{x}.$$

Under this definition $\|u\|^2_{\text{nat}} := \langle u, u\rangle = \|g\|^2_\Lambda$. Suppose the space $\mathcal{H}_{\text{nat}}(D)$ is non-degenerate, i.e., $\|v\|_{\text{nat}} = 0$ iff $v = 0$. Additionally, suppose the PDE system $A(u)(x) = g(x)$ has a unique solution $u \in \mathcal{H}_{\text{nat}}(D)$ for any $g \in \mathcal{H}_\Lambda(D)$ and a Green's function G satisfying:

$$A(G(\boldsymbol{x}, \boldsymbol{x}')) = \delta(\boldsymbol{x} - \boldsymbol{x}') \qquad \boldsymbol{x} \in D \tag{2.15}$$
$$B(G(\boldsymbol{x}, \boldsymbol{x}')) = 0 \qquad \boldsymbol{x} \in \partial D. \tag{2.16}$$

On the basis of assumption above, [1] finally defines the *natural kernel* $k_{\text{nat}} : D \times D \to \mathbb{R}$ by

$$k_{\text{nat}} := \int_D \int_D G(\boldsymbol{x}, \boldsymbol{z})G(\boldsymbol{x}', \boldsymbol{z}')\Lambda(\boldsymbol{z}, \boldsymbol{z}')\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{z}'. \tag{2.17}$$

If this natural kernel in prior **GP** of solution u is tractable we can achieve ideal result in the sense that computation in PMM is in closed form.

We briefly illustrate a simple example given in [1] to built concreteness for PMM here. Consider an real-valued one dimensional Poisson equation, namely $A = -\nabla^2 = \frac{\mathrm{d}^2}{\mathrm{d}x^2}$ and B is identical mapping in (2.8) and (2.9):

$$-\nabla^2 u(x) = g(x) \quad x \in (0, 1)$$
$$u(x) = 0 \quad x \in \{0, 1\}$$

with its Green's function:

$$G(x, x') = \begin{cases} x(x' - 1), & x > x' \\ x'(x - 1), & x < x' \end{cases}$$

Following construction of natural kernel, we assign $g \sim \mathbf{GP}(0, \Lambda)$ where Λ is a polynomial kernel function, for instance $\Lambda(x, x') = \max\left(1 - \epsilon^{-1}|x - x'|, 0\right)^2$ [see [27] for details about this kernel] so that the natural kernel (2.17) will be analytic as it is essentially a integral of polynomial. Thus, we can calculate posterior measure by PMM (Proposition 2.3.1) and achieve a computable result.

Now, a viable algorithm for implementation of PMM is fully constructed. Meanwhile, several remarks will be made here.

Remark. It has been proved by [1] that $\mathcal{H}_{\mathrm{nat}}(D)$ is a **RKHS** with its reproducing kernel k_{nat} given $\sup_{\boldsymbol{x} \in D} k_{\mathrm{nat}}(\boldsymbol{x}, \boldsymbol{x}) < \infty$. Note that the linear operator A maps $\mathcal{H}_{\mathrm{nat}}(D)$ to $\mathcal{H}_{\Lambda}(D)$ and that means $u \sim \mathbf{GP}(0, k_{\mathrm{nat}}) \iff g \sim \mathbf{GP}(0, \Lambda)$. Hence,in application we can specify either one of $k_{\mathrm{nat}}, \Lambda$ as the other one is determined simultaneously due to linearity of A. Furthermore, the natural kernel k_{nat} also simplifies the evaluation as boundary condition is encoded in this kernel by (2.16). All of the above explains why k_{nat} is natural.

Remark. In practical application, the Green's functions for complex PDEs are almost unavailable. In addition, natural kernel also requires integral (2.17) has a closed form which means extra efforts for finding another paired kernel must be made. To improve this, [1] also poses a practical prior kernel which skips over finding Green's function of PDEs by considering a wider **RKHS** $\mathcal{H}_{\tilde{k}}$ which contains $\mathcal{H}_{\mathrm{nat}}(D)$. As this construction is largely based on the nature of the PDE itself, we will not go into detail here.

Chapter 3

PNM for FIE of the First Kind

In this chapter we will develop a PNM for FIE of the first kind. In Chapter 1 we have seen that type I FIE can be written in an operator equation form (1.1). We also prove that the operator K is linear in Example 1.3.1. Inspired by two probabilistic approaches illustrated above, we consider establishing a PNM for type I FIE as the linearity of operator K can be exploited in Gaussian inference. On the basis of PMM, we substitute the differential operator in PMM by operator K associated with corresponding FIE of the first kind and that provides the straightforward idea for solving the FIE probabilistically. In addition, we will explain the reasons for advocating PNM rather than classical numerrical methods. The inner relationship between this specific PNM and common regularisation method will also be illustrated to emphasize the viability of PNM. Finally, We will apply this PNM to some concrete numerical examples which will be presented in Chapter 4.

3.1 Derivation of the Probabilistic Solver

In this section a probabilistic solver for FIE of the first kind will be derived. Consider a generic **FIEFK** on a compact measurable set $D \subset \mathbb{R}^d$ with a compact operator $K : \mathcal{H}(D) \to \mathcal{H}_{K_h(D)}$:

$$K_h(f)(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in D \tag{3.1}$$

$$K_h := \int_E h(\boldsymbol{x}, \boldsymbol{t}) f(\boldsymbol{t}) \mathrm{d}\boldsymbol{t} \tag{3.2}$$

where E is a given integral domain and $h(\boldsymbol{x}, \boldsymbol{t})$ is a measurable integral kernel function on $D \times E$. For the remainder of this subsection it will be assumed that $h(\boldsymbol{x}, \boldsymbol{t})$ is a real-valued function:

$$\int_D \int_E |h(\boldsymbol{x}, \boldsymbol{t})|^2 \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{t} < \infty,$$

this is to correspond with most practical **FIEFK**-related problems.

In (3.1), our goal is to compute (or estimate) the solution $f(\boldsymbol{x}) \in \mathcal{H}(D)$ whose form is not analytical. The information we have so far to solve the problem is the known function g and Fredholm integral kernel h. Here, we are uncertain about f previous to evaluating g at some locations. As with the other PNMs, f is regarded as our epistemic uncertainty in FIE of the first kind.

Due to the linear nature of operator K_h, we utilise **GP** again in solving **FIEFK**. Let f follow a zero-mean Gaussian Process as we rarely have extra information for f. The covariance function of this **GP** is $k : D \times D \to \mathbb{R}$. These two elements characterise f, i.e., $f \sim \mathbf{GP}(0, k)$, and also characterise the finite marginals distribution of this Gaussian process by selecting specific index set. In other words, a joint Gaussian distribution which consists of evaluated function g at specific locations is also characterised by zero-mean and the covariance matrix identified by k.

Next, we interpolate (3.1) at $n \in \mathbb{N}$ locations $X_0 = [\boldsymbol{x}_{0,1}, \dots, \boldsymbol{x}_{0,n}] \subset D$ to link unknown function f with known function g:

$$K_h(f)(X_0) = g(X_0) = \boldsymbol{g}. \tag{3.3}$$

Since Gaussian distributions are closed under linear operator K_h (Proposition 1.3.8), $\boldsymbol{g}$ is also a joint Gaussian:

$$p(\boldsymbol{g}) = N(\boldsymbol{g}; 0, K_h K_h^\dagger k(X_0, X_0))$$

Similarly, at $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \subset D$ we have:

$$p(\boldsymbol{f}, \boldsymbol{g}) = N\left(\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{g} \end{bmatrix}; \mathbf{0}, \begin{bmatrix} k(X, X) & K_h^\dagger k(X, X_0) \\ K_h k(X_0, X) & K_h K_h^\dagger k(X_0, X_0) \end{bmatrix}\right).$$

Using conditioning rule (Proposition 1.19), we instantly have the posterior which is again a Gaussian, given in Proposition below.

Proposition 3.1.1 (Probabilistic solver for **FIEFK**). *Let $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \subset D$. Denote $\boldsymbol{f}$ the $n \times 1$ vector $f(X)$. We have*

$$p(\boldsymbol{f} \mid \boldsymbol{g}) = N(\boldsymbol{f}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\mu} = K_h^\dagger k(X, X_0)[K_h K_h^\dagger k(X_0, X_0)]^{-1} \boldsymbol{g} \tag{3.4}$$

$$\boldsymbol{\Sigma} = k(X, X) - K_h^\dagger k(X, X_0)[K_h K_h^\dagger k(X_0, X_0)]^{-1} K_h k(X_0, X). \tag{3.5}$$

These calculations yield the algorithm of probabilistic solver for FIE of the first kind as follows:

Algorithm 3 Probabilistic solver for **FIEFK**

```
Construct prior: let $u \sim \mathbf{GP}(0, k)$.
Evaluate $g$ at $X_0 = [\boldsymbol{x}_{0,1}, \dots, \boldsymbol{x}_{0,n}]^\intercal \subset D$:
    set $\boldsymbol{g} = g(X_0)$.
Infer posterior: $\boldsymbol{f} \mid \boldsymbol{g} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Output MAP estimator: $\boldsymbol{\mu}$
```

This algorithm thus attains a posterior probability $p(\boldsymbol{f} \mid \boldsymbol{g})$ over the unknown function f. Its mean $\boldsymbol{\mu}$ works as the MAP estimate of true solution f. The covariance here represents the squared error of estimation which enables formal quantification of numerical error.

This solver is also a PNM complying with Definition 2.1.2. We may notice that this probabilistic solver is quite similar to PMM we introduced in last chapter. The state space $(\mathcal{H}(D), \Sigma_{\mathcal{H}(D)})$ is a Hilbert space[1] of integrable functions. The QoI, information, belief distribution and belief update operator of this solver is essentially identical as PMM. The main difference between these two PNM is the dissimilar operators which operates on the respective solution function. That leads to the prior selection of our solver different from PMM.

3.2 Prior Selection

Similar to other PNMs, a viable kernel needs to be proposed for the prior **GP** in our solver. Consider the posterior Gaussian measure $p(\boldsymbol{f} \mid \boldsymbol{g})$ characterised by its mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A viable probabilistic solver requires that the expressions of both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are tractable. More specifically, the tractability of these two expression depends on whether the integrals below have closed forms:

$$K_h k(x, y) := \int_E h(x, z) k(z, y) \mathrm{d}z \tag{3.6}$$

$$K_h^\dagger k(x, y) := \int_E k(x, z') h(z', y) \mathrm{d}z' \tag{3.7}$$

$$K_h K_h^\dagger k(x, y) := \int_E \int_E h(x, z) k(z, z') h(z', y) \mathrm{d}z \mathrm{d}z'. \tag{3.8}$$

More concisely, it can be seen that the tractablity of (3.8) guarantees the tractablity of the other two expressions. Thus, in order to make (3.8) tractable, it is required that the kernel k of prior **GP** must be chosen carefully. However, the Fredholm integral kernels are diverse. The smoother the Fredholm integral kernel is, the more unstable it is. Compared with the rule-based kernel, say natural kernel for PMM (in Section 2.3.2), it is very hard to construct a general covariance function k for this integral (3.8) unless

[1] Most commonly L^2 space.

we strengthen some conditions on Fredholm integral kernel.

Here we consider a simplest case when the Fredholm integral kernel h is degenerate:

Definition 3.2.1. The integral kernel $h(x, t)$ in (1.1) is called degenerate kernel if it can be rewritten as:

$$h(x,t) = \sum_{i=1}^{n} a_i(x) b_i(t).$$

Under this condition we simplify (3.8):

$$\begin{aligned} K_h K_h^\dagger k(x,y) &= \int_E \int_E \sum_i a_i(x) b_i(z) k(z,z') \sum_j a_j(z') b_j(y) \mathrm{d}z \mathrm{d}z' \\ &= \sum_i \sum_j a_i(x) b_j(y) \int_E \int_E a_j(z') b_i(z) k(z,z') \mathrm{d}z \mathrm{d}z'. \end{aligned}$$

We propose that a class of prior covariance functions that work well are polynomial kernels as they can reduce most of integrand to polynomials which possess a tractable form. Specifically, if functions $\{a_i(x)\}_{i=1}^n$ and $\{b_i(x)\}_{i=1}^n$ are polynomial functions as well, (3.8) is available in closed form. Nonetheless, this condition is still too strong and does not make it widely applicable. Hence, a more practical way is to select prior covariance function based on the expression of Fredholm integral kernel. We will show some examples as illustrations in next chapter.

3.3 Links to Common Numerical Methods

In this section, we will draw some connections between classic and probabilistic numerical methods. We will first introduce the advantages of using probabilistic solver. A connection between regularisation will also be establish here to show that our probabilistic solver for FIE of the first kind is effective.

3.3.1 Why PNM

The most immediate reason we may think for using PNM in solving FIE of the first kind is that it avoids computing the inverse of the operator K as calculating K^{-1} could be a great challenge for only given discrete observed data. In fact, this is where the ill-posedness of FIE of the first kind lies. Aside from this, there are actually several profound reasons as follows.

By now, the classical numerical methods have been well developed. Thus, one may question whether PNM is a better numerical method or there is any advantages of using PNM. [2] gives an affirmative answer. Classical numerical methods can also be regarded as inference in the sense that they return an approximation of latent quantities[2] based on some known conditions or data. However, this connection between inference and computation is vague.

In contrast, PNM provides a deterministic rules for inference to latent quantities. For instance, probabilistic integration, PMM and our probabilistic solver use respective priors and posterior MAP estimates instead of lax approximations in classical methods. Furthermore, PNMs not only returns the point estimate same as classical methods but also provide new measures of uncertainty (or numerical errors) captured by the rest of outcome probability distribution. [2] also establishes a link between PNMs and uncertainty quantification. As Bayesian theorem allows hierarchical inference, PNM allows these numerical errors to propagate through chains of computation, which contributes significantly to uncertainty quantification. Further details are discussed in [2].

3.3.2 Relationship with Regularisation

A. N. Tikhonov (see [8]) has proved that regularisation method is effective in solving the inverse/ill-posed problems. In this subsection we will show the connection between our probabilistic solver and regularisation method. This latent connection adds some credibility to the idea of using PNM to solve FIE of the fist kind. **FIEFK** usually appears as a typical inverse problem in many literature (see [28], [29]). Following [30], a inverse problem is defined as:

Definition 3.3.1 (Inverse Problem)**.** A problem finds $m \in \mathbb{R}^p$ from $d \in \mathbb{R}^q$ where e m and d are related by the equation

$$d = G(m) + \eta \tag{3.9}$$

is called **inverse problem**, where d and m are referred as observed data and the unknown respectively. η represents the observational noise which enters the observed data.

Consider the FIE of the first kind (3.1). It fits almost exactly with the above formula (3.9) except the noise part as it is essentially a degenerate inverse problem[3]. To accommodate this setting, let noise term $\sigma \mapsto 0$, we add the noise term in (3.1) corresponding to the practical noise contaminate:

$$\eta + K_h(f)(\boldsymbol{x}) = g(\boldsymbol{x}) + \eta = d(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in D \subset \mathbb{R}^d \tag{3.10}$$

where $\eta \sim N(0, \sigma^2)$. Recall our probabilistic solver takes epistemic uncertainty $f(x)$ as a Gaussian process with a zero mean $N(0, k(\boldsymbol{x}, \boldsymbol{x}))$. After gaining observed data at locations X_0, we have $f(X_0) \sim N(0, k(X_0, X_0))$ and $g(X_0) \sim N(0, K_h K_h^{\dagger} k(X_0, X_0))$ due

[2] This notion corresponds to epistemic uncertainty in the framework of PNM.

[3] Namely, a noise-free inverse problem.

to linearity of operator K_h. It is also assumed that noise term $\boldsymbol{\eta} = [\eta_1, \dots, \eta_d]$ follows independent and identical Gaussian distribution, i.e., $\boldsymbol{\eta} \sim N(0, \sigma^2 I_{d\times d})$.

Next, we define the prior distribution of the form:

$$\Pi_{prior}(f(X_0)) = \Pi_{prior}(\boldsymbol{f}) = \exp\left(-\frac{1}{2}\boldsymbol{f}^\intercal [k(X_0, X_0)]^{-1}\boldsymbol{f}\right).$$

By assuming noise covariance is known, we define likelihood density as follows:

$$\Pi_{likeli}(\boldsymbol{d} \mid \boldsymbol{f}) = \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{d} - K_h\boldsymbol{f}\|^2\right).$$

By Bayes theorem, we can define our a posterior density by the product of the likelihood and the a prior densities. The a posterior density defines as:

$$\begin{aligned}\Pi_{post}(\boldsymbol{f} \mid \boldsymbol{d}) &\propto \Pi_{likeli}(\boldsymbol{d} \mid \boldsymbol{f})\Pi_{prior}(\boldsymbol{f}) \\ &= \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{d} - K_h\boldsymbol{f}\|^2 - \frac{1}{2}\boldsymbol{f}^\intercal [k(X_0, X_0)]^{-1}\boldsymbol{f}\right) \\ &= \exp(-V(\boldsymbol{f} \mid \boldsymbol{d})),\end{aligned}$$

where

$$V(\boldsymbol{f} \mid \boldsymbol{d}) = \frac{1}{2\sigma^2}\|\boldsymbol{d} - K_h\boldsymbol{f}\|^2 + \frac{1}{2}\boldsymbol{f}^\intercal [k(X_0, X_0)]^{-1}\boldsymbol{f} \tag{3.11}$$

Note that matrix $h(X_0, X_0)$ is symmetric positive definite, so is its inverse. Thus, by Cholesky factorisation we have:

$$[k(X_0, X_0)]^{-1} = C^\intercal C$$

where C is a triangular matrix.

Now, we can finally establish a regularisation formulation as:

$$\begin{aligned}T(\boldsymbol{f}) = 2\sigma^2 V(\boldsymbol{f} \mid \boldsymbol{d}) &= \|\boldsymbol{d} - K_h\boldsymbol{f}\|^2 + \sigma^2 \boldsymbol{f}^\intercal C^\intercal C\boldsymbol{f} \\ &= \|\boldsymbol{d} - K_h\boldsymbol{f}\|^2 + \sigma^2\|C\boldsymbol{f}\|^2.\end{aligned} \tag{3.12}$$

Moreover, if we use a scaled covariance family, separating a simple scale $\theta \in \mathbb{R}$ from a unit kernel $\tilde{k}(x, x')$,

$$k(x, x') := \theta^2 \tilde{k}(x, x'),$$

(3.12) will have the following form

$$T(\boldsymbol{f}) = \|\boldsymbol{d} - K_h\boldsymbol{f}\|^2 + \lambda^2\|C\boldsymbol{f}\|^2, \quad \lambda = \frac{\sigma}{\theta} \tag{3.13}$$

where λ is the ratio of the noise and the a priori variances. It also plays the role of balancing the information contributions between the likelihood and the a priori densities. In [31], such a functional, or more generally, an operator T, is called Tikhonov regularisation operator.

Now we can draw a conclusion that using probabilistic solver in **FIEFK** is essentially utilising a Gaussian prior to *regularise* the ill-posed **FIEFK**. That also explain why PNM works well in solving **FIEFK**. Before presenting the numerical examples demonstrating the probabilistic solver in the next chapter a remark will be made:

Remark. Recall that in our probabilistic solver we use the MAP estimate from posterior density as an approximation to the true solution. In fact, the problems of estimating the MAP can be simplified into minimising a misfit function of the negative logarithm of the a posterior density. By [31], the negative logarithm of a posterior density resembles the general Tikhonov regularisation misfit function (3.13). The MAP estimator defines as:

$$\begin{aligned}
\pi_{MAP} &= \arg\max_{\boldsymbol{f}} \Pi_{post}(\boldsymbol{f} \mid \boldsymbol{d}) \\
&= \arg\min_{\boldsymbol{f}} - \log \Pi_{post}(\boldsymbol{f} \mid \boldsymbol{d}) \\
&= \arg\min_{\boldsymbol{f}} V(\boldsymbol{f} \mid \boldsymbol{d}) \\
&= \arg\min_{\boldsymbol{f}} T(\boldsymbol{f}).
\end{aligned}$$

This proves once again that PNM can be used to regularise ill-posed/inverse numerical problems.

Chapter 4

Numerical Examples

In this chapter we will present our numerical experiments to validate the probabilistic solver we have just derived.

4.1 Example 1

Consider the following Fredholm integral equation of the first kind

$$\int_0^1 e^x t f(t) \mathrm{d}t = \frac{e^x}{4}, \tag{4.1}$$

with exact solution $f(t) = t^2$. In (4.1), the Fredholm integral kernel is degenerate: $h(x,t) = e^x \cdot t$.

Let $f \sim \mathbf{GP}(0,k)$. Here we choose the covariance function $h(x,y)$ to be a quadratic homogeneous polynomial kernel:

$$k(x,y) = x^2 y^2.$$

According to formulation in Proposition 3.1.1, we need to calculate integral in the form of (3.8). Specifically,

$$\begin{aligned} K_h K_h^\dagger k(x,y) &= \int_0^1 \int_0^1 e^x z (z \cdot z')^2 e^{z'} y \mathrm{d}z \mathrm{d}z' \\ &= e^x y \cdot \frac{1}{4}(e-2). \end{aligned}$$

This is in a closed form. Similarly, we have the other two tractable integrals:

$$\begin{aligned} K_h k(x,y) &= \int_0^1 e^x z \cdot z^2 y^2 \mathrm{d}z = \frac{1}{4} e^x y \\ K_h^\dagger k(x,y) &= \int_0^1 x^2 z^2 e^z y \mathrm{d}z = x^2 y(e-2). \end{aligned}$$

Thus, for any fixed x_0, by Proposition 3.1.1, we have:

$$p(f \mid g(x_0)) = N(f; \mu, \Sigma)$$

where

$$\begin{aligned}
\mu &= K_h^\dagger k(x, x_0)[K_h K_h^\dagger k(x_0, x_0)]^{-1} g(x_0) = x^2 \\
\Sigma &= k(x, x) - K_h^\dagger k(x, x_0)[K_h K_h^\dagger k(x_0, x_0)]^{-1} K_h k(x_0, x) = 0.
\end{aligned}$$

Since we take μ as our MAP estimate of f, we see that $\mu = f(x)$ is exactly the true solution with zero squared error $\Sigma = 0$.

4.2 Example 2

Consider the following Fredholm integral equation of the first kind

$$\int_0^\pi \cos(x - t) f(t) \mathrm{d}t = \frac{\pi}{2} \cos x, \tag{4.2}$$

with exact solution $f(t) = \cos t$. In (4.2), the Fredholm integral kernel is degenerate: $h(x, t) = \cos(x - t) = \cos x \cos t + \sin x \sin t$.

Let $f \sim \mathbf{GP}(0, k)$. As the integral in the form of (3.8) need to be tractable, here we consider a symmetric positive definite kernel in a trigonometric form:

$$k(x, y) = \phi(x)^\intercal \phi(y) = \cos x \cos y + \sin x \sin y, \quad \phi(x) = [cosx \ sinx]^\intercal.$$

Specifically, we have

$$\begin{aligned}
K_h K_h^\dagger k(x, y) &= \int_0^\pi \int_0^\pi \cos(x - z) \cos(z - z') \cos(z' - y) \mathrm{d}z \mathrm{d}z' \\
&= \frac{\pi^2}{4} \cos(x - y).
\end{aligned}$$

This is in a closed form. Thus, in accordance with Proposition 3.1.1, the exact solution can be achieved by taking two observations $X_0 = [x_{0,1} \ x_{0,2}]$:

$$p(f \mid g(X_0)) = N(f; \mu, \Sigma)$$

where

$$\begin{aligned}
\mu &= K_h^\dagger k(x, X_0)[K_h K_h^\dagger k(X_0, X_0)]^{-1} g(X_0) = \cos x \\
\Sigma &= k(x, x) - K_h^\dagger k(x, X_0)[K_h K_h^\dagger k(X_0, X_0)]^{-1} K_h k(X_0, x) = 0.
\end{aligned}$$

4.3 Some Comments

Having displayed two examples, we now comment on several deficiencies of these examples. Notice that the posterior covariances Σ of above two examples are always zero which means the output of our probabilistic solver is the true solution exactly. This is because we have enough observed data to evaluate **FIEFK** under a one-dimensional **RKHS**. In practical application, we usually do not have sufficient information like observed data for the underlying **FIEFK**. Under this circumstance, the posterior covariance will not be zero but a function of locations we evaluated. Additionally, the case of high- or infinite-dimensional **RKHS** is not investigate here. These deficiencies should be addressed in future work as a improvement to illustration for our probabilistic solver.

Chapter 5

Discussion

5.1 Summary

In the preceding chapters the use of probabilistic numerical method as a statistical tool providing inference for the numerical solution of Fredholm Integral Equation of the First Kind was investigated. In Chapter 1 the focus was on introducing the relevant background theory needed for one to have a good grasp on PNM, in particular the theory of Reproducing Kernel Hilbert Spaces was discussed with a concentration to elicit the notion of kernel which plays a important role in another necessary ingredient, Gaussian Process. Particular attention was also paid to operator theory that guarantees the successful operation of the PNM, say linearity.

In Chapter 2 we introduced the so-called probabilistic numerical method by posing the formal definition of PNM. Two cases were presented in order to build concreteness. Intrinsically, the PNM works on the basis of a Gaussian closure under a linear projection. This chapter focused especially on explaining how PNM works in some common numerical tasks. Additionally, a necessary prior selection was discussed in illustrating the two cases.

In Chapter 3 a detailed derivation of the PNM for FIE of the first kind was presented, including the prior selection part. The advantages of using PNM instead of classical methods were also explained. By establishing the link between PNM and regularisation method, the effectiveness of PNM for such ill-posed/inverse problems has also been demonstrated.

In Chapter 4 numerical examples were presented to demonstrate the probabilistic solver derived in Chapter 3. Both of two examples showed that PNM works well in solving FIE of the first kind. Some comments were made on shortcomings of examples.

5.2 Issue Faced

In this section we will comment on the main issue which were faced while working on this project.

Looking back on Section 3.2, the biggest issue we faced was finding appropriate covariance function for the prior measure. So far, we have been unable to construct a covariance function with some generality to make integral (3.8) tractable. In most cases we may need to find these functions based on integration rules. It leads to no general applicability of this method unless a kernel can be found which is able to pair with a class of Fredholm integral kernels. In fact, finding such pairs satisfying (3.8) are hard as integrals are complicated searches that do not have some element rules in derivatives such as chain ruls and product rules. Further work will aim to establish an appropriate and general prior covariance function.

5.3 Potential Directions Forward

In this section, we shall briefly comment on potential areas that could be further investigated for this project. In particular, 2 directions appear to be the most promising: uncertainty quantification based on posterior covariance and the performance of PNM in high-dimensional state space. These shall be briefly discussed in the following two subsections.

5.3.1 Uncertainty Quantification

As we know, variance is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value. In our settings of numerical problems, this notion fits well as squared error of the values around the true solution. This contributes a lot to uncertainty quantification. It would be interesting to investigate a design rule of our probabilistic solver which minimises the variance in posterior measure. Further work will aim to address this issue.

5.3.2 Curse of dimensionality

The curse of dimensionality caused by the volume of the space increases so fast that the available data become sparse in high-dimensional space. The performance of Gaussian process regression is not ideal in this setting. In order to calibrate this, the kernel of prior measure might need to be well designed.

5.3.3 Other potential directions

Further numerical problems in the field of linear algebra and optimisation, etc., may be able to use PNMs to get a better solution.

Bibliography

[1] Jon Cockayne, Chris Oates, TJ Sullivan, and Mark Girolami. Probabilistic meshless methods for partial differential equations and bayesian inverse problems. 2016.

[2] Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015.

[3] Philipp Hennig, Michael A. Osborne, and Hans P. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

[4] Anatolii Borisovich Bakushinskii. A numerical method for solving fredholm integral equations of the 1st kind. *USSR Computational mathematics and mathematical physics*, 5(4):226–233, 1965.

[5] Aleksandr Mikhailovich Denisov. Approximation of quasi-solutions of fredholm's equation of the first kind with a kernel of special form. *USSR Computational Mathematics and Mathematical Physics*, 11(5):269–276, 1971.

[6] Aleksandr Mikhailovich Denisov. Approximation of the quasi-solutions of a fredholm integral equation of the first kind of a special form. *USSR Computational Mathematics and Mathematical Physics*, 12(6):244–248, 1972.

[7] AM Denisov. On the order of approximation when solving a fredholm equation of the first kind with a kernel of special type. *USSR Computational Mathematics and Mathematical Physics*, 13(1):255–260, 1973.

[8] Andreĭ Nikolaevich Tikhonov, Vasilij Ja Arsenin, Vasiliĭ IAkovlevich Arsenin, Vasiliy Y Arsenin, et al. *Solutions of ill-posed problems*. Vh Winston, 1977.

[9] Vitalii Pavlovich Tanana, E Yu Vishnyakov, and Anna Ivanovna Sidikova. An approximate solution of a fredholm integral equation of the first kind by the residual method. *Numerical Analysis and Applications*, 9(1):74–81, 2016.

[10] J Caldwell. Numerical study of fredholm integral equations. *International Journal of Mathematical Education in Science and Technology*, 25(6):831–836, 1994.

[11] Di Yuan and Xinming Zhang. An overview of numerical methods for the first kind fredholm integral equation. *SN Applied Sciences*, 1(10):1–12, 2019.

[12] S Yousefi and A Banifatemi. Numerical solution of fredholm integral equations by using cas wavelets. *Applied mathematics and computation*, 183(1):458–463, 2006.

[13] Khosrow Maleknejad and Saeed Sohrabi. Numerical solution of fredholm integral equations of the first kind by using legendre wavelets. *Applied Mathematics and Computation*, 186(1):836–843, 2007.

[14] Jianping Zhang, Huili Han, and X Pan. Wavelet-regularization method and extrapolation for solving fredholm integral equations of the first kind [j]. *Science technology and engineering*, 10(1):17–19, 2010.

[15] Zhongying Chen, Yuesheng Xu, and Hongqi Yang. A multilevel augmentation method for solving ill-posed operator equations. *Inverse Problems*, 22(1):155, 2006.

[16] Zhongying Chen, Yuesheng Xu, and Hongqi Yang. Fast collocation methods for solving ill-posed integral equations of the first kind. *Inverse Problems*, 24(6):065007, 2008.

[17] Xingjun Luo, Wenyu Hu, Lingjuan Xiong, and Fanchun Li. Multilevel jacobi and gauss–seidel type iteration methods for solving ill-posed integral equations. *Journal of Inverse and Ill-posed Problems*, 23(5):477–490, 2015.

[18] D. Sejdinovic and Arthur Gretton. What is an rkhs? 2012.

[19] Elias M Stein and Rami Shakarchi. *Functional analysis: introduction to further topics in analysis*, volume 4. Princeton University Press, 2011.

[20] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[21] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[22] Robert J Adler. An introduction to continuity, extrema, and related topics for general gaussian processes. IMS, 1990.

[23] Jon Cockayne, Chris J Oates, Timothy John Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM review*, 61(4):756–789, 2019.

[24] François-Xavier Briol, Chris J Oates, Mark Girolami, Michael A Osborne, and Dino Sejdinovic. Probabilistic integration: a role in statistical computation? *Statistical Science*, 34(1):1–22, 2019.

[25] FM Larkin. Gaussian measure in hilbert space and applications in numerical analysis. *The Rocky Mountain Journal of Mathematics*, pages 379–421, 1972.

[26] Xiaoyue Xi, François-Xavier Briol, and Mark Girolami. Bayesian quadrature for multiple related integrals. In *International Conference on Machine Learning*, pages 5373–5382. PMLR, 2018.

[27] Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.

[28] Qichao Que and Mikhail Belkin. Inverse density as an inverse problem: The fredholm equation approach. *Advances in neural information processing systems*, 26, 2013.

[29] Takehiko Ogawa, Yukio Kosugi, and Hajime Kanada. Neural network based solution to inverse problems. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, volume 3, pages 2471–2476. IEEE, 1998.

[30] Masoumeh Dashti and Andrew M Stuart. The bayesian approach to inverse problems. In *Handbook of uncertainty quantification*, pages 311–428. Springer, 2017.

[31] Muhammad Izzatullah, Daniel Peter, Sergey Kabanikhin, and Maxim Shishlenin. Bayes meets tikhonov: understanding uncertainty within gaussian framework for seismic inversion. In *Advanced Methods for Processing and Visualizing the Renewable Energy*, pages 121–145. Springer, 2021.