An RKHS Approach to Estimation with Gaussian Random Field

by

Shuangming Yang

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

One popular approach to estimating an unknown function from noisy data is the use of a regularized optimization over a reproducing kernel Hilbert space (RKHS). The solution belongs to a finite-dimensional function space. If we assume the additive measurement noise is Gaussian, then there is a well known statistical interpretation that the RKHS estimate represents the posterior mean (minimum variance estimate) of a Gaussian random field with covariance proportional to the kernel associated with the RKHS. In this thesis, we calculate the sharp upper bound of the error of the RKHS estimate (given unit RKHS norm of the underlying function). We also present a statistical interpretation for general loss functions, by assuming the density of prior is in exponential form in terms of RKHS norm and then give some simulation examples.

To my parents

Acknowledgements

I would like to express my appreciation to all those who helped me to complete this thesis. Most of all, I sincerely thank my supervisor Dr.Yaozhong Hu, whose great support, stimulating suggestions and encouragement helped me throughout the time of writing this thesis. This work would have been impossible without his guidance and help.

Deepest gratitude is also due to Dr.Christoph Frei and Dr.Bei Jiang for teaching me throughout my years of study and accepting to be members of my supervisory committee.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement through the process of researching and writing this thesis. Thank you.

Contents

List of Figures

Chapter 1

Introduction

In machine learning, two non-parametric approaches based on positive definite kernels have been widely used for the purpose of modeling non-linear functional relationships. On the one side, there is Bayesian machine learning with Gaussian processes (GP) which defines a hypothesis space through a GP prior distribution of the true function's realizations and then produces a posterior distribution for an unknown function of interest. On the other hand, frequentists assume the true function can be well approximated by functions in a reproducing kernel Hilbert space (RKHS) and then search for the function via regularized optimization. These two approaches have been shown to be practically powerful, and have found a wide range of practical applications in dealing with nonlinear phenomena. It is widely known in machine learning that these two approaches are closely related; for instance, the estimator of kernel ridge regression is identical to the posterior mean of Gaussian process regression In this thesis we review the existing literature and show the connection between them. In the present section, we review some basic concepts and known connections between the Bayesian and frequentist approaches in terms of reconstructing a function $f : \mathcal{X} \to \mathbb{R}$ from noisy data.

1.1 Basic definitions

In this chapter we review basic concepts and models which are related to Gaussian process prediction. We cover the well known *reproducing kernel*

Hilbert space (RKHS), which defines a class of sufficiently smooth functions with certain positive definite kernel k.

Definition 1.1.1. *Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be an **inner product** on $\mathcal{H}$ if:*

1. $\langle \lambda f, g\rangle_{\mathcal{H}} = \lambda\langle f, g\rangle_{\mathcal{H}}$
2. $\langle f_1 + f_2, g\rangle_{\mathcal{H}} = \langle f_1, g\rangle_{\mathcal{H}} + \langle f_2, g\rangle_{\mathcal{H}}$
3. $\langle f, g\rangle_{\mathcal{H}} = \langle g, f\rangle_{\mathcal{H}}$ [1]
4. $\langle f, f\rangle_{\mathcal{H}} \geq 0$ *and* $\langle f, f\rangle_{\mathcal{H}} = 0$ *if and only if* $f = 0$

With this inner product, the *norm* will be: $||f||_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$.

Definition 1.1.2. *A **inner product space** is a vector space $\mathcal{H}$ equipped with its inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.*

Definition 1.1.3. *An inner product space over the field of complex numbers are called a **unitary space**.*

Definition 1.1.4. *A sequence $\{a_n\}_{n=1}^{\infty}$ is said to be a **Cauchy sequence** if for every $\epsilon > 0$, there exists $N = N(\epsilon) \in \mathbb{N}$, such that $\forall n, m \geq N$, $||a_n - a_m||_{\mathcal{H}} < \epsilon$.*

Definition 1.1.5. *A space C is **complete** if every Cauchy sequence in C converges: it has a limit, and this limit is in C.*

Definition 1.1.6. *Given a set $\mathcal{X}$ and a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we say that the pair $\mathcal{M} = (\mathcal{X}, d)$ is a **metric space** if and only if $d(\cdot)$ satisfies the following properties:*

- *(Non-negativeness) For all $x, y \in \mathcal{X}$, $d(x, y) \geq 0$*
- *(Identification) For all $x, y \in \mathcal{X}$ we have that $d(x, y) = 0 \iff x = y$*
- *(Symmetry) For all $x, y \in \mathcal{X}$, $d(x, y) = d(y, x)$*

[1] Throughout we discuss in real-valued world only; otherwise the right hand side becomes its conjugate symmetry: $\langle f, g\rangle_{\mathcal{H}} = \overline{\langle g, f\rangle_{\mathcal{H}}}$

- *(Triangular inequality) For all $x, y, z \in \mathcal{X}$ we have that $d(x, z) \leq d(x, y) + d(y, z)$*

Definition 1.1.7. *A* ***Banach space*** *is a complete normed space, i.e., it contains the limits of all its Cauchy sequences.*

Definition 1.1.8. *A* ***Hilbert space*** *$\mathcal{H}$ is a complete metric space on which an inner product is defined and every Cauchy sequence converges to a limit in $\mathcal{H}$.*

Definition 1.1.9. ***Positive definite kernel*** *Let $\mathcal{X}$ be a non-empty set. A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite (pd) kernel, if for any $n \in \mathbb{N}, (c_1, ..., c_n) \in \mathbb{R}^n$ and $(x_1, ..., x_n) \in \mathcal{X}^n$,*

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

Remark 1. *Equivalently, it can be described by the following: a symmetric function k is positive definite if the matrix $\bar{K} \in \mathbb{R}^{n \times n}$ with elements $[\bar{K}]_{ij} = k(x_i, x_j)$ is positive semidefinite for any $(x_1, ..., x_n) \in \mathcal{X}^n$ of any size $n \in \mathbb{N}$. The matrix $\bar{K}$ is called the kernel matrix or Gram matrix.*

Proposition 1.1.1. *Let $\mathcal{X}$ be a non-empty set. If there exists an $\mathbb{R}$-Hilbert space and a map[2] $\phi : \mathcal{X} \to \mathcal{H}$, then the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $k(x_1, x_2) := \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}, \forall x_1, x_2 \in \mathcal{X}$ is a positive definite kernel.*

Proof. Let $(x_1, ..., x_n) \in \mathcal{X}^n$ and $(c_1, ..., c_n) \in \mathbb{R}^n$.

$$\begin{aligned}\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i,j=1}^{n} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \\ &= \langle \sum_{i=1}^{n} c_i \phi(x_i), \sum_{j=1}^{n} c_j \phi(x_j) \rangle_{\mathcal{H}} \qquad (1.1) \\ &= || \sum_{i=1}^{n} c_i \phi(x_i) ||_{\mathcal{H}}^2 \geq 0\end{aligned}$$

□

[2] ϕ is usually called the feature map.

Remark 2. *In fact, the reverse direction also holds: a positive definite function is guaranteed to be the inner product in a Hilbert space $\mathcal{H}$ between features $\phi(x)$ [8].*

Remark 3. *Positive definite kernels may be viewed as generalized inner products. Indeed, any inner product is naturally a pd kernel though the property of linearity does not carry over from inner products to pd kernels in general. However, another property of inner product, the Cauchy-Schwarz inequality, does have a natural extension as follows.*

Proposition 1.1.2. *If k is a positive definite kernel, and $x_1, x_2 \in \mathcal{X}$, then $k(x_1, x_2)^2 \leq k(x_1, x_1)k(x_2, x_2)$.*

Proof. The 2×2 Gram matrix with entries $\bar{K}_{ij} = k(x_i, x_j)$ is positive definite. Hence both its eigenvalues are non-negative, and so is their product, $\bar{K}$'s determinant, i.e.,

$$0 \leq \bar{K}_{11}\bar{K}_{22} - \bar{K}_{1}2\bar{K}_{21} = \bar{K}_{11}\bar{K}_{22} - \bar{K}_{12}^2$$

Substituting $k(x_i, x_j)$ for $\bar{K}_{ij}$, we get the desired inequality. □

Informally, a kernel function computes the inner product of the images under an feature map ϕ of two data points. Note that we imposed almost no conditions on $\mathcal{X}$: the inner product is defined on the features of elements of $\mathcal{X}$ rather than on $\mathcal{X}$ itself. Also, if $k(\cdot, \cdot)$ is positive definite, one can always define a family $X(t)$ (on certain index set $\mathcal{X}$) of zero-mean Gaussian random variables with covariance function k, that is, $\mathbb{E}[X(s)X(t)] = k(s, t)$, $s, t \in \mathcal{X}$. The well-definedness of classes of random variables in the continuous case is supported by the *Kolmogorov consistency theorem.*

Two popular classes of kernel functions are the Mat'ern kernels and spline kernels [9, p.6].

Definition 1.1.10. ***Mat'ern kernels*** *Let $\mathcal{X} \subset \mathbb{R}^d$. For fixed $\alpha > 0$ and $h > 0$, the Matern kernel $k_{\alpha,h} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given by:*

$$k_{\alpha,h}(x_1, x_2) = \frac{1}{2^{\alpha-1}\Gamma(\alpha)}(\frac{\sqrt{2\alpha}||x_1 - x_2||^{\alpha}}{h})K_{\alpha}(\frac{\sqrt{2\alpha}||x_1 - x_2||^{\alpha}}{h})$$

where Γ is the gamma function, and K_α is the Bessel function of the second kind of order α.

If $\alpha = 1/2$, it becomes exponential kernel:

$$k_{1/2,h}(x_1, x_2) = \exp\,(-\frac{||x_1 - x_2||}{h})$$

The limiting case where $\alpha \to \infty$ turns out to be the square-exponential kernel:

$$k_h(x_1, x_2) = \exp(-\frac{||x_1 - x_2||^2}{2h^2})$$

Definition 1.1.11. ***Spline kernels*** *Let $\mathcal{X} \subset \mathbb{R}$. For fixed $m \geq 1$ and $m \in \mathbb{N}$, the Spline kernel $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given by:*

$$k_m(x_1, x_2) = \int_0^1 \frac{((x_1 - u)_+(x_2 - u)_+)^{m-1}}{((m-1)!)^2} du$$

where $(x)_+ = x$ when $x \geq 0$ and zero elsewhere.

That Spline kernel is often used to reconstruct a real-valued function on the unit interval $[0, 1]$ with $(m-1)$th continuous derivatives and square integrable $f^{(m)}$, satisfying the boundary condition $f^{(k)}(0) = 0$ for $k = 0, ..., m-1$. In particular, in the context of cubic spline where $m = 2$, the kernel is given by:

$$\begin{aligned} k_2(x_1, x_2) &= \int_0^1 ((x_1 - u)_+(x_2 - u)_+) du \\ &= \int_0^{x_1 \wedge x_2} ((x_1 - u)(x_2 - u)) du \\ &= (u(6x_1x_2 - 3x_1u - 3x_2u + 2u^2))/6|_{u = x_1 \wedge x_2} \\ &= x_1x_2(x_1 \wedge x_2)/2 + (x_1 \wedge x_2)^3/6 \end{aligned}$$

where $x_1 \wedge x_2 = \min(x_1, x_2) = (x_1 + x_2 - |x_1 - x_2|)/2$

Some other common choices of kernels:

- polynomial
  $k(x_1, x_2) = (x_1^T x_2 + c)^M$ for $c > 0$ and $M \in \mathbb{N}$

- sigmoid
  $k(x_1, x_2) = \tanh(a x_1^T x_2 + b)$

- non-vectorial space [2, p. 297]

  $k(A_1, A_2) = 2^{|A_1 \cap A_2|}$, where ”$| \cdot |$” means the number of subsets of a set.

1.2 The reproducing kernel Hilbert space

Provided with the notation of feature spaces, and positive definite kernels on them, we are now able to define a particular class of functions on $\mathcal{X}$ in this section. The space of such functions is known as a *reproducing kernel Hilbert space*. We will see that RKHS are spaces of functions with the nice property that if a function f is close to a function g (in the sense of the distance derived from the inner product), then their evaluations $f(x)$ and $g(x)$ are close, too.

Definition 1.2.1. *Reproducing kernel Hilbert space*

Let $\mathcal{H}$ be a Hilbert space endowed with an inner product $\left\langle \cdot, \cdot \right\rangle_{\mathcal{H}}$ of real functions f defined on an index set $\mathcal{X}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:

1. *for every fixed $x_1 \in \mathcal{X}$, $k(x_1, x_2)$ as a function of x_2 belongs to $\mathcal{H}$,*

2. *k has the reproducing property:*

$$\left\langle f(\cdot), k(x, \cdot) \right\rangle_{\mathcal{H}} = f(x)$$

   for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$

Remark 4. *Note that, by the reproducing property we have the following:*

$$k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} \tag{1.2}$$

That is, an RKHS associated with a positive definite kernel k (called a reproducing kernel (RK)) gives a desired feature space. This can be further developed as a theorem [9, p. 2] below:

Proposition 1.2.1. *Every RKHS corresponds a unique RK and conversely, given a positive-definite function k on $\mathcal{X} \times \mathcal{X}$ we can construct a unique RKHS of real-valued functions on $\mathcal{X}$ with k as its RK.*

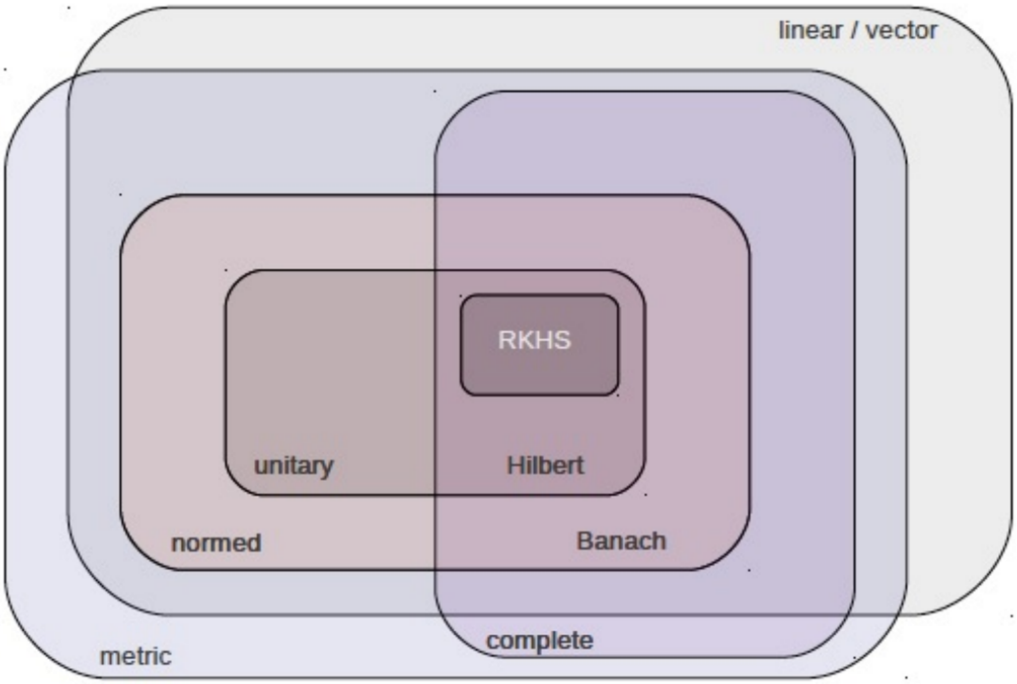

Figure 1.1: The relationship between some common abstract spaces [5, p. 19]

Proof. Assume we have two RKs k_1 and k_2, then by the linearity of inner product and the reproducing property we have $\langle f, k_1(\cdot,x)-k_2(\cdot,x)\rangle_{\mathcal{H}} = f(x) - f(x) = 0$. Letting $f(\cdot) = k_1(\cdot,x) - k_2(\cdot,x)$ yields $||k_1(\cdot,x) - k_2(\cdot,x)||^2_{\mathcal{H}} = 0$ which means k_1 and k_2 are identical by property of norm.

The converse part follows from constructing a linear manifold by taking all finite linear combinations of the form $\sum_{i=1}^{n} c_i k(x_i,\cdot)$ for all choice of n and $\{c_i\}_n^{i=1} \subset \mathbb{R}$, $\{x_i\}_n^{i=1} \subset \mathcal{X}$. □

Example 1.2.1. ***Cameron-Martin space*** *[10, p. 88] Define* $\mathcal{H} = \{f : f(0) = 0, f$ *is absolutely continuous and its derivative* f' *is square integrable on the interval* $\mathcal{X} = (0,1)\}$ *equipped with the inner product:*

$$\langle f, g\rangle_{\mathcal{H}} = \int_{\mathcal{X}} f'g'$$

Then by choosing the kernel $k(x,y) = x \wedge y$ *we see the reproducing property holds:*

$$\big\langle f(\cdot), k(x,\cdot)\big\rangle_{\mathcal{H}} = \int_0^1 f'(y)\frac{\partial}{\partial y}k(x,y)dy = \int_0^x f'(y)dy = f(x)$$

where we compute $\frac{\partial}{\partial y}(x \wedge y) = \mathbb{I}_{(0,x)}(y)$ *in the sense of weak derivative.*

Example 1.2.2. *Define* $\mathcal{H} = \{f : f$ *is absolutely continuous.* f *and* f' *are square integrable on* $\mathcal{X} = \mathbb{R}\}$ *equipped with the inner product:*

$$\langle f, g\rangle_{\mathcal{H}} = \int_{\mathcal{X}} (fg + f'g')$$

Then by choosing a (scaled) exponential kernel $k(x, y) = \frac{1}{2}\exp(-|x - y|)$ *we see the reproducing property holds:*

$$\begin{aligned}
\big\langle f(\cdot), k(x, \cdot)\big\rangle_{\mathcal{H}} &= \int_{\mathbb{R}} f(y)k(x, y) + f'(y)\frac{\partial}{\partial y}k(x, y)dy \\
&= \int_{\mathbb{R}} f(y)k(x, y)dy + \int_{-\infty}^{x} f'(y)\frac{\partial}{\partial y}k(x, y)dy + \int_{x}^{\infty} f'(y)\frac{\partial}{\partial y}k(x, y)dy \\
&= \int_{\mathbb{R}} f(y)k(x, y)dy + f(y)\frac{\partial}{\partial y}k(x, y)|_{y=-\infty}^{x} - \int_{-\infty}^{x} f(y)\frac{\partial^2}{\partial y^2}k(x, y)dy \\
&+ f(y)\frac{\partial}{\partial y}k(x, y)|_{y=x}^{\infty} - \int_{x}^{\infty} f(y)\frac{\partial^2}{\partial y^2}k(x, y)dy \\
&= \int_{\mathbb{R}} f(y)k(x, y)dy + \frac{1}{2}f(x) - 0 - \int_{-\infty}^{x} f(y)k(x, y)dy \\
&+ 0 - (-\frac{1}{2}f(x)) - \int_{x}^{\infty} f(y)k(x, y)dy \\
&= f(x)
\end{aligned} \tag{1.3}$$

where we used the identities: $\frac{\partial}{\partial y}k(x, y) = \begin{cases} k(x, y) \text{ if } y > x \\ -k(x, y) \text{ if } y < x \end{cases}$ *and* $\frac{\partial^2}{\partial y^2}k(x, y) = k(x, y)$ *if* $y = x$*.*

1.3 An equivalent definition

The evaluation functional δ_x defined by $\delta_x(f) = f(x)$ assigns a real number to the given function. In general, the evaluation functional is not continuous in the sense that $f_n \to f$ does not imply $\delta_x(f_n) \to \delta_x(f)$. For example, given $f(x) = 0$ and $f_n(x) = \sqrt{n}\mathbb{I}(x < 1/n^2)$, then we have $||f_n - f|| = (\int_0^1 (|f_n - f|)^2 dx)^{1/2} = (\int_0^{1/n^2} (\sqrt{n})^2 dx)^{1/2} = 1/\sqrt{n} \to 0$ but $\delta_0(f_n) = \sqrt{n}$ which does not converge to $\delta_0(f) = 0$. Intuitively, this is because Hilbert spaces can contain very unsmooth functions. We shall see that RKHS is a special class of Hilbert spaces where the evaluation functional is continuous, meaning that the functions in the space are well-behaved. We first introduce some basic concepts in functional analysis.

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 it satisfies the following properties:

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

Remark 5. *Clearly the evaluation functional δ_x is a liner operator, as $\delta_x(\alpha f + \beta g) = (\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x) = \alpha\delta_x(f) + \beta\delta_x(g)$.*

Definition 1.3.2. ***(Operator norm)*** *The operator norm of a linear operator A: $\mathcal{F} \to \mathcal{G}$ is defined by:*$||A|| = \sup\limits_{f\in\mathcal{F}} \frac{||Af||_{\mathcal{G}}}{||f||_{\mathcal{F}}}$.

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

Definition 1.3.4. ***(Continuity)***

A function A: $\mathcal{F} \to \mathcal{G}$ is said to be continuous at $f_0 \in \mathcal{F}$, if for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon, f_0) > 0$, such that $||f - f_0||_{\mathcal{F}} < \delta$ implies $||Af - Af_0||_{\mathcal{G}} < \epsilon$.

Definition 1.3.5. *A function A is continuous on $\mathcal{F}$, if it is continuous at every element of $\mathcal{F}$.*

In fact the boundedness and continuity turn out to be equivalent in normed linear spaces:

Proposition 1.3.1. *Let $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ and $(\mathcal{G}, ||\cdot||_{\mathcal{G}})$ be normed linear spaces. If A is a linear operator, then the following are equivalent:*

1. *A is a bounded operator.*
2. *A is continuous on $\mathcal{F}$.*
3. *A is continuous at one point of $\mathcal{F}$.*

Proof. (1)⇒(2):By definition 1.3.2 we have

$$||A(f_1 - f_2)||_{\mathcal{G}} \leq ||A||\ ||(f_1 - f_2)||_{\mathcal{F}}$$

Since $||A|| < \infty$, we know A is Lipschitz continuous with a Lipschitz constant $||A||$ and hence A is continuous over $\mathcal{F}$.
(2)⇒(3): It trivially holds true.
(3)⇒(1): For linearl operator f and f_0, let $g = f - f_0$. By definition 1.3.4, there exists a fixed $\delta > 0$, for any $||f - f_0||_{\mathcal{F}} = ||g||_{\mathcal{F}} < \delta$ we can have $||Af - Af_0||_{\mathcal{G}} = ||A(g)||_{\mathcal{G}} < 1$. Then

$$||A(g)||_{\mathcal{G}} = \delta^{-1}||g||_{\mathcal{F}}||A(\delta\frac{g}{||g||_{\mathcal{F}}})||_{\mathcal{G}} \leq \delta^{-1}||g||_{\mathcal{F}}$$

Hence $||A|| \leq \delta^{-1} < \infty$ and A is bounded. □

With the ingredients listed above we can now build the equivalent definition of RKHS:

Proposition 1.3.2. ***(Equivalence definition of RHKS)*** *A Hilbert space $\mathcal{H}$ of functions defined on non-empty $\mathcal{X}$ is an RKHS if and only if $\forall x \in \mathcal{X}$ the evaluation functional δ_x is continuous (or, equivalently, bounded) on $\mathcal{H}$.*

Proof. Suppose $\mathcal{H}$ is an RKHS with RK k. Then $\forall x \in \mathcal{X}$, applying the Cauchy-Schwarz inequality and the reproducing property of RKHS yields:

$$\begin{aligned} |\delta_x(f)| &= |\langle f, k(x,\cdot)\rangle| \\ &\leq ||f||_{\mathcal{H}}||k(x,\cdot)||_{\mathcal{H}} \\ &= ||f||_{\mathcal{H}}\langle k(x,\cdot), k(x,\cdot)\rangle_{\mathcal{H}}^{1/2} \\ &= ||f||_{\mathcal{H}}k(x,x)^{1/2} \end{aligned} \tag{1.4}$$

Hence δ_x is a bounded linear operator on $\mathcal{H}$.
The converse part immediately follows from *Riesz Representation Theorem* [3, p. 13], [3] where we let $g_{\delta_x}(\cdot) = k(\cdot, x) \in \mathcal{H}$ and write $\delta_x(f) = \langle f, g_{\delta_x}\rangle_{\mathcal{H}} = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ and hence k is the corresponding RK. □

[3] **Riesz Representation Theorem**: In a Hilbert space $\mathcal{H}$, all bounded linear operators are of the form $\langle\cdot, g\rangle_{\mathcal{H}}$, for some $g \in \mathcal{H}$.

Remark 6. *$k(\cdot, x)$ is usually called the kernel section or* ***representer*** *of evaluation at x.*

Corollary 1.3.5.1. *(Evaluation functional is continuous in $\mathcal{H}$.) If two functions converge in RKHS norm, then they converge at every point, i.e.,$||f_n - f||_{\mathcal{H}} \to 0$ implies $f_n(x) \to f(x)$, $\forall x \in \mathcal{X}$*

Proof. $|f_n(x) - f(x)| = |\langle f_n - f, k(x, \cdot)\rangle_{\mathcal{H}}| \leq ||f_n - f||_{\mathcal{H}} ||k(x, \cdot)||_{\mathcal{H}} \to 0$ □

An important property of the RKHS norm $||f||_{\mathcal{H}}$ is that it captures not only the magnitude of a function $f \in \mathcal{H}$ but also its smoothness: f gets smoother as $||f||_{\mathcal{H}}$ decreases, and vice versa. This is particularly important in understanding why regularization is required for kernel ridge regression, to avoid overfitting.

1.4 Representer Theorem

As a consequence of the last section, one of the crucial properties of kernels is that even if the input domain $\mathcal{X}$ is only a set, we can nevertheless think of the pair $(\mathcal{X}, k)$ as a (subset of a) Hilbert space, thus allowing us to explore various data structures in Hilbert spaces whose theory is very well developed. From a practical point of view, however, the new problem is that the Hilbert space is usually infinite-dimensional for many popular kernels and that causes difficulty in optimization. This issue may be addressed by the following well known theorem, which shows that a large class of optimization problems with RKHS regularizers have solutions spanned by kernel sections in terms of the training data, whose dimension is finite.

Theorem 1.4.1 (Representer Theorem). *[7] Given a positive definite kernel k on $\mathcal{X} \times \mathcal{X}$ (and an RKHS $\mathcal{H}$ induced by it), a training set $(\mathbf{x}, \mathbf{y}) \subset \mathcal{X} \times \mathbb{R}$, a strictly monotonic increasing real-valued function Ω on $[0, \infty]$, and an arbitrary cost function $c : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$, any $f \in \mathcal{H}$ minimizing the* *regularized risk functional*

$$c((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + \Omega(||f||_{\mathcal{H}})^4$$

admits a representation of the form

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

Proof. First we define $\phi(x_i) = k(\cdot, x_i)$. Since every vector space can be written as a direct sum of a subspace and its orthogonal complement, we use the orthogonal projection and decompose the function f in the following form

$$f = \sum_{i=1}^{n} \alpha_i \phi(x_i) + v,$$

where

$$\langle \phi(x_i), v \rangle_{\mathcal{H}} = 0 \quad \forall i = 1, .., n.$$

The orthogonality condition simple ensures that v is not in the span of $\{\phi(x_i)\}_{i=1}^{n}$.So for any point x_j $(j = 1, ..., n)$, applying the reproducing property yields:

$$f(x_j) = \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i) + v, \phi(x_j) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}},$$

so v has no effect on the cost function $c\left((f(x_1), y_1), ..., (f(x_n), y_n)\right)$.

We now check the regularization term, by the monotonicity of Ω and orthogonality condition, we compute:

$$\begin{aligned} \Omega(\|f\|_{\mathcal{H}}) &= \Omega\left(\left\| \sum_{i=1}^{n} \alpha_i \phi(x_i) + v \right\|_{\mathcal{H}} \right) \\ &= \Omega\left(\sqrt{ \left\| \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2 } \right) \qquad (1.5) \\ &\geq \Omega\left(\sqrt{ \left\| \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|_{\mathcal{H}}^2 } \right). \end{aligned}$$

[4] Here $||\cdot||_{\mathcal{H}}$ is the norm in the RKHS $\mathcal{H}_k$ associated with a given pd kernel k, i.e. for any $z_i \in \mathcal{X}$, $\beta_i \in \mathbb{R}(\ i \in \mathbb{N})$,

$$||\sum_{i=1}^{\infty} \beta_i k(\cdot, z_i)||^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \beta_i \beta_j k(z_i, z_j).$$

Setting $v = 0$ thus does not affect the cost function , while strictly reducing the second term. Hence, any minimizer must have $v = 0$ and the solution takes the form:

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i \phi(x_i) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

□

This theorem in fact provides a powerful way of detecting non-linear relations using standard function minimization algorithms. It dramatically simplify the regularized empirical risk minimization problem by reducing the search domain from an infinite-dimensional function space to a finite (n-dimensional) vector space. Consequently, the representer theorem provides the theoretical basis for the reduction of the general machine learning problem to algorithms that can actually be implemented on computers in practice.

1.5 An equivalent characterization of RKHS norm

If we could write a function $f \in \mathcal{H}$ as a finite sum of kernel sections as $f(\cdot) = \sum_{i=1}^{n} c_i k(\cdot, x_i)$, then there exists an a equivalent characterization of RKHS norm of such f.

Theorem 1.5.1. *Let k be a pd kernel on $\mathcal{X}$ and $\mathcal{H}$ be its RKHS. Then for any $n \in \mathbb{N}$, $x_1, ..., x_n \in \mathcal{X}$ and $c_1, ..., c_n \in \mathbb{R}$, we have:*

$$||f(\cdot)||_{\mathcal{H}} = ||\sum_{i=1}^{n} c_i k(\cdot, x_i)||_{\mathcal{H}} = \sup_{f \in \mathcal{H}, ||f||_{\mathcal{H}} \leq 1} \sum_{i=1}^{n} c_i f(x_i)$$

Proof. By the reproducing property, the right side of 1.5.1 can be written as

$$\sup_{||f||_{\mathcal{H}} \leq 1} \sum_{i=1}^{n} c_i f(x_i) = \sup_{||f||_{\mathcal{H}} \leq 1} \langle \sum_{i=1}^{n} c_i k(\cdot, x_i), f(\cdot) \rangle_{\mathcal{H}}$$

By the Cauchy-Schwartz inequality, it is upper-bounded as:

$$RHS \leq \sup_{||f||_{\mathcal{H}} \leq 1} ||\sum_{i=1}^{n} c_i k(\cdot, x_i)||_{\mathcal{H}} ||f||_{\mathcal{H}} = ||\sum_{i=1}^{n} c_i k(\cdot, x_i)||_{\mathcal{H}}$$

On the other hand, on letting $g(\cdot) = \sum_{i=1}^{n} c_i k(\cdot, x_i)/|| \sum_{i=1}^{n} c_i k(\cdot, x_i)||_{\mathcal{H}}$, we have $||g||_{\mathcal{H}} = 1$, and thus the right side of 1.5.1 can be lower-bounded as:

$$RHS \geq \langle \sum_{i=1}^{n} c_i k(\cdot, x_i), g(\cdot) \rangle_{\mathcal{H}} = || \sum_{i=1}^{n} c_i k(\cdot, x_i)||_{\mathcal{H}}$$

Hence the equality in 1.5.1 holds true. □

Chapter 2

Bayesian estimation of a Gaussian random field

We first review the problem of regression model. For a given non-empty set $\mathcal{X}$ and a *regression function* f: $\mathcal{X} \to \mathbb{R}$, assume that the training data is described by n pairs of observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ such that:

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

where ϵ_i is a zero-mean random variable, usually referred as the noise term in the output. The main goal of regression is to reconstruct f from the training sample $(x_i, y_i)_{i=1}^n$. Though f itself is a deterministic function, its evaluation will be henceforth treated as realization of a Gaussian random field. Specifically, we shall make two assumptions:

Assumption 1 (f as a Gaussian random field): For any finite collections of sample points $\{x_j : j = 1, ..., J\}$, the vector $f(\mathbf{x}) = (f(x_1), ..., f(x_J))$ is a zero-mean Gaussian random variable with covariance:

$$\mathbf{cov}(f(x_i), f(x_l)) = \lambda k(x_i, x_l) \tag{2.1}$$

where λ is a fixed positive constant and $k(\cdot, \cdot)$ is a positive definite autocovariance function on $\mathcal{X} \times \mathcal{X}$. Such function f that satisfies the above assumption is called a zero-mean **Gaussian random field** on $\mathcal{X}$.

Assumption 2 (Conditions on the noise): The independent measurement $\mathbf{y}$ given the regression function f is distributed according to an exponential form:

$$\mathbb{P}(\mathbf{y}|f(\mathbf{x})) \propto \Pi_{i=1}^{n} \exp(-\frac{V(y_i - f(x_i))}{2\sigma^2})$$

where V is a loss function and $\sigma > 0$. Note that we assume the measurement noise $\epsilon_i = y_i - f(x_i)$ are independent of $f(x_i)$.

2.1 Gaussian noised data

Suppose assumptions 1 and 2 hold with $V_i(r) = r^2$ applying the above result from Appendix A to the jointly Gaussian random vector $\mathbf{x_1} = f(x^*)$ for a new input x^* and $\mathbf{x_2} = \mathbf{y}$ yileds

$$\begin{aligned} \mathbf{cov}(f(x^*), y_j) &= \mathbf{cov}(f(x^*), f(x_j) + \epsilon_j) \\ &= \mathbf{cov}(f(x^*), f(x_j)) + \mathbf{cov}(f(x^*), \epsilon_j) \\ &= \lambda k(x^*, x_j) + 0 \end{aligned} \tag{2.2}$$

$$\begin{aligned} \mathbf{cov}(y_i, y_j) &= \mathbf{cov}(f(x_i) + \epsilon_i, f(x_j) + \epsilon_j) \\ &= \mathbf{cov}(f(x_i), f(x_j)) + \mathbf{cov}(\epsilon_i, \epsilon_j) + 0 \\ &= \lambda k(x_i, x_j) + \sigma^2 \end{aligned} \tag{2.3}$$

where the zero terms are because of independence assumption. Then the posterior conditional expectation of the prediction at new observation x^* is given by:

$$\begin{aligned} \mathbb{E}((f(x^*))|\mathbf{y})) &= 0 + (\mathbf{cov}(f(x^*), \mathbf{y}))(\mathbf{cov}(\mathbf{y}, \mathbf{y})^{-1}(\mathbf{y} - \mathbf{0}) \\ &= \lambda(k(x^*, x_1), ..., k(x^*, x_n))(\lambda\bar{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} \\ &= (k(x^*, x_1), ..., k(x^*, x_n))(\bar{K} + \gamma\mathbf{I}_n)^{-1}\mathbf{y} \end{aligned} \tag{2.4}$$

where we define $\gamma = \sigma^2/\lambda$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

The problem of inferring an underlying function $f(\cdot)$ from a noisy dataset without any additional assumptions is clearly ill-posed [6, p.129]. For example, in the noise-free case, any function that passes through the given data points is acceptable. Under a Bayesian approach our assumptions are characterized by a prior over functions, and given some data, we obtain a posterior over functions. The problem of bringing prior assumptions to bear has also been addressed under the regularization viewpoint, where these assumptions are encoded in terms of the smoothness of f.

In general, the problem can be written as follows. To reconstruct f from the noisy data, we estimate

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \sum_{i=1}^{n} V_i(y_i - f(x_i)) + \gamma||f||_{\mathcal{H}}^2 \tag{2.5}$$

One of the important features of the above approach is that, even if the dimension of $\mathcal{H}$ is infinite, the solution belongs to a finite-dimensional subspace. In fact, under mild assumptions on the loss function, according to the representer theorem 1.4.1, $\hat{f}(\cdot)$ in 2.5 is given by a finite sum of kernel sections $k(x_i, \cdot)$:

$$\hat{f}(\cdot) = \sum_{i=1}^{n} \hat{c}_i k(x_i, \cdot) \tag{2.6}$$

where $\hat{c}$ is hence estimated by:

$$\hat{c} = \arg\min_{c\in\mathbb{R}^n} \sum_{i=1}^{n} V_i(y_i - \sum_{j=1}^{n} c_j k(x_j, x_i)) + \gamma c^T \bar{K} c. \tag{2.7}$$

Notice that the squared norm with f in the form of 2.6 is then calculated by:

$$\begin{aligned} ||f||_{\mathcal{H}}^2 &= \langle \sum_{i=1}^{n} c_i k(x_i, \cdot), \sum_{j=1}^{n} c_j k(x_j, \cdot)\rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^{n} c_i c_j \langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \\ &= c^T \bar{K} c \end{aligned} \tag{2.8}$$

To see this coincides with the Gaussian random field approach in the previous example, we calculate:

$$\frac{d}{dc}[(\mathbf{y} - \bar{K}c)^T(\mathbf{y} - \bar{K}c) + \gamma c^T \bar{K} c] = 2\bar{K}^2 c - 2\bar{K}y + 2\gamma \bar{K} c \tag{2.9}$$

Since the Hessian matrix is clearly positive definite, setting the above derivative to zero we conclude the minimizer is given by:

$$\hat{c} = \arg\min_{c\in\mathbb{R}^n} (\mathbf{y} - c^T\bar{K})^T(\mathbf{y} - c^T\bar{K}) + \gamma c^T \bar{K} c = (\bar{K} + \gamma \mathbf{I}_N)^{-1} y$$

2.2 Error estimate of Posterior Variance

Apart from the estimate of the mean, we may also compute the posterior variance and do some further analysis. Throughout this section we will still assume the noise to be Gaussian with $V(r) = r^2$. As in 2.4 we would compute with help from classic conclusion from conditional multivariate Gaussian distribution. First we define: $k_x = (k(x_1, x), ..., k(x_n, x))^T$, the column $n \times 1$ vector of kernel section evaluated at x and define $k_{(\cdot)} = (k(x_1, \cdot), ..., k(x_n, \cdot))^T$ wherever the argument x can be omitted. Then the posterior variance is given by:

$$\begin{aligned}\mathbf{Var}((f(x^*))|\mathbf{y})) = (\mathbf{cov}(f(x^*), \mathbf{cov}(f(x^*))) - (\mathbf{cov}(f(x^*), \mathbf{y}))(\mathbf{cov}(\mathbf{y},\\ \mathbf{y})^{-1})(\mathbf{cov}(\mathbf{y}, f(x^*)) \quad (2.10)\\ = \lambda k(x^*, x^*) - \lambda k_{x^*}^T (\bar{K} + \gamma \mathbf{I}_n)^{-1} k_{x^*}\end{aligned}$$

From the Bayesian viewpoint, that quantity is usually interpreted as the average case error at a location x^* to measure the uncertainty of the estimates from Gaussian process regression. The next upcoming theorem is to show that there exists a frequentist interpretation of $\mathbf{Var}((f(x^*))|\mathbf{y}))$ as *the worst case error.*

Define a new kernel

$$k^\gamma(x_1, x_2) = k(x_1, x_2) + \gamma \Delta(x_1, x_2), \ x_1, x_2 \in \mathcal{X}$$

where $\Delta(\cdot, \cdot)$ is the Kronecker delta function:$\Delta(i, j) = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$.
With that kernel we have corresponding RKHS $\mathcal{H}_{k^\gamma}$ and Gram matrix $\bar{K}_\gamma = \bar{K} + \gamma \mathbf{I}_n$. Notice that the posterior mean from 2.4 can now be conveniently written as:

$$\mathbb{E}((f(x^*))|\mathbf{y})) = k_{x^*}^T \bar{K}_\gamma^{-1} \mathbf{y}$$

We then have the following theorem [4]:

Theorem 2.2.1. *For any $x^* \in \mathcal{X}$ that is independent of $(x_i)_{i=1}^n$, we have*

$$\sqrt{\mathbf{Var}((f(x^*))|\mathbf{y}))/\lambda + \gamma} = \sup_{g \in \mathcal{H}_{k^\gamma}, ||g||_{\mathcal{H}_{k^\gamma}} \leq 1} (g(x^*) - k_{x^*}^T \bar{K}_\gamma^{-1} g(\mathbf{x}))$$

Proof. By using the theorem 1.5 we have:

$$\begin{aligned} ||k^\gamma(\cdot, x^*) - k_{x^*}^T \bar{K}_\gamma^{-1} k_{(\cdot)}^\gamma||_{\mathcal{H}^\gamma} &= ||k^\gamma(\cdot, x^*) - \sum_{i=1}^n (k_{x^*}^T \bar{K}_\gamma^{-1})_i k^\gamma(\cdot, x_i)||_{\mathcal{H}^\gamma} \\ &= \sup_{g \in \mathcal{H}_{k^\gamma}, ||g||_{\mathcal{H}_{k^\gamma}} \leq 1} (g(x^*) - \sum_{i=1}^n (k_{x^*}^T \bar{K}_\gamma^{-1})_i g(x_i)) \\ &= \sup_{g \in \mathcal{H}_{k^\gamma}, ||g||_{\mathcal{H}_{k^\gamma}} \leq 1} (g(x^*) - k_{x^*}^T \bar{K}_\gamma^{-1} g(\mathbf{x})) \end{aligned} \tag{2.11}$$

Simplifying the square of the left hand side of the above equation yields:

$$\begin{aligned} (LHS)^2 &= k^\gamma(x^*, x^*) - 2\sum_{i=1}^n (k_{x^*}^T \bar{K}_\gamma^{-1})_i k^\gamma(x^*, x_i) + \sum_{i,j=1}^n (k_{x^*}^T \bar{K}_\gamma^{-1})_i (k_{x^*}^T \bar{K}_\gamma^{-1})_j k^\gamma(x_i, x_j) \\ &= k(x^*, x^*) + \gamma - 2\sum_{i=1}^n (k_{x^*}^T \bar{K}_\gamma^{-1})_i k^\gamma(x^*, x_i) + (k_{x^*}^T \bar{K}_\gamma^{-1}) \bar{K}^\gamma (k_{x^*}^T \bar{K}_\gamma^{-1})^T \\ &= k(x^*, x^*) + \gamma - 2\sum_{i=1}^n (k_{x^*}^T \bar{K}_\gamma^{-1})_i k(x^*, x_i) + k_{x^*}^T \bar{K}_\gamma^{-1} k_{x^*} \\ &= k(x^*, x^*) + \gamma - k_{x^*}^T \bar{K}_\gamma^{-1} k_{x^*} \\ &= \mathbf{Var}((f(x^*))|\mathbf{y}))/\lambda + \gamma \end{aligned} \tag{2.12}$$

where in the third line we used the property of k^γ that $k^\gamma(x_1, x_2) = k(x_1, x_2)$ if $x_1 \neq x_2$. □

Remark 7. *If $\sigma^2 = 0$ i.e. the sample data is noise-free, then the problem is called interpolation. If we further assume $\bar{K}$ is invertible, then the Theorem 2.2.1 shows that the posterior variance* $\mathbf{Var}((f(x^*))|\mathbf{y}))$ *(up to a constant multiple λ) provides an upper-bound on the squared error of kernel interpolation for any fixed target function in RKHS induced by k. Specifically, by the reproducing property and Cauchy-Schwartz inequality we have the following corollary:*

Corollary 2.2.1.1. *Given $\sigma^2 = 0$ (and thus $\gamma = 0$), and an invertible Gram matrix $\bar{K}$, for any new input $x^* \in \mathcal{X}$ we have:*

$$(m(x^*) - f(x^*))^2 \leq ||f||_{\mathcal{H}} \mathbf{Var}((f(x^*))|\mathbf{y}))/\lambda, \ \forall f \in \mathcal{H},$$

$where\ m(x^*) = k_{x^*}^T \bar{K}^{-1} f(\mathbf{x})$

Chapter 3

Statistical interpretation

Now we consider the general case when the Gaussian assumptions on the noise ϵ_i are removed. If we are given that f has a well defined prior probability density:

$$\mathbb{P}(f) \propto \exp\left(\frac{-||f||^2_{\mathcal{H}}}{2\lambda}\right) \tag{3.1}$$

Then it immediately follows from the Bayesian rule that the posterior distribution (given the assumption 1) is:

$$\mathbb{P}(f|\mathbf{y}) \propto \exp\left(-\sum_{i=1}^{n} V_i(y_i - f(x_i))/2\sigma^2 - ||f||^2_{\mathcal{H}}/2\lambda\right) \tag{3.2}$$

Then $\hat{f}$ defined by 2.5 is called a maximum a posteriori (MAP) estimator, as minimization in 2.5 is evidently equivalent to maximizing the right handside of 3.2 . Given $V_i(r) = r^2$, $\hat{f}$ can be also interpreted as the posterior mean which the minimum variance estimator (MVE) of $f(x)$, see Appendix B. It is worth to mention that in general the MAP and the MVE are different without the the special assumption of square loss function.

3.1 Connection between MVE and MAP

Nevertheless, the minimum variance estimate $\mathbb{E}[f(\cdot)|y]$ and the MAP estimate $\hat{f}$ belong to the same (finite-dimensional) subspace of $\mathcal{H}$ spanned by the kernel sections $k(x_i, \cdot)$, $i = 1, ..., n$.

Proposition 3.1.1. *[1] Assume f satisfies Assumption 1 and $\mathbb{P}(y|f)$ satisfies Assumption 2. Then the MVE of f at x^* (independent of training sample* *inputs* $\mathbf{x} = (x_i)_{i=1}^n$*) is given by:*

$$\mathbb{E}[f(x^*)|y] = k_{x^*}^T \bar{K}^{-1}\mathbb{E}(f(\mathbf{x})|y)$$

Proof.

$$\begin{aligned}
\mathbb{E}[f(x^*)|y] &= \mathbb{E}(\mathbb{E}[f(x^*)|f(\mathbf{x})]|y) \\
&= \mathbb{E}(\mathbf{cov}[f(x^*), f(\mathbf{x})]\mathbf{cov}(f(\mathbf{x}), f(\mathbf{x}))^{-1}f(\mathbf{x})|y) \\
&= \mathbb{E}(k_{x^*}^T \bar{K}^{-1}f(\mathbf{x})|y) \\
&= k_{x^*}^T \bar{K}^{-1}\mathbb{E}(f(\mathbf{x})|y)
\end{aligned} \tag{3.3}$$

□

where in the second equality we applied the result from Appendix A.

3.2 Justification of the MAP estimate on new inputs

The statistical interpretation would be perfectly valid if the density of the prior information of the Gaussian random field f (given by $\mathbb{P}(f) \propto \exp\frac{-||f||_{\mathcal{H}}^2}{2\lambda}$) was well defined. Unfortunately this is not true in general since $\mathcal{H}$ can be infinite-dimensional where the concept of probability density is not well defined and the finite-dimensional form 2.8 does not apply. Although in some special cases as in Section 2.1 one may obtain the exact same result by properties of Gaussian Process instead of relying on Bayes' rule, this may not work for general choice of loss function $V_i(\cdot)$.

However, in real applications one can never obtain infinite-size samples. After establishing the model $\hat{y}_i = \hat{f}(x_i)$ based on the training set $(x_i, y_i)_{i=1}^n$, we would like to examine its performance on the new inputs $(x_i)_{i=n+1}^{n+m}$ whose elements are independently distributed according to the same probability distribution as the training set. In fact, one can prove [1] that , the estimated function $\hat{f}$ defined by 2.5 is a legitimate MAP on $(x_i)_{i=1}^{n+m}$.

Theorem 3.2.1. *Assume the Assumption 1 and 2 are satisfied. Let* $(x_i)_{i=n+1}^{n+m}$ *be an arbitrary set of points in* $\mathcal{X}$ *where* m *is a given non-negative integer, and* *define* $f = [f(x_1), ..., f(x_{n+m})]^T$*. Then the MAP estimate for* f *given* $(y_i)_{i=1}^n$ *is:*

$$\arg\max_f \mathbb{P}(y, f) = [\hat{f}(x_1), ..., \hat{f}(x_{n+m})]^T \tag{3.4}$$

where $\hat{f}$ *is defined by 2.5, with* $\gamma = \sigma^2/\lambda$ *and* $\mathcal{H}$ *is the RKHS induced by* k *from 2.1.*

Proof. Define $g = [f(x_1), ..., f(x_n)]^T$ and $h = [f(x_{n=1}), ..., f(x_{n+m})]^T$. Then we have $f = [g^T, h^T]^T$. Note that $\mathbb{P}(y|f) = \mathbb{P}(y|g)$ by the independent observation assumption. Then applying the Lemma 1 from Appendix C and the representer theorem yields:

$$\begin{aligned} \hat{g} &= \arg\max_{g \in \mathcal{H}} \mathbb{P}(y|g)\mathbb{P}(g), \\ &= \bar{K}\hat{c} \end{aligned} \tag{3.5}$$

where $\hat{c}$ is determined by:

$$\begin{aligned} \hat{c} &= \arg\max_{c \in \mathbb{R}^n} \exp\{-(\frac{\sum_{i=1}^n V_i(y_i - \sum_{j=1}^n c_j k(x_j, x_i))}{2\sigma^2} + \frac{c^T \bar{K} c}{2\lambda})\} \\ &= \arg\min_{c \in \mathbb{R}^n} \sum_{i=1}^n V_i(y_i - \sum_{j=1}^n c_j k(x_j, x_i)) + \gamma c^T \bar{K} c \end{aligned} \tag{3.6}$$

Hence $\hat{g} = [\hat{f}(x_1), ..., \hat{f}(x_n)]$ by 2.5.

Then we used the results and from Appendix A to calculate:

$$\begin{aligned} \hat{h} &= \max_h \mathbb{P}(h|g = \hat{g}) \\ &= \mathbf{cov}(h, g)\mathbf{cov}(g, g)^{-1}\hat{g} \\ &= \mathbf{cov}(h, g)(\lambda\bar{K})^{-1}(\bar{K}\hat{c}) \\ &= \begin{bmatrix} \bar{K}(x_1, x_{n+1}) & \dots & \bar{K}(x_n, x_{n+1}) \\ \vdots & \ddots & \vdots \\ \bar{K}(x_1, x_{n+m}) & \dots & \bar{K}(x_n, x_{n+m}) \end{bmatrix} \begin{bmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_n \end{bmatrix} \\ &= [\hat{f}(x_{n+1}), ..., \hat{f}(x_{n+m})]^T \end{aligned} \tag{3.7}$$

Hence we have:

$$\begin{aligned}
\arg\max_{f} \mathbb{P}(y, f) &= \arg\max_{g,h} \mathbb{P}(y, g, h) \\
&= \arg\max_{g,h} \mathbb{P}(y|g, h)\mathbb{P}(h|g)\mathbb{P}(g) \\
&= \arg\max_{g,h} \mathbb{P}(y|g)\mathbb{P}(g)\mathbb{P}(h|g) \\
&= \arg\max_{g} [\mathbb{P}(y|g)\mathbb{P}(g) \max_{h} \mathbb{P}(h|g)] \\
&= \arg\max_{g} \mathbb{P}(y|g)\mathbb{P}(g) \max_{h} \mathbb{P}(h|g = \hat{g}) \\
&= [\hat{f}(x_1), ..., \hat{f}(x_{n+m})]^T
\end{aligned} \tag{3.8}$$

where the fifth equality follows from the fact that $\max_h \mathbb{P}(h|g)$ is constant with respect to g. □

Thus we showed a formal connection between Bayesian estimation of a Gaussian random field and the more general case prescribed by Assumption 2. Given the training sample $(x_i, y_i)_{i=1}^{n}$, for any finite set of locations $(x_i)_{i=1}^{n+m}$, the MAP estimate of f at any given locations is the RKHS estimate (which are determined only from the training sample) evaluated at these locations.

Chapter 4

Simulation Study

In this section the use of the general model 2.5 is illustrated by means of numerical examples. We basically replicate the result from [1] . The function to be estimated is given by $f(x) = \exp(\sin 8x)$ on the unit interval. The choice of kernels reflects the subjective belief on the smoothness of the underlying function. A cubic spline kernel from 1.1.11 (shifted by 1 unit to satisfy the null initial condition) is adopted to reconstruct the function from the sample generated with additive iid noise: $y = f(x)+\epsilon$, assuming the unknown function f has first order continuous derivative.The samples are taken from observations of f at $x_i = (i-1)/66$, $i = 1, ..., 65$. The measurement noise is modeled by iid centered random variable (either Gaussian or Laplace) with variance of 0.09.

We first estimate the regularization parameter γ by a 3-fold cross validation. That is, we evenly divide the sample into three subsamples and obtain the total errors with different values of γ. Then an estimate of γ is chosen at which the minimum of CV errors is attained.

We simulate four cases, each consisting of 100 function reconstructions:

- Gaussian noise
- Gaussian noise with outliers
- Laplace noise
- Laplace noise with outliers

The presence of outliers is given by tripling the true function evaluations with probability 0.1. A typical case is presented below [4].

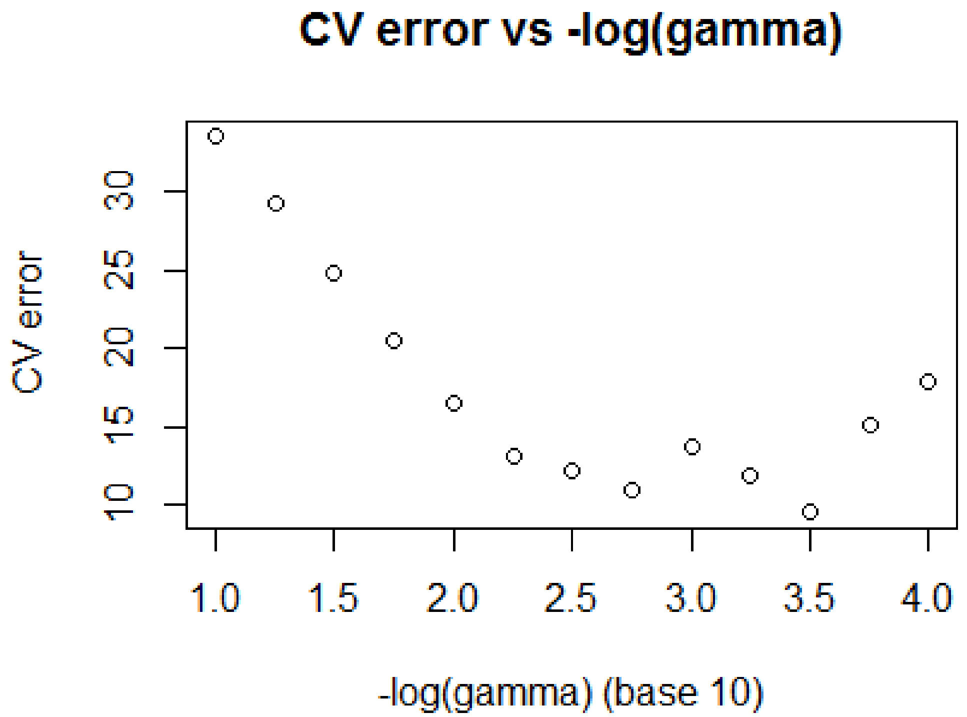

Figure 4.1: The errors in validation sets vs -$\log_{10}(\gamma)$. Note that the relationship between them may not be convex. We choose the minimum at -$\log_{10}(\gamma)$=3.5

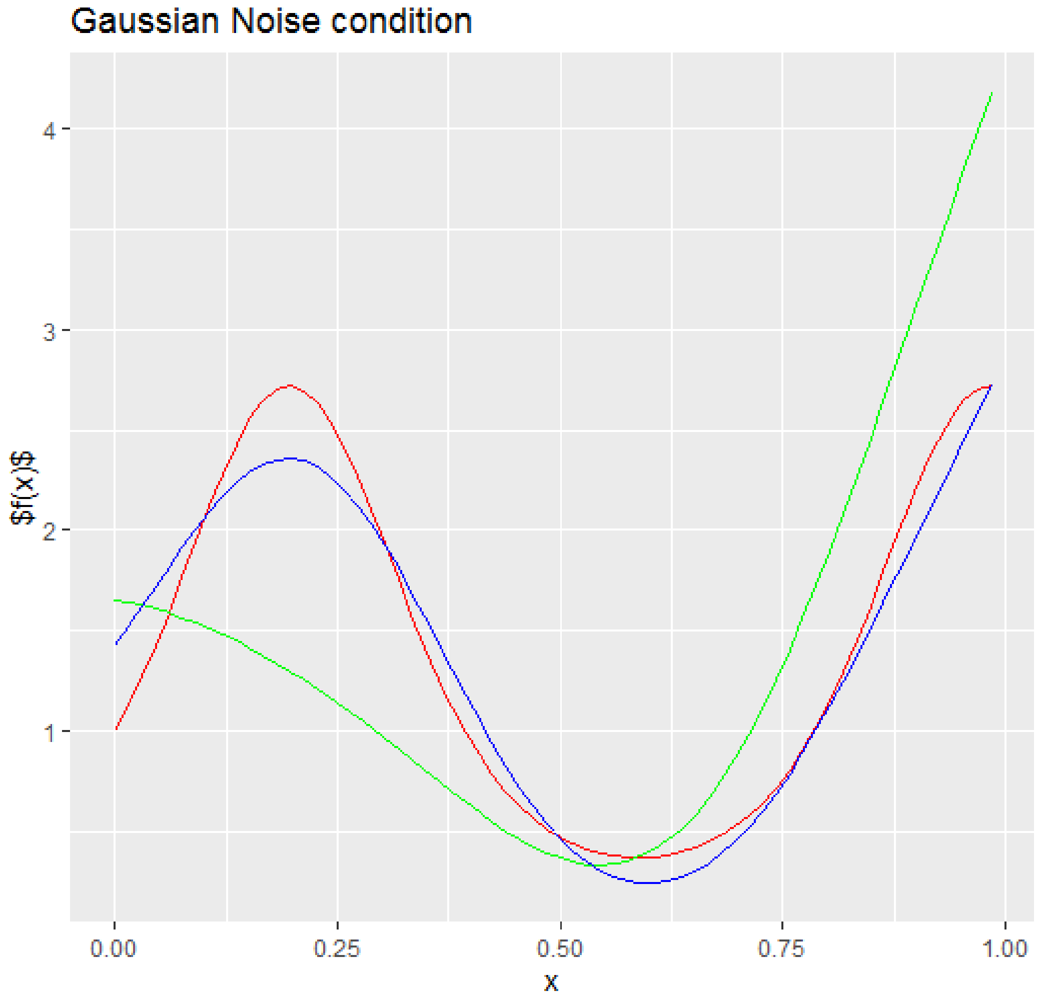

Figure 4.2: Red/blue/green lines represents true function/estimates/estimates under perturbed condition, respectively.

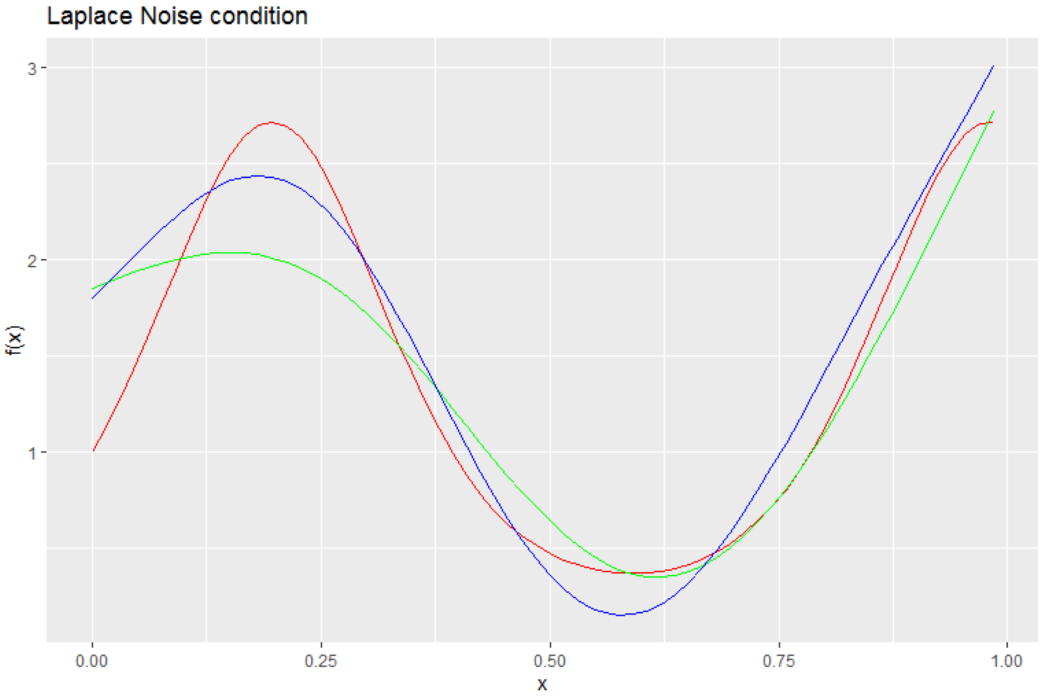

Figure 4.3: Example of simulated case under Laplace noise

It appears that under absolute value losses assumption (i.e. $V(\cdot) = |\cdot|$) the estimates suffers less from outliers than that under square losses assumption. From the box-plots [4] of the errors it is true in general, as expected.

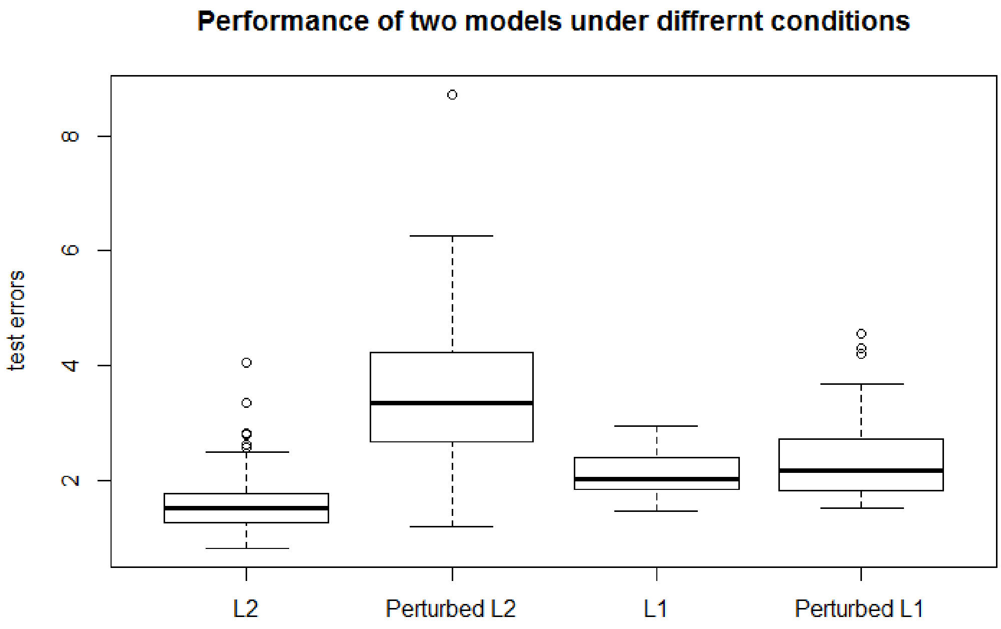

Figure 4.4: The errors are defined by $\sqrt{\sum_{i=1}^{66} \hat{f}(x_i) - f(x_i)}$

Chapter 5

Conclusions

In machine learning, statistics and numerical analysis, both the notion of a kernel and that of a Gaussian process play central roles in theoretical analysis. Kernel methods are founded on notions like regularization and optimization, while Gaussian processes are generative models operating in terms of marginal and conditional distributions. The present thesis provided a review of the intersection of these two areas, by proving the fact that the estimate computed by the regularization framework is indeed the MAP estimate of the undelying function f. We also showed that the MAP estimate and the MVE of f belong to the finite dimensional subspace of the RKHS induced by the pd kernel k. In particular, by assuming the quadratic loss function, the posterior variance of the RKHS estimate can be interpreted as the worst-case error in frequentist viewpoint.

References

[1] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, "The connection between bayesian estimation of a gaussian random field and rkhs", *IEEE transactions on neural networks and learning systems*, vol. 26, no. 7, pp. 1518–1524, 2015. 21, 22, 25

[2] C. M. Bishop, *Pattern recognition and machine learning.* springer, 2006. 6

[3] J. B. Conway, *A course in functional analysis.* Springer Science & Business Media, 2013, vol. 96. 10

[4] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, "Gaussian processes and kernel methods: A review on connections and equivalences", *arXiv preprint arXiv:1807.02582*, 2018. 18

[5] D. J. McDonald. (). Topics in mathematical statistics: Machine learning, [Online]. Available: `http://mypage.iu.edu/~dajmcdon//teaching/2014spring/s682/lectures/lec19a.pdf`. (accessed: 05.09.2014). 7

[6] C. E. Rasmussen, "Gaussian processes in machine learning", in *Summer School on Machine Learning*, Springer, 2003, pp. 63–71. 16

[7] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem", in *International conference on computational learning theory*, Springer, 2001, pp. 416–426. 11

[8] I. Steinwart and A. Christmann, *Support vector machines.* Springer Science & Business Media, 2008. 4

[9] G. Wahba, *Spline models for observational data.* Siam, 1990, vol. 59. 4, 6

[10] H. Yaozhong, *Analysis on Gaussian spaces.* World Scientific, 2016. 7

Appendix A

Gaussian random variable

Suppose $\mathbf{X} = [X_1, X_2, \ldots, X_p]^T \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ and its density function is given by:

$$\boldsymbol{\phi(x_1, \ldots, x_p) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathrm{x} - \mu)^T \Sigma^{-1}(\mathrm{x} - \mu)\}} \ .$$

We partition the vector $\mathbf{X}$ as follows:

$$\boldsymbol{\mathrm{X}^T = [X_1, X_2, \ldots, X_p] =}$$

$$\boldsymbol{[[X_1, X_2, \ldots, X_k], [X_{k+1}, X_{k+2} \ldots, X_p]] \equiv [\mathrm{X}_1^T, \mathrm{X}_2^T]}$$

Partitioning the mean vector and variance-covariance matrix in the same way,

$$\boldsymbol{\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \ ; \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}},$$

we have

$$\mathbf{X_1 \sim MVN(\boldsymbol{\mu_1}, \Sigma_{11}) \ ; \ X_2 \sim MVN(\boldsymbol{\mu_2}, \Sigma_{22})} \ .$$

Note that $\mathbf{X_1}$ and $\mathbf{X_2}$ are statistically independent vectors if and only if $\boldsymbol{\Sigma_{12} = 0_{k \times p}}$, the $\boldsymbol{k \times p}$ zero matrix.

Assume they are not statistically independent, then the conditional density of $\mathbf{X_1}$ given $\mathbf{X_2 = x_2}$; it is multivariate normal with mean vector of order $\boldsymbol{k \times 1}$:

$$\boldsymbol{\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathrm{x}_2 - \mu_2)} \ ,$$

and variance-covariance matrix of order $\boldsymbol{k \times k}$:

$$\boldsymbol{\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T} \ .$$

Proof. We introduce the auxiliary variable $\mathbf{z} = \mathbf{x_1} + \mathbf{A}\mathbf{x_2}$, where $\mathbf{A} = -\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}$. Note that:

$$\begin{aligned}
\mathbf{cov(z, x_2)} &= \mathbf{cov(x_1, x_2) + cov(Ax_2, x_2)} \\
&= \boldsymbol{\Sigma}_{\mathbf{12}} + \mathbf{Avar(x_2)} \\
&= \boldsymbol{\Sigma}_{\mathbf{12}} - \boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{\mathbf{22}} \\
&= \mathbf{0}
\end{aligned}$$

Hence $\boldsymbol{z}$ and $\boldsymbol{x_2}$ are uncorrelated and, since they are jointly Gaussian, they are independent.Then we compute:

$$\begin{aligned}
\boldsymbol{E}(\mathbf{x_1}|\mathbf{x_2}) &= \boldsymbol{E}(\mathbf{z - Ax_2}|\mathbf{x_2}) \\
&= \boldsymbol{E}(\mathbf{z}|\mathbf{x_2}) - \boldsymbol{E}(\mathbf{Ax_2}|\mathbf{x_2}) \\
&= \boldsymbol{E}(\mathbf{z}) - \mathbf{Ax_2} \\
&= \boldsymbol{\mu_1} + \mathbf{A}\boldsymbol{\mu_2} - \mathbf{Ax_2} \\
&= \boldsymbol{\mu_1} + \mathbf{A}(\boldsymbol{\mu_2} - \mathbf{x_2}) \\
&= \boldsymbol{\mu_1} + \boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}(\mathbf{x_2} - \boldsymbol{\mu_2})
\end{aligned}$$

For the variance part, first note that:

$$\begin{aligned}
\mathbf{var(x_1|x_2)} &= \mathbf{var(z - Ax_2|x_2)} \\
&= \mathbf{var(z|x_2) + var(Ax_2|x_2) - Acov(z, -x_2) - cov(z, -x_2)A}^{\boldsymbol{T}} \\
&= \mathbf{var(z|x_2)} \\
&= \mathbf{var(z)}
\end{aligned}$$

Hence we have:

$$\begin{aligned}
\mathbf{var(x_1|x_2)} = \mathbf{var(z)} &= \mathbf{var(x_1 + Ax_2)} \\
&= \mathbf{var(x_1) + Avar(x_2)A}^{\boldsymbol{T}} + \mathbf{Acov(x_1, x_2) + cov(x_2, x_1)A}^{\boldsymbol{T}} \\
&= \boldsymbol{\Sigma}_{\mathbf{11}} + \boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{\mathbf{22}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{\mathbf{21}} - \mathbf{2}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{\mathbf{21}} \\
&= \boldsymbol{\Sigma}_{\mathbf{11}} + \boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{\mathbf{21}} - \mathbf{2}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{\mathbf{21}} \\
&= \boldsymbol{\Sigma}_{\mathbf{11}} - \boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{\mathbf{21}}
\end{aligned}$$

□

It is worth to mention that the conditional variance is independent of $\boldsymbol{x_2}$; it depends only on the fixed parameters from $\boldsymbol{\Sigma}$.

Appendix B

Minimum Variance Estimate

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two jointly distributed random vectors. Then the minimum variance estimator $\hat{\boldsymbol{X}}$ of $\boldsymbol{X}$ in terms of $\boldsymbol{Y}$ is given by:

$$\hat{\boldsymbol{X}} = \mathbb{E}(\boldsymbol{X}|\boldsymbol{Y}).$$

When $\boldsymbol{Y}$ is interpreted as the prior and $\boldsymbol{X}$ is the observations, we say that the posterior mean minimizes the (error) variance.

Proof. We define $\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}]$ the conditional mean of $\boldsymbol{X}$ given that $\boldsymbol{Y} = \boldsymbol{y}$. Then we show that $\hat{\boldsymbol{x}}$ is a minimum variance estimate:

$$\begin{aligned}\mathbb{E}(||\boldsymbol{X} - \boldsymbol{z}||^{\boldsymbol{2}}|\boldsymbol{Y} = \boldsymbol{y}) &= \mathbb{E}(\boldsymbol{X^T X}|\boldsymbol{Y} = \boldsymbol{y}) - \boldsymbol{2z^T}\mathbb{E}(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}) + \boldsymbol{z^T z} \\ &= \mathbb{E}(\boldsymbol{X^T X}|\boldsymbol{Y} = \boldsymbol{y}) + ||\boldsymbol{z} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y})||^{\boldsymbol{2}} - ||\mathbb{E}(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y})||^{\boldsymbol{2}} \\ &\geq \mathbb{E}(\boldsymbol{X^T X}|\boldsymbol{Y} = \boldsymbol{y}) - ||\mathbb{E}(\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y})||^{\boldsymbol{2}}\end{aligned} \tag{B.1}$$

The minimum is attained when $\boldsymbol{z} = \hat{\boldsymbol{x}}$. Now assume $\boldsymbol{Z}$ (as a function of $\boldsymbol{Y}$) is an estimator of $\boldsymbol{X}$, then:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}(||\boldsymbol{X} - \boldsymbol{Z}(\boldsymbol{Y})||^{\boldsymbol{2}}) &= \mathbb{E}_{\boldsymbol{Y}}(\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}(||\boldsymbol{X} - \boldsymbol{Z}(\boldsymbol{Y})||^{\boldsymbol{2}}|\boldsymbol{Y} = \boldsymbol{y})) \\ &= \mathbb{E}_{\boldsymbol{Y}}(\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}(||\boldsymbol{X} - \boldsymbol{Z}(\boldsymbol{y})||^{\boldsymbol{2}}|\boldsymbol{Y} = \boldsymbol{y})) \\ &\geq \mathbb{E}_{\boldsymbol{Y}}(\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}(||\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{y})||^{\boldsymbol{2}}|\boldsymbol{Y} = \boldsymbol{y})) \qquad \text{(B.2)} \\ &= \mathbb{E}_{\boldsymbol{Y}}(\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}(||\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})||^{\boldsymbol{2}}|\boldsymbol{Y} = \boldsymbol{y})) \\ &= \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}(||\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})||^{\boldsymbol{2}})\end{aligned}$$

And the result follows.

□

Appendix C

Lemmas

Suppose that $\boldsymbol{g}$ and $\boldsymbol{h}$ are jointly Gaussian random vectors. Then the maximum of the log-likelihood of $\boldsymbol{h}|\boldsymbol{g}$ does not depend on $\boldsymbol{g}$. To see that, we calculate:

$$\boldsymbol{-2\log p(h|g) = \log\det[2\pi\mathrm{cov}(h,h|g)]+[h-\mathbb{E}(h|g)]^T\mathrm{cov}(h,h|g)^{-1}[h-\mathbb{E}(h|g)]}$$

And that quantity is clearly minimized at $\boldsymbol{h = \mathbb{E}(h|g)}$. From the basic property of conditional Gaussian distribution we see that

$$\begin{aligned}\boldsymbol{\max_h \log p(h|g)} &\boldsymbol{= -\log\det[2\pi\mathrm{cov}(h,h|g)]/2} \\ &\boldsymbol{= -\log\det[2\pi(\mathrm{cov}(h,h) - \mathrm{cov}(h,g)\mathrm{cov}(g,g)^{-1}\mathrm{cov}(g,h))]/2}\end{aligned} \tag{C.1}$$

Since the covariances in the last equality are from the fixed parameter of the joint Gaussian distribution of $\boldsymbol{(g,h)}$, the result follows.

Appendix D

Code

Estimating the regularization parameter by using 3-fold cross-validation:

```
rm(list = ls()) # clear the memory
install.packages("kernlab") # for computing the kernal matrix
install.packages("rmutil") # for generating Laplace noise
install.packages("ggplot2") # for ploting
library(ggplot2)
library(kernlab)
library(rmutil)
set.seed(2)
m<-66
xx<-(0:65)/66 # location "x" is evenly spaced on unit interval
zz<-exp(sin(8*xx)) # the example funtion
yy<-zz+rnorm(m, 0, 0.3) # Gaussian noise case

CVL2<-function(gamma,xx,yy,xtest,ytest) #Cross-validation function
{
kernel<-function(x,y){((x +1)*(y +1)*min(x +1, y +1)/2-(min(x +1, y +1))^3/6)}
Kbar<-kernelMatrix(kernel,xx)
fc<-function(c,ga,x,y)
{sum((y-c%*%kernelMatrix(kernel,x,x))^2)+ga*c%*%Kbar%*%c} # applying the square loss function
gr<-function(c,ga,x,y){Kbar%*%c-y+ga*c}
# the gradient, to improve the perforamnce of optimization

result<-optim(par=rep(0,22), fc,gr,ga=gamma, x=xx, y=yy,method = "BFGS")
c.hat<-as.vector(result$par) # estimates of the coefficients
q<-sum((ytest-c.hat%*%kernelMatrix(kernel,xtest,xtest))^2)
+gamma*c.hat%*%kernelMatrix(kernel,xtest,xtest)%*%c.hat # calculating the errors
return(q)
}

index1<-3*(1:round(m/3)) # creating 3-fold cross-validation subsamples
x1<-xx[index1]
y1<-yy[index1]

index2<-3*(1:round(m/3))-1
x2<-xx[index2]
y2<-yy[index2]

index3<-3*(1:round(m/3))-2
x3<-xx[index3]
y3<-yy[index3]

CV3foldL2<-function(gamma) #validating on test sets
  {
```

```
  CVL2(gamma,x1,y1,x2,y2)+CVL2(gamma,x1,y1,x3,y3)
  +CVL2(gamma,x2,y2,x1,y1)+CVL2(gamma,x2,y2,x3,y3)
  +CVL2(gamma,x3,y3,x1,y1)+CVL2(gamma,x3,y3,x3,y3)
}
gamma<-(16:4)/4
plot(gamma,mapply(CV3foldL2,10^(-gamma)),
     main = "CV error vs -log(gamma)",
     xlab = "-log(gamma) (base 10)",
     ylab = "CV error")
# Hence we choose the value of gamma=10^-3.5
```

An example of our simulated case:

```
kernel<-function(x,y){((x +1)*(y +1)*min(x +1, y +1)/2-(min(x +1, y +1))^3/6)}
Kbar<-kernelMatrix(kernel,xx)
fc<-function(c,ga,x,y)
{sum((y-c%*%kernelMatrix(kernel,x,x))^2)+ga*c%*%Kbar%*%c}
gamma<-10^-3.5
gr<-function(c,ga,x,y)
{Kbar%*%c-y+ga*c}
result<-optim(par=rep(0,m), fn=fc,gr,ga=gamma, x=xx, y=yy,method = "BFGS")
y2<-as.vector(result$par %*%Kbar)
df <- data.frame(xx,zz,y2)
g <- ggplot(df, aes(xx))+geom_line(aes(y=zz), colour="red")+geom_line(aes(y=y2), colour="green")
g
```

Montle Carlo simulation of the L-1 noise model with perturbed data:

```
rm(list = ls()) # clear the memory

library(ggplot2)
library(kernlab)
library(rmutil)
m<-66
xx<-(0:65)/66 # location "x" is evenly spaced on unit interval
zz0<-exp(sin(8*xx)) # the example funtion
kernel<-function(x,y){((x +1)*(y +1)*min(x +1, y +1)/2-(min(x +1, y +1))^3/6)}
Kbar<-kernelMatrix(kernel,xx)
fc<-function(c,ga,x,y)
{sum(abs(y-kernelMatrix(kernel,x,x)%*%c))+ga*c%*%Kbar%*%c}
gr<-function(c,ga,x,y)
{-kernelMatrix(kernel,x,x)%*%sign(y-kernelMatrix(kernel,x,x)%*%c)+2*ga*Kbar%*%c}
gamma<-10^-2.8


w4<-rep(0,100)
for (j in 1:100)
{
  zz<-rep(0,m) # perturbing the data
  for (i in 1:m)
    {
    zz[i]<- zz0[i]+3*zz0[i]*rbinom(1, 1, 0.1)*((2*rbinom(1, 1, 0.5))-1)
    }
  yy<-zz+rlaplace(m, 0, s=0.3/sqrt(2)) # Laplace noise case

  result<-optim(par=rep(0,m), fn=fc,gr,ga=gamma, x=xx, y=yy,method = "BFGS")
  y2<-as.vector(result$par %*%Kbar)
  w4[j]<-sqrt(sum((zz0-y2)^2))}
```

```

boxplot(w4)
```