UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

BANDLIMITED FUNCTIONS IN MACHINE LEARNING

TESIS PARA OPTAR AL GRADO DE MAGÍSTER
CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

CRISTOBAL LUIS VALEZUELA MUSURA

PROFESOR GUÍA:
FELIPE TOBAR HENRIQUEZ

MIEMBROS DE LA COMISIÓN:
SEBASTIÁN DONOSO
JORGE SILVA SÁNCHEZ
PABLO HUIJSE HEISE

Este trabajo ha sido parcialmente financiado por Fondecyt-Iniciación # 11171165: "On the relationship between Gaussian process regression and spectral estimation" y CMM ANID PIA AFB170001 Centro de Modelamiento Matemático

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE
MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS E INGENIE
POR: CRISTOBAL LUIS VALEZUELA MUSURA
FECHA: 2021
PROF. GUÍA: FELIPE TOBAR HENRIQUEZ

BANDLIMITED FUNCTIONS IN MACHINE LEARNING

En el presente trabajo se realizó un resumen de la teoría de procesamiento de señales junto con la de los procesos gaussianos y su relación con los espacios RKHS, con un enfoque al caso donde los puntos son sampleados de manera irregular. Se presentan aplicaciones de la teoría, las cuales hacen uso del kernel sinc, tanto para optimizacion de esquemas de sampleo basados en optimización bayesiana, regresión no paramétrica de funciones y filtrado de señales. Finalmente, se estudia un ejemplo aplicado a "compressed sensing", dando pie a trabajos futuros.

ABSTRACT OF THE REPORT TO QUALIFY TO THE DEGREE OF
MASTER OF SCIENCE IN ENGINEERING, MENTION APPLIED MATHEMATICS
BY: CRISTOBAL LUIS VALEZUELA MUSURA
DATE: 2021
GUIDE: FELIPE TOBAR HENRIQUEZ

BANDLIMITED FUNCTIONS IN MACHINE LEARNING

There has been an increasing interest in the study of Gaussian Processes in the machine learning literature, both for the increase in computational power in the recent years and also the development of new algorithms that reduce the computational complexity to train a Gaussian Process model. In particular, there has been an special interest in its connections with the classical signal processing literature, and its applications in studying the spectrum of time series, more specifically band-limited time series, as it is the main object of interest in signal processing.

One of the advantages of using Gaussian process in signal processing is that its probabilistic nature can naturally handle problems regarding irregular sampled points and, not only that, it allow us to study the phenomenon of irregular sampling to our advantage, as, theoretically, when sampling irregularly it is possible to obtain better results by reducing the alias error that regular sampling is often associated with.

In the first chapter we will review the basic theory regarding Fourier analysis, classical sampling theory and Gaussian processes, as well as the connections between the two, including some results which will be useful in order to understand said connections. In the second chapter, we will discuss the Nyquist-Shannon frequency and irregular sampling in order to understand the possible problems as well as the advantages given by it, in particular, we will study some known sampling schemes as well as some theoretic results.

Lastly, we will review some applications regarding Gaussian Processes and band-limited kernels. First we will develop a filtering model by developing a generative model, which can handle irregular sampled points. Next, it will be shown how different irregular sampling techniques actually work on practice using Gaussian Process regression, and what can be gained by them. Following this, we present an algorithm for finding an optimal sampling scheme of a signal based on Bayesian optimization. At the end, an application using band-limited functions for a compressed sensing problem is presented, which is what will be, hopefully, the next step of this research.

It was the detour that was our shortest path

Agradecimientos

Agradecimientos a CMM ANID PIA AFB170001.

Decir que este trabajo no hubiese sido posible sólo sería decir poco. No puedo decir que terminó como lo esperaba, pero si como necesitaba que fuese. Pues los años que pasé en la universidad siempre contaron con un punto ciego, uno que nunca quise ver y que tuve que enfrentar durante todo este proceso.

Es por eso que debo agradecer en primera instancia a mis padres, que fueron quienes siempre estuvieron allí y se que terminar este trabajo los llenará de orgullo, a la Tere que siempre se preocupaba de que estuviera bien, a todos los funcionarios del DIM (Eterin, Silvia, Don Oscar) que siempre estaban dispuestos a ayudar, y nada más que ayudar. Por supuesto también a mi profesor guía, Felipe Tobar, que me enseñó como aprovechar el potencial que entrega la intuición matemática y a como encontrar el foco en cada problema que va apareciendo.

Quiero agradecer a todos los alumnos que tuve como profesor auxiliar en distintas ocasiones, gracias por mostrarme lo apasionante que es compartir las matemáticas con otras personas y de enseñarme muchísimo más de vuelta.

Por último no quiero dejar de mencionar a los alumnos Cristobal Vicuña, Joaquín Castillo y Ricardo Aldana que, de no ser por su trágica partida, hoy serían ingenieros. Su memoria sigue presente en mi corazón y me recuerda lo afortuando que es haber llegado acá, por más difícil, decepcionante y lleno de contratiempos haya sido.

Contents

Introduction

One of the main problems in signal-processing is the signal reconstruction problem, which consists in reconstructing an analog continuous signal from a discrete number of samples. To correctly define such problem, an adequate space must be defined for the target analog function $f(t)$. Famously, the Nyquist-Shannon sampling theorem gives an explicit solution to the reconstruction problem from a set of sampling points $\{(f(t_n), t_n)\}_{n\in\mathbb{N}}$ as long as the signal spectrum is bounded, the sampling times evenly-spaced and the sampling frequency is at least two times as large as the highest frequency present in the spectrum of the signal. Thus, the bounded spectrum condition, naturally defines the function space, the space of band-limited functions or the Paley-Wiener space.

The previous theorem is, nonetheless, of limited use in practical applicaitons for two reasons. The first one being that in practical applications only a limited (i.e finite) number of samples are usually available, meanwhile the formula given by the theorem requires an infinite number of samples (And it can be shown that truncation errors cannot be disregarded easily). The second limitation is associated with the evenly-spaced sampling times condition, which in practice may not be attainable or even desired. It becomes evident the necessity to study the signal reconstruction problem when irregular sampling schemes are present. A Bayesian framework is proposed for this same reason.

Bayesian frameworks are utilized to model the trade-off between the reward of an informative view and the cost of missing a target [8] by using a prior and a likelihood function. In the signal-processing context, expecting the function to live in the Paley-Wiener space works, naturally, as an informative view or *prior* of the model. An appropiate *likelihood* function is defined in order to evaluate such trade-off. The framework is properly defined this way, and once some samples are observed, the main goal becomes to evaluate the *posterior* distribution over the set of possible candidates which fit the sampled points the best (In some correctly defined metric). Such approach turns out to outperform the limitations previously mentioned and it possseses some other advantages, such as having not just a point-estimate for the solution, but a distribution over possible answers, as well as giving reliable estimates of its own uncertainty.

The bayesian framework utilized is based on Gaussian Process [16] which is a non-parametric Bayesian framework utilized to estimate functions from a finite set of points. The likelihood and prior functions are completely determined by a mean function (Usually assumed to be zero) and a *kernel* function, which must be positive-definite. By defining a kernel, the functions generated by a GP are restricted to a certain space. In fact it can be shown

that if the kernel function is bandlimited, then the function space will be restricted to the space of band-limited functions (With possibly some modifications to its norm). This result can be used to prove a theorem which generalizes the classical Shannon-Nyquist theorem.

In the final applications of the chapter, the GP framwork is used to solve a particular isntance of the signal reconstruction problem, the filtering problem, were samples are corrupted by frequencies outside a desired range. A generative model, which views the corruption of the signal as a mix between two Gaussian Processes generated by two different kernel function, is proposed in this setting. Even further, given a continuous band-limited signal $f(t)$, bayesian optimization is utilized in this same context to find the minimal set of points which can encode the signal (Where the decoding process is done by performing GP regression on the proposed sequence).

Before advancing further, it is worth mentioning previous methods which have been used for signal reconstruction and periodic analysis in the non-uniform setting. Most notably, the Lomb-Scargle periodogram [14] is a Fourier method for finding periodic components in such setting. Many iterative approaches have been considered as well [12] , whose main idea is to apply successive approximations, which are represented by some linear operator, until convergence. Notably, the POCS (Projection onto convex sets) method considers the projection onto a convex set of some defined Hilbert space as such linear operator [27]. Least squares method involving some special basis such as the PSWF's have also been proposed in [28]. Pseudo-inverse matrix reconstruction, which is efficient when the number of sampling points and the size of the spectrum are small [22], considers the non-uniform sampling reconstruction problem from an algebraic point of view, based on the solution of systems of linear equations. Other Bayesian approaches have been proposed [7], as well as other works which also involve Gaussian Processes, such as models for periodic light curves in the astronomical setting [26], by proposing a kernel function [2] and our own previous work on filtering techiques [24]. Finally, compressive sensing (CS) is the process of creating an undetermined system of equations where there are far fewer equations than unkowns, but the solution is known to be sparse in some other basis, so reconstruction is performed in a way to maintain such sparseness.

The outline of this work is presented as the following: Chapter 1 reviews the background regarding the Fourier transform, band-limited functions, Gaussian processes and Kernel Hilbert Spaces (Which is just the frequentist approach to GP Regression) and at the end results which relates the three are shown. Chapter 2 reviews different sampling schemes, and their effect on the interpolation formula, for example some of them have an aliasing suppressing effect. Chapter 3 contains the main applications of the proposed work, such as the GPLP model, which aims to solve the filtering problem in a generative model approach as previously mentioned, study the effect of different sampling schemes and a Bayesian optimization approach for obtaining a sequence in which to encode a signal, also as mentioned. The final experiment distances itself from the GP approach, and illustrates an instance of compressive sensing in the multi-band signal restricted context, and it constitutes an example on how prior knowledge can reduce significantly the number of samples needed for reconstruction, way beyond the Nyquist limit, and it will be, hopefully, inspiration for future work. Finally, results are discussed.

Publications

Part of the work which composes this thesis has been previously presented in:

- C.Valenzuela, F.Tobar, "Low-pass Filtering as Bayesian Inference," *Proceedings of the IEEE International Conference on Multimedia*, pp. 3367-3371, 2019.

Main contributions

The main contributions of this thesis are the following:

1. Study the connections between the classical signal processing theory and the Gaussian process theory, and see how the latter can expand the former. Corollary 1.7.1 is clear example of this.

2. Design and implement a novel generative model approach for the signal filtering problem, with features such as different window shapes, priori kernels and the possibility to either have a low-pass filter as well as a high pass filter.

3. Study how different irregular sampling schemes influence in signal reconstruction, in paticular it will be shown that the jittering sampling scheme can give satisfactory results in sub-Nyquist instances.

4. Use the Gaussian Process framework to obtain a set of sampling points which optimally reduces the MSE of the original signal and its GP-interpolation along a desired interval (Although an hyperparameter which controls the trade-off between exploration and explotaition must be defined in advance). The distribution of such a set is also studied.

Chapter 1

Preliminaries

1.1 The main problem: Signal Reconstruction

One of the main problems that will be tackled in this work is the signal reconstruction problem.

In order to give a bit of context of our work, as there are many algorithms available to do similar work [25][15], the main difference is that the following work uses a Bayesian approach, specifically in the context of Gaussian Processes, to define both the problem and a generative model for it. Subsequent results are more in line with how sampling affects our results. [23]

Consider the signal reconstruction problem, pictured in 1.1, which is based on some signal or function over time $f(t)$ and some samples are retrieved from it at times $t_0, ..., t_M$, such that the values $f(t_0), ..., f(t_M)$ are observed.

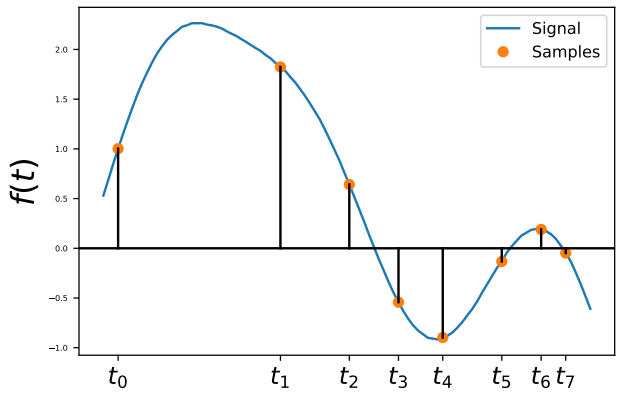

Figure 1.1: Signal $f(t)$ observed at times $t_0, ..., t_7$.

The questions to be answered in this setting are:

- How many samples are required to have an accurate approximation to the whole func-

tion $f(t)$ and, if possible, what sampling scheme will select these points optimally?

- How can an algorithm which solves the problem be implemented?

These questions require, of course, to impose some restrictions on the signal $f(t)$, otherwise the problem would be ill-posed. SO it will be assumed that $f(t)$ lives in some kind of space, which implicitly imposes an smoothness condition, or some kind of simple implicit structure. In engineering applications the most common way to impose structure is via making assumptions of the spectrum of the signal $f(t)$, in other words by making assumptions of its Fourier transform.

1.2 The Fourier Transform

In order to understand the following sections, we will step back a little and review the subject of Fourier analysis of continuous signals, with signals refering to any continuous function of time $f(t)$.

Consider a signal $f(t)$, which is just a function of time in a determined space (i.e L^1 to assure convergence) then its **Fourier transform** is given by:

$$\mathcal{F}(f)(\xi) = \hat{f}(\xi) = \int_{-\infty}^{+\infty} f(t)e^{-2\pi i\xi t}dt.$$

The Fourier transform is said to be defined in the **frequency** domain. Intuitively, this says that if, for example, the Fourier transform is concentrated in values near 0, then the signal presents slow changes through time, if, on the contrary, the Fourier transform is concentrated in values far from 0, then the signal presents quick oscilations through time.

Of upmost importance in this sense is the Fourier inversion theorem, which states that for many types of functions it is possible to recover a function from its Fourier transform. In the present work, square integrable (ie. $L_2(\mathbb{R})$) functions will be considered, were the theorem holds true via a density argument. This inverse relationship is given by:

$$f(t) = \int_{-\infty}^{\infty} \hat{f}(\xi)e^{2\pi i\xi t}d\xi.$$

Note that this theorem can be restated as:

Theorem 1.2.1 (Fourier Inversion Theorem) Let $f(t) \in L^2$ be a function defined on the real line, then we have

$$f(t) = \mathcal{F}^{-1}(\hat{f})(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{2\pi i(t-x)\xi} f(t)dtd\xi,$$

and, furthermore, if f is real valued, which will be our case on the rest of this work, by equaling the real part from both sides of the equation above, we obtain the following equation:

$$f(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cos(2\pi(t-x)\xi) f(t)dtd\xi. \quad \blacksquare$$

The functions f and $\hat{f}$ are said to be a **Fourier pair**.

The most notable and the most important properties of the Fourier transform (Which we will be using later) are summarized on Table 1.1.

Property name	**Time-domain**	**Frequency Domain**
Time-shift	$f(t-t_0)$	$\hat{f}(\xi)e^{-2\pi i t_0 \xi}$
Scaling	$f(at)$	$\frac{1}{a}\hat{f}\left(\frac{\xi}{a}\right)$
Convolution Property	$(f \star g)(t)$	$(\hat{f} \cdot \hat{g})(\xi)$
Sinc Transform	$\text{sinc}(at)$	$\frac{1}{a}\text{rect}\left(\frac{\xi}{a}\right)$
Delta Train Transform	$\sum_{n=-\infty}^{\infty} \delta(t-nT)$	$\frac{1}{T}\sum_{k=-\infty}^{\infty} \delta\left(\xi - \frac{k}{T}\right)$

Table 1.1: Fourier transform most notable properties and known transforms

Finally we announce the famous Plancherel theorem (A simplified version, as we will not be working with complex numbers)

Theorem 1.2.2 (Plancherel Theorem) If $f(t), g(t)$ are functions on the real line and $\hat{f}(\xi), \hat{g}(\xi)$ are their corresponding frequency spectrums, then we have

$$\int_{-\infty}^{\infty} f(t)g(t)dt = \int_{-\infty}^{\infty} \hat{f}(\xi)\hat{g}(\xi)d\xi. \quad \blacksquare$$

1.3 Bandlimited Functions

We define the space of **band-limited functions**, also known as Paley-Wiener spaces, as the subspace of all $L_1(\mathbb{R})$ functions f whose Fourier transform is supported in a certain compact subset of $\Omega \subset \mathbb{R}$. More formally:

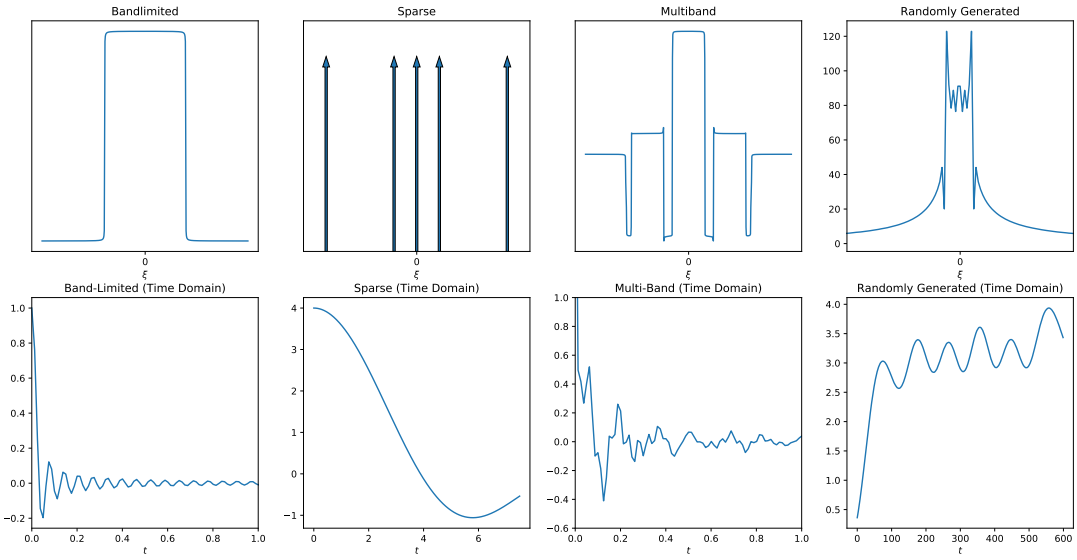

Figure 1.2: Examples of diferent bandlimited functions on time and frequency domain.

$$PW_\Omega(\mathbb{R}) = \{f \in L_1(\mathbb{R}) : \hat{f}(\xi) = 0 \text{ a.e. for } \xi \in \Omega^c\}.$$

Examples of band-limited functions are given in Fig. 1.2. These examples not only work as an illustration of band-limited functions, but they also show how one can add more structure in order to obtain even sparser representations of signals in the frequency domain.

Having band-limited functions defined this way, one must be asking, why are band-limited functions important or even interesting?. The answer is the Shannon-Nyquist sampling theorem, which states sufficient conditions in which the signal reconstruction problem can be solved, as long as the signal is band-limited.

1.4 Sinc Kernel and the Sampling Theorem

The (normalized) sinc function is defined simply by:

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}. \tag{1.1}$$

The normalized sinc function is the Fourier transform of the rectangular function with no scaling.

It is used in the concept of reconstructing a continuous bandlimited signal from uniformly spaced samples of that signal by using the next famous theorem, the Nyquist-Shannon sampling theorem.

Theorem 1.4.1 (Nyquist-Shannon sampling theorem) If a function $x(t)$ contains no frequencies higher than W (Bandlimited in the interval $[-W, W]$), then it is completely determined

by giving its ordinates at points evenly spaced at $1/2W$ apart by the formula

$$x(t) = \sum_{n\in\mathbb{Z}} x\left(\frac{n}{2W}\right) \operatorname{sinc}(2Wt - n).$$

Moreover, the convergence of the righ-hand side series is uniform.

Proof. Let

$$x(t) = \int_{-\infty}^{+\infty} X(f)e^{2\pi ift}df = \int_{-W}^{W} X(f)e^{2\pi ft}df, \tag{1.2}$$

where $X(f)$ is the Fourier transform of $x(t)$. Next $X(f)$ is expanded as a Fourier series on the interval $(-W, W)$ as:

$$X(f) = \sum_{n\in\mathbb{Z}} c_n e^{-2\pi if\frac{n}{2W}}, \tag{1.3}$$

and the coefficents c_n are obtained by Fourier analysis:

$$c_n = \frac{1}{2W}\int_{-W}^{W} X(f)e^{2\pi if\frac{n}{2W}}df = \frac{1}{2W}x\left(\frac{n}{2W}\right). \tag{1.4}$$

The last equality is consequence of Eqn. (1.2). Finally, substituing Eqn. (1.3) into Eqn. (1.2), and calculating the integral we obtain the result. □

The regular sampling theorem combines two aspects:

(a) Any band-limited function can be completely reconstructed from its sampled values over a sufficiently fine lattice by means of a simple series expansion with the sampled values as coefficients.

(b) Any band-limited function can be expanded into a series with translates of the sinc function as building blocks.

Unfortunately, this theory is asymptotic, which means that it requires infinite samples over the entire real line to build an exact interpolation even at a single point. When a finite number of samples are considered to construct the interpolation, it turns out that in order to reach a level of error no greater than ε, then $O(\frac{1}{\varepsilon})$ samples are required.

To illustrate this, consider the following example. Let M be an arbitrary big integer and let K be the set of even integers on the interval $[0, 2M]$.

Define the following function

$$y(t) = \sum_{k\in K} \frac{\sin(\pi\cdot(t-k))}{\pi\cdot(t-k)}.$$

The Fourier transform of this function is supported on the box $[\frac{-1}{2}, \frac{1}{2}]$ so the Nyquist rate is 1 and the Shannon interpolation formula reconstructs $y(t)$ as the following sum of sinc functions:

$$y(t) = \sum_{k=-\infty}^{\infty} y(k) \frac{\sin(\pi \dot{(t-k)})}{\pi \dot{(t-k)}}.$$

Naming $\tilde{y}$ as the truncated sinc-interpolation formula for $k < 2M$ and $k \notin K$ (in other words, uneven k) , then we have that $\tilde{y} \equiv 0$ on $[0, 1]$. So, we calculate the error

$$\begin{aligned} ||y - \tilde{y}||^2_{[0,1]} &= ||y||^2_{[0,1]} \\ &= \int_0^1 \left(\sum_{k \in K} \frac{\sin(\pi \cdot (t-k))}{\pi \cdot (t-k)} \right)^2 \\ &\leq \frac{1}{\mathcal{O}(M^2)} \int_0^1 \left(\sum_{k \in K} \sin(\pi t) \right)^2 dt \\ &= \mathcal{O}(1). \end{aligned}$$

This example shows that you can always construct adversarial examples to the Shannon interpolation formula. This problem was addressed by the seminal work of Slepian [20], who presented an interesting set of basis functions for interpolating bandlimited functions, when only a finite number of samples are available, they are the so called Prolate Sphreoidal Wave Functions.

1.5 Prolate Spheroidal Wave Functions

Prolate Sphreoidal Wave Functions (PSWF) constitute an orthogonal base of $L^2(-T, T)$ [1][9][11]. Thus they can serve as an interpolator on any compact interval of $\mathbb{R}$ as an alternative choice which can enjoy spectral accuracy. They also constitute an orthonormal basis of the space of bandlimited functions on the real line, just as the traslates of the sinc function.

The idea behind them is to find the most energy concentrated signals in both fixed time and frequency domains at the same time [20]. That problem can be mathematically formulated as the solutions of the integral equation

$$\int_{-\tau}^{\tau} \varphi_n(s) \frac{\sin(\sigma(t-s))}{\pi(t-s)} ds = \alpha_n \varphi_n(t), \tag{1.5}$$

where $[-\sigma, \sigma]$ and $[-\tau, \tau]$ are the fixed time and frequency domains, respectively.

Some interesting properties of the PSWF include:

- Equation Eqn. 1.5 has solutions for certain real values α_n of α and can be ordered as

  $$1 > \alpha_0 > \alpha_1 > ... > \alpha_n$$

  It must be noted that, as an intuition remark, these values often drop drastically from being close to 1 to being close to 0 [20]

- The functions $\{\varphi_n\}_{n=0}^{\infty}$ form a dual orthogonal set both in the interval $(-\infty, \infty)$ and $(-\tau, \tau)$. This is known as double orthogonality.

$$\int_{-\tau}^{\tau} \varphi_m(t)\varphi_n(t)dt = \alpha_n \delta_{mn}$$
$$\int_{-\infty}^{\infty} \varphi_m(t)\varphi_n(t)dt = \delta_{mn},$$

  where $\delta_{mn} = 1$ if $m = n$ and 0 otherwise.
- It is complete, in other words, any σ-bandlimited function $f(t)$ can be expressed as

$$f(t) = \sum_{n=0}^{\infty} c_n \varphi_n(t).$$

The last property combined with the sinc interpolation formula means that we can write any function bandlimited in $[-W, W]$ as

$$\begin{aligned} f(t) &= \sum_{k\in\mathbb{Z}} f(2Wt) \sum_{m=0}^{\infty} \varphi_m(2Wt)\varphi_m(n) \\ &= \sum_{m=0}^{\infty} \left[\sum_{k\in\mathbb{Z}} f(2Wt)\varphi_m(n)\right] \varphi_m(2Wt) \\ &= \sum_{m=0}^{\infty} \gamma_m \varphi_m(2Wt), \end{aligned} \tag{1.6}$$

where the above formula is just consequence of decomposing the $\text{sinc}(\cdot)$ function in the PSWF basis.

Eqn. 1.6 allows the truncation of the sinc series approximation, not by reducing the number of terms (time), which is known for not being an accurate option (at least in theory), but by limiting both frequencies and time (respectively number of basis functions to consider and a finite number of time samples). This technique has been done in signal processing before (it was the reason Slepian developed this set of functions) and in the Gaussian Process literature it is known as the Nystrom approximation of the kernel function, or reduced rank approximation [16]

Another way of realizing that the Slepian functions may be a better alternative to compute the regression is by looking at the eigenalues of both gram matrices. While the sinc-kernel gram matrix usually has some eigenvalues close to zero (consequence of the sinc function not being in L^1), the slepian pseudo matrix does not show this behaviour.

1.5.1 Priors and restrictions over Band-Limited functions

Saying that a function is band-limited can be understood, in Bayesian terms, as giving a certain prior over the Fourier structure of the signal. And there are many examples across

literature were such constraints are imposed over signals in order to obtain modified, or better, sampling theorems [6] [15].

For example, signals with a bandlimit F (and no further constraint) can be understood as having a uniform probability across frequencies on the interval $[-F, F]$. Signals are said to be multiband when its prior is the union of uniform measures of k intervals. Analogously, sparse signals are the union of k Dirac measures. Gaussian or Cauchy distributions, although not bandlimited in the strict sense, in practice they can be well approximated as being band limited [16]

We will develop this idea of having a prior over band-limited functions more in depth in the next section, which is about Gaussian Processes.

1.6 Gaussian Processes and Kernel Hilbert Spaces

Gaussian processes (GP) and Kernel Ridge Regression methods form two sides of the same coin, were the two are nonparametric based on positive definite kernel methos for the purpose of modeling nonlinear functional realtionships, were the main difference is that GPs are a Bayesian approach and kernel ridge regression is a more frequentist approach. In the following section, both techinques will be briefly reviewed and it will be shown how they are related with the classical signal processing theory, in particular their relationship to the Shannon sampling theory.

1.6.1 Definitions

Definition 1 (Positive Definite Kernel) Let X be a nonempty set. A symmetric function $k : X \times X \mapsto \mathbb{R}$ is called a positive definite kernel if, for any $n \in \mathbb{N}$, $(c_1, ..., c_n) \subset \mathbb{R}^n$ and $(x_1, ..., x_n) \subset X$,

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

In other words, k is positive semi-definite if the matrix Σ with elements $\Sigma_{ij} = k(x_i, x_j)$ is poositive semi-definite, for any size n and vector $(x_1, ..., x_n)$. Σ is usually called the **kernel matrix** or the **Gram matrix**.

Given this definition, it is possible to prove that the function $k(x, y) = \text{sinc}(x - y)$ is a positive definite kernel, although not quite easy. For this reason we announce the following theorem

Theorem 1.6.1 (Bochner's Theorem) Given a positive finite Borel measure μ on the real line $\mathbb{R}$, the Fourier transform K of μ is the continuous function

$$K(t) = \int_{\mathbb{R}} e^{-2\pi i t \xi} d\mu(\xi).$$

Then, the function $k(x, y) := K(x - y)$ is positive semi-definite. The converse, that any positive semi definite function is the Fourier transform of a finite positive Borel measure, is also true.

From the previous theorem it its obvious that the function $k(x, y) = \text{sinc}(x - y)$ is a positive semi-definite kernel.

Definition 2 (Gaussian Process) Let X be a nonempty set, $k : X \times X \mapsto \mathbb{R}$ be a positive definite kernel and $m : X \mapsto \mathbb{R}$ be any real-valued function. Then a random function $f : X \mapsto \mathbb{R}$ is said to be a Gaussian Process (GP) with mean function m and covariance kernel k, denoted by $\mathcal{GP}(m, k)$ if, for any finite set $\boldsymbol{x} = (x_1, ..., x_n) \subset X$ of any size $n \in \mathbb{N}$, the random vector

$$\boldsymbol{f} = (f(x_1), ..., f(x_n))^T \in \mathbb{R}^n$$

follows a multivariate normal distribution $\mathcal{N}(m_{\boldsymbol{x}}, k_{\boldsymbol{x}})$ where the covariance matrix and the mean vector are defined as $[k_{\boldsymbol{x}}]_{ij} = k(x_i, x_j)$ and $[m_{\boldsymbol{x}}]_i = m(x_i)$.

Next, we briefly review the basics of Gaussian process regression, also known as *kriging*. For more information, the reader can refer to [16].

The regression problem in this context, given a set of points $\{(x_i, y_i)\}_{i=1}^n$,is to find the best fitting function f such that:

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

for i.i.d. noise variables $\xi_1, ..., \xi_n \sim \mathcal{N}(0, \sigma^2 I)$

Being a Bayesian approach to regression, we first define a prior distribution over functions $\mathcal{GP}(m, k)$, for some function m and covariance kernel K, and a likelihood function defined by a probabilistic model $p(y_i|f(x_i))$, which is implictitly defined by the noise variables as:

$$\ell(f) = \prod_{i=1}^{n} \mathcal{N}(y_i|f(x_i), \sigma^2).$$

Note that the only difference of this problem with the signal reconstruction problem is the assumption of noisy observations. The solution of the regression problem is given in the next theorem

Theorem 1.6.2 (GP Regression) Given the set of points $M = \{(x_i, y_i)\}_{i=1}^n$ we have that the conditional distribution of f given X is also a Gaussian process:

$$f|M = \mathcal{GP}(\overline{m}, \overline{k}),$$

where the posteriori mean function and the posteriori covariance function are given by

$$\overline{m}(x) = m(x) + k_{xM}(\Sigma + \sigma I)^{-1}(\boldsymbol{y} - m_M), \quad x \in X;$$
$$\overline{k}(x, x') = k(x, x') - k_{Mx}(\Sigma + \sigma I)^{-1}k_{Mx'}, \quad x, x' \in X;$$

where X is the domain, m_M is a vector defined by $[m_M]_i = m(x_i)$, and k_{Mx} is a vector of functions defined by $[k_{Mx}]_i = k(\cdot - x_i)$.

It is interesting to note how the posterior covariance formula does not depend on the observations $\boldsymbol{y}$.

Now we will also introduce briefly an almost equivalent, but frequentist, approach to the same problem.

Definition 3 (Reproducing Kernel Hilbert Space (RKHS)) Let X be a nonempty set and k be a positive definite kernel on X. A Hilbert space $\mathcal{H}_k$ of functions on X equipped with an inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is called a reproducing kernel Hilbert space (RKHS) with reproducing kernel k, if the following are satisfied:

1. For all $x \in X$, we have that $k(\cdot, x) \in \mathcal{H}_k$.
2. For all $x \in X$ and for all $f \in \mathcal{H}_k$:
$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}, \quad x, y \in X.$$

The last property is known as the **reproducing property.**

The space of bandlimited functions mentioned in the earlier chapter is in fact an RKHS, with the $\text{sinc}(x)$ function as its reproducing kernel. For the sake of illustrating this fact, lets assume that $\Omega = [-a, a]$, then, defining $K_x = \frac{a}{\pi}\text{sinc}(a(\cdot - x))$ and choosing a function $f \in PW_\Omega$. The first property is obvious from the shift property of the Fourier transform. For the second property, we have that:

$$\begin{aligned}
\langle f, K_x \rangle_{L^2} &= \int_{-\infty}^{\infty} f(t)K_x(t)dt \\
&= \int_{-\infty}^{\infty} \hat{f}(\xi)\hat{K_x}(\xi)d\xi \\
&= \int_{-a}^{a} \hat{f}(\xi)e^{2\pi\xi x}d\xi \quad \text{(Plancherel + Shift property)} \\
&= \mathcal{F}^{-1}(\hat{f})(x) \\
&= f(x),
\end{aligned}$$

which is exactly the reproducing property of the kernel.

As it is important to find connections between the frequentist RKHS approach, the bayesian Gaussian Process approach and the classical theory of signal processing, the following result is presented beforehand as it will be usful later. The proof can be found in [13] and it is an spectral characterization of the RKHS space.

The inner product and the construction of this space as a Hilbert space is given by the following characterization: Given the kernel function k its RKHS $\mathcal{H}_k$ can be written as

$$\mathcal{H}_k = \left\{ f \in L_2(\mathbb{R}) \cap \mathcal{C}(\mathbb{R}) \text{ s.t. } f = \sum_{i=1}^{\infty} c_i k(\cdot, x_i) \text{ and } ||f||^2_{\mathcal{H}_k} := \sum_{i,j=1}^{\infty} c_i c_j k(x_i, x_j) < \infty \right\}.$$

Essentialy, it means that functions in an RKHS can be aproximated by shifts of the kernel function, and such expansion must be regular enough for the norm to converge.

But a more useful characterization for our purposes is given by the following theorem:

Theorem 1.6.3 (Spectral Characterization of RKHS) Let k be a positive definite kernel, defined on $\mathbb{R} \times \mathbb{R}$, and let K be such that $k(x, y) = K(x - y)$. Then the RKHS $\mathcal{H}_k$ associated with k is given by:

$$\mathcal{H}_k = \left\{ f \in L_2(\mathbb{R}) \cap \mathcal{C}(\mathbb{R}) : ||f||^2_{\mathcal{H}_k} = \int \frac{|\mathcal{F}[f](\xi)|^2}{\mathcal{F}[K](\xi)} d\xi < \infty \right\},$$

where, implicitly, the division inside the integral restricts the domain of the Fourier transform of f to the domain of the Fourier transform of K

This last statement is interesting, because it means that an adequate RKHS encapsulates the a priori space mentioned in the previous section. In particular, any bandlimited kernel gives rise to the same RKHS, the bandlimited space of functions, but it changes its inner product in a way that reflects our prior beliefs of were the spectrum of our signal is concentrated.

The problem to be solved in a RKHS, similarly to the GP Regression, is called **kernel ridge regression** which is solving

$$\overline{f} = \underset{f \in \mathcal{H}_k}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||^2_{\mathcal{H}_k},$$

where $\lambda > 0$ is a parameter. The solution is obtained by applying the representer theorem [17] which implies that the solution can be written as

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

and then solving by differentiation for α obtaining that

$$\boldsymbol{\alpha} = (\Sigma + n\lambda I)^{-1} \boldsymbol{y}.$$

Note that selecting $\lambda = \frac{\sigma^2}{n}$ the solution is equivalent to the one obtained in the GP regression.

There is a result which relates together the posterior variance from the ridge regression point of view with the GP point of view. This result also implies the Shannon-Whithaker interpolation formula and will helpful to tie everything together.

1.6.2 Posterior Variance as certain worst-case error

Consider the function $v(x) := \overline{k}(x,x)$, known as the *marginal posterior variance*, where $\overline{k}$ is the posterior covariance kernel, and the square root of it is interpreted, informally, as an margin error bar. By definition, the posterior variance is defined by the equation

$$v(x) = \mathbb{E}_{f\sim\mathcal{GP}(\bar{m},\bar{k})}[(f(x)-\bar{m}(x))^2].$$

So, in a way, you can say that $v(x)$ can be interpreted as the *average* case error at a location x. It is interesting that this same quantity can be viewed as *worst case error* from the kernel/frequentist point of view.

For simplicty, a zero-mean GP prior $f \sim \mathcal{GP}(0,k)$ is considered and define $\omega^\sigma(x)$ to be a vector-valued function defined by

$$\omega_\sigma := (k_{XX} + \sigma^2 I_n)^{-1} k_{Xx}, \quad x \in \mathcal{X}.$$

With this notation, the posterior mean function can be written as:

$$\overline{m}(x) = \sum_{i=1}^{n} \omega_i^\sigma y_i = Y^t \omega^\sigma.$$

Similarly, define the kernel k^σ by

$$k^\sigma(x,y) := k(x,y) + \sigma^2 \delta(x,y).$$

With both definitions in mind, the worst case error interpretation of the posterior variance can be stated in the following proposition.

Proposition 1.6.1 Let $\overline{k}$ be the posterior covariance function with noise variance σ^2. Then, for any $x \in \mathcal{X}$ with $x \neq x_i$ we have:

$$\sqrt{v(x)+\sigma^2} = \sup_{g\in\mathcal{H}_{k^\sigma}:||g||_{\mathcal{H}_{k^\sigma}}\leq 1} \left(g(x) - \sum_{i=1}^{n} \omega_i^\sigma g(x_i) \right).$$

The proof can be found in [13].

Corollary 1.6.1 Assume that $\sigma^2 = 0$, that $\mathcal{X} = \mathbb{R}$, that f is a banlimited function, that the x_i are uniformly sampled at the Nyquist frequency with $i \in \{1,..,n\}$. Then, for any x, we have

$$\sqrt{v(x)} \to 0 \text{ as } n \to \infty.$$

This result is interesting because asymptotic consistency results like this one usually use that the sampling points x_i become dense as n goes to infinity as part of the hypothesis . So this means that the sampling theorem is somewhat special, even in the context of GP/RKHS.

PROOF. Given x in $\mathbb{R}$, using the above proposition, it can be stated

$$v(x) = \sup_{g \in \mathcal{H}_k : ||g||_{\mathcal{H}_k} \leq 1} \left(g(x) - \sum_{i=1}^{n} \omega_i g(x_i) \right).$$

Note that $\omega_i(x) = \text{sinc}(x - x_i)$ because note that in the Nyquist sampling case $\Sigma = k_{XX} = I$ so the formula is really

$$v(x) = \sup_{g \in \mathcal{H}_k : ||g||_{\mathcal{H}_k} \leq 1} \left(g(x) - \sum_{i=1}^{n} \text{sinc}(x - x_i) g(x_i) \right).$$

Decomposing g with the Shannon interpolation formula and replacing it in the equation

$$v(x) = \sup_{g \in \mathcal{H}_k : ||g||_{\mathcal{H}_k} \leq 1} \sum_{i=n+1}^{\infty} \text{sinc}(x - x_i) g(x_i).$$

Now, bounding the inside term

$$\begin{aligned}
\sum_{i=n+1}^{\infty} \text{sinc}(x - x_i) g(x_i) &\leq \left| \sum_{i=n+1}^{\infty} \text{sinc}(x - x_i) g(x_i) \right| \\
&\leq \sum_{i=n+1}^{\infty} |\text{sinc}(x - x_i) g(x_i)| \\
&\leq \left(\sum_{i=n+1}^{\infty} g(x_i)^2 \right)^{\frac{1}{2}} \left(\sum_{i=n+1}^{\infty} \text{sinc}^2(x - x_i) \right)^{\frac{1}{2}} \\
&\leq ||g||_{\mathcal{H}_k^2} \left(\sum_{i=n+1}^{\infty} \text{sinc}^2(x - x_i) \right)^{\frac{1}{2}}.
\end{aligned}$$

So, finally

$$v(x) \leq \left(\sum_{i=n+1}^{\infty} \text{sinc}^2(x - x_i) \right)^{\frac{1}{2}},$$

and taking limits in both sides finishes the proof. □

A different proof of the same principle can be found in [23].

Chapter 2

Sampling schemes and how to work with real-world signals

2.1 Nyquist-Limit

The reults associated with the Shannon-Nyquist theorem requires precise definition of the signal on the interval $(-\infty, \infty)$. However, working with real world signals obviously require some kind of measurements in a very specific (bounded) period of time. The way in which such measurement times are selected, directly or indirectly, is called a sampling scheme.

Sampling schemes can be described by window function that spans the duration of the observation. The resulting data can be described as a point-wise product of the original, defined at all times, signal with said sampling scheme. For example, a Dirac-comb restricted to some interval can be seen as the sampling scheme that is sampling the signal at regular intervals of time.

So in reality the Fourier transform of the underlying signal is not attainable, instead we have acces to the transform of the point-wise product of the signal and the sampling scheme. Say if the sampling scheme is $\Pi(t)$ and the signal is $f(t)$ then we observe

$$g(t) = f(t)\Pi(t).$$

By aplying the Fourier transform and the convolution theorem we obtain

$$\hat{g}(\xi) = \hat{f}(\xi) \star \hat{\Pi}(\xi).$$

The convolution with $\hat{\Pi}$ is what will, in most cases, distort the signal observed spectrum. In fact the Nyquist-limits provides exactly a bound of the rate of sampling frequency in which the spectrum will not be distorted, given that the sampling is uniform and given that the signal in band-limited. We can see why in the next figure.

2.2 Non-uniform sampling

In practice, it is not unusual to come across examples where sampling times are distorted or that uniform sampling is simply unpractical. In Astronomy, for instance, sampling times

can be influenced by the weather, or in mining industries sudden bottlenecks or erratic work practices can also affect the periods in which samples from the mineral are taken.

This time irregular or non-uniform sampling can be seen not as a Dirac comb but a sum of distinct Dirac deltas at some time points t_n

$$\Pi(t) = \sum_{n=1}^{N} \delta(t - t_n).$$

As such, we have the following observed Fourier transform

$$\hat{g}(\xi) = \hat{f}(\xi) \star \hat{\Pi}(\xi).$$

However, this time, as $\Pi(t)$ does not have the symmetry provided by the Dirac comb has, its transform is not as straightforward.

So recovering a signal given a set of unevenly-spaced time points is a much more diffcult problem, because the uneveness of the sampled points lead to uneveness of the peaks of the transform of the sampling scheme, and as such will bias our results to some, in principle, arbitrary frequencies.

However, there is also much more to gain from a unevenly spaced sampling scheme. For instance, theoretically, in the irregularly sampled case, there is no relevant Nyquist frequency limit. In fact the only result of a non-regular sampling scheme pseudo-Nyquist limit is given in [5].

Theorem Let p be the largest value such that each sampling time t_i can be written as $t_i = t_0 + n_i p$, for integers n_i. Then the *Nyquist Frequency* is $1/2p$

In other words, in the unevenly sampled case, you can reduce to the Nyquist frequency sampling scheme as long as we can find a regular sampling scheme which will contain our sampled points, which can become pretty large for arbirtary sequences t_i, in particular if the spacing of a pair of points is irrational, then there is no such Nyquist limit (although in reality there may be one due to the precision of said time measurements)

2.3 Sampling schemes

Previously, it has been shown that the Nyquist-Shannon theorem requires uniform sampling. As it will be necessary to analyze the consequences of other sampling schemes in the framework of Gaussian Process Regression, lets review some of the most common random sampling schemes and some of their known properties in signal processing.

2.3.1 Uniform sampling

Uniform sampling for a $[-W, W]$ band-limited signal f is defined as the set of observtions points defined by:

$$t_n = \frac{n}{f_s}, \quad n \in \mathbb{Z}, \tag{2.1}$$

where f_s is the **sampling frequency.** Remember that if $f_s = 2W$, then, from Nyquist-Shannon theorem we are able to reconstruct the complete function is determined by its values at the points t_n.

The main reason for the uniform sampling requirement is the intrinsic periodicity behind the dirac-delta train derived from this sampling scheme , which makes possible the use of Fourier series. BUt when the bandwidth of the signal is not correctly specified or the signal is corrupted by high-frequency noise, then reulst may be affected by the **alias-error.**

Alias Error

In signal processing and related disciplines, aliasing is an effect that causes different signals to become indistinguishable (or aliases of one another) when sampled. This problem is not exclusive for band-limited signals. An example is given in Fig. 2.1

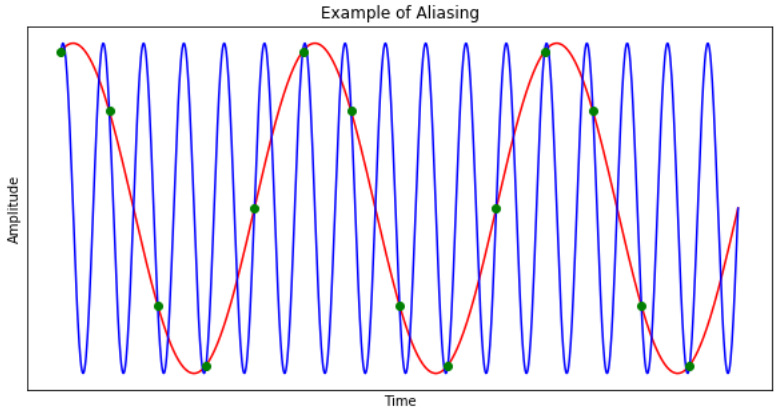

Figure 2.1: An example of alias-error, if the sampling frequency is low, high-frequencies may be mistaken by low frequency content, making our approximation

To show how aliases are cosntructed in this setting, consider the following: Let $x(t)$ be a (wide-sense) stationary Gaussian process. It will be assumed that $x(t)$ is real, zero-mean and with a covariance function $K(\tau)$. From the Bochner theorem, we know that there exists a power spectral density $S(\xi)$ such that

$$K(\tau) = \int_{-\infty}^{\infty} e^{i2\pi\xi t} S(\xi) d\xi,$$

$$S(\xi) = \int_{-\infty}^{\infty} e^{-i2\pi\xi\tau} K(\tau) d\tau.$$

Given the sequence of sampling points given in Eqn. 2.1, define the sequence $x(t_n)$ which is a stationary random sequence, with correlation function k:

$$k_{f_s}(n) = K(t_n).$$

But notice that, using the power spectrum of the kernel

$$\begin{aligned} k_{f_s}(n) &= K(t_n) \\ &= \sum_{j=-\infty}^{\infty} \int_{2\pi j f_s}^{2\pi(j+1)f_s} e^{i\xi t_n} S(\xi) d\xi \\ &= f_s \int_{-\pi}^{\pi} e^{i\xi t_n} \left[\sum_{j=-\infty}^{\infty} S(f_s(\xi + 2\pi j)) \right] d\xi. \end{aligned}$$

The inside sum in the final equation correspond to the winding of the power spectral density. The name comes from how, viewing such sum as an operator $\mathcal{W}$ such that

$$\mathcal{W}(S) = \sum_{j=-\infty}^{\infty} S(f_s(\xi + 2\pi j)),$$

then the operator folds the whole spectrum inside the domain $[-\pi, \pi]$, were superimposed frequencies (or densities) are added together.

This shows, as the covariance function determines the stationary zero-mean sequence $x(t_n)$, that any other power density $\hat{S}(\xi)$, different from $S(\xi)$, such that

$$\mathcal{W}(\hat{S}) = \mathcal{W}(S)$$

is an alias for the signal $x(t)$.

2.3.2 Randomized sampling

Randomized sampling is the process of selecting sampling points via an stochastic process of some sort. When considering random sampling schemes, then there is no Nyquist limit (there is a computational variant, but it is not useful for theoretical or experimental practices). So, in a way, randomized smapling can have its upsides and its downsides. Some of the better known sampling schemes known in the literature are the following:

2.3.3 Jittered sampling

Jittered sampling corresponds to a perturbated form of uniform sampling, that is, the sampling points t_n correspond to

$$t_n = \frac{n}{f_s} + \varepsilon_n,$$

where $\varepsilon_n \sim \mathcal{N}(0, \sigma)$ i.i.d. or $\textbf{Uniform}[-\delta, \delta]$. In the former case, for example, the covariance function of the sequence $x(t_n)$ ends up being

$$k_{f_s}(n) = \int_{-\infty}^{\infty} K(\tau) p(n f_s - \tau) d\tau,$$

where p is the density of $\mathcal{N}(0, \sigma)$.

A proof of the former can be found in [18] It is worth noting that this sampling scheme, even though it is random, it can still suffer from alias error. Nonetheless in the following chapter we will show that jittered sampling gives surprisingly low errors when compared to other sampling schemes, in the sub-Nyquist setting.

Kadec's 1/4 theorem

Related to this type of sampling there is a theorem, known as the Kadec 1/4 theorem [2] which states that

Theorem 2.3.1 (Kadec 1/4 theorem) Consider the following set of sampling points $t_n = \frac{n}{f_s} + \varepsilon_n$, with $\varepsilon \sim$ Uniform$[-f_s/4, f_s/4]$ i.i,d. Then any bandlimited function can be decomposed uniquely in the basis $\text{sinc}(t - t_n)$. The 1/4 bound is tight.

In a way, a small amount of jittering, as long as it under certain bounds, does not affect the independency of the basis, in practice , when a finite number of samples are considered, it even aliviates conditioning problems of the Gram matrix.

2.3.4 Additive Sampling

Jittered sampling does not prevent aliasing because the sampling times t_n are still attracted to the equi-spaced values [18]. In order to break symmetry, additive random sampling is considered, whose observation times are defined by:

$$t_n = t_{n-1} + \gamma_n,$$

where the family $\{\gamma_n\_n \in \mathbb{Z}\}$ is a family of independent i.i.d random variables.

In this case, the sequence $x(t_n)$ has as covariance function

$$k_{f_s}(n) = \int_{-\infty}^{\infty} S(\xi)\phi(\xi)^n d\xi, \tag{2.2}$$

where ϕ is the characteristic function associated with the random variables γ_n.

An important case is when the time samples t_n follow a Poisson process (i.e when $\gamma_n \sim \exp(\lambda)$ for some $\lambda \in \mathbb{R}$) because that particular case of additive sampling is called Poisson sampling and has been proven to be alias free [18], meaning that no two different sequences gives the same estimates for $S(\xi)$. The proof exploits the cahracteristic function in Eqn. 2.2 and uses complex analysis to find a contradiction.

Even though this sounds nice, in practice alias elimination comes at a cost, that is having large jumps between sampling points, which contribute to the error, not by aliasing, but by

making estimations harder to fit, as there is more uncertainty. An experiment was carried out and results are shown in Chapter 3.

Chapter 3

Applications

3.1 Low-pass filtering as Bayesian inference

Filtering is one of the main concerns of signal processing and time-series analysis. Applications of this technique include denoising, fault detection, edge-detection in images, infering a long term component of a time-series and many more.

In this section we will use the sinc-kernel to propose a latent-component generative model for Bayesian filtering. Its main attractive feature will be that it performs well even if the data points are sampled irregularly.

In the following section we will also extend this scheme in a way that will also handle the case when there is uncertainty of the exact separation between the spectral components in a very natural way.

The proposed generative model for a continuous-time(latent) signal $(f(t))_{t\in\mathbb{R}}$ is simply:

$$f(t) = \ell(t) + h(t),$$

where $\ell(t)$ is a low-frequency signal (bandlimited) and $h(t)$ one of high-frequency content.

We will assume that both signals can be modeled as Gaussian Processes with stationary covariance kernels denoted respectively by L and H, and, accordingly, power spectral densities $\hat{L}$ and $\hat{H}$. To discriminate between higher and lower frequencies, we impose the following restrictions over $\hat{L}$ and $\hat{S}$

1. The support of $\hat{L}$, denoted by $\text{supp}(\hat{L})$, is compact and centred around the origin, meaning that $\ell(t)$ is a process of low-frequency content.

2. The supports of $\hat{L}$ and $\hat{H}$ are non overlapping, that is, $\text{supp}(\hat{L}) \cap \text{supp}(\hat{H}) = \emptyset$. This implies that each frequency present in the signal f came, exclusively, from either ℓ or h.

3. The sum of the of the component PSD's correspond to some other known PSD of some known kernel. For the sake of the example, we chose the square-exponential kernel, that is $\hat{H} + \hat{L} = S_{\text{SE}}$ where S_{SE} is the PSD of a square exponential kernel with some hyperparameters $\sigma^2$and l, but any other stationary kernel can be used, as we will discuss later.

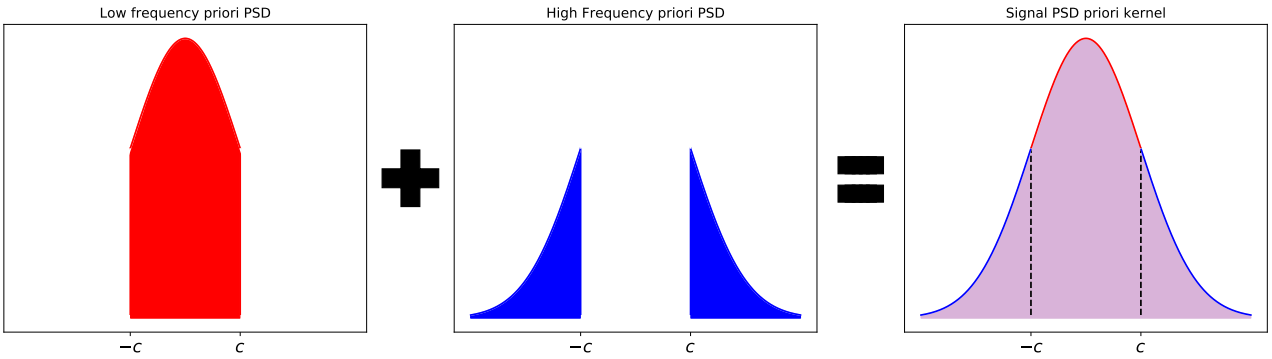

Figure 3.1: Model assumptions given in the model can be seen in this figure. The respective PSD's are such that the sum of them can be approximated by an already known kernel and their supports are non-overlaping, the Gaussian one in this case.

Here we derive a formula for the entries of the cross-covariance matrix of the (multi-Gaussian) process $\psi(t) = (\ell(t), h(t))$. This matrix is defined by the following formula:

$$\Theta(\tau) = \langle \psi(t), \psi(t)^T \rangle = \begin{pmatrix} \langle \ell(t), \ell(t+\tau) \rangle & \langle \ell(t), h(t+\tau) \rangle \\ \langle h(t), h(t+\tau) \rangle & \langle h(t), \ell(t+\tau) \rangle \end{pmatrix}.$$

The diagonal terms are, respectively, $\langle \ell(t), \ell(t+\tau) \rangle = L(\tau)$ and $\langle h(t), h(t+\tau) \rangle = H(\tau) = K(\tau) - L(\tau)$ where the last equality is consequence of our third assumption.

In contrast, the remaining terms vanish. Indeed we have that

$$\begin{aligned} \langle \ell(t), h(t+\tau) \rangle &= \hat{\ell}(\xi)\hat{h}(\xi)e^{i\tau\xi} \\ &= 0, \end{aligned}$$

where the first equality is Parseval's theorem and the second one is consequence of our second assumption and the fact that a signal generted by a band-limited kernel, will be band-limited to the same interval as its kernel.

As a consequence we deduce that both components on our model are independent from each other. This will be important in a minute.

3.1.1 Likelihood and model fitting

Assuming independent sequence of Gaussian observation noise the observations $(y(t))_{t\in\mathbb{R}}$ are then defined as

$$y(t) = f(t) + \eta(t), \quad \eta(t) \sim \mathcal{N}(0, \sigma_\eta^2).$$

Combining the observation model , with the GP-prior assumed for the spectral components, the marginal likelihood of the proposed model is Gaussian and therefore its hyperparameters can be obtained through minimization of the negative log-likelihood (NLL). Notice that despite the elaborate frequency-wise construction of the latent process f through the non-overlapping spectra of the components $\ell(t)$ and $h(t)$,he covariance kernel of f is square-exponential, thus allowing for straightforward model learning. Specifically, the NLL of the model is given by

$$NLL(y|t) = \log(2\pi|\Sigma_{\mathbf{y}}|) + \frac{1}{2}\mathbf{y}^T\Sigma_{\mathbf{y}}^{-1}\mathbf{y},$$

where $\mathbf{y} = [y_1, ..., y_N]$ is the vector of observed values acquired at time instants $\mathbf{t} = [t_1, t_2, ..., t_N]$ and $\Sigma_{\mathbf{y}}$ is the covariance matrix of $\mathbf{y}$ defined by

$$\Sigma_{\mathbf{y}} = K_{\text{SE}}(t, t) + \sigma_\eta^2\mathbf{I}.$$

Therefore, the hyperparameters are those of the K_{SE} kernel and the noise variance σ_η^2. Finally, observe that the strict non-overlapping property of the components ℓ and h is not problematic for training, in fact, the cutoff frequency does not even appear for model training.

3.1.2 Filtering as Posterior Inference

Denoting c as the required cut-off frequency of the low-pass filtering problem, using the proposed model, we can assume that this cutoff frequency is equal to the limit between the low- and high- frequency components. In this context, low-pass filtering problem is equivalent to performing inference over the low-frequency component ℓ conditional to observations of the time seires. Dure to the assumptions made on the signal we refer to this approach as GP low-pass filter (GPLP).

Denoting the observations by $\mathbf{y} \in \mathbb{R}^n$, GPLP addresses low-pass filtering by computing the posterior distribution $p(\ell|\mathbf{y})$. Due to the self-conjugacy of the Gaussian distribution and its closure under additivity, this posterior is also a GP, with mean and covariance given by

$$\begin{aligned} m_{\ell|\mathbf{y}}(\tau) &= L(\tau - \mathbf{t})\Sigma_{\mathbf{y}}^{-1}\mathbf{y}, \\ K_{\ell|\mathbf{y}}(\tau) &= L(\tau) - L(t - \mathbf{t})\Sigma_{\mathbf{y}}^{-1}L(t - \mathbf{t})^T, \end{aligned}$$

where, remember, we have assumed zero mean for h and ℓ, we denote by $L(\tau - \mathbf{t})$ the vector of translated basis functions $[L(t - t_1), L(t - t_2), ..., L(t - t_n)]$. This expression is due to the independence property mentioned before, indeed

$$\begin{aligned} \mathbb{E}[\ell(t)(\ell(t') + h(t') + \eta(t'))] &= \mathbb{E}[\ell(t)\ell(t')] \\ &= L(t - t'). \end{aligned}$$

Therefore, the only critical quantity required to compute the former equations is the kernel L. Following the model proposed, we have that the PSD of ℓ can be obtained by multiplying the PSD of f with a centered rectangular box function of width $2c$, that is

$$\hat{L}(\xi)) = S_{\text{SE}}\text{rect}\left(\frac{\xi}{2c}\right),$$

with the rect(ξ) function being equal to $\frac{1}{2}$ for $|\xi| < \frac{1}{2}$ and 0 elsewhere. As a consequence, the kernel L can be calculated using the convolution theorem:

$$\begin{aligned}
L(t) &= \mathcal{F}^{-1}(\hat{L}(\xi)) \\
&= \mathcal{F}^{-1}\left(S_{\text{SE}}\text{rect}\left(\frac{\xi}{2c}\right)\right) \\
&= \mathcal{F}^{-1}(S_{\text{SE}}) \star \mathcal{F}^{-1}\left(\text{rect}\left(\frac{\xi}{2c}\right)\right) \\
&= K_{\text{SE}}(t) \star \text{sinc}(2ct) \cdot 2c \\
&= 2c \cdot \int \sigma^2 \exp\left(-\frac{1}{2l^2}(t-\tau)^2\right) \frac{\sin(2\pi c\tau)}{2\pi c\tau} d\tau \\
&= \sigma^2 e^{-\frac{1}{2l^2}t^2} \Re\left(\text{erf}\left(\sqrt{2}cl\pi - i\frac{t}{\sqrt{2}l}\right)\right),
\end{aligned}$$

where erf(t) denotes the error function given by

$$\text{erf}(t) = \frac{1}{\sqrt{\pi}} \int_{-t}^{t} e^{-x^2} dx$$

Using Taylor exapnsions, the error function can be calculated up to an arbitrary degree of accuracy.

3.1.3 Extensions

Window shapes

So far, we have shown that the GPLP allows to construct a probabilistic version of the high-frequency component and low-frequency component of the original signal $f(t)$.

However, we know that convergence of the sinc function series is not very efficient, as you need many samples in order to obtain error ($\Omega(1/\varepsilon)$ samples to achieve an ε error, [3]) which is not desirable.

Two main ways to obtain better convergence rates is by using continuous window functions, which in the signal processing literature are known to give faster convergence in exchange of the need to oversample past the Nyquist rate [15].

Model-wise, the price for doing so is to allow a level of uncertainty in the cutoff level, meaning that now there will be a non-empty intersection between both signal components.

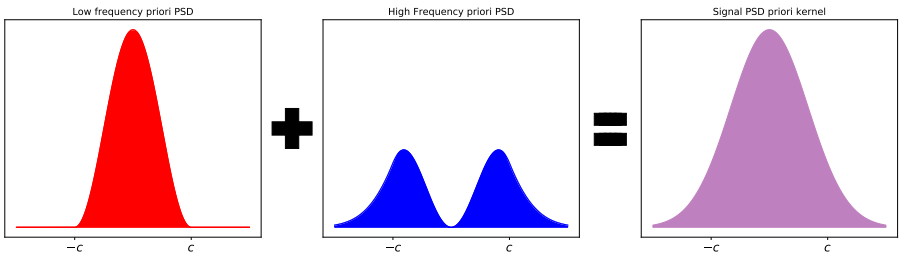

Figure 3.2: An example of a different window shape for our generative model. In this case a Hahn window is used, but triangular or eliptic can be used too.

Base Kernel

It was mentioned that the main prior is the RBF Kernel, in other words this are our priori beliefs of how the total signal (The sum of both components) is shaped. The bell shape kinds of works like an low-pass filter in and of itself (It has been compared to a sinc kernel in [16]).

So if we want a more general model we can choose the RBF parameters in such a way so as to have a wide PSD (And, in consequence, a narrow shaped bell in time domain), this is known as a diffuse prior.

However adjusting the hyperparametrs that way can be tricky, as you could end up with a bunch of non-correlated points, which means no inference will be made afterwards.

Another way of widening up our a priori beliefs of frequency bands is by using a kernel with heavier tails, such as the so popular Mattern kernel, which also has a widely known PSD. The Mattern Kernel has an hyperparametrs which controls how heavy the tails are. Actually the Mattern kernel is composed of a sum of polynomials, as the degree of the polynomial increases, so does the smoothness of the kernel, and also the PSD concentrates the most in its center. At its limit (When you consider the sum of all the polynomials) you obtain the RBF Kernel. Heavier tails are usually better in handling outliers and high-frequency noise.

3.1.4 Implementation

So the possibility of tweaking the window shape and the base kernel of our method has been included. Unfortunately, making such changes analytically require the calculation of certain integrales, which are not analytically possible to calculate. So numerical techniques must be applied when using the algorithm in this way.

If the base kernel is the RBF, then a qudrature rule, Gauss-Hermite [19], can be applied to make exact calculations of it. Even in the case of the Mattern kernel, Laguerre polynomials can be used as a quadrature rule. As Gauss-Hermite quadrature makes the following approximation

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{i=1}^{n} w_i f(x),$$

where n is the number of sample points used, and the x_i are the roots of the Hermite

polynomial $H_n(x)$ given by:

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(x_i)]^2}.$$

If we are not working with the RBF, or any other kernel which has a quadrature rule associated, then we can use an approximiation which is actually an application of the Discrete Fourier Transform.

Consider $S(\xi)$ the PSD of our base kernel, and consider $W(\xi)$ our window funciton (Hahn, semicircle, etc.). What we have to calculate is the low frequency kernel function in time which by definition is

$$L(t) = \mathcal{F}^{-1}(S(\xi) \cdot W(\xi).$$

Then, if S if regular enough, then we could approximate the product inside the inverse Fourier transform as:

$$S(\xi)W(\xi) \approx \sum_{i=1}^{n} S(x_i)W(x_i)\text{rect}[\Delta(\xi - x_i)],$$

where the $\{x_i\}_{i=1}^n$ is the uniform partition of the target interval, with spacing Δ and $\text{rect}(\dot{})$ is the indicator function of the interval $[\frac{-1}{2}, \frac{1}{2}]$.

Using this approximation to calculate $L(t)$ we obtain the formula

$$\begin{aligned} L(t) &\approx \mathcal{F}^{-1}\left(\sum_{i=1}^{n} S(x_i)W(x_i)\text{rect}[\Delta(\xi - x_i)]\right) \\ &= \sum_{i=1}^{n} S(x_i)W(x_i)e^{2\pi i t x_i} \cdot \Delta \cdot \text{sinc}(\Delta \cdot t), \end{aligned}$$

which is exactly what we used in the simulations. We even compared this approximation with the analytic one and, as long as the partition is thin enough, no further differences were found.

3.2 Simulation

The proposed model for Bayesian low-pass filtering using GPs, termed GPLP, is next validated using synthectic and real-world data. Our experimental validation aims to show that GPLP accomplished both: To successfully recover low or high frequency data from missing and noisy observations, to provide accurate point-estimates with respect to the benchmarks, places meaningful error bars.

3.2.1 A synthetic time series with line spectra

We considered the line-spectra time series given by

$$f(t) = \sum_{\omega \in F_{\text{low}}} c_\omega \cos(2\pi\omega t) + \sum_{\omega \in F_{\text{high}}} c_\omega \cos(2\pi\omega t), \tag{3.1}$$

where the sets F_{low} and F_{high} are such that $\forall \omega_i \in F_{\text{low}}, \forall \omega_j F_{\text{high}} : \omega_i < \omega_j$, note that also each component is multiplied by an individual term c_ω. Simply put, F_{low} is a set of low frequencies and F_{high} a set of high frequencies - all these frequencies are in Hertz(Hz). Signals constructed in this wat have sparse PSDs meaning that only a finite number of frequencies convey all the signal energy or information. We chose $F_{\text{low}} = \{0.1, 0.2, 0.3, 0.45\}$ and $F_{\text{high}} = \{0.55, 0.7, 0.8, 0.6, 0.65\}$ with coefficients $\{1, 2, 2, 1.5\}$ and $\{1.0, 1, 2, 1, 3\}$ respectively. Then a path of $f(t)$ was simulated as defined in Eqn. (3.1) for 200 evenly spaced time indexes in $t \in [-10, 10]$. The observation time-series $\mathbf{y}$ consisted only in a 25% of the signal (again, evenly spaced) all of which were corrupted by Gaussian noise of std. dev. $\sigma_\eta = 1.0$. Fig. 3.3 shows the latent signal and the observation considered for this experiment.

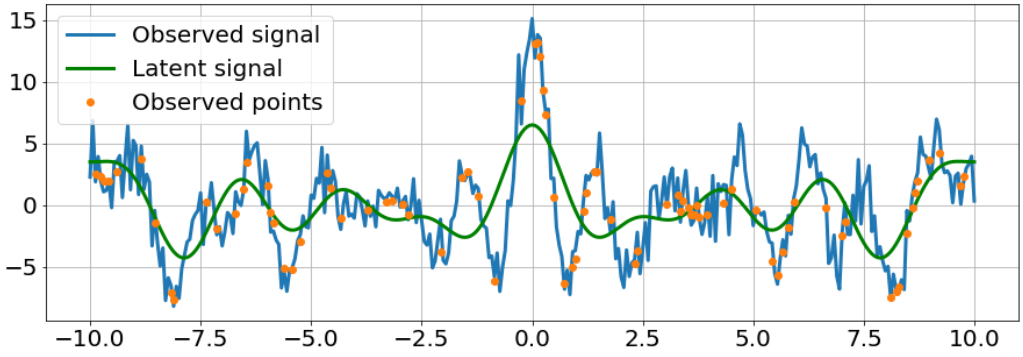

Figure 3.3: Signal used for the example of the GPLP model in the non-uniform sampled casevg.

The proposed GPLP was implemented in order to recover both the high-frequency and the low-frequency content of the original (latent) signal f only using the observations $\mathbf{y}$, the only difference being that the high-frequency content utilizes a shift-window parameter in its GPLP definition. First, the generative model is trained as explained in 3.1.1 to find estimates for the hyper parameters l, σ^2 and σ_η^2. Then, the cutoff frequency was chosen to be $b = 0.5$ HZ. Then either the low-frequency or the high frequency covariance function is computed in order to use it to compute the moments of the posterior distribution $p(f|\mathbf{y})$. Fig. 3.5 shows the learnt kernels and their corresponding PSDs. Notice how, just as illustrated in Fig. 3.1, the spectral densities of the latent low-frequency component is band-limited, suported only on $[-0.5, 0.5]$ and tightly bounded by the (unfiltered) time-series.

Next, figure 3.6 shows the GPLP estimates for both the high frequency filter and the low frequency filter, compared against the ground truth and a low-pass filter of order 10, with the same cutoff frequency; this filter is a standard in linear filtering. GPLP obtained a mean-squared error (MSE) of 0.43 while the Butterworth low-pass filter gave a mean-squared error of 0.6, in addition to this difference in performance, notice that GPLP provided accurate 95% error bars.

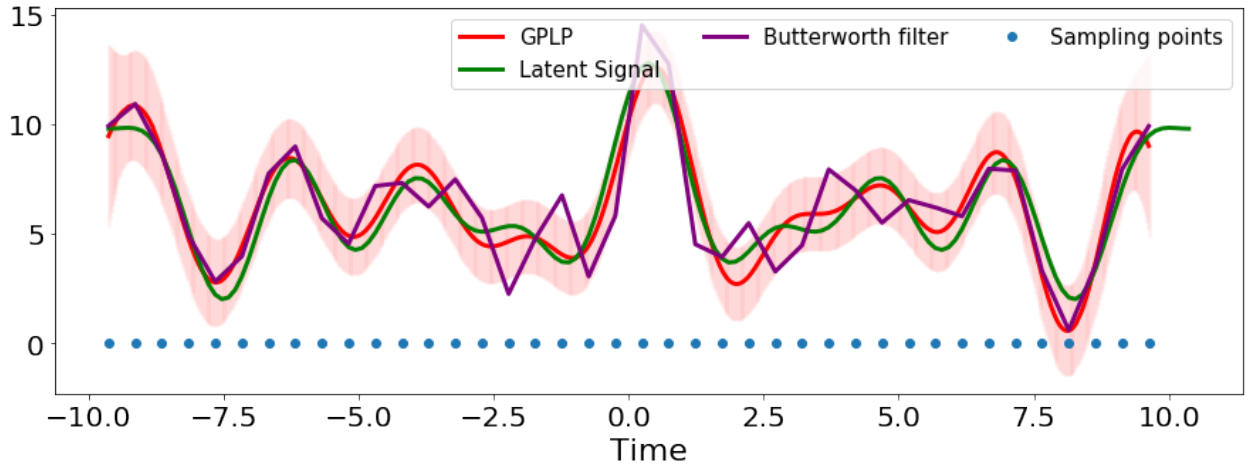

Figure 3.4: Results of the GPLP estimate versus the butterworth filter in the uniformly spaced setting.

Next we replicated the exact same setting, but considered randomly-chosen observations (Again, just 25% of the total number points). Fig 3.6 shows the posterior mean over the low-frequency component together with the 95% confidence interval and the ground truth, as well as the result in the frequency domain.

The MSE of the GPLP estimate was 0.49 thus improving over Butterworth using evenly spaced data. To further validate the ability of the proposed GPLP to filter out low-frequency spectral content, Fig 3.6 shows the Fast Fourier Transform (FFT) of the low-pass versions of GPLP and the original signal. Notice that the GPLP successfully recovered the first three spectral components and rejected the higher ones.

3.3 An experiment on sampling schemes

In this section, we evaluate the impact of the different sampling schemes mentioned in the former chapter (These are: random, jitter, additive, uniform) and test if its theoretical properties are reflected someway in practice and what kind of relationship can we conclude with respect to the classical signal processing theory.

Different sampling schemes such as additive, jitter, full-random, etc. were reviewed at the beginning and some properties and applications are known for them. In the case of the GPLP, the performance of each one of them was tested by solving the same problem as before, filtering a signal composed of line-spectra, and comparing their performance in terms of MSE and the number of points sampled.

For this, number of sampling points ranging from 10 to 80 were considered. For each one of these, 100 simulations of the line-spectra problem were performed. The MSE and the maximum separation (i.e The largest gap between sampled points) were recorded on each iteration. In Fig. 3.15 an histogram of the largest gap for each method is represented. It is interesting how additive sampling obtains the highest point, while uniform sampling has the lowest as expected. From the graph on the right one can conclude that larger gaps on time

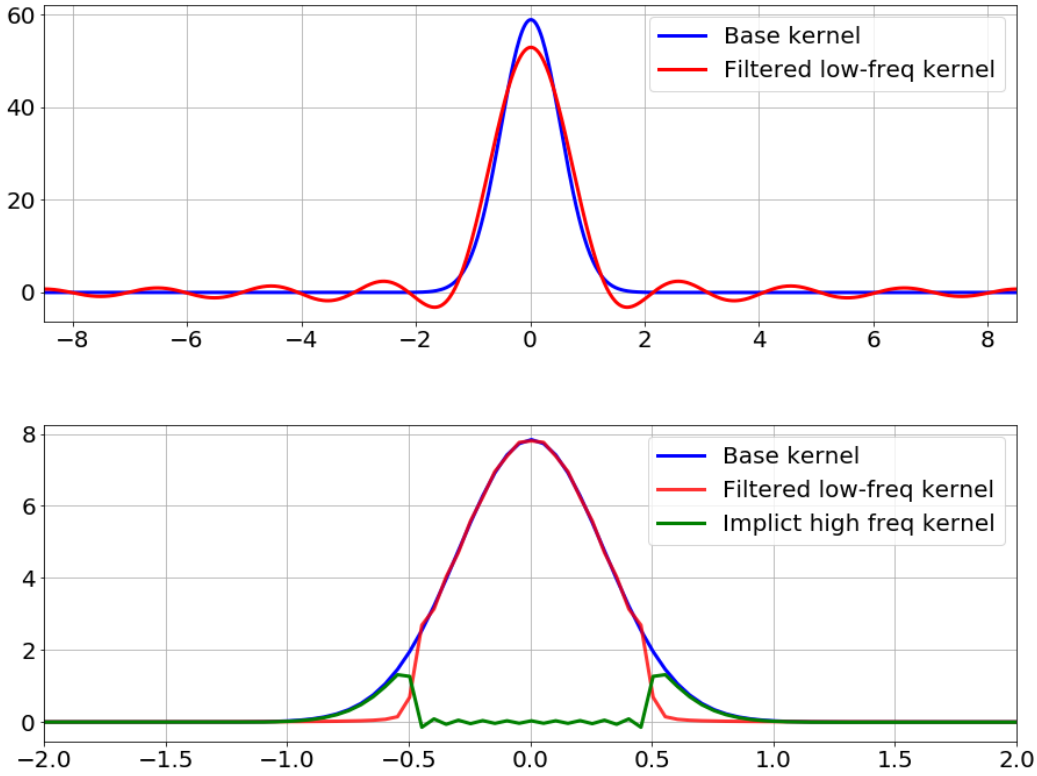

Figure 3.5: Above: Base kernel function, with the maximum likelihood parameters, with the low-frequency corresponding kernel. Below: The PSD of the base kernel, the PSD of the low freq kernel and the implicit high frequency kernel. Notice how the ringing artifact that is produced

samples are related to higher variability on the MSE statistic.

However, notice that from Fig. 3.8 it can be concluded that uniform sampling is not always the optimal sampling scheme, in fact we can see that jitter sampling is almost always better when the number of points considered is between approximately one half of the Nyquist limit and the Nyquist limit itself.

A closer inspection of the results is given in 3.9, which gives more details of the results from the simulation. The jitter gives a smoothing effect to the discontinuity of the MSE given by the uniform sampling. This is most likely due to a reduction on the alias error which is present on the uniform sampling scheme up until the cutoff point, while introducing little variance into the sampling point effect, while the other methods (full random sampling and additive sampling) may reduce the alias (In fact, we know how additive sampling reduces it completely in theory) they probably do not give a convenient tradeoff. This relationships hints that alias error vs. random sampling error,as it is seen in signal processing, can be seen, from a machine learning perspective, as the classic bias-variance trade-off.

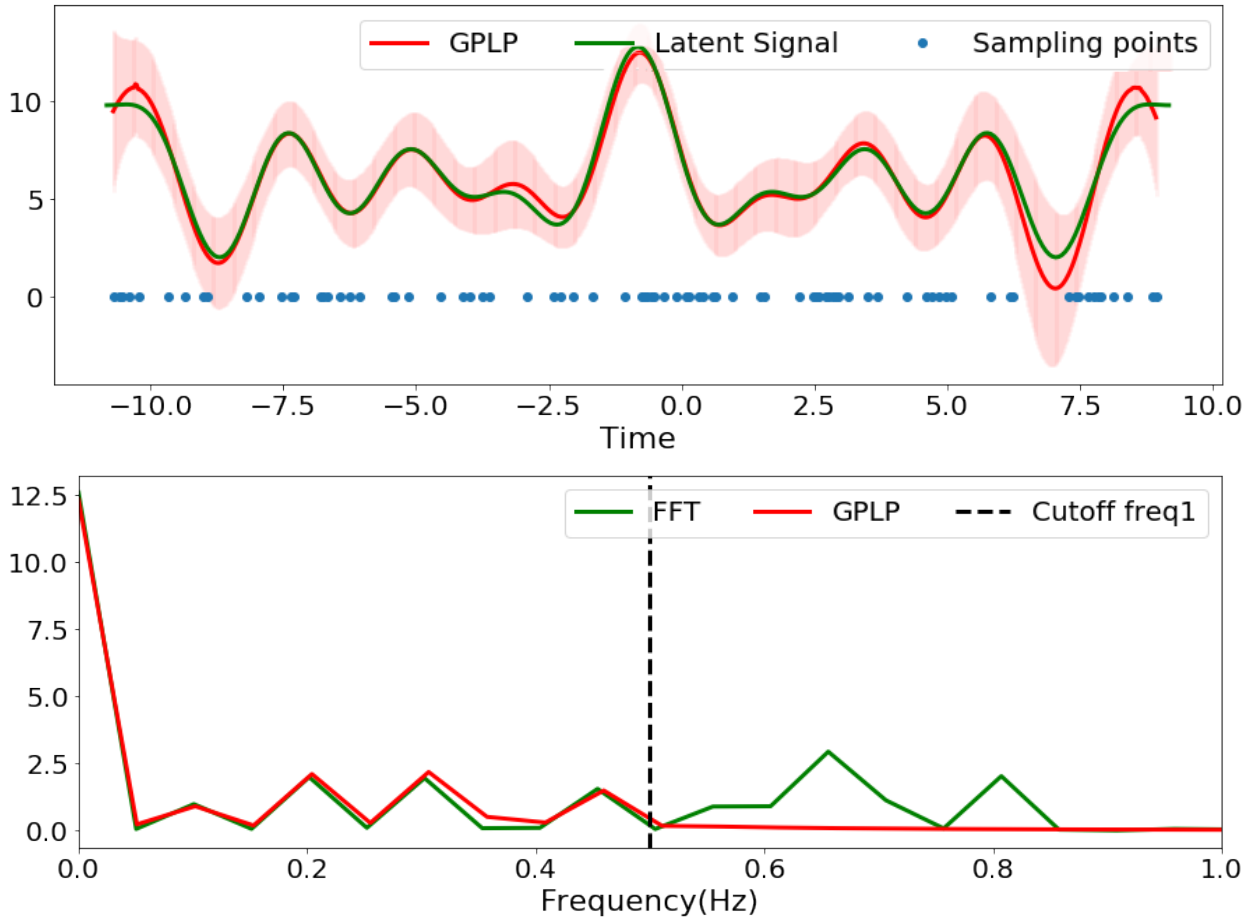

Figure 3.6: Results given when a full random sampling scheme is considered. MSE is 0.47.

3.4 Application: Finding an optimal sampling scheme using Bayesian Optimization

One of the advantages of Gaussian Process regression over classical signal processing techiques is having an estimation of the uncertainty (which is given by the posterior variance) and that we we are free to use any sampling scheme we want. In applications, such as compression, this may come handy. Remember that Nyquist bound on the number of sampling points is only valid in the uniformly spaced sampling scheme: We may save some sampling points (memory) by giving a little randomness to our sampling, as we saw in the previous experiment, where jittered sampling gave decent results halfway the Nyquist sampling rate. So we will search for an optimal sampling scheme using a procedure based on Bayesian optimization.

3.4.1 Setting

Consider an objective bandlimited signal ϕ of known bandwidth and the goal is to compress it into a number of discrete points $[x_1, .., x_M]$ in a way such that if the points $[x_1, ..., x_M]$ and its corresponding values $[\phi(x_1), ..., \phi(x_M)]$ are sent to an external agent, then it would be possible to reconstruct the whole signal (by making the corresponding regression). The number of points to be sampled may be fixed, or the points may be sampled until a certain

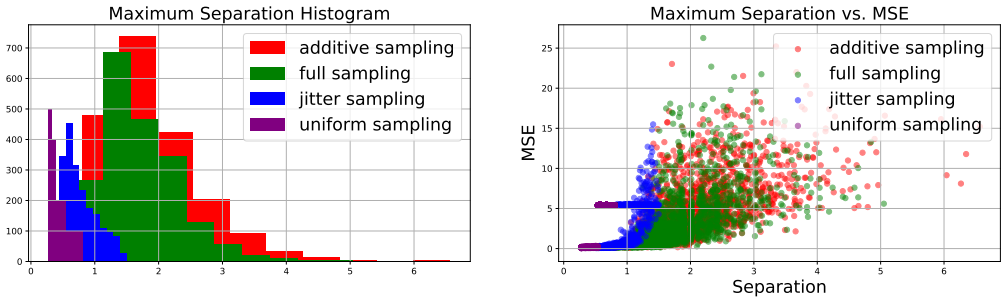

Figure 3.7: Left: Maximum separation between batch of sampling points histogram for each sampling procedure. Right: Relationship between separation and MSE for different number of sampling points.

error bound is achieved. Also, the reconstruction algorithm has to be robust against noise outside the bandwidth.

A nice way to make this possible is an iterative algorithm which will balance exploration and explotaition of the objective function.

This is made possible by defining an **acquisition function**. The one which we will use (and a popular choice by the way) is the function:

$$\alpha(x) = M(x) + \kappa\sigma(x),$$

where $M(x)$ is the mean squared error between the original signal and the posterior mean, $\sigma(x)$ the posteriori variance defined in the signal's domain and κ just a trade-off parameter between the two. If a point x^* maximizes the acquisition function that means that either the value $M(x)$ is high, so the signal may present some complexity in that zone (For instance high frequencies) or the value $\sigma(x)$ is too high, indicating an underexplored zone.

This setting, which balances exploration and explotaition via the acquisition function and approximating using Gaussian processes, is the main idea behind Bayesian optimization.

The pseudo-code would be

An instance of the algorithm can be seen in Fig. 10 where some iterations showing where the acquisition function is maxmimized and the effect the point has on the Gaussian process approximation. Interestingly, note how it seems to naturally go for the uniform sampling scheme at first, then the algorithm proceeds to exploit the extremes of the domain.

After running 100 simulations, the distribution of the sampled points outputted by the algorithm can be estimted by plotting the empirical distribution. It seems that the target distribution is uniform, except ont eh borders, where the distribution presents spikes. This is coherent with what was stated before: the algorithms tries to explore at first, and then exploits the borders. Note that a very similar optimal distribution was reported in [3]

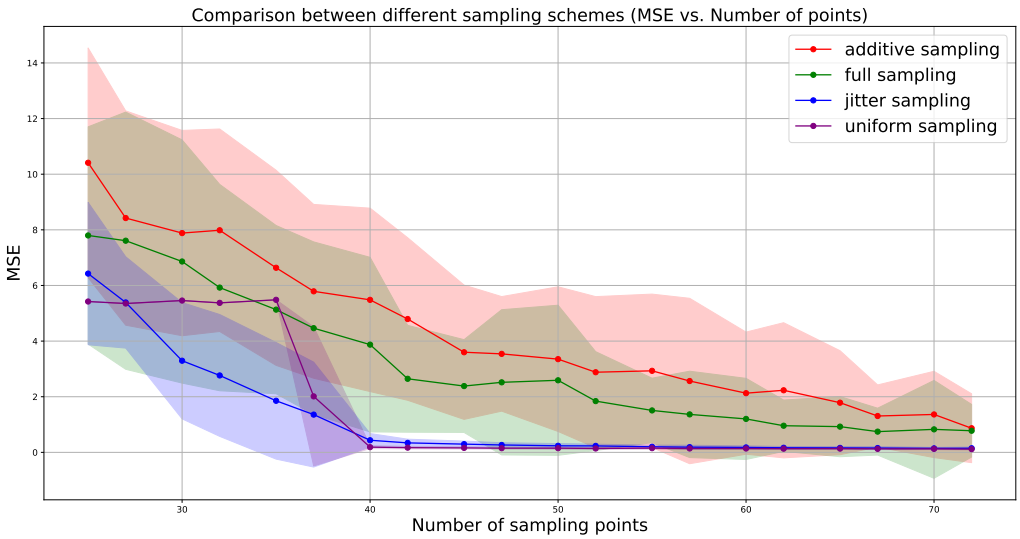

Figure 3.8: Each point corresponds to an average of 100 simulations with the given number of sampling points. Error bars are given based on standard deviation (1 unit). Surprisingly, jitter sampling (and to an extent full random sampling) perform well at a sub-nyquist level

3.5 Compressed sensing: Multi-band signal reconstruction using the sinc function

In this final section we will show an application of Compressed sensing setting, but using sinc functions. The main idea is to be even more restrictive in the Fourier structure of the signal in such a way that we will be able to reconstruct with very little data points. The purpose of this application is to expand on this topic in posterior work.

In particular, we consider the one dimensional compressed sensing setting, and, as mentioned on Chapter 2, random sampling is crucial and reconstruction is attainable with average sampling rates much lower than the Nyquist sampling rate.

The application consists in reconstructing a sum of sinc functions (this is, a multi-band signal). This is the restriction mentioned before, our prior knowledge of the signal is that it is sparse in some basis, in this case the sinc function basis.

The compressed sensing methodology requires to pose the problem as solving a system of equations:

$$Ax = b,$$

where A is an $m \times n$ matrix and $m << n$ so the system has infinite solutions. To obtain a unique solution, you add a regularization constraint to the problem, typically that the number of non-zero entries of x is limited or penalized (this is the so-called ℓ_0 norm defined by $||x||_{\ell_0} = |\{i : x_i \neq 0\}|$). In practice, however, one usually ocuppy the ℓ_1 norm, because using the ℓ_0 norm makes the problem computationally expensive most of the time.

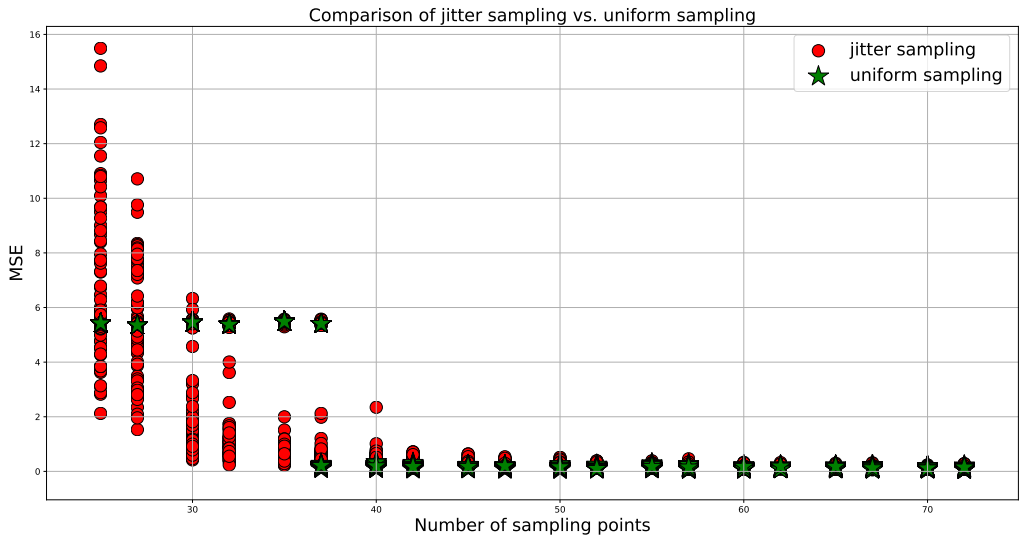

Figure 3.9: Each point correspond to a simulation of the line-spectra problem, and it is represented by its number of points and MSE score. It can be seen how jitter sampling outperforms uniform sampling at a sub-Nyquist level.

In our context, Ax is our objective, the signal we wish to reconstruct from our samples, and it will be represented by a vector x in a certain basis (a finite basis, for background in infinite-dimensional compressed sensing go to [1]). The b vector it simply is a vector of (randomly selected) samples.

In our example, the signal is a sum of two sinc functions (of width 584 and 600 respectively) multiplied by high frequency cosines (2585 Hz. and 3493 Hz). In Fig. 3.12 the full signal is composed of 5000 points along a time doimain of 1/8. From the 5000, only 5% is randomly selected. Random selection is important to mantain certain properties of the (random) matrix A which are beyond this work, but can be reviewed in [9].

Keep in mind, if these points were evenly spaced, the Nyquist frequency would be around 1kHz. making it impossible to reconstruct the signal (it would violate Shannon's theorem). Consider, for example, the signal in Fig.3.12:

In the plots above, we see that the signal has a clear pattern, yet is non-trivial. The plots in the top row are of the signal in the temporal domain at different scales. The plots in the bottom row are of the signal in the spectral domain (i.e., the signal's frequency content). Considering the frequency domain in particular, we note that the spectrum is mostly zero except for the two squares representing the two sinc functions frequencies.

Take a moment to imagine how you would reconstruct the (temporal) signal with just the red dots. You'd have a data set that, to the naked eye, would look like nonsense. One might ask if it is somehow possible to extract those two dominant from the incomplete data so that we might reconstruct the signal? The answer is yes! We just have to correctly pose the problem. We want to solve, as mentioned before, $Ax = b$, we need to specify a correct A

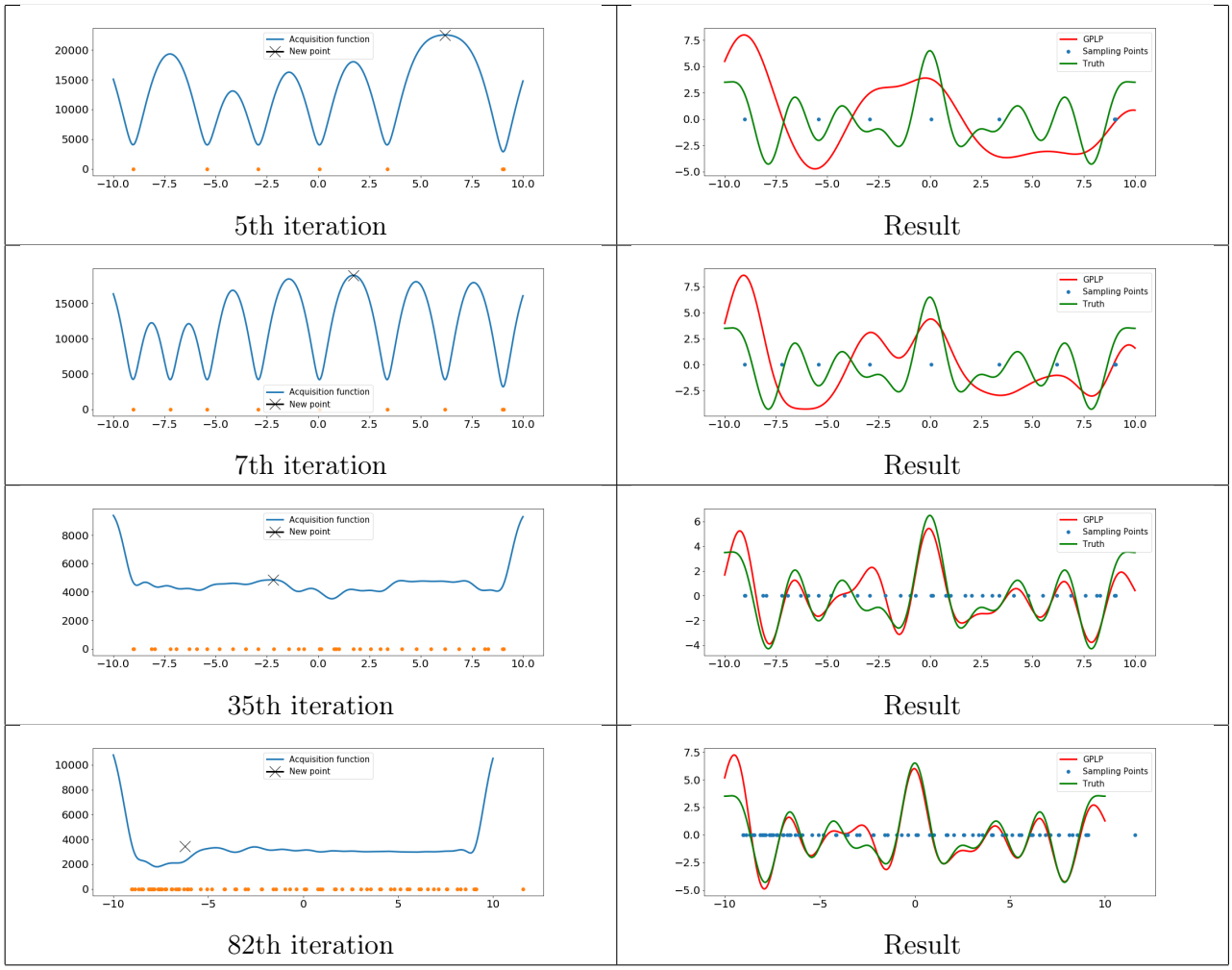

Figure 3.10: Different iterations of the optimal sampling scheme procedure. On the left, the aquisition function along with previously selected points and the newly accepted point. On the right, the actual signal vs. the fitted Gaussian process. Note how exploration is priorized on the earlier iterations, and then explotation (mostly on the borders).

Algorithm 1: Bayesian Optimization algorithm for finding optimal random sampling points

Input : A bandlimited function ϕ of known bandwith W, tolerance ε or maximum number of samples M and starting point x_0

Output: A sequence of sampling points $x_0, ...x_n$ and values $y_1, ..., y_n$

$X \leftarrow x_0$
$Y \leftarrow \phi(x_0)$
tol $\leftarrow \infty$
while $n < M$ *or* $tol > \varepsilon$ **do**
    Update the posterior probability distribution of ϕ (which is given by the GPLP procedure) over the current set of points $X$and values Y ;
    $x_{\text{new}} \leftarrow$ the minimizer of the acquisition function. ;
    $Y \leftarrow \phi(x_{\text{new}})$;
    $X \leftarrow x_{\text{new}}$ tol $\leftarrow$ the maximum value of the posterior variance.
end
return X, Y;

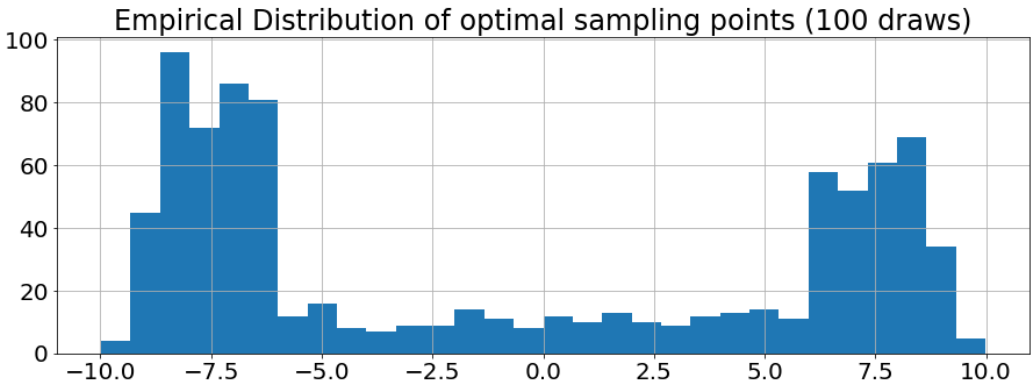

Figure 3.11: Empirical distribution of points after running 100 simulations of the Bayesian optimization sampling scheme algorithm

matrix such that the x vector of coefficients is as sparse as possible. One choice is to make $A = BC$ were C is the inverse Fourier (or cosine, it is equivalent) transform of an upper triangular matrix of ones, and B is just a sampling matrix (with sampling times according to b). That way x just need 4 non-zero entries, for each of the two components we need one entry for making a window wide enoguh to capture the component, and another window to trim off the surplus.

In Fig. 3.13 a graph of the columns of the A matrix (which are all sinc functions as they are Fourier squares after all), and below we can see how this matrix is affected when we multiply by the sampling operator. As we can see the columns are still distinguishable, but are noisier though.

After solving the ℓ_1 minimization problem, we obtain a vector x, with only 4 dominating non-zero entries. With those non-zero entries we can detect the location of the edges (in the

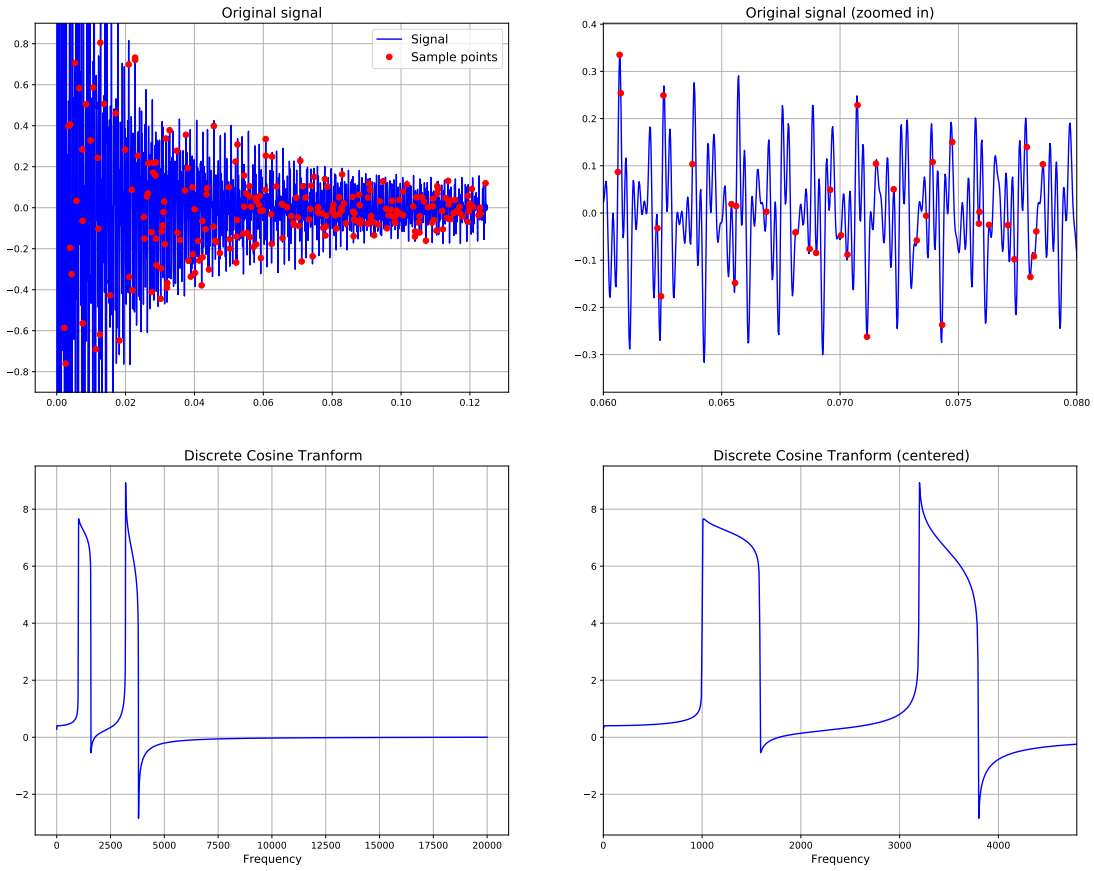

Figure 3.12: In blue, our target signal to reconstruct. In red the randomly sampled points (5% of the total). In the grap below we have a view of the signal in Fourier space.

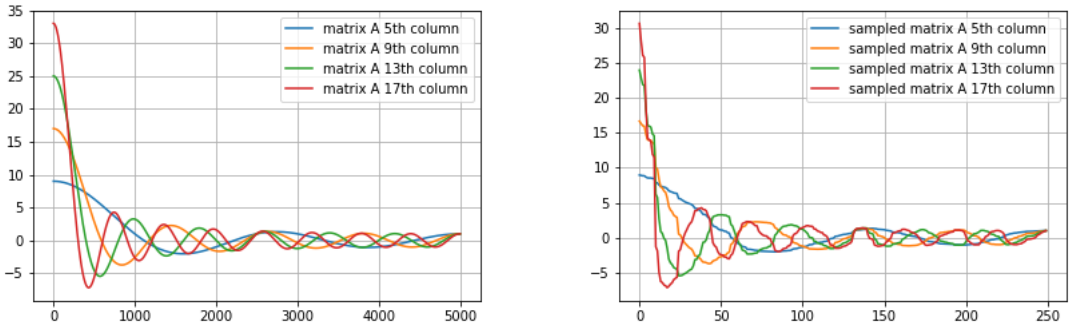

Figure 3.13: Caption

frequency domain sense of course)of the sinc components. The result is in Fig. 3.14:

Once we obtain the sinc component parameters, we perform a final adjustment of a scale

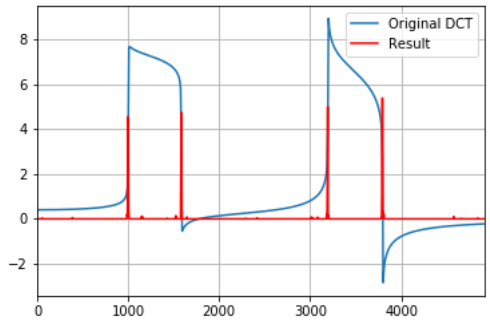

Figure 3.14:

factor to finally reconstruct the original signal. The mean squared error is absurdly low (0.006) and the reconstruction is very precise even though we are working at different scales as we move in the time domain (remember the sinc function decays as $1/x$).

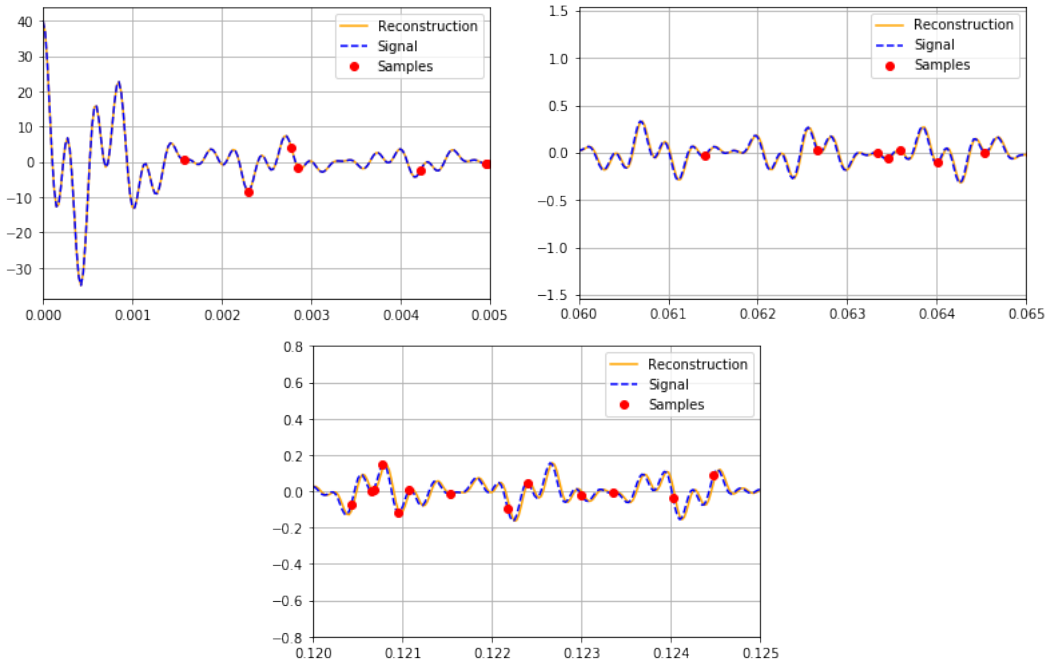

Figure 3.15: Final reconstruction of the signal given the sampled points at three diferent zones of the original domain and different scales to have a more clear visualization.

Compressed sensing is a very wide topic, and its not the main focus of this work, the purpose of this section is to encourage search in this regard. Gaussian Processes may be related in a way to such a procedure[10], but unfortunately the ℓ_1 norm minimization problem is key in all of this, and it is known that the sinc kernel is not well-suited to handle ℓ_1 problems in the Gaussian Process context [21].

Conclusion

This thesis has brought techniques from the Gaussian Process literature, the RKHS theory, non-parametric Bayesian methods as well as classical signal processing and Fourier analysis techinques in order to study the problem of signal analyisis when sampled at irregular time instants.

The questions posed at the beginning of the work , reproduced here for convenience, were:

- How many samples are required to have an accurate approximation to the whole function $f(t)$ and, if possible, what kind of scheme do we have to perform to select these points optimally?
- Can we design an algorithm which will, hopefully, solve the problem and implement it?

For the first question, the answer is that it depends on the bandwidth of the signal AND the sampling scheme being used. As we saw, it is not easy to answer that question in a way that is satisfying in practice, as you can always exploit prior knowledge of the signal in order to have reconstruction formulas with a few datapoints (as we saw on our compressed sensing example). From our results, however, we offer a solution via finding an optimal scheme using Bayesian optimization, and even study some classical sampling schemes and compare them.

For the second questiond the GPLP model was developed, which can be seen as solving the signal reconstruction problem when you have high frequency components as noise. The model is flexible and gives fairly accurate results even in the presence of totally random sampling schemes. One downside, however, is that obtaining results is computationally expensive, considering one dimensional data standards. So further research in more efficient variants is encouraged, in this sense implementations such as the oens made by [4] are promising.

Future research in this aspect which may be interesting is to find meaningful connections between the RKHS/Gaussian Process theory and the Compressed Sensing approach, which could give insights in developing even more efficeint tools.

Bibliography

[1] B. Adcock. "Infinite-dimensional compressed sensing and function interpolation". In: *Found. Comput. Math.* 18.3 (2018), pp. 661–701. DOI: 10.1007/s10208-017-9350-3.

[2] A. Avantaggiati, P. Loreti, and P. Vellucci. "Kadec-1/4 Theorem for Sinc Bases". Preprint (2016). arXiv: 1603.08762 [math.FA].

[3] H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, and A. Zandieh. "A universal sampling method for reconstructing signals with simple Fourier transforms". In: *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2019, pp. 1051–1063. DOI: 10.1145/3313276.3316363.

[4] A. Banerjee, D. B. Dunson, and S. T. Tokdar. "Efficient Gaussian process regression for large datasets". In: *Biometrika* 100.1 (2013), pp. 75–89. DOI: 10.1093/biomet/ass068.

[5] P. Bartholdi and L. Eyer. "Variable stars: Which Nyquist frequency?" In: *Astron. Astrophys. Suppl. Ser.* 135 (1999), pp. 1–3. DOI: 10.1051/aas:1999102.

[6] M. G. Beaty and M. M. Dodson. "An application of a general sampling theorem". In: *Results Math.* 34.3-4 (1998), pp. 241–254. DOI: 10.1007/BF03322054.

[7] G. L. Bretthorst. "Bayesian spectrum analysis and parameter estimation". Vol. 48. Lecture Notes in Statistics. Springer-Verlag, New York, 1988, pp. xii+209. ISBN: 0-387-96871-7. DOI: 10.1007/978-1-4684-9399-3.

[8] C.-H. Chen and R. Chellappa. "Chapter 2 – Face Recognition Using an Outdoor Camera Network". In: *Human Recognition in Unconstrained Environments*. Ed. by M. De Marsico, M. Nappi, and H. Proença. Academic Press, 2017, pp. 31–54. DOI: 10.1016/B978-0-08-100705-1.00002-6.

[9] M. A. Davenport and M. B. Wakin. "Compressive Sensing of Analog Signals Using Discrete Prolate Spheroidal Sequences". Preprint (2011). arXiv: 1109.3649 [cs.IT].

[10] S. Diamond and S. Boyd. "CVXPY: a Python-embedded modeling language for convex optimization". In: *J. Mach. Learn. Res.* 17 (2016), Paper No. 83, 5.

[11] G. Fasshauer and M. McCourt. "Kernel-based Approximation Methods using MATLAB". Vol. 19. Interdisciplinary Mathematical Sciences. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2015, pp. xviii+518. DOI: 10.1142/9335.

[12] H. G. Feichtinger and K. Gröchenig. "Iterative reconstruction of multivariate band-limited functions from irregular sampling values". In: *SIAM J. Math. Anal.* 23.1 (1992), pp. 244–261. DOI: 10.1137/0523013.

[13] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. "Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences". Preprint (2018). arXiv: 1807.02582 [stat.ML].

[14] N. R. Lomb. "Least-squares frequency analysis of unequally spaced data". In: *Astrophys. Space Sci.* 39.2 (1976), pp. 447–462. DOI: 10.1007/BF00648343.

[15] R. J. Martin. "Irregularly sampled signals: Theories and techniques for analysis". Doctoral thesis (Ph.D.) University College London (United Kingdom), Jan. 1998.

[16] C. E. Rasmussen. "Gaussian Processes in Machine Learning". In: *Advanced Lectures on Machine Learning*. Ed. by O. Bousquet, U. von Luxburg, and G. Rätsch. Vol. 3176. Lecture Notes in Comput. Sci. Springer, Berlin, 2004, pp. 63–71. DOI: 10.1007/978-3-540-28650-9_4.

[17] B. Schölkopf, R. Herbrich, and A. J. Smola. "A generalized representer theorem". In: *Computational learning theory (Amsterdam, 2001)*. Vol. 2111. Lecture Notes in Comput. Sci. Springer, Berlin, 2001, pp. 416–426. DOI: 10.1007/3-540-44581-1_27.

[18] H. S. Shapiro and R. A. Silverman. "Alias-free sampling of random noise". In: *J. Soc. Indust. Appl. Math.* 8 (1960), pp. 225–248.

[19] B. Shizgal. "A Gaussian quadrature procedure for use in the solution of the Boltzmann equation and related problems". In: *J. Comput. Phys.* 41.2 (1981), pp. 309–328. DOI: 10.1016/0021-9991(81)90099-1.

[20] D. Slepian and H. O. Pollak. "Prolate spheroidal wave functions, Fourier analysis and uncertainty. I". In: *Bell System Tech. J.* 40 (1961), pp. 43–63. DOI: 10.1002/j.1538-7305.1961.tb03976.x.

[21] G. Song, H. Zhang, and F. J. Hickernell. "Reproducing kernel Banach spaces with the ℓ^1 norm". In: *Appl. Comput. Harmon. Anal.* 34.1 (2013), pp. 96–116. DOI: 10.1016/j.acha.2012.03.009.

[22] T. Strohmer. "Numerical analysis of the non-uniform sampling problem". In: vol. 122. 1–2. Numerical analysis 2000, Vol. II: Interpolation and extrapolation. 2000, pp. 297–316. DOI: 10.1016/S0377-0427(00)00361-7.

[23] F. Tobar. "Band-Limited Gaussian Processes: The Sinc Kernel". Preprint (2019). arXiv: 1909.07279 [stat.ML].

[24] C. Valenzuela and F. Tobar. "Low-pass filtering as Bayesian inference". Preprint (2019). arXiv: 1902.03427 [stat.ML].

[25] J. T. VanderPlas. "Understanding the Lomb–Scargle Periodogram". In: *Astrophys. J. Suppl. Ser.* 236.16 (2018), pp. 1–3. DOI: 10.3847/1538-4365/aab766.

[26] Y. Wang, R. Khardon, and P. Protopapas. "Nonparametric Bayesian Estimation of Periodic Light Curves". In: *Astrophys. J.* 756.1, 67 (2012), pp. 279–289. DOI: 10.1088/0004-637X/756/1/67.

[27] S.-j. Yeh and H. Stark. "Least-squares frequency analysis of unequally spaced data". In: *J. Opt. Soc. Am. A* 7.3 (1990), pp. 491–499. DOI: 10.1364/JOSAA.7.000491.

[28] C. Zou and K. I. Kou. "Robust signal recovery using the prolate spherical wave functions and maximum correntropy criterion". In: *Mech. Syst. Signal Process.* 104 (2018), pp. 279–289. DOI: 10.1016/j.ymssp.2017.10.025.