

# Estimation of Spatial Fields from Samples obtained at Unknown Random Locations

*by*

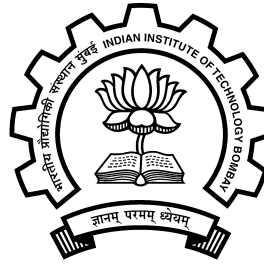
Ankur Mallick

ankur\_mallick@iitb.ac.in

*Mentor*

Prof. Animesh Kumar

animesh@ee.iitb.ac.in



Department of Electrical Engineering  
Indian Institute of Technology Bombay  
Duration: January 2015-June 2016

# 1 Introduction

The problem of localizing sensors in a wireless sensor network poses several challenges. An alternative to expensive localization schemes is to work with sensors which are location unaware. Recently, bandlimited field estimation *without* location information of the sensors in a distributed setup has been studied. This is an emerging field where the key idea is to utilize a multitude of such location-unaware sensors (oversampling) and leverage the random distribution on their spatial locations to estimate the underlying field. Due to symmetry and shift-invariance properties of bandlimited fields, it is known that uniformly distributed location-unaware sensors do not infer the field uniquely.

We study asymmetric (statistical) distributions on location-unaware sensors, that may enable bandlimited field reconstruction. If the location of each sensor is random, then a bandlimited field operating on this randomness is observed. We propose a model for estimation of periodic bandlimited fields from samples obtained by sensors whose location is unknown but restricted to a random point on an *equi-spaced discrete grid*.<sup>1</sup> Oversampling will be used to overcome location unawareness.

With oversampling, samples obtained from location-unaware sensors can be clustered together to infer which sample belongs to which spatial location on the equi-spaced grid where the sensors are present. If  $p$  is the probability with which a sensor falls at a given location, then  $\approx np$  will be the number of samples obtained from there, as  $n$  becomes large. The success of this clustering scheme will depend on the probability distribution that governs sensor placement on the grid. By assigning locations to samples based on their expected frequency, the field can be *detected*. The *main result* of this work is to find the *optimal* probability distribution on sensor locations that minimizes the detection error-probability of the underlying spatial field.

This is a new model, the likes of which has not been explored in literature to the best of our knowledge. The proposed field detection algorithm uses results from statistics and information theory. Since most real-world measurements are corrupted by noise we also include an extension of the algorithm to field estimation from noisy samples which involves topics from machine learning. The following sections present an overview of our model, main results and the simulations used to validate our work.

<sup>1</sup>This may arise in scenarios where location information is masked to preserve the identity of the sensors, or to reduce the amount of data that needs to be transmitted.

# 2 Sampling Model

Here we discuss the field model and the manner in which sensors are deployed in the field.

## 2.1 Spatial field model

The spatial field  $g(t)$  is assumed to be periodic, real-valued, bounded and bandlimited. Without loss of generality (WLOG), the period of  $g(t)$  is fixed to 1. Then, the Fourier series of  $g(t)$  is

$$g(t) = \sum_{k=-b}^b a[k] \exp(j2\pi kt) \quad (1)$$

where  $a[k]$  are the Fourier series coefficients of  $g(t)$  and  $b$  is a *known* bandwidth parameter. Since  $g(t)$  is real valued,  $a[-k] = a[k]^*$  (conjugate symmetry). For simplicity of notation, define  $s_b := 1/(2b + 1)$  as a spacing parameter. Since there are  $2b + 1$  unknown Fourier series coefficients  $a[-b], \dots, a[b]$ , the  $2b + 1$  field values  $(g(0), g(s_b), \dots, g(2bs_b))$  uniquely specify the field.

## 2.2 Sensor deployment model

A discrete-valued non-uniform distribution is considered for bandlimited field inference. It will be assumed that a sensor is at location  $T$  such that  $T = is_b$  with probability  $p_i$  where  $i = 0, 1, \dots, 2b$  and  $\sum_{i=0}^{2b} p_i = 1$ . Correspondingly,

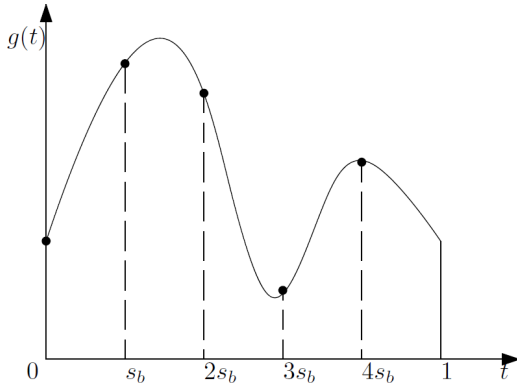
$$g(T) = g(is_b) \text{ with probability } p_i, \text{ for } 0 \leq i \leq 2b. \quad (2)$$

In our model (illustrated in Fig. 1), the sensor falls at  $is_b, 0 \leq i \leq 2b$  but its location, that is the index  $i$ , is *not known*. The parameter  $\vec{p} := p_0, p_1, \dots, p_{2b}$  will be treated as a *design choice* to optimize any performance criterion. It will be assumed that elements of  $\vec{p}$  are distinct (to break symmetry in the distribution of sensor-locations). WLOG, assume that

$$p_0 < p_1 < \dots < p_{2b}. \quad (3)$$

It will be assumed that i.i.d. samples  $g(T_1), g(T_2), \dots, g(T_n)$  are available for the detection of spatial field, where  $n$  corresponds to oversampling.<sup>2</sup>

<sup>2</sup>It is desirable to address the setup where each sensor's location  $T$  is realized from an asymmetric continuous distribution supported in  $[0, 1]$ . This problem is nonlinear and presents several difficulties, some of which have been investigated by us and are included in the detailed report.



**Fig. 1:** Sampling model for a signal  $g(t)$  with  $b = 2$  i.e.  $s_b = 1/5$

### 3 Field detection and its performance

Field detection and the choice of  $\vec{p}$  is discussed in this section.

#### 3.1 The field detection algorithm

Based on the readings  $g(T_1), g(T_2), \dots, g(T_n)$ , the field  $g(t)$  has to be detected. From (3) we know that  $p_i$  are distinct. We also assume that the field values  $g(is_b)$  are distinct<sup>3</sup>. Each sensor records  $g(is_b)$  with probability  $p_i$ . The following *clustering algorithm* will be used to ascertain the field samples  $g(is_b)$ , which specify the entire field  $g(t)$ :

1. The readings  $Y_1 := g(T_1), \dots, Y_n := g(T_n)$ , with  $T_i$  unknown and in the set  $\{0, s_b, \dots, 2bs_b\}$ , are collected.
2. The values  $Y_1, Y_2, \dots, Y_n$  are clustered into (*value, type*) pairs. Equal values (*value*) in  $Y_1, Y_2, \dots, Y_n$  are collected together and the number of equal values (*type*) is recorded.
3. Empirical probabilities  $\text{type}/n$  for each *value* are calculated. For large  $n$ , the empirical probability  $\text{type}/n$  of each *value* will be near the correct  $p_i$  in  $\vec{p}$ .
4. The *value* with smallest empirical probability is assigned to  $g(0)$ , the *value* with next smallest empirical probability is assigned to  $g(s_b)$ , and so on till  $g(2bs_b)$ .

<sup>3</sup>If  $\vec{a}$  is the realization of a continuous random distribution, then this condition will hold almost surely. A violation of this condition implies that  $\sum_{k=-b}^b a[k](\exp(j2\pi km s_b) - \exp(j2\pi kn s_b)) = 0$ . That is, a linear combination of  $\vec{a}$ -a continuous random variable-is zero with probability one.

**Example 3.1.** Consider a signal  $g_1(t)$  with bandwidth parameter  $b = 1$ , and  $s_b = \frac{1}{2b+1} = \frac{1}{3}$ . The field values are known to be  $g_1(0) = 1.06, g_1(1/3) = 1.80, g_1(2/3) = 0.14$ .

The field is sampled using  $n = 10$  randomly realized values of sensor's location in the set  $\{0, 1/3, 2/3\}$ . The 10 observed samples were 1.80, 0.14, 0.14, 1.06, 1.80, 0.14, 1.80, 1.06, 0.14, 0.14. The (*value, type*) pairs are (1.06, 2), (1.80, 3), and (0.14, 5). The above algorithm concludes that  $g_1(0) = 1.06, g_1(1/3) = 1.80, g_1(2/3) = 0.14$ , and is correct.

The field is again sampled using  $n = 10$  randomly realized values of sensor's location. This time, the 10 observed samples were 1.06, 0.14, 0.14, 1.06, 1.80, 0.14, 1.80, 1.06, 0.14, 0.14. The (*value, type*) pairs are (1.06, 3), (1.80, 2), and (0.14, 5). The above algorithm concludes that  $g_1(0) = 1.80, g_1(1/3) = 1.06, g_1(2/3) = 0.14$ , and is incorrect.

#### 3.2 Detection-Error Probability

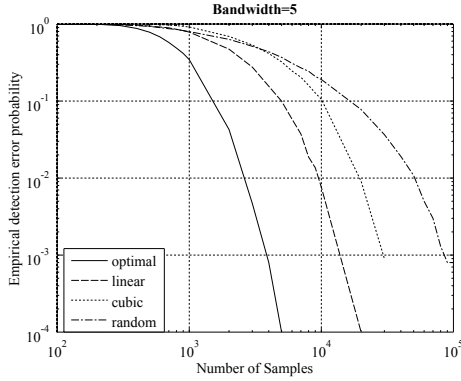
Define  $N_i := \sum_{j=1}^n \mathbb{1}[Y_j = g(is_b)]$  as the type of  $g(is_b)$  in  $n$  field observations. Then, in the above algorithm as  $n \rightarrow \infty$ , it is expected that

$$0 < N_0 < N_1 < \dots < N_{2b}. \quad (4)$$

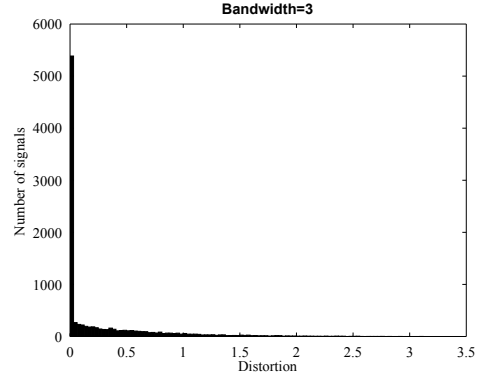
If the above event is violated, it results in erroneous field detection. We wish to maximize the probability of correct detection (or minimize the detection-error probability) in (4) which depends on the distribution  $\vec{p}$ . We formulate this problem as an optimization problem using Sanov's Theorem from the theory of large deviations, and solve it to obtain the following distribution  $\vec{p}$  that gives *minimum* detection-error probability for our field detection algorithm and is the *main result* of this work:

$$p_i = \frac{3(i+1)^2}{(b+1)(2b+1)(4b+3)} \text{ for } 0 \leq i \leq 2b. \quad (5)$$

The performance of the optimal distribution (as derived above) is verified against 3 other distributions for a randomly generated field using MATLAB. For each distribution the field is sampled as described in Section 2.2 and estimated using the proposed field detection algorithm. The results in Fig 2 (calculated over 10000 Monte Carlo trials) show that the optimal distribution performs best in terms of minimum empirical detection error probability



**Fig. 2:** Comparison of detection error-probabilities for different laws on  $\vec{p}$



**Fig. 3:** Histogram of Distortion for estimation of signals from noisy samples

## 4 Field Estimation from Noisy Samples

Each uncorrupted sample  $g(T_1), g(T_2), \dots, g(T_n)$  has a value equal to  $g(is_b)$  for some  $0 \leq i \leq 2b$  as discussed in Section 3.1. Assuming that the samples are corrupted by zero mean, i.i.d additive Gaussian noise with standard deviation  $\sigma$ , the field detection algorithm is modified as follows:

1. The readings  $Y_1 := g(T_1) + \eta_1, \dots, Y_n := g(T_n) + \eta_n$ , with  $T_i$  unknown and in the set  $\{0, s_b, \dots, 2bs_b\}$ , and  $\eta_i \sim \mathcal{N}(0, \sigma^2)$  are collected.
2. Since  $T_k = is_b$  with probability  $p_i$ , the readings  $Y_k$  form a Gaussian Mixture Model (GMM) with means  $g(is_b)$  and weights  $p_i$ .
3. The Expectation Maximization (EM) algorithm for clustering samples obtained from a GMM gives an estimate of the *weights* and *means* (analogous to *type* and *value* in the noiseless case) of the GMM ( $\sigma$  is assumed to be known.)
4. The *mean* with smallest *weight* is assigned to  $g(0)$ , the *mean* with next smallest *weight* is assigned to  $g(s_b)$ , and so on till  $g(2bs_b)$ .

The performance of the algorithm in this case is demonstrated in Fig 3 which is a histogram of distortion between 10000 randomly generated signals and their estimates from noisy samples. The distribution on the sensor locations is the one in (5). 10000 noisy samples are used to estimate each signal (as described above) and the distortion is computed as the ratio of the mean squared error between estimated and original signals to the energy of the original signal. Most of the signals are reconstructed with a low value of distortion. The signals

that give a high distortion are the ones in which the clusters formed by the noisy samples overlap to a large extent. It is known that the EM algorithm performs poorly for overlapping clusters and there exist approaches which claim to perform better in this case though their applicability to the present scenario remains to be seen.

## 5 Conclusions

This work proposes a new model for the detection and estimation of periodic bandlimited fields using location-unaware sensors. Our work lies at the intersection of signal processing, remote sensing, information theory and machine learning. Future directions include studying the effect of perturbing sensor locations and deploying sensors according to a continuous distribution in one period of the field.

## 6 Highlights

1. **Publication:** Mallick, Ankur, and Animesh Kumar. "Bandlimited field reconstruction from samples obtained on a discrete grid with unknown random locations." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
2. **Under Preparation:** Ankur Mallick, and Animesh Kumar. "On Bandlimited field reconstruction from samples obtained on a discrete grid with unknown random locations." to be submitted to IEEE Transactions on Signal Processing.
3. **Award:** Received the IIT Bombay Undergraduate Research Award (URA 03) for exceptional work in the project

# Estimation of Spatial Fields from Samples obtained at Unknown Random Locations

*Submitted in partial fulfillment of the requirements*

*of the degree of*

*Bachelor of Technology and Master of Technology*

*by*

Ankur Mallick

(Roll no. 110110013)

*Supervisor:*

Prof. Animesh Kumar



Department of Electrical Engineering

Indian Institute of Technology Bombay

2016

*Dedicated to my parents*

# Dissertation Approval

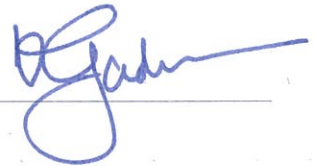
This dissertation entitled “**Estimation of Spatial Fields from Samples obtained at Unknown Random Locations**”, submitted by Ankur Mallick (Roll No. 110110013), is approved for the award of degree of Bachelor of Technology and Master of Technology in Electrical Engineering.

## Examiners

Prof. Sibi Raj Pillai



Prof. Vikram Gadre



## Supervisor

Prof. Animesh Kumar



## Chairman

Prof. Maryam Shojaei Baghini



Date: 24 June 2016

Place: IIT BOMBAY

# Declaration of Authorship

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/ data/ fact/ source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature: .....



**Ankur Mallick**

110110013

Date: 24. June 2016



## *Abstract*

Sampling and estimation of spatial fields using sensors which are location unaware is an exciting topic. Here we study this topic under the assumption that the sensors are deployed according to a known probability distribution, under different scenarios.

The initial part of this work studies detection of bandlimited fields from location-unaware sensors that are restricted to a *discrete grid*. Oversampling is used to overcome the lack of location information. The samples obtained from location-unaware sensors are clustered together to infer the field using the probability distribution that governs sensor placement on the grid. Based on this clustering algorithm, the main result of this part is to find the *optimal* probability distribution on sensor locations that minimizes the detection error-probability of the underlying spatial field. The proposed clustering algorithm is also extended to include the case of signal reconstruction in the presence of sensor noise by treating the distribution of the noisy samples as a mixture model and using clustering to estimate the mixture model parameters.

In the later part of the work the restriction that sensor locations must lie on a discrete grid is removed. It is already known that location-unaware sensors deployed according to a uniform distribution cannot infer the field uniquely in the absence of order information on the sensor locations. We strengthen this result further and give a procedure for estimating the ordering of sensor locations which is absent in related work. It is also shown that even in the case where sensors are deployed according to a general (not necessarily uniform) distribution there exist several fields that cannot be inferred uniquely. This reinforces the need for restricting the sensor locations to a discrete grid or knowing the ordering on the sensor locations. These are the main results for this part of the work.

**Index terms:** Signal Sampling, Signal Reconstruction, Wireless Sensor Networks

# Contents

<b>Dissertation Approval</b>	<b>ii</b>
<b>Declaration of Authorship</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>List of Figures</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Review of Literature</b>	<b>4</b>
<b>3 Sampling Model and Review</b>	<b>6</b>
3.1 Spatial field model . . . . .	6
3.2 Sensor deployment model . . . . .	7
3.3 Useful mathematical results . . . . .	8
<b>4 Field Detection</b>	<b>10</b>
4.1 The field detection algorithm . . . . .	10
4.2 Statistical sufficiency of type . . . . .	11
4.3 Location assignment and KL divergence . . . . .	12
4.4 Remarks on the Algorithm . . . . .	13
<b>5 Performance of the Field Detection Algorithm</b>	<b>15</b>
5.1 Detection error-probability minimization . . . . .	15
5.2 Controlling the detection-error probability . . . . .	19
5.3 Simulation results . . . . .	19

---

<b>6</b>	<b>Field Estimation from Noisy Samples</b>	<b>22</b>
6.1	The field estimation algorithm . . . . .	22
6.2	Overview of the EM Algorithm . . . . .	23
6.3	Simulation results . . . . .	24
<b>7</b>	<b>Sampling with a Uniform Continuous Distribution</b>	<b>29</b>
7.1	Field reconstruction with samples at uniformly distributed locations . . . . .	29
7.2	Order information on sample locations . . . . .	31
<b>8</b>	<b>Sampling with a General Continuous Distribution</b>	<b>36</b>
8.1	Sampling Model . . . . .	36
8.2	Optimisation Problem Formulation . . . . .	37
8.3	Non-Uniqueness of solutions . . . . .	39
<b>9</b>	<b>Conclusions</b>	<b>42</b>
	<b>List of Publications</b>	<b>45</b>
	<b>Acknowledgements</b>	<b>46</b>

# List of Figures

3.1	Sampling model for a signal $g(t)$ with $b = 2$ i.e. $s_b = 1/5$ . . . . .	8
3.2	Probability Simplex illustrating Sanov's Theorem . . . . .	9
5.1	Detection error-probabilities for different laws on $\vec{p}$ and different bandwidths are compared. The four laws used include the optimal $\vec{p}$ in (5.16), a linear law, a cubic law, and a randomly generated $\vec{p}$ . Fields of bandwidth 3, 5, 10, and 20 are studied. As expected, the law in (5.16) is the best in performance in all cases. . .	20
5.2	Number of samples required to reduce the empirical detection error probability to 1% for fields of bandwidth 3, 5, 10, and 20 . . . . .	21
6.1	Results of the sampling and estimation experiment for 10000 randomly generated signals of bandwidth 3. Histograms of the distortion are plotted for each sample size ( $n$ ) . . . . .	26
6.2	Results of the sampling and estimation experiment for 10000 randomly generated signals of bandwidth 5. Histograms of the distortion are plotted for each sample size ( $n$ ) . . . . .	27
6.3	Results of the sampling and estimation experiment for 10000 randomly generated signals of bandwidth 10. Histograms of the distortion are plotted for each sample size ( $n$ ) . . . . .	27
6.4	Minimum pairwise squared Euclidean distance ( $d_g$ ) between the signal values at the sampling locations, is compared for signals of bandwidth 3, 5 and 10. Histograms of $d_g$ are plotted using 10000 randomly generated signals for each value of bandwidth . . . . .	28
7.1	Random bandlimited, periodic field $G(t)$ (period=1) sampled at locations $t_1, t_2,$ and $t_3$ . . . . .	32

8.1	$g(t)$ is a polynomial of known degree sampled at unknown points $T_1, T_2, \dots, T_n$ with arbitrary (unknown) ordering . . . . .	36
-----	--	----

# Chapter 1

## Introduction

The problem of localizing sensors in a wireless sensor network poses several challenges [1]. An alternative to expensive localization schemes is to work with sensors which are location unaware. Recently, bandlimited field estimation *without* any location information of the sensors in a distributed setup has been studied [2], [3]. This is an emerging field where the key idea is to utilize a multitude of such location-unaware sensors (oversampling) and leverage the random distribution on their spatial locations. Henceforth, *location-unaware sensors will be simply termed as sensors*.

Due to symmetry and shift-invariance properties of bandlimited fields, it is known that uniformly distributed sensors only infer the underlying field up to a shift and a flip [2]. We will show in this work that scaling the independent variable of the underlying field also leads to ambiguous estimation by uniformly distributed sensors. This motivates the pursuit of alternate approaches for estimating spatial fields from samples in the absence of location information.

In the first part of this work we have considered estimation of bandlimited fields from samples obtained by sensors whose location is restricted to a random point on an *equi-spaced discrete grid*.<sup>1</sup> The sensors are deployed according to an asymmetric statistical distribution on the grid points. Since the location of each sensor is random, a bandlimited field operating on this randomness is observed. Oversampling is used to overcome the lack of location information.

With oversampling, samples obtained from sensors can be clustered together to infer which sample belongs to which spatial location on the equi-spaced grid where the sensors are present. If  $p$  is the probability with which a sensor falls at a given location, then  $\approx np$  will be the number of

---

<sup>1</sup>This may arise in scenarios where location information is masked to preserve the identity of the sensors, or to reduce the amount of data that needs to be transmitted.

samples obtained from there, as  $n$  (total number of samples across all locations) becomes large. By assigning locations to samples based on their expected frequency, the field can be *detected*. We show that such a scheme is optimal in the sense of minimising the KL divergence between the distribution on the sample data and the underlying distribution on the sensor locations. The success of this clustering scheme will depend on the probability distribution that governs sensor placement on the grid. The *main result* of this part is to find the *optimal* probability distribution on sensor locations that minimizes the detection error-probability of the underlying spatial field.

Since samples obtained from most real world signals are corrupted by some noise, this work also extends the proposed clustering algorithms to the case of signal reconstruction from noisy samples. The distribution of the noisy samples is modeled as a mixture model and the special case of Gaussian noise is analysed to show that our approach works fairly well in most cases even in the presence of noise.

In the latter part of the work we have considered the case where the sensors are not constrained to lie on a discrete grid but can lie anywhere in the support of the field. As stated earlier if the sensor locations are distributed uniformly then the field cannot be inferred uniquely. To overcome this most existing works in this area [2], [3], [4], [5], assume that the order in which the sensors are located in the field is known. However methods for inferring the ordering have not been explored in literature to the best of our knowledge. Hence we have proposed a scheme for inferring the order of the sensor locations based on correlation between samples. Using this scheme fields can be estimated from sensors deployed according to a uniform distribution.

Lastly we consider the case where the field is sampled according to a general (not necessarily uniform) continuous distribution on its support. We observe that a large class of fields cannot be uniquely specified by samples obtained from such a distribution. Since the underlying field is not typically known in most sampling problems and most natural fields change with time it does not appear to be possible to guard against errors in field estimation in this case.

*Notation:* Space will be denoted by  $t$ . Spatial fields will be denoted by  $g(t)$  and its variants, and the Fourier Series coefficients will be denoted by  $a[k]$  and its variants.  $a^*$  will denote the complex conjugate of  $a$ .  $j = \sqrt{-1}$ .  $\mathcal{N}(\mu, \sigma^2)$  denotes a Gaussian distribution with mean  $\mu$  and standard deviation  $\sigma$ . Vectors are column-vectors. The probability and expectation operators will be denoted by  $\mathbb{P}$  and  $\mathbb{E}$  respectively. The indicator function of a set  $A$  will be denoted by  $\mathbb{1}(x \in A)$ . Uppercase letters such as  $A, Y, X$ , and  $N$  denote random variables while lowercase letters like  $a$  and  $x$  denote deterministic quantities. Probability density functions will be denoted

by  $f_Y(y)$  and its variants, and cumulative distribution functions (CDF) will be denoted by  $F_Y(y)$  and its variants, where  $Y$  is the corresponding random variable. Independent and identically distributed will be termed as i.i.d. It is assumed that there is an underlying probability space  $\Omega, \mathcal{F}, \mathbb{P}$  over which all probability events discussed in the paper are defined.

*Organization:* Chapter 2 reviews the existing literature on the topic. Chapters 3-5 deal with the problem of field estimation from samples obtained on a discrete grid without measurement noise. The sampling model is introduced in Chapter 3, the field detection algorithm is explained in Chapter 4 and the optimal distribution on the sampling locations for the proposed field detection algorithm is derived in Chapter 5. Chapter 6 extends the scheme to include measurement noise in the samples. Chapter 7 considers the case where the samples are obtained from a uniform continuous distribution on the sensor locations and introduces the scheme for inferring the order information on the sensor locations. Chapter 8 discusses sampling with general (not necessarily uniform) continuous distributions on the sensor locations. Conclusions are presented in Chapter 9.



# Chapter 2

## Review of Literature

Interest in this topic has developed relatively recently and hence the existing body of work is fairly small.

Estimation of bandlimited fields from samples taken at unknown but statistically distributed sampling locations was studied by Kumar [2], [3]. [2] deals with bandlimited field estimation from samples obtained from uniformly distributed sensors with known ordering on the sensor locations. [3] extends the problem to the case where the distribution on the sensor locations is unknown, such as when the field is sampled at random points by a mobile sensor (thus ensuring the availability of order information).

Reconstruction of discrete-time bandlimited fields from unknown sampling locations was studied by Marziliano and Vetterli [4] in a combinatorial setting. Once again the ordering on the sensors is assumed to be known and the problem is solved as a combinatorial optimization problem.

Estimation of periodic bandlimited signals with random sampling locations has been studied by Nordio et al. [6], where the samples are obtained by perturbing sensors that are located on a deterministic equi-spaced grid with the grid points as the mean locations. The notion of a discrete grid is introduced in this work and although the actual locations of the samples are unknown, the mean value of the location (grid point) of each sensor is known.

Estimating a bandlimited signal from a finite number of ordered non-uniform samples at unknown locations has been studied by Browning [5]. The end points are assumed to be fixed and the signal values at these points are assumed to be known.

In contrast with these works our work eliminates the need for order information on sensor locations by constraining the sensors to lie on a discrete grid. The grid point at which a particular

sample is obtained is random and unknown. We have designed an *optimal distribution* for deploying sensors on the grid points that minimizes the field detection error probability.

In situations where it might not be possible to restrict sensor locations to a discrete grid it is desirable to know the ordering of the sensors so that one of the above field estimation schemes may be applied. Thus through our proposed method of estimating the ordering of the sensors we have strengthened the existing methods in this area.

# Chapter 3

## Sampling Model and Review

The next 4 chapters deal with sensor deployment on a discrete grid. The sampling and sensor deployment models are discussed, and related theoretical results are reviewed, in this chapter.

### 3.1 Spatial field model

The spatial field  $g(t)$  is assumed to be periodic, real-valued, bounded and bandlimited. Without loss of generality (WLOG), the period of  $g(t)$  is fixed to 1. Then, the Fourier series of  $g(t)$  is

$$g(t) = \sum_{k=-b}^b a[k] \exp(j2\pi kt) \quad (3.1)$$

where  $a[k]$  are the Fourier series coefficients of  $g(t)$  and  $b$  is a *known* bandwidth parameter. Since  $g(t)$  is real valued,  $a[-k] = a[k]^*$  (conjugate symmetry). For simplicity of notation, define  $s_b := 1/(2b + 1)$  as a spacing parameter and  $\phi_k := \exp(j2\pi ks_b)$ ,  $-b \leq k \leq b$ . Let  $\Phi_b$  be defined as

$$\Phi_b = \begin{bmatrix} 1 & \dots & 1 \\ \phi_{-b} & \dots & \phi_b \\ \vdots & & \vdots \\ (\phi_{-b})^{2b} & \dots & (\phi_b)^{2b} \end{bmatrix}.$$

The columns of  $\Phi_b$  are orthogonal and a sampling theorem ensures that [6, 7]:

$$\vec{a} = (\Phi_b)^{-1} \vec{g} = \frac{1}{(2b+1)} \Phi_b^\dagger \vec{g}, \quad (3.2)$$

where  $\vec{a} = (a[-b], a[-b+1], \dots, a[b])^T$ , where  $\Phi_b^\dagger$  is the conjugate transpose of  $\Phi_b$ , and  $\vec{g} = (g(0), g(s_b), \dots, g(2bs_b))^T$ . From (3.2),  $\vec{a}$  and  $g(t)$  can be obtained using the samples in  $\vec{g}$ .

It will be assumed that  $g(is_b)$  are *distinct* for different values of  $i$ . This feature will be useful during clustering.<sup>1</sup>

## 3.2 Sensor deployment model

A discrete-valued non-uniform distribution is considered for bandlimited field inference. It will be assumed that a sensor is at location  $T$  such that  $T = is_b$  with probability  $p_i$  where  $i = 0, 1, \dots, 2b$  and  $\sum_{i=0}^{2b} p_i = 1$ . Correspondingly,

$$g(T) = g(is_b) \text{ with probability } p_i, \quad i = 0, 1, \dots, 2b \quad (3.3)$$

In our model (illustrated in Fig. 3.1), the sensor falls at  $is_b, 0 \leq i \leq 2b$  but its location, that is the index  $i$ , is *not known*. The parameter  $\vec{p} := p_0, p_1, \dots, p_{2b}$  will be treated as a *design choice* to optimize any performance criterion (see Chapter 4). It will be assumed that elements of  $\vec{p}$  are distinct (to break symmetry in the distribution of sensor-locations). WLOG, assume that

$$p_0 < p_1 < \dots < p_{2b}. \quad (3.4)$$

It will be assumed that i.i.d. samples  $g(T_1), g(T_2), \dots, g(T_n)$  are available for the detection of spatial field, where  $n$  corresponds to oversampling.<sup>2</sup>

<sup>1</sup>If  $\vec{a}$  is the realization of a continuous random distribution (as is the case for Fourier coefficients of a natural signal), then this condition will hold almost surely. A violation of this condition implies that  $\sum_{k=-b}^b a[k](\exp(j2\pi km s_b) - \exp(j2\pi kn s_b)) = 0$ . That is, a linear combination of  $\vec{a}$ —a continuous random variable—is zero with probability one.

<sup>2</sup>It is desirable to address the setup where each sensor's location  $T$  is realized from an asymmetric continuous distribution supported in  $[0, 1]$ . This problem will be discussed in a later chapter

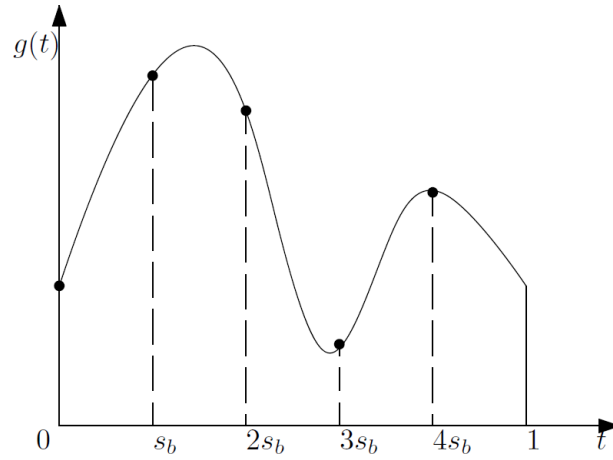


Figure 3.1: Sampling model for a signal  $g(t)$  with  $b = 2$  i.e.  $s_b = 1/5$

### 3.3 Useful mathematical results

To analyze the detection error-probability, large deviation analysis setup will be used. Sanov's theorem, which addresses the asymptotic likelihood properties with respect to an incorrect probability model, will be used [8, Chap 11.4]. Let  $X_1, \dots, X_n$  be i.i.d. random variables with discrete distribution  $\vec{p}$ . Then, the observed distribution of  $X_1, \dots, X_n$  lies in the closed set  $E$  with the following probability

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log_2 [\mathbb{P}(X_1^n \in E)] = -D(\vec{q}^* \parallel \vec{p}) \quad (3.5)$$

where  $\vec{q}^* = \arg \min_{\vec{q} \in E} D(\vec{q} \parallel \vec{p})$  is the distribution in  $E$  that is the closest to  $\vec{p}$  in the Kullback Leibler divergence or relative entropy terms. The quantity  $D(\vec{q}^* \parallel \vec{p})$  will be termed as the *error-exponent* in this work. Fig. 3.2 illustrates Sanov's theorem.

To determine if a statistic is sufficient the Fisher-Neyman Factorization Theorem [9, Chap 5.5] will be used. Let  $X_1, \dots, X_n$  be a random sample with probability density  $f_X(x|\theta)$ . The random vector  $T(X_1, \dots, X_n)$  is a sufficient statistic for  $\theta$  if and only if:

$$f_X(X_1, \dots, X_n|\theta) = f_\theta(T(X_1, \dots, X_n))H(X_1, \dots, X_n) \quad (3.6)$$

where  $f_\theta(T(X_1, \dots, X_n))$  is a function of  $T$  (depends on  $\theta$ ) and  $H(X_1, \dots, X_n)$  is independent of  $\theta$ .

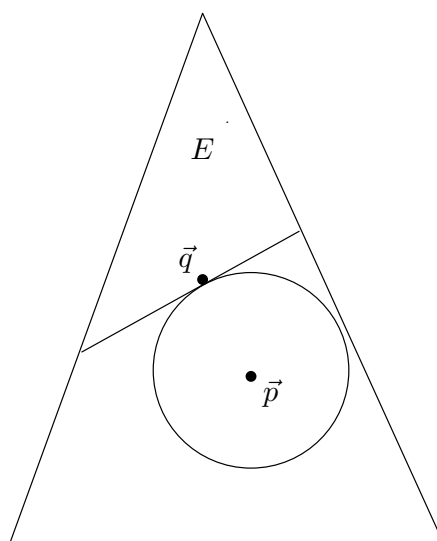


Figure 3.2: Probability Simplex illustrating Sanov's Theorem

The following inequalities will be used for optimization

$$\text{AM-GM: } \frac{x+y}{2} \geq \sqrt{xy}, \quad x, y \geq 0 \quad (3.7)$$

$$\text{Log-sum: } \sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}, \quad a_i, b_i > 0 \quad (3.8)$$

where  $a = \sum_{i=1}^n a_i$  and  $b = \sum_{i=1}^n b_i$ .

# Chapter 4

## Field Detection

A field detection algorithm is proposed in this chapter and the rationale behind choosing this algorithm is discussed.

### 4.1 The field detection algorithm

Based on the readings  $g(T_1), g(T_2), \dots, g(T_n)$ , the field  $g(t)$  has to be detected. From (3.4) and Section 3.1,  $\{g(is_b), p_i\}$  pairs are distinct in both the elements. Each sensor records  $g(is_b)$  with probability  $p_i$ . The following *clustering algorithm* will be used to ascertain the field samples  $g(is_b)$ , which specify the entire field  $g(t)$  (see (3.2)):

1. The readings  $Y_1 := g(T_1), \dots, Y_n := g(T_n)$ , with  $T_i$  unknown and in the set  $\{0, s_b, \dots, 2bs_b\}$ , are collected.
2. The values  $Y_1, Y_2, \dots, Y_n$  are clustered into (*value*, *type*) pairs. Equal values (*value*) in  $Y_1, Y_2, \dots, Y_n$  are collected together and the number of times a *value* is repeated in  $Y_1, Y_2, \dots, Y_n$  is recorded as *type*.
3. Empirical probabilities ( $\text{type}/n$ ) for each *value* are calculated. For large  $n$ , the empirical probability ( $\text{type}/n$ ) of each *value* will be near the correct  $p_i$  in  $\vec{p}$ .
4. The *value* with smallest empirical probability is assigned to  $g(0)$ , the *value* with next smallest empirical probability is assigned to  $g(s_b)$ , and so on till  $g(2bs_b)$ .

**Example 4.1.1.** Consider a signal  $g_1(t)$  with bandwidth parameter  $b = 1$ , and  $s_b = \frac{1}{2b+1} = \frac{1}{3}$ . The field values are known to be  $g_1(0) = 1.06, g_1(1/3) = 1.80, g_1(2/3) = 0.14$ .

The field is sampled using  $n = 10$  randomly realized values of sensor's location in the set  $\{0, 1/3, 2/3\}$ . The 10 observed samples were 1.80, 0.14, 0.14, 1.06, 1.80, 0.14, 1.80, 1.06, 0.14, 0.14. The (value, type) pairs are (1.06, 2), (1.80, 3), and (0.14, 5). The above algorithm concludes that  $g_1(0) = 1.06$ ,  $g_1(1/3) = 1.80$ ,  $g_1(2/3) = 0.14$ , and is correct.

The field is again sampled using  $n = 10$  randomly realized values of sensor's location. This time, the 10 observed samples were 1.06, 0.14, 0.14, 1.06, 1.80, 0.14, 1.80, 1.06, 0.14, 0.14. The (value, type) pairs are (1.80, 2), (1.06, 3), and (0.14, 5). The above algorithm concludes that  $g_1(0) = 1.80$ ,  $g_1(1/3) = 1.06$ ,  $g_1(2/3) = 0.14$ , and is incorrect.

Thus we see that the proposed field algorithm can detect the field both correctly and incorrectly. However we will show in the following two sections that:

1. The *type* as defined above, is a sufficient statistic with respect to field detection in our sampling model
2. The manner in which our algorithm assigns locations to the samples minimizes the KL divergence between the empirical and actual distributions on the sampling locations

These factors form the rationale behind our field detection algorithm.

## 4.2 Statistical sufficiency of type

Since the field is sampled at  $2b + 1$  distinct locations, the samples  $Y_1, Y_2, \dots, Y_n$  take  $2b + 1$  distinct values as discussed in Section 3.1. Let  $\vec{V} = V_0, V_1, \dots, V_{2b}$  denote the vector of these values. We observe that the field values at the sampling locations,  $g(k s_b), 0 \leq k \leq 2b$ , are an unknown (due to lack of location information) permutation of the elements of  $\vec{V}$ .

There are  $(2b + 1)!$  distinct permutations of  $\vec{V}$ . The goal of our field detection algorithm is to assign the correct location to each of the elements of  $\vec{V}$ . In other words we seek to estimate the permutation of  $\vec{V}$  that gives the correct values of the field at the sampling locations.

Let  $\vec{M} = M_0, M_1, \dots, M_{2b}$  denote the vector of types corresponding to the values in  $\vec{V}$  ( $M_k$  is the type of value  $V_k$ ). Each permutation of the values corresponds to a permutation of the types as well. Let  $\rho$  denote a permutation and let  $\vec{V}^\rho$  and  $\vec{M}^\rho$  denote the permuted versions of the values and types respectively. Our goal is to estimate the permutation  $\rho^*$  that leads to the correct assignment of values to the locations.



**Example 4.2.1.** Consider the second sampling experiment of the previous example where the (value, type) pairs are (1.80, 2), (1.06, 3), and (0.14, 5). Assigning  $V_0 = 1.80, V_1 = 1.06, V_2 = 0.14$  some possible permutations ( $\rho$ ) are the identity permutation  $\{(1.80, 2), (1.06, 3), (0.14, 5)\}$ , swapping  $V_0$  and  $V_1$ ,  $\{(1.06, 3), (1.80, 2), (0.14, 5)\}$ , swapping  $V_0$  and  $V_2$ ,  $\{(0.14, 5), (1.80, 2), (1.06, 3)\}$  and so on.

Assuming that the first element after permutation is assigned to  $g(0)$ , the second to  $g(s_b)$  and the third to  $g(2s_b)$ , we see that swapping  $V_0$  and  $V_1$  is the permutation ( $\rho^*$ ) that leads to the correct assignment of values to the locations.

The field is sampled according to a distribution  $\vec{p}$  as defined in Section 3.2. Thus, given that the field values at the sampling locations,  $g(ks_b), 0 \leq k \leq 2b$  correspond to a permutation  $\rho$  of  $\vec{V}$ , the permuted vector of types is  $\vec{M}^\rho$  and the distribution of the samples  $Y_1, Y_2, \dots, Y_n$  is given by:

$$f_Y(Y_1, Y_2, \dots, Y_n | \rho) = \prod_{k=0}^{2b} p_k^{M_k^\rho} \quad (4.1)$$

This satisfies the Fisher-Neymann Factorization Theorem. Here  $\rho$  is the parameter( $\theta$ ) that we wish to estimate,  $H(Y_1, Y_2, \dots, Y_n) = 1$  and the statistic  $T(Y_1, Y_2, \dots, Y_n)$  is  $\vec{M}$ , the vector of (types). Thus it is a sufficient statistic and no other statistic calculated from the samples can provide more information about the parameter  $\rho$ .

### 4.3 Location assignment and KL divergence

The field detection algorithm assigns values to sampling locations in  $0, s_b, \dots, 2bs_b$  in increasing order of *type*. We will show in this section that this assignment minimizes the KL divergence between the empirical and actual distributions on the data.

The data is sampled at location  $ks_b, 0 \leq k \leq 2b$  with probability  $p_k$  and we know from (3.4) that  $p_0 < p_1 < \dots < p_{2b}$ .

Consider an initial arbitrary assignment of values to sampling location such that the empirical probability(*type/n*) of the *value* assigned to location  $ks_b$  is  $q_k$ . Let  $\vec{q}$  be the vector of these empirical probabilities. Consider 2 locations  $l_1$  and  $l_2$  such that  $0 \leq l_1 < l_2 \leq 2b$ . Therefore  $p_{l_1} < p_{l_2}$ . Assume that our current location assignment is such that  $q_{l_1} > q_{l_2}$ . Thus:

$$D(\vec{q}|\vec{p}) = \sum_{k=0}^{2b} q_k \log_2 \frac{q_k}{p_k} \quad (4.2)$$

$$= D_0 + q_{l_1} \log_2 \frac{q_{l_1}}{p_{l_1}} + q_{l_2} \log_2 \frac{q_{l_2}}{p_{l_2}} \quad (4.3)$$

Consider a new assignment of the values such that the values at locations  $l_1$  and  $l_2$  are swapped and the values at all other locations are unchanged. Let the empirical probabilities for this assignment be stored in the vector  $\vec{r}$ . Thus  $r_{l_1} = q_{l_2}$ ,  $r_{l_2} = q_{l_1}$  and  $r_k = q_k$  for all  $k \neq l_1, l_2$ . Thus:

$$D(\vec{r}|\vec{p}) = \sum_{k=0}^{2b} r_k \log_2 \frac{r_k}{p_k} \quad (4.4)$$

$$= D_0 + q_{l_2} \log_2 \frac{q_{l_2}}{p_{l_1}} + q_{l_1} \log_2 \frac{q_{l_1}}{p_{l_2}} \quad (4.5)$$

The difference between the KL divergence for the two arrangements is:

$$D(\vec{q}|\vec{p}) - D(\vec{r}|\vec{p}) = (q_{l_1} - q_{l_2}) \log_2 \frac{p_{l_2}}{p_{l_1}} > 0 \quad (4.6)$$

Thus  $D(\vec{q}|\vec{p}) > D(\vec{r}|\vec{p})$  and we can see that the process of swapping values can be repeated until a state of minimum KL divergence is reached which in fact corresponds to the case where the empirical probabilities follow the same ordering as the actual probabilities or in other words when the *values* are arranged in increasing order of *type*. Due to the asymmetry in the distribution this is in fact the *unique* minimum.

## 4.4 Remarks on the Algorithm

In the previous two sections we have seen that the proposed algorithm uses all the available information from the samples for location assignment and finds the location assignment that most closely matches the empirical distribution on the data to the actual distribution on the sampling locations. Hence we expect it to yield correct location assignment in the asymptotic sense. A few concluding remarks on this algorithm will help in motivating the following chapters and our main results:

1. Although our work deals with periodic bandlimited fields, the detection algorithm can be

generalized to any finite support/periodic field which can be uniquely specified by samples at a finite number of locations. For eg-Finite degree polynomials.

2. For any distribution on sampling locations such that  $p_k = p_l$  for some  $k \neq l$  the algorithm breaks down since there will be more than one ordering of the values corresponding to the minimum KL divergence (swapping the values at assigned to locations  $k s_b$  and  $l s_b$  yields the same KL divergence but a different estimated field).
3. Even for the class of asymmetric distributions the performance of our algorithm will depend on the exact values of the elements of  $\vec{p}$  and it would be desirable to find the  $\vec{p}$  that gives the least errors in field detection. We shall look into this in detail in the next chapter.

# Chapter 5

## Performance of the Field Detection Algorithm

This chapter seeks the distribution  $\vec{p}$  on the sampling locations that gives the best performance in our field detection algorithm. For further discussions, define  $N_i := \sum_{j=1}^n \mathbb{1}[Y_j = g(is_b)]$  as the type of  $g(is_b)$  in  $n$  field observations. Then, in the above algorithm as  $n \rightarrow \infty$ , it is expected that

$$0 < N_0 < N_1 < \dots < N_{2b}. \quad (5.1)$$

If the above event is violated, it results in erroneous field detection. The probability of correct detection in (5.1) will be maximized by choosing the sensor deployment distribution  $\vec{p}$ .

### 5.1 Detection error-probability minimization

The spatial field is detected correctly when the condition in (5.1) is satisfied. Let  $P_e$  be the detection error-probability. The error-exponent (as the number of samples  $n$  gets large) in the detection error-probability will be maximized. Note that,

$$\begin{aligned} P_e &= \mathbb{P}\left[(0 < N_0 < N_1 < \dots < N_{2b})^c\right] \\ &= \mathbb{P}\left[\{N_0 = 0\} \cup \{N_0 \geq N_1\} \cup \dots \cup \{N_{2b-1} \geq N_{2b}\}\right] \end{aligned} \quad (5.2)$$

By applying the union-bound and the subset-inequality ( $A \subseteq B$  implies  $\mathbb{P}(A) \leq \mathbb{P}(B)$ ) in the above equation [10], we get

$$P_e \leq (2b + 1) \max \left\{ \mathbb{P}(N_0 = 0), \mathbb{P}(N_0 \geq N_1), \dots, \mathbb{P}(N_{2b-1} \geq N_{2b}) \right\} \quad (5.3)$$

$$\text{and } P_e \geq \max \left\{ \mathbb{P}(N_0 = 0), \mathbb{P}(N_0 \geq N_1), \dots, \mathbb{P}(N_{2b-1} \geq N_{2b}) \right\}. \quad (5.4)$$

From the above equations, the error-exponent in  $P_e$  is maximized if the error exponent of  $\max \left\{ \mathbb{P}(N_0 = 0), \mathbb{P}(N_0 \geq N_1), \dots, \mathbb{P}(N_{2b-1} \geq N_{2b}) \right\}$  is maximized. The constant factor  $(2b + 1)$  in (5.3) does not contribute to the error-exponent. The error-exponent maximization is addressed next.

A sensor falls at location 0 with probability  $p_0$ . With  $n$  randomly deployed sensors,

$$\mathbb{P}[N_0 = 0] = (1 - p_0)^n. \quad (5.5)$$

To compute  $\mathbb{P}[N_0 \geq N_1]$  and other similar events, Sanov's theorem will be used (see (3.5)). An empirical distribution  $\vec{q}$  will be found such that  $D(\vec{q} \parallel \vec{p})$  is minimum, which results in the error-exponent via Sanov's theorem (see (3.5)). The empirical distribution is  $\vec{q} = \left[ \frac{N_0}{n}, \frac{N_1}{n}, \dots, \frac{N_{2b}}{n} \right]$  and, from Sanov's theorem, the function to be minimized is

$$D(\vec{q} \parallel \vec{p}) = \sum_{i=0}^{2b} \frac{N_i}{n} \log_2 \frac{N_i}{np_i}$$

subject to  $\sum_{i=0}^{2b} \frac{N_i}{n} = 1$  and  $N_1 \leq N_0$ . (5.6)

The corresponding Lagrangian is

$$L = \sum_{i=0}^{2b} \frac{N_i}{n} \log_2 \frac{N_i}{np_i} + \lambda \left( \sum_{i=0}^{2b} N_i - n \right) + \mu(N_1 - N_0)$$

At the minima of  $D(\vec{q} \parallel \vec{p})$  in (5.6),

$$\frac{\partial L}{\partial N_i} = 0 \quad \text{for } 0 \leq i \leq 2b \quad (5.7)$$

The solutions of above equation are

$$N_0 = \frac{np_0}{e} 2^{-n(\lambda-\mu)}, N_1 = \frac{np_1}{e} 2^{-n(\lambda+\mu)}, \quad (5.8)$$

and,

$$N_i = \frac{np_i}{e} 2^{-n\lambda} \text{ for } i \geq 2. \quad (5.9)$$

The values of  $\mu$  and  $\lambda$  can be found by KKT conditions [11], but by using the log-sum and AM-GM inequalities in (3.8) and (3.7)  $\mu$  can be found directly as follows. Observe that  $\mu$  is only associated with  $N_0$  and  $N_1$ . The terms corresponding to  $N_0$  and  $N_1$  in (5.6) is lower-bounded by

$$\begin{aligned} \frac{N_0}{n} \log_2 \frac{N_0}{np_0} + \frac{N_1}{n} \log_2 \frac{N_1}{np_1} &\geq \frac{N_0 + N_1}{n} \log_2 \frac{N_0 + N_1}{n(p_0 + p_1)} \\ &\geq \frac{2\sqrt{N_0 N_1}}{n} \log_2 \frac{2\sqrt{N_0 N_1}}{n(p_0 + p_1)}. \end{aligned}$$

In the above equation, the minimum value requires that  $N_0 = N_1$ . This results in

$$\mu = \frac{1}{2n} \log_2 \frac{p_1}{p_0}, \text{ and } N_0 = N_1 = \frac{n}{e} 2^{-n\lambda} \sqrt{p_0 p_1} \quad (5.10)$$

In (5.8), the product  $N_0 N_1$  does not depend on  $\mu$ . So the minimum value of first two terms in  $D(\vec{q} \| \vec{p})$  is attained only when  $N_0 = N_1 = \frac{n}{e} 2^{-n\lambda} \sqrt{p_0 p_1}$ .

For finding  $\lambda$ , note that  $\sum_{i=0}^{2b} N_i = n$ . Using  $N_0, N_1$  from (5.10) and  $N_i$  from (5.9) results in

$$\lambda = -\frac{1}{n} \log_2 \left( \frac{e}{1 - (\sqrt{p_1} - \sqrt{p_0})^2} \right) \quad (5.11)$$

This value of  $\lambda$  gives

$$\begin{aligned} N_i &= \frac{np_i}{1 - (\sqrt{p_1} - \sqrt{p_0})^2} \\ \text{and } N_0 = N_1 &= \frac{n\sqrt{p_0 p_1}}{1 - (\sqrt{p_1} - \sqrt{p_0})^2}. \end{aligned}$$

Substitution of  $N_0, N_1, \dots, N_{2b}$  from the above equation in (5.6) results in the desired minimum

value of  $D(\vec{q}^* \parallel \vec{p})$ ,

$$D(\vec{q}^* \parallel \vec{p}) = \log_2 \frac{1}{1 - (\sqrt{p_1} - \sqrt{p_0})^2} \quad (5.12)$$

For  $N_i \geq N_{i+1}$ , the optimization constraint  $N_0 \geq N_1$  will get replaced by  $N_i \geq N_{i+1}$  in (5.6).

The analysis is identical and the result is

$$D(\vec{q}^* \parallel \vec{p}) = \log_2 \frac{1}{1 - (\sqrt{p_{i+1}} - \sqrt{p_i})^2}. \quad (5.13)$$

Let  $d_0 = \sqrt{p_0}$  and  $d_i = \sqrt{p_i} - \sqrt{p_{i-1}}$ ,  $1 \leq i \leq 2b$  and let  $d_{\min} = \min\{d_0, d_1, \dots, d_{2b}\}$ . Then  $d_{\min}$  will determine the value of the largest term in  $\max\{\mathbb{P}(N_0 = 0), \mathbb{P}(N_0 \geq N_1), \dots, \mathbb{P}(N_{2b-1} \geq N_{2b})\}$ . This is by Sanov's theorem which asserts that  $\mathbb{P}(N_i \geq N_{i+1}) \propto 2^{-nD(\vec{q}^* \parallel \vec{p})}$ . Consequently, the value of  $d_{\min}$  has to be maximized.

For maximizing  $d_{\min}$ , note that

$$(2b+1)d_{\min} \leq \sum_{i=0}^{2b} d_i = \sqrt{p_{2b}}. \quad (5.14)$$

To satisfy equality in (5.14),

$$\sqrt{p_0} = \frac{\sqrt{p_{2b}}}{2b+1} \text{ and } \sqrt{p_{i+1}} = \sqrt{p_i} + \frac{\sqrt{p_{2b}}}{2b+1}. \quad (5.15)$$

This relationship, along with  $p_0 + \dots + p_{2b} = 1$ , results in

$$p_i = \frac{3(i+1)^2}{(b+1)(2b+1)(4b+3)} \text{ for } 0 \leq i \leq 2b. \quad (5.16)$$

This law on  $\vec{p}$  ensures that the field detection error probability in (5.2) is *minimized*, and is the *main result* of this work.

For this law:

$$d_{\min} = d_0 = d_1 = \dots, d_{2b} = \frac{\sqrt{p_{2b}}}{2b+1} \quad (5.17)$$

$$\mathbb{P}(N_0 = 0) = \mathbb{P}(N_0 \geq N_1) = \dots = \mathbb{P}(N_{2b-1} \geq N_{2b}) = (1 - d_{\min}^2)^n \quad (5.18)$$

$$(1 - d_{\min}^2)^n \leq P_e \leq (2b+1)(1 - d_{\min}^2)^n \quad (5.19)$$

## 5.2 Controlling the detection-error probability

The probability law obtained in the previous section has the minimum detection error probability that converges to zero asymptotically. However asymptotic convergence has little practical applications. Instead, in a practical situation where a field is sampled at unknown locations on a discrete grid, it is desirable to find the number of samples  $n$  that can be drawn to guarantee that any field of bandwidth  $b$  can be estimated with detection-error probability,  $P_e$ , less than some threshold  $\epsilon \rightarrow 0$ .

From equation (5.19) a sufficient condition for  $P_e \leq \epsilon$  is:

$$(2b + 1)(1 - d_{\min}^2)^n \leq \epsilon \quad (5.20)$$

Taking natural logarithm on both sides and noting that  $\ln x < 0$  for  $x < 1$  gives us the condition:

$$n \ln(1 - d_{\min}^2) \geq \ln\left(\frac{\epsilon}{(2b + 1)}\right) \quad (5.21)$$

$$n \geq \frac{\ln(\hat{\epsilon})}{\ln(1 - d_{\min}^2)} \quad (5.22)$$

where  $\hat{\epsilon} = \frac{\epsilon}{(2b+1)}$

This is a sufficient condition on the number of samples required to reduce the detection error probability below a specified threshold for a field of given bandwidth.

## 5.3 Simulation results

Using MATLAB, the detection error-probability was compared for different laws on  $\vec{p}$ . Fields of bandwidth 3, 5, 10, and 20 respectively were used. The Fourier Series coefficients of each field was picked by a uniform random number generator. The number of randomly collected samples for each field was varied between 100 to  $10^5$  for the fields of bandwidth 3, 5, 10, and between 100 to  $10^6$  for the field of bandwidth 20. The empirical detection error-probability, when calculated using 10000 Monte-Carlo trials, is plotted in Fig. 5.1. The log scale on the Y-axis serves to model the error exponent. Four different methods to select  $\vec{p}$  were used for comparison. The selections include: (i) the optimal distribution in (5.16), (ii) a linear distribution  $\vec{p} = [\alpha, 2\alpha, \dots, (2b+1)\alpha]$ , (iii) a cubic distribution  $\vec{p} = [\alpha, 8\alpha, \dots, (2b+1)^3\alpha]$ , and (iv) ordered uniformly distributed



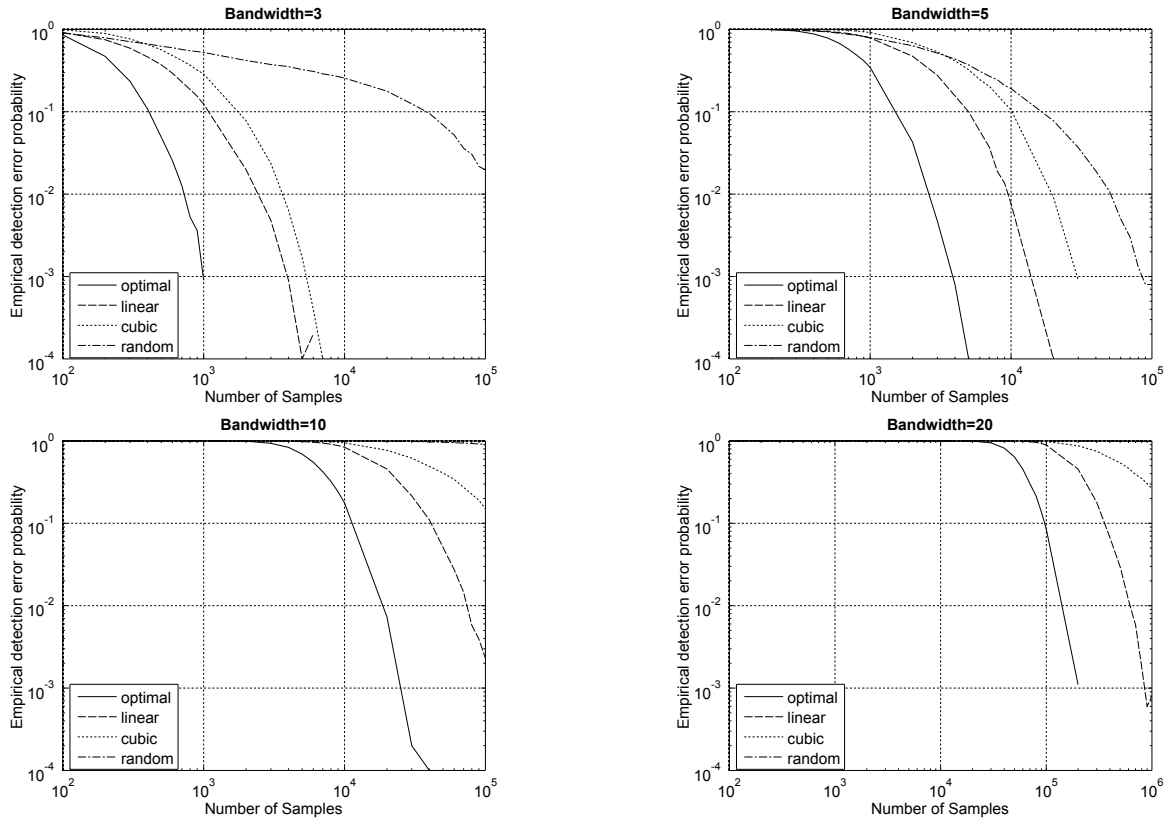


Figure 5.1: Detection error-probabilities for different laws on  $\vec{p}$  and different bandwidths are compared. The four laws used include the optimal  $\vec{p}$  in (5.16), a linear law, a cubic law, and a randomly generated  $\vec{p}$ . Fields of bandwidth 3, 5, 10, and 20 are studied. As expected, the law in (5.16) is the best in performance in all cases.

random variable realizations based distribution  $\vec{p} = \alpha[U(1), U(2), \dots, U(2b + 1)]$ . In all these cases,  $\alpha$  was selected to ensure  $\sum_{i=0}^{2b} p_i = 1$ . From the plots, the distribution discovered in (5.16) results in smallest detection error-probability (as expected) for all bandwidths. The number of samples required to reach zero detection error probability increases with increasing bandwidth but the optimal distribution in (5.16) is the one whose detection error probability decays fastest to zero in all cases.

For the optimal distribution we also simulated the number of samples required to reduce the empirical detection error probability  $P_e$  to 1% for fields of bandwidth 3, 5, 10, and 20 respectively. The Fourier Series coefficients of each field was picked by a uniform random number generator. A binary search algorithm was used to locate the sample size for  $0.01 - 0.001 \leq P_e \leq 0.01 + 0.001$ . The tolerance of 0.001 is used since the detection error probability is calculated as the fraction of incorrectly detected samples from Monte Carlo simulations and so it need not be exactly equal to 0.01. The results are plotted in Fig. 5.2.

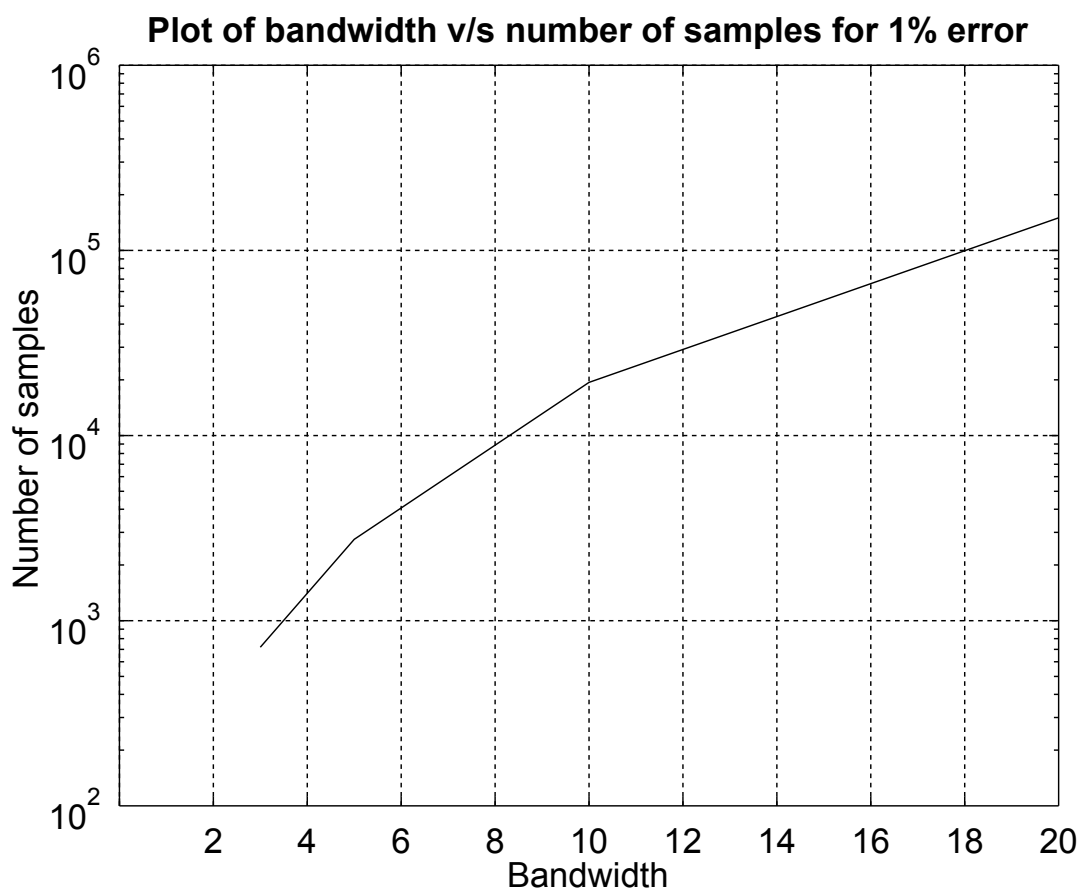


Figure 5.2: Number of samples required to reduce the empirical detection error probability to 1% for fields of bandwidth 3, 5, 10, and 20

# Chapter 6

## Field Estimation from Noisy Samples

In this chapter we will consider the case where the signal is sampled as described in Section 3.2 and the samples of the signal are then corrupted by zero mean, i.i.d, additive Gaussian noise with known standard deviation  $\sigma$ .

### 6.1 The field estimation algorithm

Each uncorrupted samples  $g(T_1), g(T_2), \dots, g(T_n)$  has a value equal to  $g(is_b)$  for some  $0 \leq i \leq 2b$  as discussed in Section 4.1. Assuming that the samples are corrupted by zero mean, i.i.d additive Gaussian noise with standard deviation  $\sigma$ , the field detection algorithm is modified as follows:

1. The readings  $Y_1 := g(T_1) + \eta_1, \dots, Y_n := g(T_n) + \eta_n$ , with  $T_i$  unknown and in the set  $\{0, s_b, \dots, 2bs_b\}$ , and  $\eta_i \sim \mathcal{N}(0, \sigma^2)$  are collected.
2. Since  $T_i = ks_b$  with probability  $p_k$ , the readings  $Y_i$  follow the probability distribution  $f_Y(y)$  given by the following Gaussian Mixture Model(GMM):

$$f_Y(y) = \sum_{k=0}^{2b} p_k G(y, g(ks_b), \sigma^2) \quad (6.1)$$

where

$$G(y, g(ks_b), \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-g(ks_b))^2}{2\sigma^2}} \quad (6.2)$$

3. The readings  $Y_1, Y_2, \dots, Y_n$  are clustered using the well known Expectation Maximization

(EM) algorithm [12], for GMM parameter estimation. The details of the algorithm are given in Section 6.2.

4. The algorithm gives an estimate of the *weights* and *means* (analogous to *type* and *value* in the noiseless case) of the GMM ( $\sigma$  is assumed to be known.) Let  $w_k$  and  $\mu_k$  be the estimated *weights* and the corresponding *means* respectively for  $0 \leq k \leq 2b$ . For large  $n$ , the estimated *weight*  $w_k$  for each *mean* will be near the correct  $p_k$  in  $\vec{p}$ .
5. The *mean* with smallest *weight* is assigned to  $g(0)$ , the *mean* with next smallest *weight* is assigned to  $g(s_b)$ , and so on till  $g(2bs_b)$ .

## 6.2 Overview of the EM Algorithm

The EM algorithm iteratively estimates the parameters of a GMM by creating a function for the expectation of the log likelihood function using the current estimate of parameters (E-step) and maximizing this expected log-likelihood to compute a new estimate of the parameters (M-step). These two steps are repeated until the algorithm converges to a maximum of the log likelihood function, to obtain an estimate of the means and the weights of each cluster.

In this work we use the EM algorithm for 'soft' segmentation of the data, as discussed in [13]. The data comprises of the readings  $Y_i$  which are segmented into clusters. A cluster  $C_k$  is defined as  $C_k = \{Y_i : Y_i = g(ks_b) + \eta_i, \eta_i \sim \mathcal{N}(0, \sigma^2)\}$ . Thus there are  $2b + 1$  clusters in this case. Define a membership matrix matrix  $\gamma$  and a random variable  $Z_i$  such that  $Z_i = k$  implies that  $Y_i$  belongs to cluster  $C_k$ . Then:

$$\gamma_{ik} = P(Z_i = k | Y_i) \quad (6.3)$$

The algorithm also requires an initial guess of the means  $\mu_k$  which is provided using the k-means++ algorithm [14]. The weights,  $w_k$  are initially assumed to be uniformly distributed. The variance of the clusters is known ( $\sigma^2$ ) and is fixed at this value. Each iteration of the algorithm involves the following two steps:-

1) E-Step:

$$\gamma_{ik} = \frac{G(Y_i, \mu_k, \sigma^2)w_k}{\sum_{k=0}^{2b} G(Y_i, \mu_k, \sigma^2)w_k} \quad (6.4)$$

2) M-Step:

$$\mu_k = \sum_{i=0}^n Y_i \gamma_{ik} \quad (6.5)$$

$$w_k = \frac{\sum_{i=0}^n \gamma_{ik}}{n} \quad (6.6)$$

The E-step computes  $\gamma_{ik}$  by applying Bayes' Theorem on the current parameter estimates. The M-step uses the computed value of  $\gamma_{ik}$  to come up with a new estimate of the parameters that maximizes the expected data log-likelihood. The values of  $\mu_k$  and  $w_k$  estimated in the M-step are substituted in the E-step of the next iteration and the process is repeated until the estimates converge.

We use a 'soft' formulation instead of a 'hard' assignment of each sample to a single cluster. The membership of a sample in a cluster is the posterior probability of the sample lying in that cluster. We use this formulation because if the values of any two or more  $g(k s_b)$  are very close (in terms of Euclidean distance) then the corresponding clusters tend to overlap. Thus, samples lying in such overlapping clusters could have originated from either of the sampling locations making it difficult to assign them to a single cluster. Moreover since we are only interested in the cluster means and weights and not in knowing which sample originated from which location, the 'soft' clustering makes more sense since it uses each sample to contribute to the estimate of the means and weights of all clusters.

The final GMM estimated by the EM algorithm is:

$$\hat{f}_Y(y) = \sum_{k=0}^{2b} w_k G(y, \mu_k, \sigma^2) \quad (6.7)$$

### 6.3 Simulation results

The field estimation algorithm of Section 6.1 was simulated using MATLAB for signals of bandwidth ( $b$ ) 3,5, and 10 respectively. For each value of  $b$ , 3 field estimation experiments were conducted with sample sizes  $n = 1000, 10000$ , and  $100000$  respectively. For each value of  $n$  10000 randomly generated signals were considered. For each signal, the samples were drawn as described in Section 3.2 and then each sample was corrupted by Gaussian noise ( $\mu = 0, \sigma = 0.05$ ). To measure the performance of the algorithm the following distortion metric was used:

$$D = \frac{\int_0^1 |\tilde{g}(t) - g(t)|^2 dt}{\int_0^1 |g(t)|^2 dt} \quad (6.8)$$

where  $D$  is the distortion,  $\tilde{g}(t)$  is the estimated signal,  $g(t)$  is the original signal and the limits of the integral are so chosen to cover 1 period of the signal.

The results of the simulation are shown in Fig. 6.1, Fig. 6.2, Fig. 6.3. Histograms of the distortion are plotted for each value of  $n$  for each bandwidth. The number of bins in each histogram is 100.

As can be seen from the histograms, the performance of the algorithm deteriorates on increasing the bandwidth. The number of signals estimated with a low value of distortion decreases as we go from bandwidth 3 to bandwidth 10. More than 50% of the signals are reconstructed with a very low value of distortion (first bin) for bandwidth 3 while the number reduces to about 18% for bandwidth 5 and less than 1% for bandwidth 10. Increasing the number of samples drawn improves the performance of the algorithm slightly especially for the higher bandwidths as attested by an increase in the height of the first bar of the histogram for bandwidths 5 and 10 on increasing the number of samples from 1000 to 10000.

We will now suggest an explanation for the deterioration in performance of the algorithm with increasing bandwidth. It is known that for a normal distribution,  $\mathcal{N}(\mu, \sigma^2)$ , 99.7% of the data lies within  $[\mu - 3\sigma, \mu + 3\sigma]$ . Consider 2 normal distributions with means  $x_1$  and  $x_2$ ,  $x_2 > x_1$ , and the same standard deviation,  $\sigma$ . If  $x_2 - 3\sigma < x_1 + 3\sigma$  or  $(x_2 - x_1)^2 < 36\sigma^2$  then the samples drawn from these 2 distributions form clusters that overlap to a large extent. In our experiments we have taken  $\sigma = 0.05$ , i.e.  $36\sigma^2 = 0.09$ .

In our experiments, samples are drawn from normal distributions with means  $g(ks_b)$ . For any field  $g(t)$  define  $d_g := \min\{(g(is_b) - g(js_b))^2, 0 \leq i, j \leq 2b, i \neq j\}$  as the minimum pairwise squared Euclidean distance between the means. We computed  $d_g$  for 10000 randomly generated signals of bandwidths 3, 5, and 10. Fig. 6.4 is a histogram of the results. It can be seen that as the bandwidth increases the value of  $d_g$  decreases and most of the values of  $d_g$  lie close to zero. In fact the percentage of signals with  $d_g$  below 0.09 ( $36\sigma^2$ ) increases from about 81% for bandwidth 3 to 98% for bandwidth 5 and 100% for bandwidth 10. This is expected since the number of sampling locations increases with bandwidth but the sampling interval remains in  $[0, 1]$  (one period) and hence the probability of the field values at 2 sampling locations lying close to each other increases. As discussed above if  $d_g$  falls below  $36\sigma^2$  then the corresponding clusters

overlap to a large extent which creates problems in the convergence of the EM algorithm.

The problem of overlapping clusters is a common problem in clustering especially with the EM algorithm. Several approaches attempting to solve this problem exist in literature such as [15], [16]. The application of these approaches to the present problem remains to be studied.

In all our analysis, the distribution of the noise is assumed to be Gaussian. If the distribution of the noise is non-Gaussian, the mixture-model and the clustering algorithm will have to be changed to suit the noise distribution.

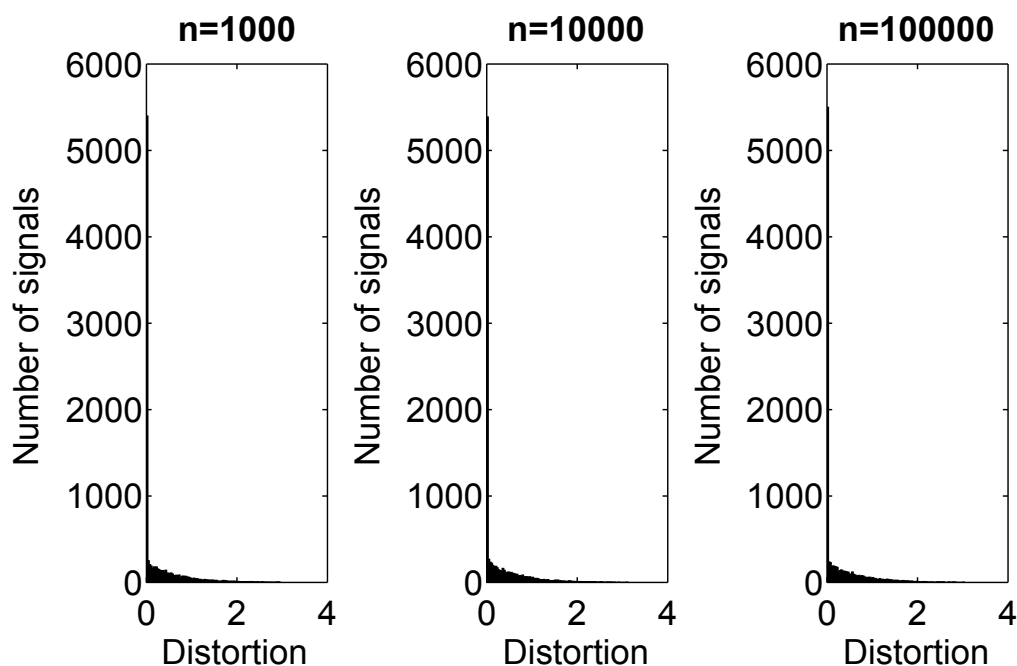


Figure 6.1: Results of the sampling and estimation experiment for 10000 randomly generated signals of bandwidth 3. Histograms of the distortion are plotted for each sample size ( $n$ )

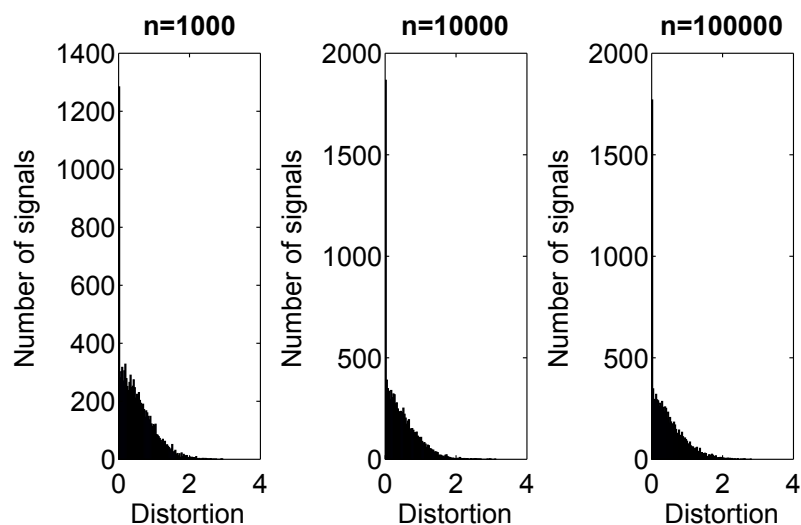


Figure 6.2: Results of the sampling and estimation experiment for 10000 randomly generated signals of bandwidth 5. Histograms of the distortion are plotted for each sample size ( $n$ )

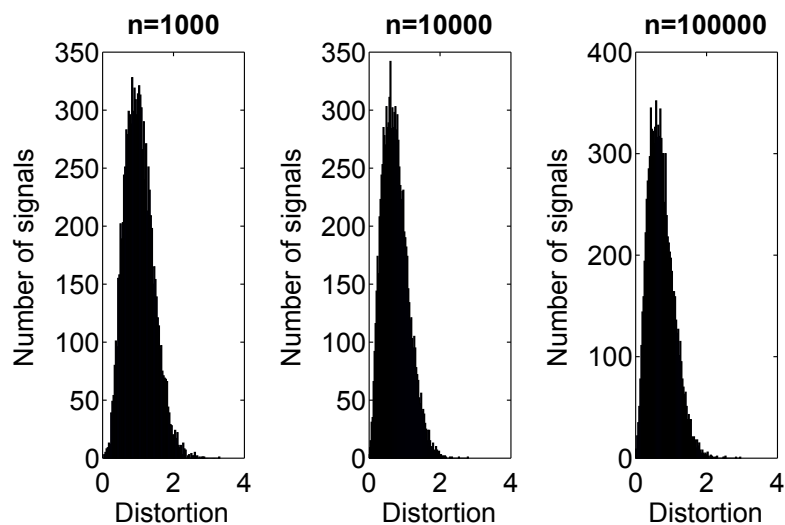


Figure 6.3: Results of the sampling and estimation experiment for 10000 randomly generated signals of bandwidth 10. Histograms of the distortion are plotted for each sample size ( $n$ )



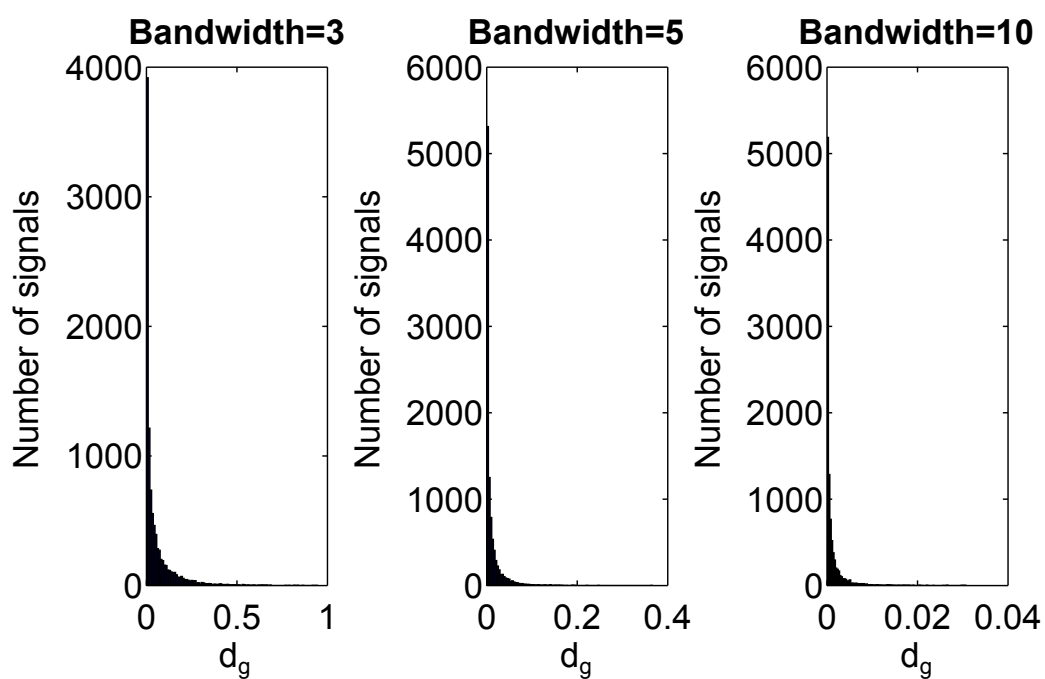


Figure 6.4: Minimum pairwise squared Euclidean distance ( $d_g$ ) between the signal values at the sampling locations, is compared for signals of bandwidth 3, 5 and 10. Histograms of  $d_g$  are plotted using 10000 randomly generated signals for each value of bandwidth

# Chapter 7

## Sampling with a Uniform Continuous Distribution

The sampling model introduced in our work had two main simplifying assumptions - the restriction of sampling locations to a discrete grid and the absence of measurement noise. The previous chapter dealt with the case where samples were corrupted by Gaussian noise. In this and the following chapter we consider the problem of estimating a field with samples obtained from a known continuous distribution on its support.

Some of the works that deal with this case assume that order information on the sample locations are known. While this is implicit in the case of sampling with a mobile sensor as in [3], it is not clear how this information may be obtained if the sensors are static. In this chapter we will also discuss a means of obtaining order information on sampling locations for static sensors.

### 7.1 Field reconstruction with samples at uniformly distributed locations

Here we consider the case where the unknown sampling locations are realized according to a uniform distribution anywhere in one period of the field, that is,  $T \sim \text{Uniform}[0, 1]$ . Let  $U_1, U_2, \dots, U_n$  be the (random) sampling locations; then, the corresponding sampled field values

are  $g(U_1), g(U_2), \dots, g(U_n)$ . The empirical cumulative distribution function

$$F_{g,n}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(g(U_i) \leq x) \quad (7.1)$$

for  $x \in [-1, 1]$  completely characterizes the field values  $g(U_1), g(U_2), \dots, g(U_n)$  (up to a permutation) and vice-versa. By the Glivenko-Cantelli theorem, for every  $x \in [-1, 1]$ ,  $F_{g,n}(x)$  in (7.1) converges almost surely to  $\mathbb{P}(g(U) \leq x)$ . For a uniform random variable,  $\mathbb{P}(g(U) \leq x)$  corresponds to length of the level set  $\{u : g(u) \leq x\}$ . It has been shown that a bandlimited field when shifted or space-reversed results in the same length of level set. That is,

$$\mathbb{P}(g(U) \leq x) = \mathbb{P}(g_1(U) \leq x) = \mathbb{P}(g_2(U) \leq x) \quad (7.2)$$

where  $g_1(t) = g(t - \theta)$  and  $g_2(t) = g(\theta - t)$  for any  $\theta \in [0, 1]$ . This means by observing the distribution  $\mathbb{P}(g(U) \leq x), x \in [-1, 1]$ , the field  $g(t)$  cannot be inferred due to ambiguity in phase and direction. However, it is not clear if these are the only ambiguities in the estimation of the field. In other words, is it possible to claim that the field can be obtained up to a delay and direction ambiguity as hinted in (7.2)? It is shown next that *scale ambiguity* is also present and this makes sampling a spatial field with uniformly distributed sensors difficult.

Let  $g(t)$  be a field with bandwidth  $2\pi$ . Consider the field  $g_3(t) = g(mt)$  for any positive integer  $m < b$ . Then  $g_3(t)$  is bandlimited with bandwidth up to  $2b\pi$ . It will be shown that

$$\mathbb{P}(g(U) \leq x) = \mathbb{P}(g_3(U) \leq x). \quad (7.3)$$

The core idea behind the proof is the accounting of the length of level set. Let

$$\{u : g(u) \leq x\} = [t_0, t_1] \cup [t_2, t_3] \cup \dots \cup [t_{N-1}, t_N] \quad (7.4)$$

where  $t_0, t_1, \dots, t_N$  depend on  $g(u)$  and  $x$ . Then, for  $m = 2$ ,

$$\{u : g_3(u) \leq x\} \quad (7.5)$$

$$= \{u : g(2u) \leq x\} \quad (7.6)$$

$$= \left[ \frac{t_0}{2}, \frac{t_1}{2} \right] \cup \dots \cup \left[ \frac{t_{N-1}}{2}, \frac{t_N}{2} \right] \cup \left[ \frac{t_0+1}{2}, \frac{t_1+1}{2} \right] \cup \dots \cup \left[ \frac{t_{N-1}+1}{2}, \frac{t_N+1}{2} \right]. \quad (7.7)$$

Observe that the lengths of level sets in (7.4) and (7.7) are equal to

$$(t_1 - t_0) + (t_3 - t_2) + \dots + (t_N - t_{N-1}) \quad (7.8)$$

This is true for any  $g(t)$  and any  $x \in [-1, 1]$ . So  $g(t)$  and  $g(2t)$  have the same level sets and consequently same distribution  $\mathbb{P}(g(U) \leq x)$ . This result can be also shown in a similar manner for  $m = 3, \dots, b$ . Thus, even if  $n \rightarrow \infty$ , the field  $g(t)$  cannot be inferred uniquely from  $F_{g,n}(x)$  which converges to  $P(g(U) \leq x)$ ,  $x \in [-1, 1]$ .

This result shows that it is not possible to uniquely estimate a field from samples obtained according to a uniform continuous distribution on its support.

## 7.2 Order information on sample locations

It is shown in [2] that it is possible to infer a field uniquely from samples obtained according to a uniform continuous distribution on its support if the order in which the samples are collected is known. Here we discuss how this order information may be obtained. Consider a randomly generated spatial field that is periodic, real-valued and bounded (period assumed to be 1) given by:

$$G(t) = \sum_{k=-b}^b A[k] \exp(j2\pi kt) \quad (7.9)$$

The Fourier Series coefficients  $A[k] = X_k + jY_k$  are generated randomly and vary with time ( $A[k] = A[-k]^*$  since the field is real valued). If sensors are deployed at points  $t_1, t_2$  and  $t_3$  in the interval  $(0, 1)$  as shown in Fig 7.1, then we expect that as the field  $G(t)$  varies randomly, the field values  $g(t_1), g(t_2)$  will be highly correlated (due to continuity) and will be uncorrelated with  $g(t_3)$ . In other words if a large number of sensors are deployed at fixed but unknown locations and are used to sample a field that changes randomly with time, then the readings of a sensor will have the highest correlation with the readings of the sensor nearest to it.

To see this mathematically we need to make the following assumptions which are consistent with our simulations in Chapter 5:

1. All  $X_k, Y_k$  are i.i.d
2.  $\mathbb{E}[X_k] = 0, \mathbb{E}[Y_k] = 0$

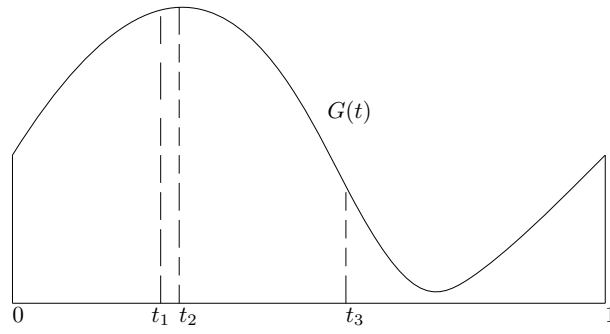


Figure 7.1: Random bandlimited, periodic field  $G(t)$  (period=1) sampled at locations  $t_1, t_2,$  and  $t_3$

$$3. \mathbb{E}[X_k^2] = S^2, \mathbb{E}[Y_k^2] = S^2$$

Let sensors be deployed at  $t_1, t_2 \in (0, 1)$  and let the corresponding field samples be  $G(t_1), G(t_2)$ . Consider the correlation coefficient between the samples:

$$r = \frac{\text{Cov}(G(t_1), G(t_2))}{\sqrt{\text{Var}(G(t_1))\text{Var}(G(t_2))}} \quad (7.10)$$

The sample mean for  $G(t_1)$  is given by:

$$\mathbb{E}[G(t_1)] = \mathbb{E}\left[\sum_{k=-b}^b A[k] \exp(j2\pi kt_1)\right] \quad (7.11)$$

$$= \sum_{k=-b}^b \mathbb{E}[X_k + jY_k] \exp(j2\pi kt) = 0 \quad (7.12)$$

Since  $\mathbb{E}[X_k] = 0, \mathbb{E}[Y_k] = 0$ . Therefore the sample variance is given by:

$$\text{Var}(G(t_1)) = \mathbb{E}[G(t_1)^2] \quad (7.13)$$

$$= \mathbb{E}\left[\sum_{k=-b}^b \sum_{l=-b}^b A[k]A[l] \exp(j2\pi(k+l)t_1)\right] \quad (7.14)$$

Since all  $X_k, Y_k$  are i.i.d and zero mean the above expression reduces to:

$$\text{Var}(G(t_1)) = \mathbb{E}[X_0^2] + T_1 + T_2 \quad (7.15)$$

$T_1$  is given by:

$$T_1 = \mathbb{E}\left[\sum_{k=-b, k \neq 0}^b A[k]^2 \exp(j4\pi kt_1)\right] = 0 \quad (7.16)$$

since  $A[k]^2 = X_k^2 - Y_k^2 + 2jX_kY_k$  and all  $X_k, Y_k$  are i.i.d and zero mean, and  $\mathbb{E}[X_k^2] = \mathbb{E}[Y_k^2] = S^2$

$T_2$  is given by:

$$T_2 = \mathbb{E}\left[\sum_{k=-b, k \neq 0}^b A[k]A[-k] \exp(j2\pi(0))\right] = 0 \quad (7.17)$$

$$= \sum_{k=-b, k \neq 0}^b \mathbb{E}[X_k^2 + Y_k^2] \quad (7.18)$$

$$= 4bS^2 \quad (7.19)$$

Thus:

$$\text{Var}(G(t_1)) = (1 + 4b)S^2 \quad (7.20)$$

Since the final expression is independent of  $t_1$  we have:

$$\text{Var}(G(t_2)) = (1 + 4b)S^2 \quad (7.21)$$

A similar procedure is followed to calculate the covariance:

$$\text{Cov}(G(t_1), G(t_2)) = \mathbb{E}[G(t_1)G(t_2)] \quad (7.22)$$

$$= \mathbb{E}\left[\sum_{k=-b}^b \sum_{l=-b}^b A[k]A[l] \exp(j2\pi(kt_1 + lt_2))\right] \quad (7.23)$$

$$= \mathbb{E}[X_0^2] + T_1 + T_2 \quad (7.24)$$

This time  $T_1$  is given by:

$$T_1 = \mathbb{E}\left[\sum_{k=-b, k \neq 0}^b A[k]^2 \exp(j2\pi k(t_1 + t_2))\right] = 0 \quad (7.25)$$

for the same reasons as before.

$T_2$  is given by:

$$T_2 = \mathbb{E}\left[\sum_{k=-b, k \neq 0}^b A[k]A[-k] \exp(j2\pi k(t_1 - t_2))\right] = 0 \quad (7.26)$$

$$= \sum_{k=-b, k \neq 0}^b (\mathbb{E}[X_k^2 + Y_k^2]) \exp(j2\pi k(t_1 - t_2)) \quad (7.27)$$

$$= 4S^2 \sum_{k=1}^b \cos[2\pi k(t_1 - t_2)] \quad (7.28)$$

Therefore the covariance is given by:

$$\text{Cov}(G(t_1), G(t_2)) = S^2(1 + 4 \sum_{k=1}^b \cos[2\pi k(t_1 - t_2)]) \quad (7.29)$$

Substituting the variances and covariance in (7.10) we get the following expression for the correlation coefficient:

$$r = C_0 + C_1 \sum_{k=1}^b \cos[2\pi kd] \quad (7.30)$$

Here  $C_0 = \frac{1}{4b+1}$ ,  $C_1 = \frac{4}{4b+1}$ ,  $d = |t_1 - t_2|$  since  $\cos(-x) = \cos(x)$ . Since  $t_1, t_2 \in (0, 1)$ ,  $d \in (0, 1)$  and so the functions  $\cos(2\pi kd)$  for each  $k$  have common maxima at  $d \rightarrow 0^+$  and  $d \rightarrow 1^-$ . Hence  $r$  is maximum for values of  $d$  close to 0 or 1. We are interested in the maxima of  $d$  near 0 since this means the sensor locations  $t_1, t_2$  are close to each other. Keeping this in mind the following procedure for finding the order informations is proposed:

1.  $n$  sensors are deployed uniformly at unknown locations  $t_1, t_2 \dots t_n$  with arbitrary ordering
2. Two other sensors  $S_0, S_1$  are fixed at the endpoints 0 and 1 respectively. Their locations are known and the remaining  $n$  sensors are grouped into two categories, those closer  $S_0$  than to  $S_1$  (Group 0) those closer  $S_1$  than to  $S_0$  (Group 1)<sup>1</sup>. respectively)
3. Let the number of sensors in Group 0 (excluding  $S_0$ ) be  $n_0$  and the number of sensors in Group 1 ((excluding  $S_1$ )) be  $n_1$  ( $n_0 + n_1 = n$ )
4. All sensors (including  $S_0, S_1$ ) take  $m$  readings of the field at  $m$  different time instants.

Since the field varies randomly with time, the underlying field at each of these  $m$  time

---

<sup>1</sup>This can be visualised as a network of sensors in mobile phones that ping off the nearest cell tower ( $S_0$  and  $S_1$ )

instants is different

5. For Group 0, let the sensor whose readings have the highest correlation with the readings of  $S_0$  be marked  $u_1$ , let the sensor with the highest correlation to the readings of sensor  $u_1$  be marked  $u_2$  and so on upto  $u_{n_0}$ . Then the final ordering of the sensors (in terms of sensor locations) in Group 0 is  $S_0 < u_1 < u_2 < \dots < u_{n_0}$
6. For Group 1, let the sensor whose readings have the highest correlation with the readings of  $S_1$  be marked  $v_{n_1}$ , let the sensor with the highest correlation to the readings of sensor  $v_{n_1}$  be marked  $v_{n_1-1}$  and so on upto  $v_1$ . Then the final ordering of the sensors (in terms of sensor locations) in Group 1 is  $v_1 < v_2 < \dots < v_{n_1} < S_1$
7. The final ordering of all the sensors deployed is  $S_0 < u_1 < u_2 < \dots < u_{n_0} < v_1 < v_2 < \dots < v_{n_1} < S_1$

Using sensors  $S_0$  and  $S_1$  as reference points allows us to give a direction to our ordering. Without this if we have a randomly chosen sensor  $w_1$ , then its nearest neighbor (in terms of maximum correlation)  $w_2$  could be located on either side of  $w_1$  since  $\cos(-x) = \cos(x)$ . Also dividing the sensors into two groups ensures that the maximum at  $d \rightarrow 1^-$  is not considered since if  $|t_1 - t_2| \rightarrow 1^-$  then the sensors at locations  $t_1$  and  $t_2$  will lie in different groups.



# Chapter 8

## Sampling with a General Continuous Distribution

In this chapter we consider the estimation of finite support polynomial fields from samples obtained at random locations according to a known (not necessarily uniform) distribution.

### 8.1 Sampling Model

Consider a polynomial,  $g(t) = a_0^{(1)} + a_1^{(1)}t + \dots + a_r^{(1)}t^r$  with support  $[0, 1]$  sampled at points  $T_1, T_2, \dots, T_n$  i.i.d  $f_T(t)$ . The distribution  $f_T(t)$ , and the degree  $r$  of the polynomial are assumed to be known but neither the sampling locations  $T_1, T_2, \dots, T_n$  nor their ordering is known. The field does not change with time. The situation is shown in Fig 8.1

The distribution of the samples  $g(T_1), g(T_2), \dots, g(T_n)$  is governed by the coefficients  $\vec{a}^1 = [a_0^{(1)}, \dots, a_r^{(1)}]$ . Hence the problem of estimating the polynomial coefficients reduces to the problem of estimating the parameters of the distribution of its samples. Since it is difficult to write a

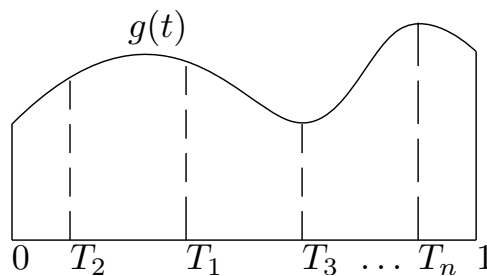


Figure 8.1:  $g(t)$  is a polynomial of known degree sampled at unknown points  $T_1, T_2, \dots, T_n$  with arbitrary (unknown) ordering

closed form expression for the distribution of the samples we use the sample moments to estimate the distribution parameters according to the well known Generalised Method of Moments [17].

Before we move on to the problem formulation it is important to highlight the reasons for changing the model from sampling of bandlimited fields to sampling of polynomial fields:

1. The Generalised Method of Moments involves computing the parameters by solving a minimization problem. As we will see in the following section the objective function will be a multivariate polynomial with the coefficients of  $g(t)$  as variables. Most of the existing literature on minimization of multivariate polynomials is applicable to real variables whereas the Fourier Series Coefficients of bandlimited fields are generally complex numbers.
2. This also improves the generalisation of our model since we know from the Weierstrass Approximation Theorem [18] that polynomials can be used to uniformly approximate any continuous function in a finite interval as closely as desired

It is to be noted that the results in Chapters 3-6 hold for any field that can be uniquely specified by its samples at a finite number of points, and this class of fields includes polynomials of finite degree so the results hold even in this case.

## 8.2 Optimisation Problem Formulation

Define  $\vec{a}^k = \vec{a}^1 * \vec{a}^1 * \dots * \vec{a}^1 = [a_0^{(k)}, \dots, a_{kr}^{(k)}]$  where  $*$  denotes convolution and it is applied  $k - 1$  times.

The  $k^{\text{th}}$  sample moment is estimated from the field samples as:

$$\hat{M}_k = \frac{1}{n} \sum_{i=1}^n (g(T_i))^k \quad (8.1)$$

$$= \frac{1}{n} \sum_{i=1}^n (a_0^{(1)} + a_1^{(1)}T_i + \dots + a_r^{(1)}T_i^r)^k \quad (8.2)$$

$$= a_0^{(k)} + a_1^{(k)}\hat{m}_1 + \dots + a_{kr}^{(k)}\hat{m}_{kr} \quad (8.3)$$

$$\hat{m}_l = \frac{1}{n} \sum_{i=1}^n (T_i)^l \quad (8.4)$$

$$(8.5)$$

Analogously the  $k^{\text{th}}$  population moment is given by:

$$M_k = \int_0^1 (g(t))^k f_T(t) dt \quad (8.6)$$

$$= a_0^{(k)} + a_1^{(k)} m_1 + \cdots + a_{kr}^{(k)} m_{kr} \quad (8.7)$$

$$m_l = \int_0^1 t^l f_T(t) dt \quad (8.8)$$

Define  $\Delta m_l = \hat{m}_l - m_l$ . Since  $\hat{M}_k$  is an unbiased estimator of  $M_k$  its variance is given by:

$$\mathbb{E}[(\hat{M}_l - M_l)^2] = \mathbb{E}\left[\left(\sum_{j=0}^{lr} a_j^{(l)} \Delta m_j\right)^2\right] \quad (8.9)$$

$$= \sum_{j=0}^{lr} \sum_{i=0}^{lr} a_j^{(l)} a_i^{(l)} \mathbb{E}[\Delta m_i \Delta m_j] \quad (8.10)$$

where:

$$\mathbb{E}[\Delta m_i \Delta m_j] = \mathbb{E}[(\hat{m}_i - m_i)(\hat{m}_j - m_j)] \quad (8.11)$$

$$= \mathbb{E}[\hat{m}_i \hat{m}_j] - m_i m_j \quad (8.12)$$

$$= \frac{1}{n^2} \sum_{l_1=1}^n \sum_{l_2=1}^n \mathbb{E}[X_{l_1}^i X_{l_2}^j] - m_i m_j \quad (8.13)$$

$$= \frac{1}{n^2} \sum_{l_1=1, l_1 \neq l_2}^n \sum_{l_2=1}^n \mathbb{E}[X_{l_1}^i] \mathbb{E}[X_{l_2}^j] + \frac{1}{n^2} \sum_{l=1}^n \mathbb{E}[X_l^{i+j}] - m_i m_j \quad (8.14)$$

$$= \frac{n(n-1)}{n^2} m_i m_j + \frac{m_{i+j}}{n} - m_i m_j \quad (8.15)$$

$$= \frac{m_{i+j} - m_i m_j}{n} \quad (8.16)$$

Thus:

$$\mathbb{E}[(\hat{M}_l - M_l)^2] = \frac{1}{n} \sum_{j=0}^{lr} \sum_{i=0}^{lr} a_j^{(l)} a_i^{(l)} (m_{i+j} - m_i m_j) \quad (8.17)$$

The term inside the summation depends only on the field being estimated and the distribution on the sampling locations which are both fixed. Thus the average squared error between the sample moments and the population moments decays linearly with sample size  $n$ . This leads us to expect that the vector of field coefficients,  $\vec{a}^{-1}$  can be estimated by solving the following optimization problem:

$$\vec{a}^1 = \operatorname{argmin}_{\vec{a}^1} \sum_{l=1}^L (\hat{M}_l - M_l)^2 \quad (8.18)$$

$$a_i^{(1)} \in [-1, 1], i = 0, 1, \dots, r \quad (8.19)$$

The polynomial coefficients are chosen to lie within a fixed (known) range to ensure that the polynomial (and its moments) are bounded.

The objective function is a multivariate polynomial with the elements  $\vec{a}^1$  as variables. It is clear that as  $n \rightarrow \infty$  the true polynomial coefficients will be a point of global minima. However such functions cannot be minimised using standard convex optimization methods such as the method of steepest descent due to the proliferation of local minima. Methods such as [19], [20] for the global optimisation of multivariate polynomials were tried but did not produce satisfactory results. One of the key difficulties was in choosing  $L$ , the number of moment estimators required to obtain a unique solution of 8.18 for all polynomials of a given degree. Since there are  $r + 1$  unknown coefficients for a polynomial of degree  $r$ , at least  $r + 1$  moments will be required to obtain a unique solution to 8.18 but even choosing  $L$  as  $r + 1$  or higher yielded convergence very far away from the true solution in several cases. This led us to suspect that there exists a class of signals for any distribution  $f_T(t)$  on the sample locations such that 8.18 does not have a unique solution. The following section deals with our analysis in this regard.

### 8.3 Non-Uniqueness of solutions

If  $f(t)$  is symmetric in  $[0, 1]$  i.e.  $f_T(t) = f_T(1 - t)$  then the fields  $g(t)$  and  $g(1 - t)$  will always have the same moments. To see this consider:

$$M_k = \int_0^1 (g(t))^k f_T(t) dt \quad (8.20)$$

$$= \int_0^1 (g(1 - t))^k f_T(1 - t) dt \quad (8.21)$$

$$= \int_0^1 (g(1 - t))^k f_T(t) dt \quad (8.22)$$

which follows from the properties of the definite integral. Thus 8.18 will never have a unique solution in this case.

For the case where the distribution  $f_T(t)$  is asymmetric in  $[0, 1]$  consider two strictly mono-

tonic (hence invertible) fields  $g_1(t)$  and  $g_2(t)$  lying in the same range. Assuming  $T \sim f_T(t)$  is a random variable, let  $Y_1 = g_1(t)$  and  $Y_2 = g_2(t)$  with distributions  $P(Y_1 \leq y) = P(g_1(T) \leq y)$ ,  $P(Y_2 \leq y) = P(g_2(T) \leq y)$  respectively. Assuming that  $g_1(t)$  is strictly increasing and  $g_2(t)$  is strictly decreasing we have ( $F_T(t)$  is the CDF of  $T$ ):

$$P(g_1(T) \leq y) = P(T \leq g_1^{-1}(y)) = F_T(g_1^{-1}(y)) \quad (8.23)$$

$$P(g_2(T) \leq y) = P(T \geq g_2^{-1}(y)) = 1 - F_T(g_2^{-1}(y)) \quad (8.24)$$

Let  $S_{12}$  denote the (common) range of  $g_1(t)$  and  $g_2(t)$  and let  $h(t) = g_1^{-1}(g_2(t))$ . Then the following condition must be satisfied for the two distributions to be identical:

$$F_T(g_1^{-1}(y)) + F_T(g_2^{-1}(y)) = 1, \forall y \in S_{12} \quad (8.25)$$

or:

$$F_T(h(t)) + F_T(t) = 1, \forall t \in [0, 1] \quad (8.26)$$

The family of solutions to this is:

$$h(t) = g_1^{-1}(g_2(t)) = F_T^{-1}(1 - F_T(t)) \quad (8.27)$$

i.e.  $g_1(t) = H(F_T(t))$  and  $g_2(t) = H(1 - F_T(t))$  where  $H(y)$  is any function that is invertible in  $y \in S_{12}$ . An obvious solution with  $S_{12} = [0, 1]$  is to put  $H(y)$  as the identity function. ( $F_T(t)$  being a CDF is strictly increasing). For other ranges one can use a linear  $H(y)$  or any other invertible transformation that maps  $F_T(t)$  and  $1 - F_T(t)$  to the desired range to satisfy the condition (8.25).

Thus the functions  $g_1(t)$  and  $g_2(t)$  of the form given above, have the same distribution (and hence the same moments) and thus for any distribution  $f_T(t)$  on the sampling locations if either of  $g_1(t)$  or  $g_2(t)$  is the field to be estimated then the optimization problem (8.18) does not have a unique solution.

Moreover consider two sequences of functions, one converging to  $g_1(t)$  and the other converging to  $g_2(t)$ . Since the random variables that represent samples drawn from these functions according to  $f_T(t)$  converge in distribution to the distributions of  $g_1(t)$  and  $g_2(t)$  respectively their moments converge to the corresponding moments (which are equal). This indicates that as

the functions move closer to  $g_1(t)$  and  $g_2(t)$  respectively more and more moments will be required to distinguish between their distributions in terms of (8.18) and thus it is not possible to put an upper bound on  $L$ , the number of moment estimators required to find a unique solution to (8.18) for all fields.

The above remarks are made for general fields but in our problem we are considering the estimation of polynomial fields. If  $f_T(t)$  is a polynomial, then so is  $F_T(t)$  and hence the above observations hold for polynomial fields of degree greater than or equal to the degree of  $F_T(t)$  since there exist fields belonging to this class, as discussed, that cannot be uniquely determined by solving (8.18). It might be possible to uniquely specify all fields of degree lower than that of  $F_T(t)$  by solving (8.18) but typically that would imply that to estimate high degree polynomial fields we require distributions that have an even higher degree and such distributions would be hard to realize.

In most of our attempts at obtaining a solution to the problem (8.18) computationally we assumed a polynomial distribution  $f_T(t)$  on the sampling locations. There are two reasons for this:

1. It is difficult to evaluate the integral  $m_l = \int_0^1 t^l f_T(t) dt$  for higher values of  $l$  if  $f_T(t)$  is not a polynomial
2. Any non-polynomial continuous  $f_T(t)$  can be uniformly approximated by a polynomial to the desired level of accuracy since it has finite support  $[0,1]$  (Weierstrass Approximation Theorem)

# Chapter 9

## Conclusions

The detection and estimation of fields using location-unaware sensors has been addressed in this work.

The first part of the work deals with the case when sensor locations were restricted to an equi-spaced discrete grid. Using an algorithm, which clusters distinct field values and records their types, field detection can be performed. It was shown that the detection error-probability decreases exponentially fast in the number of sensors deployed. The optimal distribution for maximizing the error-probability exponent was derived and was shown to perform better than other distributions for different choices of signal bandwidth. The algorithm was also extended to the case where there is measurement noise and the effects of changing the bandwidth and the sample size were studied for this case.

The second part of the work deals with the case where the sensors are deployed according to an arbitrary continuous distribution in the field's support. Additions to existing results were derived that strengthen the conclusion that if the sensors are uniformly distributed then the field cannot be uniquely inferred without order information on the sensor locations. A procedure for estimating order information when sensors are deployed uniformly was given which can aid existing works most of which deal with estimating fields from ordered samples. It was also shown that for an arbitrary continuous distribution on the sensor locations there exist a large class of fields which cannot be uniquely specified which emphasizes the necessity for restricting sensor locations to a discrete grid or knowing order information on sensor locations.

The detection of fields from samples at locations restricted to a discrete grid but with error in locations is a problem lying at the intersection of the above problems and hence is interesting to explore. This is left for a future work.

# Bibliography

- [1] Neal Patwari, Joshua N Ash, Spyros Kyperountas, Alfred O Hero III, Randolph L Moses, and Neiyer S Correal. Locating the nodes: cooperative localization in wireless sensor networks. *Signal Processing Magazine, IEEE*, 22(4):54–69, 2005.
- [2] Animesh Kumar. On bandlimited signal reconstruction from the distribution of unknown sampling locations. *Signal Processing, IEEE Transactions on*, 63(5):1259–1267, 2015.
- [3] Animesh Kumar. Bandlimited spatial field sampling with mobile sensors in the absence of location information. *arXiv preprint arXiv:1509.03966*, 2015.
- [4] Pina Marziliano and Martin Vetterli. Reconstruction of irregularly sampled discrete-time bandlimited signals with unknown sampling locations. *Signal Processing, IEEE Transactions on*, 48(12):3462–3471, 2000.
- [5] John Browning. Approximating signals from nonuniform continuous time samples at unknown locations. *Signal Processing, IEEE Transactions on*, 55(4):1549–1554, 2007.
- [6] Alessandro Nordio, Carla-Fabiana Chiasserini, and Emanuele Viterbo. Performance of linear field reconstruction techniques with noise and uncertain sensor locations. *Signal Processing, IEEE Transactions on*, 56(8):3535–3547, 2008.
- [7] Alan V Oppenheim, Ronald W Schafer, John R Buck, et al. *Discrete-time signal processing*, volume 2. Prentice hall Englewood Cliffs, NJ, 1989.
- [8] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [9] Narayan C Giri. *Introduction to probability and statistics*. M. Dekker New York, 1993.
- [10] Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.



- 
- [11] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [12] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
- [13] Alexandra Lauric and Sarah Frisken. Soft segmentation of ct brain data. *Tufts University*, 2007.
- [14] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
- [15] Arindam Banerjee, Chase Krumpelman, Joydeep Ghosh, Sugato Basu, and Raymond J Mooney. Model-based overlapping clustering. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 532–537. ACM, 2005.
- [16] Qiang Fu and Arindam Banerjee. Multiplicative mixture models for overlapping clustering. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 791–796. IEEE, 2008.
- [17] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.
- [18] Donald Estep. *Practical analysis in one variable*. Springer Science & Business Media, 2002.
- [19] Didier Henrion and Jean-Bernard Lasserre. Gloptipoly: Global optimization over polynomials with matlab and sedumi. *ACM Transactions on Mathematical Software (TOMS)*, 29(2):165–194, 2003.
- [20] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

# List of Publications

1. Mallick, Ankur, and Animesh Kumar. "Bandlimited field reconstruction from samples obtained on a discrete grid with unknown random locations." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

# *Acknowledgements*

I would like to express my sincere thanks to my supervisor Prof. Animesh Kumar whose guidance was invaluable at every stage of my work. His dynamic attitude and constant enthusiasm in coming up with new ideas has given direction to my work and has enabled me to come up with the results included in this dissertation. He has always been available for solving my doubts and his lucid explanations of the matter at hand in response to my queries have enabled me to obtain a greater understanding of the various topics that we have worked on. Thanks are also due to many of my peers in the Electrical Engineering Department of IIT Bombay as discussions with them often provided me with new insights for my research.

Signature: 

**Ankur Mallick**

110110013

Date: 24 June 2016